.

LECTURES ON STATISTICAL INFERENCES VIA CONVEX OPTIMIZATION

Anatoli Juditsky

Université Grenoble Alpes

Arkadi Nemirovski

Georgia Institute of Technology

Spring 2018

Preface

When speaking about links between Statistics and Optimization, what comes to mind first is the indispensable role played by optimization algorithms in the "computational toolbox" of Statistics (think about the numerical implementation of the fundamental Maximum Likelihood method). However, on a second thought, we should conclude that whatever high this role could be, the fact that it comes to our mind first primarily reflects the weaknesses of Optimization rather than its strengths; were optimization algorithms used in Statistics as efficient and as reliable as, say, Linear Algebra techniques used there, nobody would think about special links between Statistics and Optimization, same as nobody usually thinks about special links between Statistics and Linear Algebra. When computational, rather than methodological, issues are concerned, we start to think about links with Optimization, Linear Algebra, Numerical Analysis, etc., only when computational tools offered to us by these disciplines do not work well and need the attention of experts in these disciplines.

The goal of Lectures is to present another type of links between Optimization and Statistics, those which have nothing in common with algorithms and number-crunching. What we are speaking about, are the situations where Optimization theory (theory, not algorithms!) seems to be of methodological value in Statistics, acting as the source of statistical inferences with provably optimal, or nearly so, performance. In this context, we focus on utilizing Convex Programming theory, mainly due to its power, but also due to the desire to end up with inference routines reducing to solving convex optimization problems and thus implementable in a computationally efficient fashion. Thus, while we do not mention computational issues explicitly, we do remember that at the end of the day we need a number, and in this respect, intrinsically computationally friendly convex optimization models are the first choice.

The three topics we intend to consider are:

1. Sparsity-oriented Compressive Sensing. Here the role of Convex Optimization theory, by itself by far not negligible (it allows, e.g., to derive from "first principles" the necessary and sufficient conditions for the validity of ℓ_1 recovery) is relatively less important than in two other topics. Nevertheless, we believe that Compressive Sensing, due to its popularity and the fact that now it is one of the major "customers" of advanced convex optimization algorithms, is worthy of being considered.

2. Pairwise and Multiple Hypothesis Testing, including sequential tests, estimation of linear functionals, and some rudimentary design of experiments. This is the topic where, as of now, the approaches based on Convex Optimization theory were most successful.

3. Recovery of signals from noisy observations of their linear images.

The exposition does *not* require prior knowledge of Statistics and Optimization; as far as these disciplines are concerned, all necessary for us facts and concepts are incorporated into the text. The actual prerequisites are elementary Calculus, Probability, and Linear Algebra and (last but by far not least) general mathematical culture[1].

[1] As about optimization-related "knowledge prerequisites," all we need can be found (with proofs!) in Appendices. As a result, the Appendices are a bit longer than the main text, but why not to use paperless technology to make the Notes fully self-contained?

We are greatly indebted to Prof. H. Edwin Romeijn who initiated creating the Ph.D. course "Topics in Data Science," part of which are the lectures to follow. The second author gratefully acknowledges support from NSF Grant CCF-1523768 *Statistical Inference via Convex Optimization*; this research project is the source of basically all novel results presented in Lectures 2 – 4. Needless to say, responsibility for all drawbacks of Lectures is ours.

Anatoli Juditsky & Arkadi Nemirovski
April 11, 2018

Contents

Notational conventions

Vectors and matrices. By default, all vectors are column ones; to write them down, we use "Matlab notation:" $\begin{bmatrix}1\\2\\3\end{bmatrix}$ is written as $[1;2;3]$. More generally, for vectors/matrices $A,B,C,...,Z$ of the same "width" $[A;B;C;...;D]$ is the matrix obtained by writing B beneath of A, C beneath of B, and so on. For vectors/matrices $A,B,C,...,Z$ of the same "height," $[A,B,C,...,Z]$ denotes the matrix obtained by writing B to the right of A, C to the right of B, and so on. Examples: for what in the "normal" notation is written down as $A = \begin{bmatrix}1&2\\3&4\end{bmatrix}$, $B = \begin{bmatrix}5&6\end{bmatrix}$, $C = \begin{bmatrix}7\\8\end{bmatrix}$, we have

$$[A;B] = \begin{bmatrix}1&2\\3&4\\5&6\end{bmatrix} = [1,2;3,4;5,6],\ [A,C] = \begin{bmatrix}1&2&7\\3&4&8\end{bmatrix} = [1,2,7;3,4,8].$$

Blanks in matrices replace (blocks of) zero entries. For example,

$$\begin{bmatrix}1&&\\2&&\\3&4&5\end{bmatrix} = \begin{bmatrix}1&0&0\\2&0&0\\3&4&5\end{bmatrix}.$$

Block-diagonal matrix with diagonal blocks $A_1,A_2,...,A_k$ is denoted $\mathrm{Diag}\{A_1,A_2,...,A_k\}$. For example,

$$\mathrm{Diag}\{1,2,3\} = \begin{bmatrix}1&&\\&2&\\&&3\end{bmatrix},\ \mathrm{Diag}\{[1,2];[3;4]\} = \begin{bmatrix}1&2&\\&&3\\&&4\end{bmatrix}.$$

For an $m\times n$ matrix A, $\mathrm{diag}(A)$ is the diagonal of A – vector of dimension $\min[m,n]$ with entries A_{ii}, $1\leq i\leq \min[m,n]$.

Standard linear spaces in our course are $\mathbf{R}^n$ (the space of n-dimensional column vectors), $\mathbf{R}^{m\times n}$ (the space of $m\times n$ real matrices), and $\mathbf{S}^n$ (the space of $n\times n$ real symmetric matrices). All these linear spaces are equipped with the standard inner product:

$$\langle A,B\rangle = \sum_{i,j} A_{ij}B_{ij} = \mathrm{Tr}(AB^T) = \mathrm{Tr}(BA^T) = \mathrm{Tr}(A^TB) = \mathrm{Tr}(B^TA);$$

in the case when $A=a$ and $B=b$ are column vectors, this simplifies to $\langle a,b\rangle = a^Tb = b^Ta$, and when A,B are symmetric, there is no need to write B^T in $\mathrm{Tr}(AB^T)$.

Usually we denote vectors by lowercase, and matrices – by uppercase letters; sometimes, however, lowercase letters are used also for matrices.

Given a linear mapping $\mathcal{A}(x) : E_x \to E_y$, where E_x, E_y are standard linear spaces, one can define the *conjugate* mapping $\mathcal{A}^*(y) : E_y \to E_x$ via the identity

$$\langle \mathcal{A}(x), y\rangle = \langle x, \mathcal{A}^*(y)\rangle\ \forall (x\in E_x, y\in E_y).$$

One always has $(\mathcal{A}^*)^* = \mathcal{A}$. When $E_x = \mathbf{R}^n$, $E_y = \mathbf{R}^m$ and $\mathcal{A}(x) = Ax$, one has $\mathcal{A}^*(y) = A^T y$; when $E_x = \mathbf{R}^n$, $E_y = \mathbf{S}^m$, so that $\mathcal{A}(x) = \sum_{i=1}^n x_i A_i$, $A_i \in \mathbf{S}^m$, we have

$$\mathcal{A}^*(Y) = [\mathrm{Tr}(A_1 Y); ...; \mathrm{Tr}(A_n Y)].$$

$\mathbf{Z}^n$ is the set of n-dimensional integer vectors.

Norms. For $1 \leq p \leq \infty$ and for a vector $x = [x_1; ...; x_n] \in \mathbf{R}^n$, $\|x\|_p$ is the standard p-norm of x:

$$\|x\|_p = \left\{ \begin{array}{ll} (\sum_{i=1}^n |x_i|^p)^{1/p} & , 1 \leq p < \infty \\ \max_i |x_i| = \lim_{p' \to \infty} \|x\|_{p'} & , p = \infty \end{array} \right. ,$$

Notation for various norms of matrices is specified when used.

Standard cones. $\mathbf{R}_+$ is the nonnegative ray on the real axis, $\mathbf{R}^n_+$ stands for the *n-dimensional nonnegative orthant* – the cone comprised of all entrywise nonnegative vectors from $\mathbf{R}^n$, $\mathbf{S}^n_+$ stands for the *positive semidefinite cone in* $\mathbf{S}^n$ – the cone comprised of all positive semidefinite matrices from $\mathbf{S}^n$.

Miscellaneous. • For matrices A, B, relation $A \preceq B$, or, equivalently, $B \succeq A$, means that A, B are symmetric matrices of the same size such that $B - A$ is positive semidefinite; we write $A \succeq 0$ to express the fact that A is a symmetric positive semidefinite matrix. Strict version $A \succ B$ ($\Leftrightarrow B \prec A$) of $A \succeq B$ means that $A - B$ is positive definite (and, as above, A and B are symmetric matrices of the same size).
• Linear Matrix Inequality (LMI, a.k.a. *semidefinite constraint*) in variables x is the constraint on x stating that a symmetric matrix affinely depending on x is positive semidefinite. When $x \in \mathbf{R}^n$, LMI reads

$$a_0 + \sum_i x_i a_i \succeq 0 \qquad [a_i \in \mathbf{S}^m, 0 \leq i \leq n]$$

• $\mathcal{N}(\mu, \Theta)$ stands for the Gaussian distribution with mean μ and covariance matrix Θ.
• For a probability distribution P,

- $\xi \sim P$ means that ξ is a random variable with distribution P. Sometimes we express the same fact by writing $\xi \sim p(\cdot)$, where p is the density of P taken w.r.t. some reference measure (the latter always is fixed by the context);

- $\mathbf{E}_{\xi \sim P}\{f(\xi)\}$ is the expectation of $f(\xi)$, $\xi \sim P$; when P is clear from the context, this notation can be shortened to $\mathbf{E}_\xi\{f(\xi)\}$, or $\mathbf{E}_P\{f(\xi)\}$, or even $\mathbf{E}\{f(\xi)\}$. Similarly, $\mathrm{Prob}_{\xi \sim P}\{...\}$, $\mathrm{Prob}_\xi\{...\}$, $\mathrm{Prob}_P\{...\}$, $\mathrm{Prob}\{...\}$ denote the P-probability of the event specified inside the braces.

• $O(1)$'s stand for positive *absolute* constants – positive reals which we do not want or are too lazy to write down explicitly, like in $\sin(x) \leq O(1)|x|$.
• $\int_\Omega f(\xi)\Pi(d\xi)$ stands for the integral, taken w.r.t. measure Π over domain Ω, of function f.

About proofs

Lecture Notes are basically self-contained in terms of proofs of the statements to follow. Simple proofs usually are placed immediately after the corresponding statements; more technical proofs are transferred to dedicated sections titled “Proof of ...,” and this is where a reader should look for “missing” proofs.

Lecture 1

Sparse Recovery via ℓ_1 Minimization

In this lecture, we overview basic results of *Compressed Sensing* – a relatively new and extremely rapidly developing area in Signal Processing dealing with recovering signals (vectors x from some $\mathbf{R}^n$) from their noisy observations $Ax+\eta$ (A – given $m\times n$ *sensing matrix*, η is observation noise) in the case when the number of observations m is much smaller than the signal's dimension n, but is essentially larger than the "true" dimension – the number of nonzero entries – in the signals. This setup leads to extremely deep, elegant and highly innovative theory and possesses quite significant applied potential. It should be added that along with the plain sparsity (small number of nonzero entries), Compressed Sensing deals with other types of "low-dimensional structure" hidden in high-dimensional signals, most notably, with the case of *low rank matrix recovery*, when signal is a matrix, and sparse signals are matrices with low ranks, and the case of *block sparsity*, where signal is a block vector, and sparsity means that only small number of blocks are nonzero. In our presentation, we do *not* consider these extensions of the simplest sparsity paradigm.

1.1 Compressed Sensing: What it is about?

1.1.1 Signal Recovery Problem

One of the basic problems in Signal Processing is the problem of recovering a *signal* $x\in\mathbf{R}^n$ from noisy observations

$$y = Ax+\eta \tag{1.1.1}$$

of the affine image of the signal under a given *sensing mapping* $x\mapsto Ax:\mathbf{R}^n\to\mathbf{R}^m$; in (1.1.1), η is the *observation error*. Matrix A in (1.1.1) is called *sensing matrix*.

Recovery problem of outlined types arise in many applications, including, but *by far* not reducing to,

- *communications*, where x is the signal sent by transmitters, y is the signal recorded by receivers, A represents the communication channel (reflecting, e.g., dependencies of decays in signals' amplitude on the transmitter-receiver distances); η here typically is modeled as the standard (zero mean, unit covariance matrix) m-dimensional Gaussian noise[1];

[1] The "physical" noise usually indeed is Gaussian with zero mean, but its covariance matrix not necessarily is (proportional to) the unit matrix. Note, however, that a zero mean Gaussian noise η always can be represented as $Q\xi$ with standard Gaussian ξ; assuming Q nonsingular (which indeed is so when the covariance matrix of η is positive definite), we can rewrite (1.1.1) equivalently as

$$Q^{-1}y=[Q^{-1}A]x+\xi$$

and treat $Q^{-1}y$ and $Q^{-1}A$ as our new observation and new sensing matrix; new observation noise ξ is indeed standard. Thus, in the case of Gaussian zero mean observation noise, to assume the noise standard Gaussian is the same as to assume that its covariance matrix is known.

- *image reconstruction*, where the signal x is an image – a 2D array in the usual photography, or a 3D array in Tomography, and y is data acquired by the imaging device. Here η in many cases (although not always) can again be modeled as the standard Gaussian noise;

- *linear regression* arising in a wide range of applications. In linear regression, one is given m pairs "input $a^i \in \mathbf{R}^n$" to a "black box" — output $y_i \in \mathbf{R}$ of the black box." Sometimes we have reasons to believe that the output is a corrupted by noise version of the "existing in the nature," but unobservable, ideal output" $y_i^* = x^T a^i$ which is just a linear function of the input (this is called "linear regression model," with inputs a^i called "regressors"). Our goal is to convert actual observations (a^i, y_i), $1 \leq i \leq m$, into estimates of the *unknown* "true" vector of parameters x. Denoting by A the matrix with the rows $[a^i]^T$ and assembling individual observations y_i into a single observation $y = [y_1; ...; y_m] \in \mathbf{R}^m$, we arrive at the problem of recovering vector x from noisy observations of Ax. Here again the most popular model for η is the standard Gaussian noise.

1.1.2 Signal Recovery: parametric and non-parametric cases

Recovering signal x from observation y would be easy if there were no observation noise ($\eta = 0$) and the rank of matrix A were equal to the dimension n of signals. In this case, which can take place only when $m \geq n$ ("more observations that unknown parameters"), and is typical in this range of sizes m, n, the desired x would be the unique solution to the system of linear equation, and to find x would be a simple problem of Linear Algebra. Aside of this trivial "enough observations, no noise" case, people over the years looked at the following two versions of the recovery problem:

Parametric case: $m \gg n$, η is nontrivial noise with zero mean, say, standard Gaussian one. This is the classical statistical setup considered in thousands of papers, with the emphasis on how to use the numerous observations we have at our disposal in order to suppress in the recovery, to the extent possible, the influence of observation noise.

Nonparametric case: $m \ll n$. Literally treated, this case seems to be senseless: when the number of observations is less that the number of unknown parameters, even in the no-noise case we arrive at the necessity to solve an undetermined (less equations than unknowns) system of linear equations. Linear Algebra says that if solvable, the system has infinitely many solutions; moreover, the solution set (an affine subspace of positive dimension) is unbounded, meaning that the solutions are in no sense close to each other. Typical way to make the case of $m \ll n$ meaningful is to add to the observations (1.1.1) some a priori information on the signal. In traditional Nonparametric Statistics this additional information is summarized in a given to us in advance *bounded convex set* $X \subset \mathbf{R}^n$ known to contain the true signal x. This set usually is such that *every signal* $x \in X$ *can be approximated by a linear combination of* $s = 1, 2, ..., n$ *of vectors from properly selected and known to us in advance orthonormal basis* ("dictionary" in the slang of signal processing) *within accuracy* $\delta(s)$, where $\delta(s)$ is a known in advance function approaching 0 as $s \to \infty$. In this situation, with appropriate A (e.g., just the unit matrix, as in denoising problem), we can select somehow $s \ll m$ and try to recover x *as if* it were a vector from the linear span E_s of the first s vectors of the outlined basis. In the "ideal case" $x \in E_s$, recovering x in fact reduces to the case where the dimension of the signal is $s \ll m$ rather than $n \gg m$, and we arrive at the well-studied situation of recovering signal of low (as compared to the number of observations) dimension. In the "realistic case" of x $\delta(s)$-close to E_s, deviation of x from E_s results in additional component in the recovery error ("bias"); a typical result of traditional Nonparametric Statistics quantifies the resulting error and minimizes it in s. Of course, this outline of traditional statistical approach to "nonparametric" (with $n \gg m$) recovery problems is extremely sketchy, but it captures the most important in our context fact: with the traditional approach to nonparametric signal recovery, one assumes that after representing the signals by vectors of their coefficients in properly selected orthonormal basis, the n-dimensional signal to be recovered can be well approximated by s-sparse (at most s nonzero entries) signal, with $s \ll n$, *and this sparse approximation can be obtained by zeroing out all but the first s entries in the signal vector.*

The just formulated assumption indeed takes place for signals obtained by discretization of *smooth* uni- and multivariate functions, and this class of signals for several decades was the main, if not the only, focus of Nonparametric Statistics.

To the best of our knowledge, developments in the traditional Nonparametric Statistics had nearly nothing to do with Convex Optimization.

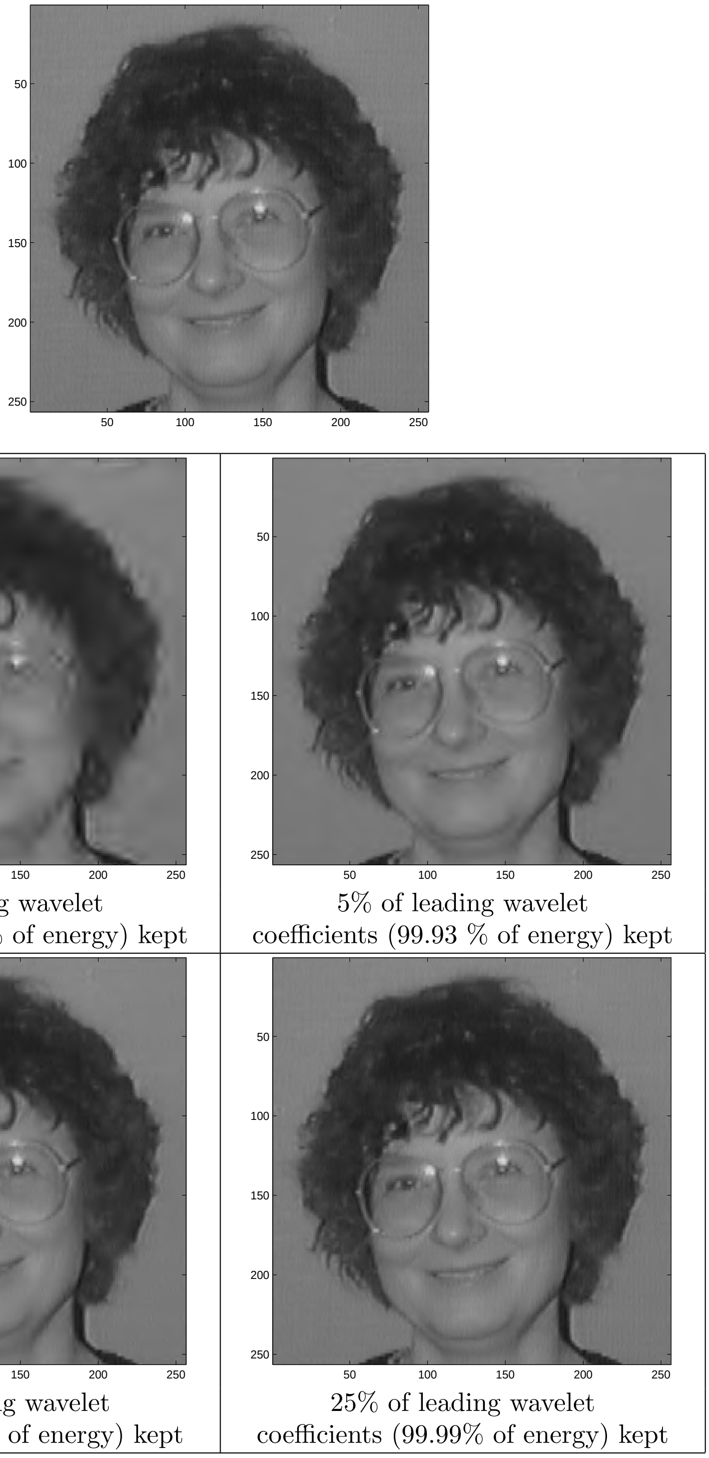
Compressed Sensing. The situation changed dramatically around Year 2000 as a consequence of the breakthroughs due to D. Donoho, T. Tao, J. Romberg, E. Candes, J. Fuchs and several other researchers; as a result of these breakthroughs, an extremely popular and completely novel area of research, called *Compressed Sensing*, emerged.

In the Compressed Sensing (CS) setup of the Signal Recovery problem, same as in the traditional Nonparametric Statistics, is assumed that after passing to an appropriate basis, the signal to be recovered is s-sparse (has $\leq s$ nonzero entries), or is well approximated by s-sparse signal. The difference with the traditional approach is that now we assume *nothing* on the location of the nonzero entries. Thus, the a priori information on the signal x both in the traditional and in the CS settings is summarized in a set X known to contain the signal x we want to recover. The difference is, that in the traditional setting, X is a bounded convex and "nice" (well approximated by its low-dimensional cross-sections) set, while in CS this set is, computationally speaking, a "monster:" already in the simplest case of recovering *exactly s-sparse* signals, X is the union of all s-dimensional coordinate planes, which is a heavily combinatorial entity.

> Note that in many applications we indeed can be sure that the true vector of parameters θ^* is sparse. Consider, e.g., the following story about signal detection. *There are n locations where signal transmitters could be placed, and m locations with the receivers. The contribution of a signal of unit magnitude originating in location j to the signal measured by receiver i is a known quantity A_{ij}, and signals originating in different locations merely sum up in the receivers; thus, if x is the n-dimensional vector with entries x_j representing the magnitudes of signals transmitted in locations $j = 1, 2, ..., n$, then the m-dimensional vector y of measurements of the m receivers is $y = Ax + \eta$, where η is the observation noise. Given y, we intend to recover x.*
>
> Now, if the receivers are hydrophones registering noises emitted by submarines in certain part of Atlantic, tentative positions of submarines being discretized with resolution 500 m, the dimension of the vector x (the number of points in the discretization grid) will be in the range of tens of thousands, if not tens of millions. At the same time, the total number of submarines (i.e., nonzero entries in x) can be safely upper-bounded by 50, if not by 20.

In order to see sparsity on our everyday life, look at the 256×256 image on the top of Figure 1.1. The image can be thought of as a $256^2 = 65536$-dimensional vector comprised of pixels' intensities in gray scale, and there is no much sparsity in this vector. However, when representing the image in the *wavelet basis*, whatever it means, we get a "nearly sparse" vector of wavelet coefficients (this is true for typical "non-pathological" images). On the bottom of Figure 1.1 we see what happens when we zero out all but a percentage of the largest in magnitude wavelet coefficients and replace the true image by its sparse, in the wavelet basis, approximations.

Our visual illustration along with numerous similar examples show the "everyday presence" of sparsity and the possibility to utilize it when compressing signals. The difficulty, however, is that simple compression – compute the coefficients of the signal in an appropriate basis and then keep,



1% of leading wavelet coefficients (99.70 % of energy) kept

5% of leading wavelet coefficients (99.93 % of energy) kept

10% of leading wavelet coefficients (99.96% of energy) kept

25% of leading wavelet coefficients (99.99% of energy) kept

Figure 1.1: Top: true 256×256 image; bottom: sparse in the wavelet basis approximations of the image. Wavelet basis is orthonormal, and a natural way to quantify near-sparsity of a signal is to look at the fraction of total energy (sum of squares of wavelet coefficients) stored in the leading coefficients; these are the "energy data" presented on the figure.

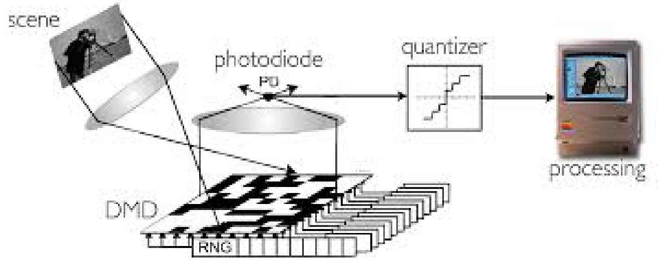


Figure 1.2: Singe-pixel camera

say, 10% of the largest in magnitude coefficients – requires to start with digitalizing the signal – representing it as an array of all its coefficients in some orthonormal basis. These coefficients are inner products of the signal with vectors of the basis; for a "physical" signal, like speech or image, these inner products are computed by analogous devices, with subsequent discretization of the results. After the measurements are discretized, processing the signal (denoising, compression, storing, etc., etc.) can be fully computerized. The major potential (to some extent, already actual) advantage of Compressed Sensing is in the possibility to reduce the "analogous effort" in the outlined process: instead of computing analogously n linear forms of n-dimensional signal x (its coefficients in a basis), we use analogous device to compute $m \ll n$ other linear forms of the signal and then use signal's sparsity in a known to us basis in order to recover the signal reasonably well from these m observations.

In our "picture illustration" this technology would work (in fact, works - it is called "single pixel camera," see Figure 1.2) as follows: in reality, the digital 256×256 image on the top of Figure 1.1 was obtained by analogous device – a digital camera which gets on input analogous signal (light of varying along the field of view intensity caught by camera's lenses) and discretizes lights's intensity in every pixel to get the digitalized image. We then can compute the wavelet coefficients of the digitalized image, compress its representation by keeping, say, just 10% of leading coefficients, etc., etc., but "the damage is already done" – we have already spent our analogous resources to get the entire digitalized image. The technology utilizing Compressed Sensing would work as follows: instead of measuring and discretizing light intensity in every one of the 65,536 pixels, we compute analogously the integral, taken over the field of view, of the product of light intensity and an analogously generated "mask," and do it for, say, 20,000 different masks, thus obtaining measurements of 20,000 linear forms of our 65,536-dimensional signal. Next we utilize, via the Compressed Sensing machinery, signal's sparsity in the wavelet basis in order to recover the signal from these 20,000 measurements. With this approach, we reduce the "analogous component" of signal processing effort, at the price of increasing the "computerized component" of the effort (instead of ready-to-use digitalized image directly given by 65,536 analogous measurements, we need to recover the image by applying computationally not so trivial decoding algorithms to our 20,000 "indirect" measurements). When taking pictures by your camera or ipad, the game is not worth the candle – analogous component of taking usual pictures is cheap enough, and decreasing it at the price of nontrivial decoding of the digitalized measurements would be counter-productive. There are, however, important applications where the advantages stemming from reduced "analogous effort" overweight significantly the drawbacks caused by the necessity to use nontrivial computerized decoding.

1.1.3 Compressed Sensing via ℓ_1 minimization: Motivation

1.1.3.1 Preliminaries

In principle there is nothing surprising in the fact that under reasonable assumption on $m \times n$

sensing matrix A we may hope to recover from noisy observations of Ax an s-sparse, with $s \ll m$, signal x. Indeed, assume for the sake of simplicity that there are no observation errors, and let $\mathrm{Col}_j[A]$ be j-th column in A. If we knew the locations $j_1 < j_2 < ... < j_s$ of the nonzero entries in x, identifying x could be reduced to solving system of linear equations $\sum_{\ell=1}^{s} x_{i_\ell} \mathrm{Col}_{j_\ell}[A] = y$ with m equations and $s \ll m$ unknowns; assuming every s columns in A linearly independent (a quite unrestrictive assumption on a matrix with $m \geq s$ rows), the solution to the above system is unique, and is exactly the signal we are looking for. Of course, the assumption that we know the locations of nonzeros in x makes the recovery problem completely trivial. However, it suggests the following course of actions: given noiseless observation $y = Ax$ of an s-sparse signal x, let us solve the combinatorial optimization problem

$$\min_z \{\|z\|_0 : Az = y\}, \tag{1.1.2}$$

where $\|z\|_0$ is the number of nonzero entries in z. Clearly, the problem has a solution with the value of the objective at most s. Moreover, it is immediately seen (verify it!) that if every $2s$ columns in A are linearly independent (which again is a very unrestrictive assumption on the matrix A provided that $m \geq 2s$), then the true signal x is the unique optimal solution to (1.1.2).

> What was said so far can be extended to the case of noisy observations and "nearly s-sparse" signals x. For example, assuming that the observation error is "uncertain-but-bounded," specifically some known norm $\|\cdot\|$ of this error does not exceed a given $\epsilon > 0$, and that the true signal is *exactly* s-sparse (think how to relax this to "near s-sparsity"), we could solve the combinatorial optimization problem
>
> $$\min_z \{\|z\|_0 : \|Az - y\| \leq \epsilon\}. \tag{1.1.3}$$
>
> Assuming that every $m \times 2s$ submatrix $\bar{A}$ of A is not just with linearly independent columns (i.e., with trivial kernel), but is reasonably well conditioned:
>
> $$\|\bar{A}w\| \geq C^{-1}\|w\|_2$$
>
> for all $(2s)$-dimensional vectors w, with some constant C, it is immediately seen that the true signal x underlying observation and the optimal solution $\widehat{x}$ of (1.1.3) are close to each other within accuracy of order of ϵ: $\|x - \widehat{x}\|_2 \leq 2C\epsilon$; it is easily seen that the resulting error bound is basically as good as it could be.

We see that the difficulties with recovering sparse signals stem not from the lack of information, they are of purely computational nature: (1.1.2) is a disastrously difficult combinatorial problem, and the only known way to process it is by "brute force" search through all guesses on where the nonzeros in x are located – by inspecting first the only option that there are no nonzeros in x at all, then by inspecting n options that there is only one nonzero, for every one of n locations of this nonzero, then $n(n-1)/2$ options that there are exactly two nonzeros, etc., etc. until the current option will result in a solvable system of linear equations $Az = y$ in variables z with entries restricted to vanish outside the locations prescribed by the option under consideration. Running time of this "brute force" search, beyond the range of small values of s and n (by far too small to be of any applied interest) is by many orders of magnitude larger than what we can afford to ourselves in reality[2].

A partial remedy is as follows. Well, if we do not know how to minimize under linear constraints, as in (1.1.2), the "bad" objective $\|z\|_0$, let us "approximate" this objective with one which we do

[2] When $s = 5$ and $n = 100$, a sharp upper bound on the number of linear systems we should process before termination in the "brute force" algorithm is $\approx 7.53\text{e}7$ — much, but perhaps doable. When $n = 200$ and $s = 20$, the number of systems to be processed jumps to $\approx 1.61\text{e}27$, which is by many orders of magnitude beyond our "computational grasp"; we would be unable to carry out that many computations even if the fate of the mankind were at stake. And from the perspective of Compressed Sensing, $n = 200$ still is a completely toy size, by 3-4 orders of magnitude less than we would like to handle.

know how to minimize. The true objective is separable: $\|z\| = \sum_{i=1}^n \xi(z_j)$, where $\xi(s)$ is the function on the axis equal to 0 at the origin and equal to 1 otherwise. As a matter of fact, the separable functions which we do know how to minimize under linear constraints are sums of *convex* functions of $z_1, ..., z_n$. The most natural candidate to the role of *convex* approximation of $\xi(s)$ is $|s|$; with this approximation, (1.1.2) converts into the ℓ_1 *minimization problem*

$$\min_z \left\{ \|z\|_1 := \sum_{i=1}^n |z_j| : Az = y \right\}, \tag{1.1.4}$$

and (1.1.3) becomes the convex optimization problem

$$\min_z \left\{ \|z\|_1 := \sum_{i=1}^n |z_j| : \|Az - y\| \leq \epsilon \right\}. \tag{1.1.5}$$

Both problems are efficiently solvable, which is nice; the question, however, is how relevant these problems are in our context – whether it is true that they do recover the "true" s-sparse signals in the noiseless case, or "nearly recover" these signals when the observation error is small. Since we want to be able to handle *whatever* s-sparse signals, the validity of ℓ_1 recovery – it ability to recover well *every* s-sparse signal – depends solely on the sensing matrix A. Our current goal is to understand what are "good" in this respect sensing matrices.

1.2 Validity of sparse signal recovery via ℓ_1 minimization

What follows is based on the standard basic results of Compressed Sensing theory originating from [36, 37, 38, 39, 40, 41, 56, 57, 58, 65, 66] and augmented by the results of [93][3].

1.2.1 Validity of ℓ_1 minimization in the noiseless case

The minimal requirement on sensing matrix A which makes ℓ_1 minimization valid is to guarantee the correct recovery of *exactly* s-sparse signals in the *noiseless* case, and we start with investigating this property.

1.2.1.1 Notational convention

From now on, for a vector $x \in \mathbf{R}^n$

- $I_x = \{j : x_j \neq 0\}$ stands for the *support* of x; we also set
$$I_x^+ = \{j : x_j > 0\},\ I_x^- = \{j : x_j < 0\} \qquad [\Rightarrow I_x = I_x^+ \cup I_x^-]$$
- for a subset I of the index set $\{1, ..., n\}$, x_I stands for the vector obtained from x by zeroing out entries with indexes *not* in I, and I^o for the complement of I:
$$I^o = \{i \in \{1, ..., n\} : i \notin I\};$$
- for $s \leq n$, x^s stands for the vector obtained from x by zeroing our all but the s largest in magnitude entries[4] Note that x^s is the best s-sparse approximation of x in any one of the ℓ_p norms, $1 \leq p \leq \infty$;

[3] in fact, in the latter source, an extension of the sparsity, the so called block sparsity, is considered; in what follows, we restrict the results of [93] to the case of plain sparsity.

[4] note that in general x^s is not uniquely defined by x and s, since the s-th largest among the magnitudes of entries in x can be achieved at several entries. In our context, it does not matter how the ties of this type are resolved; for the sake of definiteness, we can assume that when ordering the entries in x according to their magnitudes, from the largest to the smallest, entries of equal magnitude are ordered in the order of their indexes.

- for $s \leq n$ and $p \in [1, \infty]$, we set
$$\|x\|_{s,p} = \|x^s\|_p;$$
note that $\|\cdot\|_{s,p}$ is a norm (why?).

1.2.1.2 s-Goodness

Definition of s-goodness. Let us say that an $m \times n$ sensing matrix A is *s-good*, if whenever the true signal x underlying *noiseless* observations is s-sparse, this signal will be recovered *exactly* by ℓ_1 minimization. In other words, A is s-good, if whenever in y in (1.1.4) is of the form $y = Ax$ with s-sparse x, x is the unique optimal solution to (1.1.4).

Nullspace property. There is a simply-looking *necessary and sufficient* condition for a sensing matrix A to be s-good – the *nullspace property.* After this property is guessed, it is easy to see that it indeed is necessary and sufficient for s-goodness; we, however, prefer to *derive* this condition from the "first principles," which can be easily done via Convex Optimization; thus, in the case in question, same as in many other cases, there is no necessity to be smart to arrive at the truth via "'lucky guess," it suffices to be knowledgeable and use the standard tools.

Let us start with necessary and sufficient condition for A to be such that whenever x is s-sparse, x is an optimal solution (perhaps, not the unique one) of the optimization problem
$$\min_z \left\{\|z\|_1 : Az = Ax\right\}, \tag{$*$}$$
let us call the latter property of A *weak s-goodness*. Our first observation is as follows:

Proposition 1.2.1 *A is weakly s-good if and only if the following condition holds true: whenever I is a subset of $\{1, ..., n\}$ of cardinality $\leq s$, we have*
$$\forall w \in \operatorname{Ker} A : \|w_I\|_1 \leq \|w_{\bar{I}}\|_1 \tag{1.2.1}$$

Proof is immediate. In one direction: Assume A is weakly s-good, and let us verify (1.2.1). Let I be an s-element subset of $\{1, ..., n\}$, and x be s-sparse vector with support I. Since A is weakly s-good, x is an optimal solution to $(*)$. Rewriting the latter problem in the form of LP, that is, as
$$\min_{z,t}\{\sum_j t_j : t_j + z_j \geq 0, t_j - z_j \geq 0, Az = Ax\},$$
and invoking LP optimality conditions, the necessary and sufficient condition for $z = x$ to be the z-component of an optimal solution is the existence of λ_j^+, λ_j^-, $\mu \in \mathbf{R}^m$ (Lagrange multipliers for the constraints $t_j - z_j \geq 0$, $t_j + z_j \geq 0$, and $Az = Ax$, respectively) such that
$$\begin{array}{lrcl}
(a) & \lambda_j^+ + \lambda_j^- & = & 1\,\forall j,\\
(b) & \lambda^+ - \lambda^- + A^T\mu & = & 0,\\
(c) & \lambda_j^+(|x_j| - x_j) & = & 0\,\forall j,\\
(d) & \lambda_j^-(|x_j| + x_j) & = & 0\,\forall j,\\
(e) & \lambda_j^+ & \geq & 0\,\forall j,\\
(f) & \lambda_j^- & \geq & 0\,\forall j.
\end{array} \tag{1.2.2}$$

From (c, d), we have $\lambda_j^+ = 1, \lambda_j^- = 0$ for $j \in I_x^+$ and $\lambda_j^+ = 0, \lambda_j^- = 1$ for $j \in I_x^-$. From (a) and nonnegativity of $\lambda_j^\pm$ it follows that for $j \notin I_x$ we should have $-1 \leq \lambda_j^+ - \lambda_j^- \leq 1$. With this in mind, the above optimality conditions admit eliminating λ's and reduce to the following conclusion:

(!) *x is an optimal solution to $(*)$ if and only if there exists vector $\mu \in \mathbf{R}^m$ such that j-th entry of $A^T\mu$ is -1, if $x_j > 0$, $+1$, if $x_j < 0$, and a real from $[-1, 1]$, if $x_j = 0$.*

Now let $w \in \operatorname{Ker} A$ be a vector with the same signs of entries w_i, $i \in I$, as these of the entries in x. Then

$$0 = \mu^T A w = [A^T\mu]^T w = \sum_j [A^T\mu]_j w_j \Rightarrow \sum_{j\in I_x} |w_j| = \sum_{j\in I_x} [A^T\mu]_j w_j = -\sum_{j\not\in I_x} [A^T\mu]_j w_j \leq \sum_{j\not\in I_x} |w_j|$$

(we have used the fact that $[A^T\mu]_j = \operatorname{sign} x_j = \operatorname{sign} w_j$ for $j \in I_x$ and $|[A^T\mu]_j| \leq 1$ for all j). Since I can be an arbitrary s-element subset of $\{1, ..., n\}$ and the pattern of signs of an s-sparse vector x supported on I can be arbitrary, (1.2.1) holds true.

Now let us assume that (1.2.1) holds true, and let us prove that A is weakly s-sparse. Assume the opposite; then for some s-sparse x, x is not an optimal solution to $(*)$, meaning that system (1.2.2) of linear constraints in variables $\lambda^\pm$, μ has no solution. Applying Theorem on Alternative (Theorem B.2.6), we can assign the constraints (a) – (f) in (1.2.2) with respective vectors of weights $w_a, ..., w_e, w_f$, with the weights w_e, w_f of inequality constraints (e), (f) being nonnegative, such that multiplying the constraints by the weights and summing up the results, we get as a consequence of (1.2.2) a contradictory inequality – one with no solutions at all. This contradictory consequence of (1.2.2) is the linear inequality in variables $\lambda^\pm$, μ:

$$[w_a + w_b + G_+ w_c + w_e]^T \lambda^+ + [w_a - w_b + G_- w_d + w_f]^T \lambda^- + w_b^T A^T \mu \geq \sum_j (w_a)_j, \qquad (**)$$

where G_+, G_- are diagonal matrices with j-th diagonal entry equal to $|x_j| - x_j$ (G_+) and $|x_j| + x_j$ (G_-). Thus, we can find $w_a, ..., w_f$ with nonnegative w_e and w_f such that

$$w_a + w_b + G_+ w_c + w_e = 0,\ w_a - w_b + G_- w_d + w_f = 0,\ A w_b = 0,\ \sum_j (w_a)_j > 0.$$

or, equivalently, there exist w_a, w_b, w_c, w_d such that

$$\begin{array}{ll}
(p) & w_a + w_b + \underbrace{G_+ w_c}_{g} \leq 0,\\
(q) & w_a - w_b + \underbrace{G_- w_d}_{h} \leq 0,\\
(r) & A w_b = 0,\\
(s) & \sum_j (w_a)_j > 0.
\end{array}$$

Now note that when $j \in I_x^+$, we have $g_j = 0$ and thus (p) says that $|[w_b]_j| \geq [w_a]_j$, and when $j \in I_x^-$, we have $h_j = 0$ and thus (q) says that $|[w_b]_j| \geq [w_a]_j$. And when $j \not\in I_x := I_x^+ \cup I_x^-$, (p) and (q) say that $[w_a]_j \leq -|[w_b]_j|$. With this in mind, (s) implies that $-\sum_{j\not\in I_x} |[w_b]_j| + \sum_{j\in I_x} |[w_b]_j| \geq \sum_j [w_a]_j > 0$. Thus, assuming that A is not weakly s-good, we have found a set I_x of indexes of cardinality $\leq s$ and a vector $w_b \in \operatorname{Ker} A$ (see (r)) such that $\sum_{j\in I_x} |[w_b]_j| > \sum_{j\not\in I_x} |[w_b]_j|$, contradicting the condition (1.2.1). □

1.2.1.3 Nullspace property

We have established necessary and sufficient condition for A to be weakly s-good; it states that $\|w_I\|_1$ should be $\leq \|w_{I^o}\|_1$ for all $w \in \operatorname{Ker} A$ and all I of cardinality s. It may happen that this inequality holds true as equality, for some nonzero $w \in \operatorname{Ker} A$:

$$\exists (w \in \operatorname{Ker} A \backslash \{0\}, I, \operatorname{Card}(I) \leq s) : \|w_I\|_1 = \|w_{I^o}\|_1.$$

In this case matrix A clearly is not s-good, since the s-sparse signal $x = w_I$ is *not* the unique optimal solution to $(*)$ – the vector $-w_{I^o}$ is a different feasible solution to the same problem and with the same value of the objective. We conclude that for A to be s-good, a necessary condition is for the inequality in (1.2.1) to be strict whenever $w \in \operatorname{Ker} A$ is nonzero. By the standard compactness arguments, the latter condition means the existence of $\gamma \in (0,1)$ such that

$$\forall (w \in \operatorname{Ker} A, I, \operatorname{Card}(I) \leq s) : \|w_I\|_1 \leq \gamma \|w_{I^o}\|_1,$$

or, which is the same, existence of $\kappa \in (0, 1/2)$ such that

$$\forall (w \in \operatorname{Ker} A, I, \operatorname{Card}(I) \leq s) : \|w_I\|_1 \leq \kappa \|w\|_1.$$

Finally, the supremum of $\|w_I\|_1$ over I of cardinality s is what we have defined the norm $\|w\|_{1,s}$ (the sum of s largest magnitudes of entries) of w, so that the condition we are processing finally can be formulated as

$$\exists \kappa \in (0, 1/2) : \|w\|_{1,s} \leq \kappa \|w\|_1 \ \forall w \in \operatorname{Ker} A. \tag{1.2.3}$$

The resulting *nullspace condition* in fact is necessary *and sufficient* for A to be s-good:

Proposition 1.2.2 *Condition (1.2.3) is necessary and sufficient for A to be s-good.*

Proof. We have already seen that the nullspace condition is necessary for s-goodness. To verify sufficiency, let A satisfy nullspace condition, and let us prove that A is s-good. Indeed, let x be an s-sparse vector. By Proposition 1.2.1, A is weakly s-good, so that x is an optimal solution to $(*)$, and the only thing we need to prove is that if y is another optimal solution to $(*)$, then $y = x$. Assuming y optimal for $(*)$, let I be the support of x. Setting $w = y - x$, we have

$$Aw = 0 \ \& \ \underbrace{\|y_I - x\|_1}_{\|w_I\|_1} \leq \kappa \left[\|y_I - x\|_1 + \|y_{I^o} - x_{I^o}\|_1\right] = \kappa[\|w_I\|_1 + \|y_{I^o}\|_1],$$

whence

$$(1 - \kappa)\|w_I\|_1 \leq \kappa \|y_{I^o}\|_1 = \kappa(\|y\|_1 - \|y_I\|_1).$$

Since $\|w_I\|_1 = \|y_I - x\|_1 \geq \|x\|_1 - \|y_I\|_1$, we arrive at

$$(1 - \kappa)(\|x\|_1 - \|y_I\|_1) \leq \kappa(\|y\|_1 - \|y_I\|_1),$$

which, due to $\|x\|_1 = \|y\|_1$ (since x and y are optimal solutions of $(*)$) and $\kappa < 1/2$, boils down to

$$[\|y\|_1 =]\|x\|_1 \leq \|y_I\|_1,$$

implying, due to $\|x\|_1 = \|y\|_1$, that $y_I = y$, that is, y is supported on the support I of x. In other words, $w = y - x$ is supported on s-element set, and since $Aw = 0$, nullspace property implies that $y = x$. □

1.2.2 Imperfect ℓ_1 minimization

We have found a necessary and sufficient condition for ℓ_1 minimization to recover *exactly s-sparse signals* in the *noiseless* case. “In reality,” both these assumptions typically are violated: instead of s-sparse signals, we should speak about “nearly s-sparse ones,” quantifying the deviation from sparsity by the distance from the signal x underlying observations to its best s-sparse approximation x^s. Similarly, we should allow for nonzero observation noise. With noisy observations and/or imperfect sparsity, we cannot hope to recover signal exactly; all we may hope for, is to recover it with some error depending on the level of observation noise and “deviation from s-sparsity” and tending to zero as these level and deviation tend to 0. We are about to quantify the Nullspace property to allow for instructive “error analysis.”

1.2.2.1 Contrast matrices and quantifications of Nullspace property

By itself, Nullspace property says something about the signals from the kernel of the sensing matrix. We can reformulate it equivalently to say something important about *all* signals. Namely, observe that given sparsity s and $\kappa \in (0, 1/2)$, the Nullspace property

$$\|w\|_{s,1} \leq \kappa \|w\|_1 \ \forall w \in \operatorname{Ker} A \tag{1.2.4}$$

is satisfied if and only if for a properly selected constant C one has

$$\|w\|_{s,1} \leq C\|Aw\|_2 + \kappa\|w\|_1 \, \forall w. \tag{1.2.5}$$

Indeed, (1.2.5) clearly implies (1.2.4); to get the inverse implication, note that for every h orthogonal to $\operatorname{Ker} A$ it holds

$$\|Ah\|_2 \geq \sigma\|h\|_2,$$

where $\sigma > 0$ is the minimal positive singular value of A. Now, given $w \in \mathbf{R}^n$, we can decompose w into the sum of $\bar{w} \in \operatorname{Ker} A$ and $h \in (\operatorname{Ker} A)^\perp$, so that

$$\begin{array}{l}\|w\|_{s,1} \leq \|\bar{w}\|_{s,1} + \|h\|_{s,1} \leq \kappa\|\bar{w}\|_1 + \sqrt{s}\|h\|_{s,2} \leq \kappa[\|w\|_1 + \|h\|_1] + \sqrt{s}\|h\|_2 \\ \leq \kappa\|w\|_1 + [\kappa\sqrt{n} + \sqrt{s}]\|h\|_2 \leq \underbrace{\sigma^{-1}[\kappa\sqrt{n} + \sqrt{s}]}_{C} \underbrace{\|Ah\|_2}_{=\|Aw\|_2} + \kappa\|w\|_1,\end{array}$$

as required in (1.2.5).

Condition $\mathbf{Q}_1(s, \kappa)$. For our purposes, it is convenient to present the condition (1.2.5) in the following flexible form:

$$\|w\|_{s,1} \leq s\|H^T Aw\| + \kappa\|w\|_1, \tag{1.2.6}$$

where H is an $m \times N$ *contrast* matrix and $\|\cdot\|$ is some norm on $\mathbf{R}^N$. Whenever a pair $(H, \|\cdot\|)$, called *contrast pair*, satisfies (1.2.6), we say that $(H, \|\cdot\|)$ *satisfies condition* $\mathbf{Q}_1(s, \kappa)$. From what we have seen, *If* A possesses Nullspace property with some sparsity level s and some $\kappa \in (0, 1/2)$, *then* there are many ways to select pairs $(H, \|\cdot\|)$ satisfying $\mathbf{Q}_1(s, \kappa)$, e.g., to take $H = CI_m$ with appropriately large C and $\|\cdot\| = \|\cdot\|_2$.

Conditions $\mathbf{Q}_q(s, \kappa)$. As we shall see in a while, it makes sense to embed the condition $\mathbf{Q}_1(s, \kappa)$ into a parametric family of conditions $\mathbf{Q}_q(s, \kappa)$, where the parameter q runs through $[1, \infty]$. Specifically,

Given $m \times n$ sensing matrix A, sparsity level $s \leq n$ and $\kappa \in (0, 1/2)$, we say that $m \times N$ matrix H and a norm $\|\cdot\|$ on $\mathbf{R}^N$ satisfy condition $\mathbf{Q}_q(s, \kappa)$, if

$$\|w\|_{s,q} \leq s^{\frac{1}{q}}\|H^T Aw\| + \kappa s^{\frac{1}{q}-1}\|w\|_1 \, \forall w \in \mathbf{R}^n. \tag{1.2.7}$$

Let us make two immediate observations on relations between the conditions:

A. When a pair $(H, \|\cdot\|)$ satisfies condition $\mathbf{Q}_q(s, \kappa)$, the pair satisfies also all conditions $\mathbf{Q}_{q'}(s, \kappa)$ with $1 \leq q' \leq q$.

Indeed in the situation in question for $1 \leq q' \leq q$ it holds

$$\|w\|_{s,q'} \leq s^{\frac{1}{q'}-\frac{1}{q}}\|w\|_{q,s} \leq s^{\frac{1}{q'}-\frac{1}{q}}\left[s^{\frac{1}{q}}\|H^T Aw\| + \kappa s^{\frac{1}{q}-1}\|w\|_1\right] = s^{\frac{1}{q'}}\|H^T Aw\| + \kappa s^{\frac{1}{q'}-1}\|w\|_1,$$

where the first inequality is the standard inequality between ℓ_p-norms of the s-dimensional vector w^s.

B. When a pair $(H, \|\cdot\|)$ satisfies condition $\mathbf{Q}_q(s,\kappa)$ and $1 \le s' \le s$, the pair $((s/s')^{\frac{1}{q}}H, \|\cdot\|)$ satisfies the condition $\mathbf{Q}_q(s',\kappa)$.

Indeed, in the situation in question we clearly have for $1 \le s' \le s$:

$$\|w\|_{s',q} \le \|w\|_{s,q} \le (s')^{\frac{1}{q}} \| \left[(s/s')^{\frac{1}{q}} H\right] Aw\| + \kappa \underbrace{s^{\frac{1}{q}-1}}_{\le (s')^{\frac{1}{q}-1}} \|w\|_1.$$

1.2.3 Regular ℓ_1 recovery

Given observation scheme (1.1.1) with $m \times n$ sensing matrix A, we define the *regular ℓ_1 recovery* of x via observation y as

$$\widehat{x}_{\rm reg}(y) \in \operatorname*{Argmin}_u \left\{ \|u\|_1 : \|H^T(Au - y)\| \le \rho \right\}, \tag{1.2.8}$$

where the *contrast matrix* $H \in \mathbf{R}^{m\times N}$, the norm $\|\cdot\|$ on $\mathbf{R}^N$ and $\rho > 0$ are parameters of the construction.

The role of $\mathbf{Q}$-conditions we have introduced is clear from the following

Theorem 1.2.1 *Let s be a positive integer, $q \in [1,\infty]$ and $\kappa \in (0, 1/2)$. Assume that the pair $(H, \|\cdot\|)$ satisfies the condition $\mathbf{Q}_q(s,\kappa)$ associated with A, and let*

$$\Xi_\rho = \{\eta : \|H^T\eta\| \le \rho\}. \tag{1.2.9}$$

Then for all $x \in \mathbf{R}^n$ and $\eta \in \Xi_\rho$ one has

$$\|\widehat{x}_{\rm reg}(Ax + \eta) - x\|_p \le \frac{4(2s)^{\frac{1}{p}}}{1 - 2\kappa}\left[\rho + \frac{\|x - x^s\|_1}{2s}\right],\ 1 \le p \le q. \tag{1.2.10}$$

The above result can be slightly strengthened by replacing the assumption that $(H, \|\cdot\|)$ satisfies $\mathbf{Q}_q(s,\kappa)$ with some $\kappa < 1/2$, with a weaker, by observation **A** from Section 1.2.2.1, assumption that $(H, \|\cdot\|)$ satisfies $\mathbf{Q}_1(s,\varkappa)$ with $\varkappa < 1/2$ and satisfies $\mathbf{Q}_q(s,\kappa)$ with some (perhaps large) κ:

Theorem 1.2.2 *Given A, integer $s > 0$ and $q \in [1,\infty]$, assume that $(H, \|\cdot\|)$ satisfies the condition $\mathbf{Q}_1(s,\varkappa)$ with $\varkappa < 1/2$ and the condition $\mathbf{Q}_q(s,\kappa)$ with some $\kappa \ge \varkappa$, and let Ξ_ρ be given by* (1.2.9). *Then for all $x \in \mathbf{R}^n$ and $\eta \in \Xi_\rho$ it holds:*

$$\|\widehat{x}_{\rm reg}(Ax + \eta) - x\|_p \le \frac{4(2s)^{\frac{1}{p}}[1 + \kappa - \varkappa]^{\frac{q(p-1)}{p(q-1)}}}{1 - 2\varkappa}\left[\rho + \frac{\|x - x^s\|_1}{2s}\right],\ 1 \le p \le q. \tag{1.2.11}$$

Before commenting on the above results, let us present their alternative versions.

1.2.4 Penalized ℓ_1 recovery

Penalized ℓ_1 recovery of signal x from its observation (1.1.1) is

$$\widehat{x}_{\rm pen}(y) \in \operatorname*{Argmin}_u \left\{ \|u\|_1 + \lambda \|H^T(Au - y)\| \right\}, \tag{1.2.12}$$

where $H \in \mathbf{R}^{m\times N}$, a norm $\|\cdot\|$ on $\mathbf{R}^N$ and a positive real λ are parameters of the construction.

Theorem 1.2.3 *Given A, positive integer s, and $q\in[1,\infty]$, assume that $(H,\|\cdot\|)$ satisfies the conditions $\mathbf{Q}_q(s,\kappa)$ and $\mathbf{Q}_1(s,\varkappa)$ with $\varkappa<1/2$ and $\kappa\geq\varkappa$. Then*

(i) *Let $\lambda\geq 2s$. Then for all $x\in\mathbf{R}^n$, $y\in\mathbf{R}^m$ it holds:*

$$\|\widehat{x}_{\rm pen}(y)-x\|_p\leq\frac{4\lambda^{\frac{1}{p}}}{1-2\varkappa}\left[1+\frac{\kappa\lambda}{2s}-\varkappa\right]^{\frac{q(p-1)}{p(q-1)}}\left[\|H^T(Ax-y)\|+\frac{\|x-x^s\|_1}{2s}\right],\ 1\leq p\leq q. \tag{1.2.13}$$

In particular, with $\lambda=2s$ we have:

$$\|\widehat{x}_{\rm pen}(y)-x\|_p\leq\frac{4(2s)^{\frac{1}{p}}}{1-2\varkappa}\left[1+\kappa-\varkappa\right]^{\frac{q(p-1)}{p(q-1)}}\left[\|H^T(Ax-y)\|+\frac{\|x-x^s\|_1}{2s}\right]. \tag{1.2.14}$$

(ii) *Let $\rho\geq 0$, and let Ξ_ρ be given by (1.2.9). Then for all $x\in\mathbf{R}^n$ and all $\eta\in\Xi_\rho$ one has:*

$$\begin{array}{l}\lambda\geq 2s\ \ \Rightarrow\\ \qquad\|\widehat{x}_{\rm pen}(Ax+\eta)-x\|_p\leq\frac{4\lambda^{\frac{1}{p}}}{1-2\varkappa}\left[1+\frac{\kappa\lambda}{2s}-\varkappa\right]^{\frac{q(p-1)}{p(q-1)}}\left[\rho+\frac{\|x-x^s\|_1}{2s}\right],\ 1\leq p\leq q;\\ \lambda=2s\ \ \Rightarrow\\ \qquad\|\widehat{x}_{\rm pen}(Ax+\eta)-x\|_p\leq\frac{4(2s)^{\frac{1}{p}}}{1-2\varkappa}\left[1+\kappa-\varkappa\right]^{\frac{q(p-1)}{p(q-1)}}\left[\rho+\frac{\|x-x^s\|_1}{2s}\right],\ 1\leq p\leq q.\end{array} \tag{1.2.15}$$

1.2.5 Discussion

Some remarks are in order.

A. Qualitatively speaking, Theorems 1.2.1, 1.2.2, 1.2.3 say the same: under **Q**-conditions, the regular, resp., penalized recoveries are capable to reproduce the true signal *exactly* when there is no observation noise and the signal is s-sparse; in the presence of observation error η and imperfect sparsity, the signal is recovered within the error which can be upper-bounded by the sum of two terms, one proportional to the magnitude of observation noise and one proportional to the deviation $\|x-x^s\|_1$ of the signal from s-sparse ones. In the penalized recovery, the observation error is measured in the scale given by the contrast matrix and the norm $\|\cdot\|$ - as $\|H^T\eta\|$, and in the regular one – by an a priori upper bound ρ on $\|H^T\eta\|$ — when $\rho\geq\|H^T\eta\|$, η belongs to Ξ_ρ and thus the bounds (1.2.10), (1.2.11) are applicable to the actual observation error η. Clearly, in qualitative terms error bound of this type is the best we may hope for. Now let us look at the quantitative aspect. Assume that in the regular recovery we use $\rho\approx\|H^T\eta\|$, and in the penalized one use $\lambda=2s$. In this case, error bounds (1.2.10), (1.2.11), (1.2.15), up to factors C depending solely on $\varkappa$ and κ, are the same, specifically,

$$\|\widehat{x}-x\|_p\leq Cs^{1/p}[\|H^T\eta\|+\|x-x^s\|_1/s],\ 1\leq p\leq q. \tag{!}$$

Is this error bound bad or good? The answer depends on many factors, including on how well we select H and $\|\cdot\|$. To get a kind of orientation, consider the trivial case of *direct* observations, where matrix A is square and, moreover, is proportional to the unit matrix: $A=\alpha I$; assume in addition that x is exactly s-sparse. In this case, the simplest way to ensure condition $\mathbf{Q}_q(s,\kappa)$, even with $\kappa=0$, is to take $\|\cdot\|=\|\cdot\|_{s,q}$ and $H=s^{-1/q}\alpha^{-1}I$, so that (!) becomes

$$\|\widehat{x}-x\|_p\leq C\alpha^{-1}s^{1/p-1/q}\|\eta\|_{s,q},\ 1\leq p\leq q. \tag{!!}$$

As far as the dependence of the bound on the magnitude $\|\eta\|_{s,q}$ of the observation noise is concerned, this dependence is as good as it can be – even if we knew in advance the positions of the s largest in magnitude entries of x, we would be unable to recover x is q-norm with error $\leq\alpha^{-1}\|\eta\|_{s,q}$ (why?); in addition, with the equal to each other s largest magnitudes of entries in η, the $\|\cdot\|_p$-norm of the recovery error clearly cannot be guaranteed to be less than $\alpha^{-1}\|\eta\|_{s,p}=\alpha^{-1}s^{1/p-1/q}\|\eta\|_{s,q}$. Thus, at least for s-sparse signals x, our error bound is, basically, the best one can get already in the "ideal" case of direct observations.

B. Given that $(H, \|\cdot\|)$ obeys $\mathbf{Q}_1(s, \varkappa)$ with some $\varkappa < 1/2$, the larger is q such that the pair $(H, \|\cdot\|)$ obeys the condition $\mathbf{Q}_q(s, \kappa)$ with a given $\kappa \geq \varkappa$ (κ can be $\geq 1/2$) and s, the larger is the range $p \leq q$ of values of p where the error bounds (1.2.11), (1.2.15) are applicable. This is in full accordance with the fact that if a pair $(H, \|\cdot\|)$ obeys condition $\mathbf{Q}_q(s, \kappa)$, it obeys also all conditions $\mathbf{Q}_{q'}(s, \kappa)$ with $1 \leq q' \leq q$ (item **A** in Section 1.2.2.1).

C. Flexibility offered by contrast matrix H and norm $\|\cdot\|$ allows to adjust, to some extent, the recovery to the "geometry of observation errors." For example, when η is "uncertain but bounded," say, all we know is that $\|\eta\|_2 \leq \delta$ with some given δ, all what matters (on the top of the requirement for $(H, \|\cdot\|)$ to obey $\mathbf{Q}$-conditions) is how large could be $\|H^T\eta\|$ when $\|\eta\|_2 \leq \delta$. In particular, when $\|\cdot\| = \|\cdot\|_2$, the error bound "is governed" by the spectral norm of H; consequently, if we have a technique allowing *to design* H such that $(H, \|\cdot\|_2)$ obeys $\mathbf{Q}$-condition(s) with given parameters, it makes sense to look for design with as small spectral norm of H as possible. In contrast to this, in the most interesting for applications case of Gaussian noise:

$$y = Ax + \eta,\ \eta \sim \mathcal{N}(0, \sigma^2 I_m) \tag{1.2.16}$$

looking at the spectral norm of H, with $\|\cdot\|_2$ in the role of $\|\cdot\|$, is counter-productive, since a typical realization of η is of Euclidean norm of order of $\sqrt{m}\sigma$ and thus is quite large when m is large. In this case to quantify "the magnitude" of $H^T\eta$ by the product of the spectral norm of H and the Euclidean norm of η is *completely misleading* – in typical cases, this product will grow rapidly with the number of observations m, completely ignoring the fact that η is random with zero mean[5]. What is much better suited for the case of Gaussian noise, is $\|\cdot\|_\infty$ norm in the role of $\|\cdot\|$ and the norm "the maximum of $\|\cdot\|_2$-norms of the columns in H," let it be denoted by $\|H\|_{1,2}$, of H. Indeed, with $\eta \sim \mathcal{N}(0, \sigma^2 I_m)$, the entries in $H^T\eta$ are Gaussian with zero mean and variance bounded by $\sigma^2\|H\|_{1,2}^2$, so that $\|H^T\eta\|_\infty$ is the maximum of magnitudes of N zero mean Gaussian random variables with standard deviations bounded by $\sigma\|H\|_{1,2}$. As a result,

$$\mathrm{Prob}\{\|H^T\eta\|_\infty \geq \rho\} \leq N\mathrm{Erf}(\rho/\sigma)\|H\|_{1,2} \leq N\mathrm{e}^{\frac{-\rho^2}{2\sigma^2}}\|H\|_{1,2}, \tag{1.2.17}$$

where

$$\mathrm{Erf}(s) = \frac{1}{\sqrt{2\pi}}\int_s^\infty \mathrm{e}^{-t^2/2}dt$$

is the error function. It follows that the typical values of $\|H^T\eta\|_\infty$, $\eta \sim \mathcal{N}(0, \sigma^2 I_m)$ are of order of at most $\sigma\sqrt{\ln(N)}\|H\|_{1,2}$; typically, $N = O(m)$, so that with σ and $\|H\|_{1,2}$ given, typical values $\|H^T\eta\|_\infty$ are nearly independent of m. The bottom line is that ℓ_1 minimization is capable to handle large-scale Gaussian observation noise incomparably better than "uncertain-but-bounded" observation noise of similar magnitude (measured in Euclidean norm).

D. As far as comparison of regular and penalized ℓ_1 recoveries with the same pair $(H, \|\cdot\|)$ is concerned, the situation is as follows. Assume for the sake of simplicity that $(H, \|\cdot\|)$ satisfies $\mathbf{Q}_q(s, \kappa)$ with some s and some $\kappa < 1/2$, and let the observation error be random. Given $\epsilon \in (0, 1)$, let

$$\rho_\epsilon[H, \|\cdot\|] = \min\left\{\rho : \ \mathrm{Prob}\left\{\eta : \|H^T\eta\| \leq \rho\right\} \geq 1 - \epsilon\right\}; \tag{1.2.18}$$

this is nothing but the smallest ρ such that

$$\mathrm{Prob}\{\eta \in \Xi_\rho\} \geq 1 - \epsilon \tag{1.2.19}$$

[5] the simplest way to see the difference is to look at a particular entry $h^T\eta$ in $H^T\eta$. Operating with spectral norms, we upper-bound this entry by $\|h\|_2\|\eta\|_2$, and the second factor for $\eta \sim \mathcal{N}(0, \sigma^2 I_m)$ is typically as large as $\sigma\sqrt{m}$, in sharp contrast to the fact that typical values of $h^T\eta$ are of order of σ, completely independently of what m is!

(see (1.2.9)) and thus – the smallest ρ for which the error bound (1.2.10) for the regular ℓ_1 recovery holds true with probability $1-\epsilon$ (or at least the smallest ρ for which the latter claim is supported by Theorem 1.2.1). With $\rho = \rho_\epsilon[H, \|\cdot\|]$, the regular ℓ_1 recovery guarantees (and that is the best guarantee one can extract from Theorem 1.2.1) that

> (#) *For some set* Ξ, $\text{Prob}\{\eta \in \Xi\} \geq 1-\epsilon$, *of "good" realizations of* $\eta \sim \mathcal{N}(0, \sigma^2 I_m)$, *one has*
>
> $$\|\widehat{x}(Ax+\eta) - x\|_p \leq \frac{4(2s)^{\frac{1}{p}}}{1-2\kappa}\left[\rho_\epsilon[H, \|\cdot\|] + \frac{\|x-x^s\|_1}{2s}\right], \; 1 \leq p \leq q, \tag{1.2.20}$$
>
> *whenever* $x \in \mathbf{R}^n$ *and* $\eta \in \Xi_\rho$.

The error bound (1.2.14) (where we set $\varkappa = \kappa$) says that *(#) holds true for the penalized ℓ_1 recovery with $\lambda = 2s$.* The latter observation suggests that the penalized ℓ_1 recovery associated with $(H, \|\cdot\|)$ and $\lambda = 2s$ is better than its regular counterpart, the reason being twofold. First, in order to ensure (#) with the regular recovery, the "built in" parameter ρ of this recovery should be set to $\rho_\epsilon[H, \|\cdot\|]$, and the latter quantity not always is easy to identify. In contrast to this, the construction of penalized ℓ_1 recovery is completely independent of a priori assumptions on the structure of observation errors, while automatically ensuring (#) for the error model we use. Second, and more importantly, for the penalized recovery the bound (1.2.20) is no more than the "worst, with confidence $1-\epsilon$, case," while the typical values of the quantity $\|H^T\eta\|$ which indeed participates in the error bound (1.2.13) are essentially smaller than $\rho_\epsilon[H, \|\cdot\|]$. Numerical experience fully supports the above suggestion: the difference in observed performance of the two routines in question, although not dramatic, is definitely in favor of the penalized recovery. The only potential disadvantage of the latter routine is that the penalty parameter λ should be tuned to the level s of sparsity we aim at, while the regular recovery is free of any guess of this type. Of course, the "tuning" is rather loose – all we need (and experiments show that we indeed need this) is the relation $\lambda \geq 2s$, so that a rough upper bound on s will do; note, however, that bound (1.2.13) deteriorates as λ grows.

Finally, we remark that when H is $m \times N$ and $\eta \sim \mathcal{N}(0, \sigma^2 I_m)$, we have

$$\rho_\epsilon[H, \|\cdot\|_\infty] \leq \text{ErfInv}(\epsilon/N)\|H\|_{1,2} \leq \sqrt{2\ln(N/\epsilon)}\|H\|_{1,2}$$

(see 1.2.17)); here $\text{ErfInv}(\delta)$ is the inverse error function:

$$\text{Erf}(\text{ErfInv}(\delta)) = \delta, \; 0 < \delta < 1.$$

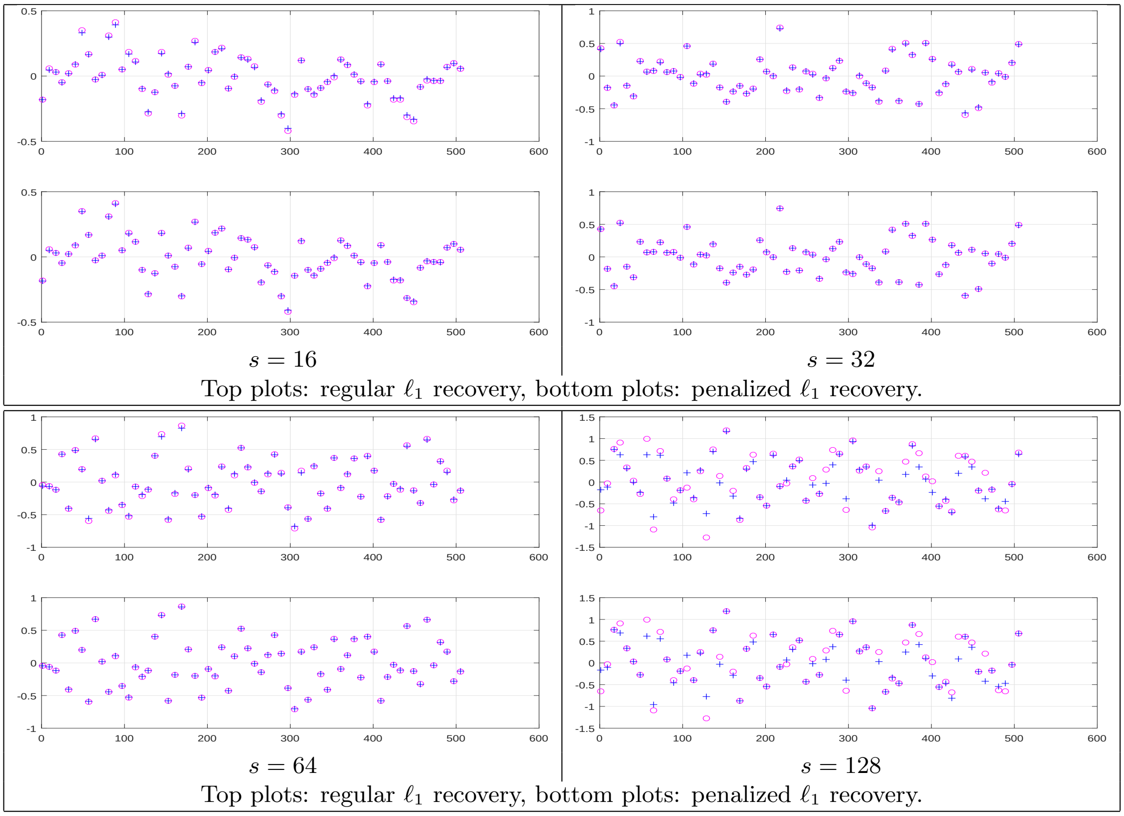
How it works. Here we present a small numerical illustration. We observe in Gaussian noise $m = n/2$ randomly selected terms in n-element "time series" $z = (z_1, ..., z_n)$ and want to recover this series under the assumption that the series is "nearly s-sparse in frequency domain," that is, that

$$z = Fx \text{ with } \|x - x^s\|_1 \leq \delta,$$

where F is the matrix of $n \times n$ Inverse Discrete Cosine Transform, x^s is the vector obtained from x by zeroing out all but s largest in magnitude entries, and δ upper-bounds the distance from x to s-sparse signals. Denoting by A the $m \times n$ submatrix of F corresponding to the time instants t where z_t is observed, our observation scheme becomes

$$y = Ax + \sigma\xi,$$

where ξ is the standard Gaussian noise. After the signal in frequency domain, that is, x, is recovered by ℓ_1 minimization, let the recovery be $\widehat{x}$, we recover the signal in the time domain as $\widehat{z} = F\widehat{x}$.



$s = 16$ $s = 32$

Top plots: regular ℓ_1 recovery, bottom plots: penalized ℓ_1 recovery.

$s = 64$ $s = 128$

Top plots: regular ℓ_1 recovery, bottom plots: penalized ℓ_1 recovery.

	$s = 16$	$s = 32$	$s = 64$	$s = 128$
$\|z - \widehat{z}\|_2$	0.170	0.220	0.365	3.772
$\|z - \widehat{z}\|_\infty$	0.0239	0.0323	0.0608	0.729

recovery errors, regular ℓ_1 recovery

	$s = 16$	$s = 32$	$s = 64$	128
$\|z - \widehat{z}\|_2$	0.0679	0.0673	0.0812	3.665
$\|z - \widehat{z}\|_\infty$	0.0095	0.0107	0.0143	0.705

recovery errors, penalized ℓ_1 recovery

Figure 1.3: Regular and penalized ℓ_1 recovery of nearly s-sparse signals. Red circles: true time series, blue crosses: recovered time series (to make the plots readable, one per eight consecutive terms in the time series is shown). Problem's sizes are $m = 256$ and $n = 2m = 512$, noise level is $\sigma = 0.01$, deviation from s-sparsity is $\|x - x^s\|_1 = 1$, contrast pair is $(H = \sqrt{n/m}A, \|\cdot\|_\infty)$. In penalized recovery, $\lambda = 2s$, parameter ρ in regular recovery is set to $\mathrm{ErfInv}(0.005/n)$.

On Figure 1.3, we present four test signals, of different (near) sparsity, along with their regular and penalized ℓ_1 recoveries. The data on Figure 1.3 clearly show how the quality of ℓ_1 recovery deteriorates as the number s of "essential nonzeros" of the signal in the frequency domain grows. It is seen also that the penalized recovery meaningfully outperforms the regular one in the range of sparsities up to 64.

1.2.6 Proofs of Theorems 1.2.1, 1.2.2, 1.2.3

1.2.6.1 Proofs of Theorem 1.2.1, 1.2.2

All we need is to prove Theorem 1.2.2, since Theorem 1.2.1 is the particular case $\varkappa = \kappa < 1/2$ of Theorem 1.2.2.

Let us fix $x \in \mathbf{R}^n$ and $\eta \in \Xi_\rho$, and let us set $\widehat{x} = \widehat{x}_{\mathrm{reg}}(Ax + \eta)$. Let also $I \subset \{1, ..., n\}$ be the set of indexes of the s largest in magnitude entries in x, I^o be the complement of I in $\{1, ..., n\}$, and let for $w \in \mathbf{R}^n$, w_I and w_{I^o} be the vectors obtained from w by zeroing entries with indexes $j \notin I$ and $j \notin I^o$, respectively, and keeping the remaining entries intact. Finally, let $z = \widehat{x} - x$.

1^0. By the definition of Ξ_ρ and due to $\eta \in \Xi_\rho$ we have

$$\|H^T([Ax+\eta]-Ax)\| \leq \rho, \tag{1.2.21}$$

so that x is a feasible solution to the optimization problem specifying $\widehat{x}$, whence $\|\widehat{x}\|_1 \leq \|x\|_1$. We therefore have

$$\begin{array}{rcl}\|\widehat{x}_{I^o}\|_1 & = & \|\widehat{x}\|_1 - \|\widehat{x}_I\|_1 \leq \|x\|_1 - \|\widehat{x}_I\|_1 = \|x_I\|_1 + \|x_{I^o}\|_1 - \|\widehat{x}_I\|_1 \\ & \leq & \|z_I\|_1 + \|x_{I^o}\|_1, \end{array} \tag{1.2.22}$$

and therefore

$$\|z_{I^o}\|_1 \leq \|\widehat{x}_{I^o}\|_1 + \|x_{I^o}\|_1 \leq \|z_I\|_1 + 2\|x_{I^o}\|_1.$$

It follows that

$$\|z\|_1 = \|z_I\|_1 + \|z_{I^o}\|_1 \leq 2\|z_I\|_1 + 2\|x_{I^o}\|_1. \tag{1.2.23}$$

Further, by definition of $\widehat{x}$ we have $\|H^T([Ax+\eta]-A\widehat{x})\| \leq \rho$, which combines with (1.2.21) to imply that

$$\|H^T A(\widehat{x}-x)\| \leq 2\rho. \tag{1.2.24}$$

2^0. Since $(H, \|\cdot\|)$ satisfies $\mathbf{Q}_1(s,\varkappa)$, we have

$$\|z\|_{s,1} \leq s\|H^T Az\| + \varkappa\|z\|_1.$$

By (1.2.24), it follows that $\|z\|_{s,1} \leq 2s\rho + \varkappa\|z\|_1$, which combines with the evident inequality $\|z_I\| \leq \|z\|_{s,1}$ (recall that $\mathrm{Card}(I)=s$) and with (1.2.23) to imply that

$$\|z_I\|_1 \leq 2s\rho + \varkappa\|z\|_1 \leq 2s\rho + 2\varkappa\|z_I\|_1 + 2\varkappa\|x_{I^o}\|_1,$$

whence

$$\|z_I\|_1 \leq \frac{2s\rho + 2\varkappa\|x_{I^o}\|_1}{1-2\varkappa}.$$

Invoking (1.2.23), we conclude that

$$\|z\|_1 \leq \frac{4s}{1-2\varkappa}\left[\rho + \frac{\|x_{I^o}\|_1}{2s}\right]. \tag{1.2.25}$$

3^0. Since $(H, \|\cdot\|)$ satisfies $\mathbf{Q}_q(s,\kappa)$, we have

$$\|z\|_{s,q} \leq s^{\frac{1}{q}}\|H^T Az\| + \kappa s^{\frac{1}{q}-1}\|z\|_1,$$

which combines with (1.2.25) and (1.2.24) to imply that

$$\|z\|_{s,q} \leq s^{\frac{1}{q}}2\rho + \kappa s^{\frac{1}{q}}\tfrac{4\rho+2s^{-1}\|x_{I^o}\|_1}{1-2\varkappa} \leq \tfrac{4s^{\frac{1}{q}}[1+\kappa-\varkappa]}{1-2\varkappa}\left[\rho + \tfrac{\|x_o\|_1}{2s}\right] \tag{1.2.26}$$

(we have taken into account that $\varkappa < 1/2$ and $\kappa \geq \varkappa$). Let θ be the $(s+1)$-st largest magnitude of entries in z, and let $w = z - z^s$. Now (1.2.26) implies that

$$\theta \leq \|z\|_{s,q}s^{-\frac{1}{q}} \leq \frac{4[1+\kappa-\varkappa]}{1-2\varkappa}\left[\rho + \frac{\|x_{I^o}\|_1}{2s}\right].$$

Hence invoking (1.2.25) we have

$$\begin{array}{rcl}\|w\|_q & \leq & \|w\|_\infty^{\frac{q-1}{q}}\|w\|_1^{\frac{1}{q}} \leq \theta^{\frac{q-1}{q}}\|z\|_1^{\frac{1}{q}} \\ & \leq & \theta^{\frac{q-1}{q}}\tfrac{(4s)^{\frac{1}{q}}}{[1-2\varkappa]^{\frac{1}{q}}}\left[\rho + \tfrac{\|x_{I^o}\|_1}{2s}\right]^{\frac{1}{q}} \\ & \leq & \tfrac{4s^{\frac{1}{q}}[1+\kappa-\varkappa]^{\frac{q-1}{q}}}{1-2\varkappa}\left[\rho + \tfrac{\|x_{I^o}\|_1}{2s}\right]. \end{array}$$

Taking into account (1.2.26) and the fact that the supports of z^s and w do not intersect, we get

$$\begin{array}{rcl}\|z\|_q & \leq & 2^{\frac{1}{q}}\max[\|z^s\|_q,\|w\|_q] = 2^{\frac{1}{q}}\max[\|z\|_{s,q},\|w\|_q]\\ & \leq & \frac{4(2s)^{\frac{1}{q}}[1+\kappa-\varkappa]}{1-2\varkappa}\left[\rho+\frac{\|x_{I^o}\|_1}{2s}\right].\end{array}$$

This bound combines with (1.2.25), the Hölder inequality and the relation $\|x_{I^o}\|_1 = \|x-x^s\|_1$ to imply (1.2.11). □

1.2.6.2 Proof of Theorem 1.2.3

Let us prove (i). Let us fix $x \in \mathbf{R}^n$ and η, and let us set $\widehat{x} = \widehat{x}_{\mathrm{pen}}(Ax+\eta)$. Let also $I \subset \{1,...,K\}$ be the set of indexes of the s largest in magnitude entries in x, I^o be the complement of I in $\{1,...,n\}$, and for $w \in \mathbf{R}^n$ let w_I, w_{I^o} be the vectors obtained from w by zeroing out all entries with indexes not in I, respectively, not in I^o. Finally, let $z = \widehat{x} - x$ and $\nu = \|H^T\eta\|$.

1^0. We have

$$\|\widehat{x}\|_1 + \lambda\|H^T(A\widehat{x} - Ax - \eta)\| \leq \|x\|_1 + \lambda\|H^T\eta\|$$

and

$$\|H^T(A\widehat{x} - Ax - \eta)\| = \|H^T(Az-\eta)\| \geq \|H^TAz\| - \|H^T\eta\|,$$

whence

$$\|\widehat{x}\|_1 + \lambda\|H^TAz\| \leq \|x\|_1 + 2\lambda\|H^T\eta\| = \|x\|_1 + 2\lambda\nu. \tag{1.2.27}$$

We have

$$\begin{array}{rcl}\|\widehat{x}\|_1 & = & \|x+z\|_1 = \|x_I + z_I\|_1 + \|x_{I^o} + z_{I^o}\|_1\\ & \geq & \|x_I\|_1 - \|z_I\|_1 + \|z_{I^o}\|_1 - \|x_{I^o}\|_1,\end{array}$$

which combines with (1.2.27) to imply that

$$\|x_I\|_1 - \|z_I\|_1 + \|z_{I^o}\|_1 - \|x_{I^o}\|_1 + \lambda\|H^TAz\| \leq \|x\|_1 + 2\lambda\nu,$$

or, which is the same,

$$\|z_{I^o}\|_1 - \|z_I\|_1 + \lambda\|H^TAz\| \leq 2\|x_{I^o}\|_1 + 2\lambda\nu. \tag{1.2.28}$$

Since $(H, \|\cdot\|)$ satisfies $\mathbf{Q}_1(s,\varkappa)$, we have

$$\|z_I\|_1 \leq \|z\|_{s,1} \leq s\|H^TAz\| + \varkappa\|z\|_1,$$

so that

$$(1-\varkappa)\|z_I\|_1 - \varkappa\|z_{I^o}\|_1 - s\|H^TAz\| \leq 0. \tag{1.2.29}$$

Taking weighted sum of (1.2.28) and (1.2.29), the weights being 1, 2, respectively, we get

$$(1-2\varkappa)\left[\|z_I\|_1 + \|z_{I^o}\|_1\right] + (\lambda - 2s)\|H^TAz\| \leq 2\|x_{I^o}\|_1 + 2\lambda\nu,$$

that is (since $\lambda \geq 2s$),

$$\|z\|_1 \leq \frac{2\lambda\nu + 2\|x_{I^o}\|_1}{1-2\varkappa} \leq \frac{2\lambda}{1-2\varkappa}\left[\nu + \frac{\|x_{I^o}\|_1}{2s}\right]. \tag{1.2.30}$$

Further, by (1.2.27) we have

$$\lambda\|H^TAz\| \leq \|x\|_1 - \|\widehat{x}\|_1 + 2\lambda\nu \leq \|z\|_1 + 2\lambda\nu,$$

which combines with (1.2.30) to imply that

$$\lambda\|HA^Tz\| \leq \frac{2\lambda\nu + 2\|x_{I^o}\|_1}{1-2\varkappa} + 2\lambda\nu = \frac{2\lambda\nu(2-2\varkappa) + 2\|x_{I^o}\|_1}{1-2\varkappa}. \tag{1.2.31}$$

From $\mathbf{Q}_q(s,\kappa)$ it follows that

$$\|z\|_{s,q} \leq s^{\frac{1}{q}}\|H^TAz\| + \kappa s^{\frac{1}{q}-1}\|z\|_1,$$

which combines with (1.2.31) and (1.2.30) to imply that

$$\begin{array}{rl}\|z\|_{s,q} & \leq s^{\frac{1}{q}-1}\left[s\|H^TAz\| + \kappa\|z\|_1\right] \leq s^{\frac{1}{q}-1}\left[\frac{4s\nu(1-\varkappa)+\frac{2s}{\lambda}\|x_{I^o}\|_1}{1-2\varkappa} + \frac{\kappa[2\lambda\nu+2\|x_{I^o}\|_1]}{1-2\varkappa}\right] \\ & = s^{\frac{1}{q}}\frac{4\nu(1-\varkappa)+2s^{-1}\lambda\kappa\nu]+2[\lambda^{-1}+s^{-1}\kappa]\|x_{I^o}\|_1}{1-2\varkappa} \leq 4\frac{s^{\frac{1}{q}}}{1-2\varkappa}\left[1+\frac{\kappa\lambda}{2s}-\varkappa\right]\left[\nu+\frac{\|x_{I^o}}{2s}\right]\end{array} \tag{1.2.32}$$

(recall that $\lambda \geq 2s$, $\kappa \geq \varkappa$, and $\varkappa < 1/2$). It remains to repeat the reasoning following (1.2.26) in item 3^0 of the proof of Theorem 1.2.2. Specifically, denoting by θ the $(s+1)$-st largest magnitude of entries in z, (1.2.32) implies that

$$\theta \leq s^{-1/q}\|z\|_{s,q} \leq \frac{4}{1-2\varkappa}[1+\kappa\frac{\lambda}{2s}-\varkappa]\left[\nu+\frac{\|x_{I^o}\|_1}{2s}\right], \tag{1.2.33}$$

so that for the vector $w = z - z^s$ one has

$$\|w\|_q \;\; \leq \;\; \theta^{1-\frac{1}{q}}\|w\|_1^{\frac{1}{q}} \leq \frac{4(\lambda/2)^{\frac{1}{q}}}{1-2\varkappa}\left[1+\kappa\frac{\lambda}{2s}-\varkappa\right]^{\frac{q-1}{q}}\left[\nu+\frac{\|x_{I^o}\|_1}{2s}\right]$$

(we have used (1.2.33), (1.2.30) and the fact that $\lambda \geq 2s$). Hence, taking into account that z^s and w have non-intersecting supports,

$$\begin{array}{rcl}\|z\|_q & \leq & 2^{\frac{1}{q}}\max[\|z^s\|_q, \|w\|_q] = 2^{\frac{1}{q}}\max[\|z\|_{s,q}, \|w\|_q] \\ & \leq & \frac{4\lambda^{\frac{1}{q}}}{1-2\varkappa}\left[1+\kappa\frac{\lambda}{2s}-\varkappa\right]\left[\nu+\frac{\|x_{I^o}\|_1}{2s}\right]\end{array}$$

(we have used (1.2.32) along with $\lambda \geq 2s$ and $\kappa \geq \varkappa$). This combines with (1.2.30) and Hölder inequality to imply (1.2.13). All remaining claims of Theorem 1.2.3 are immediate corollaries of (1.2.13). □

1.3 Verifiability and tractability issues

Good news on ℓ_1 recovery stated in Theorems 1.2.1, 1.2.2, 1.2.3 are "conditional" – we assume that we are smart enough to point out a pair $(H, \|\cdot\|)$ satisfying condition $\mathbf{Q}_1(s,\varkappa)$ with $\varkappa < 1/2$ (and condition $\mathbf{Q}_q(s,\kappa)$ with a "moderate" $\varkappa$ [6]). The related issues are twofold:

1. First, we do not know in which range of s, m, n these conditions, or even the weaker than $\mathbf{Q}_1(s,\varkappa)$, $\varkappa < 1/2$, Nullspace property can be satisfied; and without the Nullspace property, ℓ_1 minimization becomes useless, at least when we want to guarantee its validity whatever be s-sparse signal we want to recover;

2. Second, it is unclear how to verify whether a given sensing matrix A satisfies the Nullspace property for a given s, or a given pair $(H, \|\cdot\|)$ satisfies the condition $\mathbf{Q}_q(s,\kappa)$ with given parameters.

What is known on these crucial issues, can be outlined as follows.

[6] $\mathbf{Q}_q(s,\kappa)$ always is satisfied with "large" κ, namely, $\kappa = s$, but this large value of κ is of no interest: the associated bounds on p-norms of recovery error are straightforward consequences of the bounds on $\|\cdot\|_1$-norm of this error yielded by the condition $\mathbf{Q}_1(s,\varkappa)$.

1. It is known that for given m, n with $m \ll n$ (say, $m/n \leq 1/2$), there exist $m \times n$ sensing matrices which are s-good for the values of s "nearly as large as m", specifically, for $s \leq O(1)\frac{m}{\ln(n/m)}$ [7]. Moreover, there are natural families of matrices where this level of goodness "is a rule." E.g., when drawing an $m \times n$ matrix at random from the Gaussian or the ± 1 distributions (i.e., filling the matrix with independent realizations of a random variable which is either a standard (zero mean, unit variance) Gaussian one, or takes values ± 1 with probabilities 0.5), the result will be s-good, for the outlined value of s, with probability approaching 1 as m and n grow. All this remains true when instead of speaking about matrices A satisfying "plain" Nullspace properties, we are speaking about matrices A for which it is easy to point out a pair $(H, \|\cdot\|)$ satisfying the condition $\mathbf{Q}_2(s, \varkappa)$ with, say, $\varkappa = 1/4$.

   The above results can be considered as a good news. A bad news is, that we do *not* know how to check efficiently, given an s and a sensing matrix A, that the matrix is s-good, same as we do not know how to check that A admits good (i.e., satisfying $\mathbf{Q}_1(s, \varkappa)$ with $\varkappa < 1/2$) pairs $(H, \|\cdot\|)$. Even worse: we do not know an efficient recipe allowing to build, given m, an $m \times 2m$ matrix A^m which is provably s-good for s larger than $O(1)\sqrt{m}$, which is a much smaller "level of goodness" then the one promised by theory for randomly generated matrices[8]. The "common life" analogy of this pitiful situation would be as follows: you know that 90% of bricks in your wall are made of gold, and at the same time, you do not know how to tell a golden brick from a usual one.[9]

2. There exist *verifiable sufficient conditions* for s-goodness of a sensing matrix, same as verifiable sufficient conditions for a pair $(H, \|\cdot\|)$ to satisfy condition $\mathbf{Q}_q(s, \kappa)$. A bad news that when $m \ll n$, these verifiable sufficient conditions can be satisfied only when $s \leq O(1)\sqrt{m}$ – once again, in a much more narrow range of values of s than the one where typical randomly selected sensing matrices are s-good. In fact, $s = O(\sqrt{m})$ is *the best* known so far sparsity level for which we know individual s-good $m \times n$ sensing matrices with $m \leq n/2$.

[7] From now on, $O(1)$'s denote positive *absolute constants* – appropriately chosen numbers like 0.5, or 1, or perhaps 100,000. We could, in principle, replace all $O(1)$'s by specific numbers; following the standard mathematical practice, we do not do it, partly from laziness, partly because the particular values of these numbers in our context are irrelevant.

[8] Note that the naive algorithm "generate $m \times 2m$ matrices at random until an s-good, with s promised by the theory, matrix is generated" is *not* an efficient recipe, since we do not know how to check s-goodness efficiently.

[9] This phenomenon is met in many other situations. E.g., in 1938 Claude Shannon (1916-2001), "the father of Information Theory," made (in his M.Sc. Thesis!) a fundamental discovery as follows. Consider a Boolean function of n Boolean variables (i.e., both the function and the variables take values 0 and 1 only); as it is easily seen there are 2^{2^n} function of this type, and every one of them can be computed by a dedicated circuit comprised of "switches" implementing just 3 basic operations AND, OR and NOT (like computing a polynomial can be carried out on a circuit with nodes implementing just two basic operation: addition of reals and their multiplication). The discovery of Shannon was that every Boolean function of n variables can be computed on a circuit with no more than $Cn^{-1}2^n$ switches, where C is an appropriate absolute constant. Moreover, Shannon proved that "nearly all" Boolean functions of n variables require circuits with *at least* $cn^{-1}2^n$ switches, c being another absolute constant; "nearly all" in this context means that the fraction of "easy to compute" functions (i.e., those computable by circuits with less than $cn^{-1}2^n$ switches) among all Boolean functions of n variables goes to 0 as n goes to ∞. Now, computing Boolean functions by circuits comprised of switches was an important technical task already in 1938; its role in our today life can hardly be overestimated — the outlined computation is nothing but what is going on in a computer. Given this observation, it is not surprising that the Shannon discovery of 1938 was the subject of countless refinements, extensions, modifications, etc., etc. What is still missing, is a *single individual example* of a "difficult to compute" Boolean function: as a matter of fact, all multivariate Boolean functions $f(x_1, ..., x_n)$ people managed to describe explicitly are computable by circuits with just *linear* in n number of switches!

1.3.1 Restricted Isometry Property and s-goodness of random matrices

There are several sufficient conditions for s-goodness, equally difficult to verify, but provably satisfied for typical random sensing matrices. The best known of them is the *Restricted Isometry Property* (RIP) defined as follows:

Definition 1.3.1 *Let k be an integer and $\delta\in(0,1)$. We say that an $m\times n$ sensing matrix A possesses the Restricted Isometry Property with parameters δ and k, $\mathrm{RIP}(\delta,k)$, if for every k-sparse $x\in\mathbf{R}^n$ one has*

$$(1-\delta)\|x\|_2^2\leq\|Ax\|_2^2\leq(1+\delta)\|x\|_2^2. \tag{1.3.1}$$

It turns out that for natural ensembles of random $m\times n$ matrices, a typical matrix from the ensemble satisfies $\mathrm{RIP}(\delta,k)$ with small δ and k "nearly as large as m," and that $\mathrm{RIP}(\frac{1}{6},2s)$ implies Nullspace condition, and more. The simplest versions of the corresponding results are as follows.

Proposition 1.3.1 *Given $\delta\in(0,\frac{1}{5}]$, with properly selected positive $c=c(\delta)$, $d=d(\delta)$, $f=f(\delta)$ for all $m\leq n$ and all positive integers k such that*

$$k\leq\frac{m}{c\ln(n/m)+d} \tag{1.3.2}$$

the probability for a random $m\times n$ matrix A with independent $\mathcal{N}(0,\frac{1}{m})$ entries to satisfy $\mathrm{RIP}(\delta,k)$ is at least $1-\exp\{-fm\}$.

Proposition 1.3.2 *Let $A\in\mathbf{R}^{m\times n}$ satisfy $\mathrm{RIP}(\delta,2s)$ for some $\delta<1/3$ and positive integer s. Then*

(i) *The pair $\left(H=\frac{s^{-1/2}}{\sqrt{1-\delta}}I_m,\|\cdot\|_2\right)$ satisfies the condition $\mathbf{Q}_2\left(s,\frac{\delta}{1-\delta}\right)$ associated with A;*

(ii) *The pair $(H=\frac{1}{1-\delta}A,\|\cdot\|_\infty)$ satisfies the condition $\mathbf{Q}_2\left(s,\frac{\delta}{1-\delta}\right)$ associated with A.*

1.3.2 Verifiable sufficient conditions for $\mathbf{Q}_q(s,\kappa)$

When speaking about verifiable sufficient conditions for a pair $(H,\|\cdot\|)$ to satisfy $\mathbf{Q}_q(s,\kappa)$, it is convenient to restrict ourselves with the case when H, same as A, is an $m\times n$ matrix, and $\|\cdot\|=\|\cdot\|_\infty$.

Proposition 1.3.3 *Let A be an $m\times n$ sensing matrix, and $s\leq n$ be a sparsity level. Given $m\times n$ matrix H and $q\in[1,\infty]$, let us set*

$$\nu_{s,q}[H]=\max_{j\leq n}\|\mathrm{Col}_j[I-H^TA]\|_{s,q}, \tag{1.3.3}$$

where $\mathrm{Col}_j[C]$ is j-th column of matrix C. Then

$$\|w\|_{s,q}\leq s^{1/q}\|H^TAw\|_\infty+\nu_{s,q}[H]\|w\|_1\ \forall w\in\mathbf{R}^n, \tag{1.3.4}$$

implying that the pair $(H,\|\cdot\|_\infty)$ satisfies the condition $\mathbf{Q}_q(s,s^{1-\frac{1}{q}}\nu_{s,q}[H])$.

Proof is immediate. Setting $V=I-H^TA$, we have

$$\begin{array}{rcl}\|w\|_{s,q}&=&\|[H^TA+V]w\|_{s,q}\leq\|H^TAw\|_{s,q}+\|Vw\|_{s,q}\\&\leq&s^{1/q}\|H^TAw\|_\infty+\sum_j|w_j|\|\mathrm{Col}_j[V]\|_{s,q}\leq s^{1/q}\|H^TA\|_\infty+\nu_{s,q}[H]\|w\|_1.\end{array}\qquad\square$$

Observe that the function $\nu_{s,q}[H]$ is an efficiently computable convex function of H, so that the set

$$\mathcal{H}^\kappa_{s,q}=\{H\in\mathbf{R}^{m\times n}:\nu_{s,q}[H]\leq s^{\frac{1}{q}-1}\kappa\} \tag{1.3.5}$$

is a computationally tractable convex set. When this set is nonempty for some $\kappa < 1/2$, every point H in this set is a contrast matrix such that $(H, \|\cdot\|_\infty)$ satisfies the condition $\mathbf{Q}_a(s,\kappa)$, that is, we can find contrast matrices making ℓ_1 minimization valid. Moreover, we can *design* contrast matrix, e.g., by minimizing over $\mathcal{H}^\kappa_{s,q}$ the function $\|H\|_{1,2}$, thus optimizing the sensitivity of the corresponding ℓ_1 recoveries to Gaussian observation noise, see items **C**, **D** in Section 1.2.5.

Explanation. The sufficient condition for s-goodness of A stated in Proposition 1.3.3 looks as coming out of thin air; in fact it is a particular case of a simple and general construction as follows. Let $f(x)$ be a real-valued convex function on $\mathbf{R}^n$, and $X \subset \mathbf{R}^n$ be a nonempty bounded polytope represented as

$$X = \{x \in \mathrm{Conv}\{g_1, ..., g_N\} : Ax = 0\},$$

where $\mathrm{Conv}\{g_1, ..., g_N\} = \{\sum_i \lambda_i g_i : \lambda \geq 0, \sum_i \lambda_i = 1\}$ is the convex hull of $g_1, ..., g_N$. Our goal is to upper-bound the maximum $\mathrm{Opt} = \max_{x\in X} f(x)$; this is a meaningful problem, since precise maximizing a convex function over a polytope typically is a computationally intractable task. Let us act as follows: clearly, for a whatever matrix H of the same sizes as A we have $\max_{x\in X} f(x) = \max_{x\in X} f([I - H^T A]x)$, since on X we have $[I - H^A]x = x$. As a result,

$$\mathrm{Opt} := \max_{x\in X} f(x) = \max_{x\in X} f([I - H^T A]x) \leq \max_{x\in\mathrm{Conv}\{g_1,...,g_N\}} f([I - H^T A]x) = \max_{j\leq N} f([I - H^T A]g_j).$$

We get a parametric, the parameter being H, upper bound on Opt, namely, the bound $\max_{j\leq N} f([I - H^T A]g_j)$. This parametric bound is convex in the parameter H, and thus is well suited for minimization over this parameter.

The result of Proposition 1.3.3 is inspired by this construction as applied to the nullspace property: given $m \times n$ sensing matrix A and setting

$$X = \{x \in \mathbf{R}^n : \|x\|_1 \leq 1, Ax = 0\} = \{x \in \mathrm{Conv}\{\pm e_1, ..., \pm e_n\} : Ax = 0\}$$

(e_i are the basic orths in $\mathbf{R}^n$), A is s-good if and only if

$$\mathrm{Opt}_s := \max_{x\in X}\{f(x) := \|x\|_{s,1}\} < 1/2;$$

A verifiable sufficient condition for this yielded by the above construction is the existence of $m \times n$ matrix H such that

$$\max_{j\leq n} \max[f([I_n - H^T A]e_j), f(-[I_n - H^T A]e_j)] < 1/2,$$

or, which is the same,

$$\max_j \|\mathrm{Col}_j[I_n - H^T A]\|_{s,1} < 1/2,$$

bringing to our attention the matrix $I - H^T A$ with varying H and the idea to express sufficient conditions for s-goodness and related properties in terms of this matrix.

1.3.3 Tractability of $\mathbf{Q}_\infty(s,\kappa)$

As we have already mentioned, the conditions $\mathbf{Q}_q(s,\kappa)$ are intractable, in the sense that we do not know how to verify whether a given pair $(H, \|\cdot\|)$ satisfies the condition. Surprisingly, this is *not* the case with the strongest of these conditions, the one with $q = \infty$. Specifically,

Proposition 1.3.4 *Let A be an $m \times n$ sensing matrix, s be a sparsity level, and $\kappa \geq 0$. Then whenever a pair $(\bar{H}, \|\cdot\|)$ satisfies the condition $\mathbf{Q}_\infty(s,\kappa)$, there exists an $m \times n$ matrix H such that*

$$\|\mathrm{Col}_j[I_n - H^T A]\|_{s,\infty} = \|\mathrm{Col}_j[I_n - H^T A]\|_\infty \leq s^{-1}\kappa,\ 1 \leq j \leq n,$$

(so that $(H,\|\cdot\|_\infty)$ satisfies $\mathbf{Q}_\infty(s,\kappa)$ by Proposition 1.3.3) and, in addition,

$$\|H^T\eta\|_\infty \leq \|\bar{H}^T\eta\| \ \forall \eta\in\mathbf{R}^m. \tag{1.3.6}$$

In addition, $m\times n$ contrast matrix H such that the pair $(H,\|\cdot\|_\infty)$ satisfies the condition $\mathbf{Q}_\infty(s,\kappa)$ with as small κ as possible can be found as follows: we consider n LP programs

$$\mathrm{Opt}_i = \min_{\nu,h}\left\{\nu : \|A^Th-e^i\|_\infty\leq\nu\right\}, \tag{$\#_i$}$$

where e^i is i-th basic orth in $\mathbf{R}^n$, find optimal solutions Opt_i, h_i to these problems, and make h_i, $i=1,...,n$, the columns of H; the corresponding value of κ is

$$\kappa_* = s\max_i \mathrm{Opt}_i.$$

Besides this, there exists a transparent alternative description of the quantities Opt_i (and thus – of κ_*); specifically,

$$\mathrm{Opt}_i = \max_x\left\{x_i : \|x\|_1\leq 1, Ax=0\right\}. \tag{1.3.7}$$

Looking at (1.3.6) and error bounds in Theorems 1.2.1, 1.2.2, 1.2.3, Proposition 1.3.4 says that

> *As far as the condition $\mathbf{Q}_\infty(s,\kappa)$ is concerned, we lose nothing when restricting ourselves with pairs $(H\in\mathbf{R}^{m\times n},\|\cdot\|_\infty)$ and contrast matrices H satisfying the condition*
>
> $$|[I_n - H^TA]_{ij}|\leq s^{-1}\kappa \tag{1.3.8}$$
>
> *implying that $(H,\|\cdot\|_\infty)$ satisfies $\mathbf{Q}_\infty(s,\kappa)$.*

A good news is that (1.3.8) is an explicit convex constraint on H (in fact, even on H and κ), so that we can solve the *design problems*, where we want to optimize a convex function of H under the requirement that $(H,\|\cdot\|_\infty)$ satisfies the condition $\mathbf{Q}_\infty(s,\kappa)$ (and, perhaps, additional convex constraints on H and κ).

1.3.3.1 Mutual Incoherence

The simplest (and up to some point in time, the only) verifiable sufficient condition for s-goodness of a sensing matrix A is expressed in terms of *mutual incoherence* of A defined as

$$\mu(A) = \max_{i\neq j}\frac{|\mathrm{Col}_i^T[A]\mathrm{Col}_j[A]|}{\|\mathrm{Col}_i[A]\|_2^2}; \tag{1.3.9}$$

this quantity is well defined whenever A has no zero columns (otherwise A is not even 1-good). Note that when A is normalized to have all columns of equal $\|\cdot\|_2$-lengths[10], $\mu(A)$ is small when the directions of distinct columns in A are nearly orthogonal. The standard related result is that

> *Whenever A and a positive integer s are such that $\frac{2\mu(A)}{1+\mu(A)}<\frac{1}{s}$, A is s-good.*

It is immediately seen that the latter condition is weaker than what we can get with the aid of (1.3.8):

[10] as far as ℓ_1 minimization is concerned, this normalization is non-restrictive: we always can enforce it by diagonal scaling of the signal underlying observations (1.1.1), and ℓ_1 minimization in scaled variables is the same as weighted ℓ_1 minimization in original variables.

Proposition 1.3.5 *Let A be an $m\times n$ matrix, and let the columns in $m\times n$ matrix H be given by*

$$\mathrm{Col}_j(H)=\frac{1}{(1+\mu(A))\|\mathrm{Col}_j(A)\|_2^2}\mathrm{Col}_j(A),\ 1\leq j\leq n.$$

Then

$$|[I_m-H^TA]_{ij}|\leq\frac{\mu(A)}{1+\mu(A)}\ \forall i,j. \tag{1.3.10}$$

In particular, when $\frac{2\mu(A)}{1+\mu(A)}<\frac{1}{s}$, A is s-good.

Proof. With H as above, the diagonal entries in $I-H^TA$ are equal to $1-\frac{1}{1+\mu(A)}=\frac{\mu(A)}{1+\mu(A)}$, while by definition of mutual incoherence the magnitudes of the off-diagonal entries in $I-H^TA$ are $\leq\frac{\mu(A)}{1+\mu(A)}$ as well, implying (1.3.10). The "in particular" claim is given by (1.3.10) combined with Proposition 1.3.3. □

1.3.3.2 From RIP to conditions $\mathbf{Q}_q(\cdot,\kappa)$

It turns out that when A is RIP(δ,k) and $q\geq 2$, it is easy to point out pairs $(H,\|\cdot\|)$ satisfying $\mathbf{Q}_q(t,\kappa)$ with a desired $\kappa>0$ and properly selected t:

Proposition 1.3.6 *Let A be an $m\times n$ sensing matrix satisfying* RIP$(\delta,2s)$ *with some s and some $\delta\in(0,1)$, and let $q\in[2,\infty]$ and $\kappa>0$ be given. Then*

(i) *Whenever a positive integer t satisfies*

$$t\leq\min\left[\left[\frac{\kappa\sqrt{1-\delta}}{\delta\sqrt{1+\delta}}\right]^{\frac{q}{q-1}},s^{\frac{q-2}{q-1}}\right]s^{\frac{q}{2(q-1)}}, \tag{1.3.11}$$

the pair $(H=\frac{t^{-1/2}}{\sqrt{1-\delta}}I_m,\|\cdot\|_2)$ satisfies $\mathbf{Q}_q(t,\kappa)$;

(ii) *Whenever a positive integer t satisfies*

$$t\leq\min\left[\left[\frac{\kappa(1-\delta)}{\delta}\right]^{\frac{q}{q-1}},s^{\frac{q-2}{2q-2}}\right]s^{\frac{q}{2(q-1)}}, \tag{1.3.12}$$

the pair $(H=\frac{s^{\frac{1}{2}}t^{-\frac{1}{q}}}{1-\delta}A,\|\cdot\|_\infty)$ satisfies $\mathbf{Q}_q(t,\kappa)$.

The most important consequence of Proposition 1.3.6 deals with the case of $q=\infty$ and states that *when s-goodness of a sensing matrix A can be ensured by difficult to verify condition* RIP$(\delta,2s)$ *with, say, $\delta=0.2$, the somehow worse level of sparsity, $t=O(1)\sqrt{s}$ with properly selected absolute constant $O(1)$ can be certified via condition $\mathbf{Q}_\infty(t,\frac{1}{3})$ – there exists pair $(H,\|\cdot\|_\infty)$ satisfying this condition.* The point is that by Proposition 1.3.4, if the condition $\mathbf{Q}_\infty(t,\frac{1}{3})$ can at all be satisfied, a pair $(H,\|\cdot\|_\infty)$ satisfying this condition can be found efficiently.

Unfortunately, the significant "dropdown" in the level of sparsity when passing from unverifiable RIP to verifiable $\mathbf{Q}_\infty$ is inevitable; this bad news is what is on our agenda now.

1.3.3.3 Limits of performance of verifiable sufficient conditions for goodness

Proposition 1.3.7 *Let A be an $m\times n$ sensing matrix which is "essentially non-square," specifically, such that $2m\leq n$, and let $q\in[1,\infty]$. Whenever a positive integer s and an $m\times n$ matrix H are linked by the relation*

$$\|\mathrm{Col}_j[I_n-H^TA]\|_{s,q}<\frac{1}{2}s^{\frac{1}{q}-1},\ 1\leq j\leq n, \tag{1.3.13}$$

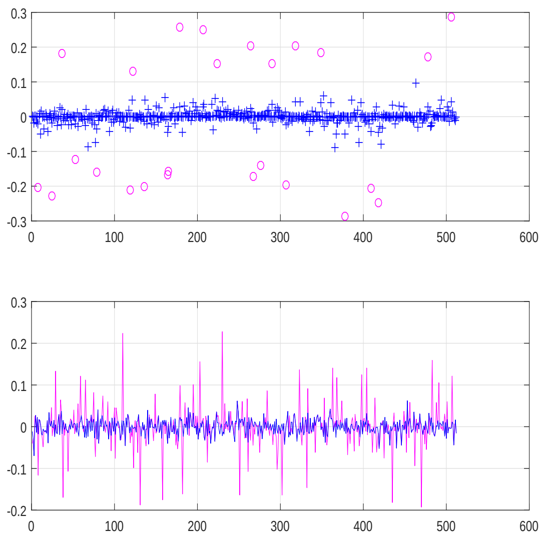


Figure 1.4: Erroneous ℓ_1 recovery of 25-sparse signal, no observation noise. Magenta: true signal, blue: ℓ_1 recovery. Top: frequency domain, bottom: time domain.

one has

$$s \leq \sqrt{2m}. \tag{1.3.14}$$

As a result, sufficient condition for the validity of $\mathbf{Q}_q(s,\kappa)$ with $\kappa < 1/2$ from Proposition 1.3.3 can never be satisfied when $s > \sqrt{2m}$. Similarly, the verifiable sufficient condition $\mathbf{Q}_\infty(s,\kappa)$, $\kappa < 1/2$ for s-goodness of A cannot be satisfied when $s > \sqrt{2m}$.

We see that unless A is "nearly square," our (same as all other known to us) verifiable sufficient conditions for s-goodness are unable to justify this property for "large" s. This unpleasant fact is in full accordance with the already mentioned fact that no individual provably s-good "essentially nonsquare" $m \times n$ matrices with $s \geq O(1)\sqrt{m}$ are known.

Matrices for which our verifiable sufficient conditions do establish s-goodness with $s \leq O(1)\sqrt{m}$ do exist.

How it works: Numerical illustration. Let us apply our machinery to the 256×512 randomly selected submatrix A of the matrix of 512×512 Inverse Discrete Cosine Transform which we used in experiments reported on Figure 1.3. These experiments exhibit nice performance of ℓ_1 minimization when recovering sparse (even nearly sparse) signals with as much as 64 nonzeros. *In fact, the level of goodness of A is at most 24*, as is witnessed by Figure 1.4.

> In order to upper-bound the level of goodness of a matrix A, one can try to maximize the convex function $\|w\|_{s,1}$ over the set $W = \{w : Aw = 0, \|w\|_1 \leq 1\}$; if, for a given s, the maximum of $\|\cdot\|_{s,1}$ over W is $\geq 1/2$, the matrix is not s-good – it does not possess the Nullspace property. Now, while global maximization of the convex function $\|w\|_{s,1}$ over W is difficult, we can try to find suboptimal solutions as follows: let us start with a vector $w_1 \in W$ of $\|\cdot\|_1$-norm 1, and let u^1 be obtained from w_1 by replacing the s largest in magnitude entries in w_1 by the signs of these entries and zeroing out all other entries, so that $w_1^T u^1 = \|w_1\|_{s,1}$. After u^1 is found, let us solve the LO program $\max_w \{[u^1]^T w : w \in W\}$. w_1 is a feasible solution to this problem, so that for the optimal solution w_2 to it we have $[u^1]^T w_2 \geq [u^1]^T w_1 = \|w_1\|_{s,1}$; this inequality, by virtue of what u^1 is, implies that $\|w_2\|_{s,1} \geq \|w_1\|_{s,1}$ and by construction $w_2 \in W$. We now can iterate the construction, with w_2 in the role of w_1, to get $w_3 \in W$ with $\|w_3\|_{s,1} \geq \|w_2\|_{s,1}$; proceeding in this way, we generate a sequence of points from W with monotonically increasing value of the

> objective $\|\cdot\|_{s,1}$ we want to maximize. Usually, people terminate this recurrence either when the achieved value of the objective becomes $\geq 1/2$ (then we know for sure that A is not s-good, and can proceed to investigating s-goodness for a smaller value of s) or when the recurrence becomes stuck – the observed progress in the objective falls below a given threshold, say, 1.e-6; after it happens we can restart this process from a new randomly selected in W starting point, after getting stuck, restart again, etc., etc., until exhausting our time budget. The output of the process is the best – with the largest $\|\cdot\|_{s,1}$ – of the points from W we have generated. Applying this approach to the matrix A in question, in a couple of minutes it turns out that the matrix is at most 24-good.

One can ask how happens that experiments with recovering 64-sparse signals went fine, when in fact some 25-sparse signals cannot be recovered by ℓ_1 minimization even in the ideal noiseless case. The answer is simple: in our experiments, we dealt with *randomly selected* signals, and, as it typically is the case, randomly selected data are much nicer, whatever be the purpose of a numerical experiment, that the worst-case data.

It is interesting to understand also which goodness we can certify with our verifiable sufficient conditions. Computation shows that the fully verifiable (and strongest in our scale of sufficient conditions for s-goodness) condition $\mathbf{Q}_\infty(s,\varkappa)$ can be satisfied with $\varkappa < 1/2$ when s is as large as 7 and $\varkappa = 0.4887$, and can*not* be satisfied with $\varkappa < 1/2$ when $s = 8$. As about Mutual Incoherence, it can justify just 3-goodness, no more. We hardly could be happy with the resulting bounds – goodness at least 7 and at most 24; however, it could be worse...

1.3.4 Proofs

1.3.4.1 Proof of Proposition 1.3.1

1^0. Assuming $k \leq m$ and selecting a set I of k distinct from each other indexes from $\{1,...,n\}$, consider an $m\times k$ submatrix A_I of A comprised of columns with indexes from I, and let u be a unit vector in $\mathbf{R}^k$. The entries in the vector $m^{1/2}A_Iu$ are independent $\mathcal{N}(0,1)$ random variables, so that for the random variable $\zeta_u = \sum_{i=1}^m (m^{1/2}A_Iu)_i^2$ and $\gamma \in (-1/2,1/2)$ it holds (in what follows,expectations and probabilities are taken w.r.t. our ensemble of random A's)

$$\ln\left(\mathbf{E}\{\exp\{\gamma\zeta\}\}\right) = m\ln\left(\frac{1}{\sqrt{2\pi}}\int \mathrm{e}^{\gamma t^2-\frac{1}{2}t^2}ds\right) = -\frac{m}{2}\ln(1-2\gamma).$$

Given $\alpha\in(0,0.1]$ and selecting γ in such a way that $1-2\gamma = \frac{1}{1+\alpha}$, we get $0<\gamma<1/2$ and therefore

$$\begin{array}{l}\mathrm{Prob}\{\zeta_u > m(1+\alpha)\} \leq \mathbf{E}\{\exp\{\gamma\zeta_u\}\}\exp\{-m\gamma(1+\alpha)\} = \exp\{-\frac{m}{2}\ln(1-2\gamma)-m\gamma(1+\alpha)\}\\ = \exp\{\frac{m}{2}\left[\ln(1+\alpha)-\alpha\right]\} \leq \exp\{-\frac{m}{5}\alpha^2\},\end{array}$$

and similarly, selecting γ in such a way that $1-2\gamma=\frac{1}{1-\alpha}$, we get $-1/2<\gamma<0$ and therefore

$$\begin{array}{l}\mathrm{Prob}\{\zeta_u < m(1-\alpha)\} \leq \mathbf{E}\{\exp\{\gamma\zeta_u\}\}\exp\{-m\gamma(1-\alpha)\} = \exp\{-\frac{m}{2}\ln(1-2\gamma)-m\gamma(1-\alpha)\}\\ = \exp\{\frac{m}{2}\left[\ln(1-\alpha)+\alpha\right]\} \leq \exp\{-\frac{m}{5}\alpha^2\},\end{array}$$

and we end up with

$$u\in\mathbf{R}^k, \|u\|_2=1 \Rightarrow \left\{\begin{array}{l}\mathrm{Prob}\{A: \|A_Iu\|_2^2 > 1+\alpha\} \leq \exp\{-\frac{m}{5}\alpha^2\}\\ \mathrm{Prob}\{A: \|A_Iu\|_2^2 < 1-\alpha\} \leq \exp\{-\frac{m}{5}\alpha^2\}\end{array}\right. \tag{1.3.15}$$

2^0. Same as above, let $\alpha \in (0, 0.1]$, let

$$M = 1 + 2\alpha, \epsilon = \frac{\alpha}{2(1+2\alpha)},$$

and let us build an ϵ-net on the unit sphere S in $\mathbf{R}^k$ as follows. We start with a point $u_1 \in S$; after $\{u_1, ..., u_t\} \subset S$ is already built, we check whether there is a point in S at the $\|\cdot\|_2$-distance from all points of the set $> \epsilon$. If it is the case, we add such a point to the net built so far and proceed with building the net, otherwise we terminate with the net $\{u_1, ..., u_t\}$. By compactness of S and due to $\epsilon > 0$, this process eventually terminates; upon termination, we have at our disposal collection $\{u_1, ..., u_N\}$ of unit vectors such that every two of them are at the $\|\cdot\|_2$-distance $> \epsilon$ from each other, and every point from S is at the distance at most ϵ from some point of the collection. We claim that the cardinality N of the resulting set can be bounded as

$$N \leq \left[\frac{2+\epsilon}{\epsilon}\right]^k = \left[\frac{4+9\alpha}{\alpha}\right]^k \leq \left(\frac{5}{\alpha}\right)^k. \tag{1.3.16}$$

Indeed, the interiors of the $\|\cdot\|_2$-balls of radius $\epsilon/2$ centered at the points $u_1, ..., u_N$ are mutually disjoint, and their union is contained in the $\|\cdot\|_2$-ball of radius $1 + \epsilon/2$ centered at the origin; comparing the volume of the union and the one of the ball, we arrive at (1.3.16).

3^0. Consider event E comprised of all realizations of A such that for all k-element subsets I of $\{1, ..., n\}$ and all $t \leq n$ it holds

$$1 - \alpha \leq \|A_I u_t\|_2^2 \leq 1 + \alpha. \tag{1.3.17}$$

By (1.3.15) and the union bound,

$$\text{Prob}\{A \notin E\} \leq 2N \left(\begin{array}{c} n \\ k \end{array}\right) \exp\{-\frac{m}{5}\alpha^2\}. \tag{1.3.18}$$

We claim that

$$A \in E \Rightarrow (1 - 2\alpha) \leq \|A_I u\|_2^2 \leq 1 + 2\alpha \;\forall (I \subset \{1, ..., n\} : \text{Card}(I) = k) \forall (u \in \mathbf{R}^k : \|u\|_2 = 1). \tag{1.3.19}$$

Indeed, let $A \in E$, let us fix $I \in \{1, ..., n\}$, $\text{Card}(I) = k$, and let M be the maximal value of the quadratic form $f(u) = u^T A_I^T A_I u$ on the unit $\|\cdot\|_2$-ball B, centered at the origin, in $\mathbf{R}^k$. In this ball, f is Lipschitz continuous with constant $2M$ w.r.t. $\|\cdot\|_2$; denoting by $\bar{u}$ a maximizer of the form on B, we lose nothing when assuming that $\bar{u}$ is a unit vector. Now let u_s be the point of our net which is at the $\|\cdot\|_2$-distance from $\bar{u}$ at most ϵ. We have

$$M = f(\bar{u}) \leq f(u_s) + 2M\epsilon \leq 1 + \alpha + 2M\epsilon,$$

whence

$$M \leq \frac{1+\alpha}{1-2\epsilon} = 1 + 2\alpha,$$

implying the right inequality in (1.3.19). Now let u be unit vector in $\mathbf{R}^k$, and u_s be a point in the net at the $\|\cdot\|$-distance $\leq \epsilon$ from u. We have

$$f(u) \geq f(u_s) - 2M\epsilon \geq 1 - \alpha - 2\frac{1+\alpha}{1-2\epsilon}\epsilon = 1 - 2\alpha,$$

justifying the first inequality in (1.3.19).

The bottom line is:

$$\begin{array}{l}\delta\in(0,0.2], 1\leq k\leq n\\ \Rightarrow \mathrm{Prob}\{A: A \hbox{ does not satisfy RIP}(\delta,k)\}\leq 2\underbrace{\left(\frac{10}{\delta}\right)^k}_{\leq(\frac{20}{\delta})^k}\left(\begin{array}{c}n\\k\end{array}\right)\exp\{-\frac{m\delta^2}{20}\}.\end{array}\tag{1.3.20}$$

Indeed, setting $\alpha=\delta/2$, we have seen that whenever $A\not\in E$, we have $(1-\delta)\leq\|Au\|_2^2\leq(1+\delta)$ for all unit k-sparse u, which is nothing but RIP(δ,k); with this in mind, (1.3.20) follows from (1.3.18) and (1.3.16).

4^0. It remains to verify that with properly selected, depending solely on δ, positive quantities c,d,f, for every $k\geq 1$ satisfying (1.3.2) the right hand side in (1.3.20) is at most $\exp\{-fm\}$. Passing to logarithms, our goal is to ensure the relation

$$\begin{array}{c}G:=a(\delta)m-b(\delta)k-\ln\left(\begin{array}{c}n\\k\end{array}\right)\geq mf(\delta)>0\\ \left[a(\delta)=\frac{\delta^2}{20}, b(\delta)=\ln\left(\frac{20}{\delta}\right)\right]\end{array}\tag{1.3.21}$$

provided that $k\geq 1$ satisfies (1.3.2).

Let k satisfy (1.3.2) with some c,d to be specified later, and let $y=k/m$. Assuming $d\geq 3$, we have $0\leq y\leq 1/3$. Now, it is well known that

$$C:=\ln\left(\begin{array}{c}n\\k\end{array}\right)\leq n\left[\frac{k}{n}\ln(\frac{n}{k})+\frac{n-k}{n}\ln(\frac{n}{n-k})\right],$$

whence

$$\begin{array}{l}C\leq n\big[\frac{m}{n}y\ln(\frac{n}{my})+\frac{n-k}{n}\underbrace{\ln(1+\frac{k}{n-k})}_{\leq\frac{k}{n-k}}\big]\\ \leq n\big[\frac{m}{n}y\ln(\frac{n}{my})+\frac{k}{n}\big]=m\big[y\ln(\frac{n}{my})+y\big]\leq 2my\ln(\frac{n}{my})\end{array}$$

(recall that $n\geq m$ and $y\leq 1/3$). It follows that

$$G=a(\delta)m-b(\delta)k-C\geq a(\delta)m-b(\delta)ym-2my\ln(\frac{n}{my})=m\underbrace{\left[a(\delta)-b(\delta)y-2y\ln(\frac{n}{m})-2y\ln(\frac{1}{y})\right]}_{H},$$

and all we need is to select c,d in such a way that (1.3.2) would imply that $H\geq f$ with some positive $f=f(\delta)$. This is immediate: we can find $u(\delta)>0$ such that when $0\leq y\leq u(\delta)$, we have $2y\ln(1/y)+b(\delta)y\leq\frac{1}{3}a(\delta)$; selecting $d(\delta)\geq 3$ large enough, (1.3.2) would imply $y\leq u(\delta)$, and thus would imply

$$H\geq\frac{2}{3}a(\delta)-2y\ln(\frac{n}{m}).$$

Now we can select $c(\delta)$ large enough for (1.3.2) to ensure that $2y\ln(\frac{n}{m})\leq\frac{1}{3}a(\delta)$. With just specified c, d, (1.3.2) implies that $H\geq\frac{1}{3}a(\delta)$, and we can take the latter quantity as $f(\delta)$. □

1.3.4.2 Proof of Propositions 1.3.2, 1.3.6

Let us prove Proposition 1.3.6; as a byproduct of our reasoning, we shall prove Proposition 1.3.2 as well.

Let $x \in \mathbf{R}^n$, and let $x^1, ..., x^q$ be obtained from x by the following construction: x^1 is obtained from x by zeroing all but the s largest in magnitude entries; x^2 is obtained by the same procedure applied to $x - x^1$, x^3 – by the same procedure applied to $x - x^1 - x^2$, and so on; the process is terminated at the first step q when it happens that $x = x^1 + ... + x^q$. Note that for $j \geq 2$ we have $\|x^j\|_\infty \leq s^{-1}\|x^{j-1}\|_1$ and $\|x^j\|_1 \leq \|x^{j-1}\|_1$, whence also $\|x^j\|_2 \leq \sqrt{\|x^j\|_\infty \|x^j\|_1} \leq s^{-1/2}\|x^{j-1}\|_1$. It is easily seen that if A is RIP$(\delta, 2s)$, then for every two s-sparse vectors u, v with non-overlapping supports we have

$$|v^T A^T A u| \leq \delta \|u\|_2 \|v\|_2. \tag{$*$}$$

> Indeed, for s-sparse u, v, let I be the index set of cardinality $\leq 2s$ containing the supports of u and v, so that, denoting by A_I the submatrix of A comprised of columns with indexes from I, we have $v^T A^T A u = v_I^T [A_I^T A_I] u_I$. By RIP, the eigenvalues $\lambda_i = 1 + \mu_i$ of the symmetric matrix $Q = A_I^T A_I$ are in-between $1 - \delta$ and $1 + \delta$; representing u_I and v_I by vectors w, z of their coordinates in the orthonormal eigenbasis of Q, we get $|v^T A^T A u| = |\sum_i \lambda_i w_i z_i| = |\sum_i w_i z_i + \sum_i \mu_i w_i z_i| \leq |w^T z| + \delta \|w\|_2 \|z\|_2$. It remains to note that $w^T z = u_I^T v_I = 0$ and $\|w\|_2 = \|u\|_2$, $\|z\|_2 = \|v\|_2$.

(i): We have

$$\begin{array}{rl}
& \|Ax^1\|_2 \|Ax\|_2 \geq [x^1]^T A^T A x = \|Ax^1\|_2^2 - \sum_{j=2}^q [x^1]^T A^T A x^j \\
& \geq \|Ax^1\|_2^2 - \delta \sum_{j=2}^q \|x^1\|_2 \|x^j\|_2 \text{ [by } (*)] \\
& \geq \|Ax^1\|_2^2 - \delta s^{-1/2} \|x^1\|_2 \sum_{j=2}^q \|x^{j-1}\|_1 \geq \|Ax^1\|_2^2 - \delta s^{-1/2} \|x^1\|_2 \|x\|_1 \\
\Rightarrow & \|Ax^1\|_2^2 \leq \|Ax^1\|_2 \|Ax\|_2 + \delta s^{-1/2} \|x^1\|_2 \|x\|_1 \\
\Rightarrow & \|x^1\|_2 = \frac{\|x^1\|_2}{\|Ax^1\|_2^2} \|Ax^1\|_2^2 \leq \frac{\|x^1\|_2}{\|Ax^1\|_2} \|Ax\|_2 + \delta s^{-1/2} \left(\frac{\|x^1\|_2}{\|Ax^1\|_2}\right)^2 \|x\|_1 \\
\Rightarrow & \|x\|_{s,2} = \|x^1\|_2 \leq \frac{1}{\sqrt{1-\delta}} \|Ax\|_2 + \frac{\delta s^{-1/2}}{1-\delta} \|x\|_1 \text{ [by RIP}(\delta, 2s)]
\end{array}$$

and we see that the pair $\left(H = \frac{s^{-1/2}}{\sqrt{1-\delta}} I_m, \|\cdot\|_2\right)$ satisfies $\mathbf{Q}_2(s, \frac{\delta}{1-\delta})$, as claimed in (i).

In addition, the relation after the first $\Rightarrow$ implies that

$$\|Ax^1\|_2 \leq \|Ax\|_2 + \delta s^{-1/2} \left[\frac{\|x_1\|_2}{\|Ax^1\|_2}\right] \|x\|_1.$$

By RIP, the left hand side in this inequality is $\geq \|x^1\|_2 \sqrt{1-\delta}$, while the ratio of norms in the right hand side is $\leq \frac{1}{\sqrt{1-\delta}}$, so that

$$\|x\|_{s,2} = \|x^1\|_2 \leq \frac{1}{\sqrt{1-\delta}} \|Ax\|_2 + \frac{\delta s^{-1/2}}{1-\delta} \|x\|_1,$$

implying Proposition 1.3.2.i. Moreover, when $q \geq 2$, $\kappa > 0$ and integer $t \geq 1$ satisfy $t \leq s$ and $\kappa t^{1/q-1} \geq \frac{\delta s^{-1/2}}{1-\delta}$, we have

$$\|x\|_{t,q} \leq \|x\|_{s,q} \leq \|x\|_{s,2} \leq \frac{1}{\sqrt{1-\delta}} \|Ax\|_2 + \kappa t^{1/q-1} \|x\|_1,$$

or, equivalently,

$$\begin{array}{rl}
& 1 \leq t \leq \min \left[\left[\frac{\kappa(1-\delta)}{\delta}\right]^{\frac{q}{q-1}}, s^{\frac{q-2}{2q-2}} \right] s^{\frac{q}{2(q-1)}} \\
\Rightarrow & (H = \frac{t^{-1/2}}{\sqrt{1-\delta}} I_m, \|\cdot\|_2) \text{ satisfies } \mathbf{Q}_q(t, \kappa),
\end{array}$$

as required in item (i) of Proposition 1.3.6.

(ii): We have

$$\begin{array}{rl} & \|x^1\|_1\|A^TAx\|_\infty \geq [x^1]^TA^TAx = \|Ax^1\|_2^2 - \sum_{j=2}^q [x^1]^TA^TAx^j \\ & \geq \|Ax^1\|_2^2 - \delta s^{-1/2}\|x^1\|_2\|x\|_1 \text{ [exactly as above]} \\ \Rightarrow & \|Ax^1\|_2^2 \leq \|x^1\|_1\|A^TAx\|_\infty + \delta s^{-1/2}\|x^1\|_2\|x\|_1 \\ \Rightarrow & (1-\delta)\|x^1\|_2^2 \leq \|x^1\|_1\|A^TAx\|_\infty + \delta s^{-1/2}\|x^1\|_2\|x\|_1 \text{ [by RIP}(\delta, 2s)] \\ & \leq s^{1/2}\|x^1\|_2\|A^TAx\|_\infty + \delta s^{-1/2}\|x^1\|_2\|x\|_1 \\ \Rightarrow & \|x\|_{s,2} = \|x^1\|_2 \leq \frac{s^{1/2}}{1-\delta}\|A^TAx\|_\infty + \frac{\delta}{1-\delta}s^{-1/2}\|x\|_1, \end{array}$$

and we see that the pair $\left(H = \frac{1}{1-\delta}A, \|\cdot\|_\infty\right)$ satisfies the condition $\mathbf{Q}_2\left(s, \frac{\delta}{1-\delta}\right)$, as required in Proposition 1.3.2.ii.

In addition, the inequality after the second ⇒ implies that

$$\|x^1\|_2 \leq \frac{1}{1-\delta}\left[s^{1/2}\|A^TAx\|_\infty + \delta s^{-1/2}\|x\|_1\right],$$

Consequently, when $q \geq 2$, $\kappa > 0$ and integer $t \geq 1$ satisfy $t \leq s$ and $\kappa t^{1/q-1} \geq \frac{\delta}{1-\delta}s^{-1/2}$, we have

$$\|x\|_{t,q} \leq \|x\|_{s,q} \leq \|x\|_{s,2} \leq \frac{1}{1-\delta}s^{1/2}\|A^TAx\|_\infty + \kappa t^{1/q-1}\|x\|_1,$$

or, equivalently,

$$1 \leq t \leq \min\left[\left[\frac{\kappa(1-\delta)}{\delta}\right]^{\frac{q}{q-1}}, s^{\frac{q-2}{2q-2}}\right] s^{\frac{q}{2(q-1)}} \Rightarrow (H = \frac{s^{\frac{1}{2}}t^{-\frac{1}{q}}}{1-\delta}A, \|\cdot\|_\infty) \text{ satisfies } \mathbf{Q}_q(t,\kappa),$$

as required in item (ii) of Proposition 1.3.6. □

1.3.4.3 Proof of Proposition 1.3.4

(i): Let $\bar{H} \in \mathbf{R}^{m\times N}$ and $\|\cdot\|$ satisfy $\mathbf{Q}_\infty(s,\kappa)$. Then for every $k \leq n$ we have

$$|x_k| \leq \|\bar{H}^TAx\| + s^{-1}\kappa\|x\|_1,$$

or, which is the same by homogeneity,

$$\min_x \left\{\|\bar{H}^TAx\| - x_k : \|x\|_1 \leq 1\right\} \geq -s^{-1}\kappa.$$

In other words, the optimal value Opt_k of the conic optimization problem[11]

$$\mathrm{Opt}_k = \min_{x,t}\left\{t - [e^k]^Tx : \|\bar{H}^TAx\| \leq t, \|x\|_1 \leq 1\right\},$$

where $e^k \in \mathbf{R}^n$ is k-th basic orth, is $\geq -s^{-1}\kappa$. Since the problem clearly is strictly feasible, this is the same as to say that the dual problem

$$\max_{\mu\in\mathbf{R}, g\in\mathbf{R}^n, \eta\in\mathbf{R}^N}\left\{-\mu : A^T\bar{H}\eta + g = e^k, \|g\|_\infty \leq \mu, \|\eta\|_* \leq 1\right\},$$

where $\|\cdot\|_*$ is the norm conjugate to $\|\cdot\|$:

$$\|u\|_* = \max_{\|h\|\leq 1} h^Tu$$

[11] For summary on conic programming, see Section 4.1.2.

has a feasible solution with the value of the objective $\geq -s^{-1}\kappa$. It follows that there exists $\eta = \eta^k$ and $g = g^k$ such that

$$\begin{array}{l} (a): e^k = A^T h^k + g^k, \\ (b): h^k := \bar{H}\eta^k, \|\eta^k\|_* \leq 1, \\ (c): \|g^k\|_\infty \leq s^{-1}\kappa. \end{array} \tag{1.3.22}$$

Denoting $H = [h^1, ..., h^n]$, $V = I - H^T A$, we get

$$\mathrm{Col}_k[V^T] = e^k - A^T h^k = g^k,$$

implying that $\|\mathrm{Col}_k[V^T]\|_\infty \leq s^{-1}\kappa$. Since the latter inequality it is true for all $k \leq n$, we conclude that

$$\|\mathrm{Col}_k[V]\|_{s,\infty} = \|\mathrm{Col}_k[V]\|_\infty \leq s^{-1}\kappa,\ 1 \leq k \leq n,$$

whence, by Proposition 1.3.3, $(H, \|\cdot\|_\infty)$ satisfies $\mathbf{Q}_\infty(s,\kappa)$. Moreover, for every $\eta \in \mathbf{R}^m$ and every $k \leq n$ we have, in view of (b) and (c),

$$|[h^k]^T\eta| = |[\eta^k]^T \bar{H}^T \eta| \leq \|\eta^k\|_* \|\bar{H}^T\eta\|,$$

whence $\|H^T\eta\|_\infty \leq \|\bar{H}^T\eta\|$.

Now let us prove the "In addition" part of Proposition. Let $H = [h_1, ..., h_n]$ be the contrast matrix specified in this part. We have

$$|[I_m - H^T A]_{ij}| = |[[e^i]^T - h_i^T A]_j| \leq \|[e^i]^T - h_i^T A\|_\infty = \|e^i - A^T h_i\|_\infty \leq \mathrm{Opt}_i,$$

implying by Proposition 1.3.3 that $(H, \|\cdot\|_\infty)$ does satisfy the condition $\mathbf{Q}_\infty(s, \kappa_*)$ with $\kappa_* = s\max_i \mathrm{Opt}_i$. Now assume that there exists a matrix H' which, taken along with some norm $\|\cdot\|$, satisfies the condition $\mathbf{Q}_\infty(s,\kappa)$ with $\kappa < \kappa_*$, and let us lead this assumption to a contradiction. By the already proved first part of Proposition 1.3.4, our assumption implies that there exists $m \times n$ matrix $\bar{H} = [\bar{h}_1, ..., \bar{h}_n]$ such that $\|\mathrm{Col}_j[I_n - \bar{H}^T A]\|_\infty \leq s^{-1}\kappa$ for all $j \leq n$, implying that $|[[e^i]^T - \bar{h}_i^T A]_j| \leq s^{-1}\kappa$ for all i and j, or, which is the same, $\|e^i - A^T\bar{h}_i\|_\infty \leq s^{-1}\kappa$ for all i. Due to the origin of Opt_i, we have $\mathrm{Opt}_i \leq \|e^i - A^T\bar{h}_i\|_\infty$ for all i, and we arrive at $s^{-1}\kappa_* = \max_i \mathrm{Opt}_i \leq s^{-1}\kappa$, that is, $\kappa_* \leq \kappa$, which is a desired contradiction.

It remains to prove (1.3.7), which is just an exercise on LP duality: denoting by $\mathbf{e}$ n-dimensional all-ones vector, we have

$$\begin{array}{rcl} \mathrm{Opt}_i & := & \min_h \|e^i - A^T h\|_\infty = \min_{h,t}\left\{t : e^i - A^T h \leq t\mathbf{e}, A^T h - e^i \leq t\mathbf{e}\right\} \\ & = & \max_{\lambda,\mu}\left\{\lambda_i - \mu_i : \lambda, \mu \geq 0, A[\lambda - \mu] = 0, \sum_i \lambda_i + \sum_i \mu_i = 1\right\} \\ & & \text{[LP duality]} \\ & = & \max_{x:=\lambda-\mu}\left\{x_i : Ax = 0, \|x\|_1 \leq 1\right\} \end{array}$$

where the concluding equality follows from the fact that vectors x representable as $\lambda - \mu$ with $\lambda, \mu \geq 0$ satisfying $\|\lambda\|_1 + \|\mu\|_1 = 1$ are exactly vectors x with $\|x\|_1 \leq 1$. □

1.3.4.4 Proof of Proposition 1.3.7

Let H satisfy (1.3.13). Since $\|v\|_{s,1} \leq s^{1-1/q}\|v\|_{s,q}$, it follows that H satisfies for some $\alpha < 1/2$ the condition

$$\|\mathrm{Col}_j[I_n - H^T A]\|_{s,1} \leq \alpha,\ 1 \leq j \leq n. \tag{1.3.23}$$

whence, as we know,

$$\|x\|_{s,1} \leq s\|H^T Ax\|_\infty + \alpha\|x\|_1\ \forall x \in \mathbf{R}^n$$

It follows that $s \leq m$, since otherwise there exists a nonzero s-sparse vector x with $Ax = 0$; for this x, the inequality above cannot hold true.

Let us set $\bar{n} = 2m$, so that $\bar{n} \leq n$, and let $\bar{H}$ and $\bar{A}$ be the $m \times \bar{n}$ matrices comprised of the first $2m$ columns of H, respectively, A. Relation (1.3.23) implies that the matrix $V = I_{\bar{n}} - \bar{H}^T\bar{A}$ satisfies

$$\|\mathrm{Col}_j[V]\|_{s,1} \leq \alpha < 1/2, 1 \leq j \leq \bar{n}. \tag{1.3.24}$$

Now, $V = I_{\bar{n}} - \bar{H}^T\bar{A}$, and since the rank of $\bar{H}^T\bar{A}$ is $\leq m$, at least $\bar{n} - m$ singular values of V are ≥ 1, and therefore the squared Frobenius norm $\|V\|_F^2$ of V is at least $\bar{n} - m$. On the other hand, we can upper-bound this squared norm as follows. Observe that for every $\bar{n}$-dimensional vector f one has

$$\|f\|_2^2 \leq \max\left[\frac{\bar{n}}{s^2}, 1\right] \|f\|_{s,1}^2. \tag{1.3.25}$$

Indeed, by homogeneity it suffices to verify the inequality when $\|f\|_{s,1} = 1$; besides, we can assume w.l.o.g. that the entries in f are nonnegative, and that $f_1 \geq f_2 \geq \ldots \geq f_{\bar{n}}$. We have $f_s \leq \|f\|_{s,1}/s = \frac{1}{s}$; in addition, $\sum_{j=s+1}^{\bar{n}} f_j^2 \leq (\bar{n}-s)f_s^2$. Now, due to $\|f\|_{s,1} = 1$, for fixed $f_s \in [0, 1/s]$ we have

$$\sum_{j=1}^{s} f_j^2 \leq f_s^2 + \max_t\{\sum_{j=1}^{s-1} t_s^2 : t_j \geq f_s, j \leq s-1, \sum_{j=1}^{s-1} t_j = 1 - f_s\}.$$

The maximum in the right hand side is the maximum of a convex function over a bounded polytope; it is achieved at an extreme point, that is, at a point where one of the t_j is equal to $1 - (s-1)f_s$, and all remaining t_j are equal to f_s. As a result,

$$\sum_j f_j^2 \leq \left[(1-(s-1)f_s)^2 + (s-1)f_s^2\right] + (\bar{n}-s)f_s^2 \leq (1-(s-1)f_s)^2 + (\bar{n}-1)f_s^2.$$

The right hand side in the latter inequality is convex in f_s and thus achieves its maximum over the range $[0, 1/s]$ of allowed values of f_s at an endpoint, yielding $\sum_j f_j^2 \leq \max[1, \bar{n}/s^2]$, as claimed.

Applying (1.3.25) to the columns of V and recalling that $\bar{n} = 2m$, we get

$$\|V\|_F^2 = \sum_{j=1}^{2m} \|\mathrm{Col}_j[V]\|_2^2 \leq \max[1, 2m/s^2] \sum_{j=1}^{2m} \|\mathrm{Col}_j[V]\|_{s,1}^2 \leq 2\alpha m \max[1, 2m/s^2].$$

The left hand side in this inequality, as we remember, is $\geq \bar{n} - m = m$, and we arrive at

$$m \leq 2\alpha m \max[1, 2m/s^2].$$

Since $\alpha < 1/2$, this inequality implies $2m/s^2 \geq 1$, whence $s \leq \sqrt{2m}$.

It remains to prove that when $m \leq n/2$, the condition $\mathbf{Q}_\infty(s,\kappa)$ with $\kappa < 1/2$ can be satisfied only when $s \leq \sqrt{2m}$. This is immediate: by Proposition 1.3.4, assuming $\mathbf{Q}_\infty(s,\kappa)$ satisfiable, there exists $m \times n$ contrast matrix H such that $|[I_n - H^TA]_{ij}| \leq \kappa/s$ for all i, j, which, by the already proved part of Proposition 1.3.7, is impossible when $s > \sqrt{2m}$. □

1.4 Exercises for Lecture 1

Exercise 1.1 k-th Hadamard matrix, $\mathcal{H}_k$ (here k is nonnegative integer) is the $n_k \times n_k$ matrix, $n_k = 2^k$, given by the recurrence

$$\mathcal{H}_0 = [1]; \mathcal{H}_{k+1} = \left[\begin{array}{c|c} \mathcal{H}_k & \mathcal{H}_k \\ \hline \mathcal{H}_k & -\mathcal{H}_k \end{array}\right] \tag{1.4.1}$$

In the sequel, we assume that $k > 0$. Now goes the exercise:

1. *Check that $\mathcal{H}_k$ is symmetric matrix with entries ± 1, and columns of the matrix are mutually orthogonal, so that $\mathcal{H}_k/\sqrt{n_k}$ is an orthogonal matrix.*

2. *Check that when $k > 0$, $\mathcal{H}_k$ has just two distinct eigenvalues, $\sqrt{n_k}$ and $-\sqrt{n_k}$, each of multiplicity $m_k := 2^{k-1} = n_k/2$.*

3. *Prove that whenever f is an eigenvector of $\mathcal{H}_k$, one has*
$$\|f\|_\infty \leq \|f\|_1/\sqrt{n_k}.$$
Derive from this observation the conclusion as follows:

   > *Let $a_1, ..., a_{m_k} \in \mathbf{R}^{n_k}$ be orthogonal to each other unit vectors which are eigenvectors of $\mathcal{H}_k$ with eigenvalues $\sqrt{n_k}$* (by the above, the dimension of the eigenspace of $\mathcal{H}_k$ associated with the eigenvalue $\sqrt{n_k}$ is m_k, so that the required $a_1, ..., a_{m_k}$ do exist), *and let A be the $m_k \times n_k$ matrix with the rows $a_1^T, ..., a_{m_k}^T$. For every $x \in \operatorname{Ker} A$ it holds*
   > $$\|x\|_\infty \leq \frac{1}{\sqrt{n_k}}\|x\|_1,$$
   > *whence A satisfies the nullspace property whenever the sparsity s satisfies $2s < \sqrt{n_k} = \sqrt{2m_k}$. Moreover, there exists (and can be found efficiently) an $m_k \times n_k$ contrast matrix $H = H_k$ such that for every $s < \frac{1}{2}\sqrt{n_k}$, the pair $(H_k, \|\cdot\|_\infty)$ satisfies the associated with A condition $\mathbf{Q}_\infty(s, \kappa_s = \underbrace{s/\sqrt{n_k}}_{<1/2})$, and the $\|\cdot\|_2$-norms of columns of H_k do not exceed $\sqrt{2\frac{\sqrt{n_k}+1}{\sqrt{n_k}}}$.*

   Note that the above conclusion yields a sequence of individual ($m_k = 2^{k-1}$) × ($n_k = 2^k$) sensing matrices, $k = 1, 2, ...$, with "size ratio" $n_k/m_k = 2$, which make an efficiently verifiable condition for s-goodness, say, $\mathbf{Q}_\infty(s, \frac{1}{3})$ satisfiable in basically the entire range of values of s allowed by Proposition 1.3.7. It would be interesting to get similar "fully constructive" results for other size ratios, like $m : n = 1 : 4$, $m : n = 1 : 8$, etc.

Exercise 1.2 [Follow-up to Exercise 1.1] Exercise 1.1 provides us with an explicitly given ($m = 512$) × ($n = 1024$) sensing matrix $\bar{A}$ such that the efficiently verifiable condition $\mathbf{Q}_\infty(15, \frac{15}{32})$ is satisfiable; in particular, $\bar{A}$ is 15-good. With all we know about limits of performance of verifiable sufficient conditions for goodness, how should we evaluate this specific sensing matrix? *Could we point out a sensing matrix of the same size which is provably s-good for a larger (or "much larger") than 15 value of s?*

We do not know the answer, and you are requested to explore some possibilities, including (but not reducing to – you are welcome to investigate more options!) the following ones.

1. *Generate at random a sample of $m \times n$ sensing matrices A, compute their mutual incoherences and look how large goodness levels they justify. What happens when the matrices are the Gaussian (independent $\mathcal{N}(0, 1)$ entries) and the Rademacher ones (independent entries taking values ± 1 with probabilities $1/2$)?*

2. *Generate at random a sample of $m \times n$ matrices with independent $\mathcal{N}(0, 1/m)$ entries. Proposition 1.3.1 suggests that a sampled matrix A has good chances to satisfy* $\mathrm{RIP}(\delta, k)$ *with some $\delta < 1/3$ and some k, and thus to be s-good (and even more than this, see Proposition 1.3.2) for every $s \leq k/2$. Of course, given A we cannot check whether the matrix indeed satisfies* $\mathrm{RIP}(\delta, k)$ *with given δ, k; what we can try to do is to certify that* $\mathrm{RIP}(\delta, k)$ *does <u>not</u> take place. To this end, it suffices to select at random, say, 200 $m \times k$ submatrices $\tilde{A}$ of A and compute the eigenvalues of $\tilde{A}^T\tilde{A}$; if A possesses* $\mathrm{RIP}(\delta, k)$*, all these eigenvalues should belong to the segment $[1 - \delta, 1 + \delta]$, and if in reality this does not happen, A definitely is not* $\mathrm{RIP}(\delta, k)$*.*

Exercise 1.3 Let us start with preamble. Consider a finite Abelian group; the only thing which matters for us is that such a group G is specified by a collection of a $k \geq 1$ of positive integers $\nu_1, ..., \nu_k$ and is comprised of all collections $\omega = (\omega_1, ..., \omega_k)$ where every ω_i is an integer from the range $\{0, 1, ..., \nu_k - 1\}$; the group operation, denoted by $\oplus$, is

$$(\omega_1, ..., \omega_k) \oplus (\omega'_1, ..., \omega'_k) = ((\omega_1 + \omega'_1) \bmod \nu_1, ..., (\omega_k + \omega'_k) \bmod \nu_k),$$

where $a \bmod b$ is the remainder, taking values in $\{0, 1, ..., b-1\}$, in the division of an integer a by positive integer b; say, $5 \bmod 3 = 2$, and $6 \bmod 3 = 0$. Clearly, the cardinality of the above group G is $n_k = \nu_1\nu_2...\nu_k$. A *character* of group G is a homomorphism acting from G into the multiplicative group of complex numbers of modulus 1, or, in simple words, a complex-valued function $\chi(\omega)$ on G such that $|\chi(\omega)| = 1$ for all $\omega \in G$ and $\chi(\omega \oplus \omega') = \chi(\omega)\chi(\omega')$ for all $\omega, \omega' \in G$. Note that characters themselves form a group w.r.t. pointwise multiplication; clearly, all characters of our G are functions of the form

$$\chi((\omega_1, ..., \omega_k)) = \mu_1^{\omega_1}...\mu_k^{\omega_k},$$

where μ_i are restricted to be roots of degree ν_i from 1: $\mu_i^{\nu_i} = 1$. It is immediately seen that the group G_* of characters of G is of the same cardinality $n_k = \nu_1...\nu_k$ as G. We can associate with G the matrix $\mathcal{F}$ of size $n_k \times n_k$; the columns in the matrix are indexed by the elements ω of G, the rows – by the characters $\chi \in G_*$ of G, and the element in cell (χ, ω) is $\chi(\omega)$. The standard example here corresponds to $k = 1$, in which case $\mathcal{F}$ clearly is the $\nu_1 \times \nu_1$ matrix of Discrete Fourier Transform.

Now goes the exercise:

1. *Verify that the above $\mathcal{F}$ is, up to factor $\sqrt{n_k}$, a unitary matrix: denoting by $\overline{a}$ the complex conjugate of a complex number a, $\sum_{\omega \in G} \chi(\omega)\overline{\chi'}(\omega)$ is n_k or 0 depending on whether $\chi = \chi'$ or $\chi \neq \chi'$.*

2. *Let $\bar{\omega}, \bar{\omega}'$ be two elements of G. Prove that there exists a permutation Π of elements of G which maps $\bar{\omega}$ into $\bar{\omega}'$ and is such that*

$$\mathrm{Col}_{\Pi(\omega)}[\mathcal{F}] = D\mathrm{Col}_{\omega}[\mathcal{F}] \;\forall \omega \in G,$$

where D is diagonal matrix with diagonal entries $\chi(\bar{\omega}')/\chi(\bar{\omega})$, $\chi \in G_*$.

3. *Consider the special case of the above construction where $\nu_1 = \nu_2 = ... = \nu_k = 2$. Verify that in this case $\mathcal{F}$, up to permutation of rows and permutation of columns (these permutations depend on how we assign the elements of G and of G_* their serial numbers) is exactly the Hadamard matrix $\mathcal{H}_k$.*

4. *Extract from the above the following fact: let m, k be positive integers such that $m \leq n_k := 2^k$, and let sensing matrix A be obtained from $\mathcal{H}_k$ by selecting m distinct rows. Assume we want to find an $m \times n_k$ contrast matrix H such that the pair $(H, \|\cdot\|_\infty)$ satisfies the condition $\mathbf{Q}_\infty(s, \kappa)$ with as small κ as possible; by Proposition 1.3.4, to this end we should solve n LP programs*

$$\mathrm{Opt}_i = \min_h \|e^i - A^T h\|_\infty,$$

where e^i is i-th basic orth in $\mathbf{R}^n$. Prove that with A coming from $\mathcal{H}_k$, all these problems have the same optimal value, and optimal solutions to all of the problems are readily given by the optimal solution to just one of them.

Exercise 1.4 Proposition 1.3.7 states that the verifiable condition $\mathbf{Q}_\infty(s, \kappa)$ can certify s-goodness of "essentially nonsquare" (with $m \leq n/2$) $m \times n$ sensing matrix A only when s is small as compared to m, namely, $s \leq \sqrt{2m}$. The exercise to follow is aimed at investigating what happens when $m \times n$ "low" (with $m < n$) sensing matrix A is "nearly square", meaning that $m^o = n - m$ is small as compared to n. Specifically, you should prove that for properly selected individual $(n - m^o) \times n$ matrices A the condition $\mathbf{Q}_\infty(s, \kappa)$ with $\kappa < 1/2$ is satisfiable when s is as large as $O(1)n/\sqrt{m^o}$.

1. *Let $n = 2^k p$ with positive integer p and integer $k \geq 1$, and let $m^o = 2^{k-1}$. Given $2m^o$-dimensional vector u, let u^+ be n-dimensional vector built as follows: we split indexes from $\{1, ..., n = 2^k p\}$ into 2^k consecutive groups $I_1, ..., I_{2^k}$, p elements per group, and all entries of u^+ with indexes from I_i are equal to i-th entry, u_i, of vector u. Now let U be the linear subspace in $\mathbf{R}^{2^k}$ comprised of all eigenvectors, with eigenvalue $\sqrt{2^k}$, of the Hadamard matrix $\mathcal{H}_k$, see Exercise 1.1, so that the dimension of U is $2^{k-1} = m^o$, and let L be given by*

$$L = \{u^+ : u \in U\} \subset \mathbf{R}^n.$$

Clearly, L is a linear subspace in $\mathbf{R}^n$ of dimension m^o. Prove that

$$\forall x \in L : \|x\|_\infty \leq \frac{\sqrt{2m^o}}{n} \|x\|_1.$$

Conclude that if A is $(n - m^o) \times n$ sensing matrix with $\mathrm{Ker}\, A = L$, *then the verifiable sufficient condition $\mathbf{Q}_\infty(s, \kappa)$ does certify s-goodness of A whenever*

$$1 \leq s < \frac{n}{2\sqrt{2m^o}}.$$

2. *Let L be m^o-dimensional subspace in $\mathbf{R}^n$. Prove that L contains nonzero vector x with*

$$\|x\|_\infty \geq \frac{\sqrt{m^o}}{n} \|x\|_1,$$

so that the condition $\mathbf{Q}_\infty(s, \kappa)$ cannot certify s-goodness of $(n - m^o) \times n$ sensing matrix A whenever $s > O(1)n/\sqrt{m^o}$, for properly selected absolute constant $O(1)$.

Exercise 1.5 *Utilize the results of Exercise 1.3 in a numerical experiment as follows.*

- *select n as an integer power 2^k of 2, say, set $n = 2^{10} = 1024$*
- *select a "representative" sequence M of values of m, $1 \leq m < n$, including values of m close to n and "much smaller" than n, say, use*

$$M = \{2, 5, 8, 16, 32, 64, 128, 256, 512, 7, 896, 960, 992, 1008, 1016, 1020, 1022, 1023\}$$

- *for every $m \in M$,*
  - *generate at random an $m \times n$ submatrix A of the $n \times n$ Hadamard matrix $\mathcal{H}_k$ and utilize the result of item 4 of Exercise 1.3 in order to find the largest s such that s-goodness of A can be certified via the condition $\mathbf{Q}_\infty(\cdot, \cdot)$; call $s(m)$ the resulting value of s.*
  - *generate a moderate sample of Gaussian $m \times n$ sensing matrices A_i with independent $\mathcal{N}(0, 1/m)$ entries and use the construction from Exercise 1.2 to upper-bound the largest s for which a matrix from the sample satisfies* $\mathrm{RIP}(1/3, 2s)$; *call $\widehat{s}(m)$ the largest, over your A_i's, of the resulting upper bounds.*

The goal of the exercise is to compare the computed values of $s(m)$ and $\widehat{s}(m)$; in other words, we again want to understand how "theoretically perfect" RIP compares to "conservative restricted scope" condition $\mathbf{Q}_\infty$.

Lecture 2

Hypothesis Testing

Disclaimer for experts

In what follows, we allow for "general" probability and observation spaces, general probability distributions, etc., which, formally, would make it necessary to address the related measurability issues. In order to streamline our exposition, and taking into account that we do not expect from our target audience to be experts in formal nuances of the measure theory, we decided to omit in the text comments (always self-evident for an expert) on measurability and replace them with a "disclaimer" as follows:
Below, unless the opposite is explicitly stated,

- all probability and observation spaces are Polish (complete separable metric) spaces equipped by σ-algebras of Borel sets;
- all random variables (i.e., functions from a probability space to some other space) take values in Polish spaces; these variables, same as other functions we deal with, are Borel;
- all probability distributions we are dealing with are σ-additive Borel measures on the respective probability spaces; the same is true for all reference measures and probability densities taken w.r.t. these measures.

When an entity (a random variable, or a probability density, or a function, say, a test) is part of the data, the Borel property is a default assumption; e.g., the sentence "Let random variable η be a deterministic transformation of random variable ξ" should be read as "let $\eta = f(\xi)$ for some Borel function f", and the sentence "Consider test $\mathcal{T}$ deciding on hypotheses $H_1, ..., H_L$ via observation $\omega \in \Omega$" should be read as "Consider a Borel function $\mathcal{T}$ on Polish space Ω, the values of the function being subsets of the set $\{1, ..., L\}$." When an entity is built by us rather than being part of the data, the Borel property is (always straightforwardly verifiable) property of the construction. For example, the statement "The test $\mathcal{T}$ given by... is such that..." should be read as "The test $\mathcal{T}$ given by... is a Borel function of observations and is such that..."

On several occasions, we still use the word "Borel;" those not acquainted with the notion are welcome to just ignore this word.

2.1 Preliminaries from Statistics: Hypotheses, Tests, Risks

2.1.1 Hypothesis Testing Problem

Hypothesis Testing is one of the most basic problems of Statistics. Informally, this is the problem where one is given an *observation* – a realization of random variable with unknown (at least partially) probability distribution and want to decide, based on this observation, on two or more hypotheses on the actual distribution of the observed variable. A convenient for us formal setting is as follows:

> Given are:
>
> - *Observation space* Ω, where the observed random variable (r.v.) takes its values;
> - *L families* $\mathcal{P}_\ell$ of probability distributions on Ω. We associate with these families L hypotheses $H_1, ..., H_L$, with H_ℓ stating that the probability distribution P of the observed r.v. belongs to the family $\mathcal{P}_\ell$ (shorthand: $H_\ell : P \in \mathcal{P}_\ell$). We shall say that the distributions from $\mathcal{P}_\ell$ *obey* hypothesis H_ℓ.
>
>   Hypothesis H_ℓ is called *simple*, if $\mathcal{P}_\ell$ is a singleton, and is called *composite* otherwise.
>
> Our goal is, given an observation – a realization ω of the r.v. in question – to decide which one of the hypotheses is true.

2.1.2 Tests

Informally, a *test* is an inference procedure one can use in the above testing problem. Formally, a test for this testing problem is a function $\mathcal{T}(\omega)$ of $\omega \in \Omega$; the value $\mathcal{T}(\omega)$ of this function at a point ω is some subset of the set $\{1, ..., L\}$:

$$\mathcal{T}(\Omega) \subset \{1, ..., L\}.$$

Given observation ω, the test accepts all hypotheses H_ℓ with $\ell \in \mathcal{T}(\omega)$ and rejects all hypotheses H_ℓ with $\ell \notin \mathcal{T}(\omega)$. We call a test *simple*, if $\mathcal{T}(\omega)$ is a singleton for every ω, that is, whatever be the observation, the test accepts exactly one of the hypotheses $H_1, ..., H_L$ and rejects all other hypotheses.

Note: what we have defined is a *deterministic* test. Sometimes we shall consider also *randomized* tests, where the set of accepted hypotheses is a (deterministic) function of observation ω *and* of a realization θ of independent of ω random parameter (which w.l.o.g. can be assumed to be uniformly distributed on $[0, 1]$). Thus, in a randomized test, the inference depends both on the observation ω and the outcome θ of "flipping a coin," while in a deterministic test the inference depends on observation only. In fact, randomized testing can be reduced to deterministic one. To this end it suffices to pass from our "actual" observation ω to new observation $\omega_+ = (\omega, \theta)$, where $\theta \sim \text{Uniform}[0, 1]$ is independent of ω; the ω-component of our new observation ω_+ is, as before, generated by "the nature," and the θ-component is generated by ourselves. Now, given families $\mathcal{P}_\ell$, $1 \le \ell \le L$, of probability distributions on the original observation space Ω, we can associate with them families $\mathcal{P}_{\ell,+} = \{P \times \text{Uniform}[0, 1] : P \in \mathcal{P}_\ell\}$ of probability distributions on our new observation space $\Omega_+ = \Omega \times [0, 1]$; clearly, to decide on the hypotheses associated with the families $\mathcal{P}_\ell$ via observation ω is the same as to decide on the hypotheses associated with the families $\mathcal{P}_{\ell,+}$ of our new observation ω_+, and deterministic tests for the latter testing problem are exactly the randomized tests for the former one.

2.1.3 Testing from repeated observations

There are situations where an inference can be based on several observations $\omega_1, ..., \omega_K$ rather than on a single one. Our related setup is as follows:

> We are given L families $\mathcal{P}_\ell$, $\ell = 1, ..., L$, of probability distributions on observation space Ω and a collection
>
> $$\omega^K = (\omega_1, ..., \omega_K) \in \Omega^K = \underbrace{\Omega \times ... \times \Omega}_{K}$$

and want to make conclusions on how the distribution of ω^K "is positioned" w.r.t. the families $\mathcal{P}_\ell$, $1 \leq \ell \leq L$.

Specifically, we are interested in three situations of this type, specifically, as follows.

2.1.3.1 Stationary K-repeated observations

In the case of stationary K-repeated observations $\omega_1, ..., \omega_K$ are *independently of each other* drawn from a distribution P. Our goal is to decide, given ω^K, on the hypotheses $P \in \mathcal{P}_\ell$, $\ell = 1, ..., L$.

Equivalently: Families $\mathcal{P}_\ell$ of probability distributions of $\omega \in \Omega$, $1 \leq \ell \leq L$, give rise to the families

$$\mathcal{P}_\ell^{\odot,K} = \{P^K = \underbrace{P \times ... \times P}_{K} : P \in \mathcal{P}_\ell\}$$

of probability distributions on Ω^K; we refer to the families $\mathcal{P}_\ell^{\odot,K}$ as to *K-th diagonal powers* of the family $\mathcal{P}_\ell$. Given observation $\omega^K \in \Omega^K$, we want to decide on the hypotheses

$$H_\ell^{\odot,K} : \omega^K \sim P^K \in \mathcal{P}_\ell^{\odot,K},\ 1 \leq \ell \leq L.$$

2.1.3.2 Semi-stationary K-repeated observations

In the case of semi-stationary K-repeated observations, "the nature" selects somehow a sequence $P_1, ..., P_K$ of distributions on Ω, and then draws, *independently across k*, observations ω_k, $k = 1, ..., K$, from these distributions:

$$\omega_k \sim P_k \text{ are independent across } k \leq K$$

Our goal is to decide, given $\omega^K = (\omega_1, ..., \omega_K)$, on the hypotheses $\{P_k \in \mathcal{P}_\ell, 1 \leq k \leq K\}$, $\ell = 1, ..., L$.

Equivalently: Families $\mathcal{P}_\ell$ of probability distributions of $\omega \in \Omega$, $1 \leq \ell \leq L$, give rise to the families

$$\mathcal{P}_\ell^{\oplus,K} = \{P^K = P_1 \times ... \times P_K : P_k \in \mathcal{P}_\ell,\ 1 \leq k \leq K\}$$

of probability distributions on Ω^K. Given observation $\omega^K \in \Omega^K$, we want to decide on the hypotheses

$$H_\ell^{\oplus,K} : \omega^K \sim P^K \in \mathcal{P}_\ell^{\oplus,K},\ 1 \leq \ell \leq L.$$

In the sequel, we refer to families $\mathcal{P}_\ell^{\oplus,K}$ as to *K-th direct powers* of the families $\mathcal{P}_\ell$. A closely related notion is the one of *direct product*

$$\mathcal{P}_\ell^{\oplus,K} = \bigoplus_{k=1}^{K} \mathcal{P}_{\ell,k}$$

of K families $\mathcal{P}_{\ell,k}$, of probability distributions on Ω_k, over $k = 1, ..., K$. By definition,

$$\mathcal{P}_\ell^{\oplus,K} = \{P^K = P_1 \times ... \times P_K : P_k \in \mathcal{P}_{\ell,k}, 1 \leq k \leq K\}.$$

2.1.3.3 Quasi-stationary K-repeated observations

Quasi-stationary K-repeated observations $\omega_1 \in \Omega, ..., \omega_K \in \Omega$ stemming from a family $\mathcal{P}$ of probability distributions on an observation space Ω are generated as follows:

> "In the nature" there exists random sequence $\zeta^K = (\zeta_1, ..., \zeta_K)$ of "driving factors" such that for every k, ω_k is a deterministic function of $\zeta_1, ..., \zeta_k$:
>
> $$\omega_k = \theta_k(\zeta_1, ..., \zeta_k)$$
>
> and the conditional, $\zeta_1, ..., \zeta_{k-1}$ given, distribution $P_{\omega_k|\zeta_1,...,\zeta_{k-1}}$ of ω_k always (i.e., for all $\zeta_1, ..., \zeta_{k-1}$) belongs to $\mathcal{P}$.

With the above mechanism, the collection $\omega^K = (\omega_1, ..., \omega_K)$ has some distribution P^K which depends on the distribution of driving factors and on functions $\theta_k(\cdot)$. We denote by $\mathcal{P}^{\otimes,K}$ the family of all distributions P^K which can be obtained in this fashion and we refer to random observations ω^K with distribution P^K of the just define type as to *quasi-stationary K-repeated observations stemming from $\mathcal{P}$*. The quasi-stationary version of our hypothesis testing problem reads: Given L families $\mathcal{P}_\ell$ of probability distributions $\mathcal{P}_\ell$, $\ell = 1, ..., L$, on Ω and an observation $\omega^K \in \Omega^K$, decide on the hypotheses

$$H_\ell^{\otimes,K} = \{P^K \in \mathcal{P}_\ell^{\otimes,K}\},\ 1 \leq \ell \leq K$$

on the distribution P^K of the observation ω^K.

A closely related notion is the one of *quasi-direct product*

$$\mathcal{P}_\ell^{\otimes,K} = \bigotimes_{k=1}^{K} \mathcal{P}_{\ell,k}$$

of K families $\mathcal{P}_{\ell,k}$, of probability distributions on Ω_k, over $k = 1, ..., K$. By definition, $\mathcal{P}_\ell^{\otimes,K}$ is comprised of all probability distributions of random sequences $\omega^K = (\omega_1, ..., \omega_K)$, $\omega_k \in \Omega_k$, which can be generated as follows: "in the nature" there exists a random sequence $\zeta^K = (\zeta_1, ..., \zeta_K)$ of "driving factors" such that for every $k \leq K$, ω_k is a deterministic function of $\zeta^k = (\zeta_1, ..., \zeta_k)$, and conditional, ζ^{k-1} being given, distribution of ω_k always belongs to $\mathcal{P}_{\ell,k}$.

The above description of quasi-stationary K-repeated observations seems to be too complicated; well, this is what happens in some important applications, e.g., in *hidden Markov chain*. Here $\Omega = \{1, ..., d\}$ is a finite set, and $\omega_k \in \Omega$, $k = 1, 2, ...$, are generated as follows: "in the nature there" exists a Markov chain with D-element state space $\mathcal{S}$ split into d non-overlapping bins, and ω_k is the serial number $\beta(\eta)$ of the bin to which the state η_k of the chain belongs. Now, every column Q^j of the transition matrix Q of the chain (this column is a probability distribution on $\{1, ..., D\}$) generates a probability distribution P_j on Ω, specifically, the distribution of $\beta(\eta)$, $\eta \sim Q^j$. Now, a family $\mathcal{P}$ of distributions on Ω induces a family $\mathcal{Q}[\mathcal{P}]$ of all $D \times D$ stochastic matrices Q for which all D distributions P^j, $j = 1, ..., D$, belong to $\mathcal{P}$. When $Q \in \mathcal{Q}[\mathcal{P}]$, observations ω_k, $k = 1, 2, ...$ clearly are given by the above "quasi-stationary mechanism" with η_k in the role of driving factors and $\mathcal{P}$ in the role of $\mathcal{P}_\ell$. Thus, in the situation in question, given L families $\mathcal{P}_\ell$, $\ell = 1, ..., L$ of probability distributions on $\mathcal{S}$, deciding on hypotheses $Q \in \mathcal{Q}[\mathcal{P}_\ell]$, $\ell = 1, ..., L$, on the transition matrix Q of the Markov chain underlying our observations reduces to hypothesis testing via quasi-stationary K-repeated observations.

2.1.4 Risk of a simple test

Let $\mathcal{P}_\ell$, $\ell = 1, ..., L$, be families of probability distributions on observation space Ω; these families give rise to hypotheses

$$H_\ell : P \in \mathcal{P}_\ell,\ \ell = 1, ..., L$$

on the distribution P of a random observation $\omega \sim P$. We are about to define the *risks* of a *simple* test $\mathcal{T}$ deciding on the hypotheses H_ℓ, $\ell = 1,...,L$, via observation ω; recall that simplicity means that as applied to an observation, our test accepts exactly one hypothesis and rejects all other hypotheses.

Partial risks $\mathrm{Risk}_\ell(\mathcal{T}|H_1,...,H_L)$ are the worst-case, over $P \in \mathcal{P}_\ell$, P-probabilities for $\mathcal{T}$ to reject ℓ-th hypothesis when it is true, that is, when $\omega \sim P$:

$$\mathrm{Risk}_\ell(\mathcal{T}|H_1,...,H_L) = \sup_{P\in\mathcal{P}_\ell} \mathrm{Prob}_{\omega\sim P}\left\{\omega : \mathcal{T}(\omega) \neq \{\ell\}\right\},\ \ell = 1,...,L.$$

Note that for ℓ fixed, ℓ-th partial risk depends on how we order the hypotheses; when reordering them, we should reorder risks as well. In particular, for a test $\mathcal{T}$ deciding on two hypotheses H, H' we have

$$\mathrm{Risk}_1(\mathcal{T}|H,H') = \mathrm{Risk}_2(\mathcal{T}|H',H).$$

Total risk $\mathrm{Risk}_{\rm tot}(\mathcal{T}|H_1,...,H_L)$ is the sum of all L partial risks:

$$\mathrm{Risk}_{\rm tot}(\mathcal{T}|H_1,...,H_L) = \sum_{\ell=1}^{L} \mathrm{Risk}_\ell(\mathcal{T}|H_1,...,H_L).$$

Risk $\mathrm{Risk}(\mathcal{T}|H_1,...,H_L)$ is the maximum of all L partial risks:

$$\mathrm{Risk}(\mathcal{T}|H_1,...,H_L) = \max_{1\leq\ell\leq L} \mathrm{Risk}_\ell(\mathcal{T}|H_1,...,H_L).$$

Note that *at the first glance*, we have defined risks for single-observation tests only; in fact, we have defined them for tests based on stationary, semi-stationary, and quasi-stationary K-repeated observations as well, since, as we remember from Section 2.1.3, the corresponding testing problems, after redefining observations and families of probability distributions (ω^K in the role of ω and, say, $\mathcal{P}_\ell^{\oplus,K} = \bigoplus\limits_{k=1}^{K} \mathcal{P}_\ell$ in the role of $\mathcal{P}_\ell$), become single-observation testing problems.

Pay attention to the following two important observations:

- Partial risks of a simple test are defined in the worst-case-oriented fashion: as the worst, over the true distributions P of observations compatible with the hypothesis in question, probability to reject this hypothesis

- Risks of a simple test say what happens, statistically speaking, when the true distribution P of observation obeys one of the hypotheses in question, and *say nothing on what happens when P does not obey neither one of the L hypotheses.*

Remark 2.1.1 "The smaller are hypotheses, the less are risks." Specifically, given families of probability distributions $\mathcal{P}_\ell \subset \mathcal{P}'_\ell$, $\ell = 1,...,L$, on observation space Ω, along with hypotheses $H_\ell : P \in \mathcal{P}_\ell$, $H'_\ell : P \in \mathcal{P}'_\ell$ on the distribution P of an observation $\omega \in \Omega$, every test $\mathcal{T}$ deciding on the "larger" hypotheses $H'_1,...,H'_L$ can be considered as a test deciding on smaller hypotheses $H_1,...,H_L$ as well, and the risks of the test when passing from larger hypotheses to smaller ones can only drop down:

$$\mathcal{P}_\ell \subset \mathcal{P}'_\ell, 1 \leq \ell \leq L \Rightarrow \mathrm{Risk}(\mathcal{T}|H_1,...,H_L) \leq \mathrm{Risk}(\mathcal{T}|H'_1,...,H'_L).$$

For example, families of probability distributions $\mathcal{P}_\ell$, $1 \leq \ell \leq L$, on Ω and a positive integer K induce three families of hypotheses on a distribution P^K of K-repeated observations:

$$H_\ell^{\odot,K}K : P^K \in \mathcal{P}_\ell^{\odot,K},\ H_\ell^{\oplus,K} : P^K \in \mathcal{P}_\ell^{\oplus,K} = \bigoplus_{k=1}^K \mathcal{P}_\ell,\ H_\ell^{\otimes,K} : P^K \in \mathcal{P}_\ell^{\otimes,K} = \bigotimes_{k=1}^K \mathcal{P}_\ell,\ 1 \leq \ell \leq L,$$

(see Section 2.1.3), and clearly

$$\mathcal{P}_\ell^K \subset \mathcal{P}_\ell^{\oplus,K} \subset \mathcal{P}_\ell^{\otimes,K};$$

it follows that when passing from quasi-stationary K-repeated observations to semi-stationary K-repeated, and then to stationary K-repeated observations, the risks of a test can only go down.

2.1.5 Two-point lower risk bound

The following observation is nearly evident:

Proposition 2.1.1 *Consider two simple hypotheses $H_1 : P = P_1$ and $H_2 : P = P_2$ on the distribution P of observation $\omega \in \Omega$, and assume that P_1, P_2 have densities p_1, p_2 w.r.t. some reference measure Π on Ω [1]. Then for any simple test $\mathcal{T}$ deciding on H_1, H_2 it holds*

$$\mathrm{Risk}_{\mathrm{tot}}(\mathcal{T}|H_1,H_2) \geq \int_\Omega \min[p_1(\omega), p_2(\omega)]\Pi(d\omega). \tag{2.1.1}$$

Note that the right hand side in this relation is independent of how Π is selected.

Proof. Consider a simple test $\mathcal{T}$, perhaps a randomized one, and let $\pi(\omega)$ be the probability for this test to accept H_1 and reject H_2 when the observation is ω; since the test is simple, the probability for $\mathcal{T}$ to accept H_2 and to reject H_1, observation being ω, is $1 - \pi(\omega)$. Consequently,

$$\begin{array}{rcl}\mathrm{Risk}_1(\mathcal{T}|H_1,H_2) & = & \int_\Omega (1-\pi(\omega))p_1(\omega)\Pi(d\omega),\\ \mathrm{Risk}_2(\mathcal{T}|H_1,H_2) & = & \int_\Omega \pi(\omega)p_2(\omega)\Pi(d\omega),\end{array}$$

whence

$$\mathrm{Risk}_{\mathrm{tot}}(\mathcal{T}|H_1,H_2) = \int_\Omega [(1-\pi(\omega))p_1(\omega) + \pi(\omega)p_2(\omega)]\Pi(d\omega) \geq \int_\Omega \min[p_1(\omega), p_2(\omega)]\Pi(d\omega). \quad \square$$

Remark 2.1.2 *Note that the lower risk bound* (2.1.1) *is achievable; given an observation ω, the corresponding test $\mathcal{T}$ accepts H_1 with probability 1 (i.e., $\pi(\omega) = 1$ when $p_1(\omega) > p_2(\omega)$), accepts H_2 when $p_1(\omega) < p_2(\omega)$ (i.e., $\pi(\omega) = 0$ when $p_1(\omega) < p_2(\omega)$) and accepts H_1 and H_2 with probabilities 1/2 in the case of tie (i.e., $\pi(\omega) = 1/2$ when $p_1(\omega) = p_2(\omega)$); this is nothing but maximum likelihood test naturally adjusted to account for ties.*

Example 2.1 Let $\Omega = \mathbf{R}^d$, let the reference measure Π be the Lebesgue measure on $\mathbf{R}^d$, and let $p_\chi(\cdot) = \mathcal{N}(\mu_\chi, I_d)$, be the Gaussian densities on $\mathbf{R}^d$ with unit covariance and means μ_χ, $\chi = 1, 2$. In this case, assuming $\mu_1 \neq \mu_2$, the recipe from Remark 2.1.2 reduces to the following:

Let

$$\phi_{1,2}(\omega) = \frac{1}{2}[\mu_1 - \mu_2]^T[\omega - w],\ w = \frac{1}{2}[\mu_1 + \mu_2]. \tag{2.1.2}$$

[1] This assumption is w.l.o.g. – we can take, as Π, the sum of the measures P_1 and P_2.

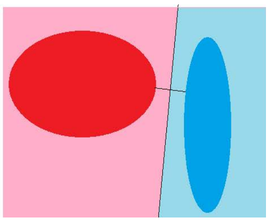


Figure 2.1: "Gaussian Separation" (Example 2.2): Optimal test deciding on whether the mean of Gaussian r.v. belongs to the dark red (H_1) or to the dark blue (H_2) domains. Dark and light red: acceptance domain for H_1. Dark and light blue: acceptance domain for H_2.

Consider the simple test $\mathcal{T}$ which, given an observation ω, accepts $H_1 : p = p_1$ and rejects $H_2 : p = p_2$ when $\phi_{1,2}(\omega) \geq 0$, otherwise accepts H_2 and rejects H_1. For this test,

$$\begin{array}{l}\mathrm{Risk}_1(\mathcal{T}|H_1,H_2) = \mathrm{Risk}_2(\mathcal{T}|H_1,H_2) = \mathrm{Risk}(\mathcal{T}|H_1,H_2)\\ \quad = \frac{1}{2}\mathrm{Risk}_{\mathrm{tot}}(\mathcal{T}|H_1,H_2) = \mathrm{Erf}(\frac{1}{2}\|\mu_1-\mu_2\|_2),\end{array} \tag{2.1.3}$$

where

$$\mathrm{Erf}(\delta) = \frac{1}{\sqrt{2\pi}}\int_\delta^\infty \mathrm{e}^{-s^2/2}ds \tag{2.1.4}$$

is the error function, and the test is optimal in terms of its risk and its total risk.

Note that optimality of $\mathcal{T}$ in terms of total risk is given by Proposition 2.1.1 and Remark 2.1.2; optimality in terms of risk is ensured by optimality in terms of total risk combined with the first equality in (2.1.3).

Example 2.1 admits an immediate and useful extension:

Example 2.2 *Let $\Omega = \mathbf{R}^d$, let the reference measure Π be the Lebesgue measure on $\mathbf{R}^d$, and let M_1, M_2 be two nonempty closed convex sets in $\mathbf{R}^d$ with empty intersection and such that the convex optimization program*

$$\min_{\mu_1,\mu_2}\{\|\mu_1-\mu_2\|_2 : \mu_\chi \in M_\chi,\ \chi = 1,2\} \tag{$*$}$$

has an optimal solution μ_1^*, μ_2^* (this definitely is the case when at least one of the sets M_1, M_2 is bounded). Let

$$\phi_{1,2}(\omega) = \frac{1}{2}[\mu_1^*-\mu_2^*]^T[\omega - w],\ w = \frac{1}{2}[\mu_1^*+\mu_2^*], \tag{2.1.5}$$

and let the simple test $\mathcal{T}$ deciding on the hypotheses

$$H_1 : p = \mathcal{N}(\mu, I_d) \text{ with } \mu\in M_1, \quad H_2 : p = \mathcal{N}(\mu, I_d) \text{ with } \mu\in M_2$$

be as follows (see Figure 2.1): given an observation ω, $\mathcal{T}$ accepts H_1 and rejects H_2 when $\phi_{1,2}(\omega) \geq 0$, otherwise accepts H_2 and rejects H_1. Then

$$\begin{array}{l}\mathrm{Risk}_1(\mathcal{T}|H_1,H_2) = \mathrm{Risk}_2(\mathcal{T}|H_1,H_2) = \mathrm{Risk}(\mathcal{T}|H_1,H_2)\\ \quad = \frac{1}{2}\mathrm{Risk}_{\mathrm{tot}}(\mathcal{T}|H_1,H_2) = \mathrm{Erf}(\frac{1}{2}\|\mu_1^*-\mu_2^*\|_2),\end{array} \tag{2.1.6}$$

and the test is optimal in terms of its risk and its total risk.

Justification of Example 2.2 is immediate. Let e be the $\|\cdot\|_2$-unit vector with the same direction as the one of $\mu_1^*-\mu_2^*$, and let $\xi[\omega] = e^T(\omega - w)$. From optimality conditions for $(*)$ it follows that

$$e^T\mu \geq e^T\mu_1^*\ \forall \mu\in M_1\ \&\ e^T\mu \leq e^T\mu_2^*\ \forall\mu\in M_2.$$

As a result, if $\mu \in M_1$ and the density of ω is $p_\mu = \mathcal{N}(\mu, I_d)$, the random variable $\xi[\omega]$ is scalar Gaussian random variable with unit variance and expectation $\geq \delta := \frac{1}{2}\|\mu_1^* - \mu_2^*\|_2$, implying that p_μ-probability for $\xi[\omega]$ to be negative (which is exactly the same as the p_μ-probability for $\mathcal{T}$ to reject H_1 and accept H_2) is at most $\mathrm{Erf}(\delta)$. Similarly, when $\mu \in M_2$ and the density of ω is $p_\mu = \mathcal{N}(\mu, I_d)$, $\xi[\omega]$ is scalar Gaussian random variable with unit variance and expectation $\leq -\delta$, implying that the p_μ-probability for $\xi[\omega]$ to be nonnegative (which is exactly the same as the probability for $\mathcal{T}$ to reject H_2 and accept H_1) is at most $\mathrm{Erf}(\delta)$. These observations imply the validity of (2.1.6); optimality in terms of risks follows from the fact that risks of a simple test deciding on our now – composite – hypotheses H_1, H_1 on the density p of observation ω can be only larger than the risks of a simple test deciding on two simple hypotheses $p = p_{\mu_1^*}$ and $p = p_{\mu_2^*}$, that is, the quantity $\mathrm{Erf}(\frac{1}{2}\|\mu_1^* - \mu_2^*\|_2)$, see Example 2.1, is a lower bound on the risk and half of the total risk of a test deciding on H_1, H_2; with this in mind, the announced optimalities of $\mathcal{T}$ in terms of risks are immediate consequences of (2.1.6).

We remark that the (nearly self-evident) result stated in Example 2.2 seems first been noticed in [31].

Example 2.2 allows for substantial extensions in two directions: first. it turns out that the "Euclidean separation" underlying the test built in this example can be used to decide on hypotheses on location of a "center" of d-dimensional distribution far beyond the Gaussian observation model considered in this example; this extension will be our goal in the next Section, based on recent paper [78]. A less straightforward and, we believe, more instructive extensions, originating from [73], will be considered in Section 2.3.

2.2 Hypothesis Testing via Euclidean Separation

2.2.1 Situation

In this section, we will be interested in testing hypotheses

$$H_\ell : P \in \mathcal{P}_\ell, \ell = 1, ..., L \tag{2.2.1}$$

on the probability distribution of a random observation ω in the situation where the families of distributions $\mathcal{P}_\ell$ are obtained from the probability distributions from a given family $\mathcal{P}$ by shifts. Specifically, we are given

- A family $\mathcal{P}$ of probability distributions on $\Omega = \mathbf{R}^d$ such that all distributions from $\mathcal{P}$ possess densities with respect to the Lebesgue measure on $\mathbf{R}^n$, and these densities are even functions on $\mathbf{R}^d$ [2];

- A collection $X_1, ..., X_L$ of nonempty closed and convex subsets of $\mathbf{R}^d$, with at most one of the sets unbounded.

These data specify L families $\mathcal{P}_\ell$ of distributions on $\mathbf{R}^d$; $\mathcal{P}_\ell$ is comprised of distributions of random vectors of the form $x + \xi$, where $x \in X_\ell$ is deterministic, and ξ is random with distribution from $\mathcal{P}$. Note that with this setup, deciding upon hypotheses (2.2.1) via observation $\omega \sim P$ is exactly the same as to decide, given observation

$$\omega = x + \xi, \tag{2.2.2}$$

where x is a deterministic "signal" and ξ is random noise with distribution P known to belong to $\mathcal{P}$, on the "position" of x w.r.t. $X_1, ..., X_L$; ℓ-th hypothesis H_ℓ merely says that $x \in H_\ell$. The latter

[2] Allowing for a slight abuse of notation, we write $P \in \mathcal{P}$, where P is a probability distribution, to express the fact that P belongs to $\mathcal{P}$ (no abuse of notation so far), and write $p(\cdot) \in \mathcal{P}$ (this is the abuse of notation), where $p(\cdot)$ is the density of a probability distribution P, to express the fact that $P \in \mathcal{P}$.

allows us to write down ℓ-th hypothesis as $H_\ell : x \in X_\ell$ (of course, this shorthand makes sense only within the scope of our current "signal plus noise" setup).

2.2.2 Pairwise Hypothesis Testing via Euclidean Separation

2.2.2.1 The simplest case

Consider nearly the simplest case of the situation from Section 2.2.1, one with $L = 2$, $X_1 = \{x^1\}$ and $X_2 = \{x^2\}$, $x^1 \neq x^2$, are singletons, and $\mathcal{P}$ also is a singleton; moreover, the probability density of the only distribution from $\mathcal{P}$ is of the form

$$p(u) = f(\|u\|_2),\ f(\cdot) \text{ is a strictly monotonically increasing function on the nonnegative ray.} \tag{2.2.3}$$

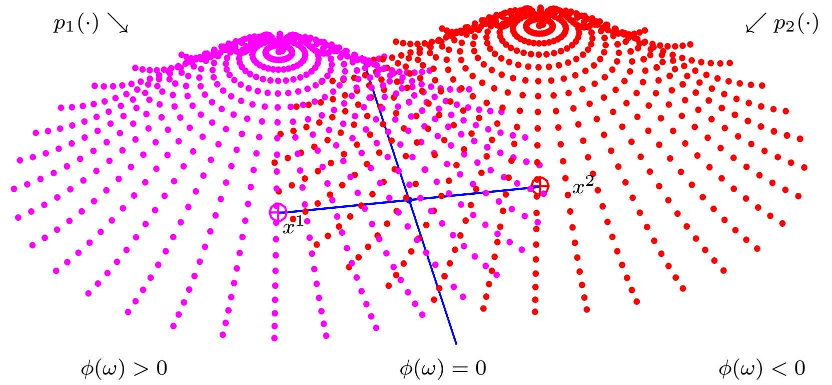
This situation is a generalization of the one considered in Example 2.1, where we dealt with the special case of f, namely, with

$$p(u) = (2\pi)^{-d/2}\mathrm{e}^{-u^Tu/2}.$$

In the case in question our goal is to decide on two simple hypotheses $H_\chi : p(u) = f(\|u - x^\chi\|_2)$, $\chi = 1, 2$, on the density of observation (2.2.2). Let us set

$$\delta = \frac{1}{2}\|x^1 - x^2\|_2,\ e = \frac{x^1 - x^2}{\|x^1 - x^2\|_2},\ \phi(\omega) = e^T\omega - \underbrace{\frac{1}{2}e^T[x^1 + x^2]}_{c}, \tag{2.2.4}$$

and consider the test $\mathcal{T}$ which, given observation $\omega = x + \xi$, accepts the hypothesis $H_1 : x = x^1$ when $\phi(\omega) \geq 0$, and accepts the hypothesis $H_2 : x = x^2$ otherwise.



We have (cf. Example 2.1)

$$\mathrm{Risk}_1(\mathcal{T}|H_1, H_2) = \int\limits_{\omega:\phi(\omega)<0} p_1(\omega)d\omega = \int\limits_{u:e^Tu\geq\delta} f(\|u\|_2)du = \int\limits_{\omega:\phi(\omega)\geq 0} p_2(\omega)d\omega = \mathrm{Risk}_2(\mathcal{T}|H_1, H_2)$$

Since $p(u)$ is strictly decreasing function of $\|u\|_2$, we have also

$$\min[p_1(u), p_2(u)] = \left\{\begin{array}{ll} p_1(u), & \phi(u)\geq 0 \\ p_2(u), & \phi(u)\leq 0 \end{array}\right.,$$

whence

$$\begin{array}{rcl} \mathrm{Risk}_1(\mathcal{T}|H_1, H_2) + \mathrm{Risk}_2(\mathcal{T}|H_1, H_2) & = & \int\limits_{\omega:\phi(\omega)<0} p_1(\omega)d\omega + \int\limits_{\omega:\phi(\omega)\geq 0} p_2(\omega)d\omega) \\ & = & \int\limits_{\mathbf{R}^d} \min[p_1(u), p_2(u)]du \end{array}$$

Invoking Proposition 2.1.1, we conclude that *the test $\mathcal{T}$ is the minimum risk simple test deciding on H_1, H_2, and the risk of this test is*

$$\mathrm{Risk}(\mathcal{T}|H_1,H_2)=\int\limits_{u:e^Tu\geq\delta} f(\|u\|_2)du. \tag{2.2.5}$$

2.2.2.2 Extension

Now consider a slightly more complicated case of the situation from Section 2.2.1, the one with $L=2$ and nonempty and nonintersecting closed convex sets X_1, X_2, one of the sets being bounded; as about $\mathcal{P}$, we still assume that it is a singleton, and the density of the only distribution from $\mathcal{P}$ is of the form 2.2.3. Our now situation is an extension of the one from Example 2.2. By the same reasons as in the case of the latter Example, with X_1, X_2 as above, the convex minimization problem

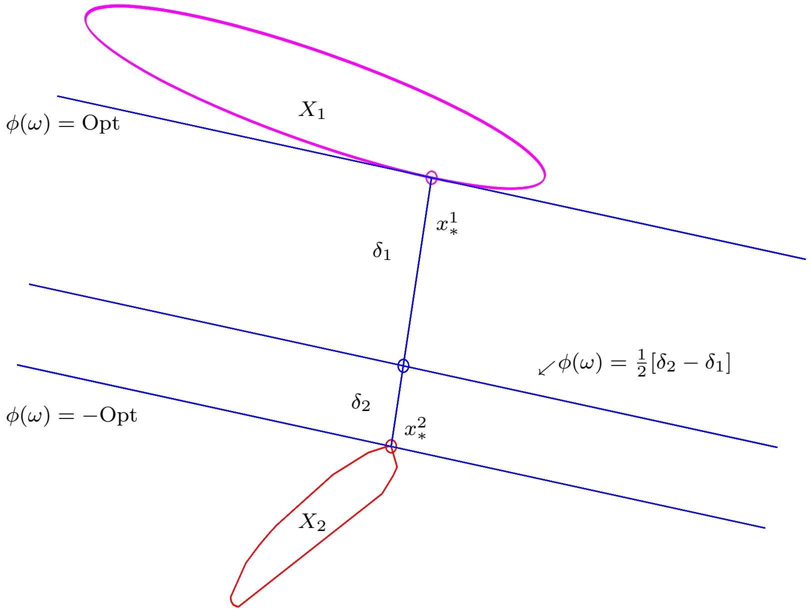
$$\mathrm{Opt}=\min_{x^1\in X_1,x^2\in X_2}\frac{1}{2}\|x^1-x^2\|_2 \tag{2.2.6}$$

is solvable, and denoting by (x^1_*,x^2_*) an optimal solution and setting

$$\phi(\omega)=e^T\omega-c,\ e=\frac{x^1_*-x^2_*}{\|x^1_*-x^2_*\|_2},\ c=\frac{1}{2}e^T[x^1_*+x^2_*] \tag{2.2.7}$$

the stripe $\{\omega:-\mathrm{Opt}\leq\phi(x)\leq\mathrm{Opt}\}$ separates X_1 and X_2:

$$\phi(x^1)\geq\phi(x^1_*)=\mathrm{Opt}\,\forall x^1\in X_1\ \&\ \phi(x^2)\geq\phi(x^2_*)=-\mathrm{Opt}\,\forall x^2\in X_2 \tag{2.2.8}$$


X_1
$\phi(\omega)=\mathrm{Opt}$
x^1_*
δ_1
$\swarrow\phi(\omega)=\frac{1}{2}[\delta_2-\delta_1]$
δ_2
$\phi(\omega)=-\mathrm{Opt}$
x^2_*
X_2

Proposition 2.2.1 *Let X_1,X_2 be nonempty and nonintersecting closed convex sets in $\mathbf{R}^d$, one of the sets being bounded. With Opt and $\phi(\cdot)$ given by (2.2.6) – (2.2.7), let us split the width $2\mathrm{Opt}$ of the stripe $\{\omega:-\mathrm{Opt}\leq\phi(\omega)\leq\mathrm{Opt}\}$ separating X_1 and X_2 into two nonnegative parts:*

$$\delta_1\geq 0,\delta_2\geq 0,\ \delta_1+\delta_2=2\mathrm{Opt} \tag{2.2.9}$$

and consider simple test $\mathcal{T}$ deciding on the hypotheses $H_1:x\in X_1$, $H_2:x\in X_2$ via observation (2.2.2) by accepting H_1 when

$$\phi(\omega)\geq\frac{1}{2}[\delta_2-\delta_1]$$

and accepting H_2 otherwise. Then

$$\mathrm{Risk}_\chi(\mathcal{T}|H_1,H_2) \leq \int\limits_{\delta_\chi}^{\infty} \gamma(s)ds,\ \chi=1,2, \tag{2.2.10}$$

where $\gamma(\cdot)$ is the univariate marginal density of ξ, that is, probability density of the scalar random variable $h^T\xi$, where $\|h\|_2=1$ (note that due to (2.2.3), $\gamma(\cdot)$ is independent of how we select h with $\|h\|_2=1$).

In addition, when $\delta_1=\delta_2=\mathrm{Opt}$, $\mathcal{T}$ is the minimum risk test deciding on H_1, H_2. The risk of this test is

$$\mathrm{Risk}(\mathcal{T}|H_1,H_2) = \int\limits_{\mathrm{Opt}}^{\infty} \gamma(s)ds. \tag{2.2.11}$$

Proof. By (2.2.3) and (2.2.8), for $x\in X_1$ we have (see the picture above):

$$\mathrm{Prob}_{\xi\sim p(\cdot)}\left\{\phi(x+\xi)<\frac{1}{2}[\delta_2-\delta_1]\right\} \leq \mathrm{Prob}_{\xi\sim p(\cdot)}\left\{[-e]^T\xi\geq\delta_1\right\} = \int\limits_{\delta_1}^{\infty}\gamma(s)ds;$$

by "symmetric" reasoning, for $x\in X_2$ we have

$$\mathrm{Prob}_{\xi\sim p(\cdot)}\left\{\phi(x+\xi)\geq\frac{1}{2}[\delta_2-\delta_1]\right\}] \leq \mathrm{Prob}_{\xi\sim p(\cdot)}\left\{e^T\xi\geq\delta_2\right\} = \int\limits_{\delta_2}^{\infty}\gamma(s)ds,$$

and we arrive at (2.2.10). The fact that in the case of $\delta_1=\delta_2=\mathrm{Opt}$ our test $\mathcal{T}$ becomes the minimum risk test deciding on composite hypotheses H_1,H_2, same as (2.2.10), are readily given by the fact that due to the analysis in Section 2.2.2.1, the minimal, over all possible tests, risk of deciding on two simple hypotheses $H_1':x=x_*^1$, $H_2':x=x_*^2$ is given by (2.2.5), that is, is equal to $\int_{\mathrm{Opt}}^\infty\gamma(s)ds$ (note that e in (2.2.5) by construction is a $\|\cdot\|_2$-unit vector), that is, it is equal to the already justified upper bound (2.2.11) on the risk of the test $\mathcal{T}$ deciding on the larger than H_χ' composite hypotheses H_χ, $\chi=1,2$. □

2.2.2.3 Further extensions: spherical families of distributions

Now let us assume that we are in the situation of Section 2.2.1 with $L=2$ and nonempty closed, convex and non-intersecting X_1, X_2, one of the sets being bounded, exactly what we have assumed in Section 2.2.2.2. What we intend to do now, is to relax the restrictions on the family $\mathcal{P}$ of noise distributions, which in Section 2.2.2.2 was just a singleton with density which is a strictly decreasing function of the $\|\cdot\|_2$-norm. Observe that as far as the density $p(\cdot)$ of noise is concerned, justification of the upper risk bound (2.2.10) in Proposition 2.2.1 used the only fact that whenever $h\in\mathbf{R}^d$ is a $\|\cdot\|_2$-unit vector and $\delta\geq0$, we have $\int_{h^Tu\geq\delta}p(u)du\leq\int_\delta^\infty\gamma(s)ds$, with the even univariate probability density $\gamma(\cdot)$ specified in Proposition. We use this observation to extend our construction to *spherical families of probability densities*.

Spherical family of probability densities. Let $\gamma(\cdot)$ be an even probability density on the axis such that there is no neighbourhood of the origin where $\gamma=0$ almost surely. We associate with γ the *spherical family of densities* $\mathcal{P}_\gamma^d$ comprised of all probability densities $p(\cdot)$ on $\mathbf{R}^d$ such that

A. $p(\cdot)$ is even

B. Whenever $e \in \mathbf{R}^d$, $\|e\|_2 = 1$, and $\delta \geq 0$, we have

$$\mathrm{Prob}_{\xi\sim P}\{\xi : e^T\xi \geq \delta\} \leq \Gamma(\delta) := \int\limits_{\delta}^{\infty} \gamma(s)ds. \tag{2.2.12}$$

Geometrically: $p(\cdot)$-probability for $\xi \sim p(\cdot)$ to belong to a half-space not containing origin does not exceed $\Gamma(\delta)$, where δ is the $\|\cdot\|_2$-distance from the origin to the half-space.

Note that density (2.2.3) belongs to the family $\mathcal{P}^d_\gamma$ with $\gamma(\cdot)$ defined in Proposition 2.2.1; the resulting γ, in addition to being an even density, is strictly monotonically decreasing on the nonnegative ray. When speaking about general-type spherical families $\mathcal{P}^d_\gamma$, we do *not* impose monotonicity requirements on $\gamma(\cdot)$. If a spherical family $\mathcal{P}^d_\gamma$ includes a density $p(\cdot)$ of the form (2.2.3) *such that* $\gamma(\cdot)$ *is the induced by* $p(\cdot)$ *univariate marginal density*, as in Proposition 2.2.1, we say that $\mathcal{P}^d_\gamma$ *has a cap*, and this cap is $p(\cdot)$.

Example: Gaussian mixtures. Let $\eta \sim \mathcal{N}(0,\Theta)$, where the $d\times d$ covariance matrix Θ satisfies $\Theta \preceq I_d$, and let Z be an independent of η positive scalar random variable. *Gaussian mixture* of Z and η (or, better to say, of the distribution P_Z of Z and the distribution $\mathcal{N}(0,\Theta)$) is the probability distribution of the random vector $\xi = \sqrt{Z}\eta$. Examples of Gaussian mixtures include

- Gaussian distribution $\mathcal{N}(0,\Theta)$ (take Z identically equal to 1),
- multidimensional Student's t-distribution with $\nu \in \{1,2,...\}$ degrees of freedom and "covariance structure" Θ; here Z is given by the requirement that ν/Z has χ^2-distribution with ν degrees of freedom.

An immediate observation (check it!) is that with γ given by the distribution P_Z of Z according to

$$\gamma_Z(s) = \int_{z>0} \frac{1}{\sqrt{2\pi z}} \mathrm{e}^{-\frac{s^2}{2z}} P_Z(dz), \tag{2.2.13}$$

the distribution of random variable $\sqrt{Z}\eta$, with $\eta \sim \mathcal{N}(0,\Theta)$, $\Theta \preceq I_d$, independent of Z, belongs to the family $\mathcal{P}^d_{\gamma_Z}$, and the family $\mathcal{P}^d_{\gamma_Z}$ has a cap, specifically, the Gaussian mixture of P_Z and $\mathcal{N}(0,I_d)$.

Another example of this type: Gaussian mixture of a distribution P_Z of random variable Z taking values in $(0,1]$ and a distribution $\mathcal{N}(0,\Theta)$ with $\Theta \preceq I_d$ belongs to the spherical family $\mathcal{P}^d_{\gamma_\mathcal{G}}$ associated with the standard univariate Gaussian density

$$\gamma_\mathcal{G}(s) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-s^2/2};$$

This family has a cap, specifically, the standard Gaussian d-dimensional distribution $\mathcal{N}(0,I_d)$. Looking at the proof of Proposition 2.2.1, we arrive at the following

Proposition 2.2.2 *Let X_1, X_2 be nonempty and nonintersecting closed convex sets in $\mathbf{R}^d$, one of the sets being bounded, and let $\mathcal{P}^d_\gamma$ be a spherical family of probability distributions. With* Opt *and $\phi(\cdot)$ given by* (2.2.6) – (2.2.7)*, let us split the width* 2Opt *of the stripe $\{\omega : -\mathrm{Opt} \leq \phi(\omega) \leq \mathrm{Opt}\}$ separating X_1 and X_2 into two nonnegative parts:*

$$\delta_1 \geq 0, \delta_2 \geq 0,\ \delta_1 + \delta_2 = 2\mathrm{Opt} \tag{2.2.14}$$

and consider simple test $\mathcal{T}$ deciding on the hypotheses $H_1 : x \in X_1$, $H_2 : x \in X_2$ via observation (2.2.2) *by accepting H_1 when*

$$\phi(\omega) \geq \frac{1}{2}[\delta_2 - \delta_1]$$

and accepting H_2 otherwise. Then

$$\mathrm{Risk}_\chi(\mathcal{T}|H_1,H_2) \leq \int_{\delta_\chi}^{\infty} \gamma(s)ds,\ \chi=1,2 \tag{2.2.15}$$

In addition, when $\delta_1=\delta_2=\mathrm{Opt}$ and $\mathcal{P}_\gamma^d$ has a cap, $\mathcal{T}$ is the minimum risk test deciding on H_1, H_2. The risk of this test is

$$\mathrm{Risk}(\mathcal{T}|H_1,H_2)=\Gamma(\mathrm{Opt}):=\int_{\mathrm{Opt}}^{\infty}\gamma(s)ds. \tag{2.2.16}$$

To illustrate the power of Proposition 2.2.2, consider the case when γ is the function (2.2.13) stemming from Student's t-distribution on $\mathbf{R}^d$ with t degrees of freedom. It is known that in this case γ is the density of univariate Student's t-distribution with t degrees of freedom:

$$\gamma(s)=\frac{1}{\sqrt{\nu}\mathrm{B}(\frac{1}{2},\frac{\nu}{2})}(1+s^2/\nu)^{-\frac{\nu+1}{2}},$$

where $\mathrm{B}(\cdot,\cdot)$ is Beta function. When $\nu=1$, $\gamma_Z(\cdot)$ is just the heavy tailed (no expectation!) standard Cauchy density $\frac{1}{\pi}(1+s^2)^{-1}$. Same as in this "extreme case," multidimensional Student's distributions have relatively heavy tails (the heavier the less is ν) and as such are of interest in Finance.

2.2.3 Euclidean Separation and Repeated Observations

2.2.3.1 The simplest case

Assume that X_1, X_2, $\mathcal{P}_\gamma^d$ are as in the premise of Proposition 2.2.2 and K-repeated observations are allowed, $K>1$. An immediate attempt to reduce the situation to the single-observation case by calling K-repeated observation $\omega^K=(\omega_1,...,\omega_K)$ our new observation and thus reducing testing via repeated observations to the single-observation case seemingly fails: already in the simplest case of stationary K-repeated observations this reduction would require replacing the family $\mathcal{P}_\gamma^d$ with the family of product distributions $\underbrace{P\times...\times P}_{K}$ stemming from $P\in\mathcal{P}_\gamma^d$, and it is unclear how to apply to the resulting single-observation testing problem our machinery based on Euclidean separation. Instead, let us use *K-step majority test.*

2.2.3.2 Preliminaries: Repeated observations in "signal plus noise" observation noise

We are in the situation when our inference should be based on observations

$$\omega^K=(\omega_1,\omega_2,...,\omega_K), \tag{2.2.17}$$

and decide on hypotheses $\mathcal{H}_1$, $\mathcal{H}_2$ on the distribution Q^K of ω^K, and we are interested in the following 3 cases:

S [*stationary K-repeated observations*, cf. Section 2.1.3.1]: $\omega_1,...,\omega_K$ are drawn independently of each other from the same distribution Q, that is, Q^K is the product distribution $Q\times...\times Q$. Further, under hypothesis $\mathcal{H}_\chi$, $\chi=1,2$, Q is the distribution of random variable $\omega=x+\xi$, where $x\in X_\chi$ is deterministic, and the distribution P of ξ belongs to the family $\mathcal{P}_\gamma^d$;

SS [*semi-stationary K-repeated observations*, cf. Section 2.1.3.2]: there are two deterministic sequences, one of signals $\{x_k\}_{k=1}^K$, another of distributions $\{P_k\in\mathcal{P}_\gamma^d\}_{k=1}^K$, and $\omega_k=x_k+\xi_k$, $1\leq k\leq K$, with $\xi_k\sim P_k$ independent across k. Under hypothesis $\mathcal{H}_\chi$, *all* signals x_k, $k\leq K$, belong to X_χ.

QS [*quasi-stationary K-repeated observations*, cf. Section 2.1.3.3]: "in the nature" there exists a random sequence of driving factors $\zeta^k=(\zeta_1,...,\zeta_K)$ such that observation ω_k, for every k, is a deterministic function of $\zeta^k=(\zeta_1,...,\zeta_k)$: $\omega_k=\theta_k(\zeta^k)$. On the top of it, under ℓ-th hypothesis $\mathcal{H}_\ell$, for all $k\leq K$ and all ζ^{k-1} the conditional, ζ^{k-1} being given, distribution of ω_k belong to the family $\mathcal{P}_\ell$ of distributions of all random vectors of the form $x+\xi$, where $x\in X_\ell$ is deterministic, and ξ is random noise with distribution from $\mathcal{P}_\gamma^d$.

2.2.3.3 Majority Test

2.2.3.4 The simplest case

The construction of K-observation majority test is very natural.

Building block. We use Euclidean separation to build simple single-observation test $\mathcal{T}$ deciding on hypotheses $H_\chi: x\in X_\chi$, $\chi=1,2$, via observation $\omega=x+\xi$, where x is deterministic, and the distribution of noise ξ belongs to $\mathcal{P}_\gamma^d$. $\mathcal{T}$ is given by the construction from Proposition 2.2.2 applied with $\delta_1=\delta_2=\mathrm{Opt}$. The summary of our actions is as follows:

$$\begin{array}{cc}
X_1,X_2\Rightarrow & \left\{\begin{array}{c}\mathrm{Opt}=\min_{x^1\in X_1,x^2\in X_2}\frac{1}{2}\|x^1-x^2\|_2\\ (x^1_*,x^2_*)\in\mathrm{Argmin}_{x^1\in X_1,x^2\in X_2}\frac{1}{2}\|x^1-x^2\|_2\end{array}\right.\\
\Rightarrow & e=\frac{x^1_*-x^2_*}{\|x^1_*-x^2_*\|_2},\ c=\frac{1}{2}e^T[x^1_*+x^2_*]\\
\Rightarrow & \phi(\omega)=e^T\omega-c
\end{array}\tag{2.2.18}$$

Majority test $\mathcal{T}_K^{\mathrm{maj}}$, as applied to K-repeated observation $\omega^K=(\omega_1,...,\omega_K)$ builds the K reals $v_k=\phi(\omega_k)$. If at least $K/2$ of these reals are nonnegative, the test accepts $\mathcal{H}_1$ and rejects $\mathcal{H}_2$; otherwise the test accepts $\mathcal{H}_2$ and rejects $\mathcal{H}_1$.

2.2.3.5 Risk analysis

We intend to carry out the risk analysis for the case **QS** of quasi-stationary K-repeated observations; this analysis automatically applies to the cases **SS** of stationary and **S** of K-repeated stationary/semi-stationary observations, which are special cases of **QS**.

Proposition 2.2.3 *With $X_1,X_2,\mathcal{P}_\gamma^d$ obeying the premise of Proposition 2.2.2, in the case* **OS** *of quasi-stationary observations the risk of K-observation Majority test $\mathcal{T}_K^{\mathrm{maj}}$ can be bounded as*

$$\mathrm{Risk}(\mathcal{T}_K^{\mathrm{maj}}|\mathcal{H}_1,\mathcal{H}_2)\leq\epsilon_K\equiv\sum_{K/2\leq k\leq K}\binom{K}{k}\epsilon_\star^k(1-\epsilon_\star)^{K-k},\ \epsilon_\star=\int_{\mathrm{Opt}}^{\infty}\gamma(s)ds.\tag{2.2.19}$$

Proof. *Here we restrict ourselves to the case* **SS** *of semi-stationary K-repeated observations.* In "full generality," that is, in the case **QS** of quasi-stationary K-repeated observations, Proposition will be proved in Section 2.2.5.

Assume that $\mathcal{H}_1$ takes place, so that (recall that we are in the **SS** case!) $\omega_k=x_k+\xi_k$ with some deterministic $x_k\in X_1$ and independent across k noises $\xi_k\sim P_k$, for some deterministic sequence

$P_k \in \mathcal{P}^d_\gamma$. Let us fix $\{x_k \in X_1\}_{k=1}^K$ and $\{P_k \in \mathcal{P}^d_\gamma\}_{k=1}^K$. Then the random reals $v_k = \phi(\omega_k = x_k + \xi_k)$ are independent across k, and so are the Boolean random variables

$$\chi_k = \left\{ \begin{array}{ll} 1, & v_i < 0 \\ 0, & v_i \geq 0 \end{array} \right.$$

$\chi_k = 1$ if and only if test $\mathcal{T}$, as applied to observation ω_k, rejects hypothesis $H_1 : x_k \in X_1$. By Proposition 2.2.2, P_k-probability p_k of the event $\chi_k = 1$ is at most $\epsilon_\star$. Further, by construction of the Majority test, if $\mathcal{T}_K^{\rm maj}$ rejects the true hypothesis $\mathcal{H}_1$, then the number of k's with $\chi_k = 1$ is $\geq K/2$. Thus, with $\{x_k \in X_1\}$ and $P_k \in \mathcal{P}^d_\gamma$ the probability to reject $\mathcal{H}_1$ is not greater than the probability of the event

> In K independent coin tosses, with probability $p_k \leq \epsilon_*$ to get head in k-th toss, the total number of heads is $\geq K/2$.

The probability of this event clearly does not exceed the right hand side in (2.2.19), implying that $\text{Risk}_1(\mathcal{T}_K^{\rm maj}|\mathcal{H}_1, \mathcal{H}_2) \leq \epsilon_k$. "Symmetric" reasoning yields $\text{Risk}_2(\mathcal{T}_K^{\rm maj}|\mathcal{H}_1, \mathcal{H}_2) \leq \epsilon_k$, completing the proof of (2.2.19). □

Corollary 2.2.1 *Under the premise of Proposition 2.2.3, the upper bounds ϵ_K on the risk of the K-observation Majority test goes to 0 exponentially fast as $K \to \infty$.*

Indeed, we are in the situation of Opt > 0, so that $\epsilon_\star < \frac{1}{2}$ [3].

Remark 2.2.1 When proving (**SS**-version of) Proposition 2.2.3, we have used "evident" observation as follows:

> (#) Let χ_1,..., χ_K be independent random variables taking values 0 and 1, and let the probabilities p_k for χ_k to be 1 be upper-bounded by some $\epsilon \in [0, 1]$ for all k. Then for every fixed M the probability of the event *"at least M of $\chi_1, ..., \chi_K$ are equal to 1"* is upper-bounded by the probability $\sum_{M \leq k \leq K} \binom{K}{k} \epsilon^k (1 - \epsilon)^{K-k}$ of the same event in the case when $p_k = \epsilon$ for all k.

If there are evident facts in Math, (#) definitely is one of them. Nevertheless, why (#) is true? Reader is kindly asked to prove (#). For your information: design of proof took about 10-minute effort of the authors (a bit too much for an evident statement); the results of their effort can be found in Section 2.2.5.

2.2.4 From Pairwise to Multiple Hypotheses Testing

2.2.4.1 Situation

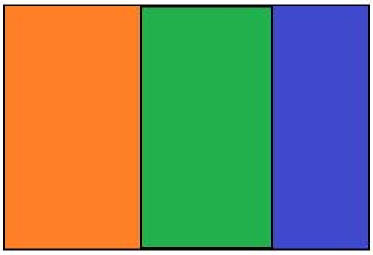
Assume we are given L families of probability distributions $\mathcal{P}_\ell$, $1 \leq \ell \leq L$, on observation space Ω, and observe a realization of random variable $\omega \sim P$ taking values in Ω. Given ω, we want to decide on the L hypotheses

$$H_\ell : P \in \mathcal{P}_\ell, \ 1 \leq \ell \leq L. \tag{2.2.20}$$

Our *ideal goal* would be to find a low-risk simple test deciding on the hypotheses. However, it may happen that the " ideal goal" is not achievable, for example, when some pairs of families $\mathcal{P}_\ell$ have nonempty intersections. When $\mathcal{P}_\ell \cap \mathcal{P}_{\ell'} \neq \emptyset$ for some $\ell \neq \ell'$, there is no way to decide on the hypotheses with risk $< 1/2$.

[3] Recall that we have assumed from the very beginning that γ is an even probability density on the axis, and there is no neighbourhood of the origin where $\gamma = 0$ a.s.

But: *Impossibility to decide reliably on all L hypotheses "individually" does not mean that no meaningful inferences can be done.*
For example, consider the 3 colored rectangles on the plane:



and 3 hypotheses, with H_ℓ, $1 \le \ell \le 3$, stating that our observation is $\omega = x + \xi$ with deterministic "signal" x belonging to ℓ-th rectangle and $\xi \sim \mathcal{N}(0, \sigma^2 I_2)$. Whatever small σ be, no test can decide on the 3 hypotheses with risk $< 1/2$; e.g., there is no way to decide reliably on H_1 vs. H_2. However, we may hope that when σ is small (or when repeated observations are allowed), observations allow us to discard reliably some of the hypotheses; for example, when the signal "is brown" (i.e., H_1 holds true), we hardly can discard reliably the hypothesis H_2 stating that the signal "is green," but hopefully can discard reliably H_3 (that is, infer that the signal is not blue).

When handling multiple hypotheses which cannot be reliably decided upon "as they are," it makes sense to speak about *testing the hypotheses "up to closeness."*

2.2.4.2 Closeness relation and "up to closeness" risks

Closeness relation, or simply *closeness* $\mathcal{C}$ on a collection of L hypotheses $H_1, ..., H_L$ is defined as some *set of pairs* (ℓ, ℓ') with $1 \le \ell, \ell' \le L$. We interpret the relation $(\ell, \ell') \in \mathcal{C}$ as the fact that the hypotheses H_ℓ and H'_ℓ are close to each other. Sometimes we shall use the words "ℓ and ℓ' are/are not $\mathcal{C}$-close to each other" as an equivalent form of "hypotheses H_ℓ, $H_{\ell'}$ are/are not $\mathcal{C}$-close to each other."
We always assume that

- $\mathcal{C}$ contains all "diagonal pairs" (ℓ, ℓ), $1 \le \ell \le L$ ("every hypothesis is close to itself");
- $(\ell, \ell') \in \mathcal{C}$ is and only if $(\ell', \ell) \in \mathcal{C}$ ("closeness is a symmetric relation").

Note that by symmetry of $\mathcal{C}$, the relation $(\ell, \ell') \in \mathcal{T}$ is in fact a property of *un*ordered pair $\{\ell, \ell'\}$.

"Up to closeness" risks. Let $\mathcal{T}$ be a test deciding on L hypotheses $H_1, ..., H_L$, see (2.2.20); given observation ω, $\mathcal{T}$ accepts all hypotheses H_ℓ with indexes $\ell \in \mathcal{T}(\omega)$ and rejects all other hypotheses. We say that *ℓ-th partial $\mathcal{C}$-risk of test $\mathcal{T}$ is* $\le \epsilon$, if *whenever H_ℓ is true: $\omega \sim P \in \mathcal{P}_\ell$, the P-probability of the event*

$$\boxed{\begin{array}{c}\mathcal{T} \textit{ accepts } H_\ell\textit{:}\;\; \ell \in \mathcal{T}(\omega)\\ \text{and}\\ \textit{all hypotheses } H_{\ell'} \textit{ accepted by } \mathcal{T} \textit{ are } \mathcal{C}\textit{-close to } H_\ell\textit{:}\;\; (\ell, \ell') \in \mathcal{C}, \forall \ell' \in \mathcal{T}(\omega)\end{array}}$$

is *at least* $1 - \epsilon$.

ℓ-th partial $\mathcal{C}$-risk $\mathrm{Risk}^{\mathcal{C}}_\ell(\mathcal{T}|H_1, ..., H_L)$ *of $\mathcal{T}$* is the smallest ϵ with the outlined property, or, equivalently,

$$\mathrm{Risk}^{\mathcal{C}}_\ell(\mathcal{T}|H_1, ..., H_L) = \sup_{P \in \mathcal{P}_\ell} \mathrm{Prob}_{\omega \sim P}\left\{[\ell \notin \mathcal{T}(\omega)] \text{ or } [\exists \ell' \in \mathcal{T}(\omega) : (\ell, \ell') \notin \mathcal{C}]\right\}$$

$\mathcal{C}$-risk $\mathrm{Risk}^{\mathcal{C}}(\mathcal{T}|H_1,...,H_L)$ *of* $\mathcal{T}$ is the largest of the partial $\mathcal{C}$-risks of the test:

$$\mathrm{Risk}^{\mathcal{C}}(\mathcal{T}|H_1,...,H_L) = \max_{1\leq\ell\leq L} \mathrm{Risk}^{\mathcal{C}}_{\ell}(\mathcal{T}|H_1,...,H_L).$$

Observe that when $\mathcal{C}$ is the "strictest possible" closeness, that is, $(\ell,\ell')\in\mathcal{C}$ if and only if $\ell=\ell'$, then a test $\mathcal{T}$ deciding on $H_1,...,H_L$ up to closeness $\mathcal{C}$ with risk ϵ is, *basically*, the same as a simple test deciding on $H_1,...,H_L$ with risk $\leq\epsilon$. Indeed, a test with the latter property clearly decides on $H_1,...,H_L$ with $\mathcal{C}$-risk $\leq\epsilon$. The inverse statement, *taken literally*, is not true, since even with our "as strict as possible" closeness, a test $\mathcal{T}$ with $\mathcal{C}$-risk $\leq\epsilon$ not necessarily is simple. However, we can enforce $\mathcal{T}$ to be simple, specifically, to accept a once for ever fixed hypothesis, say, H_1, and only it, when the set of hypotheses accepted by $\mathcal{T}$ "as is" is not a singleton, otherwise accept exactly the same hypothesis as $\mathcal{T}$. The modified test already is simple, and clearly its $\mathcal{C}$-risk does not exceed the one of $\mathcal{T}$.

2.2.4.3 Multiple Hypothesis Testing via pairwise tests

Assume that for every *un*ordered pair $\{\ell,\ell'\}$ with $(\ell,\ell')\not\in\mathcal{C}$ we are given a *simple* test $\mathcal{T}_{\{\ell,\ell'\}}$ deciding on H_ℓ vs. $H_{\ell'}$ via observation ω.
Our goal is to "assemble" the tests $\mathcal{T}_{\{\ell,\ell'\}}$, $(\ell,\ell')\not\in\mathcal{C}$, into a test $\mathcal{T}$ deciding on $H_1...,H_L$ up to closeness $\mathcal{C}$.

The construction we intend to use is as follows:

- For $1\leq\ell,\ell'\leq L$, we define functions $T_{\ell\ell'}(\omega)$ as follows:
  - when $(\ell,\ell')\in\mathcal{C}$, we set $T_{\ell\ell'}(\cdot)\equiv 0$.
  - when $(\ell,\ell')\not\in\mathcal{C}$, so that $\ell\neq\ell'$, we set

$$T_{\ell\ell'}(\omega) = \left\{\begin{array}{rl} 1, & \mathcal{T}_{\{\ell,\ell'\}}(\omega)=\{\ell\} \\ -1, & \mathcal{T}_{\{\ell,\ell'\}}(\omega)=\{\ell'\}\end{array}\right. . \tag{2.2.21}$$

    Note that $\mathcal{T}_{\{\ell,\ell'\}}$ is a simple test , so that $T_{\ell\ell'}(\cdot)$ is well defined and takes values ±1 when $(\ell,\ell')\not\in\mathcal{C}$ and 0 when $(\ell,\ell')\in\mathcal{C}$.

  Note that by construction and since $\mathcal{C}$ is symmetric, we have

$$T_{\ell\ell'}(\omega)\equiv -T_{\ell'\ell}(\omega),\ 1\leq\ell,\ell'\leq L. \tag{2.2.22}$$

- The test $\mathcal{T}$ is as follows: *given observation ω, we build the $L\times L$ matrix $T(\omega)=[T_{\ell\ell'}(\omega)]$ and accept exactly those of the hypotheses H_ℓ for which ℓ-th row in $T(\omega)$ is nonnegative.*

Observation 2.2.1 *When $\mathcal{T}$ accepts some hypothesis H_ℓ, all hypotheses accepted by $\mathcal{T}$ are $\mathcal{C}$-close to H_ℓ.*

Indeed, if ω is such that $\ell\in\mathcal{T}(\omega)$, then the ℓ-th row in $T(\omega)$ is nonnegative. If now ℓ' is *not* $\mathcal{C}$-close to ℓ, we have $T_{\ell\ell'}(\omega)\geq 0$ and $T_{\ell\ell'}(\omega)\in\{-1,1\}$, whence $T_{\ell\ell'}(\omega)=1$. Consequently, by (2.2.22) it holds $T_{\ell'\ell}(\omega)=-1$, implying that ℓ'-th row in $T(\omega)$ is *not* nonnegative, and thus $\ell'\not\in\mathcal{T}(\omega)$. □

Risk analysis. For $(\ell,\ell')\not\in\mathcal{C}$, let

$$\begin{array}{rcl}\epsilon_{\ell\ell'} &=& \mathrm{Risk}_1(\mathcal{T}_{\{\ell,\ell'\}}|H_\ell,H_{\ell'}) = \sup\limits_{P\in\mathcal{P}_\ell}\mathrm{Prob}_{\omega\sim P}\{\ell\not\in\mathcal{T}_{\{\ell,\ell'\}}(\omega)\}\\ &=& \sup\limits_{P\in\mathcal{P}_\ell}\mathrm{Prob}_{\omega\sim P}\{T_{\ell\ell'}(\omega)=-1\} = \sup\limits_{P\in\mathcal{P}_\ell}\mathrm{Prob}_{\omega\sim P}\{T_{\ell'\ell}(\omega)=1\}\\ &=& \sup\limits_{P\in\mathcal{P}_\ell}\mathrm{Prob}_{\omega\sim P}\{\ell'\in\mathcal{T}_{\{\ell,\ell'\}}(\omega)\}\\ &=& \mathrm{Risk}_2(\mathcal{T}_{\{\ell,\ell'\}}|H_{\ell'},H_\ell).\end{array}\tag{2.2.23}$$

Proposition 2.2.4 *For the just defined test $\mathcal{T}$ it holds*

$$\forall \ell\leq L:\mathrm{Risk}_\ell^{\mathcal{C}}(\mathcal{T}|H_1,...,H_L)\leq\epsilon_\ell:=\sum_{\ell':(\ell,\ell')\not\in\mathcal{C}}\epsilon_{\ell\ell'}.\tag{2.2.24}$$

Proof. Let us fix ℓ, let H_ℓ be true, and let $P\in\mathcal{P}_\ell$ be the distribution of observation ω. Set $I=\{\ell'\leq L:(\ell,\ell')\not\in\mathcal{C}\}$. For $\ell'\in I$, let $E_{\ell'}$ be the event

$$\{\omega:T_{\ell\ell'}(\omega)=-1\}.$$

We have $\mathrm{Prob}_{\omega\sim P}(E_{\ell'})\leq\epsilon_{\ell\ell'}$ (by definition of $\epsilon_{\ell\ell'}$), whence

$$\mathrm{Prob}_{\omega\sim P}\big(\underbrace{\cup_{\ell'\in I}E_{\ell'}}_{E}\big)\leq\epsilon_\ell.$$

When the event E does *not* take place, we have $T_{\ell\ell'}(\omega)=1$ for all $\ell'\in I$, so that $T_{\ell\ell'}(\omega)\geq0$ for all ℓ', $1\leq\ell'\leq L$, whence $\ell\in\mathcal{T}(\omega)$. By Observation 2.2.1, the latter inclusion implies that

$$\{\ell\in\mathcal{T}(\omega)\,\&\,\{(\ell,\ell')\in\mathcal{C}\,\forall\ell'\in\mathcal{T}(\omega)\}.$$

Invoking the definition of partial $\mathcal{C}$-risk, we get

$$\mathrm{Risk}_\ell^{\mathcal{C}}(\mathcal{T}|H_1,...,H_L)\leq\mathrm{Prob}_{\omega\sim P}(E)\leq\epsilon_\ell.$$ □

2.2.4.4 Testing Multiple Hypotheses via Euclidean separation

Situation. We are given L nonempty and closed convex sets $X_\ell\subset\Omega=\mathbf{R}^d$, $1\leq\ell\leq L$, with at least $L-1$ of the sets being bounded, and a spherical family of probability distributions $\mathcal{P}_\gamma^d$. These data define L families $\mathcal{P}_\ell$ of probability distributions on $\mathbf{R}^d$; the family $\mathcal{P}_\ell$, $1\leq\ell\leq L$, is comprised of probability distributions of all random vectors of the form $x+\xi$, where deterministic x ("signal") belongs to X_ℓ, and ξ is random noise with distribution from $\mathcal{P}_\gamma^d$. Given positive integer K, we can speak about L hypotheses on the distribution P^K of K-repeated observation $\omega^K=(\omega_1,...,\omega_K)$, with $\mathcal{H}_\ell$ stating that ω^K is a quasi-stationary K-repeated observation associated with $\mathcal{P}_\ell$. In other words $\mathcal{H}_\ell=H_\ell^{\otimes,K}$, see Section 2.1.3.3. Finally, we are given a closeness $\mathcal{C}$.

Our goal is to decide on the hypotheses $\mathcal{H}_1,...,\mathcal{H}_L$ up to closeness $\mathcal{C}$ via K-repeated observation ω^K. Note that this is a natural extension of the case **QS** of pairwise testing from repeated observations considered in Section 2.2.3 (there $L=2$ and $\mathcal{C}$ is the only meaningful closeness on a two-hypotheses set: $(\ell,\ell')\in\mathcal{C}$ is and only if $\ell=\ell'$).

Standing Assumption which is by default in force everywhere in this Section is:

Whenever ℓ,ℓ' are not $\mathcal{C}$-close: $(\ell,\ell')\not\in\mathcal{C}$, the sets X_ℓ, $X_{\ell'}$ do not intersect.

Strategy: We intend to attack the above testing problem by assembling pairwise Euclidean separation Majority tests via the construction from Section 2.2.4.3.

Building blocks to be assembled are Euclidean separation K-observation pairwise Majority tests built for the pairs $\mathcal{H}_\ell$, $\mathcal{H}_{\ell'}$ of hypotheses with *not* close to each other ℓ and ℓ', that is, with $(\ell,\ell')\not\in\mathcal{C}$. These tests are built as explained in Section 2.2.3.3; for reader's convenience, here is the construction. For a pair $(\ell,\ell')\not\in\mathcal{C}$, we

1. Find the optimal value $\mathrm{Opt}_{\ell\ell'}$ and an optimal solution $(u_{\ell\ell'}, v_{\ell\ell'})$ to the convex optimization problem
$$\mathrm{Opt}_{\ell\ell'}=\min_{u\in X_\ell, v\in X_{\ell'}}\frac{1}{2}\|u-v\|_2, \tag{2.2.25}$$
The latter problem is solvable, since we have assumed from the very beginning that X_ℓ, X'_ℓ are nonempty, closed and convex, and that at least one of these sets is bounded;

2. Set
$$e_{\ell\ell'}=\frac{u_{\ell\ell'}-v_{\ell\ell'}}{\|u_{\ell\ell'}-v_{\ell\ell'}\|_2},\ c_{\ell\ell'}=\frac{1}{2}e^T_{\ell\ell'}[u_{\ell\ell'}+v_{\ell\ell'}],\ \phi_{\ell\ell'}(\omega)=e^T_{\ell\ell'}\omega-c_{\ell\ell'}.$$
Note that the construction makes sense, since by our Standing Assumption for ℓ, ℓ' in question X_ℓ and $X_{\ell'}$ do not intersect. Further, $e_{\ell\ell'}$ and $c_{\ell\ell'}$ clearly depend solely on (ℓ,ℓ'), but not on how we select an optimal solution $(u_{\ell\ell'}, v_{\ell\ell'})$ to (2.2.25). Finally, we have
$$e_{\ell\ell'}=-e_{\ell'\ell}, c_{\ell\ell'}=-c_{\ell'\ell}, \phi_{\ell\ell'}(\cdot)\equiv-\phi_{\ell'\ell}(\cdot).$$

3. We consider separately the case of $K=1$ and the case of $K>1$. Specifically,

   (a) when $K=1$, we select somehow nonnegative reals $\delta_{\ell\ell'}$, $\delta_{\ell'\ell}$ such that
$$\delta_{\ell\ell'}+\delta_{\ell'\ell}=2\mathrm{Opt}_{\ell\ell'} \tag{2.2.26}$$
and specify the single-observation simple test $\mathcal{T}_{\ell\ell'}$ deciding on the hypotheses $\mathcal{H}_\ell$, $\mathcal{H}_{\ell'}$ according to
$$\mathcal{T}_{\ell\ell'}(\omega)=\left\{\begin{array}{ll}\{\ell\}, & \phi_{\ell\ell'}(\omega)\geq\frac{1}{2}[\delta_{\ell'\ell}-\delta_{\ell\ell'}]\\ \{\ell'\}, & \text{otherwise}\end{array}\right.;$$
Note that by Proposition 2.2.2, setting
$$\Gamma(\delta)=\int_\delta^\infty\gamma(s)ds, \tag{2.2.27}$$
we have
$$\begin{array}{rcl}\mathrm{Risk}_1(\mathcal{T}_{\ell\ell'}|\mathcal{H}_\ell,\mathcal{H}_{\ell'}) & \leq & \Gamma(\delta_{\ell\ell'})\\ \mathrm{Risk}_2(\mathcal{T}_{\ell\ell'}|\mathcal{H}_\ell,\mathcal{H}_{\ell'}) & \leq & \Gamma(\delta_{\ell'\ell})\\ \mathrm{Risk}_1(\mathcal{T}_{\ell'\ell}|\mathcal{H}_{\ell'},\mathcal{H}_\ell) & \leq & \Gamma(\delta_{\ell'\ell})\\ \mathrm{Risk}_2(\mathcal{T}_{\ell'\ell}|\mathcal{H}_{\ell'},\mathcal{H}_\ell) & \leq & \Gamma(\delta_{\ell\ell'})\end{array} \tag{2.2.28}$$

   (b) when $K>1$, we specify K-observation simple test $\mathcal{T}_{\ell\ell'K}$ deciding on $\mathcal{H}_\ell$, $\mathcal{H}_{\ell'}$ according to
$$\mathcal{T}_{\ell\ell'}(\omega^k=(\omega_1,...,\omega_k))=\left\{\begin{array}{ll}\{\ell\}, & \mathrm{Card}\{k\leq K:\phi_{\ell\ell'}\geq 0\}\geq K/2,\\ \{\ell'\}, & \text{otherwise}\end{array}\right..$$
Note that by Proposition 2.2.3 we have
$$\mathrm{Risk}(\mathcal{T}_{\ell\ell'K}|\mathcal{H}_\ell,\mathcal{H}'_\ell)\leq\epsilon_{\ell\ell'K}:=\sum_{K/2\leq k\leq K}\binom{K}{k}\epsilon^k_{\star\ell\ell'}(1-\epsilon_{\star\ell\ell'})^{K-k},\ \epsilon_{\star\ell\ell'}=\Gamma(\mathrm{Opt}_{\ell\ell'})=\epsilon_{\star\ell'\ell}. \tag{2.2.29}$$

Assembling building blocks, case of $K=1$. In the case of $K=1$, we specify the simple pairwise tests $\mathcal{T}_{\{\ell,\ell'\}}$, $(\ell,\ell')\not\in\mathcal{C}$, participating in the construction of the multi-hypothesis test presented in Section 2.2.4.3, as follows. Given unordered pair $\{\ell,\ell'\}$ with $(\ell,\ell')\not\in\mathcal{C}$ (which is exactly the same as $(\ell',\ell)\not\in\mathcal{C}$), we arrange ℓ,ℓ' in ascending order, thus arriving at ordered pair $(\bar\ell,\bar\ell')$, and set

$$\mathcal{T}_{\{\ell,\ell'\}}(\cdot)=\mathcal{T}_{\bar\ell\bar\ell'}(\cdot),$$

with the right hand side tests defined as explained above. We then assemble, as explained in Section 2.2.4.3, the tests $\mathcal{T}_{\{\ell,\ell'\}}$ into a single-observation test $\mathcal{T}_1$ deciding on hypotheses $\mathcal{H}_1,...,\mathcal{H}_L$. Looking at (2.2.23) and (2.2.28), we conclude that for the just defined tests $\mathcal{T}_{\{\ell,\ell'\}}$ and the associated with the tests $\mathcal{T}_{\{\ell,\ell'\}}$, via (2.2.23), quantities $\epsilon_{\ell\ell'}$ it holds

$$(\ell,\ell')\not\in\mathcal{C}\Rightarrow\epsilon_{\ell\ell'}\leq\Gamma(\delta_{\ell\ell'}).\tag{2.2.30}$$

Invoking Proposition 2.2.4, we get

Proposition 2.2.5 *In the situation described in the beginning of Section 2.2.4.4 and under Standing Assumption, the $\mathcal{C}$-risks of the just defined test $\mathcal{T}_1$, whatever be the choice of nonnegative $\delta_{\ell\ell'}$, $(\ell,\ell')\not\in\mathcal{C}$, satisfying* (2.2.26), *can be upper-bounded as*

$$\mathrm{Risk}^{\mathcal{C}}_\ell(\mathcal{T}_1|\mathcal{H}_1,...,\mathcal{H}_L)\leq\sum_{\ell':(\ell,\ell')\not\in\mathcal{C}}\Gamma(\delta_{\ell\ell'}).\tag{2.2.31}$$

with $\Gamma(\cdot)$ *given by* (2.2.27).

Case of $K=1$ (continued): Optimizing the construction. We can try to optimize the risk bounds (2.2.31) over the parameters $\delta_{\ell\ell'}$ of the construction. The first question to be addressed here is what to minimize – we have several risks! A natural model here is as follows. Let us fix a nonnegative $M\times L$ *weight matrix* W and M-dimensional positive *profile vector* w, and solve the optimization problem

$$\min_{t,\{\delta_{\ell\ell'}:(\ell,\ell')\not\in\mathcal{C}}\left\{t:\begin{array}{l}W\cdot\left[\sum_{\ell':(\ell,\ell')\not\in\mathcal{C}}\Gamma(\delta_{\ell\ell'})\right]_{\ell=1}^L\leq tw\\ \delta_{\ell\ell'}\geq0,\delta_{\ell\ell'}+\delta_{\ell'\ell}=2\mathrm{Opt}_{\ell\ell'},\ (\ell,\ell')\not\in\mathcal{C}\end{array}\right\}.\tag{2.2.32}$$

For example, when $M=1$ and $w=1$, we are minimizing weighted sum of (upper bounds on) partial $\mathcal{C}$-risks of our test, and when W is a diagonal matrix with positive diagonal entries and w is the all-ones vector, we are minimizing the largest of scaled partial risks. Note that when $\Gamma(\cdot)$ is convex on $\mathbf{R}_+$, or, which is the same, $\gamma(\cdot)$ is nonincreasing in $\mathbf{R}_+$, (2.2.32) is a convex, and thus efficiently solvable, problem.

Assembling building blocks, case of $K>1$. We again pass from our building blocks – K-observation simple pairwise tests $\mathcal{T}_{\ell\ell' K}$, $(\ell,\ell')\not\in\mathcal{C}$, we have already specified, to tests $\mathcal{T}_{\{\ell,\ell'\}}=\mathcal{T}_{\bar\ell\bar\ell' K}$, with $\bar\ell=\min[\ell,\ell']$ and $\bar\ell'=\max[\ell,\ell']$, and then apply to the resulting tests the construction from Section 2.2.4.3, arriving at K-observation multi-hypothesis test $\mathcal{T}_K$. By Proposition 2.2.3, the quantities $\epsilon_{\ell\ell'}$ associated with the tests $\mathcal{T}_{\{\ell,\ell'\}}$ via (2.2.23) satisfy the relation

$$(\ell,\ell')\not\in\mathcal{C}\Rightarrow\epsilon_{\ell\ell'}\leq\sum_{K/2\leq k\leq K}\binom{K}{k}[\Gamma(\mathrm{Opt}_{\ell\ell'})]^k[1-\Gamma(\mathrm{Opt}_{\ell\ell'})]^{K-k},\tag{2.2.33}$$

which combines with Proposition 2.2.4 to imply

Proposition 2.2.6 *Let the situation described in the beginning of Section 2.2.4.4 take place, and let $K > 1$. Under Standing Assumption, the $\mathcal{C}$-risks of the just defined test $\mathcal{T}_K$ can be upper-bounded as*

$$\mathrm{Risk}_\ell^{\mathcal{C}}(\mathcal{T}_1|\mathcal{H}_1,...,\mathcal{H}_L) \leq \sum_{\ell':(\ell,\ell')\notin\mathcal{C}} \Gamma(\delta_{\ell\ell'}) \sum_{K/2\leq k\leq K} \binom{K}{k} [\Gamma(\mathrm{Opt}_{\ell\ell'})]^k [1-\Gamma(\mathrm{Opt}_{\ell\ell'})]^{K-k}, \qquad (2.2.34)$$

with $\Gamma(\cdot)$ given by (2.2.27) and $\mathrm{Opt}_{\ell\ell'}$ given by (2.2.25).

Note that by Standing Assumption the quantities $\Gamma(\mathrm{Opt}_{\ell\ell'})$ are $< 1/2$, so that the risks $\mathrm{Risk}_\ell^{\mathcal{C}}(\mathcal{T}_K|H_1,...,H_L)$ go to 0 exponentially fast as $K\to\infty$.

2.2.5 Paying debts: missing proofs

2.2.5.1 Quasi-stationary K-repeated observations

Proof of Claim in Remark 2.2.1. What we should prove is that is $p=[p_1;...;p_K]\in B=[0,1]^K$, then the probability $P_M(p)$ of the event

> The total number of heads in K independent coin tosses, with probability p_k to get head in k-th toss, is at least M

is a nondecreasing function of p: if $p'\leq p''$, $p',p''\in B$, then $P_M(p')\leq P_M(p'')$. To see it, let us associate with $p\in B$ a subset of B, specifically, $B_p=\{x\in B: 0\leq x_k\leq p_k, 1\leq k\leq K\}$, and a function $\chi_p(x): B\to\{0,1\}$ which is equal to 0 at every point $x\in B$ where the number of entries x_k satisfying $x_k\leq p_k$ is less than M, and is equal to 1 otherwise. It is immediately seen that

$$P_M(p)\equiv\int_B \chi_p(x)dx \qquad (2.2.35)$$

(since with respect to the uniform distribution on B, the events $E_k=\{x\in B: x_k\leq p_k\}$ are independent across k and have probabilities p_k, and the right hand side in (2.2.35) is exactly the probability, taken w.r.t. the uniform distribution on B, of the event "at least M of the events E_1,..., E_K take place"). But the right hand side in (2.2.20) clearly is nondecreasing in $p\in B$, since χ_p, by construction, is the characteristic function of the set

$$B[p]=\{x: \text{ at least } M \text{ of the entries } x_k \text{ in } x \text{ satisfy } x_k\leq p_k\},$$

and these sets clearly grow when p is entrywise increased. □

Proof of Proposition 2.2.3 in the case of quasi-stationary K-repeated observations

Situation and goal. We are in the case **QS**, see Section 2.2.3.2, of the situation described in the beginning of Section 2.2.3; it suffices to verify that if $\mathcal{H}_\ell$, $\ell\in\{1,2\}$, is true, then the probability for $\mathcal{T}_K^{\mathrm{maj}}$ to reject $\mathcal{H}_\ell$ is at most the quantity ϵ_K defined in (2.2.19). Let us verify this statement in the case of $\ell=1$; the reasoning for $\ell=2$ "mirrors" the one to follow.

It is clear that our situation and goal can be formulated as follows:

- "In the nature" there exists a random sequence $\zeta^K=(\zeta_1,...,\zeta_K)$ of driving factors and collection of deterministic functions $\theta_k(\zeta^k=(\zeta_1,...,\zeta_k))$ [4] such that our k-th observation is $\omega_k=\theta_k(\zeta^k)$. Besides this, the conditional, ζ^{k-1} given, distribution $P_{\omega_k|\zeta^{k-1}}$ of ω_k always belongs to the family $\mathcal{P}_1$ comprised of distributions of random vectors of the form $x+\xi$, where deterministic x belongs to X_1 and the distribution of ξ belongs to $\mathcal{P}_\gamma^d$.

[4] as always, given a K-element sequence, say, $\zeta_1,...,\zeta_K$, we write ζ^t, $t\leq K$, as a shorthand for the fragment $\zeta_1,...,\zeta_t$ of this sequence.

- There exist deterministic functions $\chi_k : \Omega \to \{0,1\}$ and integer M, $1 \leq M \leq K$, such that the test $\mathcal{T}_K^{\rm maj}$, as applied to observation $\omega^K = (\omega_1, ..., \omega_K)$, rejects $\mathcal{H}_1$ if and only if the number of ones among the quantities $\chi_k(\omega_k)$, $1 \leq k \leq K$, is at least M.

  In the situation of Proposition 2.2.3, $M = \rfloor K/2 \lfloor$ and $\chi_k(\cdot)$ are in fact independent of k: $\chi_k(\omega) = 1$ if and only if $\phi(\omega) \leq 0$ [5].

- What we know is that the conditional, ζ^{k-1} being given, probability of the event $\chi_k(\omega_k = \theta_k(\zeta^k)) = 1$ is at most $\epsilon_\star$:

  $$P_{\omega_k|\zeta^{k-1}}\{\omega_k : \chi_k(\omega_k) = 1\} \leq \epsilon_\star \, \forall \zeta^{k-1}.$$

  Indeed, $P_{\omega_k|\zeta^{k-1}} \in \mathcal{P}_1$, that is $P_{\omega_k|\zeta^{k-1}} \in \mathcal{P}_1$. As a result,

  $$P_{\omega_k|\zeta^{k-1}}\{\omega_k : \phi_k(\omega_k) = 1\} = P_{\omega_k|\zeta^{k-1}}\{\omega_k : \phi(\omega_k) \leq 0\} = P_{\omega_k|\zeta^{k-1}}\{\omega_k : \phi(\omega_k) < 0\} \leq \epsilon_\star,$$

  where the second equality is due to the fact that $\phi(\omega)$ is a nonconstant affine function and $P_{\omega_k|\zeta^{k-1}}$, along with all distributions from $\mathcal{P}_1$, has density, and the inequality is given by the origin of $\epsilon_\star$ which upper-bounds the risk of the single-observation test underlying $\mathcal{T}_K^{\rm maj}$.

What we want to prove is that under the circumstances we have just summarized, we have

$$P_{\omega^K}\{\omega^K = (\omega_1, ..., \omega_K) : \mathrm{Card}\{k \leq K : \chi_k(\omega_k) = 1\} \geq M\} \leq \epsilon_M = \sum_{M \leq k \leq K} \binom{K}{k} \epsilon_\star^k (1-\epsilon_\star)^{K-k}, \tag{2.2.36}$$

where P_{ω^K} is the distribution of $\omega^K = \{\omega_k = \theta_k(\zeta^{k-1})\}_{k=1}^K$ induced by the distribution of hidden factors. There is nothing to prove when $\epsilon_\star = 1$, since in this case $\epsilon_M = 1$. Thus, we assume from now on that $\epsilon_\star < 1$.

Achieving the goal, step 1. Our reasoning, inspired by the one we used to justify Remark 2.2.1, is as follows. Consider a sequence of random variables η_k, $1 \leq k \leq K$, uniformly distributed on $[0,1]$ and independent of each other and of ζ^K, and consider new driving factors $\lambda_k = [\zeta_k; \eta_k]$ and new observations

$$\mu_k = [\omega_k = \theta_k(\zeta^k); \eta_k] = \Theta_k(\lambda^k = (\lambda_1, ..., \lambda_k)) \quad {}^{6} \tag{2.2.37}$$

driven by these new driving factors, and let

$$\psi_k(\mu_k = [\omega_k; \eta_k]) = \chi_k(\omega_k).$$

It is immediately seen that

- $\mu_k = [\omega_k = \theta_k(\zeta^k); \eta_k]$ is a deterministic function, $\Theta_k(\lambda^k)$, of λ^k, and the conditional, $\lambda^{k-1} = [\zeta^{k-1}; \eta^{k-1}]$ given, distribution $P_{\mu_k|\lambda^{k-1}}$ of μ_k is the product distribution $P_{\omega_k|\zeta^{k-1}} \times U$ on $\Omega \times [0,1]$, where U is the uniform distribution on $[0,1]$. In particular,

  $$\pi_k(\lambda^{k-1}) := P_{\mu_k|\lambda^{k-1}}\{\mu_k = [\omega_k; \eta_k] : \chi_k(\omega_k) = 1\} = P_{\omega_k|\zeta^{k-1}}\{\omega_k : \chi_k(\omega_k) = 1\} \leq \epsilon_\star. \tag{2.2.38}$$

[5] in fact, we need to write $\phi(\omega) < 0$ instead of $\phi(\omega) \leq 0$; we replace the strict inequality with its nonstrict version in order to make our reasoning applicable to the case of $\ell = 2$, where nonstrict inequalities do arise. Clearly, replacing in the definition of χ_k strict inequality with the nonstrict one, we only increase the "rejection domain" of $\mathcal{H}_1$, so that upper bound on the probability of this domain we are about to get automatically is valid for the true rejection domain.

[6] in the formula, same as in what follows, whenever some of the variables $\lambda, \omega, \zeta, \eta, \mu$ appear in the same context, it should always be understood that ζ_t and η_t are components of $\lambda_t = [\zeta_t; \eta_t]$, $\mu_t = [\omega_t; \eta_t] = \Theta_t(\lambda^t)$, and $\omega_t = \theta_t(\zeta^t)$. To remind about these "hidden relations," we sometimes write someting like $\phi(\omega_k = \theta_k(\zeta^k))$ to stress that we are speaking about the value of function ϕ at the point $\omega_k = \theta_k(\zeta^k)$.

- We have

$$\begin{array}{l}P_{\lambda^K}\{\lambda^K:\mathrm{Card}\{k\leq K:\psi_k(\mu_k=\Theta_k(\lambda^k))=1\}\geq M\}\\ \qquad\qquad =P_{\omega^K}\{\omega^K=(\omega_1,...,\omega_K):\mathrm{Card}\{k\leq K:\chi_k(\omega_k)=1\}\geq M\},\end{array}\tag{2.2.39}$$

  where P_{ω^K} is as in (2.2.36), and $\Theta_k(\cdot)$ is defined in (2.2.37).

Now let us define $\psi_k^+(\lambda^k)$ as follows:

- when $\psi_k(\Theta_k(\lambda^k))=1$, or, which is the same, $\chi_k(\omega_k=\theta_k(\zeta^k)))=1$, we set $\psi_k^+(\lambda^k)=1$ as well;
- when $\psi_k(\Theta_k(\lambda^k))=0$, or, which is the same, $\chi_k(\omega_k=\theta_k(\zeta^k))=0$, we set $\psi_k^+(\lambda^k)=1$ whenever

$$\eta_k\leq\gamma_k(\lambda^{k-1}):=\frac{\epsilon_\star-\pi_k(\lambda^{k-1})}{1-\pi_k(\lambda^{k-1})}$$

  and $\psi_k^+(\lambda^k)=0$ otherwise.

Let us make the following immediate observations:

(A) Whenever λ^k is such that $\psi_k(\mu_k=\Theta_k(\lambda^k))=1$, we have also $\psi_k^+(\lambda^k)=1$;

(B) The conditional, $\lambda^{k-1}=[\zeta^{k-1};\eta^{k-1}]$ being fixed, probability of the event $\psi_k^+(\lambda^k)=1$ is exactly $\epsilon_\star$.
Indeed, let $P_{\lambda_k|\lambda^{k-1}}$ be the conditional, λ^{k-1} being fixed, distribution of λ_k. Let us fix λ^{k-1}. The event $E=\{\lambda_k:\psi_k^+(\lambda^k)=1\}$, by construction, is the union of two nonoverlapping events:

$$E_1=\{\lambda_k=[\zeta_k;\eta_k]:\chi_k(\theta_k(\zeta^k))=1\};\ E_2=\{\lambda_k=[\zeta_k;\eta_k]:\chi_k(\theta_k(\zeta^k))=0,\eta_k\leq\gamma_k(\lambda^k)\}.$$

Taking into account that the conditional, λ^{k-1} fixed, distribution of $\mu_k=[\omega_k=\theta_k(\zeta^k);\eta_k]$ is the product distribution $P_{\omega_k|\zeta^{k-1}}\times U$, we conclude in view of (2.2.38) that

$$\begin{array}{l}P_{\lambda_k|\lambda^{k-1}}\{E_1\}=P_{\omega_k|\zeta^{k-1}}\{\omega_k:\chi_k(\omega_k)=1\}=\pi_k(\lambda^{k-1}),\\ P_{\lambda_k|\lambda^{k-1}}\{E_2\}=P_{\omega_k|\zeta^{k-1}}\{\omega_k:\chi_k(\omega_k)=0\}U\{\eta\leq\gamma_k(\lambda^{k-1})\}=(1-\pi_k(\lambda^{k-1}))\gamma_k(\lambda^{k-1}),\end{array}$$

which combines with the definition of $\gamma_k(\cdot)$ to imply (B).

Achieving the goal, step 2. By (A) combined with (2.2.39) we have

$$\begin{array}{l}P_{\omega^K}\{\omega^K:\mathrm{Card}\{k\leq K:\chi_k(\omega_k)=1\}\geq M\}\\ \quad =P_{\lambda^K}\{\lambda^K:\mathrm{Card}\{k\leq K:\psi_k(\mu_k=\Theta_k(\lambda^k))=1\}\geq M\}\\ \quad \leq P_{\lambda^K}\{\lambda^K:\mathrm{Card}\{k\leq K:\psi_k^+(\lambda^k)=1\}\geq M\},\end{array}$$

and all we need to verify is that the first quantity in this chain is upper-bounded by the quantity ϵ_M given by (2.2.36). Invoking the chain and (B), it is enough to justify the following claim:

> (!) Let $\lambda^K=(\lambda_1,...,\lambda_K)$ be a random sequence with probability distribution P, let $\psi_k(\lambda^k)$ take values 0 and 1 only, and let for every $k\leq K$ the conditional, λ^{k-1} being fixed, probability for $\psi_k^+(\lambda^k)$ to take value 1 is, for all λ^{k-1}, equal to $\epsilon_\star$. Then the P-probability of the event
>
> $$\{\lambda^K:\mathrm{Card}\{k\leq K:\psi_k^+(\lambda_k)=1\}\geq M\}$$
>
> is exactly equal to ϵ_M given by (2.2.36).

This is immediate. For integers k, m, $1 \leq k \leq K$, $m \geq 0$, Let $\chi^k_m(\lambda^k)$ be the characteristic function of the event

$$\{\lambda^k : \mathrm{Card}\{t \leq k : \psi^+_t(\lambda^t) = 1\} = m\},$$

and let

$$\pi^k_m = P\{\lambda^K : \chi^k_m(\lambda^k) = 1\}.$$

We have the following evident recurrence:

$$\chi^k_m(\lambda^k) = \chi^{k-1}_m(\lambda^{k-1})(1 - \psi^+_k(\lambda^k)) + \chi^{k-1}_{m-1}(\lambda^{k-1})\psi^+_k(\lambda^k),\ k = 1, 2, ...$$

augmented by the "boundary conditions" $\chi^0_m = 0$, $m > 0$, $\chi^0_0 = 1$, $\chi^{k-1}_{-1} = 0$ for all $k \geq 1$. Taking expectation w.r.t. P and utilizing the fact that conditional, λ^{k-1} being given, expectation of $\psi^+_k(\lambda^k)$ is, identically in λ^{k-1}, equal to $\epsilon_\star$, we get

$$\pi^k_m = \pi^{k-1}_m(1 - \epsilon_\star) + \pi^{k-1}_{m-1}\epsilon_\star,\ k = 1, ..., K,\ \pi^0_m = \left\{\begin{array}{ll} 1, & m = 0 \\ 0, & m > 0 \end{array}\right.,\ \pi^{k-1}_{-1} = 0,\ k = 1, 2, ...$$

whence

$$\pi^k_m = \left\{\begin{array}{ll} \binom{k}{m}\epsilon^m_\star(1 - \epsilon_\star)^{k-m}, & m \leq k \\ 0, & m > k \end{array}\right.$$

and therefore

$$P\{\lambda^K : \mathrm{Card}\{k \leq K : \psi^+_k(\lambda^k) = 1\} \geq M\} = \sum_{M \leq k \leq K} \pi^K_k = \epsilon_M,$$

as required. □

2.3 Detectors and Detector-Based Tests

2.3.1 Detectors and their risks

Let Ω be an observation space, and $\mathcal{P}_\chi$, $\chi = 1, 2$, be two families of probability distributions on Ω. By definition a *detector* associated with Ω is a real-valued function $\phi(\omega)$ of Ω. We associate with a detector ϕ and families $\mathcal{P}_\chi$, $\chi = 1, 2$, *risks* defined as follows:

$$\begin{array}{rcll} \mathrm{Risk}_-[\phi|\mathcal{P}_1] & = & \sup_{P \in \mathcal{P}_1} \int_\Omega \exp\{-\phi(\omega)\}P(d\omega) & (a) \\ \mathrm{Risk}_+[\phi|\mathcal{P}_2] & = & \sup_{P \in \mathcal{P}_2} \int_\Omega \exp\{\phi(\omega)\}P(d\omega) & (b) \\ \mathrm{Risk}[\phi|\mathcal{P}_1, \mathcal{P}_2] & = & \max[\mathrm{Risk}_-[\phi|\mathcal{P}_1], \mathrm{Risk}_+[\phi|\mathcal{P}_2]] & (c) \end{array} \tag{2.3.1}$$

Given a detector ϕ, we can associate with it simple test $\mathcal{T}_\phi$ deciding, via observation $\omega \sim P$, on the hypotheses

$$H_1 : P \in \mathcal{P}_1,\ H_2 : P \in \mathcal{P}_2; \tag{2.3.2}$$

specifically, given observation $\omega \in \Omega$, the test $\mathcal{T}_\phi$ accepts H_1 and rejects H_2 whenever $\phi(\omega) \geq 0$, otherwise the test accepts H_2 and rejects H_1.

Let us make the following immediate observation:

Proposition 2.3.1 *Let Ω be an observation space, $\mathcal{P}_\chi$, $\chi = 1, 2$, be two families of probability distributions on Ω, and ϕ be a detector. The risks of the test $\mathcal{T}_\phi$ associated with this detector satisfy*

$$\begin{array}{rcl} \mathrm{Risk}_1(\mathcal{T}_\phi|H_1, H_2) & \leq & \mathrm{Risk}_-[\phi|\mathcal{P}_1]; \\ \mathrm{Risk}_2(\mathcal{T}_\phi|H_1, H_2) & \leq & \mathrm{Risk}_+[\phi|\mathcal{P}_2]. \end{array} \tag{2.3.3}$$

Proof. Let $\omega \sim P \in \mathcal{P}_1$. Then the P-probability of the event $\{\omega : \phi(\omega) < 0\}$ does not exceed $\mathrm{Risk}_-[\phi|\mathcal{P}_1]$, since on the set $\{\omega : \phi(\omega) < 0\}$ the integrand in (2.3.1.a) is > 1, and this integrand is nonnegative everywhere, so that the integral in (2.3.1.a) is $\geq P\{\omega : \phi(\omega) < 0\}$. Recalling what $\mathcal{T}_\phi$ is, we see that the P-probability to reject H_1 is at most $\mathrm{Risk}_-[\phi|\mathcal{P}_1]$, implying the first relation in (2.3.3). By similar argument, with (2.3.1.b) in the role of (2.3.1.a), when $\omega \sim P \in \mathcal{P}_2$, the P-probability of the event $\{\omega : \phi(\omega) \geq 0\}$ is upper-bounded by $\mathrm{Risk}_+[\phi|\mathcal{P}_2]$, implying the second relation in (2.3.3). □

2.3.2 Detector-based tests

Our current goal is to establish some basic properties of detector-based tests.

2.3.2.1 Structural properties of risks

Observe that the fact that ϵ_1 and ϵ_2 are upper bounds on the risks of a detector are expressed by system of *convex* constraints

$$\begin{array}{ll} \sup_{P\in\mathcal{P}_1} \int_\Omega \exp\{-\phi(\omega)\}P(d\omega) \leq \epsilon_1 & (a) \\ \sup_{P\in\mathcal{P}_2} \int_\Omega \exp\{\phi(\omega)\}P(d\omega) \leq \epsilon_2 & (b) \end{array} \tag{2.3.4}$$

on ϵ_1, ϵ_2 and $\phi(\cdot)$; this observation is useful, but not too useful, since the convex constraints in question usually are infinite-dimensional when $\phi(\cdot)$ is so, and are semi-infinite (suprema, over parameter ranging in infinite set, of parametric families of convex constraints), provided $\mathcal{P}_1$ or $\mathcal{P}_2$ are of infinite cardinalities; constraints of this type can be intractable computationally.

Another important observation is that the distributions P enter the constraints linearly; as a result, *when passing from families of probability distributions $\mathcal{P}_1$, $\mathcal{P}_2$ to their convex hulls, the risks of a detector remain intact.*

2.3.2.2 Renormalization

Let Ω, $\mathcal{P}_1$, $\mathcal{P}_2$ be the same as in Section 2.3.1, and let ϕ be a detector. When shifting this detector by a real a – passing from ϕ to the detector

$$\phi_a(\omega) = \phi(\omega) - a$$

– the risks clearly are updated as follows:

$$\begin{array}{rcl} \mathrm{Risk}_-[\phi_a|\mathcal{P}_1] & = & \mathrm{e}^a \mathrm{Risk}_-[\phi|\mathcal{P}_1], \\ \mathrm{Risk}_+[\phi_a|\mathcal{P}_2] & = & \mathrm{e}^{-a} \mathrm{Risk}_+[\phi|\mathcal{P}_2]. \end{array} \tag{2.3.5}$$

We see that

> *When speaking about risks of a detector, what matters is the product*
>
> $$\mathrm{Risk}_\odot[\phi|\mathcal{P}_1,\mathcal{P}_2] := \mathrm{Risk}_-[\phi|\mathcal{P}_1]\mathrm{Risk}_+[\phi|\mathcal{P}_2]$$
>
> *of the risks, not these risks individually: by shifting the detector, we can redistribute this product between the factors in any way we want. In particular, we can always shift a detector to make it balanced, i.e., satisfying*
>
> $$\mathrm{Risk}_-[\phi|\mathcal{P}_1] = \mathrm{Risk}_+[\phi|\mathcal{P}_2] = \mathrm{Risk}[\phi|\mathcal{P}_1,\mathcal{P}_2].$$
>
> *When deciding on the hypotheses*
>
> $$H_1 : P \in \mathcal{P}_1,\ H_2 : P \in \mathcal{P}_2$$

on the distribution P of observation, the risk of the test $\mathcal{T}_\phi$ associated with a balanced detector ϕ is bounded by the risk $\mathrm{Risk}[\phi|\mathcal{P}_1,\mathcal{P}_2]$ *of the detector:*

$$\mathrm{Risk}(\mathcal{T}_\phi|H_1,H_2):=\max\left[\mathrm{Risk}_1(\mathcal{T}_\phi|H_1,H_2),\mathrm{Risk}_2(\mathcal{T}_\phi|H_1,H_2)\right]\leq\mathrm{Risk}[\phi|\mathcal{P}_1,\mathcal{P}_2].$$

2.3.2.3 Detector-based testing from repeated observations

We are about to show that detector-based tests are perfectly well suited for passing from inferences based on *single* observation to those based on *repeated* observations.

Given K observation spaces Ω_k, $1\leq k\leq K$, each equipped with pair $\mathcal{P}_{k,1}$, $\mathcal{P}_{k,2}$ of families of probability distributions, we can build a new observation space

$$\Omega^K=\Omega_1\times\ldots\times\Omega_K=\{\omega^K=(\omega_1,\ldots,\omega_K):\omega_k\in\Omega_k,k\leq K\}$$

and equip it with two families $\mathcal{P}^K_\chi$, $\chi=1,2$, of probability distributions; distributions from $\mathcal{P}^K_\chi$ are exactly the product-type distributions $P=P_1\times\ldots\times P_K$ with all factors P_k taken from $\mathcal{P}_{k,\chi}$. Observations $\omega^K=(\omega_1,\ldots,\omega_K)$ from Ω^K drawn from a distribution $P=P_1\times\ldots.\times P_K\in\mathcal{P}^K_\chi$ are nothing but collections of observations ω_k, $k=1,\ldots,K$, drawn, independently of each other, from distributions P_k. Now, given detectors $\phi_k(\cdot)$ on observation spaces Ω_k and setting

$$\phi^{(K)}(\omega^K)=\sum_{k=1}^K\phi_k(\omega_k):\Omega^K\to\mathbf{R},$$

we clearly have

$$\begin{array}{rcl}\mathrm{Risk}_-[\phi^{(K)}|\mathcal{P}^K_1]&=&\prod\limits_{k=1}^K\mathrm{Risk}_-[\phi_k|\mathcal{P}_{k,1}],\\ \mathrm{Risk}_+[\phi^{(K)}|\mathcal{P}^K_2]&=&\prod\limits_{k=1}^K\mathrm{Risk}_+[\phi_k|\mathcal{P}_{k,2}].\end{array}\tag{2.3.6}$$

Let us look at some useful consequences of (2.3.6).

Stationary K-repeated observations. Consider the case of Section 2.1.3.1: we are given an observation space Ω and a positive integer K, and what we observe, is a sample $\omega^K=(\omega_1,\ldots,\omega_K)$ with $\omega_1,\ldots,\omega_K$ drawn, independently of each other, from some distribution P on Ω. Let now $\mathcal{P}_1$, $\mathcal{P}_2$, be two families of probability distributions on Ω; we can associate with these families two hypotheses, $H_1^{\odot,K}$, $H_2^{\odot,K}$, on the distribution of K-repeated observation $\omega^K=(\omega_1,\ldots,\omega_K)$, with $H_\chi^{\odot,K}$ stating that $\omega_1,\ldots,\omega_K$ are drawn, independently of each other, from a distribution $P\in\mathcal{P}_\chi$. Given a detector ϕ on Ω, we can associate with it the detector

$$\phi^{(K)}(\omega^K)=\sum_{k=1}^K\phi(\omega_k)$$

on

$$\Omega^K:\underbrace{\Omega\times\ldots\times\Omega}_{K}.$$

Combining (2.3.6) and Proposition 2.3.1, we arrive at the following nice result:

Proposition 2.3.2 *The risks of the simple test $\mathcal{T}_{\phi^{(K)}}$ deciding, given K-repeated observation $\omega^K=(\omega_1,\ldots,\omega_K)$ on the hypotheses*

$$\begin{array}{l}H_1^{\odot,K}:\omega_k,\ k\leq K,\ \textit{are independently of each other drawn from a distribution}\ P\in\mathcal{P}_1\\ H_2^{\odot,K}:\omega_k,\ k\leq K,\ \textit{are independently of each other drawn from a distribution}\ P\in\mathcal{P}_2\end{array}$$

according to the rule

$$\phi^{(K)}(\omega^K) := \sum_{k=1}^K \phi(\omega_k) \left\{ \begin{array}{lcl} \geq 0 & \Rightarrow & \textit{accept } H_1^{\odot,K} \\ < 0 & \Rightarrow & \textit{accept } H_2^{\odot,K} \end{array} \right.$$

admit the upper bounds

$$\begin{array}{rcl} \mathrm{Risk}_1(\mathcal{T}_{\phi^{(K)}}|H_1^{\odot,K}, H_2^{\odot,K}) & \leq & (\mathrm{Risk}_-[\phi|\mathcal{P}_1])^K \\ \mathrm{Risk}_2(\mathcal{T}_{\phi^{(K)}}|H_1^{\odot,K}, H_2^{\odot,K}) & \leq & (\mathrm{Risk}_+[\phi|\mathcal{P}_2])^K \end{array} \tag{2.3.7}$$

Semi- and Quasi-Stationary K-repeated observations. Recall that Semi-Stationary and Quasi-Stationary K-repeated observations associated with a family $\mathcal{P}$ of distributions on observation space Ω were defined in Sections 2.1.3.2 and 2.1.3.3, respectively. It turns out that Proposition 2.3.2 extends to quasi-stationary K-repeated observations:

Proposition 2.3.3 *Let Ω be an observation space, $\mathcal{P}_\chi$, $\chi = 1,2$ be families of probability distributions on Ω, $\phi : \Omega \to \mathbf{R}$ be a detector, and K be a positive integer.*

Families $\mathcal{P}_\chi$, $\chi = 1,2$, give rise to two hypotheses on the distribution P^K of quasi-stationary K-repeated observation ω^K:

$$H_\chi^{\otimes,K} : P^K \in \mathcal{P}_\chi^{\otimes,K} = \bigotimes_{k=1}^K \mathcal{P}_\chi,\ \chi = 1,2$$

(see Section 2.1.3.3), and ϕ gives rise to the detector

$$\phi^{(K)}(\omega^K) := \sum_{k=1}^K \phi(\omega_k).$$

The risks of the detector $\phi^{(K)}$ on the families $\mathcal{P}_\chi^{\otimes,K}$, $\chi = 1,2$, can be upper-bounded as follows:

$$\begin{array}{rcl} \mathrm{Risk}_-[\phi^{(K)}|\mathcal{P}_1^{\otimes,K}] & \leq & (\mathrm{Risk}_-[\phi|\mathcal{P}_1])^K\,, \\ \mathrm{Risk}_+[\phi^{(K)}|\mathcal{P}_2^{\otimes,K}] & \leq & (\mathrm{Risk}_-[\phi|\mathcal{P}_2])^K\,. \end{array} \tag{2.3.8}$$

Further, the detector $\phi^{(K)}$ induces simple test $\mathcal{T}_{\phi^{(K)}}$ deciding on $H_\chi^{\otimes,K}$, $\chi = 1,2$ as follows: given ω^K, the test accepts $H_1^{\otimes,K}$ when $\phi^{(K)}(\omega^K) \geq 0$, and accepts $H_2^{\otimes,K}$ otherwise. The risks of this test can be upper-bounded as

$$\begin{array}{rcl} \mathrm{Risk}_1(\mathcal{T}_{\phi^{(K)}}|H_1^{\otimes,K}, H_2^{\otimes,K}) & \leq & (\mathrm{Risk}_-[\phi|\mathcal{P}_1])^K\,, \\ \mathrm{Risk}_2(\mathcal{T}_{\phi^{(K)}}|H_1^{\otimes,K}, H_2^{\otimes,K}) & \leq & (\mathrm{Risk}_+[\phi|\mathcal{P}_2])^K\,. \end{array} \tag{2.3.9}$$

Finally, the above results remain intact when passing from quasi-stationary to semi-stationary K-repeated observations (that is, when replacing $\mathcal{P}_\chi^{\otimes,K}$ with $\mathcal{P}_\chi^{\oplus,K} = \bigoplus_{k=1}^K \mathcal{P}_\chi$ and $H_\chi^{\otimes,K}$ with the hypotheses $H_\chi^{\oplus,K}$ stating that the distribution of ω^K belongs to $\mathcal{P}_\chi^{\oplus,K}$, $\chi = 1,2$).

Proof. All we need is to verify (2.3.8) – in view of Proposition 2.3.1, all other claims in Proposition 2.3.3 are immediate consequences of (2.3.8) and the inclusions $\mathcal{P}_\chi^{\oplus,K} \subset \mathcal{P}_\chi^{\otimes,K}$, $\chi = 1,2$. Verification of (2.3.8) is as follows. Let $P^K \in \mathcal{P}_1^{\otimes,K}$, and let P^K be the distribution of random sequence $\omega^K = (\omega_1,..,.,\omega_K)$ generated as follows: there exists a random sequence of driving factors $\zeta_1,...,\zeta_K$ such that ω_k is a deterministic function of $\zeta^k = (\zeta_1,...,\zeta_k)$:

$$\omega_k = \theta_k(\zeta_1,...,\zeta_k),$$

and the conditional, $\zeta_1, ..., \zeta_{k-1}$ being given, distribution $P_{\omega_k|\zeta^{k-1}}$ belongs to $\mathcal{P}_1$. Let P_{ζ^k} be the distribution of the first k driving factors, and $P_{\zeta_k|\zeta^{k-1}}$ be the conditional, $\zeta_1, ...\zeta_{k-1}$ being given, distribution of ζ_k. Let us set

$$\psi^{(k)}(\zeta_1, ..., \zeta_k) = \sum_{t=1}^{k} \phi(\theta_t(\zeta_1, ..., \zeta_t)),$$

so that

$$\int\limits_{\Omega^K} \exp\{-\phi^{(K)}(\omega^k)\}P^K(d\omega^K) = \int \exp\{-\psi^{(K)}(\zeta^K)\}P_{\zeta^K}(d\zeta^K). \tag{2.3.10}$$

On the other hand, denoting $C_0 = 1$, we have

$$\begin{array}{l}
C_k := \int \exp\{-\psi^{(k)}(\zeta^k)\}P_{\zeta^k}(d\zeta^k) = \int \exp\{-\psi^{(k-1)}(\zeta^{k-1}) - \phi(\theta_k(\zeta^k))\}P_{\zeta^k}(d\zeta^k) \\
= \int \exp\{-\psi^{(k-1)}(\zeta^{k-1})\} \underbrace{\left[\int \exp\{-\phi(\theta_k(\zeta^k))\}P_{\zeta_k|\zeta^{k-1}}(d\zeta_k)\right]}_{= \int\limits_{\Omega} \exp\{-\phi(\omega_k)\}P_{\omega_k|\zeta^{k-1}}(d\omega_k)} P_{\zeta^{k-1}}(d\zeta^{-1}) \\
\underbrace{\leq}_{(*)} \text{Risk}_-[\phi|\mathcal{P}_1] \int \exp\{-\psi^{(k-1)}(\zeta^{k-1})\}P_{\zeta^{k-1}}(d\zeta^{k-1}) = \text{Risk}_-[\phi|\mathcal{P}_1]C_{k-1},
\end{array}$$

where $(*)$ is due to the fact that the distribution $P_{\omega_k|\zeta^{k-1}}$ belongs to $\mathcal{P}_1$. From the resulting recurrence we get

$$C_K \leq (\text{Risk}_-[\phi|\mathcal{P}_1])^K,$$

which combines with (2.3.10) to imply that

$$\int\limits_{\Omega^K} \exp\{-\phi^{(K)}(\omega^k)\}P^K(d\omega^K) \leq (\text{Risk}_-[\phi|\mathcal{P}_1])^K.$$

The latter inequality holds true for every distribution $P^K \in \mathcal{P}_\chi^{\otimes,K}$, and the first inequality in (2.3.8) follows. The second inequality in (2.3.8) is given by completely similar reasoning, with $\mathcal{P}_2$ in the role of $\mathcal{P}_1$, and $-\phi$, $-\phi^{(K)}$ in the roles of ϕ, $\phi^{(K)}$, respectively. □

The fact that observations ω_k under hypotheses $H_\ell^{\otimes,K}$, $\ell = 1, 2$ are related to "constant in time" families $\mathcal{P}_\ell$ has no importance here, and in fact the proof of Proposition 2.3.3 after absolutely evident modifications of wording allows to justify the following "non-stationary" version of Proposition:

Proposition 2.3.4 *For $k = 1, ..., K$, let Ω_k be observation spaces, $\mathcal{P}_{\chi,k}$, $\chi = 1, 2$ be families of probability distributions on Ω_k, and $\phi_k : \Omega_k \to \mathbf{R}$ be detectors.*

Families $\mathcal{P}_{\chi,k}$, $\chi = 1, 2$, give rise to quasi-direct products (see Section 2.1.3.3) $\mathcal{P}_\chi^{\otimes,K} = \bigotimes_{k=1}^{K} \mathcal{P}_{\chi,k}$ of the families $\mathcal{P}_{\chi,k}$ over $1 \leq k \leq K$, and thus to two hypotheses on the distribution P^K of observation $\omega^K = (\omega_1, ..., \omega_K) \in \Omega^K = \Omega_1 \times ... \times \Omega_K$:

$$H_\chi^{\otimes,K} : P^K \in \mathcal{P}_\chi^{\otimes,K}, \ \chi = 1, 2,$$

and detectors ϕ_k, $1 \leq k \leq K$, give rise to the detector

$$\phi^K(\omega^K) := \sum_{k=1}^{K} \phi_k(\omega_k).$$

The risks of the detector ϕ^K on the families $\mathcal{P}_\chi^{\otimes,K}$, $\chi = 1, 2$, can be upper-bounded as follows:

$$\begin{array}{rcl}
\text{Risk}_-[\phi^K|\mathcal{P}_1^{\otimes,K}] & \leq & \prod_{k=1}^{K} \text{Risk}_-[\phi|\mathcal{P}_{1,k}], \\
\text{Risk}_+[\phi^K|\mathcal{P}_2^{\otimes,K}] & \leq & \prod_{k=1}^{K} \text{Risk}_+[\phi|\mathcal{P}_{2,K}].
\end{array} \tag{2.3.11}$$

Further, the detector ϕ^K induces simple test $\mathcal{T}_{\phi(K)}$ deciding on $H_\chi^{\otimes,K}$, $\chi = 1,2$ as follows: given ω^K, the test accepts $H_1^{\otimes,K}$ when $\phi^K(\omega^K) \geq 0$, and accepts $H_2^{\otimes,K}$ otherwise. The risks of this test can be upper-bounded as

$$\begin{array}{rcl}\mathrm{Risk}_1(\mathcal{T}_{\phi^K}|H_1^{\otimes,K}, H_2^{\otimes,K}) & \leq & \prod_{k=1}^K \mathrm{Risk}_-[\phi|\mathcal{P}_{1,k}],\\ \mathrm{Risk}_2(\mathcal{T}_{\phi(K)}|H_1^{\otimes,K}, H_2^{\otimes,K}) & \leq & \prod_{k=1}^K \mathrm{Risk}_+[\phi|\mathcal{P}_{2,k}].\end{array} \tag{2.3.12}$$

Finally, the above results remain intact when passing from quasi-direct products to direct products of the families of distributions in question (that is, when replacing $\mathcal{P}_\chi^{\otimes,K}$ with $\mathcal{P}_\chi^{\oplus,K} = \bigoplus\limits_{k=1}^{K} \mathcal{P}_{\chi,k}$ and $H_\chi^{\otimes,K}$ with the hypotheses $H_\chi^{\oplus,K}$ stating that the distribution of ω^K belongs to $\mathcal{P}_\chi^{\oplus,K}$, $\chi = 1,2$).

2.3.2.4 Limits of performance of detector-based tests

We are about to demonstrate that as far as limits of performance of pairwise simple detector-based tests are concerned, these tests are nearly as good as simple tests can be.

Proposition 2.3.5 *Let Ω be an observation space, and $\mathcal{P}_\chi$, $\chi = 1,2$, be families of probability distributions on Ω. Assume that for some $\epsilon \in (0,1/2)$ "in the nature" there exists a simple test (deterministic or randomized) deciding on the hypotheses*

$$H_1 : P \in \mathcal{P}_1,\;\; H_2 : P \in \mathcal{P}_2$$

on the distribution P of observation ω with risks $\leq \epsilon$:

$$\mathrm{Risk}_1(\mathcal{T}|H_1, H_2) \leq \epsilon \;\&\; \mathrm{Risk}_2(\mathcal{T}|H_1, H_2) \leq \epsilon.$$

Then there exists a detector-based test $\mathcal{T}_\phi$ deciding on the same pair of hypotheses with risk "comparable" with ϵ:

$$\mathrm{Risk}_1(\mathcal{T}_\phi|H_1, H_2) \leq \epsilon^+ \;\&\; \mathrm{Risk}_2(\mathcal{T}_\phi|H_1, H_2) \leq \epsilon^+,\;\; \epsilon^+ = 2\sqrt{\epsilon(1-\epsilon)}. \tag{2.3.13}$$

Proof. Let us prove the claim in the case when the test $\mathcal{T}$ is deterministic; the case when this test is randomized is the subject of Exercise 2.9.

Let Ω_χ, $\chi = 1,2$, be the sets of $\omega \in \Omega$ such that $\mathcal{T}$ as "feeded" by observation ω accepts H_χ. Since $\mathcal{T}$ is simple, Ω_1, Ω_2 split Ω into two non-overlapping parts, and since the risks of $\mathcal{T}$ are $\leq \epsilon$, we have

$$\begin{array}{ll}(a) & \epsilon_2(P) := P\{\Omega_2\} \leq \epsilon \,\forall P \in \mathcal{P}_1\\ (a) & \epsilon_1(P) := P\{\Omega_1\} \leq \epsilon \,\forall P \in \mathcal{P}_2\end{array}$$

Let $\delta = \sqrt{(1-\epsilon)/\epsilon}$, so that $\delta \geq 1$ due to $0 < \epsilon \leq 1/2$, and let

$$\psi(\omega) = \left\{\begin{array}{ll}\delta, & \omega \in \Omega_1\\ 1/\delta, & \omega \in \Omega_2\end{array}\right.,\;\; \phi(\omega) = \ln(\psi(\omega)).$$

When $P \in \mathcal{P}_1$, we have

$$\int_\Omega \exp\{-\phi(\omega)\} P(d\omega) = \frac{1}{\delta} P\{\Omega_1\} + \delta P\{\Omega_2\} = \frac{1}{\delta} + \underbrace{\left[\delta - \frac{1}{\delta}\right]}_{\geq 0} \epsilon_2(P) \leq \frac{1}{\delta} + \left[\delta - \frac{1}{\delta}\right]\epsilon = \epsilon^+,$$

whence $\mathrm{Risk}_-[\phi|\mathcal{P}_1] \leq \epsilon^+$. Similarly, when $P \in \mathcal{P}_2$, we have

$$\int_\Omega \exp\{\phi(\omega)\} P(d\omega) = \delta P\{\Omega_1\} + \frac{1}{\delta} P\{\Omega_2\} = \underbrace{\left[\delta - \frac{1}{\delta}\right]}_{\geq 0} \epsilon_1(P) + \frac{1}{\delta} \leq \left[\delta - \frac{1}{\delta}\right]\epsilon + \frac{1}{\delta} = \epsilon^+,$$

whence $\mathrm{Risk}_+[\phi|\mathcal{P}_2] \leq \epsilon^+$. □

Discussion. Proposition 2.3.5 states that we can restrict ourselves with detector-based tests at the price of passing from risk ϵ exhibited by "the best test existing in the nature" to "comparable" risk $\epsilon^+ = 2\sqrt{\epsilon(1-\epsilon)}$. What we buy when sticking to detector-based tests are nice properties listed in Sections 2.3.2.1 – 2.3.2.3 and possibility to compute *under favorable circumstances*, see below, the best, in terms of their risk, among the detector-based tests; optimizing risk of a detector-based test turns out to be an essentially more realistic task than optimizing risk of a general-type test. This being said, one can argue that treating ϵ and ϵ^+ "comparable" is a too optimistic attitude; for example, risk level $\epsilon = 0.01$ seems to be much more attractive than $[0.01]^+ \approx 0.2$. While passing from a test $\mathcal{T}$ with risk 0.01 to a detector-based test $\mathcal{T}_\phi$ with risk 0.2 could indeed be a "heavy toll," there is some comfort in the fact that passing from a single observation to three of them (i.e., to 3-repeated, stationary or non-stationary alike, version of the original observation scheme), we can straightforwardly convert $\mathcal{T}_\phi$ into a test with risk $(0.2)^3 = 0.008 < 0.01$, and passing to 6 observations, to make the risk less than 0.0001. On the other hand, seemingly the only way to convert a general-type single-observation test $\mathcal{T}$ with risk 0.01 into a multi-observation test with essentially smaller risk is to pass to a Majority version of $\mathcal{T}$, see Section 2.2.3.3 [7]. Computation shows that with $\epsilon_\star = 0.01$, to make the risk of the majority test ≤ 0.0001 takes 5 observations, which is only marginally better than the 6 observations needed in the detector-based construction.

2.4 Simple observation schemes

2.4.1 Simple observation schemes – Motivation

A natural conclusion one can extract from the previous Section is that it makes sense, to say the least, to learn how to build detector-based tests with minimal risk. Thus, we arrive at the following design problem:

> Given an observation space Ω and two families, $\mathcal{P}_1$ and $\mathcal{P}_2$, of probability distributions on Ω, solve the optimization problem
>
> $$\mathrm{Opt} = \min_{\phi:\Omega\to\mathbf{R}} \max\Big[\underbrace{\sup_{P\in\mathcal{P}_1}\int_\Omega \mathrm{e}^{-\phi(\omega)}P(d\omega)}_{F[\phi]}, \underbrace{\sup_{P\in\mathcal{P}_2}\int_\Omega \mathrm{e}^{\phi(\omega)}P(d\omega)}_{G[\phi]}\Big] \qquad (2.4.1)$$

While being convex, problem (2.4.1) typically is computationally intractable. First, it is infinite-dimensional – candidate solutions are multivariate functions; how to represent them in a computer, not speaking of how to optimize over them? Besides, the objective to be optimized is expressed in terms of suprema of infinitely many (provided $\mathcal{P}_1$ and/or $\mathcal{P}_2$ are infinite) expectations, and computing just a single expectation can be a difficult task... We are about to consider "favorable" cases – *simple observation schemes* – where (2.4.1) is efficiently solvable.

To arrive at the notion of a simple observation scheme, consider the case when all distributions from $\mathcal{P}_1$, $\mathcal{P}_2$ admit densities taken w.r.t. some reference measure Π on Ω, and these densities are parameterized by "parameter" μ running through some parameter space $\mathcal{M}$, so that $\mathcal{P}_1$ is comprised of all distributions with densities $p_\mu(\cdot)$ and μ belonging to some subset M_1 of $\mathcal{M}$, while $\mathcal{P}_2$ is comprised of distributions with densities $p_\mu(\cdot)$ and μ belonging to another subset, M_2, of $\mathcal{M}$. To save words, we shall identify distributions with their densities taken w.r.t. Π, so that

$$\mathcal{P}_\chi = \{p_\mu : \mu \in M_\chi\},\ \chi = 1, 2,$$

[7] In Section 2.2.3.3, we dealt with "signal plus noise" observations and with specific test $\mathcal{T}$ given by Euclidean separation. Straightforward inspection of the construction and the proof of Proposition 2.2.3 makes it clear that the construction is applicable to a whatever simple test $\mathcal{T}$, and that the risk of the resulting multi-observation test obeys the upper bound in (2.2.19), with the risk of $\mathcal{T}$ in the role of $\epsilon_\star$.

where $\{p_\mu(\cdot) : \mu \in \mathcal{M}\}$ is a given "parametric" family of probability densities. Quotation marks in "parametric" reflect the fact that at this point in time, the "parameter" μ can be infinite-dimensional (e.g, we can parameterise a density by itself), so that assuming "parametric" representation of the distributions from $\mathcal{P}_1$, $\mathcal{P}_2$ in fact does not restrict generality.

Our first observation is that in our "parametric" setup, we can rewrite problem (2.4.1) equivalently as

$$\ln(\mathrm{Opt}) = \min_{\phi:\Omega\to\mathbf{R}} \sup_{\mu\in M_1,\nu\in M_2} \underbrace{\frac{1}{2}\left[\ln\left(\int_\Omega \mathrm{e}^{-\phi(\omega)}p_\mu(\omega)\Pi(d\omega)\right) + \ln\left(\int_\Omega \mathrm{e}^{\phi(\omega)}p_\nu(\omega)\Pi(d\omega)\right)\right]}_{\Phi(\phi;\mu,\nu)}. \qquad (2.4.2)$$

> Indeed, when shifting ϕ by a constant: $\phi(\cdot) \mapsto \phi(\cdot) - a$, the positive quantities $F[\phi]$ and $G[\phi]$ participating in (2.4.1) are multiplied by e^a and e^{-a}, respectively, and their product remains intact. It follows that to minimize over ϕ the maximum of $F[\phi]$ and $G[\phi]$ (this is what (2.4.1) wants of us) is exactly the same as to minimize over ϕ the quantity $H[\phi] := \sqrt{F[\phi]G[\phi]}$. Indeed, a candidate solution ϕ to the problem $\min_\phi H[\phi]$ can be *balanced* – shifted by a constant to ensure $F[\phi] = G[\phi]$, and this balancing does not change $H[\cdot]$; as a result, minimizing H over all ϕ is the same as minimizing H over balanced ϕ, and the latter problem clearly is equivalent to (2.4.1). It remains to note that (2.4.2) is nothing but the problem of minimizing $\ln(H[\phi])$.

Now, (2.4.2) is a min-max problem – a problem of the generic form

$$\min_{u\in U}\max_{v\in V}\Psi(u,v).$$

Problems of this type (at least, finite-dimensional ones) are computationally tractable when the domain of the minimization argument is convex and the cost function Ψ is convex in the minimization argument (this indeed is the case for (2.4.2)), and the domain of the maximization argument is convex, and the cost function is concave in this argument (this not necessarily is the case for (2.4.2)). *Simple observation schemes* we are about to define are, essentially, the schemes where the just outlined requirements of finite dimensionality and convexity-concavity indeed are met.

2.4.2 Simple observation schemes – Definition

Consider the situation where we are given

1. A Polish (complete separable metric) *observation space* Ω equipped with σ-finite σ-additive Borel reference measure Π such that the support of Π is the entire Ω.
   Those not fully comfortable with some of the notions from the previous sentence can be assured that the only observation spaces we indeed shall deal with are pretty simple:
   - $\Omega = \mathbf{R}^d$ equipped with the Lebesgue measure Π, and
   - a finite or countable set Ω which is discrete (distances between distinct points are equal to 1) and is equipped with the counting measure Π.
2. A parametric family $\{p_\mu(\cdot) : \mu \in \mathcal{M}\}$ of probability densities, taken w.r.t. Π, such that
   - the space $\mathcal{M}$ of parameters is a convex set in some $\mathbf{R}^n$ which coincides with its relative interior,
   - the function $p_\mu(\omega) : \mathcal{M}\times\Omega \to \mathbf{R}$ is continuous in (μ,ω) and positive everywhere.
3. A finite-dimensional linear subspace $\mathcal{F}$ of the space of continuous functions on Ω such that

- $\mathcal{F}$ contains constants,
- all functions of the form $\ln(p_\mu(\omega)/p_\nu(\omega))$ with $\mu, \nu \in \mathcal{M}$ are contained in $\mathcal{F}$;
- for every $\phi(\cdot) \in \mathcal{F}$, the function
$$\ln\left(\int_\Omega \mathrm{e}^{\phi(\omega)} p_\mu(\omega)\Pi(d\omega)\right)$$
is real-valued and *concave* on $\mathcal{M}$.

In this situation we call the collection
$$(\Omega, \Pi; \{p_\mu : \mu \in \mathcal{M}\}; \mathcal{F})$$
a *simple observation scheme* (s.o.s. for short).

Nondegenerate simple o.s. We call a simple observation scheme *nondegenerate*, if the mapping $\mu \mapsto p_\mu$ is an embedding: whenever $\mu, \mu' \in \mathcal{M}$ and $\mu \neq \mu'$, we have $p_\mu \neq p_{\mu'}$.

2.4.3 Simple observation schemes – Examples

We are about to list basic examples of s.o.s.'s.

2.4.3.1 Gaussian observation scheme

In Gaussian o.s.,

- the observation space (Ω, Π) is the space $\mathbf{R}^d$ with Lebesgue measure,
- the family $\{p_\mu(\cdot) : \mu \in \mathcal{M}\}$ is the family of Gaussian densities $\mathcal{N}(\mu, \Theta)$, with fixed positive definite covariance matrix Θ, distributions from the family are parameterized by their expectations μ. Thus,
$$\mathcal{M} = \mathbf{R}^d,\ p_\mu(\omega) = \frac{1}{(2\pi)^{d/2}\sqrt{\mathrm{Det}(\Theta)}} \exp\{-\frac{(\omega-\mu)^T\Theta^{-1}(\omega-\mu)}{2}\};$$
- the family $\mathcal{F}$ is the family of all affine functions on $\mathbf{R}^d$.

It is immediately seen that Gaussian o.s. meets all requirements imposed on a simple o.s. For example,
$$\ln(p_\mu(\omega)/p_\nu(\omega)) = (\nu-\mu)^T\Theta^{-1}\omega + \frac{1}{2}\left[\nu^T\Theta^{-1}\nu - \mu^T\Theta^{-1}\mu\right]$$
is an affine function of ω and thus belongs to $\mathcal{F}$. Besides this, a function $\phi(\cdot) \in \mathcal{F}$ is affine: $\phi(\omega) = a^T\omega + b$, implying that
$$\begin{array}{c} f(\mu) := \ln\left(\int_{\mathbf{R}^d} \mathrm{e}^{\phi(\omega)} p_\mu(\omega) d\omega\right) = \ln\left(\mathbf{E}_{\xi\sim\mathcal{N}(0,I_d)}\left\{\exp\{a^T(\Theta^{1/2}\xi + \mu) + b\}\right\}\right) = a^T\mu + b + \mathrm{const},\\ \mathrm{const} = \ln\left(\mathbf{E}_{\xi\sim\mathcal{N}(0,I_d)}\left\{\exp\{a^T\Theta^{1/2}\xi\}\right\}\right) = \frac{1}{2}a^T\Theta a \end{array}$$
is affine (and thus concave) function of μ.

As we remember from Lecture 1, Gaussian o.s. is responsible for the standard *signal processing* model where one is given a noisy observation
$$\omega = Ax + \xi \qquad [\xi \sim \mathcal{N}(0, \Theta)]$$
of the image Ax of unknown signal $x \in \mathbf{R}^n$ under linear transformation with known $d \times n$ *sensing matrix*, and the goal is to infer from this observation some knowledge about x. In this situation, a hypothesis that x belongs to some set X translates into the hypothesis that the observation ω is drawn from Gaussian distribution with known covariance matrix Θ and expectation known to belong to the set $M = \{\mu = Ax : x \in X\}$, so that deciding on various hypotheses on where x is located reduces to deciding on hypotheses on the distribution of observation in Gaussian o.s.

2.4.3.2 Poisson observation scheme

In Poisson observation scheme,

- the observation space Ω is the set $\mathbf{Z}^d_+$ of d-dimensional vectors with nonnegative integer entries, and this set is equipped with the counting measure,

- the family $\{p_\mu(\cdot) : \mu \in \mathcal{M}\}$ is the family of product-type Poisson distributions with positive parameters. In other words,

$$\mathcal{M} = \{\mu \in \mathbf{R}^d : \mu > 0\}, p_\mu(\omega) = \frac{\mu_1^{\omega_1}\mu_2^{\omega_2}...\mu_d^{\omega_d}}{\omega_1!\omega_2!...\omega_d!}\mathrm{e}^{-\mu_1-\mu_2-...-\mu_d},\ \omega \in \mathbf{Z}^d_+,$$

  that is, random variable $\omega \sim p_\mu$, $\mu \in \mathcal{M}$, is d-dimensional vector with independent random entries, and i-th of the entries is $\omega_i \sim \text{Poisson}(\mu_i)$;

- the space $\mathcal{F}$ is comprised of affine functions on $\mathbf{Z}_d^+$.

It is immediately seen that Poisson o.s. is simple. For example,

$$\ln(p_\mu(\omega)/p_\nu(\omega)) = \sum_{i=1}^d \ln(\mu_i/\nu_i)\omega_i - \sum_{i=1}^d [\mu_i - \nu_i]$$

is affine function of ω and thus belongs to $\mathcal{F}$. Besides this, a function $\phi \in \mathcal{F}$ is affine: $\phi(\omega) = a^T\omega + b$, implying that the function

$$\begin{array}{l} f(\mu) := \ln\left(\int_\Omega \mathrm{e}^{\phi(\omega)}p_\mu(\omega)\Pi(d\omega)\right) = \ln\left(\sum_{\omega\in\mathbf{Z}^d_+} \mathrm{e}^{a^T\omega+b}\prod_{i=1}^d \frac{\mu_i^{\omega_i}\mathrm{e}^{-\mu_i}}{\omega_i!}\right) \\ = b + \ln\left(\prod_{i=1}^d \left[\mathrm{e}^{-\mu_i}\sum_{s=0}^\infty \frac{[\mathrm{e}^{a_i}\mu_i]^s}{s!}\right]\right) = b + \sum_{i=1}^d \ln(\exp\{\mathrm{e}^{a_i}\mu_i - \mu_i\}) = \sum_i [\mathrm{e}^{a_i} - 1]\mu_i + b \end{array}$$

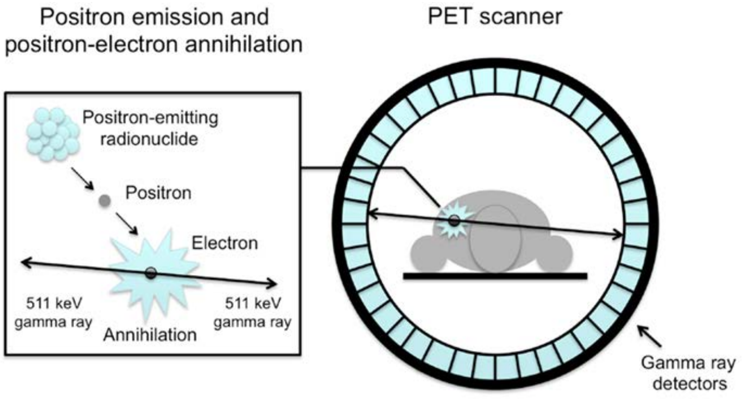
is affine (and thus concave) function of μ.

Poisson observation scheme is responsible for *Poisson Imaging*. This is the situation where there are n "sources of customers;" arrivals of customers at source i are independent of what happens at other sources, and inter-arrival times at source j are independent random variables with exponential, with parameter λ_j, random variables, so that the number of customers arriving at source j in a unit time interval is Poisson random variable with parameter λ_j. Now, there are d "servers", and a customer arrived at source j is dispatched to server i with some given probability A_{ij}, $\sum_i A_{ij} \leq 1$; with probability $1 - \sum_i A_{ij}$, such a customer leaves the system. Needless to say, the dispatches are independent of each other and of the arrival processes. What we observe is the vector $\omega = (\omega_1, ..., \omega_d)$, where ω_i is the number of customers dispatched to server i on the time horizon $[0, 1]$. It is easy to verify that in the just described situation, the entries ω_i in ω indeed are independent of each other Poisson random variables with Poisson parameters

$$\mu_i = \sum_{j=1}^n A_{ij}\lambda_j.$$

In what is called *Poisson Imaging*, one is given a random observation ω of the above type along with *sensing matrix* $A = [A_{ij}]$, and the goal is to use the observation to infer conclusions on the parameter $\mu = A\lambda$ and underlying this parameter "signal" λ.

Poisson imaging is has several important applications[8], for example, in Positron Emission Tomography (PET).



In PET, a patient is injected radioactive tracer and is placed in PET tomograph, which can be thought of as a cylinder with surface split into small detector cells. The tracer disintegrates, and every disintegration act produces a positron which immediately annihilates with a nearby electron, producing two γ-quants flying at the speed of light in two opposite directions along a line ("line of response" – LOR) with completely random orientation. Eventually, each of the γ-quants hits its own detector cell. When two detector cells are "simultaneously" hit (in fact - hit within a short time interval, like 10^{-8} sec), this event – *coincidence* – and the serial number of the *bin* (pair of detectors) where the hits were observed are registered; observing a coincidence in some bin, we know that somewhere on the line linking the detector cells from the bin a disintegration act took place. The data collected in a PET study are the numbers of coincidences registered in every one of the bins; discretizing the field of view (patient's body) into small 3D cubes (voxels), an accurate enough model of the data is a realization ω of random vector with independent Poisson entries $\omega_i \sim \text{Poisson}(\mu_i)$, with μ_i given by

$$\mu_i = \sum_{j=1}^{n} p_{ij}\lambda_j,$$

where λ_j is proportional to the amount of tracer in voxel j, and p_{ij} is the probability for LOR emanating from voxel j to be registered in bin i (these probabilities can be computed given the geometry of PET device). The tracer is selected in such a way that in the body it concentrates in the areas of interest (say, the areas of high metabolic activity when tumor is sought), and the goal of the study is to infer from the observation ω conclusions on the density of the tracer. The characteristic feature of PET as compared to other types of tomography is that with properly selected tracer, this technique allows to visualize metabolic activity, and not only the anatomy of tissues in the body. Now, PET fits perfectly well the above "dispatching customers" story, with disintegration acts taking place in voxel j in the role of customers arriving in location j and bins in the role of servers; the arrival intensities are (proportional to) the amounts λ_j of tracer in voxels, and the random dispatch of customers to servers corresponds to random orientation of LOR's (in reality, the nature draws their directions from the uniform distribution on the unit sphere in 3D).

It is worthy of noting that there are two other real life applications of Poisson Imaging: Large Binocular Telescope and Nanoscale Fluorescent Microscopy [9].

[8] in all these applications, the signal λ we ultimately are interested in is an image, this is where "Imaging" comes from.

[9] Large Binocular Telescope is a cutting edge instrument for high-resolution optical/infrared astronomical imaging; it is the subject of huge ongoing international project, see `http://www.lbto.org`. Nanoscale Fluorescent Microscopy (a.k.a. Poisson Biophotonics) is a revolutionary tool for cell imaging trigged by the advent of techniques [18, 81, 83, 143] (2014 Nobel Prize in Chemistry) allowing to break the diffraction barrier and to view biological molecules "at work" at a resolution 10-20 nm, yielding entirely new insights into the signalling and transport processes within cells.

2.4.3.3 Discrete observation scheme

In Discrete observation scheme,

- the observation space is a finite set $\Omega = \{1, ..., d\}$ equipped with counting measure,
- the family $\{p_\mu(\cdot) : \mu \in \mathcal{M}\}$ is comprised of all non-vanishing distributions on Ω, that is,

$$\mathcal{M} = \{\mu \in \mathbf{R}^d : \mu > 0, \sum_{\omega \in \Omega} \mu_\omega = 1\},\ p_\mu(\omega) = \mu_\omega, \omega \in \Omega;$$

- $\mathcal{F}$ is the space of all real-valued functions on the finite set Ω.

Clearly, Discrete o.s. is simple; for example, the function

$$f(\mu) := \ln\left(\int_\Omega \mathrm{e}^{\phi(\omega)} p_\mu(\omega) \Pi(d\omega)\right) = \ln\left(\sum_{\omega \in \Omega} \mathrm{e}^{\phi(\omega)} \mu_\omega\right)$$

indeed is concave in $\mu \in \mathcal{M}$.

2.4.3.4 Direct products of simple observation schemes

Given K simple observation schemes

$$\mathcal{O}_k = (\Omega_k, \Pi_k; \{p_{\mu,k}(\cdot) : \mu \in \mathcal{M}_k\}; \mathcal{F}_k),\ 1 \leq k \leq K,$$

we can define their *direct product*

$$\mathcal{O}^K = \prod_{k=1}^K \mathcal{O}_k = (\Omega^K, \Pi^K; \{p_\mu : \mu \in \mathcal{M}^K\}; \mathcal{F}^K)$$

by modeling the situation where our observation is a tuple $\omega^K = (\omega_1, ..., \omega_K)$ with components ω_k yielded, independently of each other, by observation schemes $\mathcal{O}_k$, namely, as follows:

- The observation space Ω^K is the direct product of observations spaces $\Omega_1, ..., \Omega_K$, and the reference measure Π^K is the product of the measures $\Pi_1, ..., \Pi_K$;
- The parameter space $\mathcal{M}^K$ is the direct product of partial parameter spaces $\mathcal{M}_1, ..., \mathcal{M}_K$, and the distribution $p_\mu(\omega^K)$ associated with parameter $\mu = (\mu_1, \mu_2, ..., \mu_K) \in \mathcal{M}^K = \mathcal{M}_1 \times ... \times \mathcal{M}_K$ is the probability distribution on Ω^K with the density

  $$p_\mu(\omega^K) = \prod_{k=1}^K p_{\mu,k}(\omega_k)$$

  w.r.t. Π^K. In other words, random observation $\omega^K \sim p_\mu$ is a sample of observations $\omega_1, ..., \omega_K$, drawn, independently of each other, from the distributions $p_{\mu_1,1}$, $p_{\mu_2,2}$,...,$p_{\mu_K,K}$;
- The space $\mathcal{F}^K$ is comprised of all *separable* functions

  $$\phi(\omega^K) = \sum_{k=1}^K \phi_k(\omega_k)$$

  with $\phi_k(\cdot) \in \mathcal{F}_k$, $1 \leq k \leq K$.

It is immediately seen that the direct product of simple observation o.s.'s is simple.

When all factors $\mathcal{O}_k$, $1 \leq k \leq K$, are identical to simple o.s.

$$\mathcal{O} = (\Omega, \Pi; \{p_\mu : \mu \in \mathcal{M}\}; \mathcal{F}),$$

the direct product of the factors can be "truncated" to yield the *K-th power* (called also the *stationary K-repeated version*) of $\mathcal{O}$, denoted

$$[\mathcal{O}]^K = (\Omega^K, \Pi^K; \{p_\mu^{(K)} : \mu \in \mathcal{M}\}; \mathcal{F}^{(K)})$$

and defined as follows:

- Ω^K and Π^K are exactly the same as in the direct product:
$$\Omega^K = \underbrace{\Omega \times ... \times \Omega}_{K},\ \Pi^K = \underbrace{\Pi \times ... \times \Pi}_{K};$$
- the parameter space is $\mathcal{M}$ rather than the direct product of K copies of $\mathcal{M}$, and the densities are
$$p_\mu^{(K)}(\omega^K = (\omega_1, ..., \omega_K)) = \prod_{k=1}^K p_\mu(\omega_k);$$
in other words, random observations $\omega^K \sim p_\mu^{(K)}$ are K-element samples with components drawn, independently of each other, from p_μ;
- the space $\mathcal{F}^{(K)}$ is comprised of separable functions
$$\phi^{(K)}(\omega^K) = \sum_{k=1}^K \phi(\omega_k)$$
with identical components belonging to $\mathcal{F}$ (i.e., $\phi \in \mathcal{F}$).

It is immediately seen that a power of simple o.s. is simple.

Remark 2.4.1 *Gaussian, Poisson and Discrete o.s.'s clearly are nondegenerate. It is also clear that the direct product of nondegenerate o.s.'s is nondegenerate.*

2.4.4 Simple observation schemes – Main result

We are about to demonstrate that when deciding on *convex*, in some precise sense to be specified below, hypotheses in *simple* observation schemes, optimal detectors can be found efficiently by solving *convex-concave saddle point problems.*

We start with "executive summary" on convex-concave saddle point problems.

2.4.4.1 Executive summary of convex-concave saddle point problems

The results to follow are absolutely standard, and their proofs can be found in all textbooks on the subject; for reader's convenience, we present the proofs in Appendix, Section D.4.

Let U and V be nonempty sets, and $\Phi : U \times V \to \mathbf{R}$ be a function. These data define an *antagonistic game* of two players, I and II, where player I selects a point $u \in U$, and player II selects a point $v \in V$; as an outcome of these selections, player I pays to player II the sum $\Phi(u, v)$. Clearly, the player I is interested to minimize this payment, and player II – to maximize the payment. The data U, V, Φ are known to the players in advance, and the question is, what should be their selections.

When the player I makes his selection u first, and player II makes his selection v with u already known, player I should be ready to pay for a selection $u \in U$ the toll as large as

$$\overline{\Phi}(u) = \sup_{v \in V} \Phi(u, v).$$

In this situation, a risk-averse player I would select u by minimizing the above worst-case payment, by solving the *primal* problem

$$\mathrm{Opt}(P) = \inf_{u \in U} \overline{\Phi}(u) = \inf_{u \in U} \sup_{v \in V} \Phi(u, v) \tag{P}$$

associated with the data U, V, Φ.

Similarly, if player II makes his selection v first, and player I selects u after v becomes known, player II should be ready to get, as a result of selecting $v \in V$, the amount as small as

$$\underline{\Phi}(v) = \inf_{u \in U} \phi(u, v).$$

In this situation, a risk-averse player II would select v by maximizing the above worst-case payment, by solving the *dual* problem

$$\mathrm{Opt}(D) = \sup_{v \in V} \underline{\Phi}(v) = \sup_{v \in V} \inf_{u \in U} \Phi(u, v) \tag{D}$$

Intuitively, the first situation is less preferable for player I than the second one, so that his guaranteed payment in the first situation, that is, Opt(P), should be $\geq$ his guaranteed payment, Opt(D), in the second situation:

$$\mathrm{Opt}(P) := \inf_{u \in U} \sup_{v \in V} \Phi(u, v) \geq \sup_{v \in V} \inf_{u \in U} \Phi(u, v) =: \mathrm{Opt}(D); \tag{2.4.3}$$

this fact, called *Weak Duality*, indeed is true.

The central question related to the game is what should the players do when making their selections simultaneously, with no knowledge of what is selected by the adversary. There is a case when this question has a completely satisfactory answer – this is the case where Φ has a *saddle point* on $U \times V$.

Definition 2.4.1 *A point $(u_*, v_*) \in U \times V$ is called a saddle point* [10] *of function $\Phi(u, v) : U \times V \to$ $\mathbf{R}$, if Φ as a function of $u \in U$ attains at this point its minimum, and as a function of $v \in V$ – its maximum, that is, if*

$$\Phi(u, v_*) \geq \Phi(u_*, v_*) \geq \Phi(u_*, v) \; \forall (u \in U, v \in V).$$

From the viewpoint of our game, a saddle point (u_*, v_*) is an equilibrium: when one of the players sticks to the selection stemming from this point, the other one has no incentive to deviate from his selection stemming from the point: if player II selects v_*, there is no reason for player I to deviate from selecting u_*, since with another selection, his loss (the payment) can only increase; similarly, when player I selects u_*, there is no reason for player II to deviate from v_*, since with any other selection, his gain (the payment) can only decrease. As a result, if the cost function Φ has a saddle point on $U \times V$, this saddle point (u_*, v_*) can be considered as a solution to the game, as the pair of preferred selections of rational players. It can be easily seen that while Φ can have many saddle points, the values of Φ at all these points are equal to each other, let us denote their common

[10] more precisely, "saddle point (min in $u \in U$, max in $v \in V$);" we will usually skip the clarification in parentheses, since it always will be clear from the context what are the minimization variables and what are the maximization ones.

value by SadVal. If (u_*, v_*) is a saddle point and player I selects $u = u_*$, his worst, over selections $v \in V$ of player II, loss is SadVal, and if player I selects a whatever $u \in U$, his worst-case, over the selections of player II, loss can be only $\geq$ SadVal. Similarly, when player II selects $v = v_*$, his worst-case, over the selections of player I, gain is SadVal, and if player II selects a whatever $v \in V$, his worst-case, over the selections of player I, gain can be only $\leq$ SadVal.

Existence of saddle points of Φ (min in $u \in U$, max in $v \in V$) can be expressed in terms of the primal problem (P) and the dual problem (P):

Proposition 2.4.1 *Φ has saddle point* (min *in* $u \in U$, max *in* $v \in V$) *if and only if problems* (P) *and* (D) *are solvable with equal optimal values:*

$$\mathrm{Opt}(P) := \inf_{u\in U} \sup_{v\in V} \Phi(u, v) = \sup_{v\in V} \inf_{u\in U} \Phi(u, v) =: \mathrm{Opt}(D). \tag{2.4.4}$$

Whenever this is the case, the saddle points of Φ are exactly the pairs (u_*, v_*) comprised of optimal solutions to problems (P) and (D), and the value of Φ at every one of these points is the common value SadVal *of* $\mathrm{Opt}(P)$ *and* $\mathrm{Opt}(D)$.

Existence of a saddle point of a function is “rare commodity;” the standard sufficient condition for it is convexity-concavity of Φ coupled with convexity of U and V; the precise statement is as follows:

Theorem 2.4.1 [Sion-Kakutani, see Section D.4.2] *Let $U \subset \mathbf{R}^m, V \subset \mathbf{R}^n$ be nonempty closed convex sets, with W bounded, and let $\Phi : U \times V \to \mathbf{R}$ be continuous function which is convex in $u \in U$ for every fixed $v \in V$, and is concave in $v \in V$ for every fixed $u \in U$. Then the equality* (2.4.4) *holds true (although it may happen that* $\mathrm{Opt}(P) = \mathrm{Opt}(D) = -\infty$*).*

If, in addition, Φ is coercive in u, meaning that the level sets

$$\{u \in U : \Phi(u, v) \leq a\}$$

are bounded for every $a \in \mathbf{R}$ and $v \in V$ (equivalently: for every $v \in V$, $\Phi(u_i, v) \to +\infty$ along every sequence $u_i \in U$ going to ∞: $\|u_i\| \to \infty$ as $i \to \infty$), then Φ admits saddle points (min *in* $u \in U$, max *in* $v \in V$).

Note that the “true” Sion-Kakutani Theorem is a bit stronger than Theorem 2.4.1; the latter, however, covers all our related needs.

2.4.4.2 Main Result

Theorem 2.4.2 *Let*

$$\mathcal{O} = (\Omega, \Pi; \{p_\mu : \mu \in \mathcal{M}\}; \mathcal{F})$$

be a simple observation scheme, and let M_1, M_2 be nonempty compact convex subsets of $\mathcal{M}$. Then

(i) *The function*

$$\Phi(\phi, [\mu; \nu]) = \tfrac{1}{2}\left[\ln\left(\int_\Omega \mathrm{e}^{-\phi(\omega)} p_\mu(\omega)\Pi(d\omega)\right) + \ln\left(\int_\Omega \mathrm{e}^{\phi(\omega)} p_\nu(\omega)\Pi(d\omega)\right)\right] : \mathcal{F} \times (M_1 \times M_2) \to \mathbf{R} \tag{2.4.5}$$

is continuous on its domain, is convex in $\phi(\cdot) \in \mathcal{F}$, concave in $[\mu; \nu] \in M_1 \times M_2$, and possesses a saddle point (min *in* $\phi \in \mathcal{F}$, max *in* $[\mu; \nu] \in M_1 \times M_2$) $(\phi_*(\cdot), [\mu_*; \nu_*])$ *on* $\mathcal{F} \times (M_1 \times M_2)$. *$\phi_*$ w.l.o.g. can be assumed to satisfy the relation*[11]

$$\int_\Omega \exp\{-\phi_*(\omega)\} p_{\mu_*}(\omega)\Pi(d\omega) = \int_\Omega \exp\{\phi_*(\omega)\} p_{\mu_*}(\omega)\Pi(d\omega). \tag{2.4.6}$$

[11] Note that $\mathcal{F}$ contains constants, and shifting by a constant the ϕ-component of a saddle point of Φ and keeping its $[\mu; \nu]$-component intact, we clearly get another saddle point of Φ.

Denoting the common value of the two quantities in (2.4.6) *by* $\varepsilon_\star$*, the saddle point value*

$$\min_{\phi\in\mathcal{F}}\max_{[\mu;\nu]\in M_1\times M_2}\Phi(\phi,[\mu;\nu])$$

is $\ln(\varepsilon_\star)$*. Besides this, setting* $\phi_*^a(\cdot)=\phi_*(\cdot)-a$*, one has*

$$\begin{array}{lrcl}(a) & \int_\Omega \exp\{-\phi_*^a(\omega)\}p_\mu(\omega)\Pi(d\omega) & \leq & \exp\{a\}\varepsilon_\star\ \forall \mu\in M_1,\\ (b) & \int_\Omega \exp\{\phi_*^a(\omega)\}p_\nu(\omega)\Pi(d\omega) & \leq & \exp\{-a\}\varepsilon_\star\ \forall \nu\in M_2.\end{array}, \tag{2.4.7}$$

implying, in view of Proposition 2.3.1, that when deciding via an observation $\omega\in\Omega$ *on the hypotheses*

$$H_\chi:\omega\sim p_\mu \text{ with } \mu\in M_\chi,\quad \chi=1,2,$$

the risks of the simple test $\mathcal{T}_{\phi_*^a}$ *based on the detector* ϕ_*^a *can be upper-bounded as follows:*

$$\mathrm{Risk}_1(\mathcal{T}_{\phi_*^a}|H_1,H_2)\leq\exp\{a\}\varepsilon_\star,\ \mathrm{Risk}_2(\mathcal{T}_{\phi_*^a}|H_1,H_2)\leq\exp\{-a\}\varepsilon_\star. \tag{2.4.8}$$

Besides this, $\phi_*,\varepsilon_\star$ *form an optimal solution to the optimization problem*

$$\min_{\phi,\epsilon}\left\{\epsilon:\begin{array}{l}\int_\Omega \mathrm{e}^{-\phi(\omega)}p_\mu(\omega)\Pi(d\omega)\leq\epsilon\,\forall\mu\in M_1\\ \int_\Omega \mathrm{e}^{\phi(\omega)}p_\mu(\omega)\Pi(d\omega)\leq\epsilon\,\forall\mu\in M_2\end{array}\right\} \tag{2.4.9}$$

(the minimum in (2.4.9) *is taken over all* $\epsilon>0$ *and all* Π*-measurable functions* $\phi(\cdot)$*, not just over* $\phi\in\mathcal{F}$*).*

(ii) *The dual problem associated with the saddle point data* Φ*,* $\mathcal{F}$*,* $M_1\times M_2$ *is*

$$\max_{\mu\in M_1,\nu\in M_2}\underline{\Phi}(\mu,\nu):=\inf_{\phi\in\mathcal{F}}\Phi(\phi;[\mu;\nu]). \tag{D}$$

The objective in this problem is in fact the logarithm of Hellinger affinity of p_μ *and* p_ν*:*

$$\underline{\Phi}(\mu,\nu)=\ln\left(\int_\Omega\sqrt{p_\mu(\omega)p_\nu(\omega)}\Pi(d\omega)\right), \tag{2.4.10}$$

and this objective is concave and continuous on $M_1\times M_2$*.*

The (μ,ν)*-components of saddle points of* Φ *are exactly the maximizers* (μ_*,ν_*) *of the concave function* $\underline{\Phi}$ *on* $M_1\times M_2$*. Given such a maximizer* $[\mu_*;\nu_*]$ *and setting*

$$\phi_*(\omega)=\frac{1}{2}\ln(p_{\mu_*}(\omega)/p_{\nu_*}(\omega)) \tag{2.4.11}$$

we get a saddle point $(\phi_*,[\mu_*;\nu_*])$ *of* Φ *satisfying* (2.4.6)*.*

(iii) *Let* $[\mu_*;\nu_*]$ *be a maximizer of* $\underline{\Phi}$ *over* $M_1\times M_2$*. Let, further,* $\epsilon\in[0,1/2]$ *be such that there exists a (whatever, perhaps randomized) test for deciding via observation* $\omega\in\Omega$ *on two simple hypotheses*

$$(A):\omega\sim p(\cdot):=p_{\mu_*}(\cdot),\quad (B):\omega\sim q(\cdot):=p_{\nu_*}(\cdot) \tag{2.4.12}$$

with total risk $\leq 2\epsilon$*. Then*

$$\varepsilon_\star\leq 2\sqrt{\epsilon(1-\epsilon)}.$$

In other words, if the simple hypotheses (A)*,* (B) *can be decided, by a whatever test, with total risk* 2ϵ*, then the risks of the simple test with detector* ϕ_* *given by* (2.4.11) *on the composite hypotheses* H_1*,* H_2 *do not exceed* $2\sqrt{\epsilon(1-\epsilon)}$*.*

Proof. 1^0. Since $\mathcal{O}$ is a simple o.s., the function $\Phi(\phi,[\mu;\nu])$ given by (2.4.5) is a well defined real-valued function on $\mathcal{F}\times(\mathcal{M}\times\mathcal{M})$ which is concave in $[\mu;\nu]$; convexity of the function in $\phi\in\mathcal{F}$ is evident. Since both $\mathcal{F}$ and $\mathcal{M}$ are convex sets coinciding with their relative interiors, convexity-concavity and real valuedness of Φ on $\mathcal{F}\times(\mathcal{M}\times\mathcal{M})$ imply the continuity of Φ on the indicated domain. As a consequence, Φ is convex-concave continuous real-valued function on $\mathcal{F}\times(M_1\times M_2)$.

Now let

$$\underline{\Phi}(\mu,\nu)=\inf_{\phi\in\mathcal{F}}\Phi(\phi,[\mu;\nu]). \tag{2.4.13}$$

Note that $\underline{\Phi}$, being the infimum of a family of concave functions of $[\mu;\nu]\in\mathcal{M}\times\mathcal{M}$, is concave on $\mathcal{M}\times\mathcal{M}$. We claim that for $\mu,\nu\in\mathcal{M}$ the function

$$\phi_{\mu,\nu}(\omega)=\frac{1}{2}\ln(p_\mu(\omega)/p_\nu(\omega))$$

(which, by definition of a simple o.s., belongs to $\mathcal{F}$) is an optimal solution to the right hand side minimization problem in (2.4.13), so that

$$\forall(\mu\in M_1,\nu\in M_2):\underline{\Phi}([x;y]):=\inf_{\phi\in\mathcal{F}}\Phi(\phi,[\mu;\nu])=\Phi(\phi_{\mu,\nu},[\mu;\nu])=\ln\left(\int_\Omega\sqrt{p_\mu(\omega)p_\nu(\omega)}\Pi(d\omega)\right). \tag{2.4.14}$$

Indeed, we have

$$\exp\{-\phi_{\mu,\nu}(\omega)\}p_\mu(\omega)=\exp\{\phi_{\mu,\nu}(\omega)\}p_\nu(\omega)=g(\omega):=\sqrt{p_\mu(\omega)p_\nu(\omega)},$$

whence $\Phi(\phi_{\mu,\nu},[\mu;\nu])=\ln\left(\int_\Omega g(\omega)\Pi(d\omega)\right)$. On the other hand, for $\phi(\cdot)=\phi_{\mu,\nu}(\cdot)+\delta(\cdot)\in\mathcal{F}$ we have

$$\begin{array}{rl}
& \int_\Omega g(\omega)\Pi(d\omega)=\int_\Omega\left[\sqrt{g(\omega)}\exp\{-\delta(\omega)/2\}\right]\left[\sqrt{g(\omega)}\exp\{\delta(\omega)/2\}\right]\Pi(d\omega)\\
(a) & \leq\left(\int_\Omega g(\omega)\exp\{-\delta(\omega)\}\Pi(d\omega)\right)^{1/2}\left(\int_\Omega g(\omega)\exp\{\delta(\omega)\}\Pi(d\omega)\right)^{1/2}\\
& =\left(\int_\Omega \exp\{-\phi(\omega)\}p_\mu(\omega)\Pi(d\omega)\right)^{1/2}\left(\int_\Omega \exp\{\phi(\omega)\}p_\nu(\omega)\Pi(d\omega)\right)^{1/2}\\
(b) & \Rightarrow\ln\left(\int_\Omega g(\omega)\Pi(d\omega)\right)\leq\Phi(\phi,[\mu;\nu]),
\end{array}$$

and thus $\Phi(\phi_{\mu,\nu},[\mu;\nu])\leq\Phi(\phi,[\mu;\nu])$ for every $\phi\in\mathcal{F}$.

Remark 2.4.2 Note that the above reasoning did not use the fact that the minimization in the right hand side of (2.4.13) is over $\phi\in\mathcal{F}$; in fact, this reasoning shows that $\phi_{\mu,\nu}(\cdot)$ minimizes $\Phi(\phi,[\mu;\nu])$ over all functions ϕ for which the integrals $\int_\Omega\exp\{-\phi(\omega)\}p_\mu(\omega)\Pi(d\omega)$ and $\int_\Omega\exp\{\phi(\omega)\}p_\nu(\omega)\Pi(d\omega)$ exist.

Remark 2.4.3 Note that the inequality in (b) can be equality only when the inequality in (a) is so. In other words, if $\bar\phi$ is a minimizer of $\Phi(\phi,[\mu;\nu])$ over $\phi\in\mathcal{F}$, setting $\delta(\cdot)=\bar\phi(\cdot)-\phi_{\mu,\nu}(\cdot)$, the functions $\sqrt{g(\omega)}\exp\{-\delta(\omega)/2\}$ and $\sqrt{g(\omega)}\exp\{\delta(\omega)/2\}$, considered as elements of $L_2[\Omega,\Pi]$, are proportional to each other. Since g is positive and g,δ are continuous, while the support of Π is the entire Ω, this "L_2-proportionality" means that the functions in question differ by a constant factor, or, which is the same, that $\delta(\cdot)$ is constant. Thus, *the minimizers of $\Phi(\phi,[\mu;\nu])$ over $\phi\in\mathcal{F}$ are exactly the functions of the form $\phi(\omega)=\phi_{\mu,\nu}(\omega)+\text{const}$.*

2^0. We are about to verify that $\Phi(\phi,[\mu;\nu])$ has a saddle point (min in $\phi\in\mathcal{F}$, max in $[\mu;\nu]\in M_1\times M_2$). Indeed, observe, first, that on the domain of Φ it holds

$$\Phi(\phi(\cdot)+a,[\mu;\nu])=\Phi(\phi(\cdot),[\mu;\nu])\ \forall(a\in\mathbf{R},\phi\in\mathcal{F}). \tag{2.4.15}$$

Let us select somehow $\bar\mu\in\mathcal{M}$, and let $\bar\Pi$ be the measure on Ω with density $p_{\bar\mu}$ w.r.t. Π. For $\phi\in\mathcal{F}$, the integrals $\int_\Omega e^{\pm\phi(\omega)}\bar\Pi(d\omega)$ are finite (since $\mathcal{O}$ is simple), implying that $\phi\in L_1[\Omega,\bar\Pi]$; note also that $\bar\Pi$ is a probabilistic measure. Let now $\mathcal{F}_0=\{\phi\in\mathcal{F}:\int_\Omega\phi(\omega)\bar\Pi(d\omega)=0\}$, so that $\mathcal{F}_0$ is a linear subspace in $\mathcal{F}$, and all functions $\phi\in\mathcal{F}$ can be obtained by shifts of functions from $\mathcal{F}_0$ by constants. Invoking (2.4.15), to prove the existence of a saddle point of Φ on $\mathcal{F}\times(M_1\times M_2)$ is exactly the same as to prove the existence of a saddle point of Φ on $\mathcal{F}_0\times(M_1\times M_2)$. Let us verify that $\Phi(\phi,[\mu;\nu])$ indeed has a saddle point on $\mathcal{F}_0\times(M_1\times M_2)$. $M_1\times M_2$ is a convex compact set, and Φ is continuous on $\mathcal{F}_0\times(M_1\times M_2)$ and convex-concave; invoking Sion-Kakutani Theorem we see that all we need in order to verify the existence of a saddle point is to show that Φ is coercive in the first argument, that is, for every fixed $[\mu;\nu]\in M_1\times M_2$ one has $\Phi(\phi,[\mu;\nu])\to+\infty$ as $\phi\in\mathcal{F}_0$ and $\|\phi\|\to\infty$ (whatever be the norm $\|\cdot\|$ on $\mathcal{F}_0$; recall that $\mathcal{F}_0$ is a finite-dimensional linear space). Setting

$$\Theta(\phi)=\Phi(\phi,[\mu;\nu])=\frac{1}{2}\left[\ln\left(\int_\omega \mathrm{e}^{-\phi(\omega)}p_\mu(\omega)\Pi(d\omega)\right)+\ln\left(\int_\omega \mathrm{e}^{\phi(\omega)}p_\nu(\omega)\Pi(d\omega)\right)\right]$$

and taking into account that Θ is convex and finite on $\mathcal{F}_0$, in order to prove that Θ is coercive, it suffices to verify that $\Theta(t\phi)\to\infty$, $t\to\infty$, for every nonzero $\phi\in\mathcal{F}_0$, which is evident: since $\int_\Omega\phi(\omega)\bar{\Pi}(d\omega)=0$ and ϕ is nonzero, we have $\int_\Omega\max[\phi(\omega),0]\bar{\Pi}(d\omega)=\int_\Omega\max[-\phi(\omega),0]\bar{\Pi}(d\omega)>0$, whence $\phi>0$ and $\phi<0$ on sets of Π-positive measure, so that $\Theta(t\phi)\to\infty$ as $t\to\infty$ due to the fact that both $p_\mu(\cdot)$ and $p_\nu(\cdot)$ are positive everywhere.

3^0. Now let $(\phi_*(\cdot);[\mu_*;\nu_*])$ be a saddle point of Φ on $\mathcal{F}\times(M_1\times M_2)$. Shifting, if necessary, $\phi_*(\cdot)$ by a constant (by (2.4.15), this does not affect the fact that $(\phi_*,[\mu_*;\nu_*])$ is a saddle point of Φ), we can assume that

$$\varepsilon_\star:=\int_\Omega\exp\{-\phi_*(\omega)\}p_{\mu_*}(\omega)\Pi(d\omega)=\int_\Omega\exp\{\phi_*(\omega)\}p_{\nu_*}(\omega)\Pi(d\omega),\tag{2.4.16}$$

so that the saddle point value of Φ is

$$\Phi_*:=\max_{[\mu;\nu]\in M_1\times M_2}\min_{\phi\in\mathcal{F}}\Phi(\phi,[\mu;\nu])=\Phi(\phi_*,[\mu_*;\nu_*])=\ln(\varepsilon_\star).\tag{2.4.17}$$

as claimed in item (i) of Theorem.

Now let us prove (2.4.7). For $\mu\in M_1$, we have

$$\begin{array}{rcl}\ln(\varepsilon_\star)&=&\Phi_*\geq\Phi(\phi_*,[\mu;\nu_*])\\&=&\frac{1}{2}\ln\left(\int_\Omega\exp\{-\phi_*(\omega)\}p_\mu(\omega)\Pi(d\omega)\right)+\frac{1}{2}\ln\left(\int_\Omega\exp\{\phi_*(\omega)\}p_{\nu_*}(\omega)\Pi(d\omega)\right)\\&=&\frac{1}{2}\ln\left(\int_\Omega\exp\{-\phi_*(\omega)\}p_\mu(\omega)P(d\omega)\right)+\frac{1}{2}\ln(\varepsilon_\star),\end{array}$$

whence $\ln\left(\int_\Omega\exp\{-\phi_*^a(\omega)\}p_\mu(\omega)\Pi(d\omega)\right)=\ln\left(\int_\Omega\exp\{-\phi_*(\omega)\}p_\mu(\omega)P(d\omega)\right)+a\leq\ln(\varepsilon_\star)+a$, and (2.4.7.$a$) follows. Similarly, when $\nu\in M_2$, we have

$$\begin{array}{rcl}\ln(\varepsilon_\star)&=&\Phi_*\geq\Phi(\phi_*,[\mu_*;\nu])\\&=&\frac{1}{2}\ln\left(\int_\Omega\exp\{-\phi_*(\omega)\}p_{\mu_*}(\omega)\Pi(d\omega)\right)+\frac{1}{2}\ln\left(\int_\Omega\exp\{\phi_*(\omega)\}p_\nu(\omega)\Pi(d\omega)\right)\\&=&\frac{1}{2}\ln(\varepsilon_\star)+\frac{1}{2}\ln\left(\int_\Omega\exp\{\phi_*(\omega)\}p_\nu(\omega)\Pi(d\omega)\right),\end{array}$$

so that $\ln\left(\int_\Omega\exp\{\phi_*^a(\omega)\}p_\nu(\omega)\Pi(d\omega)\right)=\ln\left(\int_\Omega\exp\{\phi_*(\omega)\}p_\nu(\omega)\Pi(d\omega)\right)-a\leq\ln(\varepsilon_\star)-a$, and (2.4.7.$b$) follows.

We have proved all claims in item (i), except for the claim that the just defined $\phi_*,\varepsilon_\star$ form an optimal solution to (2.4.9). Note that by (2.4.7) as applied with $a=0$, the pair in question is feasible for (2.4.9). Assuming that the problem admits a feasible solution $(\bar{\phi},\epsilon)$ with $\epsilon<\varepsilon_\star$, let us lead this assumption to a contradiction. Note that $\bar{\phi}$ should be such that

$$\int_\Omega\mathrm{e}^{-\bar{\phi}(\omega)}p_{\mu_*}(\omega)\Pi(d\omega)<\varepsilon_\star\ \&\ \int_\Omega\mathrm{e}^{\bar{\phi}(\omega)}p_{\nu_*}(\omega)\Pi(d\omega)<\varepsilon_\star,$$

and consequently $\Phi(\bar{\phi},[\mu_*;\nu_*])<\ln(\varepsilon_\star)$. On the other hand, Remark 2.4.2 says that $\Phi(\bar{\phi},[\mu_*;\nu_*])$ cannot be less than $\min_{\phi\in\mathcal{F}}\Phi(\phi,[\mu_*;\nu_*])$, and the latter quantity is $\Phi(\phi_*,[\mu_*;\nu_*])$ due to the fact that $(\phi_*,[\mu_*;\nu_*])$ is a saddle point of Φ on $\mathcal{F}\times(M_1\times M_2)$. Thus, assuming that the optimal value in (2.4.9) is $<\varepsilon_\star$, we conclude that $\Phi(\phi_*,[\mu_*;\nu_*])\leq\Phi(\bar{\phi},[\mu_*;\nu_*])<\ln(\varepsilon_\star)$, contradicting (2.4.17). Item (i) of Theorem 2.4.2 is proved.

4^0. Let us prove item (ii) of Theorem 2.4.2. Relation (2.4.10) and concavity of the right hand side of this relation in $[\mu;\nu]$ were already proved; moreover, these relations were proved in the range $\mathcal{M}\times\mathcal{M}$ of $[\mu;\nu]$. Since this range coincides with its relative interior, the real-valued concave function $\underline{\Phi}$ is continuous in $\mathcal{M}\times\mathcal{M}$ and thus is continuous in $M_1\times M_2$. Next, let ϕ_* be the ϕ-component of a saddle point of Φ on $\mathcal{F}\times(M_1\times M_2)$ (we already know that a saddle point exists). Invoking Proposition 2.4.1, the $[\mu;\nu]$-components of saddle points of Φ on $\mathcal{F}\times(M_1\times M_2)$ are exactly the maximizers of $\underline{\Phi}$ on $M_1\times M_2$. Let $[\mu_*;\nu_*]$ be such a maximizer; by the same Proposition 2.4.1, $(\phi_*,[\mu_*;\nu_*])$ is a saddle point of Φ, whence $\Phi(\phi,[\mu_*;\nu_*])$ attains its minimum over $\phi\in\mathcal{F}$ at $\phi=\phi_*$. We have also seen that $\Phi(\phi,[\mu_*;\nu_*])$ attains its minimum over $\phi\in\mathcal{F}$ at $\phi=\phi_{\mu_*,\nu_*}$. These observations combine with Remark 2.4.3 to imply that ϕ_* and ϕ_{μ_*,ν_*} differ by a constant, which, in view of (2.4.15), means that $(\phi_{\mu_*,\nu_*},[\mu_*;\nu_*])$ is a saddle point of Φ along with $(\phi_*,[\mu_*;\nu_*])$. (ii) is proved.

5^0. It remains to prove item (iii) of Theorem 2.4.2. In the notation from (iii), simple hypotheses (A) and (B) can be decided with the total risk $\leq 2\epsilon$, and therefore, by Proposition 2.1.1,

$$2\bar{\epsilon} := \int_\Omega \min[p(\omega), q(\omega)]\Pi(d\omega) \leq 2\epsilon.$$

On the other hand, we have seen that the saddle point value of Φ is $\ln(\varepsilon_\star)$; since $[\mu_*;\nu_*]$ is a component of a saddle point of Φ, it follows that $\min_{\phi\in\mathcal{F}}\Phi(\phi,[\mu_*;\nu_*]) = \ln(\varepsilon_\star)$. The left hand side in this equality, as we know from item 1^0, is $\Phi(\phi_{x_*,y_*},[x_*;y_*])$, and we arrive at $\ln(\varepsilon_\star) = \Phi(\frac{1}{2}\ln(p_{\mu_*}(\cdot)/p_{\nu_*}(\cdot)),[\mu_*;\nu_*]) = \ln\left(\int_\Omega \sqrt{p_{\mu_*}(\omega)p_{\nu_*}(\omega)}\Pi(d\omega)\right)$, so that $\varepsilon_\star = \int_\Omega \sqrt{p_{\mu_*}(\omega)p_{\nu_*}(\omega)}\Pi(d\omega) = \int_\Omega \sqrt{p(\omega)q(\omega)}\Pi(d\omega)$. We now have

$$\begin{array}{l}
\varepsilon_\star = \int_\Omega \sqrt{p(\omega)q(\omega)}\Pi(d\omega) = \int_\Omega \sqrt{\min[p(\omega),q(\omega)]}\sqrt{\max[p(\omega),q(\omega)]}\Pi(d\omega)\\
\leq \left(\int_\Omega \min[p(\omega),q(\omega)]\Pi(d\omega)\right)^{1/2}\left(\int_\Omega \max[p(\omega),q(\omega)]\Pi(d\omega)\right)^{1/2}\\
= \left(\int_\Omega \min[p(\omega),q(\omega)]\Pi(d\omega)\right)^{1/2}\left(\int_\Omega (p(\omega)+q(\omega)-\min[p(\omega),q(\omega)])\Pi(d\omega)\right)^{1/2}\\
= \sqrt{2\bar{\epsilon}(2-2\bar{\epsilon})} \leq 2\sqrt{(1-\epsilon)\epsilon},
\end{array}$$

where the concluding inequality is due to $\bar{\epsilon} \leq \epsilon \leq 1/2$. (iii) is proved, and the proof of Theorem 2.4.2 is complete. □

Remark 2.4.4 *Assume that we are under the premise of Theorem 2.4.2 and that the simple o.s. in question is nondegenerate (see Section 2.4.2). Then $\varepsilon_\star < 1$ if and only if the sets M_1 and M_2 do not intersect.*

Indeed, by Theorem 2.4.2.i, $\ln(\varepsilon_\star)$ is the saddle point value of $\Phi(\phi,[\mu;\nu])$ on $\mathcal{F}\times(M_1\times M_2)$, or, which is the same by Theorem 2.4.2.ii, the maximum of the function (2.4.10) on $M_1\times M_2$; since saddle points exist, this maximum is achieved at some pair $[\mu;\nu]\in M_1\times M_2$. Since (2.4.10) clearly is ≤ 0, we conclude that $\varepsilon_\star \leq 1$ and the equality takes place if and only if $\int_\Omega \sqrt{p_\mu(\omega)p_\nu(\omega)}\Pi(d\omega) = 1$ for some $\mu\in M_1$ and $\nu\in M_2$, or, which is the same, $\int_\Omega(\sqrt{p_\mu(\omega)}-\sqrt{p_\nu(\omega)})^2\Pi(d\omega) = 0$ for these μ and ν. Since $p_\mu(\cdot)$ and $p_\nu(\cdot)$ are continuous and the support of Π is the entire Ω, the latter can happen if and only if $p_\mu = p_\nu$ for our μ, ν, or, by nondegeneracy of $\mathcal{O}$, if and only if $M_1\cap M_2 \neq \emptyset$. □

2.4.5 Simple observation schemes – Examples of optimal detectors

Theorem 2.4.2.i states that when the observation scheme

$$\mathcal{O} = (\Omega,\Pi;\{p_\mu : \mu\in\mathcal{M}\};\mathcal{F})$$

is simple and we are interested to decide on a pair of hypotheses on the distribution of observation $\omega\in\Omega$,

$$H_\chi : \omega \sim p_\mu \text{ with } \mu\in M_\chi,\ \chi = 1,2$$

and *the hypotheses are convex*, meaning that the underlying parameter sets M_χ are convex and compact, building optimal, in terms of its risk, detector ϕ_* – that is, solving (in general, semi-infinite and infinite-dimensional) optimization problem (2.4.9) reduces to solving the usual finite-dimensional convex problem. Specifically, an optimal solution $(\phi_*,\varepsilon_\star)$ can be built as follows:

1. We solve optimization problem

$$\mathrm{Opt} = \max_{\mu\in M_1,\nu\in M_2}\left[\underline{\Phi}(\mu,\nu) := \ln\left(\int_\Omega \sqrt{p_\mu(\omega)p_\nu(\omega)}\Pi(d\omega)\right)\right]; \qquad (2.4.18)$$

of maximizing Hellinger affinity (the quantity under the logarithm) of a pair of distributions obeying H_1 and H_2, respectively; for a simple o.s., the objective in this problem is concave and continuous, and optimal solutions do exist;

2. (Any) optimal solution $[\mu_*;\nu_*]$ to (2.4.18) gives rise to an optimal detector ϕ_* and its risk $\varepsilon_\star$, according to
$$\phi_*(\omega)=\frac{1}{2}\ln\left(\frac{p_{\mu_*}(\omega)}{p_{\nu_*}(\omega)}\right),\ \varepsilon_\star=\exp\{\mathrm{Opt}\}. \tag{2.4.19}$$
The risks of the simple test $\mathcal{T}_{\phi_*}$ associated with the above detector and deciding on H_1, H_2, satisfy the bounds
$$\max\left[\mathrm{Risk}_1(\mathcal{T}_{\phi_*}|H_1,H_2),\mathrm{Risk}_2(\mathcal{T}_{\phi_*}|H_1,H_2)\right]\leq\varepsilon_\star, \tag{2.4.20}$$
and the test is *near-optimal*, meaning that whenever the hypotheses H_1, H_2 (and in fact – even two simple hypotheses stating that $\omega\sim p_{\mu_*}$ and $\omega\sim p_{\nu_*}$, respectively) can be decided upon by a test with total risk $\leq 2\epsilon$, $\mathcal{T}_{\phi_*}$ exhibits “comparable” risk:
$$\varepsilon_\star\leq 2\sqrt{\epsilon(1-\epsilon)}. \tag{2.4.21}$$
Note that *the test $\mathcal{T}_{\phi_*}$ is just the maximum likelihood test induced by the probability densities p_{μ_*} and p_{ν_*}.*

Note that after we know that $(\phi_*,\varepsilon_\star)$ form an optimal solution to (2.4.9), some kind of near-optimality of the test $\mathcal{T}_{\phi_*}$ is guaranteed already by Proposition 2.3.5; specifically, by this Proposition, whenever in the nature there exists a test $\mathcal{T}$ which decides on H_1,H_2 with risks $\mathrm{Risk}_1,\mathrm{Risk}_2$ bounded by some $\epsilon\leq 1/2$, the upper bound $\varepsilon_\star$ on the risks of $\mathcal{T}_{\phi_*}$ can be bounded according to (2.4.21). Our now near-optimality statement is a bit stronger: first, we allow $\mathcal{T}$ to have the total risk $\leq 2\epsilon$, which is weaker than to have both risks $\leq\epsilon$; second, and more important, now 2ϵ should upper-bound the total risk of $\mathcal{T}$ on a pair of *simple* hypotheses “embedded” into the hypotheses H_1,H_2; both these modifications extend the family of tests $\mathcal{T}$ to which we compare the test $\mathcal{T}_{\phi_*}$, and thus enrich the comparison.

Let us look how the above recipe works for our basic simple o.s.'s.

2.4.5.1 Gaussian o.s.

When $\mathcal{O}$ is a Gaussian o.s., that is, $\{p_\mu:\mu\in\mathcal{M}\}$ are Gaussian densities with expectations $\mu\in\mathcal{M}=\mathbf{R}^d$ and common positive definite covariance matrix Θ, and $\mathcal{F}$ is the family of affine functions on $\Omega=\mathbf{R}^d$,

- M_1, M_2 can be arbitrary nonempty convex compact subsets of $\mathbf{R}^d$,
- problem (2.4.18) becomes the convex optimization problem
$$\mathrm{Opt}=-\min_{\mu\in M_1,\nu\in M_2}\frac{(\mu-\nu)^T\Theta^{-1}(\mu-\nu)}{8} \tag{2.4.22}$$
- the optimal detector ϕ_* and the upper bound $\varepsilon_\star$ on its risks given by an optimal solution (μ_*,ν_*) to (2.4.22) are
$$\begin{array}{rcl}\phi_*(\omega)&=&\frac{1}{2}[\mu_*-\nu_*]^T\Theta^{-1}[\omega-w],\ w=\frac{1}{2}[\mu_*+\nu_*]\\ \varepsilon_\star&=&\exp\{-\frac{[\mu_*-\nu_*]\Theta^{-1}[\mu_*-\nu_*]}{8}\}\end{array} \tag{2.4.23}$$

Note that when $\Theta=I_d$, the test $\mathcal{T}_{\phi_*}$ becomes exactly the optimal test from Example 2.1. The upper bound on the risks of this test established in Example 2.1 (in our present notation, this bound is $\mathrm{Erf}(\frac{1}{2}\|\mu_*-\nu_*\|_2)$) is slightly better than the bound $\varepsilon_\star=\exp\{-\|\mu_*-\nu_*\|_2^2/8\}$ given by (2.4.23) when $\Theta=I_d$. Note, however, that when speaking about the distance $\delta=\|\mu_*-\nu_*\|_2$ between M_1 and M_2 allowing for a test with risks $\leq\epsilon\ll 1$, the results of Example 2.1) and (2.4.23) say nearly the same: Example 2.1 says that δ should be $\geq 2\mathrm{ErfInv}(\epsilon)$, where $\mathrm{ErfInv}(\epsilon)$ is the Inverse Error function:

$$\mathrm{Erf}(\mathrm{ErfInv}(\epsilon)) \equiv \epsilon,\ 0 < \epsilon < 1,$$

and (2.4.23) says that δ should be $\geq 2\sqrt{2\ln(1/\epsilon)}$. When $\epsilon \to +0$, the ratio of these two lower bounds on δ tends to 1.

It should be noted that our general construction of optimal detectors as applied to Gaussian o.s. and a pair of convex hypotheses results in *exactly* optimal test and can be analyzed directly, without any "science" (see Example 2.1).

2.4.5.2 Poisson o.s.

When $\mathcal{O}$ is a Poisson o.s., that is, $\mathcal{M} = \mathbf{R}^d_{++}$ is the interior of nonnegative orthant in $\mathbf{R}^d$, and p_μ, $\mu \in \mathcal{M}$, is the density

$$p_\mu(\omega) = \prod_i \left(\frac{\mu_i^{\omega_i}}{\omega_i!}\mathrm{e}^{-\mu_i}\right),\ \omega = (\omega_!, ..., \omega_d) \in \mathbf{Z}^d_+$$

taken w.r.t. the counting measure Π on $\Omega = \mathbf{Z}^d_+$, and $\mathcal{F}$ is the family of affine functions on Ω, the recipe from the beginning of Section 2.4.5 reads as follows:

- M_1, M_2 can be arbitrary nonempty convex compact subsets of $\mathbf{R}^d_{++} = \{x \in \mathbf{R}^d : x > 0\}$;
- problem (2.4.18) becomes the convex optimization problem

$$\mathrm{Opt} = -\min_{\mu\in M_1,\nu\in M_2} \frac{1}{2}\sum_{i=1}^d (\sqrt{\mu_i} - \sqrt{\nu_i})^2\,; \tag{2.4.24}$$

- the optimal detector ϕ_* and the upper bound $\varepsilon_\star$ on its risks given by an optimal solution (μ^*, ν^*) to (2.4.24) are

$$\begin{array}{rcl}\phi_*(\omega) & = & \frac{1}{2}\sum_{i=1}^d \ln\left(\frac{\mu_i^*}{\nu_i^*}\right)\omega_i + \frac{1}{2}\sum_{i=1}^d[\nu_i^* - \mu_i^*],\\ \varepsilon_\star & = & \mathrm{e}^{\mathrm{Opt}}\end{array} \tag{2.4.25}$$

2.4.5.3 Discrete o.s.

When $\mathcal{O}$ is a Discrete o.s., that is, $\Omega = \{1, ..., d\}$, Π is a counting measure on Ω, $\mathcal{M} = \{\mu \in \mathbf{R}^d : \mu > 0, \sum_i \mu_i = 1\}$ and

$$p_\mu(\omega) = \mu_\omega,\ \omega = 1, ..., d,\ \mu \in \mathcal{M},$$

the recipe from the beginning of Section 2.4.5 reads as follows:

- M_1, M_2 can be arbitrary nonempty convex compact subsets of the relative interior $\mathcal{M}$ of the probabilistic simplex,
- problem (2.4.18) *is equivalent* to the convex program

$$\varepsilon_\star = \max_{\mu\in M_1,\nu\in M_2} \sum_{i=1}^d \sqrt{\mu_i\nu_i}; \tag{2.4.26}$$

- the optimal detector ϕ_* given by an optimal solution (μ^*, ν^*) to (2.4.24) is

$$\phi_*(\omega) \;=\; \frac{1}{2}\ln\left(\frac{\mu_\omega^*}{\nu_\omega^*}\right), \tag{2.4.27}$$

and the upper bound $\varepsilon_\star$ on the risks of this detector is given by (2.4.26).

2.4.5.4 K-th power of simple o.s.

Recall that K-th power of a simple o.s. $\mathcal{O} = (\Omega, \Pi; \{p_\mu : \mu \in \mathcal{M}\}; \mathcal{F})$ (see Section 2.4.3.4) is the o.s.

$$[\mathcal{O}]^K = (\Omega^K, \Pi^K; \{p_\mu^{(K)} : \mu \in \mathcal{M}\}; \mathcal{F}^{(K)})$$

where Ω^K is the direct product of K copies of Ω, Π^K is the product of K copies of Π, the densities $p_\mu^{(K)}$ are product densities induced by K copies of density p_μ, $\mu \in \mathcal{M}$:

$$p_\mu^{(K)}(\omega^K = (\omega_1, ..., \omega_K)) = \prod_{k=1}^{K} p_\mu(\omega_k),$$

and $\mathcal{F}^{(K)}$ is comprised of functions

$$\phi^{(K)}(\omega^K = (\omega_1, ..., \omega_K)) = \sum_{k=1}^{K} \phi(\omega_k)$$

stemming from functions $\phi \in \mathcal{F}$. Clearly, $[\mathcal{O}]^K$ is the observation scheme describing the stationary K-repeated observations $\omega^K = (\omega_1, ..., \omega_K)$ with ω_k stemming from the o.s. $\mathcal{O}$, see Section 2.3.2.3. As we remember, $[\mathcal{O}]^K$ is simple provided that $\mathcal{O}$ is so.

Assuming $\mathcal{O}$ simple, it is immediately seen that as applied to the o.s. $[\mathcal{O}]^K$, the recipe from the beginning of Section 2.4.5 reads as follows:

- M_1, M_2 can be arbitrary nonempty convex compact subsets of $\mathcal{M}$, and the corresponding hypotheses, H_χ^K, $\chi = 1, 2$, state that the components ω_k of observation $\omega^K = (\omega_1, ..., \omega_K)$ are independently of each other drawn from distribution p_μ with $\mu \in M_1$ (hypothesis H_1^K) or $\mu \in M_2$ (hypothesis H_2^K).

- problem (2.4.18) is the convex program

$$\mathrm{Opt}(K) = \max_{\mu \in M_1, \nu \in M_2} \underbrace{\ln\left(\int_{\Omega^K} \sqrt{p_\mu^{(K)}(\omega^K) p_\nu^{(K)}(\omega^K)} \Pi^K(d\Omega)\right)}_{\equiv K \ln\left(\int_\Omega \sqrt{p_\mu(\omega) p_\nu(\omega)} \Pi(d\omega)\right)} \qquad (D_K)$$

  implying that any optimal solution to the "single-observation" problem (D_1) associated with M_1, M_2 is optimal for the "K-observation" problem (D_K) associated with M_1, M_2, and $\mathrm{Opt}(K) = K\mathrm{Opt}(1)$;

- the optimal detector $\phi_*^{(K)}$ given by an optimal solution (μ_*, ν_*) to (D_1) (this solution is optimal for (D_K) as well) is

$$\begin{array}{rcl} \phi_*^{(K)}(\omega^K) & = & \sum_{k=1}^K \phi_*(\omega_k), \\ \phi_*(\omega) & = & \frac{1}{2} \ln\left(\frac{p_{\mu_*}(\omega)}{p_{\nu_*}(\omega)}\right), \end{array} \qquad (2.4.28)$$

  and the upper bound $\varepsilon_\star(K)$ on the risks of the detector $\phi_*^{(K)}$ on the pair of families of distributions obeying hypotheses H_1^K, resp., H_2^K, is

$$\varepsilon_\star(K) = \mathrm{e}^{\mathrm{Opt}(K)} = \mathrm{e}^{K\mathrm{Opt}(1)} = [\epsilon_\star(1)]^K. \qquad (2.4.29)$$

The just outlined results on powers of simple observation schemes allow to express near-optimality of detector-based tests in simple o.s.'s in a nicer form, specifically, as follows.

Proposition 2.4.2 *Let $\mathcal{O} = (\Omega, \Pi; \{p_\mu : \mu \in \mathcal{M}\}; \mathcal{F})$ be a simple observation scheme, M_1, M_2 be two nonempty convex compact subsets of $\mathcal{M}$, and let (μ_*, ν_*) be an optimal solution to the convex optimization problem (cf. Theorem 2.4.2)*

$$\mathrm{Opt} = \max_{\mu \in M_1, \nu \in M_2} \ln \left(\int_\Omega \sqrt{p_\mu(\omega) p_\nu(\omega)} \Pi(d\omega) \right).$$

Let ϕ_* and ϕ_*^K be single- and K-observation detectors induced by (μ_*, ν_*) via (2.4.28).

Let $\epsilon \in (0, 1/2)$, and assume that for some positive integer K in the nature exists a simple test $\mathcal{T}^K$ deciding via K i.i.d. observations $\omega^K = (\omega_1, ..., \omega_K)$ with $\omega_k \sim p_\mu$, for some unknown $\mu \in \mathcal{M}$, on the hypotheses

$$H_\chi^{(K)} : \mu \in M_\chi, \ \chi = 1, 2,$$

with risks Risk_1, Risk_2 not exceeding ϵ. Then setting

$$K_+ = \rfloor \frac{2}{1 - \ln(4(1-\epsilon))/\ln(1/\epsilon)} K \lfloor,$$

the simple test $\mathcal{T}_{\phi_*^{(K_+)}}$ utilizing K_+ i.i.d. observations decides on $H_1^{(K_+)}$, $H_2^{(K_+)}$ with risks $\leq \epsilon$. Note that K_+ "is of order of K:" $K_+/K \to 2$ as $\epsilon \to +0$.

Proof. Applying item (iii) of Theorem 2.4.2 to the simple o.s. $[\mathcal{O}]^K$, we see that what above was called $\varepsilon_\star(K)$ satisfies

$$\varepsilon_\star(K) \leq 2\sqrt{\epsilon(1-\epsilon)}.$$

By (2.4.29), we conclude that $\varepsilon_\star(1) \leq \left(2\sqrt{\epsilon(1-\epsilon)}\right)^{1/K}$, whence, by the same (2.4.29), $\varepsilon_\star(T) \leq \left(2\sqrt{\epsilon(1-\epsilon)}\right)^{T/K}$, $T = 1, 2, ...$; plugging in this bound $T = K_+$, we get (check it!) the inequality $\varepsilon_\star(K_+) \leq \epsilon$. It remains to recall that $\varepsilon_\star(K_+)$ upper-bounds the risks of the test $\mathcal{T}_{\phi_*^{(K_+)}}$ when deciding on $H_1^{(K_+)}$ vs. $H_2^{(K_+)}$. □

2.5 Testing multiple hypotheses

So far, we focused on detector-based tests deciding on pairs of hypotheses, and our "constructive" results were restricted to pairs of *convex* hypotheses dealing with a simple o.s.

$$\mathcal{O} = (\Omega, \Pi; \{p_\mu : \mu \in \mathcal{M}\}; \mathcal{F}), \tag{2.5.1}$$

convexity of a hypothesis meaning that the family of probability distributions obeying the hypothesis is $\{p_\mu : \mu \in X\}$ associated with a convex (in fact, convex compact) set $X \subset \mathcal{M}$.

In this Section, we will be interested in pairwise testing *unions* of convex hypotheses and testing *multiple* (more than two) hypotheses.

2.5.1 Testing unions

2.5.1.1 Situation and goal

Let Ω be an observation space, and assume we are given two finite collections of families of probability distributions on Ω: families of *red* distributions $\mathcal{R}_i$, $1 \leq i \leq r$, and families of *blue* distributions $\mathcal{B}_j$, $1 \leq j \leq b$. These families give rise to r red and b blue hypotheses on the distribution P of an observation $\omega \in \Omega$, specifically,

$$R_i : P \in \mathcal{R}_i \text{ (red hypotheses) and } B_j : P \in \mathcal{B}_j \text{ (blue hypotheses)}$$

Assume that for every $i \leq r$, $j \leq B$ we have at our disposal a simple detector-based test $\mathcal{T}_{ij}$ capable to decide on R_i vs B_j; what we want is to assemble these tests into a test $\mathcal{T}$ deciding on the union R of red hypotheses vs. the union B of blue ones:

$$R : P \in \mathcal{R} := \bigcup_{i=1}^{r} \mathcal{R}_i, \;\; B : P \in \mathcal{B} := \bigcup_{j=1}^{b} \mathcal{B}_j,$$

where P, as always, stands for the probability distribution of observation $\omega \in \Omega$.

> Our motivation primarily stems from the case where R_i and B_j are convex hypotheses in a simple o.s. (2.5.1):
>
> $$\mathcal{R}_i = \{p_\mu : \mu \in M_i\}, \; \mathcal{B}_j = \{p_\mu : \mu \in N_j\},$$
>
> where M_i and N_j are convex compact subsets of $\mathcal{M}$. In this case we indeed know how to build near-optimal tests deciding on R_i vs. B_j, and the question we have posed becomes, how to assemble these tests into a test deciding on R vs. B, with
>
> $$\begin{array}{l} R : P \in \mathcal{R} = \{p_\mu : \mu \in X\}, \; X = \bigcup_i M_i, \\ B : P \in \mathcal{B} = \{p_\mu : \mu \in Y\}, \; Y = \bigcup_j N_j; \end{array}$$
>
> while structure of R, B is similar to the one of R_i, B_j, there is a significant difference: the sets X, Y are, in general, non-convex, and therefore the techniques we have developed fail to address testing R vs. B directly.

2.5.1.2 The construction

In the just described situation, let ϕ_{ij} be the detectors underlying the tests $\mathcal{T}_{ij}$; w.l.o.g., we can assume these detectors balanced (see Section 2.3.2.2) with some risks ϵ_{ij}:

$$\left.\begin{array}{l} \int_\Omega \mathrm{e}^{-\phi_{ij}(\omega)} P(d\omega) \leq \epsilon_{ij} \; \forall P \in \mathcal{R}_i \\ \int_\Omega \mathrm{e}^{\phi_{ij}(\omega)} P(d\omega) \leq \epsilon_{ij} \; \forall P \in \mathcal{B}_j \end{array}\right\}, \; 1 \leq i \leq r, 1 \leq j \leq b. \tag{2.5.2}$$

Let us assemble the detectors ϕ_{ij} into a detector for R, B as follows:

$$\phi(\omega) = \max_{1 \leq i \leq r} \min_{1 \leq j \leq b} [\phi_{ij} - \alpha_{ij}], \tag{2.5.3}$$

where the *shifts* α_{ij} are construction's parameters.

Proposition 2.5.1 *The risks of ϕ on R, B can be bounded as*

$$\begin{array}{ll} \forall P \in \mathcal{R} : & \int_\Omega \mathrm{e}^{-\phi(\omega)} P(d\omega) \leq \max_{i \leq r} \left[\sum_{j=1}^b \epsilon_{ij} \mathrm{e}^{\alpha_{ij}}\right] \\ \forall P \in \mathcal{B} : & \int_\Omega \mathrm{e}^{\phi(\omega)} P(d\omega) \leq \max_{j \leq b} \left[\sum_{i=1}^r \epsilon_{ij} \mathrm{e}^{-\alpha_{ij}}\right] \end{array} \tag{2.5.4}$$

Thus, the risks of ϕ on R, B are upper-bounded by the quantity

$$\varepsilon_\star = \max\left[\max_{i \leq r} \left[\sum\nolimits_{j=1}^b \epsilon_{ij} \mathrm{e}^{\alpha_{ij}}\right], \max_{j \leq b} \left[\sum\nolimits_{i=1}^r \epsilon_{ij} \mathrm{e}^{-\alpha_{ij}}\right]\right], \tag{2.5.5}$$

whence the risks of the based on the detector ϕ simple test $\mathcal{T}_\phi$ deciding on R, B are upper-bounded by $\varepsilon_\star$.

Proof. Let $P \in \mathcal{R}$, so that $P \in \mathcal{R}_{i_*}$ for some $i_* \leq r$. Then

$$\begin{array}{l} \int_\Omega \mathrm{e}^{-\phi(\omega)} P(d\omega) = \int_\Omega \mathrm{e}^{\min_{i \leq r} \max_{j \leq b} [-\phi_{ij}(\omega) + \alpha_{ij}]} P(d\omega) \\ \leq \int_\Omega \mathrm{e}^{\max_{j \leq b} [-\phi_{i_*j}(\omega) + \alpha_{i_*j}]} P(d\omega) \leq \sum_{j=1}^b \int_\Omega \mathrm{e}^{-\phi_{i_*j}(\omega) + \alpha_{i_*j}} P(d\omega) = \sum_{j=1}^b \exp^{\alpha_{i_*j}} \int_\Omega \mathrm{e}^{-\phi_{i_*j}(\omega)} P(d\omega) \\ \leq \sum_{j=1}^b \epsilon_{i_*j} \mathrm{e}^{\alpha_{i_*j}} \; [\text{by (2.5.2) due to } P \in \mathcal{R}_{i_*}] \\ \leq \max_{i \leq r} \left[\sum_{j=1}^b \epsilon_{ij} \mathrm{e}^{\alpha_{ij}}\right] \end{array}$$

Now let $P\in\mathcal{B}$, so that $P\in\mathcal{B}_{j_*}$ for some j_*. We have

$$\begin{array}{l}
\int_\Omega \mathrm{e}^{\phi(\omega)}P(d\omega)=\int_\Omega \mathrm{e}^{\max_{i\leq r}\min_{j\leq b}[\phi_{ij}(\omega)-\alpha_{ij}]}P(d\omega)\\
\leq \int_\Omega \mathrm{e}^{\max_{i\leq r}[\phi_{ij*}(\omega)-\alpha_{ij*}]}P(d\omega)\leq \sum_{i=1}^r\int_\Omega \mathrm{e}^{\phi_{ij*}(\omega)-\alpha_{ij*}}P(d\omega)=\sum_{i=1}^r\exp^{-\alpha_{ij*}}\int_\Omega \mathrm{e}^{\phi_{ij*}(\omega)}P(d\omega)\\
\leq \sum_{i=1}^r\epsilon_{ij_*}\mathrm{e}^{\alpha_{ij*}}\ [\text{by (2.5.2) due to } P\in\mathcal{B}_{j*}]\\
\leq \max_{j\leq b}\left[\sum_{i=1}^r\epsilon_{ij}\mathrm{e}^{-\alpha_{ij}}\right]
\end{array}$$

(2.5.4) is proved. The remaining claims in Proposition are readily given by (2.5.4) combined with Proposition 2.3.1. □

Optimal choice of shift parameters. The detector and the test considered in Proposition 2.5.1, same as the resulting risk bound $\varepsilon_\star$, depend on the shifts α_{ij}. We are about to optimize the risk bound w.r.t. these shifts. To this end, consider the $r\times b$ matrix

$$E=[\epsilon_{ij}]_{\substack{i\leq r\\ j\leq b}}$$

and the symmetric $(r+b)\times(r+b)$ matrix

$$\mathcal{E}=\left[\begin{array}{c|c} & E\\ \hline E^T & \end{array}\right]$$

As it is well known, the eigenvalues of the symmetric matrix $\mathcal{E}$ are comprised of the pairs $(\sigma_s,-\sigma_s)$, where σ_s are the singular values of E, and several zeros; in particular, the leading eigenvalue of $\mathcal{E}$ is the spectral norm $\|E\|_{2,2}$ (the largest singular value) of matrix E. Further, E is a matrix with positive entries, so that $\mathcal{E}$ is a symmetric entrywise nonnegative matrix. By Perron-Frobenius Theorem, the leading eigenvector of this matrix can be selected to be nonnegative. Denoting this nonnegative eigenvector $[g;h]$ with r-dimensional g and b-dimensional h, and setting $\rho=\|E\|_{2,2}$, we have

$$\begin{array}{rcl}\rho g&=&Eh\\ \rho h&=&E^Tg\end{array}\tag{2.5.6}$$

Observe that $\rho>0$ (evident), whence both g and h are nonzero (since otherwise (2.5.6) would imply $g=h=0$, which is impossible – the eigenvector $[g;h]$ is nonzero). Since h and g are nonzero nonnegative vectors, $\rho>0$ and E is entrywise positive, (2.5.6) says that g and h are strictly positive vectors. The latter allows to define shifts α_{ij} according to

$$\alpha_{ij}=\ln(h_j/g_i).\tag{2.5.7}$$

With these shifts, we get

$$\max_{i\leq r}\left[\sum_{j=1}^b\epsilon_{ij}\mathrm{e}^{\alpha_{ij}}\right]=\max_{i\leq r}\sum_{j=1}^b\epsilon_{ij}h_j/g_i=\max_{i\leq r}(Eh)_i/g_i=\max_{i\leq r}\rho=\rho$$

(we have used the first relation in (2.5.6)) and

$$\max_{j\leq b}\left[\sum_{i=1}^r\epsilon_{ij}\mathrm{e}^{-\alpha_{ij}}\right]=\max_{j\leq b}\sum_{i=1}^r\epsilon_{ij}g_i/h_j=\max_{j\leq b}[E^Tg]_j/h_j=\max_{j\leq b}\rho=\rho$$

(we have used the second relation in (2.5.6)). The bottom line is as follows:

Proposition 2.5.2 *In the situation and the notation from Section 2.5.1.1, the risks of the detector* (2.5.3) *with shifts* (2.5.6), (2.5.7) *on the families $\mathcal{R}$, $\mathcal{B}$ do not exceed the quantity*

$$\|E:=[\epsilon_{ij}]_{i\leq r,j\leq b}\|_{2,2}.$$

As a result, the risks of the simple test $\mathcal{T}_\phi$ deciding on the hypotheses R, B, does not exceed $\|E\|_{2,2}$ as well.

In fact, the shifts in the above proposition are the best possible; this is an immediate consequence of the following simple fact:

Proposition 2.5.3 *Let $\mathcal{E} = [e_{ij}]$ be nonzero entrywise nonnegative $n \times n$ symmetric matrix. Then the optimal value in the optimization problem*

$$\mathrm{Opt} = \min_{\alpha_{ij}} \left\{ \max_{i \leq n} \sum_{j=1}^{n} e_{ij} \mathrm{e}^{\alpha_{ij}} : \alpha_{ij} = -\alpha_{ji} \right\} \tag{$*$}$$

is equal to $\|\mathcal{E}\|_{2,2}$. When the Perron-Frobenius eigenvector f of $\mathcal{E}$ can be selected positive, the problem is solvable, and an optimal solution is given by

$$\alpha_{ij} = \ln(f_j/f_i),\ 1 \leq i,j \leq n. \tag{2.5.8}$$

Proof. Let us prove, first, that $\mathrm{Opt} \leq \rho := \|\mathcal{E}\|_{2,2}$. Given $\epsilon > 0$, we clearly can find an entrywise nonnegative symmetric matrix $\mathcal{E}'$ with entries e'_{ij} in-between e_{ij} and $e_{ij} + \epsilon$ such that the Perron-Frobenius eigenvector f of $\mathcal{E}'$ can be selected positive (it suffices, e.g., to set $e'_{ij} = e_{ij} + \epsilon$). Selecting α_{ij} according to (2.5.8), we get a feasible solution to $(*)$ such that

$$\forall i : \sum_j e_{ij} \mathrm{e}^{\alpha_{ij}} \leq \sum_j e'_{ij} f_j / f_i = \|\mathcal{E}'\|_{2,2},$$

implying that $\mathrm{Opt} \leq \|\mathcal{E}'\|_{2,2}$. Passing to limit as $\epsilon \to +0$, we get $\mathrm{Opt} \leq \|\mathcal{E}\|_{2,2}$. As a byproduct of our reasoning, we see that if $\mathcal{E}$ admits a positive Perron-Frobenius eigenvector f, then (2.5.8) yields a feasible solution to $(*)$ with the value of the objective equal to $\|\mathcal{E}\|_{2,2}$.

It remain to prove that $\mathrm{Opt} \geq \|\mathcal{E}\|_{2,2}$. Assume that this is not the case, so that $(*)$ admits a feasible solution $\widehat{\alpha}_{ij}$ such that

$$\widehat{\rho} := \max_i \sum_j e_{ij} \mathrm{e}^{\widehat{\alpha}_{ij}} < \rho := \|\mathcal{E}\|_{2,2}.$$

Perturbing $\mathcal{E}$ a little bit, we can make this matrix symmetric and entrywise positive, and still satisfying the above strict inequality; to save notation, assume that already the original $\mathcal{E}$ is entrywise positive. Let f be a positive Perron-Frobenius eigenvector of $\mathcal{E}$, and let, as above, $\alpha_{ij} = \ln(f_j/f_i)$, so that

$$\sum_j e_{ij} \mathrm{e}^{\alpha_{ij}} = \sum_j e_{ij} f_j / f_i = \rho\ \forall i.$$

Setting $\delta_{ij} = \widehat{\alpha}_{ij} - \alpha_{ij}$, we conclude that the convex functions

$$\theta_i(t) = \sum_j e_{ij} \mathrm{e}^{\alpha_{ij} + t\delta_{ij}}$$

all are equal to ρ as $t = 0$, and all are $\leq \widehat{\rho} < \rho$ as $t = 1$, implying that $\theta_i(1) < \theta_i(0)$ for every i. The latter, in view of convexity of $\theta_i(\cdot)$, implies that

$$\theta'_i(0) = \sum_j e_{ij} \mathrm{e}^{\alpha_{ij}} \delta_{ij} = \sum_j e_{ij} (f_j/f_i) \delta_{ij} < 0\ \forall i.$$

Multiplying the resulting inequalities by f_i^2 and summing up over i, we get

$$\sum_{i,j} e_{ij} f_i f_j \delta_{ij} < 0,$$

which is impossible: we have $e_{ij} = e_{ji}$ and $\delta_{ij} = -\delta_{ji}$, implying that the left hand side in the latter inequality is 0. □

2.5.2 Testing multiple hypotheses "up to closeness"

So far, we have considered detector-based simple tests deciding on pairs of hypotheses, specifically, convex hypotheses in simple o.s.'s (Section 2.4.4) and unions of convex hypotheses (Section 2.5.1)[12]. Now we intend to consider testing of multiple (perhaps more than 2) hypotheses "up to closeness;" the latter notion was introduced in Section 2.2.4.2.

2.5.2.1 Situation and goal

Let Ω be an observation space, and let a collection $\mathcal{P}_1, ..., \mathcal{P}_L$ of families of probability distributions on Ω be given. As always, families $\mathcal{P}_\ell$ give rise to hypotheses

$$H_\ell : P \in \mathcal{P}_\ell$$

on the distribution P of observation $\omega \in \Omega$. Assume also that we are given a *closeness relation* $\mathcal{C}$ on $\{1, ..., L\}$; recall that a closeness relation, formally, is some set of pairs of indexes $(\ell, \ell') \in \{1, ..., L\}$; we interpret the inclusion $(\ell, \ell') \in \mathcal{C}$ as the fact that hypothesis H_ℓ "is close" to hypothesis $H_{\ell'}$. When $(\ell, \ell') \in \mathcal{C}$, we say that ℓ' is close (or $\mathcal{C}$-close) to ℓ. We always assume that

- $\mathcal{C}$ contains the diagonal: $(\ell, \ell) \in \mathcal{C}$ for every $\ell \leq L$ ("each hypothesis is close to itself"), and
- $\mathcal{C}$ is symmetric: whenever $(\ell, \ell') \in \mathcal{C}$, we have also $(\ell', \ell) \in \mathcal{C}$ ("if ℓ-th hypothesis is close to ℓ'-th one, then ℓ'-th hypothesis is close to ℓ-th one").

Recall that a test $\mathcal{T}$ deciding on the hypotheses $H_1, ..., H_L$ via observation $\omega \in \Omega$ is a procedure which, given on input $\omega \in \Omega$, builds some set $\mathcal{T}(\omega) \subset \{1, ..., L\}$, accepts all hypotheses H_ℓ with $\ell \in \mathcal{T}(\omega)$, and rejects all other hypotheses.

Risks of an "up to closeness" test. The notion of $\mathcal{C}$-risk of a test was introduced in Section 2.2.4.2; we reproduce it here for reader's convenience. Given closeness $\mathcal{C}$ and a test $\mathcal{T}$, we define the *$\mathcal{C}$-risk*

$$\mathrm{Risk}^{\mathcal{C}}(\mathcal{T}|H_1, ..., H_L)$$

of $\mathcal{T}$ as the smallest $\epsilon \geq 0$ such that

> *Whenever an observation ω is drawn from a distribution $P \in \bigcup_\ell \mathcal{P}_\ell$, and ℓ_* is such that $P \in \mathcal{P}_{\ell_*}$ (i.e., hypothesis H_{ℓ_*} is true), the P-probability of the event "$\ell_* \notin \mathcal{T}(\omega)$ ("true hypothesis H_{ℓ_*} is not accepted") $\underline{\text{or}}$ there exists ℓ' $\underline{\text{not close to } \ell}$ such that $H_{\ell'}$ is accepted" is $\underline{\text{at most}}$ ϵ.*

Equivalently:

> $\mathrm{Risk}^{\mathcal{C}}(\mathcal{T}|H_1, ..., H_L) \leq \epsilon$ if and only if the following takes place:
>
> *Whenever an observation ω is drawn from a distribution $P \in \bigcup_\ell \mathcal{P}_\ell$, and ℓ_* is such that $P \in \mathcal{P}_{\ell_*}$ (i.e., hypothesis H_{ℓ_*} is true), the P-probability of the event*
>
> > *$\ell_* \in \mathcal{T}(\omega)$ ("the true hypothesis H_{ℓ_*} is accepted") $\underline{\text{and}}$ $\ell' \in \mathcal{T}(\omega)$ implies that $(\ell, \ell') \in \mathcal{C}$ ("all accepted hypotheses are $\mathcal{C}$-close to the true hypothesis H_{ℓ_*}") is $\underline{\text{at least}}$ $1 - \epsilon$.*

For example, consider 11 colored polygons presented on Figure 2.2 and associate with them 11 hypotheses on 2D "signal plus noise" observation $\omega = x + \xi$, $\xi \sim \mathcal{N}(0, I_2)$, with ℓ-th hypothesis

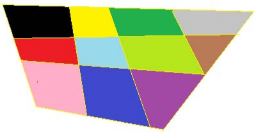


Figure 2.2: 11 hypotheses on the location of the mean μ of observation $\omega \sim \mathcal{N}(\mu, I_2)$, each stating that μ belongs to the polygon of specific color.

stating that x belongs to ℓ-th polygon. When defining closeness $\mathcal{C}$ on the collection of 11 hypotheses presented on Figure 2.2 as

"*two hypotheses are close if and only if the corresponding color polygons intersect*"

the fact that a test $\mathcal{T}$ has $\mathcal{C}$-risk ≤ 0.01 implies, in particular, that if the probability distribution P underlying the observed ω "is black," (i.e., the mean of ω belongs to the black polygon), then with P-probability at least 0.99 the list of accepted hypotheses will include the black one, and the only other hypotheses in this list will be among the red, yellow and light-blue ones.

2.5.2.2 "Building blocks" and construction

The construction we are about to present is, essentially, the one used in Section 2.2.4.3 as applied to detector-generated tests; this being said, the presentation to follow is self-contained.

Building blocks for our construction are pairwise detectors $\phi_{\ell\ell'}(\omega)$, $1 \leq \ell \leq \ell' \leq L$, for pairs $\mathcal{P}_\ell$, $\mathcal{P}_{\ell'}$ along with (upper bounds on) the risks $\epsilon_{\ell\ell'}$ of these detectors:

$$\left.\begin{array}{ll}\forall (P \in \mathcal{P}_\ell): & \int_\Omega \mathrm{e}^{-\phi_{\ell\ell'}(\omega)} P(d\omega) \leq \epsilon_{\ell\ell'} \\ \forall (P \in \mathcal{P}_{\ell'}): & \int_\Omega \mathrm{e}^{\phi_{\ell\ell'}(\omega)} P(d\omega) \leq \epsilon_{\ell\ell'}\end{array}\right\},\ 1 \leq \ell < \ell' \leq L.$$

Setting

$$\phi_{\ell'\ell}(\omega) = -\phi_{\ell\ell'}(\omega),\ \epsilon_{\ell'\ell} = \epsilon_{\ell\ell'},\ 1 \leq \ell < \ell' \leq L,\ \phi_{\ell\ell}(\omega) \equiv 0, \epsilon_{\ell\ell} = 1,\ 1 \leq \ell \leq L,$$

we get what we shall call *balanced system of detectors $\phi_{\ell\ell'}$ and risks $\epsilon_{\ell\ell'}$, $1 \leq \ell, \ell' \leq L$, for the collection $\mathcal{P}_1, ..., \mathcal{P}_L$*, meaning that

$$\begin{array}{ll}(a): & \phi_{\ell\ell'}(\omega) + \phi_{\ell'\ell}(\omega) \equiv 0,\ \epsilon_{\ell\ell'} = \epsilon_{\ell'\ell},\ 1 \leq \ell, \ell' \leq L \\ (b): & \forall P \in \mathcal{P}_\ell : \int_\Omega \mathrm{e}^{-\phi_{\ell\ell'}(\omega)} P(d\omega) \leq \epsilon_{\ell\ell'},\ 1 \leq \ell, \ell' \leq L.\end{array} \tag{2.5.9}$$

Given closeness $\mathcal{C}$, we associate with it the symmetric $L \times L$ matrix $\mathbf{C}$ given by

$$\mathbf{C}_{\ell\ell'} = \left\{\begin{array}{ll} 0, & (\ell, \ell') \in \mathcal{C} \\ 1, & (\ell, \ell') \notin \mathcal{C}\end{array}\right. \tag{2.5.10}$$

Test $\mathcal{T}_\mathcal{C}$. Let a collection of shifts $\alpha_{\ell\ell'} \in \mathbf{R}$ satisfying the relation

$$\alpha_{\ell\ell'} = -\alpha_{\ell'\ell},\ 1 \leq \ell, \ell' \leq L \tag{2.5.11}$$

be given. The detectors $\phi_{\ell\ell'}$ and the shifts $\alpha_{\ell\ell'}$ specify a test $\mathcal{T}_\mathcal{C}$ deciding on hypotheses $H_1, ..., H_L$; specifically, given an observation ω, the test $\mathcal{T}_\mathcal{C}$ accepts exactly those hypotheses H_ℓ for which $\phi_{\ell\ell'}(\omega) - \alpha_{\ell\ell'} > 0$ whenever ℓ' is *not* $\mathcal{C}$-close to ℓ:

$$\mathcal{T}_\mathcal{C}(\omega) = \{\ell : \phi_{\ell\ell'}(\omega) - \alpha_{\ell\ell'} > 0\ \forall (\ell' : (\ell, \ell') \notin \mathcal{C})\}. \tag{2.5.12}$$

[12] strictly speaking, in Section 2.5.1 it was not explicitly stated that the unions under consideration involve convex hypotheses in simple o.s.'s; our emphasis was on how to decide on a pair of union-type hypotheses *given pairwise detectors for "red" and "blue" components of the unions from the pair.* Note, however, that as of now, the only situation where we indeed have at our disposal good pairwise detectors for red and blue components is the one where these components are convex hypotheses in a good o.s.

Proposition 2.5.4 (i) *The $\mathcal{C}$-risk of the just defined test $\mathcal{T}_\mathcal{C}$ is upper-bounded by the quantity*

$$\varepsilon[\alpha]=\max_{\ell\leq L}\sum_{\ell'=1}^L \epsilon_{\ell\ell'}\mathbf{C}_{\ell\ell'}\mathrm{e}^{\alpha_{\ell\ell'}}$$

with $\mathbf{C}$ *given by* (2.5.10).

(ii) *The infimum, over shifts α satisfying* (2.5.11)*, of the risk bound $\varepsilon[\alpha]$ is the quantity*

$$\varepsilon_\star=\|\mathcal{E}\|_{2,2},$$

where the $L\times L$ symmetric entrywise nonnegative matrix $\mathcal{E}$ is given by

$$\mathcal{E}=[e_{\ell\ell'}:=\epsilon_{\ell\ell'}\mathbf{C}_{\ell\ell'}]_{\ell,\ell'\leq L}\,.$$

Assuming $\mathcal{E}$ admits a strictly positive Perron-Frobenius vector f, an optimal choice of the shifts is

$$\alpha_{\ell\ell'}=\ln(f_{\ell'}/f_\ell),1\leq\ell,\ell'\leq L,$$

resulting in $\varepsilon[\alpha]=\varepsilon_\star=\|\mathcal{E}\|_{2,2}$.

Proof. (i): Setting

$$\bar\phi_{\ell\ell'}(\omega)=\phi_{\ell\ell'}(\omega)-\alpha_{\ell\ell'},\ \bar\epsilon_{\ell\ell'}=\epsilon_{\ell\ell'}\mathrm{e}^{\alpha_{\ell\ell'}},$$

(2.5.9), (2.5.11) imply that

$$\begin{array}{ll}(a): & \bar\phi_{\ell\ell'}(\omega)+\bar\phi_{\ell'\ell}(\omega)\equiv0,1\leq\ell,\ell'\leq L\\ (b): & \forall P\in\mathcal{P}_\ell:\int_\Omega \mathrm{e}^{-\bar\phi_{\ell\ell'}(\omega)}P(d\omega)\leq\bar\epsilon_{\ell\ell'},\ 1\leq\ell,\ell'\leq L.\end{array}\tag{2.5.13}$$

Now let ℓ_* be such that the distribution P of observation ω belongs to $\mathcal{P}_{\ell_*}$. Then the P-probability of the event $\bar\phi_{\ell_*\ell'}(\omega)\leq0$ is, for every ℓ', $\leq\bar\epsilon_{\ell_*\ell'}$ by (2.5.13.b), whence the P-probability of the event

$$E_*=\{\omega:\exists\ell':(\ell_*,\ell')\not\in\mathcal{C}\ \&\ \bar\phi_{\ell_*\ell'}(\omega)\leq0\}$$

is upper-bounded by

$$\sum_{\ell':(\ell_*,\ell')\not\in\mathcal{C}}\bar\epsilon_{\ell_*\ell'}=\sum_{\ell'=1}^L\mathbf{C}_{\ell_*\ell'}\epsilon_{\ell_*\ell'}\mathrm{e}^{\alpha_{\ell_*\ell'}}\leq\varepsilon[\alpha].$$

Assume that E_* does not take place (as we have seen, this indeed is so with P-probability $\geq1-\varepsilon[\alpha]$). Then $\bar\phi_{\ell_*\ell'}(\omega)>0$ for all ℓ' such that $(\ell_*,\ell')\not\in\mathcal{C}$, implying, first, that H_{ℓ_*} is accepted by our test. Second, $\bar\phi_{\ell'\ell_*}(\omega)=-\bar\phi_{\ell_*\ell'}(\omega)<0$ whenever $(\ell_*,\ell')\not\in\mathcal{C}$, or, which is the same due to the symmetry of closeness, whenever $(\ell',\ell_*)\not\in\mathcal{C}$, implying that the test $\mathcal{T}_\mathcal{C}$ rejects the hypothesis $H_{\ell'}$ when ℓ' is not $\mathcal{C}$-close to ℓ_*. Thus, the P-probability of the event "H_{ℓ_*} is accepted, and all accepted hypotheses are $\mathcal{C}$-close to H_{ℓ_*}" is at least $1-\varepsilon[\alpha]$. We conclude that the $\mathcal{C}$-risk $\mathrm{Risk}^\mathcal{C}(\mathcal{T}_\mathcal{C}|H_1,...,H_L)$ of the test $\mathcal{T}_\mathcal{C}$ is at most $\varepsilon[\alpha]$. (i) is proved. (ii) is readily given by Proposition 2.5.3. □

2.5.2.3 Testing multiple hypotheses via repeated observations

In the situation of Section 2.5.2.1, given a balanced system of detectors $\phi_{\ell\ell'}$ and risks $\epsilon_{\ell\ell'}$, $1\leq\ell,\ell'\leq L$ for the collection $\mathcal{P}_1,...,\mathcal{P}_L$ (see (2.5.9)) *and* a positive integer K, we can

- pass from detectors $\phi_{\ell\ell'}$ and risks $\epsilon_{\ell\ell'}$ to the entities

$$\phi^{(K)}_{\ell\ell'}(\omega^K=(\omega_1,...,\omega_K))=\sum_{k=1}^K\phi_{\ell\ell'}(\omega_k),\ \epsilon^{(K)}_{\ell\ell'}=\epsilon^K_{\ell\ell'},\ 1\leq\ell,\ell'\leq L$$

- associate with the families $\mathcal{P}_\ell$ families $\mathcal{P}_\ell^{(K)}$ of probability distributions underlying quasi-stationary K-repeated versions of observations $\omega \sim P \in \mathcal{P}_\ell$, see Section 2.3.2.3, and thus arrive at hypotheses $H_\ell^K = \mathcal{H}_\ell^{\otimes,K}$ stating that the distribution P^K of K-repeated observation $\omega^K = (\omega_1, ..., \omega_K)$, $\omega_k \in \Omega$, belongs to the family $\mathcal{P}_\ell^{\otimes,K} = \bigotimes\limits_{k=1}^{K} \mathcal{P}_\ell$, see Section 2.1.3.3, associated with $\mathcal{P}_\ell$.

Invoking Proposition 2.3.3 and (2.5.9), we arrive at the following analogy of (2.5.9):

$$\begin{array}{ll} (a): & \phi_{\ell\ell'}^{(K)}(\omega^K) + \phi_{\ell'\ell}^{(K)}(\omega^K) \equiv 0,\ \epsilon_{\ell\ell'}^{(K)} = \epsilon_{\ell'\ell}^{(K)} = \epsilon_{\ell\ell'}^K,\ 1 \leq \ell, \ell' \leq L \\ (b): & \forall P^K \in \mathcal{P}_\ell^{(K)} : \int_{\Omega^K} \mathrm{e}^{-\phi_{\ell\ell'}^{(K)}(\omega^K)} P^K(d\omega^K) \leq \epsilon_{\ell\ell'}^{(K)},\ 1 \leq \ell, \ell' \leq L. \end{array} \tag{2.5.14}$$

Given shifts $\alpha_{\ell\ell'}$ satisfying (2.5.11) and applying the construction from Section 2.5.2.2 to these shifts and our new detectors and risks, we arrive at the test test $\mathcal{T}_\mathcal{C}^K$ deciding on hypotheses $H_1^K, ..., H_L^K$ via K-repeated observation ω^K; specifically, given an observation ω^K, the test $\mathcal{T}_\mathcal{C}^K$ accepts exactly those hypotheses H_ℓ^K for which $\phi_{\ell\ell'}^{(K)}(\omega^K) - \alpha_{\ell\ell'} > 0$ whenever ℓ' is *not* $\mathcal{C}$-close to ℓ:

$$\mathcal{T}_\mathcal{C}^K(\omega^K) = \{\ell : \phi_{\ell\ell'}^{(K)}(\omega^K) - \alpha_{\ell\ell'} > 0\ \forall(\ell' : (\ell, \ell') \notin \mathcal{C})\}, \tag{2.5.15}$$

Invoking Proposition 2.5.4, we arrive at

Proposition 2.5.5 (i) *The $\mathcal{C}$-risk of the just defined test $\mathcal{T}_\mathcal{C}^K$ is upper-bounded by the quantity*

$$\varepsilon[\alpha, K] = \max_{\ell \leq L} \sum_{\ell'=1}^{L} \epsilon_{\ell\ell'}^K \mathbf{C}_{\ell\ell'} \mathrm{e}^{\alpha_{\ell\ell'}}.$$

(ii) *The infimum, over shifts α satisfying* (2.5.11), *of the risk bound $\varepsilon[\alpha, K]$ is the quantity*

$$\varepsilon_\star(K) = \|\mathcal{E}^{(K)}\|_{2,2},$$

where the $L \times L$ symmetric entrywise nonnegative matrix $\mathcal{E}^{(K)}$ is given by

$$\mathcal{E}^{(K)} = \left[e_{\ell\ell'}^{(K)} := \epsilon_{\ell\ell'}^K \mathbf{C}_{\ell\ell'}\right]_{\ell,\ell' \leq L}.$$

Assuming $\mathcal{E}^{(K)}$ admits a strictly positive Perron-Frobenius vector f, an optimal choice of the shifts is

$$\alpha_{\ell\ell'} = \ln(f_\ell/f_{\ell'}), 1 \leq \ell, \ell' \leq L,$$

resulting in $\varepsilon[\alpha, K] = \varepsilon_\star(K) = \|\mathcal{E}^{(K)}\|_{2,2}$.

2.5.2.4 Consistency and near-optimality

Observe that when the closeness $\mathcal{C}$ is such that $\epsilon_{\ell\ell'} < 1$ whenever ℓ, ℓ' are *not* $\mathcal{C}$-close to each other, the entries on the matrix $\mathcal{E}^{(K)}$ exponentially fast go to 0 as $K \to \infty$, whence the $\mathcal{C}$-risk of test $\mathcal{T}_\mathcal{C}^K$ also goes to 0 as $K \to \infty$; this is called *consistency*. When, in addition, $\mathcal{P}_\ell$ correspond to convex hypotheses in a simple o.s., the test $\mathcal{T}_\mathcal{C}^K$ possesses certain near-optimality properties similar to those stated in Proposition 2.4.2

Proposition 2.5.6 *Consider the special case of the situation from Section 2.5.2.1 where, given a simple o.s. $\mathcal{O} = (\Omega, \Pi; \{p_\mu : \mu \in \mathcal{M}\}; \mathcal{F})$, the families $\mathcal{P}_\ell$ of probability distributions are of the form $\mathcal{P}_\ell = \{p_\mu : \mu \in N_\ell\}$, where N_ℓ, $1 \leq \ell \leq L$, are nonempty convex compact subsets of $\mathcal{M}$. Let also* *the pairwise detectors $\phi_{\ell\ell'}$ and their risks $\epsilon_{\ell\ell'}$ underlying the construction from Section 2.5.2.2 be obtained by applying Theorem 2.4.2 to the pairs $N_\ell, N_{\ell'}$, so that for $1 \leq \ell < \ell' \leq L$ one has*

$$\phi_{\ell\ell'}(\omega) = \frac{1}{2}\ln(p_{\mu_{\ell,\ell'}}(\omega)/p_{\nu_{\ell,\ell'}}(\omega)),\ \epsilon_{\ell\ell'} = \exp\{\mathrm{Opt}_{\ell\ell'}\},$$

where

$$\mathrm{Opt}_{\ell\ell'} = \min_{\mu\in N_\ell,\nu\in N_{\ell'}} \ln\left(\int_\Omega \sqrt{p_\mu(\omega)p_\nu(\omega)}\Pi(d\omega)\right)$$

and $(\mu_{\ell\ell'}, \nu_{\ell\ell'})$ form an optimal solution to the right hand side optimization problem.

Assume that for some positive integer K_* in the nature there exists a test $\mathcal{T}^{K_*}$ capable to decide with $\mathcal{C}$-risk $\epsilon \in (0,1/2)$, via stationary K_*-repeated observation ω^{K_*}, on the hypotheses $H_\ell^{(K_*)}$, stating that the components in ω^{K_*} are drawn, independently of each other, from a distribution $P \in \mathcal{P}_\ell$, $\ell = 1,...,L$, and let

$$K = \rfloor 2\frac{1+\ln(L-1)/\ln(1/\epsilon)}{1-\ln(4(1-\epsilon))/\ln(1/\epsilon)}K_*\lfloor. \tag{2.5.16}$$

Then the test $\mathcal{T}_\mathcal{C}^K$ yielded by the construction from Section 2.5.2.2 as applied to the above $\phi_{\ell\ell'}$, $\epsilon_{\ell\ell'}$ and trivial shifts $\alpha_{\ell\ell'} \equiv 0$ decides on the hypotheses H_ℓ^K, see Section 2.5.2.3, via quasi-stationary K-repeated observations ω^K, with $\mathcal{C}$-risk $\leq \epsilon$.

Note that $K/K_* \to 2$ as $\epsilon \to +0$.

Proof. Let

$$\bar\epsilon = \max_{\ell,\ell'}\left\{\epsilon_{\ell\ell'} : \ell < \ell' \text{ and } \ell, \ell' \text{ are not } \mathcal{C}\text{-close to each other}\right\}.$$

Denoting by (ℓ_*, ℓ'_*) the maximizer in the right hand side maximization, note that $\mathcal{T}^{K_*}$ induces a simple test $\mathcal{T}$ capable to decide via stationary K_*-repeated observations ω^K on the pair of hypotheses $H_{\ell_*}^{(K_*)}$, $H_{\ell'_*}^{(K_*)}$ with risks $\leq \epsilon$ (it suffices to make $\mathcal{T}$ to accept the first of the hypotheses in the pair and reject the second one whenever $\mathcal{T}^{K_*}$ on the same observation accepts $H_{\ell_*}^{(K_*)}$, otherwise $\mathcal{T}$ rejects the first hypothesis in the pair and accepts the second one). This observation, by the same argument as in the proof of Proposition 2.4.2, implies that $\bar\epsilon^{K_*} \leq 2\sqrt{\epsilon(1-\epsilon)} < 1$, whence all entries in the matrix $\mathcal{E}^{(K)}$ do not exceed $\bar\epsilon^{(K/K_*)}$, implying by Proposition 2.5.4 that the $\mathcal{C}$-risk of the test $\mathcal{T}_\mathcal{C}^K$ does not exceed

$$\epsilon(K) := (L-1)[2\sqrt{\epsilon(1-\epsilon)}]^{K/K_*}.$$

It remains to note that for K given by (2.5.16) one has $\epsilon(K) \leq \epsilon$. □

Remark 2.5.1 *Note that the tests $\mathcal{T}_\mathcal{C}$ and $\mathcal{T}_\mathcal{C}^K$ we have built, may, depending on observations, accept no hypotheses at all, which sometimes is undesirable. Clearly, every test deciding on multiple hypotheses up to $\mathcal{C}$-closeness always can be modified to ensure that a hypothesis always is accepted; to this end, it suffices to accept exactly those hypotheses, if any, which are accepted by our original test, and accept, say, hypothesis # 1 when the original test accepts no hypotheses. It is immediate to see that the $\mathcal{C}$-risk of the modified test cannot be larger than the one of the original test.*

2.5.3 Illustration: Selecting the best among a family of estimates

Let us illustrate our machinery for multiple hypothesis testing by applying it to the situation as follows:

We are given:

- a simple nondegenerate observation scheme $\mathcal{O} = (\Omega, \Pi; \{p_\mu(\cdot) : \mu \in \mathcal{M}\}; \mathcal{F})$,
- a seminorm $\|\cdot\|$ on $\mathbf{R}^n$, [13]
- a convex compact set $X \subset \mathbf{R}^n$ along with a collection of M points $x_i \in \mathbf{R}^n$, $1 \leq i \leq M$ and a positive D such that the $\|\cdot\|$-diameter of the set $X^+ = X \cup \{x_i : 1 \leq i \leq M\}$ is at most D:
$$\|x - x'\| \leq D \; \forall (x, x' \in X^+),$$
- an affine mapping $x \mapsto A(x)$ from $\mathbf{R}^n$ into the embedding space of $\mathcal{M}$ such that $A(x) \in \mathcal{M}$ for all $x \in \mathcal{M}$,
- a tolerance $\epsilon \in (0,1)$.

We observe K-element sample $\omega^K = (\omega_1, ..., \omega_K)$ of independent across k observations
$$\omega_k \sim p_{A(x_*)}, \; 1 \leq k \leq K, \tag{2.5.17}$$
where $x_* \in \mathbf{R}^n$ is unknown signal known to belong to X. Our "ideal goal" is to use ω^K in order to identify, with probability $\geq 1 - \epsilon$, the $\|\cdot\|$-closest to x_* point among the points $x_1, ..., x_M$.

This just outlined goal often is too ambitious, and in the sequel we focus on the relaxed goal as follows:

Given a positive integer N and a "resolution" $\theta > 1$, consider the grid
$$\Gamma = \{r_j = D\theta^{-j}, 0 \leq j \leq N\}$$
and let
$$\rho(x) = \min \left\{ \rho_j \in \Gamma : \rho_j \geq \min_{1 \leq i \leq M} \|x - x_i\| \right\}.$$
Given design parameters $\alpha \geq 1, \beta \geq 0$, we want to specify volume of observations K and an inference routine $\omega^K \mapsto i_{\alpha,\beta}(\omega^K) \in \{1, ..., M\}$ such that
$$\forall (x_* \in X) : \mathrm{Prob}\{\|x_* - x_{i_{\alpha,\beta}(\omega^K)}\| > \alpha\rho(x_*) + \beta\} \geq 1 - \epsilon. \tag{2.5.18}$$

Note that when passing from the "ideal" to the relaxed goal, the simplification is twofold: first, we do not care about the precise distance $\min_i \|x_* - x_i\|$ from x_* to $\{x_1, ..., x_M\}$, all we look at is the best upper bound $\rho(x_*)$ on this distance from the grid Γ; second, we allow factor α and additive term β in mimicking the (discretized) distance $\rho(x_*)$ by $\|x_* - x_{i_{\alpha,\beta}(\omega^K)}\|$.

The problem we have posed is rather popular in Statistics; its origin usually looks as follows: x_i are candidate estimates of x_* yielded by a number of a priori "models" of x_* and perhaps some preliminary noisy observations of x_*. Given x_i and a matrix B, we want to select among the vectors Bx_i the (nearly) best, w.r.t. a given norm $\|\cdot\|_o$, approximation of Bx_*, utilizing additional observations ω^K of the signal. To bring this problem into our framework, it suffices to specify the seminorm as $\|x\| = \|Bx\|_o$. We shall see in the mean time that in the context of this problem, the above "discretization of distances" is, for all practical purposes, irrelevant: the dependence of the volume of observations on N is just logarithmic, so that we can easily handle fine grid, like the one with $\theta = 1.001$ and $\theta^{-N} = 10^{-10}$. As about factor α and additive term β, they indeed could be "expensive in terms of applications," but the "nearly ideal" goal of making α close to 1 and β close to 0 is in many cases too ambitious to be achievable.

[13] A seminorm on $\mathbf{R}^n$ is defined by exactly the same requirements as a norm, except that now we allow zero seminorms for some nonzero vectors. Thus, a seminorm on $\mathbf{R}^n$ is a nonnegative function $\|\cdot\|$ which is even and homogeneous: $\|\lambda x\| = |\lambda|\|x\|$ and satisfies the triangle inequality $\|x + y\| \leq \|x\| + \|y\|$. A universal example is $\|x\| = \|Bx\|_o$, where $\|\cdot\|_o$ is a norm on some $\mathbf{R}^m$ and B is an $m \times n$ matrix; whenever this matrix has a nontrivial kernel, $\|\cdot\|$ is a seminorm rather than a norm.

2.5.3.1 The construction

Let us associate with $i \leq M$ and j, $0 \leq j \leq N$, hypothesis H_{ij} stating that the independent across k observations ω_k, see (2.5.17), stem from $x_* \in X_{ij} = \{x \in X : \|x - x_i\| \leq r_j\}$. Note that the sets X_{ij} are convex and compact. We denote by $\mathcal{J}$ the set of all pairs (i,j), for which $i \in \{1,...,M\}$, $j \in \{0,1,...,N\}$, and $X_{ij} \neq \emptyset$. Further, we define closeness $\mathcal{C}_{\alpha,\beta}$ on the set of hypotheses H_{ij}, $(i,j) \in \mathcal{J}$, as follows:

$(ij, i'j') \in \mathcal{C}_{\alpha\beta}$ *if and only if*

$$\|x_i - x_{i'}\| \leq \bar{\alpha}(r_j + r_{j'}) + \beta, \; \bar{\alpha} = \frac{\alpha - 1}{2}. \tag{2.5.19}$$

(here and in what follows, $k\ell$ denotes the ordered pair (k, ℓ)).

Applying Theorem 2.4.2, we can build, in a computation-friendly fashion, the system $\phi_{ij,i'j'}(\omega)$, $ij, i'j' \in \mathcal{J}$, of optimal balanced detectors for the hypotheses H_{ij} along with the risks of these detectors, so that

$$\begin{array}{ll} (a) & \phi_{ij,i'j'}(\omega) \equiv -\phi_{i'j',ij}(\omega) \, \forall (ij, i'j' \in \mathcal{J}) \\ (b) & \int_\Omega \mathrm{e}^{-\phi_{ij,i'j'}(\omega)} p_{A(x)}(\omega) \Pi(d\omega) \leq \epsilon_{ij,i'j'} \; \forall (ij \in \mathcal{J}, i'j' \in \mathcal{J}, x \in X_{ij}) \end{array} \tag{2.5.20}$$

Let us say that a pair (α, β) is *admissible*, if $\alpha \geq 1$, $\beta \geq 0$ and

$$\forall ((i,j) \in \mathcal{J}, (i',j') \in \mathcal{J}, (ij, i'j') \notin \mathcal{C}_{\alpha,\beta}) : A(X_{ij}) \cap A(X_{i'j'}) = \emptyset. \tag{2.5.21}$$

Note that checking admissibility of a given pair (α, β) is a computationally tractable task.

Given an admissible par (α, β), we associate with it positive integer $K = K(\alpha, \beta)$ and inference $\omega^K \mapsto i_{\alpha,\beta}(\omega^K)$ as follows:

1. $K = K(\alpha, \beta)$ is the smallest integer such that the detector-based test $\mathcal{T}^K_{\mathcal{C}_{\alpha,\beta}}$ yielded by the machinery of Section 2.5.2.3 decides on the hypotheses H_{ij}, $ij \in \mathcal{J}$, with $\mathcal{C}_{\alpha,\beta}$-risk not exceeding ϵ. Note that by admissibility, $\epsilon_{ij,i'j'} < 1$ whenever $(ij, i'j') \notin \mathcal{C}_{\alpha,\beta}$, so that $K(\alpha, \beta)$ is well defined.

2. Given observation ω^K, $K = K(\alpha, \beta)$, we define $i_{\alpha,\beta}(\omega^K)$ as follows:

   (a) We apply to ω^K the test $\mathcal{T}^K_{\mathcal{C}_{\alpha,\beta}}$. If the test accepts no hypothesis (case A), $i_{\alpha\beta}(\omega^K)$ is undefined. The observations ω^K resulting in case A comprise some set, which we denote by $\mathcal{B}$; given ω^K, we can recognize whether or not $\omega^K \in \mathcal{B}$.

   (b) When $\omega^K \notin \mathcal{B}$, the test $\mathcal{T}^K_{\mathcal{C}_{\alpha,\beta}}$ accepts some of the hypotheses H_{ij}, let the set of their indexes ij be $\mathcal{J}(\omega^K)$; we select from the pairs $ij \in \mathcal{J}(\omega^K)$ one with the largest j, and set $i_{\alpha,\beta}(\omega^K)$ to be equal to the first component, and $j_{\alpha,\beta}(\omega^K)$ to be equal to the second component of the selected pair.

We are about to prove the following

Proposition 2.5.7 *Assuming (α, β) admissible, for the just defined inference $\omega^K \mapsto i_{\alpha,\beta}(\omega^K)$ and for every $x_* \in X$, denoting by $P^K_{x_*}$ the distribution of stationary K-repeated observation ω^K stemming from x_* one has*

$$\|x_* - x_{i_{\alpha,\beta}(\omega^K)}\| \leq \alpha\rho(x_*) + \beta. \tag{2.5.22}$$

with $P^K_{x_*}$-probability at least $1 - \epsilon$.

Proof. Let us fix $x_* \in X$, let $j_* = j_*(x_*)$ be the largest $j \leq N$ such that $r_j \geq \min_{i \leq M} \|x_* - x_i\|$; note that j_* is well defined due to $r_0 = D \geq \|x_* - x_1\|$. We set

$$r_{j_*} = \min_j \{r_j : r_j \geq \min_i \|x_* - x_i\|\} = \rho(x_*)$$

and specify $i_* = i_*(x_*) \leq M$ in such a way that

$$\|x_* - x_{i_*}\| \leq r_{j_*}. \tag{2.5.23}$$

Note that i_* is well defined and that observations (2.5.17) stemming from x_* obey the hypothesis $H_{i_* j_*}$.

Let $\mathcal{E}$ be the set of those ω^K for which the predicate

> $\mathcal{P}$: *As applied to observation ω^K, the test $\mathcal{T}^K_{\mathcal{C}\alpha,\beta}$ accepts $H_{i_* j_*}$, and all hypotheses accepted by the test are $\mathcal{C}_{\alpha,\beta}$-close to $H_{i_* j_*}$*

holds true. Taking into account that the $\mathcal{C}_{\alpha,\beta}$-risk of $\mathcal{T}^K_{\mathcal{C}_{\alpha,\beta}}$ does not exceed ϵ and that the hypothesis $H_{i_* j_*}$ is true, the $P^K_{x_*}$-probability of the event $\mathcal{E}$ is at least $1 - \epsilon$.

Let observation ω^K satisfy

$$\omega^K \in \mathcal{E}. \tag{2.5.24}$$

Then

1. The test $\mathcal{T}^K_{\mathcal{C}_{\alpha,\beta}}$ accepts the hypothesis $H_{i_* j_*}$, that is, $\omega^K \not\in \mathcal{B}$. By construction of $i_{\alpha,\beta}(\omega^K) j_{\alpha,\beta}(\omega^K)$ (see the rule 2b above) and due to the fact that $\mathcal{T}^K_{\mathcal{C}_{\alpha,\beta}}$ accepts $H_{i_* j_*}$, we have $j_{\alpha,\beta}(\omega^K) \geq j_*$.

2. The hypothesis $H_{i_{\alpha,\beta}(\omega^K) j_{\alpha,\beta}(\omega^K)}$ is $\mathcal{C}_{\alpha,\beta}$-close to $H_{i_* j_*}$, so that

$$\|x_{i_*} - x_{i_{\alpha,\beta}(\omega^K)}\| \leq \bar{\alpha}(r_{j_*} + r_{j_{\alpha,\beta}(\omega^K)}) + \beta \leq 2\bar{\alpha} r_{j_*} + \beta = 2\bar{\alpha}\rho(x_*) + \beta, \tag{2.5.25}$$

   where the concluding inequality is due to the fact that, as we have already seen, $j_{\alpha,\beta}(\omega^K) \geq j_*$ when (2.5.24) takes place.

Invoking (2.5.23), we conclude that with $P^K_{x_*}$-probability at least $1 - \epsilon$ it holds

$$\|x_* - x_{i_{\alpha,\beta}(\omega^K)}\| \leq (2\bar{\alpha} + 1)\rho(x_*) + \beta = \alpha\rho(x_*) + \beta, \tag{2.5.26}$$

where the concluding equality is due to the definition of $\bar{\alpha}$. □

2.5.3.2 A modification

From the computational viewpoint, a shortcoming of the construction presented in the previous Section is the necessity to operate with $M(N+1)$ hypotheses, which could require computing as many as $O(M^2N^2)$ detectors. We are about to present a modified construction, where we deal at most $N+1$ times with just M hypotheses at a time (i.e., with the total of at most $O(M^2N)$ detectors). The idea is to replace simultaneous processing of all hypotheses H_{ij}, $ij \in \mathcal{J}$, with processing them in *stages* $j = 0, 1, ...,$, with stage j operating only with the hypotheses H_{ij}, $i = 1, ..., M$.

The implementation of this idea is as follows. In the situation of Section 2.5.3, given the same entities Γ, (α, β), H_{ij}, X_{ij}, $ij \in \mathcal{J}$, as in the beginning of Section 2.5.3.1 and specifying closeness $\mathcal{C}_{\alpha,\beta}$ according to (2.5.19), we now act as follows.

Preprocessing. We look, one by one, at $j = 0, 1, ..., N$, and for such a j,

1. identify the set $\mathcal{I}_j = \{i \leq M : X_{ij} \neq \emptyset\}$ and stop if this set is empty. If this set is nonempty, we

2. specify closeness $\mathcal{C}^j_{\alpha\beta}$ on the set of hypotheses H_{ij}, $i \in \mathcal{I}_j$ as a "slice" of the closeness $\mathcal{C}_{\alpha,\beta}$:

   H_{ij} and $H_{i'j}$ (equivalently, i and i') are $\mathcal{C}^j_{\alpha,\beta}$-close to each other if $(ij, i'j)$ are $\mathcal{C}_{\grave{\alpha},\beta}$-close, that is,

   $$\|x_i - x_{i'}\| \leq 2\bar{\alpha} r_j + \beta,\ \bar{\alpha} = \frac{\alpha - 1}{2}.$$

3. build the optimal detectors $\phi_{ij,i'j}$, along with their risks $\epsilon_{ij,i'j}$, for all $i, i' \in \mathcal{I}_j$ such that $(i, i') \notin \mathcal{C}^j_{\alpha,\beta}$.

   If for a pair i, i' of this type it happens that $\epsilon_{ij,i'j} = 1$, that is, $A(X_{ij}) \cap A(X_{i'j}) \neq \emptyset$, we claim that (α, β) is inadmissible and stop. Otherwise we find the smallest $K = K_j$ such that the spectral norm of the symmetric $M \times M$ matrix E^{jK} with the entries

   $$E^{jK}_{ii'} = \left\{ \begin{array}{ll} \epsilon^K_{ij,i'j}, & i \in \mathcal{I}_j, i' \in \mathcal{I}_j, (i, i') \notin \mathcal{C}^j_{\alpha,\beta} \\ 0, & \text{otherwise} \end{array} \right.$$

   does not exceed $\bar{\epsilon} = \epsilon/(N+1)$. We then use the machinery of Section 2.5.2.3 to build detector-based test $\mathcal{T}^{K_j}_{\mathcal{C}^j_{\alpha,\beta}}$ which decides on the hypotheses H_{ij}, $i \in \mathcal{I}_j$, with $\mathcal{C}^j_{\alpha,\beta}$-risk not exceeding $\bar{\epsilon}$.

It may happen that the outlined process stops when processing some value $\bar{j}$ of j; if this does not happen, we set $\bar{j} = N + 1$. Now, if the process does stop, and stops with the claim that (α, β) is inadmissible, we call (α, β) inadmissible and terminate – in this case we fail to produce a desired inference; note that if this is the case, (α, β) is inadmissible in the sense of Section 2.5.3.1 as well. When we do not stop with inadmissibility claim, we call (α, β) admissible, and in this case we do produce an inference, specifically, as follows.

Processing observations.

1. We set $\bar{\mathcal{J}} = \{0, 1, ..., \widehat{j} = \bar{j} - 1\}$, $K = K(\alpha, \beta) = \max\limits_{0 \leq j \leq \widehat{j}} K^j$. Note that $\bar{\mathcal{J}}$ is nonempty due to $\bar{j} > 0$. [14]

2. Given observation ω^K with independent across k components stemming from unknown signal $x_* \in X$ according to (2.5.17), we act as follows.

   (a) We set $\widehat{\mathcal{I}}_{-1}(\omega^K) = \{1, ..., M\} = \mathcal{I}_0$.

   (b) We look, one by one, at the values $j = 0, 1, ..., \widehat{j}$. When processing j, we already have at our disposal subsets $\widehat{\mathcal{I}}_k(\omega^K) \subset \{1, ..., M\}$, $-1 \leq k < j$, and act as follows:

      i. we apply the test $\mathcal{T}^{K_j}_{\mathcal{C}^j_{\alpha,\beta}}$ to the initial K_j components of the observation ω^K. Let $\mathcal{I}^+_j(\omega^K)$ be the set of hypotheses H_{ij}, $i \in \mathcal{I}_j$, accepted by the test.

      ii. it may happen that $\mathcal{I}^+_j(\omega^K) = \emptyset$; if it is so, we terminate.

[14] All the sets X_{i0} contain X and thus are nonempty, so that $\mathcal{I}_0 = \{1, ..., M\} \neq \emptyset$, and thus we cannot stop at step $j = 0$ due to $\mathcal{I}_0 = \emptyset$; and another possibility to stop at step $j = 0$ is ruled out by the fact that we are in the case when (α, β) is admissible.

iii. if $\mathcal{I}_j^+(\omega^K)$ is nonempty, we look, one after one, at indexes $i\in\mathcal{I}_j^+(\omega^K)$ and for such an i, check, for every $\ell\in\{-1,0,...,j-1\}$, whether $i\in\widehat{\mathcal{I}}_\ell(\omega^K)$. If it is the case for every $\ell\in\{-1,0,...,j-1\}$, we call index i good.

iv. if good indexes in $\mathcal{I}_j^+(\omega^K)$ are discovered, we define $\widehat{\mathcal{I}}_j(\omega^K)$ as the set of these good indexes and process to the next value of j (if $j<\widehat{j}$), or terminate (if $j=\widehat{j}$). If there are no good indexes in $\mathcal{I}_j^+(\omega^K)$, we terminate.

(c) Upon termination, we have at our disposal a collection $\widehat{\mathcal{I}}_j(\omega^K)$, $0\leq j\leq\widetilde{j}(\omega^K)$, of all sets $\widehat{\mathcal{I}}_j(\omega^K)$ we have built (this collection can be empty, which we encode by setting $\widetilde{j}(\omega^K)=-1$). When $\widetilde{j}(\omega^K)=-1$, our inference remains undefined. Otherwise we select from the set $\widehat{\mathcal{I}}_{\widetilde{j}(\omega^K)}(\omega^K)$ an index $i_{\alpha,\beta}(\omega^K)$, say, the smallest one, and claim that the point $x_{i_{\alpha,\beta}(\omega^K)}$ is the "nearly closest" to x_* point among $x_1,...,x_M$.

We have the following analogy of Proposition 2.5.7:

Proposition 2.5.8 *Assuming (α,β) admissible, for the just defined inference $\omega^K\mapsto i_{\alpha,\beta}(\omega^K)$ and for every $x_*\in X$, denoting by $P_{x_*}^K$ the distribution of stationary K-repeated observation ω^K stemming from x_* one has*

$$P_{x_*}^K\left\{\omega^K : i_{\alpha,\beta}(\omega^K) \text{ is well defined and } \|x_*-x_{i_{\alpha,\beta}(\omega^K)}\|\leq\alpha\rho(x_*)+\beta\right\}\geq 1-\epsilon. \tag{2.5.27}$$

Proof. Let us fix the signal $x_*\in X$ underlying observations ω^K. Same as in the proof of Proposition 2.5.7, let j_* be such that $\rho(x_*)=r_{j_*}$, and let $i_*\leq M$ be such that $x_*\in X_{i_*j_*}$; clearly, i_* and j_* are well defined, and the hypotheses H_{i_*j}, $0\leq j\leq j_*$, are true. In particular, $X_{i_*j}\neq\emptyset$ when $j\leq j_*$, implying that $i_*\in\mathcal{I}_j$, $0\leq j\leq j_*$, whence also $\widehat{j}\geq j_*$.

For $0\leq j\leq j_*$, let $\mathcal{E}_j$ be the set of all realizations of ω^K such that

$$i_*\in\mathcal{I}_j^+(\omega^K)\ \&\ \{(i_*,i)\in\mathcal{C}_{\alpha,\beta}^j\,\forall i\in\mathcal{I}_j^+(\omega^K)\}.$$

Since $\mathcal{C}_{\alpha,\beta}^j$-risk of the test $\mathcal{T}_{\mathcal{C}_{\alpha,\beta}^j}^{K_j}$ is $\leq\bar{\epsilon}$, we conclude that the $P_{x_*}^K$-probability of $\mathcal{E}_j$ is at least $1-\bar{\epsilon}$, whence the $P_{x_*}^K$-probability of the event

$$\mathcal{E}=\bigcap_{j=0}^{j_*}\mathcal{E}_j$$

is at least $1-(N+1)\bar{\epsilon}=1-\epsilon$.

Now let

$$\omega^K\in\mathcal{E}.$$

Then, by the definition of $\mathcal{E}_j$, $j\leq j_*$,

- When $j\leq j_*$, we have $i_*\in\mathcal{I}_j^+(\omega^K)$, whence, by evident induction in j, $i_*\in\widehat{\mathcal{I}}_j(\omega^K)$ for all $j\leq j_*$.
- From the above item, $\widetilde{j}(\omega^K)\geq j_*$; in particular, $i:=i_{\alpha,\beta}(\omega^K)$ is well defined and turned out to be good at step $\widetilde{j}\geq j_*$, implying that $i\in\widehat{\mathcal{I}}_{j_*}(\omega^K)\subset\mathcal{I}_{j_*}^+(\omega^K)$.

Thus, $i\in\mathcal{I}_{j_*}^+(\omega^K)$, which combines with the definition of $\mathcal{E}_{j_*}$ to imply that i and i_* are $\mathcal{C}_{\alpha,\beta}^{j_*}$-close to each other, whence

$$\|x_{i(\alpha,\beta)(\omega^K)}-x_{i_*}\|\leq 2\bar{\alpha}r_{j_*}+\beta=2\bar{\alpha}\rho(x_*)+\beta,$$

resulting in the desired relation

$$\|x_{i(\alpha,\beta)(\omega^K)}-x_*\|\leq 2\bar{\alpha}\rho(x_*)+\beta+\|x_{i_*}-x_*\|\leq[2\bar{\alpha}+1]\rho(x_*)+\beta=\alpha\rho(x_*)+\beta. \qquad\square$$

2.5.3.3 "Near-optimality"

We augment the above simple constructions with the following

Proposition 2.5.9 *Let in the nature for some positive integer $\bar{K}$, $\epsilon\in(0,1/2)$ and a pair $(a,b)\geq 0$ there exists an inference $\omega^{\bar{K}}\mapsto i(\omega^{\bar{K}})\in\{1,...,M\}$ such that whenever $x_*\in X$, we have*

$$\mathrm{Prob}_{\omega^{\bar{K}}\sim P_{x_*}^{\bar{K}}}\{\|x_*-x_{i(\omega^{\bar{K}})}\|\leq a\rho(x_*)+b\}\geq 1-\epsilon.$$

Then the pair $(\alpha=2a+3,\beta=2b)$ is admissible in the sense of Section 2.5.3.1 (and thus – in the sense of Section 2.5.3.2), and for both our constructions – the one from Section 2.5.3.1 and the one from Section 2.5.3.2) – one has

$$K(\alpha,\beta)\leq \mathrm{Ceil}\left(2\frac{1+\ln(M(N+1))/\ln(1/\epsilon)}{1-\frac{\ln(4(1-\epsilon))}{\ln(1/\epsilon)}}\bar{K}\right); \tag{2.5.28}$$

Proof. Consider the situation of Section 2.5.3.1 (the situation of Section 2.5.3.2 can be processed in a completely similar fashion). Observe that with α,β as above, there exists a simple test deciding on a pair of hypotheses H_{ij}, $H_{i'j'}$ which are *not* $\mathcal{C}_{\alpha,\beta}$-close to each other via stationary $\bar{K}$-repeated observation $\omega^{\bar{K}}$ with risk $\leq\epsilon$. Indeed, the desired test $\mathcal{T}$ is as follows: given $ij\in\mathcal{J}$, $i'j'\in\mathcal{J}$, and observation $\omega^{\bar{K}}$, we compute $i(\omega^{\bar{K}})$ and accept H_{ij} if and only if $\|x_{i(\omega^{\bar{K}})}-x_i\|\leq(a+1)r_j+b$, and accept $H_{i'j'}$ otherwise. Let us check that the risk of this test indeed is at most ϵ. Assume, first, that H_{ij} takes place. The $P_{x_*}^{\bar{K}}$-probability of the event $\mathcal{E}:\|x_{i(\omega^{\bar{K}})}-x_*\|\leq a\rho(x_*)+b$ is at lest $1-\epsilon$ due to the origin of $i(\cdot)$, and $\|x_i-x_*\|\leq r_j$ since H_{ij} takes place, implying that $\rho(x_*)\leq r_j$ by the definition of $\rho(\cdot)$. Thus, in the case of $\mathcal{E}$ it holds

$$\|x_{i(\omega^{\bar{K}})}-x_i\|\leq\|x_{i(\omega^{\bar{K}})}-x_*\|+\|x_i-x_*\|\leq a\rho(x_*)+b+r_j\leq(a+1)r_j+b.$$

We conclude that if H_{ij} is true and $\omega^{\bar{K}}\in\mathcal{E}$, then the test $\mathcal{T}$ accepts H_{ij}, and thus the $P_{x_*}^{\bar{K}}$-probability for the simple test $\mathcal{T}$ not to accept H_{ij} when the hypothesis takes place is $\leq\epsilon$.

Now let $H_{i'j'}$ take place, and let $\mathcal{E}$ be the same event as above. When $\omega^{\bar{K}}\in\mathcal{E}$, which happens with the $P_{x_*}^{\bar{K}}$-probability at least $1-\epsilon$, we by exactly the same reasons as above have $\|x_{i(\omega^{\bar{K}})}-x_{i'}\|\leq(a+1)r_{j'}+b$. It follows that when $H_{i'j'}$ takes place and $\omega^{\bar{K}}\in\mathcal{E}$, we have $\|x_{i(\omega^{\bar{K}})}-x_i\|>(a+1)r_j+b$, since otherwise we would have

$$\begin{aligned}\|x_i-x_{i'}\|\leq\|x_{i(\omega^{\bar{K}})}-x_i\|+\|x_{i(\omega^{\bar{K}})}-x_{i'}\|&\leq(a+1)r_j+b+(a+1)r_{j'}+b\\ &\leq(a+1)(r_j+r_{j'})+2b=\tfrac{\alpha-1}{2}(r_j+r_{j'})+\beta,\end{aligned}$$

which contradicts the fact that ij and $i'j'$ are not $\mathcal{C}_{\alpha,\beta}$-close. Thus, whenever $H_{i'j'}$ holds true and $\mathcal{E}$ takes place, we have $\|x_{i(\omega^{\bar{K}})}-x_i\|>(a+1)r_j+b$, implying that $\mathcal{T}$ accepts $H_{i'j'}$. Thus, the $P_{x_*}^{\bar{K}}$-probability not to accept $H_{i'j'}$ when the hypotheses if true is at most ϵ. From the just established fact that whenever $(ij,i'j')\not\in\mathcal{C}_{\alpha,\beta}$, the hypotheses H_{ij}, $H_{i'j'}$ can be decided upon, via $\bar{K}$ observations, with risk $\leq\epsilon<0.5$ it follows that for $ij,i'j'$ in question, the sets $A(X_{ij})$ and $A(X_{i'j'})$ do not intersect, so that (α,β) is an admissible pair.

Same as in the proof of Proposition 2.5.6, by basic properties of simple observation schemes, the fact that the hypotheses H_{ij}, $H_{i'j'}$ with $(ij,i'j')\not\in\mathcal{C}_{\alpha,\beta}$ can be decided upon via $\bar{K}$-repeated observations (2.5.17) with risk $\leq\epsilon<1/2$ implies that $\epsilon_{ij,i'j'}\leq[2\sqrt{\epsilon(1-\epsilon)}]^{1/\bar{K}}$, whence, again by basic results on simple observation scheme (look once again at the proof of Proposition 2.5.6), the $\mathcal{C}_{\alpha,\beta}$-risk of K-observation detector-based test $\mathcal{T}_K$ deciding on the hypotheses H_{ij}, $ij\in\mathcal{J}$, up to closeness $\mathcal{C}_{\alpha,\beta}$ does not exceed $\mathrm{Card}(\mathcal{J})[2\sqrt{\epsilon(1-\epsilon)}]^{K/\bar{K}}\leq M(N+1)[2\sqrt{\epsilon(1-\epsilon)}]^{K/\bar{K}}$, and (2.5.28) follows. □

Comment. Proposition 2.5.9 says that in our problem, the "statistical toll" for quite large values of N and M is quite moderate: with $\epsilon = 0.01$, resolution $\theta = 1.001$ (which for all practical purposes is the same as no discretization of distances at all), D/r_N as large as 10^{10}, and M as large as 10,000, (2.5.28) reads $K = \text{Ceil}(10.7\bar{K})$ – not a disaster! The actual statistical toll in our construction is in replacing the "existing in the nature" a and b with $\alpha = 2\alpha + 3$ and $\beta = 2b$. And of course there is a huge computational toll for large M and N: we need to operate with large (albeit polynomial in M, N) number of hypotheses and detectors.

2.5.3.4 Numerical illustration

The toy problem we use to illustrate the approach presented in this Section is as follows:

> A signal $x_* \in \mathbf{R}^n$ (it makes sense to think of x_* as of the restriction on the equidistant n-point grid in $[0, 1]$ of a function of continuous argument $t \in [0, 1]$) is observed according to
> $$\omega = Ax_* + \xi,\ \xi \sim \mathcal{N}(0, \sigma^2 I_n), \tag{2.5.29}$$
> where A is "discretized integration:"
> $$(Ax)_s = \frac{1}{n}\sum_{j=1}^{s} x_s,\ s = 1, ..., n.$$
> We want to approximate x in the discrete version of L_1-norm
> $$\|y\| = \frac{1}{n}\sum_{s=1}^{n} |y_s|,\ y \in \mathbf{R}^n$$
> by a low order polynomial.

In order to build the approximation, we use a single observation ω, stemming from x_* according to (2.5.29), to build 5 candidate estimates x_i, $i = 1, ..., 5$ of x_*. Specifically, x_i is the Least Squares polynomial, of degree $\leq i - 1$, approximation of x:

$$x_i = \operatorname*{argmin}_{y \in \mathcal{P}_{i-1}} \|Ay - \omega\|_2^2,$$

where $\mathcal{P}_\kappa$ is the linear space of algebraic polynomials, of degree $\leq \kappa$, of discrete argument s varying in $\{1, 2, ..., n\}$. After the candidate estimates are built, we use additional K observations (2.5.29) "to select the model" – to select among our estimates the $\|\cdot\|$-closest to x_*.

In the experiment to be reported we used $n = 128$ and $\sigma = 0.01$. The true signal x_* is plotted in magenta on the top of Figure 2.3; it is discretization of function of continuous argument $t \in [0, 1]$ which is linear, with slope 1, to the left of $t = 0.5$, and is linear, with slope -1, to the right of $t = 0.5$; at $t = 0.5$, the function has a jump. A priori information on the true signal is that it belongs to the box $\{x \in \mathbf{R}^n : \|x\|_\infty \leq 1\}$. Sample polynomial approximations x_i of x_*, $1 \leq i \leq 5$, are plotted in blue on the top of Figure 2.3; their actual $\|\cdot\|$-distances to x_* are as follows:

i	1	2	3	4	5
$\|x_i - x_*\|$	0.534	0.354	0.233	0.161	0.172

As usual, the reliability tolerance ϵ was set to 0.01. We used $N = 22$ and $\theta = 2^{1/4}$, $\alpha = 3$, $\beta = 0.05$, resulting in $K = 3$. In a series of 1000 simulations of the resulting inference, *all* 1000 results correctly identified the $\|\cdot\|$-closest to x_* candidate estimate, specifically, x_4, in spite of the factor $\alpha = 3$ in (2.5.22). Surprisingly, the same holds true when we use the resulting inference with the reduced values of K, namely, $K = 1$ and $K = 2$, although the theoretical reliability guarantees deteriorate: with $K = 1$ and $K = 2$, theory guarantees the validity of (2.5.22) with probabilities 0.77 and 0.97, respectively.

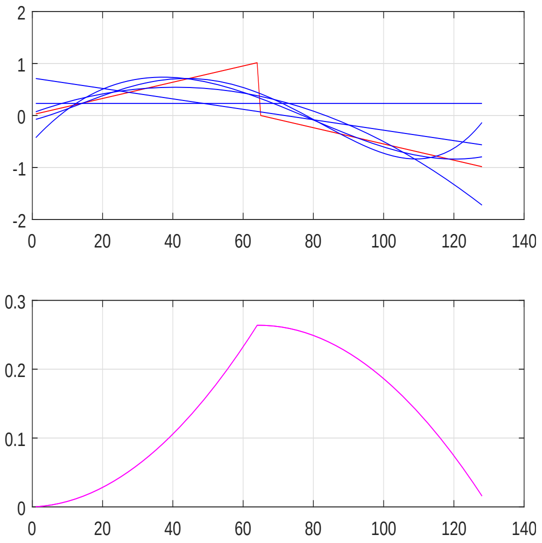


Figure 2.3: Signal (top, magenta) and its candidate estimates (top,blue). Bottom: the primitive of the signal.

2.6 Sequential Hypothesis Testing

2.6.1 Motivation: Election Polls

Consider the question as follows:

> One of L candidates for an office is about to be selected by population-wide majority vote. Every member of the population votes for exactly one of the candidates. How to predict the winner via an opinion poll?

A (naive) model of situation could be as follows. Let us represent the preference of a particular voter by his *preference vector* – basic orth e in $\mathbf{R}^L$ with unit entry in a position ℓ meaning that the voter is about to vote for the ℓ-th candidate. The entries μ_ℓ in the average μ, over the population, of these vectors are the fractions of votes in favor of ℓ-th candidate, and the elected candidate is the one "indexing" the largest of μ_ℓ's. Now assume that we select at random, from the uniform distribution, a member of the population and observe his preference vector. Our observation ω is a realization of discrete random variable taking values in the set $\Omega=\{e_1,...,e_L\}$ of basic orths in $\mathbf{R}^L$, and μ is the distribution of ω (technically, the density of this distribution w.r.t. the counting measure Π on Ω). Selecting a small threshold δ and assuming that the true – unknown to us – μ is such that the largest entry in μ is at least by δ larger than every other entry and that $\mu_\ell\geq\frac{1}{N}$ for all ℓ, N being the population size[15], the fact that ℓ-th candidate wins the elections means that

$$\mu\in M_\ell=\{\mu\in\mathbf{R}^d:\mu_i\geq\frac{1}{N},\sum_i\mu_i=1,\mu_\ell\geq\mu_i+\delta\,\forall(i\neq\ell)\}\subset\mathcal{M}=\{\mu\in\mathbf{R}^d:\mu>0,\sum_i\mu_i=1\}.$$

In an (idealized) poll, we select at random a number K of voters and observe their preferences, thus arriving at a sample $\omega^K=(\omega_1,...,\omega_K)$ of observations drawn, independently of each other, from unknown distribution μ on Ω, with μ known to belong to $\bigcup_{\ell=1}^L M_\ell$, and to predict the winner is the same as to decide on L convex hypotheses, $H_1,...,H_L$, in the Discrete o.s., with H_ℓ stating that $\omega_1,...,\omega_K$ are drawn, independently of each other, from a distribution $\mu\in M_\ell$. What we end

[15] with the size N of population in the range of tens of thousands and δ like $1/N$, both these assumptions seem to be quite realistic.

up with, is the problem of deciding on L convex hypotheses in the Discrete o.s. with L-element Ω via stationary K-repeated observations.

Illustration. Consider two-candidate elections; now the goal of a poll is, given K independent of each other realizations $\omega_1, ..., \omega_K$ of random variable ω taking value $\chi = 1, 2$ with probability μ_χ, $\mu_1 + \mu_2 = 1$, to decide what is larger, μ_1 or μ_2. As explained above, we select somehow a threshold δ and impose on the unknown μ a priori assumption that the gap between the largest and the next largest (in our case – just the smallest) entry of μ is at least δ, thus arriving at two hypotheses:

$$H_1 : \mu_1 \geq \mu_2 + \delta, \quad H_2 : \mu_2 \geq \mu_1 + \delta,$$

which is the same as

$$\begin{array}{l} H_1 : \mu \in M_1 = \{\mu : \mu_1 \geq \frac{1+\delta}{2}, \mu_2 \geq 0, \mu_1 + \mu_2 = 1\}, \\ H_2 : \mu \in M_2 = \{\mu : \mu_2 \geq \frac{1+\delta}{2}, \mu_1 \geq 0, \mu_1 + \mu_2 = 1\}. \end{array}$$

We now want to decide on these two hypotheses from stationary K-repeated observations. We are in the case of simple (specifically, Discrete) o.s.; the optimal detector as given by Theorem 2.4.2 stems from the optimal solution (μ^*, ν^*) to the convex optimization problem

$$\varepsilon_\star = \max_{\mu \in M_1, \nu \in M_2} \left[\sqrt{\mu_1 \nu_1} + \sqrt{\mu_2 \nu_2}\right], \tag{2.6.1}$$

the optimal balanced single-observation detector is

$$\phi_*(\omega) = f_*^T \omega, \;\; f_* = \frac{1}{2}[\ln(\mu_1^*/\nu_1^*); \ln(\mu_2^*/\nu_2^*)]$$

(recall that we encoded observations ω_k by basic orths from $\mathbf{R}^2$), the risk of this detector being $\varepsilon_\star$. In other words,

$$\begin{array}{l} \mu^* = [\frac{1+\delta}{2}; \frac{1-\delta}{2}], \; \nu^* = [\frac{1-\delta}{2}; \frac{1+\delta}{2}], \; \varepsilon_\star = \sqrt{1-\delta^2}, \\ f_* = \frac{1}{2}\left[\ln((1+\delta)/(1-\delta)); \ln((1-\delta)/(1+\delta))\right]. \end{array}$$

The optimal balanced K-observation detector and its risk are

$$\phi_*^{(K)}(\underbrace{\omega_1, ..., \omega_K}_{\omega^K}) = f_*^T(\omega_1 + ... + \omega_K), \;\; \varepsilon_\star^{(K)} = (1-\delta^2)^{K/2}.$$

The near-optimal K-observation test $\mathcal{T}_{\phi_*}^K$ accepts H_1 and rejects H_2 when $\phi_*^{(K)}(\omega^K) \geq 0$, otherwise it accepts H_2 and rejects H_1. Both risks of this test do not exceed $\varepsilon_\star^{(K)}$.

Given risk level ϵ, we can identify the minimal "poll size" K for which the risks Risk$_1$, Risk$_2$ of the test $\mathcal{T}_{\phi*}^K$ do not exceed ϵ. This poll size depends on ϵ and on our a priory "hypotheses separation" parameter $\delta : K = K_\epsilon(\delta)$. Some impression on this size can be obtained from Table 2.1, where, as in all subsequent "election illustrations," ϵ is set to 0.01. We see that while poll sizes for "landslide" elections are surprisingly low, reliable prediction of the results of "close run" elections requires surprisingly high sizes of the polls. Note that this phenomenon reflects reality (to the extent at which the reality is captured by our model[16]); indeed, from Proposition 2.4.2 we know

[16] in actual opinion polls, additional information is used; for example, in reality voters can be split into groups according to their age, sex, education, income, etc., etc., with variability of preferences within a group essentially lower than across the entire population; when planning a poll, respondents are selected at random within these groups, with a prearranged number of selections in every group, and their preferences are properly weighted, yielding more accurate predictions as compared to the case when the respondents are selected from the uniform distribution. In other words, in actual polls a non-trivial a priori information on the "true" distribution of preferences is used – something we do not have in our naive model.

δ	0.5623	0.3162	0.1778	0.1000	0.0562	0.0316	0.0177	0.0100
$K_{0.01}(\delta), L=2$	25	88	287	917	2908	9206	29118	92098
$K_{0.01}(\delta), L=5$	32	114	373	1193	3784	11977	37885	119745

Table 2.1: Sample of values of poll size $K_{0.01}(\delta)$ as a function of δ for 2-candidate ($L = 2$) and 5-candidate ($L = 5$) elections. Values of δ form a decreasing geometric progression with ratio $10^{-1/4}$.

that our poll size is within an explicit factor, depending solely on ϵ, from the "ideal" poll sizes – the smallest ones which allow to decide upon H_1, H_2 with risk $\leq \epsilon$. For $\epsilon = 0.01$, this factor is about 2.85, meaning that when $\delta = 0.01$, the ideal poll size is larger than 32,000. In fact, we can build more accurate lower bounds on the sizes of ideal polls, specifically, as follows. When computing the optimal detector ϕ_*, we get, as a byproduct, two distributions, μ^*, ν^* obeying H_1, H_2, respectively. Denoting by μ^*_K, ν^*_K the distributions of K-element i.i.d. samples drawn from μ^* and ν^*, the risk of deciding on two simple hypotheses on the distribution of ω^K, stating that this distribution is μ^*_K, respectively, ν^*_K can be only smaller than the risk of deciding on H_1, H_2 via K-repeated stationary observations. On the other hand, the former risk can be lower-bounded by one half of the total risk of deciding on our two simple hypotheses, and the latter risk admits a sharp lower bound given by Proposition 2.1.1, namely,

$$\sum_{i_1,\dots,i_K\in\{1,2\}} \min\left[\prod_\ell \mu^*_{i_\ell}, \prod_\ell \nu^*_{i_\ell}\right] = \mathbf{E}_{(i_1,\dots,i_K)}\left\{\min\left[\prod_\ell (2\mu^*_{i_\ell}), \prod_\ell (2\nu^*_{i_\ell})\right]\right\},$$

with the expectation taken w.r.t independent tuples of K integers taking values 1 and 2 with probabilities 1/2. Of course, when K is in the range of few tens and more, we cannot compute the above 2^K-term sum exactly; however, we can use Monte Carlo simulation in order to estimate the sum reliably within moderate accuracy, like 0.005, and use this estimate to lower-bound the value of K for which "ideal" K-observation test decides on H_1, H_2 with risks ≤ 0.01. Here are the resulting lower bounds (along with upper bounds stemming from the data in Table 2.1):

δ	0.5623	0.3162	0.1778	0.1000	0.0562	0.0316	0.0177	0.0100
$\underline{K}, \overline{K}$	$14, 25$	$51, 88$	$166, 287$	$534, 917$	$1699, 2908$	$5379, 9206$	$17023, 29122$	$53820, 92064$

Lower ($\underline{K}$) and upper ($\overline{K}$) bounds on the "ideal" poll sizes

We see that the poll sizes as yielded by our machinery are within factor 2 of the "ideal" poll sizes.

Clearly, the outlined approach can be extended to L-candidate elections with $L \geq 2$. We model the corresponding problem as the one where we need to decide, via stationary K-repeated observations drawn from unknown probability distribution μ on L-element set, on L hypotheses

$$H_\ell : \mu \in M_\ell = \{\mu \in \mathbf{R}^d : \mu_i \geq \frac{1}{N}, i \leq L, \sum_i \mu_i = 1, \mu_\ell \geq \mu_{\ell'} + \delta\,\forall(\ell' \neq \ell)\},\ \ell \leq L; \tag{2.6.2}$$

here $\delta > 0$ is a selected in advance threshold small enough to believe that the actual preferences of the voters correspond to $\mu \in \bigcup_\ell M_\ell$. Defining closeness $\mathcal{C}$ in the strongest possible way – H_ℓ is close to $H_{\ell'}$ if and only if $\ell = \ell'$, predicting the outcome of elections with risk ϵ becomes the problem of deciding upon our multiple hypotheses with $\mathcal{C}$-risk $\leq \epsilon$, and we can use the pairwise detectors yielded by Theorem 2.4.2 to identify the smallest possible $K = K_\epsilon$ such that the test $\mathcal{T}^K_{\mathcal{C}}$ from Section 2.5.2.3 is capable to decide upon our L hypotheses with $\mathcal{C}$-risk $\leq \epsilon$. Numerical illustration of the performance of this approach in 5-candidate elections is presented in Table 2.1 (where ϵ is set to 0.01).

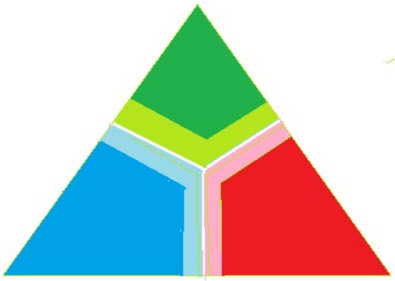


Figure 2.4: 3-candidate hypotheses in probabilistic simplex $\mathbf{\Delta}_3$:
[green] M_1 dark green + light green: candidate A wins with margin $\geq \delta_S$
[green] M_1^s dark green: candidate A wins with margin $\geq \delta_s > \delta_S$
[red] M_2 dark red + pink: candidate B wins with margin $\geq \delta_S$
[red] M_2^s dark red: candidate B wins with margin $\geq \delta_s > \delta_S$
[blue] M_3 dark blue + light blue: candidate C wins with margin $\geq \delta_S$
[blue] M_3^s dark blue: candidate C wins with margin $\geq \delta_s > \delta_S$
$\mathcal{C}_s$ closeness: hypotheses in the tuple $\{G_{2\ell-1}^s : \mu \in M_\ell, G_{2\ell}^s : \mu \in M_\ell^s, 1 \leq \ell \leq 3\}$ are *not* $\mathcal{C}_s$-close to each other, if the corresponding M-sets are of different colors and at least one the sets is dark-painted, like M_1^s and M_2, but not M_1 and M_2.

2.6.2 Sequential hypothesis testing

In view of the above analysis, when predicting outcomes of "close run" elections, huge poll sizes are a must. It, however, does not mean that nothing can be done in order to build more reasonable opinion polls. The classical related statistical idea, going back to Wald [159] is to pass to *sequential tests* where the observations are processed one by one, and at every time we either accept some of our hypotheses and terminate, or conclude that the observations obtained so far are insufficient to make a reliable inference and pass to the next observation. The idea is that a properly built sequential test, while still ensuring a desired risk, will be able to make "early decisions" in the case when the distribution underlying observations is "well inside" the true hypothesis and thus is far from the alternatives. Let us show how to utilize our machinery in building a sequential test for the problem of predicting the outcome of L-candidate elections; thus, our goal is, given a small threshold δ, to decide upon L hypotheses (2.6.2). Let us act as follows.

1. We select a factor $\theta \in (0,1)$, say, $\theta = 10^{-1/4}$, and consider thresholds $\delta_1 = \theta$, $\delta_2 = \theta\delta_1$, $\delta_3 = \theta\delta_2$, and so on, until for the first time we get a threshold $\leq \delta$; to save notation, we assume that this threshold is exactly δ, and let the number of the thresholds be S.

2. We split somehow (e.g., equally) the risk ϵ which we want to guarantee into S portions ϵ_s, $1 \leq s \leq S$, so that ϵ_s are positive and
$$\sum_{s=1}^{S} \epsilon_s = \epsilon.$$

3. For $s \in \{1, 2, ..., S\}$, we define, along with the hypotheses H_ℓ, the hypotheses
$$H_\ell^s : \mu \in M_\ell^s = \{\mu \in M_\ell : \mu_\ell \geq \mu_{\ell'} + \delta_s, \forall (\ell' \neq \ell)\}, \ell = 1, ..., L,$$
see Figure 2.4, and introduce $2L$ hypotheses $G_{2\ell-1}^s = H_\ell$, and $G_{2\ell}^s = H_\ell^s$, $1 \leq \ell \leq L$. It is convenient to color these hypotheses in L colors, with $G_{2\ell-1}^s = H_\ell$ and $G_{2\ell}^s = H_\ell^s$ assigned color ℓ. We define also *s-th closeness* $\mathcal{C}_s$ as follows:

When $s < S$, hypotheses G^s_i and G^s_j are $\mathcal{C}_s$-close to each other if either they are of the same color, or they are of different colors and both of them have odd indexes (that is, one of them is H_ℓ, and another one is $H_{\ell'}$ with $\ell \neq \ell'$).
When $s = S$ (in this case $G^S_{2\ell-1} = H_\ell = G^S_{2\ell}$), hypotheses G^S_ℓ and $G^S_{\ell'}$ are $\mathcal{C}_S$-close to each other if and only if they are of the same color, i.e., both coincide with the same hypothesis H_ℓ.

Observe that G^s_i is a convex hypothesis:

$$G^s_i : \mu \in Y^s_i \qquad\qquad [Y^s_{2\ell-1} = M_\ell, Y^s_{2\ell} = M^s_\ell]$$

The key observation is that when G^s_i and G^s_j are *not* $\mathcal{C}_s$-close, the sets Y^s_i and Y^s_j are "separated" by at least δ_s, meaning that for some vector $e \in \mathbf{R}^L$ with just two nonnegative entries, equal to 1 and -1, we have

$$\min_{\mu \in Y^s_i} e^T\mu \geq \delta_s + \max_{\mu \in Y^s_j} e^T\mu. \tag{2.6.3}$$

Indeed, let G^s_i and G^s_j be not $\mathcal{C}_s$-close to each other. That means that the hypotheses are of different colors, say, ℓ and $\ell' \neq \ell$, and at least one of them has even index; w.l.o.g. we can assume that the even-indexed hypothesis is G^s_i, so that

$$Y^s_i \subset \{\mu : \mu_\ell - \mu_{\ell'} \geq \delta_s\},$$

while Y^s_j is contained in the set $\{\mu : \mu_{\ell'} \geq \mu_\ell\}$. Specifying e as the vector with just two nonzero entries, ℓ-th equal to 1 and ℓ'-th equal to -1, we ensure (2.6.3).

4. For $1 \leq s \leq S$, we apply the construction from Section 2.5.2.3 to identify the smallest $K = K(s)$ for which the test $\mathcal{T}_s$ yielded by this construction as applied to stationary K-repeated observation allows to decide on the hypotheses $G^s_1, ..., G^s_{2L}$ with $\mathcal{C}_s$-risk $\leq \epsilon_s$; the required K exists due to the already mentioned separation of members in a pair of not $\mathcal{C}_s$-close hypotheses G^s_i, G^s_j. It is easily seen that $K(1) \leq K(2) \leq ... \leq K(S-1)$; however, it may happen that $K(S-1) > K(S)$, the reason being that $\mathcal{C}_S$ is defined differently than $\mathcal{C}_s$ with $s < S$. We set
$$\mathcal{S} = \{s \leq S : K(s) \leq K(S)\}.$$
For example, this is what we get in L-candidate Opinion Poll problem when $S = 8$, $\delta = \delta_S = 0.01$, and for properly selected ϵ_s with $\sum_{s=1}^8 \epsilon_s = 0.01$:

L	$K(1)$	$K(2)$	$K(3)$	$K(4)$	$K(5)$	$K(6)$	$K(7)$	$K(8)$
2	177	617	1829	5099	15704	49699	153299	160118
5	208	723	2175	6204	19205	60781	188203	187718

$S = 8$, $\delta_s = 10^{-s/4}$.
$\mathcal{S} = \{1, 2, ..., 8\}$ when $L = 2$ and $\mathcal{S} = \{1, 2, ..., 6\} \cup \{8\}$ when $L = 5$.

5. Our sequential test $\mathcal{T}_{\rm seq}$ works in *attempts* $s \in \mathcal{S}$ – it tries to make conclusions after observing $K(s)$, $s \in \mathcal{S}$, realizations ω_k of ω. At s-th attempt, we apply the test $\mathcal{T}_s$ to the collection $\omega^{K(s)}$ of observations obtained so far to decide on hypotheses $G^s_1, ..., G^s_{2L}$. If $\mathcal{T}_s$ accepts some of these hypotheses *and all accepted hypotheses are of the same color*, let it be ℓ, the sequential test accepts the hypothesis H_ℓ and terminates, otherwise we continue to observe the realizations of ω (when $s < S$) or terminate with no hypotheses accepted/rejected (when $s = S$).

It is easily seen that the risk of the outlined sequential test $\mathcal{T}_{\rm seq}$ does not exceed ϵ, meaning that whatever be a distribution $\mu \in \bigcup_{\ell=1}^L M_\ell$ underlying observations $\omega_1, \omega_2, ...\omega_{K(S)}$ and ℓ_* such that $\mu \in M_{\ell_*}$, the μ-probability of the event

$\mathcal{T}_{\rm seq}$ *accepts exactly one hypothesis, namely,* H_{ℓ_*}

is at least $1 - \epsilon$.

Indeed, observe, first, that the sequential test always accepts at most one of the hypotheses $H_1, ..., H_L$. Second, let $\omega_k \sim \mu$ with μ obeying H_{ℓ_*}. Consider events E_s, $s \in \mathcal{S}$, defined as follows:

- when $s < S$, E_s is the event "the test $\mathcal{T}_s$ as applied to observation $\omega^{K(s)}$ does not accept the true hypothesis $G^s_{2\ell_*-1} = H_{\ell_*}$";
- E_S is the event "as applied to observation $\omega^{K(S)}$, the test $\mathcal{T}_S$ does not accept the true hypothesis $G^S_{2\ell_*-1} = H_{\ell_*}$ or accepts a hypothesis not $\mathcal{C}_S$-close to $G^S_{2\ell_*-1}$."

Note that by our selection of $K(s)$'s, the μ-probability of E_s does not exceed ϵ_s, so that the μ-probability of *no one* of the events E_s, $s \in \mathcal{S}$, taking place is at least $1-\epsilon$. To justify the above claim on the risk of our sequential test, all we need is to verify that *when no one of the events E_s, $s \in \mathcal{S}$, takes place, then the sequential test accepts the true hypothesis H_{ℓ_*}*. Verification is immediate: let the observations be such that no one of the events E_s takes place. We claim that in this case

(a) The sequential test does accept a hypothesis – if this does not happen at s-th attempt with some $s < S$, it definitely happens at S-th attempt.

Indeed, since E_S does not take place, $\mathcal{T}_S$ accepts $G^S_{2\ell_*-1}$ and all other hypotheses, if any, accepted by $\mathcal{T}_S$ are $\mathcal{C}_S$-close to $G^S_{2\ell_*-1}$, implying by construction of $\mathcal{C}_S$ that $\mathcal{T}_S$ does accept hypotheses, and all these hypotheses are of the same color, that is, the sequential test at S-th attempt does accept a hypothesis.

(b) The sequential test does *not* accept a wrong hypothesis.

Indeed, assume that the sequential test accepts a wrong hypothesis, $H_{\ell'}$, $\ell' \neq \ell_*$, and it happens at s-th attempt, and let us lead this assumption to a contradiction. Observe that under our assumption the test $\mathcal{T}_s$ as applied to observation $\omega^{K(s)}$ does accept some hypothesis G^s_i, but does *not* accept the true hypothesis $G^s_{2\ell_*-1} = H_{\ell_*}$ (indeed, assuming the latter hypothesis to be accepted, its color, which is ℓ_*, should be the same as the color ℓ' of G^s_i (we are in the case when the sequential test accepts $H_{\ell'}$ at s-th attempt!); since in fact $\ell' \neq \ell_*$, the above assumption leads to a contradiction). On the other hand, we are in the case when E_s does not take place, that is, $\mathcal{T}_s$ does accept the true hypothesis $G^s_{2\ell_*-1}$, and we arrive at the desired contradiction.

(a) and (b) provide us with a verification we were looking for.

Discussion and illustration. It can be easily seen that when $\epsilon_s = \epsilon/S$ for all s, the worst-case duration $K(S)$ of our sequential test is within a logarithmic in SL factor of the duration of any other test capable to decide on our L hypotheses with risk ϵ. At the same time it is easily seen that when the distribution μ of our observation is "deeply inside" some set M_ℓ, specifically, $\mu \in M^s_\ell$ for some $s \in \mathcal{S}$, $s < S$, then the μ-probability to terminate not later than after just $K(s)$ realizations ω_k of $\omega \sim \mu$ are observed and to infer correctly what is the true hypothesis is at least $1 - \epsilon$. Informally speaking, in the case of "landslide" elections, a reliable prediction of elections' outcome will be made after a relatively small number of respondents are interviewed.

Indeed, let $s \in \mathcal{S}$ and $\omega_k \sim \mu \in M^s_\ell$, so that μ obeys the hypothesis $G^s_{2\ell}$. Consider the s events E_t, $1 \leq t \leq s$, defined as follows:

- For $t < s$, E_t occurs when the sequential test terminates at attempt t with accepting, instead of H_ℓ, wrong hypothesis $H_{\ell'}$, $\ell' \neq \ell$. Note that E_t can take place only when $\mathcal{T}_t$ does not accept the true hypothesis $G^s_{2\ell} = H^s_\ell$ (why?), and μ-probability of this outcome is $\leq \epsilon_t$.
- E_s occurs when $\mathcal{T}_s$ does not accept the true hypothesis $G^s_{2\ell}$ or accepts it along with some hypothesis G^s_j, $1 \leq j \leq 2L$, of color different from ℓ. Note that we are in the situation where the hypothesis $G^s_{2\ell}$ is true, and, by construction of $\mathcal{C}_s$, all hypotheses $\mathcal{C}_s$-close to $G^s_{2\ell}$ are of the same color ℓ as $G^s_{2\ell}$. Recalling what $\mathcal{C}_s$-risk is and that the $\mathcal{C}_s$-risk of $\mathcal{T}_s$ is $\leq \epsilon_s$, we conclude that the μ-probability of E_s is at most ϵ_s.

The bottom line is that μ-probability of the event $\bigcup_{t\leq s} E_t$ is at most $\sum_{t=1}^{s}\epsilon_t \leq \epsilon$; by construction of the sequential test, if the event $\bigcup_{t\leq s} E_t$ does *not* take place, the test terminates in course of the first s attempts with accepting the correct hypothesis H_ℓ. Our claim is justified.

Numerical illustration. To get an impression of the "power" of sequential hypothesis testing, here are the data on the durations of non-sequential and sequential tests with risk $\epsilon = 0.01$ for various values of δ; in the sequential tests, $\theta = 10^{-1/4}$ is used. The worst-case data for 2-candidate and 5-candidate elections are as follows (below "volume" stands for the number of observations used by test)

δ	0.5623	0.3162	0.1778	0.1000	0.0562	0.0316	0.0177	0.0100
$K, L = 2$	25	88	287	917	2908	9206	29118	92098
S & $K(S), L = 2$	1&25	2&152	3&499	4&1594	5&5056	6&16005	7&50624	8&160118
$K, L = 5$	32	114	373	1193	3784	11977	37885	119745
S & $K(S), L = 5$	1&32	2&179	3&585	4&1870	5&5931	6&18776	7&59391	8&187720

Volume of non-sequential test (K), number of stages (S) and worst-case volume ($K(S)$) of sequential test as functions of threshold $\delta = \delta_S$. Risk ϵ is set to 0.01.

As it should be, the worst-case volume of sequential test is essentially worse than the volume of the non-sequential test[17]. This being said, let us look what happens in the "average," rather than the worst, case, specifically, let us look at the empirical distribution of the volume when the distribution μ of observations is selected in the L-dimensional probabilistic simplex $\mathbf{\Delta}_L = \{\mu \in \mathbf{R}^L : \mu \geq 0, \sum_\ell \mu_\ell = 1\}$ at random. Here is the empirical statistics of test volume obtained when drawing μ from the uniform distribution on $\bigcup_{\ell\leq L} M_\ell$ and running the sequential test[18] on observations drawn from the selected μ:

L	risk	median	mean	50%	55%	60%	65%	70%	75%	80%	85%	90%	95%	100%
2	0.0010	177	9182	177	177	177	397	617	617	1223	1829	8766	87911	160118
5	0.0040	1449	18564	1449	2175	2175	4189	6204	12704	19205	39993	60781	124249	187718

The data on empirical risk (column "risk") and volume (columns "median...100%") of Sequential test Column "X%": empirical X%-quantile of test volume.

The data in the table are obtained from 1,000 experiments. We see that with the Sequential test, "typical" numbers of observations before termination are much less than the worst-case values of these numbers. For example, in as much as 80% of experiments these numbers were below quite reasonable levels, at least in the case $L = 2$. Of course, what is "typical," and what is not, depends on how we generate μ's (scientifically speaking, this is called "prior Bayesian distribution"); were our generation more likely to produce "close run" distributions, the advantages of sequential decision making would be reduced. This ambiguity is, however, unavoidable when attempting to go beyond worst-case-oriented analysis.

2.6.3 Concluding remarks

Application of our machinery to sequential hypothesis testing is in no sense restricted to the simple election model considered so far. A natural general setup we can handle is as follows:

> We are given a simple observation scheme $\mathcal{O}$ and a number L of related convex hypotheses, colored in d colors, on the distribution of an observation, with distributions obeying hypotheses of different colors being distinct from each other. Given risk level ϵ, we want to infer $(1-\epsilon)$-reliably the color of the distribution underlying observations

[17] the reason is twofold: first, for $s < S$ we pass from deciding on L hypotheses to deciding on $2L$ of them; second, the desired risk ϵ is now distributed among several tests, so that each of them should be more reliable than the non-sequential test with risk ϵ.

[18] corresponding to $\delta = 0.01$, $\theta = 10^{-1/4}$ and $\epsilon = 0.01$

(i.e., the color of the hypothesis obeyed by this distribution) from stationary K-repeated observations, utilizing as small number of observations as possible.

For detailed description of our related constructions and results, an interested reader is referred to [94].

2.7 Measurement Design in simple observation schemes

2.7.1 Motivation: Opinion Polls revisited

Consider the same situation as in Section 2.6.1 – we want to use opinion poll to predict the winner in a population-wide elections with L candidates. When addressing this situation earlier, no essential a priori information on the distribution of voters' preferences was available. Now consider the case when the population is split into I groups (according to age, sex, income, etc., etc.), with i-th group forming fraction θ_i of the entire population, and we have at our disposal, at least for some i, a nontrivial a priori information about the distribution p^i of the preferences across group # i (ℓ-th entry p^i_ℓ in p^i is the fraction of voters of group i voting for candidate ℓ). For example, we could know in advance that at least 90% of members of group #1 vote for candidate #1, and at least 85% of members of group #2 vote for candidate #2; no information of this type for group #3 is available. In this situation it would be wise to select respondents in the poll via two-stage procedure, first – selecting at random, with probabilities $q_1, ..., q_I$, the group from which the next respondent will be picked, and second – selecting the respondent from this group at random according to the uniform distribution on the group. When q_i are proportional to the sizes of the groups (i.e., $q_i = \theta_i$ for all i), we come back to selecting respondents at random from the uniform distribution over the entire population; the point, however, is that in the presence of a priori information, it makes sense to use q_i different from θ_i, specifically, to make the ratios q_i/θ_i "large" or "small" depending on whether a priori information on group #i is poor or rich.

The story we just have told is an example of situation when we can "design measurements" – draw observations from a distribution which partly is under our control. Indeed, what in fact happens in the story, is the following. "In the nature" there exist I probabilistic vectors $p^1, ..., p^I$ of dimension L representing distributions of voting preferences within the corresponding groups; the distribution of preferences across the entire population is $p = \sum_i \theta_i p^i$. With two-stage selection of respondents, the outcome of a particular interview becomes a pair (i, ℓ), with i identifying the group to which the respondent belongs, and ℓ identifying the candidate preferred by this respondent. In subsequent interviews, the pairs (i, ℓ) – these are our observations – are drawn, independently of each other, from the probability distribution on the pairs (i, ℓ), $i \leq I$, $\ell \leq L$, with the probability of an outcome (i, ℓ) equal to

$$p(i, \ell) = q_i p^i_\ell.$$

Thus, we find ourselves in the situation of stationary repeated observations stemming from the Discrete o.s. with observation space Ω of cardinality IL; the distribution from which the observations are drawn is a probabilistic vector μ of the form

$$\mu = Ax,$$

where

- $x = [p^1; ...; p^I]$ is the "signal" underlying our observations and representing the preferences of the population; this signal is selected by the nature in the known to us set $\mathcal{X}$ defined in terms of our a priori information on $p^1, ..., p^I$:

$$\mathcal{X} = \{x = [x^1; ...; x^I] : x^i \in \Pi_i, 1 \leq i \leq I\}, \tag{2.7.1}$$

where Π_i are the sets, given by our a priori information, of possible values of the preference vectors p^i of the voters from i-th group. In the sequel, we assume that Π_i are convex compact subsets in the positive part $\mathbf{\Delta}_L^o = \{p \in \mathbf{R}^L : p > 0, \sum_\ell p_\ell = 1\}$ of the L-dimensional probabilistic simplex;

- A is "sensing matrix" which, to some extent, is under our control; specifically,
$$A[x^1; ...; x^I] = [q_1 x^1; q_2 x^2; ...; q_I x^I], \tag{2.7.2}$$
with $q = [q_1; ...; q_I]$ fully controlled by us (up to the fact that q must be a probabilistic vector).

Note that in the situation under consideration the hypotheses we want to decide upon can be represented by convex sets *in the space of signals,* with particular hypothesis stating that the observations stem from a distribution μ on Ω, with μ belonging to the image of some convex compact set $X_\ell \subset \mathcal{X}$ under the mapping $x \mapsto \mu = Ax$. For example, the hypotheses

$$H_\ell : \mu \in M_\ell = \{\mu \in \mathbf{R}^L : \sum_i \mu_i = 1, \mu_i \geq \frac{1}{N}, \mu_\ell \geq \mu'_\ell + \delta, \ell' \neq \ell\}, \, 1 \leq \ell \leq L$$

considered in Section 2.6.1 can be expressed in terms of the signal $x = [x^1; ...; x^I]$:

$$H_\ell : \mu = Ax, \, x \in X_\ell = \left\{ x = [x^1; ...; x^I] : \begin{array}{c} x^i \geq 0, \sum_\ell x^i_\ell = 1 \forall i \leq I \\ \sum_i \theta_i x^i_\ell \geq \sum_i \theta_i x^i_{\ell'} + \delta \, \forall (\ell' \neq \ell) \\ \sum_i \theta_i x^i_j \geq \frac{1}{N}, \forall j \end{array} \right\}. \tag{2.7.3}$$

The challenge we intend to address is as follows: so far, we were interested in inferences from observations drawn from distributions selected "by nature." Now our goal is to make inferences from observations drawn from a distribution selected partly by the nature and partly by us: the nature selects the signal x, we select from some set matrix A, and the observations are drawn from the distribution Ax. As a result, we arrive at a completely new for us question: how to utilize the freedom in selecting A in order to improve our inferences (this is somehow similar to what in statistics is called "design of experiments.")

2.7.2 Measurement Design: SetUp

In what follows we address measurement design in simple observation schemes, and our setup is as follows (to make our intensions transparent, we illustrate our general setup by explaining how it should be specified to cover the outlined two-stage Design of Opinion Polls – DOP for short).

Given are

- simple observation scheme $\mathcal{O} = (\Omega, \Pi; \{p_\mu : \mu \in \mathcal{M}\}; \mathcal{F})$, specifically, Gaussian, Poisson or Discrete one, with $\mathcal{M} \subset \mathbf{R}^d$.
  In DOP, $\mathcal{O}$ is the Discrete o.s. with $\Omega = \{(i, \ell) : 1 \leq i \leq I, 1 \leq \ell \leq L\}$, that is, points of Ω are the potential outcomes "reference group, preferred candidate" of individual interviews.
- a nonempty closed convex *signal space* $\mathcal{X} \subset \mathbf{R}^n$, along with L nonempty convex compact subsets X_ℓ of $\mathcal{X}$, $\ell = 1, ..., L$.
  In DOP, $\mathcal{X}$ is the set (2.7.1) comprised by tuples of allowed distributions of voters' preferences from various groups, and X_ℓ are the sets (2.7.3) of signals associated with the hypotheses H_ℓ we intend to decide upon.
- a nonempty convex compact set $\mathcal{Q}$ in some $\mathbf{R}^N$ along with a continuous mapping $q \mapsto A_q$ acting from $\mathcal{Q}$ into the space of $d \times n$ matrices such that
$$\forall (x \in \mathcal{X}, q \in \mathcal{Q}) : A_q x \in \mathcal{M}. \tag{2.7.4}$$

In DOP, $\mathcal{Q}$ is the set of probabilistic vectors $q = [q_1; ...; q_I]$ specifying our measurement design, and A_q is the matrix of the mapping (2.7.2).

- a closeness $\mathcal{C}$ on the set $\{1, ..., L\}$ (that is, a set $\mathcal{C}$ of pairs (i, j) with $1 \leq i, j \leq L$ such that $(i, i) \in \mathcal{C}$ for all $i \leq L$ and $(j, i) \in \mathcal{C}$ whenever $(i, j) \in \mathcal{C}$), and a positive integer K.
  In DOP, the closeness $\mathcal{S}$ is as strict as it could be – i is close to j if and only if $i = j$ [19], and K is the total number of interviews in the poll.

We can associate with $q \in \mathcal{Q}$ and every one of X_ℓ, $\ell \leq L$, nonempty convex compact sets M_ℓ^q in the space $\mathcal{M}$:

$$M_\ell^q = \{A_q x : x \in X_\ell\}$$

and hypotheses H_ℓ^q on K-repeated stationary observations $\omega^K = (\omega_1, ..., \omega_K)$, with H_ℓ^q stating that ω_k, $k = 1, ..., K$, are drawn, independently of each other, from a distribution $\mu \in M_\ell^q$, $\ell = 1, ..., L$. Closeness $\mathcal{C}$ can be thought of as closeness on the collection of hypotheses $H_1^q, H_2^q, ..., H_L^q$. Given $q \in \mathcal{Q}$, we can use the construction from Section 2.5.2 in order to build the test $\mathcal{T}_{\phi_*}^K$ deciding on the hypotheses H_ℓ^q up to closeness $\mathcal{C}$, the $\mathcal{C}$-risk of the test being the smallest allowed by the construction. Note that this $\mathcal{C}$-risk depends on q; the "Measurement Design" (MD for short) problem we are about to consider is to select $q \in \mathcal{Q}$ which minimizes the $\mathcal{C}$-risk of the associated test $\mathcal{T}_{\phi_*}^K$.

2.7.3 Formulating the MD problem

By Proposition 2.5.5, the $\mathcal{C}$-risk of the test $\mathcal{T}_{\phi_*}^K$ is upper-bounded by the spectral norm of the symmetric entrywise nonnegative $L \times L$ matrix

$$E^{(K)}(q) = [\epsilon_{\ell\ell'}(q)]_{\ell,\ell'},$$

and this is what we intend to minimize in our MD problem. In the above formula, $\epsilon_{\ell\ell'}(q) = \epsilon_{\ell'\ell}(q)$ are zeros when $(\ell, \ell') \in \mathcal{C}$; when $(\ell, \ell') \notin \mathcal{C}$ and $1 \leq \ell < \ell' \leq L$, the quantities $\epsilon_{\ell\ell'}(q) = \epsilon_{\ell'\ell}(q)$ are defined depending on what is the simple o.s. $\mathcal{O}$. Specifically,

- In the case of *Gaussian* observation scheme (see Section 2.4.5.1), restriction (2.7.4) does not restrict the dependence A_q on q at all (modulo the default restriction that A_q is a continuous in $q \in \mathcal{Q}$ $d \times n$ matrix), and

  $$\epsilon_{\ell\ell'}(q) = \exp\{K\mathrm{Opt}_{\ell\ell'}(q)\}$$

  where

  $$\mathrm{Opt}_{\ell\ell'}(q) = \max_{x \in X_\ell, y \in X_{\ell'}} -[A_q(x-y)]^T \Theta^{-1}[A_q(x-y)] \qquad (G_q)$$

  and Θ is the common covariance matrix of the Gaussian densities forming the family $\{p_\mu : \mu \in \mathcal{M}\}$;

- In the case of Poisson o.s. (see Section 2.4.5.2), restriction (2.7.4) requires from $A_q x$ to be positive vector whenever $q \in \mathcal{Q}$ and $x \in \mathcal{X}$, and

  $$\epsilon_{\ell\ell'}(q) = \exp\{K\mathrm{Opt}_{\ell\ell'}(q)\},$$

  where

  $$\mathrm{Opt}_{\ell\ell'}(q) = \max_{x \in X_\ell, y \in X_{\ell'}} \left[\sum_i \sqrt{[A_q x]_i [A_q y]_i} - \frac{1}{2}\sum_i [A_q x]_i - \frac{1}{2}\sum_i [A_q y]_i\right]; \qquad (P_q)$$

[19] this closeness makes sense when the goal of the poll is to predict the winner; less ambitious goal, like to decide whether the winner will or will not belong to a particular set of candidates, would require weaker closeness.

- In the case of Discrete o.s. (see Section 2.4.5.3), restriction (2.7.4) requires from A_qx to be a positive probabilistic vector whenever $q \in \mathcal{Q}$ and $x \in \mathcal{X}$, and
$$\epsilon_{\ell\ell'}(q) = [\mathrm{Opt}_{\ell\ell'}(q)]^K,$$
where
$$\mathrm{Opt}_{\ell\ell'}(q) = \max_{x\in X_\ell, y\in X_{\ell'}} \sum_i \sqrt{[A_qx]_i[A_qy]_i}. \tag{D_q}$$

The summary of above observations is as follows. The norm $\|E^{(K)}\|_{2,2}$ – the quantity we are interested to minimize in $q \in \mathcal{Q}$ – as a function of $q \in \mathcal{Q}$ is of the form

$$\Psi(q) = \psi(\underbrace{\{\mathrm{Opt}_{\ell\ell'}(q) : (\ell,\ell') \notin \mathcal{C}\}}_{\overline{\mathrm{Opt}}(q)}) \tag{2.7.5}$$

where the outer function ψ is real-valued convex and nondecreasing in every one of its arguments explicitly given function on $\mathbf{R}^N$ (N is the cardinality of the set of pairs (ℓ,ℓ'), $1 \leq \ell,\ell' \leq L$, with $(\ell,\ell') \notin \mathcal{C}$). Indeed, denoting by $\Gamma(S)$ the spectral norm of $d \times d$ matrix S, note that Γ is convex function of S, and this function is nondecreasing in every one of the entries of S, provided that S is restricted to be entrywise nonnegative[20]. $\psi(\cdot)$ is obtained from $\Gamma(S)$ by substitution, instead of entries $S_{\ell\ell'}$ of S, everywhere convex, nonnegative and nondecreasing functions of new variables $\vec{z} = \{z_{\ell\ell'} : (\ell,\ell') \notin \mathcal{C}, 1 \leq \ell,\ell' \leq L\}$, specifically

- when $(\ell,\ell') \in \mathcal{C}$, we set $S_{\ell\ell'}$ to zero;
- when $(\ell,\ell') \notin \mathcal{C}$, we set $S_{\ell\ell'} = \exp\{Kz_{\ell\ell'}\}$ in the case of Gaussian and Poisson o.s.'s, and set $S_{\ell\ell'} = \max[0, z_{\ell\ell'}]^K$, in the case of Discrete o.s.

As a result, we indeed get a convex and nondecreasing in every one of its arguments function ψ of $\vec{z} \in \mathbf{R}^N$.

Now, the Measurement Design problem we want to solve reads

$$\mathrm{Opt} = \min_{q\in\mathcal{Q}} \psi(\overline{\mathrm{Opt}}(q)); \tag{2.7.6}$$

As we remember, the entries in the inner function $\overline{\mathrm{Opt}}(q)$ are optimal values of solvable *convex* optimization problems and as such are efficiently computable. When these entries are also *convex* functions of $q \in \mathcal{Q}$, the objective in (2.7.6), due to the already established convexity and monotonicity properties of ψ, is a convex function of q, meaning that (2.7.6) is a convex and thus efficiently solvable problem. On the other hand, when some of the entries in $\overline{\mathrm{Opt}}(q)$ are nonconvex in q, we hardly could expect (2.7.6) to be an easy-to-solve problem. Unfortunately, convexity of the entries in $\overline{\mathrm{Opt}}(q)$ in q turns out to be a "rare commodity." For example, we can verify by inspection that the objectives in (G_q), (P_q), (D_q) *as a functions of* A_q (not of q!) are *concave* rather than convex, so that the optimal values in the problems, as a functions of q, are maxima, over the parameters, of parametric families of concave functions of A_q (the parameter in these parametric families are the optimization variables in (G_q) – (D_q)) and as such *as a functions of* A_q hardly are convex. And indeed, as a matter of fact, the MD problem usually is nonconvex and difficult to solve. We intend to consider "Simple case" where this difficulty does not arise, specifically, the case where the

[20] monotonicity follows from the fact that for an entrywise nonnegative S, we have
$$\|S\|_{2,2} = \max_{x,y}\{x^TSy : \|x\|_2 \leq 1, \|y\|_2 \leq 1\} = \max_{x,y}\{x^TSy : \|x\|_2 \leq 1, \|y\|_2 \leq 1, x \geq 0, y \geq 0\}$$

objectives of the optimization problems specifying $\mathrm{Opt}_{\ell\ell'}(q)$ are *affine in* q; in this case, $\mathrm{Opt}_{\ell\ell'}(q)$ as a function of q is the maximum, over the parameters (optimization variables in the corresponding problems), of parametric families of affine functions of q and as such is convex.

Our current goal is to understand what our sufficient condition for tractability of the MD problem – affinity in q of the objectives in the respective problems (G_q), (P_q), (D_q) – actually means, and to show that this, by itself quite restrictive, assumption indeed takes place in some important applications.

2.7.3.1 Simple case, Discrete o.s.

Looking at the optimization problem (D_q), we see that the simplest way to ensure that its objective is affine in q is to assume that

$$A_q = \mathrm{Diag}\{Bq\}A, \tag{2.7.7}$$

where A is some fixed $d \times n$ matrix, and B is some fixed $d \times (\dim q)$ matrix such that Bq is positive whenever $q \in \mathcal{Q}$. On the top of this, we should ensure that when $q \in \mathcal{Q}$ and $x \in \mathcal{X}$, $A_q x$ is a positive probabilistic vector; this amounts to some restrictions linking $\mathcal{Q}$, $\mathcal{X}$, A, and B.

Illustration. An instructive example of the Simple case of Measurement Design in Discrete o.s. is the "Opinion Poll" problem with a priori information presented in Section 2.7.1: the voting population is split into I groups, with i-th group constituting fraction θ_i of the entire population. In i-th group, the distribution of voters' preferences is represented by unknown L-dimensional probabilistic vector $x^i = [x^i_1; ...; x^i_L]$ (L is the number of candidates, x^i_ℓ is the fraction of voters in i-th group intending to vote for ℓ-th candidate), known to belong to a given convex compact subset Π_i of the "positive part" $\mathbf{\Delta}^o_L = \{x \in \mathbf{R}^L : x > 0, \sum_\ell x_\ell = 1\}$ of the L-dimensional probabilistic simplex. We are given threshold $\delta > 0$ and want to decide on L hypotheses H_1,..., H_L, with H_ℓ stating that the population-wide vector $y = \sum_{i=1}^I \theta_i x^i$ of voters' preferences belongs to the closed convex set

$$Y_\ell = \{y = \sum_{i=1}^I \theta_i x^i : x^i \in \Pi_i,\, 1 \leq i \leq I, y_\ell \geq y_{\ell'} + \delta,\, \forall(\ell' \neq \ell)\};$$

note that Y_ℓ is the image, under the linear mapping

$$[x^1; ...; x^I] \mapsto y(x) = \sum_i \theta_i x^i$$

of the compact convex set

$$X_\ell = \{x = [x^1; ...; x^I] : x^i \in \Pi_i,\, 1 \leq i \leq I, y_\ell(x) \geq y_{\ell'}(x) + \delta,\, \forall(\ell' \neq \ell)\}$$

which is a subset of the convex compact set

$$\mathcal{X} = \{x = [x^1; ...; x^I] : x^i \in \Pi_i,\, 1 \leq i \leq I\}.$$

k-th poll interview is organized as follows:

> We draw at random a group among the I groups of voters, with probability q_i to draw i-th group, and then draw at random, from the uniform distribution on the group, the respondent to be interviewed. The outcome of the interview – our observation ω_k – is the pair (i, ℓ), where i is the group to which the respondent belongs, and ℓ is the candidate preferred by the respondent.

This results in a sensing matrix A_q, see (2.7.2), which is in the form of (2.7.7), specifically,

$$A_q = \text{Diag}\{q_1 I_L, q_2 I_L, ..., q_I I_L\} \qquad [q \in \mathbf{\Delta}_I]$$

the outcome of k-th interview is drawn at random from the discrete probability distribution $A_q x$, where $x \in \mathcal{X}$ is the "signal" summarizing voters' preferences in the groups.

Given total number of observations K, our goal is to decide with a given risk ϵ on our L hypotheses; whether this goal is or is not achievable, it depends on K and on A_q. What we want, is to find q for which the above goal is achievable with as small K as possible; in the case in question, this reduces to solving, for various trial values of K, problem (2.7.6), which under the circumstances is an explicit *convex* optimization problem.

To get an impression of the potential of Measurement Design, we present a sample of numerical results. In all reported experiments, we used $\delta = 0.05$ and $\epsilon = 0.01$. The sets Π_i, $1 \leq i \leq I$, were generated as follows: we pick at random a probabilistic vector $\bar{p}^i$ of dimension L, and Π_i was the intersection of the box $\{p : \bar{p}_\ell - u_i \leq p_\ell \leq \bar{p}_\ell + u_i\}$ centered at $\bar{p}$ with the probabilistic simplex $\mathbf{\Delta}_L$, where u_i, $i = 1, ..., I$, are prescribed "uncertainty levels;" note that uncertainty level $u_i \geq 1$ is the same as absence of any a priori information on the preferences of voters from i-th group.

The results of our numerical experiments are as follows:

L	I	Uncertainty levels u	Group sizes θ	$K_{\rm ini}$	$q_{\rm opt}$	$K_{\rm opt}$
2	2	$[0.03; 1.00]$	$[0.500; 0.500]$	1212	$[0.437; 0.563]$	1194
2	2	$[0.02; 1.00]$	$[0.500; 0.500]$	2699	$[0.000; 1.000]$	1948
3	3	$[0.02; 0.03; 1.00]$	$[0.333; 0.333; 0.333]$	3177	$[0.000; 0.455; 0.545]$	2726
5	4	$[0.02; 0.02; 0.03; 1.00]$	$[0.250; 0.250; 0.250; 0.250]$	2556	$[0.000; 0.131; 0.322; 0.547]$	2086
5	4	$[1.00; 1.00; 1.00; 1.00]$	$[0.250; 0.250; 0.250; 0.250]$	4788	$[0.250; 0.250; 0.250; 0.250]$	4788

Effect of measurement design. $K_{\rm ini}$ and $K_{\rm opt}$ are the poll sizes required for 0.99-reliable prediction of the winner when $q = \theta$ and $q = q_{\rm opt}$, respectively.

We see that measurement design allows to reduce (for some data – quite significantly) the volume of observations as compared to the straightforward selecting the respondents uniformly across the entire population. To compare our current model and results with those from Section 2.6.1, note that now we have more a priori information on the true distribution of voting preferences due to some a priori knowledge of preferences within groups, which allows to reduce the poll sizes with both straightforward and optimal measurement design[21]. The differences between $K_{\rm ini}$ and $K_{\rm opt}$ is fully due to measurement design.

Comparative drug study. A related to DOP and perhaps more interesting Simple case of the Measurement Design in Discrete o.s. is as follows. Let us speak about L competing drugs rather than L competing candidates running for an office, and population of patients the drugs are aimed to help rather than population of voters. For the sake of simplicity, assume that when a particular drug is administered to a particular patient, the outcome is binary: (positive) "effect" or "no effect" (what follows can be easily extended to the case of non-binary categorial outcomes, like "strong positive effect," "weak positive effect," "negative effect," and alike). Our goal is to organize a clinical study in order to make inferences on comparative drug efficiency, measured by the percentage of patients on which a particular drug has effect. The difference with organizing opinion poll is that now we cannot just ask a respondent what are his/her preferences; we are supposed to administer to a participant of the study a single drug on our choice and to look at the result.

As in the DOP problem, we assume that the population of patients is split into I groups, with i-th group comprising fraction θ_i of the entire population.

[21] To illustrate this point, look at the last two lines in the table: utilizing a priori information allows to reduce the poll size from 4788 to 2556 even with the straightforward measurement design.

We model the situation as follows. We associate with a patient Boolean vector of dimension $2L$, with ℓ-th entry in the vector equal to 1 or 0 depending on whether drug # ℓ has effect on the patient, and the $(L+\ell)$-th entry complementing the ℓ-th one to 1 (that is, if ℓ-th entry is χ, then $(L+\ell)$-th entry is $1-\chi$). Let x^i be the average of these vectors over patients from group i. We define "signal" x underlying our measurements as the vector $[x^1;...;x^I]$ and assume that our a priori information allows to localize x in a closed convex subset $\mathcal{X}$ of the set

$$\mathcal{Y}=\{x=[x^1;...;x^I]:x^i\geq 0, x^i_\ell+x^i_{L+\ell}=1,\, 1\leq i\leq I, 1\leq \ell\leq L\}$$

to which all our signals belong by construction. Note that the vector

$$y=Bx=\sum_i \theta_i x^i$$

can be treated as "population-wise distribution of drug effects:" y_ℓ, $\ell\leq L$, is the fraction, in the entire population of patients, of those patients on whom drug ℓ has effect, and $y_{L+\ell}=1-y_\ell$. As a result, typical hypotheses related to comparison of the drugs, like "drug ℓ has effect on a larger, at least by margin δ, percentage of patients than drug ℓ'," become convex hypotheses on the signal x. In order to test hypotheses of this type, we can use two-stage procedure for observing drug effects, namely, as follows.

To get a particular observation, we select at random, with probability $q_{i\ell}$, pair (i,ℓ) from the set $\{(i,\ell):1\leq i\leq I, 1\leq \ell\leq L\}$, select a patient from group i according to the uniform distribution on the group, administer the patient drug ℓ and check whether the drug has effect on the patient. Thus, a single observation is a triple (i,ℓ,χ), where $\chi=0$ when the administered drug has no effect on the patient, and $\chi=1$ otherwise. The probability to get observation $(i,\ell,1)$ is $q_{i\ell}x^i_\ell$, and the probability to get observation $(i,\ell,0)$ is $q_{i\ell}x^i_{L+\ell}$. Thus, we arrive at the Discrete o.s. where the distribution μ of observations is of the form $\mu=A_qx$, with the rows in A_q indexed by triples $\omega=(i,\ell,\chi)\in\Omega:=\{1,2,...,I\}\times\{1,2,...,L\}\times\{0,1\}$ and given by

$$(A_q[x^1;...;x^I])_{i,\ell,\chi}=\left\{\begin{array}{ll} q_{i\ell}x^i_\ell, & \chi=1\\ q_{i\ell}x^i_{L+\ell}, & \chi=0\end{array}\right.$$

Specifying the set $\mathcal{Q}$ of allowed measurement designs q as a closed convex subset of the set of all non-vanishing discrete probability distributions on the set $\{1,2,...,I\}\times\{1,2,...,L\}$, we find ourselves in the Simple case, as defined by (2.7.7), of Discrete o.s., and A_qx is a probabilistic vector whenever $q\in\mathcal{Q}$ and $x\in\mathcal{Y}$.

2.7.3.2 Simple case, Poisson o.s.

Looking at the optimization problem (P_q), we see that the simplest way to ensure that its objective is, same as in the case of Discrete o.s., to assume that

$$A_q=\mathrm{Diag}\{Bq\}A,$$

where A is some fixed $d\times n$ matrix, and B is some fixed $d\times(\dim q)$ matrix such that Bq is positive whenever $q\in\mathcal{Q}$. On the top of this, we should ensure that when $q\in\mathcal{Q}$ and $x\in\mathcal{X}$, A_qx is a positive vector; this amounts to some restrictions linking $\mathcal{Q}$, $\mathcal{X}$, A, and B.

Application Example: PET with time control. Positron Emission Tomography was already mentioned, as an example of Poisson o.s., in Section 2.4.3.2. As explained in the latter Section, in PET we observe a random vector $\omega\in\mathbf{R}^d$ with independent entries $[\omega]_i\sim\mathrm{Poisson}(\mu_i)$, $1\leq i\leq d$, where the vector of parameters $\mu=[\mu_1;...\mu_d]$ of the Poisson distributions is the linear image $\mu=A\lambda$ of unknown "signal" (tracer's density in patient's body) λ belonging to some known subset Λ of

$\mathbf{R}^D_+$, with entrywise nonnegative matrix A; our goal is to make inferences about λ. Now, in actual PET scan, patient's position w.r.t. the scanner is not the same during the entire study; the position is kept fixed within i-th time period, $1 \leq i \leq I$, and changes from period to period in order to expose to the scanner the entire "area of interest"

For example, with the scanner shown on the picture, during PET study the imaging table with the patient will be shifted several times along the axis of the scanning ring. As a result, observed vector ω can be split into blocks ω^i, $i = 1, ..., I$, of data acquired during i-th period, $1 \leq i \leq I$; on the closest inspection, the corresponding block μ^i in μ is

$$\mu^i = q_i A_i \lambda,$$

where A_i is a known in advance entrywise nonnegative matrix, and q_i is the duration of i-th period. In principle, q_i could be treated as nonnegative design variables subject to the "budget constraint" $\sum_{i=1}^I q_i = T$, where T is the total duration of the study[22], and perhaps some other convex constraints, say, positive lower bounds on q_i. It is immediately seen that the outlined situation is exactly as is required in the Simple case of Poisson o.s.

2.7.3.3 Simple case, Gaussian o.s.

Looking at the optimization problem (G_q), we see that the simplest way to ensure that its objective is affine in q is to assume that the covariance matrix Θ is diagonal, and

$$A_q = \mathrm{Diag}\{\sqrt{q_1}, ..., \sqrt{q_d}\} A \tag{2.7.8}$$

where A is a fixed $d \times n$ matrix, and q runs through a convex compact subset of $\mathbf{R}^d_+$.

It turns out that there are situations where assumption (2.7.8) makes perfect sense. Let us start with preamble. In Gaussian o.s.

$$\begin{array}{c} \omega = Ax + \xi \\ \left[A \in \mathbf{R}^{d\times n}, \xi \sim \mathcal{N}(0, \Sigma), \, \Sigma = \mathrm{Diag}\{\sigma_1^2, ..., \sigma_d^2\}\right] \end{array} \tag{2.7.9}$$

the "physics" behind the observations in many cases is as follows. There are d sensors (receivers), i-th registering continuous time analogous input depending linearly on the underlying observations signal x; on the time horizon on which the measurements are taken, this input is constant in time and is registered by i-th sensor on time interval Δ_i. The deterministic component of the measurement registered by sensor i is the integral of the corresponding input taken over Δ_i, and the stochastic component of the measurement is obtained by integrating over the same interval white Gaussian noise. As far as this noise is concerned, the only thing which matters is that when the white noise affecting i-th sensor is integrated over a time interval Δ, the result is random Gaussian variable

[22] T cannot be too large; aside of other considerations, the tracer disintegrates, and its density can be considered as nearly constant only on a properly restricted time horizon.

with zero mean and variance $\sigma_i^2|\Delta|$ ($|\Delta|$ is the length of Δ), and the random variables obtained by integrating white noise over non-overlapping segments are independent. Besides this, we assume that the noisy components of measurements are independent across the sensors.

Now, there could be two basic versions of the just outlined situation, both leading to the same observation model (2.7.9). In the first, "parallel," version, all d sensors work in parallel on the same time horizon of duration 1. In the second, "sequential," version, the sensors are activated and scanned one by one, each working unit time; thus, here the full time horizon is d, and the sensors are registering their respective inputs on consecutive time intervals of duration 1 each. In this second "physical" version of Gaussian o.s., we can, in principle, allow for sensors to register their inputs on consecutive time segments of varying durations $q_1 \geq 0$, $q_2 \geq 0$,..., $q_d \geq 0$, with the additional to nonnegativity restriction that our total time budget is respected: $\sum_i q_i = d$ (and perhaps with some other convex constraints on q_i). Let us look what is the observation scheme we end up with. Assuming that (2.7.9) represents correctly our observations in the reference case where all $|\Delta_i|$ are equal to 1, the deterministic component of the measurement registered by sensor i in time interval of duration q_i will be $q_i \sum_j a_{ij}x_j$, and the standard deviation of the noisy component will be $\sigma_i\sqrt{q_i}$, so that the measurements become

$$z_i = \sigma_i\sqrt{q_i}\zeta_i + q_i\sum_j a_{ij}x_j,\ i = 1,...,d,$$

with independent of each other standard (zero mean, unit variance) Gaussian noises ζ_i. Now, since we know q_i, we can scale the latter observations by making the standard deviation of the noisy component the same σ_i as in the reference case; specifically, we lose nothing when assuming that our observations are

$$\omega_i = z_i/\sqrt{q_i} = \underbrace{\sigma_i\zeta_i}_{\xi_i} + \sqrt{q_i}\sum_j a_{ij}x_j,$$

or, equivalently,

$$\omega = \xi + \underbrace{\mathrm{Diag}\{\sqrt{q_1},...,\sqrt{q_d}\}A}_{A_q}x,\ \xi \sim \mathcal{N}(0, \mathrm{Diag}\{\sigma_1^2,...,\sigma_d^2\}) \qquad [A = [a_{ij}]]$$

where q is allowed to run through a convex compact subset $\mathcal{Q}$ of the simplex $\{q \in \mathbf{R}^d_+ : \sum_i q_i = d\}$. Thus, if the "physical nature" of a Gaussian o.s. is sequential, then, making, as is natural under the circumstances, the activity times of the sensors our design variables, we arrive at (2.7.8), and, as a result, end up with easy-to-solve Measurements Design problem.

2.8 Affine detectors beyond simple observation schemes

On a closer inspection, the "common denominator" of our basic simple o.s.'s – Gaussian, Poisson and Discrete ones, is that in all these cases the minimal risk detector for a pair of convex hypotheses is *affine*. At the first glance, this indeed is so for the Gaussian and the Poisson o's"s, where $\mathcal{F}$ is comprised of affine functions on the corresponding observation space Ω ($\mathbf{R}^d$ for Gaussian o.s., and $\mathbf{Z}^d_+ \subset \mathbf{R}^d$ for Poisson o.s.), but is *not* so for the Discrete o.s. – in the latter case, $\Omega = \{1,...,d\}$, and $\mathcal{F}$ is comprised of all functions on Ω, while "affine functions on $\Omega = \{1,...,d\}$" merely make no sense. Note, however, that we can encode (and from now on indeed encode) the points $i = 1,...,d$ of d-element set by basic orths $e_i = [0;...;0;1;0;...;0] \in \mathbf{R}^d$ in $\mathbf{R}^d$, thus making our observation space $\Omega = \{1,...,d\}$ a subset of $\mathbf{R}^d$. With this encoding, *every* real valued function on $\{1,...,d\}$ becomes restriction on Ω of an affine function. Note that when passing from our basic simple o.s.'s to their direct products, the minimum risk detectors for pairs of convex hypotheses remain affine.

Now, good in our context news about simple o.s.'s state that

A) the best – with the smallest possible risk – *affine* detector, same as its risk, can be efficiently computed;

B) the smallest risk *affine* detector from A) is the best, in terms of risk, detector available under the circumstances, so that the associated test is near-optimal.

Note that as far as practical applications of the detector-based hypothesis testing are concerned, one "can survive" without B) (near-optimality of the constructed detectors), while A) *is a must.*

In this Section we focus on families of probability distributions obeying A). This class turns out to be incomparably larger than what was defined as simple o.s.'s in Section 2.4; in particular, it includes nonparametric families of distributions. Staying within this much broader class, we still are able to construct in a computationally efficient way the best affine detectors for a pair of "convex", in certain precise sense, hypotheses, along with valid upper bounds on the risks of the detectors. What we, in general, can*not* claim anymore, is that the tests associated with the above detectors are near-optimal. This being said, we believe that investigating possibilities for building tests and quantifying their performance in a computationally friendly manner is of value even when we cannot provably guarantee near-optimality of these tests. The results to follow originate from [95, 99].

2.8.1 Situation

In what follows, we fix *observation space* $\Omega = \mathbf{R}^d$, and let $\mathcal{P}_j$, $1 \leq j \leq J$, be given families of probability distributions on Ω. Put broadly, our goal still is, given a random observation $\omega \sim P$, where $P \in \bigcup\limits_{j\leq J} \mathcal{P}_j$, to decide upon the hypotheses $H_j : P \in \mathcal{P}_j$, $j = 1, ..., J$. We intend to address this goal in the case when the families $\mathcal{P}_j$ are *simple* – they are comprised of distributions for which moment-generating functions admit an explicit upper bound.

2.8.1.1 Preliminaries: Regular data and associated families of distributions

Regular data is defined as a triple $\mathcal{H}, \mathcal{M}, \Phi(\cdot,\cdot)$, where

- $\mathcal{H}$ is a nonempty closed convex set in $\Omega = \mathbf{R}^d$ symmetric w.r.t. the origin,
- $\mathcal{M}$ is a closed convex set in some $\mathbf{R}^n$,
- $\Phi(h;\mu) : \mathcal{H} \times \mathcal{M} \to \mathbf{R}$ is a continuous function convex in $h \in \mathcal{H}$ and concave in $\mu \in \mathcal{M}$.

Regular data $\mathcal{H}, \mathcal{M}, \Phi(\cdot,\cdot)$ define two families of probability distributions on Ω:

- the family of *regular* distributions

$$\mathcal{R} = \mathcal{R}[\mathcal{H}, \mathcal{M}, \Phi]$$

  comprised of all probability distributions P on Ω such that

$$\forall h \in \mathcal{H}\ \exists \mu \in \mathcal{M} : \ln\left(\int_\Omega \exp\{h^T\omega\}P(d\omega)\right) \leq \Phi(h;\mu). \tag{2.8.1}$$

- the family of *simple* distributions

$$\mathcal{S} = \mathcal{S}[\mathcal{H}, \mathcal{M}, \Phi]$$

  comprised of probability distributions P on Ω such that

$$\exists \mu \in \mathcal{M} : \forall h \in \mathcal{H} : \ln\left(\int_\Omega \exp\{h^T\omega\}P(d\omega)\right) \leq \Phi(h;\mu). \tag{2.8.2}$$

Recall that beginning with Section 2.3, the starting point in all our constructions is a "plausibly good" detector-based test which, given two families $\mathcal{P}_1$ and $\mathcal{P}_2$ of distributions with common observation space, and repeated observations $\omega_1, ..., \omega_t$ drawn from a distribution $P \in \mathcal{P}_1 \cup \mathcal{P}_2$, decides whether $P \in \mathcal{P}_1$ or $P \in \mathcal{P}_2$. Our interest in the families of regular/simple distributions stems from the fact that when the families $\mathcal{P}_1$ and $\mathcal{P}_2$ are of this type, building such a test reduces to solving a convex-concave saddle point problem and thus can be carried out in a computationally efficient manner. We postpone the related construction and analysis to Section 2.8.2, and continue with presenting some basic examples of families of simple and regular distributions along with a simple "calculus" of these families.

2.8.1.2 Basic examples of simple families of probability distributions

2.8.1.2.A. Sub-Gaussian distributions: Let $\mathcal{H} = \Omega = \mathbf{R}^d$, $\mathcal{M}$ be a closed convex subset of the set $\mathcal{G}_d = \{\mu = (\theta, \Theta) : \theta \in \mathbf{R}^d, \Theta \in \mathbf{S}^d_+\}$, where $\mathbf{S}^d_+$ is cone of positive semidefinite matrices in the space $\mathbf{S}^d$ of symmetric $d \times d$ matrices, and let

$$\Phi(h; \theta, \Theta) = \theta^T h + \frac{1}{2} h^T \Theta h.$$

In this case, $\mathcal{S}[\mathcal{H}, \mathcal{M}, \Phi]$ contains all *sub-Gaussian* distributions P on $\mathbf{R}^d$ with sub-Gaussianity parameters from $\mathcal{M}$.

> Recall that a distributions P on $\Omega = \mathbf{R}^d$ is called sub-Gaussian with sub-Gaussianity parameters $\theta \in \mathbf{R}^d$ and $\Theta \in \mathbf{S}^d_+$, if
>
> $$\mathbf{E}_{\omega \sim P}\{\exp\{h^T \omega\}\} \leq \exp\{\theta^T h + \frac{1}{2} h^T \Theta h\} \;\; \forall h \in \mathbf{R}^d. \tag{2.8.3}$$
>
> Whenever this is the case, θ is the expected value of P. We shall use the notation $\xi \sim \mathcal{SG}(\theta, \Theta)$ as a shortcut for the sentence "random vector ξ is sub-Gaussian with parameters θ, Θ." It is immediately seen that when $\xi \sim \mathcal{N}(\theta, \Theta)$, we have also $\xi \sim \mathcal{SG}(\theta, \Theta)$, and (2.8.3) in this case is an identity rather than inequality.

In particular, $\mathcal{S}[\mathcal{H}, \mathcal{M}, \Phi]$ contains all Gaussian distributions $\mathcal{N}(\theta, \Theta)$ with $(\theta, \Theta) \in \mathcal{M}$.

2.8.1.2.B. Poisson distributions: Let $\mathcal{H} = \Omega = \mathbf{R}^d$, let $\mathcal{M}$ be a closed convex subset of d-dimensional nonnegative orthant $\mathbf{R}^d_+$, and let

$$\Phi(h = [h_1; ...; h_d]; \mu = [\mu_1; ...; \mu_d]) = \sum_{i=1}^d \mu_i [\exp\{h_i\} - 1] : \mathcal{H} \times \mathcal{M} \to \mathbf{R}.$$

The family $\mathcal{S}[\mathcal{H}, \mathcal{M}, \Phi]$ contains all product-type Poisson distributions Poisson$[\mu]$ with vectors μ of parameters belonging to $\mathcal{M}$; here Poisson$[\mu]$ is the distribution of random d-dimensional vector with independent of each other entries, i-th entry being Poisson random variable with parameter μ_i.

2.8.1.2.C. Discrete distributions. Consider a discrete random variable taking values in d-element set $\{1, 2, ..., d\}$, and let us think of such a variable as of random variable taking values $e_i \in \mathbf{R}^d$, $i = 1, ..., d$, where $e_i = [0; ...; 0; 1; 0; ...; 0]$ (1 in position i) are standard basic orths in $\mathbf{R}^d$. Probability distribution of such a variable can be identified with a point $\mu = [\mu_1; ...; \mu_d]$ from the d-dimensional probabilistic simplex

$$\mathbf{\Delta}_d = \{\nu \in \mathbf{R}^d_+ : \sum_{i=1}^d \nu_i = 1\},$$

where μ_i is the probability for the variable to take value e_i. With these identifications, setting $\mathcal{H}=\mathbf{R}^d$, specifying $\mathcal{M}$ as a closed convex subset of $\mathbf{\Delta}_d$ and setting

$$\Phi(h=[h_1;...;h_d];\mu=[\mu_1;...;\mu_d])=\ln\left(\sum_{i=1}^d \mu_i\exp\{h_i\}\right),$$

the family $\mathcal{S}[\mathcal{H},\mathcal{M},\Phi]$ contains distributions of all discrete random variables taking values in $\{1,...,d\}$ with probabilities $\mu_1,...,\mu_d$ comprising a vector from $\mathcal{M}$.

2.8.1.2.D. Distributions with bounded support. Consider the family $\mathcal{P}[X]$ of probability distributions supported on a closed and bounded convex set $X\subset\Omega=\mathbf{R}^d$, and let

$$\phi_X(h)=\max_{x\in X} h^Tx$$

be the support function of X. We have the following result (to be refined in Section 2.8.1.3):

Proposition 2.8.1 *For every $P\in\mathcal{P}[X]$ it holds*

$$\forall h\in\mathbf{R}^d:\ \ln\left(\int_{\mathbf{R}^d}\exp\{h^T\omega\}P(d\omega)\right)\leq h^Te[P]+\frac{1}{8}\left[\phi_X(h)+\phi_X(-h)\right]^2, \tag{2.8.4}$$

where $e[P]=\int_{\mathbf{R}^d}\omega P(d\omega)$ is the expectation of P, and the right hand side function in (2.8.4) *is convex. As a result, setting*

$$\mathcal{H}=\mathbf{R}^d,\ \mathcal{M}=X,\ \Phi(h;\mu)=h^T\mu+\frac{1}{8}\left[\phi_X(h)+\phi_X(-h)\right]^2,$$

we get regular data such that $\mathcal{P}[X]\subset\mathcal{S}[\mathcal{H},\mathcal{M},\Phi]$.

For proof, see Section 2.8.4.1

2.8.1.3 Calculus of regular and simple families of probability distributions

Families of regular and simple distributions admit "fully algorithmic" calculus, with the main calculus rules as follows.

2.8.1.3.A. Direct summation. For $1\leq\ell\leq L$, let regular data $\mathcal{H}_\ell\subset\Omega_\ell=\mathbf{R}^{d_\ell}$, $\mathcal{M}_\ell\subset\mathbf{R}^{n_\ell}$, $\Phi_\ell(h_\ell;\mu_\ell):\mathcal{H}_\ell\times\mathcal{M}_\ell\to\mathbf{R}$ be given. Let us set

$$\begin{array}{rcl}
\Omega &=& \Omega_1\times...\times\Omega_L=\mathbf{R}^d,\ d=d_1+...+d_L,\\
\mathcal{H} &=& \mathcal{H}_1\times...\times\mathcal{H}_L=\{h=[h^1;...;h^L]:h^\ell\in\mathcal{H}_\ell,\ell\leq L\},\\
\mathcal{M} &=& \mathcal{M}_1\times...\times\mathcal{M}_L=\{\mu=[\mu^1;...;\mu^L]:\mu^\ell\in\mathcal{M}^\ell,\ell\leq L\}\subset\mathbf{R}^n,\ n=n_1+...+n_L,\\
\multicolumn{3}{l}{\Phi(h=[h^1;...;h^L];\mu=[\mu^1;...;\mu^L])=\sum_{\ell=1}^L\Phi_\ell(h^\ell;\mu^\ell):\mathcal{H}\times\mathcal{M}\to\mathbf{R}.}
\end{array}$$

Then $\mathcal{H}$ is a symmetric w.r.t. the origin closed convex set in $\Omega=\mathbf{R}^d$, $\mathcal{M}$ is a nonempty closed convex set in $\mathbf{R}^n$, $\Phi:\ \mathcal{H}\times\mathcal{M}\to\mathbf{R}$ is a continuous convex-concave function, and clearly

- the family $\mathcal{R}[\mathcal{H},\mathcal{M},\Phi]$ contains all product-type distributions $P=P_1\times...\times P_L$ on $\Omega=\Omega_1\times...\times\Omega_L$ with $P_\ell\in\mathcal{R}[\mathcal{H}_\ell,\mathcal{M}_\ell,\Phi_\ell]$, $1\leq\ell\leq L$;
- the family $\mathcal{S}[\mathcal{H},\mathcal{M},\Phi]$ contains all product-type distributions $P=P_1\times...\times P_L$ on $\Omega=\Omega_1\times...\times\Omega_L$ with $P_\ell\in\mathcal{S}[\mathcal{H}_\ell,\mathcal{M}_\ell,\Phi_\ell]$, $1\leq\ell\leq L$.

2.8.1.3.B. Mixing. For $1 \leq \ell \leq L$, let regular data $\mathcal{H}_\ell \subset \Omega = \mathbf{R}^d$, $\mathcal{M}_\ell \subset \mathbf{R}^{n_\ell}$, $\Phi_\ell(h_\ell;\mu_\ell) : \mathcal{H}_\ell \times \mathcal{M}_\ell \to \mathbf{R}$ be given, with compact $\mathcal{M}_\ell$. Let also $\nu = [\nu_1;...;\nu_L]$ be a probabilistic vector. For a tuple $P^L = \{P_\ell \in \mathcal{R}[\mathcal{H}_\ell, \mathcal{M}_\ell, \Phi_\ell]\}_{\ell=1}^L$, let $\Pi[P^L,\nu]$ be the ν-mixture of distributions $P_1,...,P_L$ defined as the distribution of random vector $\omega \sim \Omega$ generated as follows: we draw at random, from probability distribution ν on $\{1,...,L\}$, index ℓ, and then draw ω at random from the distribution P_ℓ. Finally, let $\mathcal{P}$ be the set of all probability distributions on Ω which can be obtained as $\Pi[P^L,\nu]$ from the outlined tuples P^L and vectors ν running through the probabilistic simplex $\mathbf{\Delta}_L = \{\mu \in \mathbf{R}^L : \nu \geq 0, \sum_\ell \nu_\ell = 1\}$.

Let us set

$$\begin{array}{rcl} \mathcal{H} & = & \bigcap\limits_{\ell=1}^{L} \mathcal{H}_\ell, \\ \Psi_\ell(h) & = & \max\limits_{\mu_\ell \in \mathcal{M}_\ell} \Phi_\ell(h;\mu_\ell) : \mathcal{H}_\ell \to \mathbf{R}, \\ \Phi(h;\nu) & = & \ln\left(\sum_{\ell=1}^L \nu_\ell \exp\{\Psi_\ell(h)\}\right) : \mathcal{H} \times \mathbf{\Delta}_L \to \mathbf{R}. \end{array} \tag{2.8.5}$$

Then $\mathcal{H}, \mathbf{\Delta}_L, \Phi$ clearly is a regular data (recall that all $\mathcal{M}_\ell$ are compact sets), and for every $\nu \in \mathbf{\Delta}_L$ and tuple P^L of the above type one has

$$P = \Pi[P^L,\nu] \Rightarrow \ln\left(\int_\Omega \mathrm{e}^{h^T\omega} P(d\omega)\right) \leq \Phi(h;\nu) \;\forall h \in \mathcal{H}, \tag{2.8.6}$$

implying that $\mathcal{P} \subset \mathcal{S}[\mathcal{H}, \mathbf{\Delta}_L, \Phi]$, ν being a parameter of a distribution $P = \Pi[P^L,\nu] \in \mathcal{P}$.

Indeed,(2.8.6) is readily given by the fact that for $P = \Pi[P^L,\nu] \in \mathcal{P}$ and $h \in \mathcal{H}$ it holds

$$\ln\left(\mathbf{E}_{\omega\sim P}\left\{\mathrm{e}^{h^T\omega}\right\}\right) = \ln\left(\sum_{\ell=1}^L \nu_\ell \mathbf{E}_{\omega\sim P_\ell}\{\mathrm{e}^{h^T\omega}\}\right) \leq \ln\left(\sum_{\ell=1}^L \nu_\ell \exp\{\Psi_\ell(h)\}\right) = \Phi(h;\nu),$$

with the concluding inequality given by $h \in \mathcal{H} \subset \mathcal{H}_\ell$ and $P_\ell \in \mathcal{R}[\mathcal{H}_\ell, \mathcal{M}_\ell, \Phi_\ell]$, $1 \leq \ell \leq L$.

We have build a simple family of distributions $\mathcal{S} := \mathcal{S}[\mathcal{H}, \mathbf{\Delta}_L, \Phi]$ which contains all mixtures of distributions from given regular families $\mathcal{R}_\ell := \mathcal{R}[\mathcal{H}_\ell, \mathcal{M}_\ell, \Phi_\ell]$, $1 \leq \ell \leq L$, which makes $\mathcal{S}$ a simple outer approximation of mixtures of distributions from the simple families $\mathcal{S}_\ell := \mathcal{S}[\mathcal{H}_\ell, \mathcal{M}_\ell, \Phi_\ell]$, $1 \leq \ell \leq L$. In this latter capacity, $\mathcal{S}$ has a drawback – the only parameter of the mixture $P = \Pi[P^L,\nu]$ of distributions $P_\ell \in \mathcal{S}_\ell$ is ν, while the parameters of P_ℓ's disappear. In some situations, this makes outer approximation $\mathcal{S}$ of $\mathcal{P}$ too conservative. We are about to get rid, to come extent, of this drawback.

A modification. In the situation described in the beginning of 2.8.1.3.B, let a vector $\bar{\nu} \in \mathbf{\Delta}_L$ be given, and let

$$\bar{\Phi}(h;\mu_1,...,\mu_L) = \sum_{\ell=1}^L \bar{\nu}_\ell \Phi_\ell(h;\mu_\ell) : \mathcal{H} \times (\mathcal{M}_1 \times ... \times \mathcal{M}_L) \to \mathbf{R}. \tag{2.8.7}$$

Let $d \times d$ matrix $Q \succeq 0$ satisfy

$$\left(\Phi_\ell(h;\mu_\ell) - \bar{\Phi}(h;\mu_1,...,\mu_L)\right)^2 \leq h^TQh \;\forall (h \in \mathcal{H}, \ell \leq L, \mu \in \mathcal{M}_1 \times ... \times \mathcal{M}_L), \tag{2.8.8}$$

and let

$$\Phi(h;\mu_1,...,\mu_L) = \frac{3}{5}h^TQh + \bar{\Phi}(h;\mu_1,...,\mu_L) : \mathcal{H} \times (\mathcal{M}_1 \times ... \times \mathcal{M}_L) \to \mathbf{R}. \tag{2.8.9}$$

Φ clearly is convex-concave and continuous on its domain, whence $\mathcal{H} = \bigcap_\ell \mathcal{H}_\ell, \mathcal{M}_1 \times ... \times \mathcal{M}_L, \Phi$ is regular data.

Proposition 2.8.2 *In the just defined situation, denoting by $\mathcal{P}_{\bar\nu}$ the family of all probability distributions $P = \Pi[P^L, \bar\nu]$, stemming from tuples*

$$P^L = \{P_\ell \in \mathcal{S}[\mathcal{H}_\ell, \mathcal{M}_\ell, \Phi_\ell]\}_{\ell=1}^L, \tag{2.8.10}$$

one has

$$\mathcal{P}_{\bar\nu} \subset \mathcal{S}[\mathcal{H}, \mathcal{M}_1 \times ... \times \mathcal{M}_L, \Phi]. \tag{2.8.11}$$

As a parameter of distribution $P = \Pi[P^L, \bar\nu] \in \mathcal{P}_{\bar\nu}$ with P^L as in (2.8.10)*, one can take $\mu^L = [\mu_1; ....; \mu_L]$.*

Proof. It is easily seen that

$$\mathrm{e}^a \leq a + \mathrm{e}^{\frac{3}{5}a^2}, \ \forall a.$$

As a result, when a_ℓ, $\ell = 1, ..., L$, satisfy $\sum_\ell \bar\nu_\ell a_\ell = 0$, we have

$$\sum_\ell \bar\nu_\ell \mathrm{e}^{a_\ell} \leq \sum_\ell \bar\nu_\ell a_\ell + \sum_\ell \bar\nu_\ell \mathrm{e}^{\frac{3}{5}a_\ell^2} \leq \mathrm{e}^{\frac{3}{5}\max_\ell a_\ell^2}. \tag{2.8.12}$$

Now let P^L be as in (2.8.10), and let $h \in \mathcal{H} = \bigcap\limits_L \mathcal{H}_\ell$. Setting $P = \Pi[P^L, \bar\nu]$, we have

$$\begin{array}{l}
\ln\left(\int_\Omega \mathrm{e}^{h^T\omega} P(d\omega)\right) = \ln\left(\sum_\ell \bar\nu_\ell \int_\Omega \mathrm{e}^{h^T\omega} P_\ell(d\omega)\right) = \ln\left(\sum_\ell \bar\nu_\ell \exp\{\Phi_\ell(h, \mu_\ell)\}\right) \\
= \bar\Phi(h; \mu_1, ...\mu_L) + \ln\left(\sum_\ell \bar\nu_\ell \exp\{\Phi_\ell(h, \mu_\ell) - \bar\Phi(h; \mu_1, ...\mu_L)\}\right) \\
\underbrace{\leq}_{a} \bar\Phi(h; \mu_1, ...\mu_L) + \frac{3}{5}\max_\ell[\Phi_\ell(h, \mu_\ell) - \bar\Phi(h; \mu_1, ...\mu_L)]^2 \underbrace{\leq}_{b} \Phi(h; \mu_1, ..., \mu_L),
\end{array}$$

where a is given by (2.8.12) as applied to $a_\ell = \Phi_\ell(h, \mu_\ell) - \bar\Phi(h; \mu_1, ...\mu_L)$, and b is due to (2.8.8), (2.8.9). The resulting inequality, which holds true for all $h \in \mathcal{H}$, is all we need. □

2.8.1.3.C. I.I.D summation. Let $\Omega = \mathbf{R}^d$ be an observation space, $(\mathcal{H}, \mathcal{M}, \Phi)$ be regular data on this space, and let $\lambda = \{\lambda_\ell\}_{\ell=1}^K$ be a collection of reals. We can associate with the outlined entities a new data $(\mathcal{H}_\lambda, \mathcal{M}, \Phi_\lambda)$ on Ω by setting

$$\mathcal{H}_\lambda = \{h \in \Omega : \|\lambda\|_\infty h \in \mathcal{H}\}, \ \ \Phi_\lambda(h; \mu) = \sum_{\ell=1}^L \Phi(\lambda_\ell h; \mu) : \mathcal{H}_\lambda \times \mathcal{M} \to \mathbf{R}.$$

Now, given a probability distribution P on Ω, we can associate with it and with the above λ a new probability distribution P^λ on Ω as follows: P^λ is the distribution of $\sum_\ell \lambda_\ell \omega_\ell$, where $\omega_1, \omega_2, ..., \omega_L$ are drawn, independently of each other, from P. An immediate observation is that the data $(\mathcal{H}_\lambda, \mathcal{M}, \Phi_\lambda)$ is regular, and

- whenever a probability distribution P belongs to $\mathcal{S}[\mathcal{H}, \mathcal{M}, \Phi]$, the distribution P^λ belongs to $\mathcal{S}[\mathcal{H}_\lambda, \mathcal{M}, \Phi_\lambda]$. In particular, when $\omega \sim P \in \mathcal{S}[\mathcal{H}, \mathcal{M}, \Phi]$, then the distribution P^L of the sum of L independent copies of ω belongs to $\mathcal{S}[\mathcal{H}, \mathcal{M}, L\Phi]$.

2.8.1.3.D. Semi-direct summation. For $1 \leq \ell \leq L$, let regular data $\mathcal{H}_\ell \subset \Omega_\ell = \mathbf{R}^{d_\ell}$, $\mathcal{M}_\ell$, Φ_ℓ be given. To avoid complications, we assume that for every ℓ,

- $\mathcal{H}_\ell = \Omega_\ell$,

- $\mathcal{M}_\ell$ is bounded.

Let also an $\epsilon > 0$ be given. We assume that ϵ is small, namely, $L\epsilon < 1$.

Let us aggregate the given regular data into a new one by setting

$$\mathcal{H} = \Omega := \Omega_1 \times ... \times \Omega_L = \mathbf{R}^d,\ d = d_1 + ... + d_L,\ \ \mathcal{M} = \mathcal{M}_1 \times ... \times \mathcal{M}_L,$$

and let us define function $\Phi(h;\mu) : \Omega^d \times \mathcal{M} \to \mathbf{R}$ as follows:

$$\begin{array}{l}\Phi(h = [h^1; ...; h^L]; \mu = [\mu^1; ...; \mu^L]) = \inf_{\lambda\in\mathbf{\Delta}^\epsilon} \sum_{\ell=1}^d \lambda_\ell \Phi_\ell(h^\ell/\lambda_\ell; \mu^\ell),\\ \mathbf{\Delta}^\epsilon = \{\lambda \in \mathbf{R}^d : \lambda_\ell \geq \epsilon\,\forall \ell\ \&\ \sum_{\ell=1}^L \lambda_\ell = 1\}.\end{array} \tag{2.8.13}$$

By evident reasons, the infimum in the description of Φ is achieved, and Φ is continuous. In addition, Φ is convex in $h \in \mathbf{R}^d$ and concave in $\mu \in \mathcal{M}$. Postponing for a moment verification, the consequences are that $\mathcal{H} = \Omega = \mathbf{R}^d$, $\mathcal{M}$ and Φ form a regular data. We claim that

> *Whenever $\omega = [\omega^1; ...; \omega^L]$ is a random variable taking values in $\Omega = \mathbf{R}^{d_1} \times ... \times \mathbf{R}^{d_L}$, and the marginal distributions P_ℓ, $1 \leq \ell \leq L$, of ω belong to the families $\mathcal{S}[\mathbf{R}^{d_\ell}, \mathcal{M}_\ell, \Phi_\ell]$ for all $1 \leq \ell \leq L$, the distribution P of ω belongs to $\mathcal{S}[\mathbf{R}^d, \mathcal{M}, \Phi]$.*

Indeed, since $P_\ell \in \mathcal{S}[\mathbf{R}^{d_\ell}, \mathcal{M}_\ell, \Phi_\ell]$, there exists $\widehat{\mu}^\ell \in \mathcal{M}_\ell$ such that

$$\ln(\mathbf{E}_{\omega^\ell\sim P_\ell}\{\exp\{g^T\omega^\ell\}\}) \leq \Phi_\ell(g; \widehat{\mu}^\ell)\ \forall g \in \mathbf{R}^{d_\ell}.$$

Let us set $\widehat{\mu} = [\widehat{\mu}^1; ...; \widehat{\mu}^L]$, and let $h = [h^1; ...; h^L] \in \Omega$ be given. We can find $\lambda \in \mathbf{\Delta}^\epsilon$ such that

$$\Phi(h; \widehat{\mu}) = \sum_{\ell=1}^L \lambda_\ell \Phi_\ell(h^\ell/\lambda_\ell; \widehat{\mu}^\ell).$$

Applying Hölder inequality, we get

$$\mathbf{E}_{[\omega^1;...;\omega^L]\sim P}\left\{\exp\{\sum_\ell [h^\ell]^T\omega^\ell\}\right\} \leq \prod_{\ell=1}^L \left(\mathbf{E}_{\omega^\ell\sim P_\ell}\left\{[h^\ell]^T\omega^\ell/\lambda_\ell\right\}\right)^{\lambda_\ell},$$

whence

$$\ln\left(\mathbf{E}_{[\omega^1;...;\omega^L]\sim P}\left\{\exp\{\sum_\ell [h^\ell]^T\omega^\ell\}\right\}\right) \leq \sum_{\ell=1}^L \lambda_\ell \Phi_\ell(h^\ell/\lambda_\ell; \widehat{\mu}^\ell) = \Phi(h; \widehat{\mu}).$$

We see that

$$\ln\left(\mathbf{E}_{[\omega^1;...;\omega^L]\sim P}\left\{\exp\{\sum_\ell [h^\ell]^T\omega^\ell\}\right\}\right) \leq \Phi(h; \widehat{\mu})\ \forall h \in \mathcal{H} = \mathbf{R}^d,$$

and thus $P \in \mathcal{S}[\mathbf{R}^d, \mathcal{M}, \Phi]$, as claimed.

It remains to verify that the function Φ defined by (2.8.13) indeed is convex in $h \in \mathbf{R}^d$ and concave in $\mu \in \mathcal{M}$. Concavity in μ is evident. Further, functions $\lambda_\ell \Phi_\ell(h^\ell/\lambda_\ell; \mu)$ (as perspective transformation of convex functions $\Phi_\ell(\cdot; \mu)$) are jointly convex in λ and h^ℓ, and so is $\Psi(\lambda, h; \mu) = \sum_{\ell=1}^L \lambda_\ell \Phi_\ell(h^\ell/\lambda_\ell, \mu)$. Thus $\Phi(\cdot; \mu)$, obtained by partial minimization of Ψ in λ, indeed is convex.

2.8.1.3.E. Affine image. Let $\mathcal{H}$, $\mathcal{M}$, Φ be regular data, Ω be the embedding space of $\mathcal{H}$, and $x \mapsto Ax + a$ be an affine mapping from Ω to $\bar{\Omega} = \mathbf{R}^{\bar{d}}$, and let us set

$$\bar{\mathcal{H}} = \{\bar{h} \in \mathbf{R}^{\bar{d}} : A^T\bar{h} \in \mathcal{H}\},\ \bar{\mathcal{M}} = \mathcal{M},\ \bar{\Phi}(\bar{h}; \mu) = \Phi(A^T\bar{h}; \mu) + a^T\bar{h} : \bar{\mathcal{H}} \times \bar{M} \to \mathbf{R}.$$

Note that $\bar{\mathcal{H}}$, $\bar{\mathcal{M}}$ and $\bar{\Phi}$ are regular data. It is immediately seen that

> *Whenever the probability distribution of a random variable ω belongs to $\mathcal{R}[\mathcal{H}, \mathcal{M}, \Phi]$ (or belongs to $\mathcal{S}[\mathcal{H}, \mathcal{M}, \Phi]$), the distribution $\bar{P}[P]$ of the random variable $\bar{\omega} = A\omega + a$ belongs to $\mathcal{R}[\bar{\mathcal{H}}, \bar{\mathcal{M}}, \bar{\Phi}]$ (respectively, belongs to $\mathcal{S}[\bar{\mathcal{H}}, \bar{\mathcal{M}}, \bar{\Phi}]$).*

2.8.1.3.F. Incorporating support information. Consider the situation as follows. We are given regular data $\mathcal{H} \subset \Omega = \mathbf{R}^d, \mathcal{M}, \Phi$ and are interested in a family $\mathcal{P}$ of distributions known to belong to $\mathcal{S}[\mathcal{H}, \mathcal{M}, \Phi]$. In addition, we know that all distributions P from $\mathcal{P}$ are supported on a given closed convex set $X \subset \mathbf{R}^d$. How could we incorporate this domain information to pass from the family $\mathcal{S}[\mathcal{H}, \mathcal{M}, \Phi]$ containing $\mathcal{P}$ to a smaller family of the same type still containing $\mathcal{P}$? We are about to give an answer in the simplest case of $\mathcal{H} = \Omega$. Specifically, denoting by $\phi_X(\cdot)$ the support function of X and selecting somehow a closed convex set $G \subset \mathbf{R}^d$ containing the origin, let us set

$$\widehat{\Phi}(h;\mu) = \inf_{g\in G}\left[\Phi^+(h,g;\mu) := \Phi(h-g;\mu) + \phi_X(g)\right],$$

where $\Phi(h;\mu) : \mathbf{R}^d \times \mathcal{M} \to \mathbf{R}$ is the continuous convex-concave function participating in the original regular data. Assuming that $\widehat{\Phi}$ is real-valued and continuous on the domain $\mathbf{R}^d \times \mathcal{M}$ (which definitely is the case when G is a compact set such that ϕ_X is finite and continuous on G), note that $\widehat{\Phi}$ is convex-concave on this domain, so that $\mathbf{R}^d, \mathcal{M}, \widehat{\Phi}$ is a regular data. We claim that

> *The family $\mathcal{S}[\mathbf{R}^d, \mathcal{M}, \widehat{\Phi}]$ contains $\mathcal{P}$, provided the family $\mathcal{S}[\mathbf{R}^d, \mathcal{M}, \Phi]$ does so, and the first of these two families is smaller than the second one.*

Verification of the claim is immediate. Let $P \in \mathcal{P}$, so that for properly selected $\mu = \mu_P \in \mathcal{M}$ and for all $e \in \mathbf{R}^d$ it holds

$$\ln\left(\int_{\mathbf{R}^d} \exp\{e^T\omega\}P(d\omega)\right) \le \Phi(e;\mu_P).$$

Besides this, for every $g \in G$ we have $\phi_X(g) - g^T\omega \ge 0$ on the support of P, whence for every $h \in \mathbf{R}^d$ one has

$$\ln\left(\int_{\mathbf{R}^d} \exp\{h^T\omega\}P(d\omega)\right) \le \ln\left(\int_{\mathbf{R}^d} \exp\{h^T\omega + \phi_X(g) - g^T\omega\}P(d\omega)\right) \le \phi_X(g) + \Phi(h-g;\mu_P).$$

Since the resulting inequality holds true for all $g \in G$, we get

$$\ln\left(\int_{\mathbf{R}^d} \exp\{h^T\omega\}P(d\omega)\right) \le \widehat{\Phi}(h;\mu_P)\ \forall h \in \mathbf{R}^d,$$

implying that $P \in \mathcal{S}[\mathbf{R}^d, \mathcal{M}, \widehat{\Phi}]$; since $P \in \mathcal{P}$ is arbitrary, the first part of the claim is justified. The inclusion $\mathcal{S}[\mathbf{R}^d, \mathcal{M}, \widehat{\Phi}] \subset \mathcal{S}[\mathbf{R}^d, \mathcal{M}, \Phi]$ is readily given by the inequality $\widehat{\Phi} \le \Phi$, and the latter is due to $\widehat{\Phi}(h,\mu) \le \Phi(h-0,\mu) + \phi_X(0)$.

Illustration: distributions with bounded support revisited. In Section 2.8.1.2, given a convex compact set $X \subset \mathbf{R}^d$ with support function ϕ_X, we checked that the data $\mathcal{H} = \mathbf{R}^d$, $\mathcal{M} = X$, $\Phi(h;\mu) = h^T\mu + \frac{1}{8}[\phi_X(h) + \phi_X(-h)]^2$ are regular and the family $\mathcal{S}[\mathbf{R}^d, \mathcal{M}, \Phi]$ contains the family $\mathcal{P}[X]$ of all probability distributions supported on X. Moreover, for every $\mu \in \mathcal{M} = X$, the family $\mathcal{S}[\mathbf{R}^d, \{\mu\}, \Phi\big|_{\mathbf{R}^d\times\{\mu\}}]$ contains all supported on X distributions with the expectations $e[P] = \mu$. Note that $\Phi(h;e[P])$ describes well the behaviour of the logarithm $F_P(h) = \ln\left(\int_{\mathbf{R}^d} e^{h^T\omega}P(d\omega)\right)$ of the moment-generating function of $P \in \mathcal{P}[X]$ when h is small (indeed, $F_P(h) = h^Te[P] + O(\|h\|^2)$ as $h \to 0$), and by far overestimates $F_P(h)$ when h is large. Utilizing the above construction, we replace Φ with the real-valued, convex-concave and continuous on $\mathbf{R}^d \times \mathcal{M} = \mathbf{R}^d \times X$ (see Exercise 2.20) function

$$\widehat{\Phi}(h;\mu) = \inf_g\left[\widehat{\Psi}(h,g;\mu) := (h-g)^T\mu + \frac{1}{8}[\phi_X(h-g) + \phi_X(-h+g)]^2 + \phi_X(g)\right] \le \Phi(h;\mu). \tag{2.8.14}$$

It is easy to see that $\widehat{\Phi}(\cdot;\cdot)$ still ensures the inclusion $P \in \mathcal{S}[\mathbf{R}^d, \{e[P]\}, \widehat{\Phi}|_{\mathbf{R}^d\times\{e[P]\}}]$ for every distribution $P \in \mathcal{P}[X]$ and "reproduces $F_P(h)$ reasonably well" for both small and large h. Indeed, since $F_P(h) \leq \widehat{\Phi}(h;e[P]) \leq \Phi(h;e[P])$, for small h $\widehat{\Phi}(h;e[P])$ reproduces $F_P(h)$ even better than $\Phi(h;e[P])$, and we clearly have

$$\widehat{\Phi}(h;\mu) \leq \left[(h-h)^T\mu + \frac{1}{8}[\phi_X(h-h) + \phi_X(-h+h)]^2 + \phi_X(h)\right] = \phi_X(h)\ \forall \mu,$$

and $\phi_X(h)$ is a correct description of $F_P(h)$ for large h.

2.8.2 Main result

2.8.2.1 Situation & Construction

Assume we are given two collections of regular data with common $\Omega = \mathbf{R}^d$ and common $\mathcal{H}$, specifically, the collections $(\mathcal{H}, \mathcal{M}_\chi, \Phi_\chi)$, $\chi = 1,2$. We start with constructing a specific detector for the associated families of regular probability distributions

$$\mathcal{P}_\chi = \mathcal{R}[\mathcal{H}, \mathcal{M}_\chi, \Phi_\chi],\ \chi = 1,2.$$

When building the detector, we impose on the regular data in question the following

Assumption I: *The regular data $(\mathcal{H}, \mathcal{M}_\chi, \Phi_\chi)$, $\chi = 1,2$, are such that the convex-concave function*

$$\Psi(h;\mu_1,\mu_2) = \frac{1}{2}\left[\Phi_1(-h;\mu_1) + \Phi_2(h;\mu_2)\right] : \mathcal{H} \times (\mathcal{M}_1 \times \mathcal{M}_2) \to \mathbf{R} \tag{2.8.15}$$

has a saddle point ($\min$ *in $h \in \mathcal{H}$,* $\max$ *in $(\mu_1,\mu_2) \in \mathcal{M}_1 \times \mathcal{M}_2$).*

A simple sufficient condition for existence of a saddle point of (2.8.15) is

Condition A: *The sets $\mathcal{M}_1$ and $\mathcal{M}_2$ are compact, and the function*

$$\overline{\Phi}(h) = \max_{\mu_1\in\mathcal{M}_1,\mu_2\in\mathcal{M}_2} \Phi(h;\mu_1,\mu_2)$$

is coercive on $\mathcal{H}$, meaning that $\overline{\Phi}(h_i) \to \infty$ along every sequence $h_i \in \mathcal{H}$ with $\|h_i\|_2 \to \infty$ as $i \to \infty$.

Indeed, under Condition A by Sion-Kakutani Theorem (Theorem 2.4.1) it holds

$$\mathrm{SadVal}[\Phi] := \inf_{h\in\mathcal{H}} \underbrace{\max_{\mu_1\in M_1,\mu_2\in\mathcal{M}_2} \Phi(h;\mu_1,\mu_2)}_{\overline{\Phi}(h)} = \sup_{\mu_1\in M_1,\mu_2\in\mathcal{M}_2} \underbrace{\inf_{h\in\mathcal{H}} \Phi(h;\mu_1,\mu_2)}_{\underline{\Phi}(\mu_1,\mu_2)},$$

so that the optimization problems

$$\begin{array}{ll}(P): & \mathrm{Opt}(P) = \min\limits_{h\in\mathcal{H}} \overline{\Phi}(h)\\ (D): & \mathrm{Opt}(D) = \max\limits_{\mu_1\in\mathcal{M}_1,\mu_2\in\mathcal{M}_2} \underline{\Phi}(\mu_1,\mu_2)\end{array}$$

have equal optimal values. Under Condition A, problem (P) clearly is a problem of minimizing a continuous coercive function over a closed set and as such is solvable; thus, $\mathrm{Opt}(P) = \mathrm{Opt}(D)$ is a real. Problem (D) clearly is the problem of maximizing over a compact set an upper semi-continuous (since Φ is continuous) function taking real values and, perhaps, value $-\infty$, and not identically equal to $-\infty$ (since $\mathrm{Opt}(D)$ is a real), and thus (D) is solvable. Thus, (P) and (D) are solvable with common optimal values, and therefore Φ has a saddle point.

2.8.2.2 Main Result

An immediate (and crucial!) observation is as follows:

Proposition 2.8.3 *In the situation of Section 2.8.2.1, let $h\in\mathcal{H}$ be such that the quantities*

$$\Psi_1(h)=\sup_{\mu_1\in\mathcal{M}_1}\Phi_1(-h;\mu_1),\ \ \Psi_2(h)=\sup_{\mu_2\in\mathcal{M}_2}\Phi_2(h;\mu_2)$$

are finite. Consider the affine detector

$$\phi_h(\omega)=h^T\omega+\underbrace{\frac{1}{2}[\Psi_1(h)-\Psi_2(h)]}_{\varkappa}.$$

Then

$$\mathrm{Risk}[\phi_h|\mathcal{R}[\mathcal{H},\mathcal{M}_1,\Phi_1],\mathcal{R}[\mathcal{H},\mathcal{M}_2,\Phi_2]]\leq\exp\{\frac{1}{2}[\Psi_1(h)+\Psi_2(h)]\}. \tag{2.8.16}$$

Proof. Let h satisfy the premise of Proposition. For every $\mu_1\in\mathcal{M}_1$, we have $\Phi_1(-h;\mu_1)\leq\Psi_1(h)$, and for every $P\in\mathcal{R}[\mathcal{H},\mathcal{M}_1,\Phi_1]$, we have

$$\int_\Omega\exp\{-h^T\omega\}P(d\omega)\leq\exp\{\Phi_1(-h;\mu_1)\}$$

for properly selected $\mu_1\in\mathcal{M}_1$. Thus,

$$\int_\Omega\exp\{-h^T\omega\}P(d\omega)\leq\exp\{\Psi_1(h)\}\ \forall P\in\mathcal{R}[\mathcal{H},\mathcal{M}_1,\Phi_1],$$

whence also

$$\int_\Omega\exp\{-h^T\omega-\varkappa\}P(d\omega)\leq\exp\{\Psi_1(h)-\varkappa\}=\exp\{\frac{1}{2}[\Psi_1(h)+\Psi_2(h)]\}\ \forall P\in\mathcal{R}[\mathcal{H},\mathcal{M}_1,\Phi_1].$$

Similarly, for every $\mu_2\in\mathcal{M}_2$, we have $\Phi_2(h;\mu_2)\leq\Psi_2(h)$, and for every $P\in\mathcal{R}[\mathcal{H},\mathcal{M}_2,\Phi_2]$, we have

$$\int_\Omega\exp\{h^T\omega\}P(d\omega)\leq\exp\{\Phi_2(h;\mu_2)\}$$

for properly selected $\mu_2\in\mathcal{M}_2$. Thus,

$$\int_\Omega\exp\{h^T\omega\}P(d\omega)\leq\exp\{\Psi_2(h)\}\ \forall P\in\mathcal{R}[\mathcal{H},\mathcal{M}_2,\Phi_2],$$

whence also

$$\int_\Omega\exp\{h^T\omega+\varkappa\}P(d\omega)\leq\exp\{\Psi_2(h)+\varkappa\}\exp\{\frac{1}{2}[\Psi_1(h)+\Psi_2(h)]\}\ \forall P\in\mathcal{R}[\mathcal{H},\mathcal{M}_2,\Phi_2]\qquad\square$$

An immediate corollary is as follows:

Proposition 2.8.4 *In the situation of Section 2.8.2.1 and under Assumption I, let us associate with a saddle point $(h_*;\mu_1^*,\mu_2^*)$ of the convex-concave function* (2.8.15) *the following entities:*

- *the risk*
$$\epsilon_\star:=\exp\{\Psi(h_*;\mu_1^*,\mu_2^*)\}; \tag{2.8.17}$$
this quantity is uniquely defined by the saddle point value of Ψ and thus is independent of how we select a saddle point;
- *the detector $\phi_*(\omega)$ – the affine function of $\omega\in\mathbf{R}^d$ given by*
$$\phi_*(\omega)=h_*^T\omega+a,\ \ a=\frac{1}{2}\left[\Phi_1(-h_*;\mu_1^*)-\Phi_2(h_*;\mu_2^*)\right]. \tag{2.8.18}$$

Then

$$\mathrm{Risk}[\phi_*|\mathcal{R}[\mathcal{H},\mathcal{M}_1,\Phi_1],\mathcal{R}[\mathcal{H},\mathcal{M}_2,\Phi_2]]\leq\epsilon_\star. \tag{2.8.19}$$

Consequences. Assume we are given L collections $(\mathcal{H}, \mathcal{M}_\ell, \Phi_\ell)$ of regular data on a common observation space $\Omega = \mathbf{R}^d$ and with common $\mathcal{H}$, and let

$$\mathcal{P}_\ell = \mathcal{R}[\mathcal{H}, \mathcal{M}_\ell, \Phi_\ell]$$

be the corresponding families of regular distributions. Assume also that for every pair (ℓ, ℓ'), $1 \leq \ell < \ell' \leq L$, the pair of regular data $(\mathcal{H}, \mathcal{M}_\ell, \Phi_\ell)$, $(\mathcal{H}, \mathcal{M}_{\ell'}, \Phi_{\ell'})$ satisfies Assumption I, so that the convex-concave functions

$$\Psi_{\ell\ell'}(h; \mu_\ell, \mu_{\ell'}) = \frac{1}{2}\left[\Phi_\ell(-h; \mu_\ell) + \Phi_{\ell'}(h; \mu_{\ell'})\right] : \mathcal{H} \times (\mathcal{M}_\ell \times \mathcal{M}_{\ell'}) \to \mathbf{R} \qquad [1 \leq \ell < \ell' \leq L]$$

have saddle points $(h^*_{\ell\ell'}; (\mu^*_\ell, \mu^*_{\ell'}))$ (min in $h \in \mathcal{H}$, max in $(\mu_\ell, \mu_{\ell'}) \in \mathcal{M}_\ell \times \mathcal{M}_{\ell'}$). These saddle points give rise to affine detectors

$$\phi_{\ell\ell'}(\omega) = [h^*_{\ell\ell'}]^T\omega + \frac{1}{2}\left[\Phi_\ell(-h^*_{\ell\ell'}; \mu^*_\ell) - \Phi_{\ell'}(h_*; \mu^*_{\ell'})\right] \qquad [1 \leq \ell < \ell' \leq L]$$

and the quantities

$$\epsilon_{\ell\ell'} = \exp\left\{\frac{1}{2}\left[\Phi_\ell(-h^*_{\ell\ell'}; \mu^*_\ell) + \Phi_{\ell'}(h_*; \mu^*_{\ell'})\right]\right\}; \qquad [1 \leq \ell < \ell' \leq L]$$

by Proposition 2.8.4, $\epsilon_{\ell\ell'}$ are upper bounds on the risks, taken w.r.t. $\mathcal{P}_\ell, \mathcal{P}_{\ell'}$, of the detectors $\phi_{\ell\ell'}$:

$$\int_\Omega \mathrm{e}^{-\phi_{\ell\ell'}(\omega)} P(d\omega) \leq \epsilon_{\ell\ell'}\ \forall P \in \mathcal{P}_\ell\ \&\ \int_\Omega \mathrm{e}^{\phi_{\ell\ell'}(\omega)} P(d\omega) \leq \epsilon_{\ell\ell'}\ \forall P \in \mathcal{P}_{\ell'}. \qquad [1 \leq \ell < \ell' \leq L].$$

Setting $\phi_{\ell\ell'}(\cdot) = -\phi_{\ell'\ell}(\cdot)$ and $\epsilon_{\ell\ell'} = \epsilon_{\ell'\ell}$ when $L \geq \ell > \ell' \geq 1$ and $\phi_{\ell\ell}(\cdot) \equiv 0$, $\epsilon_{\ell\ell} = 1$, $1 \leq \ell \leq L$, we get a system of detectors and risks satisfying (2.5.9) and, consequently, can use these "building blocks" in the developed so far machinery for pairwise- and multiple hypothesis testing from single and repeated observations (stationary, semi-stationary, and quasi-stationary).

Numerical example. To get some impression of how Proposition 2.8.4 extends the grasp of our computation-friendly test design machinery. consider a toy problem as follows:

> We are given observation
>
> $$\omega = Ax + \sigma A\mathrm{Diag}\left\{\sqrt{x_1}, ..., \sqrt{x_n}\right\}\xi, \qquad (2.8.20)$$
>
> where
>
> - unknown signal x is known to belong to a given convex compact subset M of the *interior* of $\mathbf{R}^n_+$;
> - A is a given $n \times n$ matrix of rank n, $\sigma > 0$ is a given noise intensity, and $\xi \sim \mathcal{N}(0, I_n)$.
>
> Our goal is to decide via K-repeated version of observations (2.8.20) on the pair of hypotheses $x \in X_\chi$, $\chi = 1, 2$, where X_1, X_2 are given nonempty convex compact subsets of M.

Note that an essential novelty, as compared to the standard Gaussian o.s., is that now we deal with zero mean Gaussian noise with covariance matrix

$$\Theta(x) = \sigma^2 A\mathrm{Diag}\{x\}A^T$$

depending on the true signal – the larger signal, the larger noise.

We can easily process the situation in question via the machinery developed in this Section. Specifically, let us set

$$\mathcal{H}_\chi = \mathbf{R}^n,\ \mathcal{M}_\chi = \{(x, \mathrm{Diag}\{x\}) : x \in X_\chi\} \subset \mathbf{R}^n \times \mathbf{S}^n_+,\ \Phi_\chi(h; x, \Xi) = h^T x + \frac{\sigma^2}{2} h^T [A\Xi A^T] h : \mathcal{M}_\chi \to \mathbf{R} \qquad [\chi = 1, 2]$$

It is immediately seen that for $\chi = 1, 2$, $\mathcal{H}, \mathcal{M}_\chi, \Phi_\chi$ is regular data, and that the distribution P of observation (2.8.20) stemming from a signal $x \in X_\chi$ belongs to $\mathcal{S}[\mathcal{H}, \mathcal{M}_\chi, \Phi_\chi]$, so that we can use Proposition 2.8.4 to build an affine detector for the families $\mathcal{P}_\chi$, $\chi = 1, 2$, of distributions of observations (2.8.20) stemming from signals $x \in X_\chi$. The corresponding recipe boils down to the necessity to find a saddle point $(h_*; x_*, y_*)$ of the simple convex-concave function

$$\Psi(h; x, y) = \frac{1}{2}\left[h^T(y - x) + \frac{\sigma^2}{2} h^T A \mathrm{Diag}\{x + y\} A^T h\right] \tag{2.8.21}$$

(min in $h \in \mathbf{R}^n$, max in $(x, y) \in X_1 \times X_2$); such a point clearly exists and is easily found, and gives rise to affine detector

$$\phi_*(\omega) = h_*^T \omega + \underbrace{\frac{\sigma^2}{4} h_*^T A \mathrm{Diag}\{x_* - y_*\} A^T h_* - \frac{1}{2} h_*^T [x_* + y_*]}_{a}$$

such that

$$\mathrm{Risk}[\phi_* | \mathcal{P}_1, \mathcal{P}_2] \le \exp\left\{\frac{1}{2}\left[h_*^T [y_* - x_*] + \frac{\sigma^2}{2} h_*^T A \mathrm{Diag}\{x_* + y_*\} A^T h_*\right]\right\}. \tag{2.8.22}$$

Note that we could also process the situation when defining the regular data as $\mathcal{H}, \mathcal{M}^+_\chi = X_\chi, \Phi^+_\chi$, $\chi = 1, 2$, where

$$\Phi^+_\chi(h; x) = h^T x + \frac{\sigma^2 \theta}{2} h^T A A^T h \qquad [\theta = \max_{x \in X_1 \cup X_2} \|x\|_\infty]$$

which, basically, means passing from our actual observations (2.8.20) to the "more noisy" observations given by the Gaussian o.s.

$$\omega = Ax + \eta,\ \eta \sim \mathcal{N}(0, \sigma^2 \theta A A^T). \tag{2.8.23}$$

It is easily seen that the risk $\mathrm{Risk}[\phi_\# | \mathcal{P}_1, \mathcal{P}_2]$ of the optimal, for this Gaussian o.s., detector $\phi_\#$, can be upper-bounded by the known to us risk $\mathrm{Risk}[\phi_\# | \mathcal{P}^+_1, \mathcal{P}^+_2]$, where $\mathcal{P}^+_\chi$ is the family of distributions of observations (2.8.23) induced by signals $x \in X_\chi$. Note that were we staying within the realm of detector-based tests in simple o.s.'s, $\mathrm{Risk}[\phi_\# | \mathcal{P}^+_1, \mathcal{P}^+_2]$ would be seemingly the best risk bound available for us. The goal of the small numerical experiment we are about to report was to understand how our new risk bound (2.8.22) compares to the "old" bound $\mathrm{Risk}[\phi_\# | \mathcal{P}^+_1, \mathcal{P}^+_2]$. We used

$$n = 16,\ X_1 = \left\{x \in \mathbf{R}^{16} : \begin{array}{l} 0.001 \le x_1 \le \delta \\ 0.001 \le x_i \le 1,\ 2 \le i \le 16 \end{array}\right\},\ X_2 = \left\{x \in \mathbf{R}^{16} : \begin{array}{l} 2\delta \le x_1 \le 1 \\ 0.001 \le x_i \le 1,\ 2 \le i \le 16 \end{array}\right\}$$

and $\sigma = 0.1$. The "separation parameter" δ was set to 0.1. Finally, 16×16 matrix A was generated to have condition number 100 (singular values $0.01^{(i-1)/15}$, $1 \le i \le 16$) with randomly oriented systems of left- and right singular vectors. With this setup, a typical numerical result is as follows:

- the right hand side in (2.8.22) is 0.4346, implying that with detector ϕ_*, 6-repeated observation is sufficient to decide on our two hypotheses with risk ≤ 0.01;

- the quantity $\mathrm{Risk}[\phi_\# | \mathcal{P}^+_1, \mathcal{P}^+_2]$ is 0.8825, meaning that with detector $\phi_\#$, we need at least 37-repeated observation to guarantee risk ≤ 0.01.

When the separation parameter δ participating in the descriptions of X_1, X_2 was reduced to 0.01, the risks in question grew to 0.9201 and 0.9988, respectively (56-repeated observation to decide on the hypotheses with risk 0.01 when ϕ_* is used vs. 3685-repeated observation needed when $\phi_\#$ is used). The bottom line is that our new developments could improve quite significantly the performance of our inferences.

2.8.2.3 Illustration: sub-Gaussian and Gaussian cases

For $\chi = 1, 2$, let U_χ be nonempty closed convex set in $\mathbf{R}^d$, and $\mathcal{V}_\chi$ be a compact convex subset of the interior of the positive semidefinite cone $\mathbf{S}^d_+$. We assume that U_1 is compact. Setting

$$\mathcal{H}_\chi = \Omega = \mathbf{R}^d,\ \mathcal{M}_\chi = U_\chi \times \mathcal{V}_\chi,\ \ \Phi_\chi(h;\theta,\Theta) = \theta^T h + \frac{1}{2}h^T\Theta h : \mathcal{H}_\chi \times \mathcal{M}_\chi \to \mathbf{R}, \chi = 1,2, \quad (2.8.24)$$

we get two collections $(\mathcal{H}, \mathcal{M}_\chi, \Phi_\chi)$, $\chi = 1, 2$, of regular data. As we know from Section 2.8.1.2, for $\chi = 1, 2$, the families of distributions $\mathcal{S}[\mathbf{R}^d, \mathcal{M}_\chi, \Phi_\chi]$ contain the families $\mathcal{SG}[U_\chi, \mathcal{V}_\chi]$ of sub-Gaussian distributions on $\mathbf{R}^d$ with sub-Gaussianity parameters $(\theta, \Theta) \in U_\chi \times \mathcal{V}_\chi$ (see (2.8.3)), as well as families $\mathcal{G}[U_\chi, \mathcal{V}_\chi]$ of Gaussian distributions on $\mathbf{R}^d$ with parameters (θ, Θ) (expectation and covariance matrix) running through $U_\chi \times \mathcal{V}_\chi$. Besides this, the pair of regular data in question clearly satisfies Condition A. Consequently, the test $\mathcal{T}^K_*$ given by the above construction as applied to the collections of regular data (2.8.24) is well defined and allows to decide on hypotheses

$$H_\chi : P \in \mathcal{R}[\mathbf{R}^d, U_\chi, \mathcal{V}_\chi],\ \chi = 1, 2,$$

on the distribution P underlying K-repeated observation ω^K. The same test can be also used to decide on stricter hypotheses H^G_χ, $\chi = 1, 2$, stating that the observations $\omega_1, ..., \omega_K$ are i.i.d. and drawn from a Gaussian distribution P belonging to $\mathcal{G}[U_\chi, \mathcal{V}_\chi]$. Our goal now is to process in detail the situation in question and to refine our conclusions on the risk of the test $\mathcal{T}^1_*$ when the *Gaussian* hypotheses H^G_χ are considered and the situation is *symmetric*, that is, when $\mathcal{V}_1 = \mathcal{V}_2$.

Observe, first, that the convex-concave function Ψ from (2.8.15) in the situation under consideration becomes

$$\Psi(h; \theta_1, \Theta_1, \theta_2, \Theta_2) = \frac{1}{2}h^T[\theta_2 - \theta_1] + \frac{1}{4}h^T\Theta_1 h + \frac{1}{4}h^T\Theta_2 h. \quad (2.8.25)$$

We are interested in solutions to the saddle point problem

$$\min_{h\in\mathbf{R}^d} \max_{\substack{\theta_1\in U_1,\theta_2\in U_2\\ \Theta_1\in\mathcal{V}_1,\Theta_2\in\mathcal{V}_2}} \Psi(h; \theta_1, \Theta_1, \theta_2, \Theta_2) \quad (2.8.26)$$

associated with the function (2.8.25). From the structure of Ψ and compactness of U_1, $\mathcal{V}_1$, $\mathcal{V}_2$, combined with the fact that $\mathcal{V}_\chi$, $\chi = 1, 2$, are comprised of positive definite matrices, it immediately follows that saddle points do exist, and a saddle point $(h_*; \theta^*_1, \Theta^*_1, \theta^*_2, \Theta^*_2)$ satisfies the relations

$$\begin{array}{ll} (a) & h_* = [\Theta^*_1 + \Theta^*_2]^{-1}[\theta^*_1 - \theta^*_2], \\ (b) & h^T_*(\theta_1 - \theta^*_1) \geq 0\ \forall \theta_1 \in U_1,\ \ h^T_*(\theta^*_2 - \theta_2) \geq 0\ \forall \theta_2 \in U_2, \\ (c) & h^T_*\Theta_1 h_* \leq h^T_*\Theta^*_1 h_*\ \forall \Theta_1 \in \mathcal{V}_1,\ \ h^T_*\Theta^*_2 h_* \leq h_*\Theta^*_2 h_*\ \forall \Theta_2 \in \mathcal{V}_2. \end{array} \quad (2.8.27)$$

From (2.8.27.a) it immediately follows that the affine detector $\phi_*(\cdot)$ and risk $\epsilon_\star$, as given by (2.8.17) and (2.8.18), are

$$\begin{array}{rcl} \phi_*(\omega) & = & h^T_*[\omega - w_*] + \frac{1}{2}h^T_*[\Theta^*_1 - \Theta^*_2]h_*,\ w_* = \frac{1}{2}[\theta^*_1 + \theta^*_2]; \\ \epsilon_\star & = & \exp\{-\frac{1}{4}[\theta^*_1 - \theta^*_2]^T[\Theta^*_1 + \Theta^*_2]^{-1}[\theta^*_1 - \theta^*_2]\} \\ & = & \exp\{-\frac{1}{4}h^T_*[\Theta^*_1 + \Theta^*_2]h_*\}. \end{array} \quad (2.8.28)$$

Note that in the *symmetric case* (where $\mathcal{V}_1 = \mathcal{V}_2$), there always exists a saddle point of Ψ with $\Theta_1^* = \Theta_2^*$ [23], and the test $\mathcal{T}_*^1$ associated with such saddle point is quite transparent: it is the maximum likelihood test for two Gaussian distributions, $\mathcal{N}(\theta_1^*, \Theta_*)$, $\mathcal{N}(\theta_2^*, \Theta_*)$, where Θ_* is the common value of Θ_1^* and Θ_2^*, and the bound $\epsilon_\star$ on the risk of the test is nothing but the Hellinger affinity of these two Gaussian distributions, or, equivalently,

$$\epsilon_\star = \exp\left\{-\tfrac{1}{8}[\theta_1^* - \theta_2^*]^T \Theta_*^{-1} [\theta_1^* - \theta_2^*]\right\}. \tag{2.8.29}$$

We arrive at the following result:

Proposition 2.8.5 *In the symmetric sub-Gaussian case (i.e., in the case of* (2.8.24) *with $\mathcal{V}_1 = \mathcal{V}_2$), saddle point problem* (2.8.25), (2.8.26) *admits a saddle point of the form $(h_*; \theta_1^*, \Theta_*, \theta_2^*, \Theta_*)$, and the associated affine detector and its risk are given by*

$$\begin{array}{rcl}\phi_*(\omega) & = & h_*^T[\omega - w_*],\ w_* = \frac{1}{2}[\theta_1^* + \theta_2^*];\\ \epsilon_\star & = & \exp\{-\frac{1}{8}[\theta_1^* - \theta_2^*]^T \Theta_*^{-1} [\theta_1^* - \theta_2^*]\}.\end{array} \tag{2.8.30}$$

As a result, when deciding, via ω^K, on "sub-Gaussian hypotheses" H_χ, $\chi = 1, 2$, the risk of the test $\mathcal{T}_*^K$ associated with $\phi_*^{(K)}(\omega^K) := \sum_{t=1}^K \phi_*(\omega_t)$ is at most $\epsilon_\star^K$.

In the symmetric single-observation Gaussian case, that is, when $\mathcal{V}_1 = \mathcal{V}_2$ and we apply the test $\mathcal{T}_* = \mathcal{T}_*^1$ to observation $\omega \equiv \omega_1$ in order to decide on the hypotheses H_χ^G, $\chi = 1, 2$, the above risk bound can be improved:

Proposition 2.8.6 *Consider symmetric case $\mathcal{V}_1 = \mathcal{V}_2 = \mathcal{V}$, let $(h_*; \theta_1^*; \Theta_1^*, \theta_2^*, \Theta_2^*)$ be "symmetric" – with $\Theta_1^* = \Theta_2^* = \Theta_*$ – saddle point of function Ψ given by* (2.8.25)*, and let ϕ_* be the affine detector given by* (2.8.27) and (2.8.28)*:*

$$\phi_*(\omega) = h_*^T[\omega - w_*],\ \ h_* = \frac{1}{2}\Theta_*^{-1}[\theta_1^* - \theta_2^*],\ \ w_* = \frac{1}{2}[\theta_1^* + \theta_2^*].$$

Let also

$$\delta = \sqrt{h_*^T \Theta_* h_*} = \frac{1}{2}\sqrt{[\theta_1^* - \theta_2^*]^T \Theta_*^{-1} [\theta_1^* - \theta_2^*]}, \tag{2.8.31}$$

so that

$$\delta^2 = h_*^T[\theta_1^* - w_*] = h_*^T[w_* - \theta_2^*] \text{ and } \epsilon_\star = \exp\{-\frac{1}{2}\delta^2\}. \tag{2.8.32}$$

Let, further, $\alpha \le \delta^2$, $\beta \le \delta^2$. Then

$$\begin{array}{ll}(a) & \forall(\theta \in U_1, \Theta \in \mathcal{V}) : \mathrm{Prob}_{\omega \sim \mathcal{N}(\theta,\Theta)}\{\phi_*(\omega) \le \alpha\} \le \mathrm{Erf}(\delta - \alpha/\delta)\\ (b) & \forall(\theta \in U_2, \Theta \in \mathcal{V}) : \mathrm{Prob}_{\omega \sim \mathcal{N}(\theta,\Theta)}\{\phi_*(\omega) \ge -\beta\} \le \mathrm{Erf}(\delta - \beta/\delta),\end{array} \tag{2.8.33}$$

where

$$\mathrm{Erf}(s) = \frac{1}{\sqrt{2\pi}} \int_s^\infty \exp\{-r^2/2\} dr$$

is the normal error function. In particular, when deciding, via a single observation ω, on Gaussian hypotheses H_χ^G, $\chi = 1, 2$, with H_χ^G stating that $\omega \sim \mathcal{N}(\theta, \Theta)$ with $(\theta, \Theta) \in U_\chi \times \mathcal{V}$, the risk of the test $\mathcal{T}_*^1$ associated with ϕ_* is at most $\mathrm{Erf}(\delta)$.

[23] Indeed, from (2.8.25) it follows that when $\mathcal{V}_1 = \mathcal{V}_2$, the function $\Psi(h; \theta_1, \Theta_1, \theta_2, \Theta_2)$ is symmetric w.r.t. Θ_1, Θ_2, implying similar symmetry of the function $\underline{\Psi}(\theta_1, \Theta_1, \theta_2, \Theta_2) = \min_{h \in \mathcal{H}} \Psi(h; \theta_1, \Theta_1, \theta_2, \Theta_2)$. Since $\underline{\Psi}$ is concave, the set M of its maximizers over $\mathcal{M}_1 \times \mathcal{M}_2$ (which, as we know, is nonempty) is symmetric w.r.t. the swap of Θ_1 and Θ_2 and is convex, implying that if $(\theta_1, \Theta_1, \theta_2, \Theta_2) \in M$, then $(\theta_1, \frac{1}{2}[\Theta_1 + \Theta_2], \theta_2, \frac{1}{2}[\Theta_1 + \Theta_2]) \in M$ as well, and the latter point is the desired component of saddle point of Ψ with $\Theta_1 = \Theta_2$.

Proof. Let us prove (a) (the proof of (b) is completely similar). For $\theta \in U_1$, $\Theta \in \mathcal{V}$ we have

$$\begin{array}{l}
\mathrm{Prob}_{\omega\sim\mathcal{N}(\theta,\Theta)}\{\phi_*(\omega) \leq \alpha\} = \mathrm{Prob}_{\omega\sim\mathcal{N}(\theta,\Theta)}\{h_*^T[\omega - w_*] \leq \alpha\}\\
= \mathrm{Prob}_{\xi\sim\mathcal{N}(0,I)}\{h_*^T[\theta + \Theta^{1/2}\xi - w_*] \leq \alpha\}\\
= \mathrm{Prob}_{\xi\sim\mathcal{N}(0,I)}\{[\Theta^{1/2}h_*]^T\xi \leq \alpha - \underbrace{h_*^T[\theta - w_*]}_{\substack{\geq h_*^T[\theta_1^* - w_*] = \delta^2\\ \text{by } (2.8.27.b),(2.8.32)}}\} \leq \mathrm{Prob}_{\xi\sim\mathcal{N}(0,I)}\{[\Theta^{1/2}h_*]^T\xi \leq \alpha - \delta^2\}\\
= \mathrm{Erf}([\delta^2 - \alpha]/\|\Theta^{1/2}h_*\|_2)\\
\leq \mathrm{Erf}([\delta^2 - \alpha]/\|\Theta_*^{1/2}h_*\|_2) \text{ [since } \delta^2 - \alpha \geq 0 \text{ and } h_*^T\Theta h_* \leq h_*^T\Theta_* h_* \text{ by } (2.8.27.c)]\\
= \mathrm{Erf}([\delta^2 - \alpha]/\delta).
\end{array}$$

The “in particular” part of Proposition is readily given by (2.8.33) as applied with $\alpha = \beta = 0$. □

Note that the progress, as compared to our results on the minimum risk detectors for convex hypotheses in Gaussian o.s. is that we do *not* assume anymore that the covariance matrix is once for ever fixed. Now both the mean *and* the covariance matrix of Gaussian random variable we are observing are not known in advance, the mean is allowed to run through a closed convex set (depending on the hypothesis), the covariance is allowed to run, independently of the mean, through a given convex compact subset of the interior of the positive definite cone, and this subset should be common for both hypotheses we are deciding upon.

2.8.3 Beyond the scope of affine detectors: lifting observations

2.8.3.1 Motivation

The detectors considered so far in this Section were affine functions of observations. Note, however, that what is an observation, it to some extent depends on us. To give an instructive example, consider the Gaussian observation

$$\zeta = A[u; 1] + \xi \in \mathbf{R}^n,$$

where u is unknown signal known to belong to a given set $U \subset \mathbf{R}^n$, $u \mapsto A[u; 1]$ is a given affine mapping from $\mathbf{R}^n$ into the observation space $\mathbf{R}^d$, and ξ is zero mean Gaussian observation noise with covariance matrix Θ known to belong to a given convex compact subset $\mathcal{V}$ of the interior of the positive semidefinite cone $\mathbf{S}^d_+$. Treating observation “as is”, affine in observation detector is affine in $[u; \xi]$. On the other hand, we can treat as our observation the image of the actual observation ζ under a whatever deterministic mapping, e.g., the “quadratic lift” $\zeta \mapsto (\zeta, \zeta\zeta^T)$. A detector affine in the new observation is quadratic in u and ξ – we get access to a wider set of detectors as compared to those affine in ζ! At the first glance, applying our “affine detectors” machinery to appropriate “nonlinear lifts” of actual observations we can handle quite complicated detectors, e.g., polynomial, of arbitrary degree, in ζ. The bottleneck here stems from the fact that in general it is really difficult to “cover” the distribution of “nonlinearly lifted” actual observation ζ (even as simple as the above Gaussian observation) by an explicitly defined family of regular distributions; and such a “covering” is what we need in order to apply to the lifted observation our affine detector machinery. It turns out, however, that in some important cases the desired covering is achievable. We are about to demonstrate that this favorable situation indeed takes place when speaking about the quadratic lifting $\zeta \mapsto (\zeta, \zeta\zeta^T)$ of (sub)Gaussian observation ζ, and the resulting quadratic detectors allow to handle some important inference problems which are far beyond the grasp of “genuinely affine” detectors.

2.8.3.2 Quadratic lifting: Gaussian case

Given positive integer d, we define $\mathcal{E}^d$ as the linear space $\mathbf{R}^d \times \mathbf{S}^d$ equipped with the inner product

$$\langle (z, S), (z', S') \rangle = s^T z' + \frac{1}{2}\mathrm{Tr}(SS').$$

Note that the quadratic lifting $z \mapsto (z, zz^T)$ maps the space $\mathbf{R}^d$ into $\mathcal{E}^d$.

In the sequel, an instrumental role is played by the following result.

Proposition 2.8.7

(i) *Assume we are given*

- *a nonempty and bounded subset U of $\mathbf{R}^n$,*
- *a convex compact set $\mathcal{V}$ contained in the interior of the cone $\mathbf{S}^d_+$ of positive semidefinite $d \times d$ matrices*
- *a $d \times (n+1)$ matrix A.*

These data specify the family $\mathcal{G}_A[U, \mathcal{V}]$ of distributions of quadratic lifts $(\zeta, \zeta\zeta^T)$ of Gaussian random vectors $\zeta \sim \mathcal{N}(A[u;1], \Theta)$ stemming from $u \in U$ and $\Theta \in \mathcal{V}$.

Let us select somehow

1. *$\gamma \in (0,1)$,*
2. *convex compact subset $\mathcal{Z}$ of the set $\mathcal{Z}^n = \{Z \in \mathbf{S}^{n+1} : Z \succeq 0, Z_{n+1,n+1} = 1\}$ such that*
$$Z(u) := [u;1][u;1]^T \in \mathcal{Z} \ \forall u \in U, \tag{2.8.34}$$
3. *positive definite $d \times d$ matrix $\Theta_* \in \mathbf{S}^d_+$ and $\delta \in [0,2]$ such that*
$$\Theta_* \succeq \Theta \ \forall \Theta \in \mathcal{V} \ \& \ \|\Theta^{1/2}\Theta_*^{-1/2} - I_d\| \leq \delta \ \ \forall \Theta \in \mathcal{V}, \tag{2.8.35}$$
where $\|\cdot\|$ is the spectral norm,[24]

and set

$$\mathcal{H} = \mathcal{H}^\gamma := \{(h,H) \in \mathbf{R}^d \times \mathbf{S}^d : -\gamma\Theta_*^{-1} \preceq H \preceq \gamma\Theta_*^{-1}\}, \tag{2.8.36}$$

$$\begin{array}{rcl}\Phi_{A,\mathcal{Z}}(h,H;\Theta) & = & -\frac{1}{2}\ln\mathrm{Det}(I - \Theta_*^{1/2}H\Theta_*^{1/2}) + \frac{1}{2}\mathrm{Tr}([\Theta - \Theta_*]H) \\ & & + \frac{\delta(2+\delta)}{2(1-\|\Theta_*^{1/2}H\Theta_*^{1/2}\|)}\|\Theta_*^{1/2}H\Theta_*^{1/2}\|_F^2 \\ & & + \frac{1}{2}\phi_{\mathcal{Z}}\left(B^T\left[\left[\begin{array}{c|c} H & h \\ \hline h^T & \end{array}\right] + [H,h]^T[\Theta_*^{-1} - H]^{-1}[H,h]\right]B\right) : \mathcal{H} \times \mathcal{V} \to \mathbf{R},\end{array} \tag{2.8.37}$$

where B is given by

$$B = \left[\begin{array}{c} A \\ [0, ..., 0, 1] \end{array}\right], \tag{2.8.38}$$

the function

$$\phi_{\mathcal{Z}}(Y) := \max_{Z \in \mathcal{Z}} \mathrm{Tr}(ZY) \tag{2.8.39}$$

is the support function of $\mathcal{Z}$, and $\|\cdot\|_F$ is the Frobenius norm.

Function $\Phi_{A,\mathcal{Z}}$ is continuous on its domain, convex in $(h,H) \in \mathcal{H}$ and concave in $\Theta \in \mathcal{V}$, so that $(\mathcal{H}, \mathcal{V}, \Phi_{A,\mathcal{Z}})$ is a regular data. Besides this,

> *(#) Whenever $u \in \mathbf{R}^n$ is such that $[u;1][u;1]^T \in \mathcal{Z}$ and $\Theta \in \mathcal{V}$, the Gaussian random vector $\zeta \sim \mathcal{N}(A[u;1], \Theta)$ satisfies the relation*
> $$\forall (h,H) \in \mathcal{H}: \ \ \ln\left(\mathbf{E}_{\zeta\sim\mathcal{N}(A[u;1],\Theta)}\left\{e^{\frac{1}{2}\zeta^T H\zeta + h^T\zeta}\right\}\right) \leq \Phi_{A,\mathcal{Z}}(h,H;\Theta). \tag{2.8.40}$$

[24] It is easily seen that with $\delta = 2$, the second relation in (2.8.35) is satisfied for all Θ such that $0 \preceq \Theta \preceq \Theta_*$, so that the restriction $\delta \leq 2$ is w.l.o.g.

which combines with (2.8.34) *to imply that*

$$\mathcal{G}_A[U,\mathcal{V}] \subset \mathcal{S}[\mathcal{H},\mathcal{V},\Phi_{A,\mathcal{Z}}]. \tag{2.8.41}$$

In addition, $\Phi_{A,\mathcal{Z}}$ *is coercive in* (h,H)*:* $\Phi_{A,\mathcal{Z}}(h_i,H_i;\Theta)\to+\infty$ *as* $i\to\infty$ *whenever* $\Theta\in\mathcal{V}$, $(h_i,H_i)\in\mathcal{H}$ *and* $\|(h_i,H_i)\|\to\infty$, $i\to\infty$.

(ii) *Let two collections of entities from* (i), $(\mathcal{V}_\chi,\Theta_*^{(\chi)},\delta_\chi,\gamma_\chi,A_\chi,\mathcal{Z}_\chi)$, $\chi=1,2$, *with common* d *be given, giving rise to the sets* $\mathcal{H}_\chi$, *matrices* B_χ, *and functions* $\Phi_{A_\chi,\mathcal{Z}_\chi}(h,H;\Theta)$, $\chi=1,2$. *These collections specify the families of normal distributions*

$$\mathcal{G}_\chi=\{\mathcal{N}(v,\Theta):\Theta\in\mathcal{V}_\chi\ \&\ \exists u\in U: v=A_\chi[u;1]\},\ \chi=1,2.$$

Consider the convex-concave saddle point problem

$$\mathcal{SV}=\min_{(h,H)\in\mathcal{H}_1\cap\mathcal{H}_2}\ \max_{\Theta_1\in\mathcal{V}_1,\Theta_2\in\mathcal{V}_2}\underbrace{\frac{1}{2}\left[\Phi_{A_1,\mathcal{Z}_1}(-h,-H;\Theta_1)+\Phi_{A_2,\mathcal{Z}_2}(h,H;\Theta_2)\right]}_{\Phi(h,H;\Theta_1,\Theta_2)}. \tag{2.8.42}$$

A saddle point $(H_*,h_*;\Theta_1^*,\Theta_2^*)$ *in this problem does exist, and the induced quadratic detector*

$$\phi_*(\omega)=\frac{1}{2}\omega^TH_*\omega+h_*^T\omega+\underbrace{\frac{1}{2}[\Phi_{A_1,\mathcal{Z}_1}(-h_*,-H_*;\Theta_1^*)-\Phi_{A_2,\mathcal{Z}_2}(h_*,H_*;\Theta_2^*)]}_{a}, \tag{2.8.43}$$

when applied to the families of Gaussian distributions $\mathcal{G}_\chi$, $\chi=1,2$, *has the risk*

$$\mathrm{Risk}[\phi_*|\mathcal{G}_1,\mathcal{G}_2]\le\epsilon_\star:=\mathrm{e}^{\mathcal{SV}},$$

that is,

$$\begin{array}{lll}(a) & \int_{\mathbf{R}^d}\mathrm{e}^{-\phi_*(\omega)}P(d\omega)\le\epsilon_\star & \forall P\in\mathcal{G}_1,\\ (b) & \int_{\mathbf{R}^d}\mathrm{e}^{\phi_*(\omega)}P(d\omega)\le\epsilon_\star & \forall P\in\mathcal{G}_2.\end{array} \tag{2.8.44}$$

For proof, see Section 2.8.4.2.

Remark 2.8.1 *Note that the computational effort to solve* (2.8.42) *reduces dramatically in the "easy case" of the situation described in item (ii) of Proposition 2.8.7, specifically, in the case where*

- *the observations are direct, meaning that* $A_\chi[u;1]\equiv u$, $u\in\mathbf{R}^d$, $\chi=1,2$;
- *the sets* $\mathcal{V}_\chi$ *are comprised of positive definite diagonal matrices, and matrices* $\Theta_*^{(\chi)}$ *are diagonal as well,* $\chi=1,2$;
- *the sets* $\mathcal{Z}_\chi$, $\chi=1,2$, *are convex compact sets of the form*

  $$\mathcal{Z}_\chi=\{Z\in\mathbf{S}_+^{d+1}:Z\succeq0,\ \mathrm{Tr}(ZQ_j^\chi)\le q_j^\chi,\ 1\le j\le J_\chi\}$$

  with diagonal matrices Q_j^χ, [25] *and these sets intersect the interior of the positive semidefinite cone* $\mathbf{S}_+^{d+1}$.

In this case, the convex-concave saddle point problem (2.8.42) *admits a saddle point* $(h_*,H_*;\Theta_1^*,\Theta_2^*)$ *where* $h_*=0$ *and* H_* *is diagonal.*

[25] In terms of the sets U_χ, this assumption means that the latter sets are given by linear inequalities on the *squares* of entries in u,

Justifying the remark. In the easy case, we have $B_\chi = I_{d+1}$ and therefore

$$\begin{array}{rcl}
M_\chi(h,H) & := & B_\chi^T \left[\left[\begin{array}{c|c} H & h \\ \hline h^T & \end{array}\right] + [H,h]^T \left[[\Theta_*^{(\chi)}]^{-1} - H\right]^{-1} [H,h]\right] B_\chi \\
& = & \left[\begin{array}{c|c} H + H\left[[\Theta_*^{(\chi)}]^{-1} - H\right]^{-1} H & h + H[[\Theta_*^{(\chi)}]^{-1} - H]^{-1} h \\ \hline h^T + h^T\left[[\Theta_*^{(\chi)}]^{-1} - H\right]^{-1} H & h^T\left[[\Theta_*^{(\chi)}]^{-1} - H\right]^{-1} h \end{array}\right]
\end{array}$$

and

$$\begin{array}{rcl}
\phi_{\mathcal{Z}_\chi}(Z) & = & \max\limits_W \left\{\mathrm{Tr}(ZW) : W \succeq 0,\, \mathrm{Tr}(WQ_j^\chi) \leq q_j^\chi,\, 1 \leq j \leq J_\chi\right\} \\
& = & \min_\lambda \left\{\sum_j q_j^\chi \lambda_j : \lambda \geq 0,\, Z \preceq \sum_j \lambda_j Q_j^\chi\right\},
\end{array}$$

where the last equality is due to semidefinite duality[26]. From the second representation of $\phi_{\mathcal{Z}_\chi}(\cdot)$ and the fact that all Q_j^χ are diagonal it follows that $\phi_{\mathcal{Z}_\chi}(M_\chi(0,H)) \leq \phi_{\mathcal{Z}_\chi}(M_\chi(h,H))$ (indeed, with diagonal Q_j^χ, if λ is feasible for the minimization problem participating in the representation when $Z = M_\chi(h,H)$, it clearly remains feasible when Z is replaced with $M_\chi(0,H)$). This, in turn, combines straightforwardly with (2.8.37) to imply that when replacing h_* with 0 in a saddle point $(h_*, H_*; \Theta_1^*, \Theta_2^*)$ of (2.8.42), we end up with another saddle point of (2.8.42). In other words, when solving (2.8.42), we can from the very beginning set h to 0, thus converting (2.8.42) into the convex-concave saddle point problem

$$\mathcal{SV} = \min_{H:(0,H)\in\mathcal{H}_1\cap\mathcal{H}_2} \max_{\Theta_1\in\mathcal{V}_1,\Theta_2\in\mathcal{V}_2} \Phi(0,H;\Theta_1,\Theta_2). \tag{2.8.45}$$

Taking into account the fact that we are in the case where all matrices from the sets $\mathcal{V}_\chi$, same as the matrices $\Theta_*^{(\chi)}$ and all the matrices Q_j^χ, $\chi = 1,2$, are diagonal, it is immediate to verify that if E is a $d \times d$ diagonal matrix with diagonal entries ±1, then $\Phi(0,H;\Theta_1,\Theta_2) = \Phi(0,EHE;\Theta_1,\Theta_2)$. Due to convexity-concavity of Φ this implies that (2.8.45) admits a saddle point $(0,H_*;\Theta_1^*,\Theta_2^*)$ with H_* invariant w.r.t. transformations $H_* \mapsto EH_*E$ with the above E, that is, with diagonal H_*, as claimed. □

2.8.3.3 Quadratic lifting – does it help?

Assume that for $\chi = 1,2$, we are given

- affine mappings $u \mapsto \mathcal{A}_\chi(u) = A_\chi[u;1] : \mathbf{R}^{n_\chi} \to \mathbf{R}^d$,
- nonempty convex compact sets $U_\chi \subset \mathbf{R}^{n_\chi}$,
- nonempty convex compact sets $\mathcal{V}_\chi \subset \mathrm{int}\,\mathbf{S}_+^d$.

These data define families $\mathcal{G}_\chi$ of Gaussian distributions on $\mathbf{R}^d$: $\mathcal{G}_\chi$ is comprised of all distributions $\mathcal{N}(\mathcal{A}_\chi(u),\Theta)$ with $u \in U_\chi$ and $\Theta \in \mathcal{V}_\chi$. The data define also families $\mathcal{SG}_\chi$ of sub-Gaussian distributions on $\mathbf{R}^d$: $\mathcal{SG}_\chi$ is comprised of all sub-Gaussian distributions with parameters $(\mathcal{A}_\chi(u),\Theta)$ with $(u,\Theta) \in U_\chi \times \mathcal{V}_\chi$.

Assume we observe random variable $\zeta \in \mathbf{R}^d$ drawn from a distribution P known to belong to $\mathcal{G}_1 \cup \mathcal{G}_2$, and our goal is to decide from stationary K-repeated version of our observation on the pair of hypotheses $H_\chi : P \in \mathcal{G}_\chi$, $\chi = 1,2$; we refer to this situation as to *Gaussian case*. We could also speak about *sub-Gaussian case*, where the hypotheses we would decide upon state that $P \in \mathcal{SG}_\chi$. In retrospect, all we are about to establish for the Gaussian case can be word by word repeated for the sub-Gaussian one, so that from now on, we assume we are in Gaussian case.

At present, we have developed two approaches to building detector-based tests for H_1, H_2:

[26] see Section 4.1.2 (or Section E.3 for more details).

A. Utilizing the *affine* in ζ detector $\phi_{\rm aff}$ given by solution to the saddle point problem (see (2.8.25), (2.8.26) and set $\theta_\chi = \mathcal{A}_\chi(u_\chi)$ with u_χ running through U_χ)

$$\mathrm{SadVal}_{\rm aff} = \min_{h\in\mathbf{R}^d} \max_{\substack{u_1\in U_1,u_2\in U_2\\ \Theta_1\in\mathcal{V}_1,\Theta_2\in\mathcal{V}_2}} \frac{1}{2}\left[h^T[\mathcal{A}_2(u_2)-\mathcal{A}_1(u_1)] + \frac{1}{2}h^T\left[\Theta_1+\Theta_2\right]h\right]; \qquad (2.8.46)$$

this detector satisfies risk bound

$$\mathrm{Risk}[\phi_{\rm aff}|\mathcal{G}_1,\mathcal{G}_2] \leq \exp\{\mathrm{SadVal}_{\rm aff}\}. \qquad (2.8.47)$$

Q. Utilizing the quadratic in ζ detector $\phi_{\rm lift}$ given by Proposition 2.8.7.ii, with the risk bound

$$\mathrm{Risk}[\phi_{\rm lift}|\mathcal{G}_1,\mathcal{G}_2] \leq \exp\{\mathrm{SadVal}_{\rm lift}\}, \qquad (2.8.48)$$

with $\mathrm{SadVal}_{\rm lift}$ given by (2.8.42).

A natural question is, which one of these options results in a better risk bound. Note that we cannot just say "clearly, the second option is better, since there are more quadratic detectors than affine ones" – the difficulty is that the key, in the context of Proposition 2.8.7, relation (2.8.40) is inequality rather than equality[27]. We are about to show that under reasonable assumptions, the second option indeed is better:

Proposition 2.8.8 *In the situation in question, assume that the sets $\mathcal{V}_\chi$, $\chi=1,2$, contain the $\succeq$-largest elements, and that these elements are taken as the matrices $\Theta_*^{(\chi)}$ participating in Proposition 2.8.7.ii. Let, further, the convex compact sets $\mathcal{Z}_\chi$ participating in Proposition 2.8.7.ii satisfy*

$$\mathcal{Z}_\chi \subset \bar{\mathcal{Z}}_\chi := \{Z = \left[\begin{array}{c|c} W & u\\ \hline u^T & 1\end{array}\right] \succeq 0, u\in U_\chi\} \qquad (2.8.49)$$

(this assumption does not restrict generality, since $\bar{\mathcal{Z}}_\chi$ is, along with U_χ, a closed convex set which clearly contains all matrices $[u;1][u;1]^T$ with $u\in U_\chi$). Then

$$\mathrm{SadVal}_{\rm lift} \leq \mathrm{SadVal}_{\rm aff}, \qquad (2.8.50)$$

that is, option **Q** *is at least as efficient as option* **A**.

Proof. Let $A_\chi = [\bar{A}_\chi, a_\chi]$. Looking at (2.8.25), where one should substitute $\theta_\chi = \mathcal{A}_\chi(u_\chi)$ with u_χ running through U_χ) and taking into account that $\Theta_\chi \preceq \Theta_*^{(\chi)} \in \mathcal{V}_\chi$ when $\Theta_\chi\in\mathcal{V}_\chi$, we conclude that

$$\mathrm{SadVal}_{\rm aff} = \min_h \max_{u_1\in U_1,u_2\in U_2} \frac{1}{2}\left[h^T[\bar{A}_2u_2 - \bar{A}_1u_1 + a_2 - a_1] + \frac{1}{2}h^T\left[\Theta_*^{(1)}+\Theta_*^{(2)}\right]h\right]. \qquad (2.8.51)$$

[27] One cannot make (2.8.40) an equality by redefining the right hand side function – it will lose the required in our context convexity-concavity properties.

			unrestricted		
ρ	σ_1	σ_2	H and h	$H=0$	$h=0$
0.5	2	2	0.31	0.31	1.00
0.5	1	4	0.24	0.39	0.62
0.01	1	4	0.41	1.00	0.41

Table 2.2: Risk of quadratic detector $\phi(\zeta) = h^T\zeta + \frac{1}{2}\zeta^T H\zeta + \varkappa$

At the same time, we have by Proposition 2.8.7.ii:

$$
\begin{array}{rcl}
\mathrm{SadVal}_{\mathrm{lift}} & = & \min\limits_{(h,H)\in\mathcal{H}_1\cap\mathcal{H}_2} \max\limits_{\Theta_1\in\mathcal{V}_1,\Theta_2\in\mathcal{V}_2} \frac{1}{2}\left[\Phi_{A_1,\mathcal{Z}_1}(-h,-H;\Theta_1) + \Phi_{A_2,\mathcal{Z}_2}(h,H;\Theta_2)\right] \\
& \leq & \min\limits_{h\in\mathbf{R}^d} \max\limits_{\Theta_1\in\mathcal{V}_1,\Theta_2\in\mathcal{V}_2} \frac{1}{2}\left[\Phi_{A_1,\mathcal{Z}_1}(-h,0;\Theta_1) + \Phi_{A_2,\mathcal{Z}_2}(h,0;\Theta_2)\right] \\
& = & \min\limits_{h\in\mathbf{R}^d} \max\limits_{\Theta_1\in\mathcal{V}_1,\Theta_2\in\mathcal{V}_2} \frac{1}{2}\Bigg[\frac{1}{2}\max\limits_{Z_1\in\mathcal{Z}_1} \mathrm{Tr}\left(Z_1\left[\begin{array}{c|c} & -\bar{A}_1^T h \\ \hline -h^T\bar{A}_1 & -2h^T a_1 + h^T\Theta_*^{(1)}h\end{array}\right]\right) \\
& & \qquad\qquad + \frac{1}{2}\max\limits_{Z_2\in\mathcal{Z}_2} \mathrm{Tr}\left(Z_2\left[\begin{array}{c|c} & \bar{A}_2^T h \\ \hline h^T\bar{A}_2 & 2h^T a_2 + h^T\Theta_*^{(2)}h\end{array}\right]\right)\Bigg] \\
& & \qquad\qquad\qquad\text{[by direct computation utilizing (2.8.37)]} \\
& \leq & \min\limits_{h\in\mathbf{R}^d} \frac{1}{2}\Bigg[\frac{1}{2}\max\limits_{u_1\in U_1}\left[-2u_1^T\bar{A}_1^T h - 2a_1^T h + h^T\Theta_*^{(1)}h\right] + \\
& & \qquad\qquad \frac{1}{2}\max\limits_{u_2\in U_2}\left[2u_2^T\bar{A}_2^T h + 2a_2^T h + h^T\Theta_*^{(2)}h\right]\Bigg] \\
& & \qquad\qquad\qquad\text{[due to (2.8.49)]} \\
& = & \mathrm{SadVal}_{\mathrm{aff}},
\end{array}
$$

where the concluding equality is due to (2.8.51). □

Numerical illustration. To get an impression of the performance of quadratic detectors as compared to affine ones under the premise of Proposition 2.8.8, we present here the results of experiment where $U_1 = U_1^\rho = \{u \in \mathbf{R}^{12} : u_i \geq \rho, 1 \leq i \leq 12\}$, $U_2 = U_2^\rho = -U_1^\rho$, $A_1 = A_2 \in \mathbf{R}^{8\times 13}$, and $\mathcal{V}_\chi = \{\Theta_*^{(\chi)} = \sigma_\chi^2 I_8\}$ are singletons. The risks of affine, quadratic and "purely quadratic" (with h set to 0) detectors on the associated families $\mathcal{G}_1, \mathcal{G}_2$ are given in Table 2.2.

We see that

- when deciding on families of Gaussian distributions with common covariance matrix and expectations varying in associated with the families convex sets, passing from affine detectors described by Proposition 2.8.5 to quadratic detectors does not affect the risk (first row in the table). This should be expected: we are in the scope of Gaussian o.s., where minimum risk affine detectors are optimal among all possible detectors.

- when deciding on families of Gaussian distributions in the case where distributions from different families can have close expectations (third row in the table), affine detectors are useless, while the quadratic ones are not, provided that $\Theta_*^{(1)}$ differs from $\Theta_*^{(2)}$. This is how it should be – we are in the case where the first moments of the distribution of observation bear no definitive information on the family this distribution belongs to, which makes affine detectors useless. In contrast, quadratic detectors are able to utilize information (valuable when $\Theta_*^{(1)} \neq \Theta_*^{(2)}$) "stored" in the second moments of the observation.

- "in general" (second row in the table), both affine and purely quadratic components in a quadratic detector are useful; suppressing one of them can increase significantly the attainable risk.

2.8.3.4 Quadratic lifting: sub-Gaussian case

Sub-Gaussian version of Proposition 2.8.7 is as follows:

Proposition 2.8.9

(i) *Assume we are given*

- *a nonempty and bounded subset U of $\mathbf{R}^n$,*
- *a convex compact set $\mathcal{V}$ contained in the interior of the cone $\mathbf{S}^d_+$ of positive semidefinite $d\times d$ matrices*
- *a $d\times(n+1)$ matrix A.*

These data specify the family $\mathcal{SG}_A[U,\mathcal{V}]$ of distributions of quadratic lifts $(\zeta,\zeta\zeta^T)$ of sub-Gaussian random vectors ζ with sub-Gaussianity parameters $A[u;1],\Theta$ stemming from $u\in U$ and $\Theta\in\mathcal{V}$.

Let us select somehow

1. *reals γ,γ^+ such that $0<\gamma<\gamma^+<1$,*
2. *convex compact subset $\mathcal{Z}$ of the set $\mathcal{Z}^n=\{Z\in\mathbf{S}^{n+1}: Z\succeq 0, Z_{n+1,n+1}=1\}$ such that relation (2.8.34) takes place,*
3. *positive definite $d\times d$ matrix $\Theta_*\in\mathbf{S}^d_+$ and $\delta\in[0,2]$ such that (2.8.35) takes place.*

These data specify the closed convex sets

$$\begin{array}{rcl}\mathcal{H} &=& \mathcal{H}^\gamma := \{(h,H)\in\mathbf{R}^d\times\mathbf{S}^d: -\gamma\Theta_*^{-1}\preceq H\preceq\gamma\Theta_*^{-1}\},\\ \widehat{\mathcal{H}} &=& \widehat{\mathcal{H}}^{\gamma,\gamma^+} = \left\{(h,H,G)\in\mathbf{R}^d\times\mathbf{S}^d\times\mathbf{S}^d: \left\{\begin{array}{l}-\gamma\Theta_*^{-1}\preceq H\preceq\gamma\Theta_*^{-1}\\ 0\preceq G\preceq\gamma^+\Theta_*^{-1},\ H\preceq G\end{array}\right\}\right.\end{array}\tag{2.8.52}$$

and the functions

$$\begin{array}{rcl}\Psi_{A,\mathcal{Z}}(h,H,G) &=& -\frac{1}{2}\ln\mathrm{Det}(I-\Theta_*^{1/2}G\Theta_*^{1/2})+\frac{1}{2}\phi_{\mathcal{Z}}\left(B^T\left[\left[\begin{array}{c|c}H & h\\ \hline h^T & \end{array}\right]+[H,h]^T[\Theta_*^{-1}-G]^{-1}[H,h]\right]B\right):\\ && \qquad\qquad\qquad\qquad \widehat{\mathcal{H}}\times\mathcal{Z}\to\mathbf{R},\\ \Psi^\delta_{A,\mathcal{Z}}(h,H,G;\Theta) &=& -\frac{1}{2}\ln\mathrm{Det}(I-\Theta_*^{1/2}G\Theta_*^{1/2})+\frac{1}{2}\mathrm{Tr}([\Theta-\Theta_*]G)+\frac{\delta(2+\delta)}{2(1-\|\Theta_*^{1/2}G\Theta_*^{1/2}\|)}\|\Theta_*^{1/2}G\Theta_*^{1/2}\|_F^2\\ && +\frac{1}{2}\phi_{\mathcal{Z}}\left(B^T\left[\left[\begin{array}{c|c}H & h\\ \hline h^T & \end{array}\right]+[H,h]^T[\Theta_*^{-1}-G]^{-1}[H,h]\right]B\right):\widehat{\mathcal{H}}\times\{0\preceq\Theta\preceq\Theta_*\}\to\mathbf{R},\\ \Phi_{A,\mathcal{Z}}(h,H) &=& \min\limits_G\left\{\Psi_{A,\mathcal{Z}}(h,H,G):(h,H,G)\in\widehat{\mathcal{H}}\right\}:\mathcal{H}\to\mathbf{R},\\ \Phi^\delta_{A,\mathcal{Z}}(h,H;\Theta) &=& \min\limits_G\left\{\Psi^\delta_{A,\mathcal{Z}}(h,H,G;\Theta):(h,H,G)\in\widehat{\mathcal{H}}\right\}:\mathcal{H}\times\{0\preceq\Theta\preceq\Theta_*\}\to\mathbf{R},\end{array}\tag{2.8.53}$$

where B is given by (2.8.38) and $\phi_{\mathcal{Z}}(\cdot)$ is the support function of $\mathcal{Z}$ given by (2.8.39).

Function $\Phi_{A,\mathcal{Z}}(h,H)$ is convex and continuous on its domain, while function $\Phi^\delta_{A,\mathcal{Z}}(h,H;\Theta)$ is continuous on its domain, convex in $(h,H)\in\mathcal{H}$ and concave in $\Theta\in\{0\preceq\Theta\preceq\Theta_*\}$. Besides this,

> (##) *Whenever $u\in\mathbf{R}^n$ is such that $[u;1][u;1]^T\in\mathcal{Z}$ and $\Theta\in\mathcal{V}$, the sub-Gaussian, with parameters $(A[u;1],\Theta)$, random vector ζ satisfies the relation*
>
> $$\begin{array}{ll}\forall(h,H)\in\mathcal{H}: & \\ (a) & \ln\left(\mathbf{E}_\zeta\left\{\mathrm{e}^{\frac{1}{2}\zeta^TH\zeta+h^T\zeta}\right\}\right)\le\Phi_{A,\mathcal{Z}}(h,H),\\ (b) & \ln\left(\mathbf{E}_\zeta\left\{\mathrm{e}^{\frac{1}{2}\zeta^TH\zeta+h^T\zeta}\right\}\right)\le\Phi^\delta_{A,\mathcal{Z}}(h,H;\Theta).\end{array}\tag{2.8.54}$$

which combines with (2.8.34) to imply that

$$\mathcal{SG}_A[U,\mathcal{V}] \subset \mathcal{S}[\mathcal{H},\mathcal{V},\Phi_{A,\mathcal{S}}]. \tag{2.8.55}$$

In addition, $\Phi_{A,\mathcal{Z}}$ and $\Phi^\delta_{A,\mathcal{Z}}$ are coercive in (h,H): $\Phi_{A,\mathcal{Z}}(h_i,H_i)\to+\infty$ and $\Phi^\delta_{A,\mathcal{Z}}(h_i,H_i;\Theta)\to+\infty$ as $i\to\infty$ whenever $\Theta\in\mathcal{V}$, $(h_i,H_i)\in\mathcal{H}$ and $\|(h_i,H_i)\|\to\infty$, $i\to\infty$.

(ii) *Let two collections of data from* (i)*: $(\mathcal{V}_\chi,\Theta^{(\chi)}_*,\delta_\chi,\gamma_\chi,\gamma^+_\chi,A_\chi,\mathcal{Z}_\chi)$, $\chi=1,2$, with common d be given, giving rise to the sets $\mathcal{H}_\chi$, matrices B_χ, and functions $\Phi_{A_\chi,\mathcal{Z}_\chi}(h,H)$, $\Phi^{\delta_\chi}_{A_\chi,\mathcal{Z}_\chi}(h,H;\Theta)$, $\chi=1,2$. These collections specify the families $\mathcal{SG}_\chi=\mathcal{SG}_{A_\chi}[U_\chi,\mathcal{V}_\chi]$ of sub-Gaussian distributions.*

Consider the convex-concave saddle point problem

$$\mathcal{SV} = \min_{(h,H)\in\mathcal{H}_1\cap\mathcal{H}_2}\ \max_{\Theta_1\in\mathcal{V}_1,\Theta_2\in\mathcal{V}_2} \underbrace{\frac{1}{2}\left[\Phi^{\delta_1}_{A_1,\mathcal{Z}_1}(-h,-H;\Theta_1)+\Phi^{\delta_2}_{A_2,\mathcal{Z}_2}(h,H;\Theta_2)\right]}_{\Phi^{\delta_1,\delta_2}(h,H;\Theta_1,\Theta_2)}. \tag{2.8.56}$$

A saddle point $(H_*,h_*;\Theta^*_1,\Theta^*_2)$ in this problem does exist, and the induced quadratic detector

$$\phi_*(\omega)=\frac{1}{2}\omega^TH_*\omega+h_*^T\omega+\underbrace{\frac{1}{2}\left[\Phi^{\delta_1}_{A_1,\mathcal{Z}_1}(-h_*,-H_*;\Theta^*_1)-\Phi^{\delta_2}_{A_2,\mathcal{Z}_2}(h_*,H_*;\Theta^*_2)\right]}_{a}, \tag{2.8.57}$$

when applied to the families of sub-Gaussian distributions $\mathcal{SG}_\chi$, $\chi=1,2$, has the risk

$$\mathrm{Risk}[\phi_*|\mathcal{SG}_1,\mathcal{SG}_2]\le\epsilon_\star:=\mathrm{e}^{\mathcal{SV}}.$$

As a result,

$$\begin{array}{lrl}(a) & \int_{\mathbf{R}^d}\mathrm{e}^{-\phi_*(\omega)}P(d\omega)\le\epsilon_\star & \forall P\in\mathcal{SG}_1,\\ (b) & \int_{\mathbf{R}^d}\mathrm{e}^{\phi_*(\omega)}P(d\omega)\le\epsilon_\star & \forall P\in\mathcal{SG}_2.\end{array} \tag{2.8.58}$$

Similarly, the convex minimization problem

$$\mathrm{Opt}=\min_{(h,H)\in\mathcal{H}_1\cap\mathcal{H}_2}\underbrace{\frac{1}{2}\left[\Phi_{A_1,\mathcal{Z}_1}(-h,-H)+\Phi_{A_2,\mathcal{Z}_2}(h,H)\right]}_{\Phi(h,H)}. \tag{2.8.59}$$

is solvable, and the induced by its optimal solution (h_*,H_*) quadratic detector

$$\phi_*(\omega)=\frac{1}{2}\omega^TH_*\omega+h_*^T\omega+\underbrace{\frac{1}{2}\left[\Phi_{A_1,\mathcal{Z}_1}(-h_*,-H_*)-\Phi_{A_2,\mathcal{Z}_2}(h_*,H_*)\right]}_{a}, \tag{2.8.60}$$

when applied to the families of sub-Gaussian distributions $\mathcal{SG}_\chi$, $\chi=1,2$, has the risk

$$\mathrm{Risk}[\phi_*|\mathcal{SG}_1,\mathcal{SG}_2]\le\epsilon_\star:=\mathrm{e}^{\mathrm{Opt}},$$

so that for just defined ϕ_* and $\epsilon_\star$ relation (2.8.59) takes place.

For proof, see Section 2.8.4.3.

Remark 2.8.2 *Proposition 2.8.9 offers two options for building quadratic detectors for the families $\mathcal{SG}_1$, $\mathcal{SG}_2$, those based on saddle point of (2.8.56) and on optimal solution to (2.8.59). Inspecting the proof, the number of options can be increased to 4: we can replace any one of the functions $\Phi^{\delta_\chi}_{A_\chi,\mathcal{Z}_\chi}$, $\chi=1,2$ (or both these functions simultaneously) with $\Phi_{A_\chi,\mathcal{Z}_\chi}$. The second of the original two options is exactly what we get when replacing both $\Phi^{\delta_\chi}_{A_\chi,\mathcal{Z}_\chi}$, $\chi=1,2$, with $\Phi_{A_\chi,\mathcal{Z}_\chi}$. It is easily seen that depending on the data, every one of these 4 options can be the best – result in the smallest risk bound. Thus, it makes sense to keep all these options in mind and to use the one which, under the circumstances, results in the best risk bound. Note that the risk bounds are efficiently computable, so that identifying the best option is easy.*

2.8.3.5 Generic application: quadratically constrained hypotheses

Propositions 2.8.7, 2.8.9 operate with Gaussian/sub-Gaussian observations ζ with matrix parameters Θ running through convex compact subsets $\mathcal{V}$ of $\mathrm{int}\,\mathbf{S}^d_+$, and means of the form $A[u;1]$, with "signals" u running through given sets $U \subset \mathbf{R}^n$. The constructions, however, involved additional entities – convex compact sets $\mathcal{Z} \subset \mathcal{Z}^n := \{Z \in \mathbf{S}^{n+1}_+ : Z_{n+1,n+1} = 1\}$ containing quadratic lifts $[u;1][u;1]^T$ of all signals $u \in U$; other things being equal, the smaller is $\mathcal{Z}$, the smaller is the associated function $\Phi_{A,\mathcal{Z}}$ (or $\Phi^\delta_{A,\mathcal{Z}}$), and consequently, the smaller are the (upper bounds on the) risks of quadratic in ζ detectors we end up with. In order to apply these propositions, we should understand how to build the required sets $\mathcal{Z}$ in an "economical" way. There exists a relatively simple case when it is easy to get reasonable candidates to the role of $\mathcal{Z}$ – the case of *quadratically constrained* signal set U:

$$U = \{u \in \mathbf{R}^n : f_k(u) := u^TQ_ku + 2q_k^Tu \leq b_k,\ 1 \leq k \leq K\}. \tag{2.8.61}$$

Indeed, the constraints $f_k(u) \leq b_k$ are just linear constraints on the quadratic lifting $[u;1][u;1]^T$ of u:

$$u^TQ_ku + 2q_k^Tu \leq b_k \Leftrightarrow \mathrm{Tr}(F_k[u;1][u;1]^T) \leq b_k,\ F_k = \left[\begin{array}{c|c} Q_k & q_k \\ \hline q_k^T & \end{array}\right] \in \mathbf{S}^{n+1}.$$

Consequently, in the case of (2.8.61), the simplest candidate on the role of $\mathcal{Z}$ is the set

$$\mathcal{Z} = \{Z \in \mathbf{S}^n : Z \succeq 0, Z_{n+1,n+1} = 1, \mathrm{Tr}(F_kZ) \leq b_k,\ 1 \leq k \leq K\}. \tag{2.8.62}$$

This set clearly is closed and convex (the latter – even when U itself is not convex), and indeed contains the quadratic lifts $[u;1][u;1]^T$ of all points $u \in U$. We need also the compactness of $\mathcal{Z}$; the latter definitely takes place when the quadratic constraints describing U contain constraint of the form $u^Tu \leq R^2$, which, in turn, can be ensured, basically "for free," when U is bounded. It should be stressed that the "ideal" choice of $\mathcal{Z}$ would be the convex hull $\mathcal{Z}[U]$ of all rank 1 matrices $[u;1][u;1]^T$ with $u \in U$ – this definitely is the smallest convex set which contains the quadratic lifts of all points from U; moreover, $\mathcal{Z}[U]$ is closed and bounded, provided U is so. The difficulty is that $\mathcal{Z}[U]$ can be computationally intractable (and thus useless in our context) already for pretty simple sets U of the form (2.8.61). The set (2.8.62) is a simple outer approximation of $\mathcal{Z}[U]$, and this approximation can be very loose; for example, when $U = \{u : -1 \leq u_k \leq 1, 1 \leq k \leq d\}$ is just the unit box in $\mathbf{R}^d$, the set (2.8.62) is

$$\{Z \in \mathbf{S}^{n+1} : Z \succeq 0, Z_{n+1,n+1} = 1, |Z_{k,n+1}| \leq 1,\ 1 \leq k \leq n\};$$

this set even is not bounded, while $\mathcal{Z}[U]$ clearly is bounded. There is, essentially, just one generic case when the set (2.8.62) is *exactly equal* to $\mathcal{Z}[U]$ – the case where

$$U = \{u : u^TQu \leq c\}, Q \succ 0$$

is an ellipsoid centered at the origin; the fact that in this case the set given by (2.8.62) is *exactly* $\mathcal{Z}[U]$ is a consequence of what is called $\mathcal{S}$-Lemma.

The fact that, in general, the set $\mathcal{Z}$ could be a very loose outer approximation of $\mathcal{Z}[U]$ does not mean that we cannot improve this construction. As an instructive example, let $U = \{u \in \mathbf{R}^n : \|u\|_\infty \leq 1\}$. We get a much better that above approximation of $\mathcal{Z}[U]$ when applying (2.8.62) to equivalent description of the box by quadratic constraints:

$$U := \{u \in \mathbf{R}^n : \|u\|_\infty \leq 1\} = \{u \in \mathbf{R}^n : u_k^2 \leq 1, 1 \leq k \leq n\}.$$

Applying recipe (2.8.62) to the second description of U, we arrive at a significantly less conservative outer approximation of $\mathcal{Z}[U]$, specifically,

$$\mathcal{Z} = \{Z \in \mathbf{S}^{n+1} : Z \succeq 0, Z_{n+1,n+1} = 1, Z_{kk} \leq 1,\ 1 \leq k \leq n\}.$$

Not only the resulting set $\mathcal{Z}$ is bounded; we can get a reasonable "upper bound" on the discrepancy between $\mathcal{Z}$ and $\mathcal{Z}[U]$. Namely, denoting by Z^o the matrix obtained from a symmetric $n \times n$ matrix Z by zeroing out the South-Eastern entry (the one in the cell $(n+1, n+1)$) and keeping the remaining entries intact, we have

$$\mathcal{Z}^o[U] := \{Z^o : Z \in \mathcal{Z}[U]\} \subset \mathcal{Z}^o := \{Z^o : Z \in \mathcal{Z}\} \subset O(1)\ln(n+1)\mathcal{Z}^o.$$

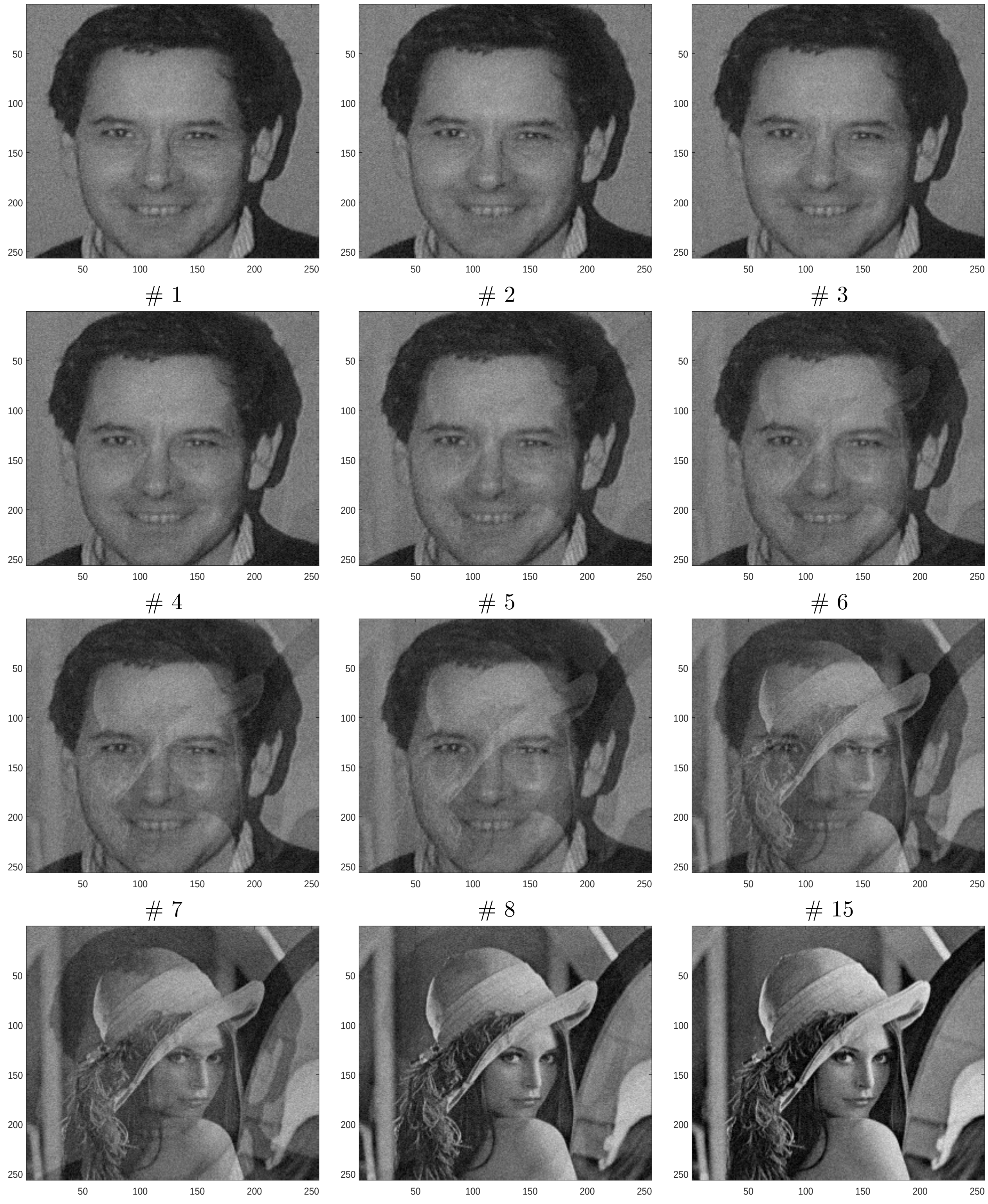
This is a particular case of a general result (going back to [124]; we shall get this result as a byproduct of our forthcoming considerations, specifically, Proposition 4.3.3) as follows:

> Let U be a bounded set given by a system of convex quadratic constraints without linear terms:
>
> $$U = \{u \in \mathbf{R}^n : u^T Q_k u \leq c_k,\, 1 \leq k \leq K\},\, Q_k \succeq 0, 1 \leq k \leq K,$$
>
> and let $\mathcal{Z}$ be the associated set (2.8.62):
>
> $$\mathcal{Z} = \{Z \in \mathbf{S}^{n+1} : Z \succeq 0, Z_{n+1,n+1} = 1,\, \mathrm{Tr}(Z\mathrm{Diag}\{Q_k, 1\}) \leq c_k,\, 1 \leq k \leq K\}$$
>
> Then
>
> $$\mathcal{Z}^o[U] := \{Z^o : Z \in \mathcal{Z}[U]\} \subset \mathcal{Z}^o := \{Z^o : Z \in \mathcal{Z}\} \subset 4\ln(5(K+1))\mathcal{Z}^o[U].$$
>
> Note that when $K = 1$ (i.e., U is an ellipsoid centered at the origin), the factor $4\ln(5(K+1))$, as it was already mentioned, can be replaced by 1.

One can think that the factor $4\ln(5(K+1))$ is too large to be of interest; well, this is nearly the best factor one can get under the circumstances, and a nice fact is that the factor is "nearly independent" of K.

Finally, we remark that, same as in the case of a box, we can try to reduce the conservatism of outer approximation (2.8.62) of $\mathcal{Z}[U]$ by passing from the initial description of U to an equivalent one. The standard recipe here is to replace linear constraints in the description of U by their quadratic consequences; for example, we can augment a pair of linear constraints $q_i^T u \leq c_i$, $q_j^T u \leq c_j$, assuming there is such a pair, with the quadratic constraint $(c_i - q_i^T u)(c_j - q_j^T u) \geq 0$. While this constraint is redundant, as far as the description of U itself is concerned, adding this constraint reduces, and sometimes significantly, the set given by (2.8.62). Informally speaking, transition from (2.8.61) to (2.8.62) is by itself "too stupid" to utilize the fact (known to every kid) that the product of two nonnegative quantities is nonnegative; when augmenting linear constraints in the description of U by their pairwise products, we somehow compensate for this stupidity. Unfortunately, "computationally tractable" assistance of this type perhaps allows to reduce the conservatism of (2.8.62), but usually does not allow to eliminate it completely: a grave "fact of life" is that even in the case of unit box U, the set $\mathcal{Z}[U]$ is computationally intractable. Scientifically speaking: maximizing quadratic forms over the unit box U provably is an NP-hard problem; were we able to get a computationally tractable description of $\mathcal{Z}[U]$, we would be able to solve this NP-hard problem efficiently, implying that P=NP. While we do not know for sure that the latter is not the case, "the informal odds" are strongly against this possibility.

The bottom line is that while the approach we are discussing in *some* situations could result in quite conservative tests, "some" is by far not the same as "always;" on the positive side, this approach allows to process some important problems. We are about to present a simple and instructive illustration.



1   # 2   # 3

4   # 5   # 6

7   # 8   # 15

20   # 28   # 34

Figure 2.5: Frames from a “movie”

2.8.3.6 Simple change detection

On Figure 2.5, you see a sample of frames from a "movie" where noisy picture of a gentleman gradually transforms into noisy picture of a lady; several initial frames differ just by realizations of noise, and starting with some instant, the "signal" (the deterministic component of the image) starts to drift from the gentleman towards the lady. What, in your opinion, is the change point – the first time instant where the signal component of the image differs from the signal component of the initial image?

A simple model of the situation is as follows: we observe, one by one, vectors (in fact, 2D arrays, but we can "vectorize" them)

$$\omega_t = x_t + \xi_t,\ t = 1, 2, ..., K, \tag{2.8.63}$$

where x_t are deterministic components of the observations and ξ_t are random noises. It may happen that for some $\tau \in \{2, 3, ..., K\}$, the vectors x_t are independent of t when $t < \tau$, and x_τ differs from $x_{\tau-1}$ ("τ is a change point"); if it is the case, τ is uniquely defined by $x^K = x_1, ..., x_K$. An alternative is that x_t is independent of t, $1 \leq t \leq K$ ("no change"). The goal is to decide, based on observation $\omega^K = (\omega_1, .., \omega_K)$, whether there was a change point, and if yes, then, perhaps, to localize it.

The model we have just described is the simplest case of "change detection," where, given noisy observations on some time horizon, one is interested to detect a "change" in some time series underlying the observations. In our simple model, this time series is comprised by deterministic components x_t of observations, and "change at time τ" is understood in the most straightforward way - as the fact that x_τ differs from equal to each other preceding x_t's. In more complicated situations, our observations are obtained from the underlying time series $\{x_t\}$ by a non-anticipative transformation, like

$$\omega_t = \sum_{s=1}^{t} A_{ts} x_s + \xi_t,\ t = 1, ..., K,$$

and we still want to detect the change, if any, in the time series $\{x_t\}$. As an instructive example, consider observations, taken along equidistant time grid, of the positions of an aircraft which "normally" flies with constant velocity, but at some time instant can start to maneuver. In this situation, the underlying time series is comprised of the velocities of the aircraft at consecutive time instants, observations are obtained from this time series by integration, and to detect a maneuver means to detect that on the observation horizon, there was a change in the series of velocities.

Change detection is the subject of huge literature dealing with a wide range of models differing from each other in

- whether we deal with direct observations of the time series of interest, as in (2.8.63), or with indirect ones (in the latter case, there is a wide spectrum of options related to how the observations depend on the underlying time series),
- what are assumptions on noise,
- what happens with x_t's after the change takes place – do they jump from their common value prior to time τ to a new common value starting with this time, or start to depend on time (and if yes, then how), etc., etc.

A significant role in change detection is played by hypothesis testing; as far as affine/quadratic-detector-based techniques developed in this Section are concerned, their applications in the context of change detection are discussed in [42]. In what follows, we focus on the simplest of these applications.

Situation and goal. We consider the situation as follows:

1. Our observations are given by (2.8.63) with independent across $t = 1, ..., K$ noises $\xi_t \sim \mathcal{N}(0, \sigma^2 I_d)$. We do not known σ a priori, what we know is that σ is independent of t and belongs to a given segment $[\underline{\sigma}, \overline{\sigma}]$, with $0 < \underline{\sigma} \leq \overline{\sigma}$;

2. Observations (2.8.63) arrive one by one, so that at time t, $2 \leq t \leq K$ we have at our disposal observation $\omega^t = \omega_1, ..., \omega_t$. Our goal is to build a system of inferences $\mathcal{T}_t$, $2 \leq t \leq K$, such that $\mathcal{T}_t$ as applied to ω^t either infers that there was a change at time t or earlier, in which case we terminate, or infers that so far there was no change, in which case we either proceed to time $t+1$ (if $t < K$), or terminate (if $t = K$) with "no change" conclusion.

   We are given $\epsilon \in (0,1)$ and want from our collection of inferences to make the probability of *false alarm* (i.e., terminating somewhere on time horizon $2, 3, ..., K$ with "there was a change" conclusion in the situation when there was no change: $x_1 = ... = x_K$) at most ϵ. Under this restriction, we want to make as small as possible the probability of a *miss* (of not detecting the change at all in the situation where there was a change).

The "small probability of a miss" desire should be clarified. When the noise is nontrivial, we have no chances to detect very small changes *and* respect the bound on the probability of false alarm. A realistic goal is to make as small as possible the probability of missing a *not too small* change, which can be formalized as follows. Given $\rho > 0$, and tolerances $\epsilon, \varepsilon \in (0,1)$, let us look for a system of inferences $\{\mathcal{T}_t : 2 \leq t \leq K\}$ such that

- the probability of false alarm is at most ϵ, and
- the probability of "ρ-miss" – the probability to detect no change when there was a change of energy $\geq \rho^2$ (i.e., when there was a change point τ, and, moreover, at this point it holds $\|x_\tau - x_1\|_2^2 \geq \rho^2$) is at most ε.

What we are interested in, is to achieve the just formulated goal with as small ρ as possible.

Construction. Let us select a large "safety parameter" R, like $R = 10^8$ or even $R = 10^{80}$, so that we can assume that for all time series we are interested in it holds $\|x_t - x_\tau\|_2^2 \leq R^2$ [28]. Let us associate with $\rho > 0$ "signal hypotheses" H_t^ρ, $t = 2, 3, ..., K$, on the distribution of observation ω^K given by (2.8.63), with H_t^ρ stating that in the time series $\{x_t\}_{t=1}^K$ underlying observation ω^K there is a change, of energy at least ρ^2, at time t:

$$x_1 = x_2 = ... = x_{t-1} \ \& \ \|x_t - x_{t-1}\|_2^2 = \|x_t - x_1\|_2^2 \geq \rho^2$$

(and on the top of it, $\|x_t - x_\tau\|_2^2 \leq R^2$ for all t, τ). Let us augment these hypotheses by the null hypothesis H_0 stating that there is no change at all – the observation ω^K stems from a stationary time series $x_1 = x_2 = ... = x_K$. We are about to use our machinery of detector-based tests in order to build a system of tests deciding, with partial risks ϵ, ε, on the null hypothesis vs. the "signal alternative" $\bigcup_t H_t^\rho$ for as small ρ as possible.

The implementation is as follows. Given $\rho > 0$ such that $\rho^2 < R^2$, consider two hypotheses, G_1 and G_2^ρ, on the distribution of observation

$$\zeta = x + \xi \in \mathbf{R}^d. \tag{2.8.64}$$

Both hypotheses state that $\xi \sim \mathcal{N}(0, \sigma^2 I_d)$ with unknown σ known to belong to a given segment $\Delta := [\sqrt{2}\underline{\sigma}, \sqrt{2}\overline{\sigma}]$. In addition, G_1 states that $x = 0$, and G_2^ρ - that $\rho^2 \leq \|x\|_2^2 \leq R^2$. We can use

[28] R is needed by the only reason – to make the domains we are working with bounded, thus allowing to apply the theory we have developed so far. The actual value of R does *not* enter our constructions and conclusions.

the result of Proposition 2.8.7.ii to build a quadratic in ζ detector for the families of distributions $\mathcal{P}_1$, $\mathcal{P}_2^\rho$ obeying the hypotheses G_1, G_2^ρ, respectively. To this end it suffices to apply Proposition to the collections of data

$$\mathcal{V}_\chi = \{\sigma^2 I_d : \sigma \in \Delta\}, \Theta_*^{(\chi)} = 2\overline{\sigma}^2 I_d, \delta_\chi = 1 - \underline{\sigma}/\overline{\sigma}, \gamma_\chi = 0.999, A_\chi = I_d, \mathcal{Z}_\chi, \qquad [\chi = 1, 2]$$

where

$$\begin{array}{rcl}\mathcal{Z}_1 & = & \{[0; ...; 0; 1][0; ...; 0; 1]^T\} \subset \mathbf{S}_+^{d+1},\\ \mathcal{Z}_2 & = & \mathcal{Z}_2^\rho = \{Z \in \mathbf{S}_+^{d+1} : Z_{d+1,d+1} = 1, 1 + R^2 \geq \mathrm{Tr}(Z) \geq 1 + \rho^2\}.\end{array}$$

The (upper bound on the) risk of the quadratic in ζ detector yielded by a saddle point of function (2.8.42), as given by Proposition 2.8.7.ii, is immediate: by the same argument as used when justifying Remark 2.8.1, in the situation in question one can look for saddle point with $h = 0$, $H = \eta I_d$, and identifying the required η reduces to solving univariate convex problem

$$\begin{array}{c}\mathrm{Opt}(\rho) = \min\limits_\eta \frac{1}{2}\left\{-\frac{d}{2}\ln(1 - \widehat{\sigma}^4\eta^2) - \frac{d}{2}\widehat{\sigma}^2(1 - \underline{\sigma}^2/\overline{\sigma}^2)\eta + \frac{d\delta(2+\delta)\widehat{\sigma}^4\eta^2}{1+\widehat{\sigma}^2\eta} + \frac{\rho^2\eta}{2(1-\widehat{\sigma}^2\eta)} : -\gamma \leq \widehat{\sigma}^2\eta \leq 0\right\}\\ \left[\widehat{\sigma} = \sqrt{2}\overline{\sigma},\ \delta = 1 - \underline{\sigma}/\overline{\sigma}\right]\end{array}$$

which can be done in no time by Bisection. The resulting detector and the upper bound on its risk are given by optimal solution $\eta(\rho)$ to the latter problem according to

$$\begin{array}{rcl}\phi_\rho^*(\zeta) & = & \frac{1}{2}\eta(\rho)\zeta^T\zeta + \underbrace{\frac{d}{4}\left[\ln\left(\frac{1 - \widehat{\sigma}^2\eta(\rho)}{1 + \widehat{\sigma}^2\eta(\rho)}\right) - \widehat{\sigma}^2(1 - \underline{\sigma}^2/\overline{\sigma}^2)\eta(\rho) - \frac{\rho^2\eta(\rho)}{1 - \widehat{\sigma}^2\eta(\rho)}\right]}_{a(\rho)},\\ \mathrm{Risk}[\phi_\rho^*|\mathcal{P}_1, \mathcal{P}_2] & \leq & \mathrm{Risk}(\rho) := \mathrm{e}^{\mathrm{Opt}(\rho)}.\end{array} \tag{2.8.65}$$

Observe that R does not appear in (2.8.65) at all. Now, it is immediately seen that $\mathrm{Opt}(\rho) \to 0$ as $\rho \to +0$ and $\mathrm{Opt}(\rho) \to -\infty$ as $\rho \to +\infty$, implying that given $\kappa \in (0, 1)$, we can easily find by bisection $\rho = \rho(\kappa)$ such that $\mathrm{Risk}(\rho) = \kappa$; in what follows, we assume w.l.o.g. that $R > \rho(\kappa)$ for the value of κ we end with, see below. Next, let us pass from the detector $\phi_{\rho(\kappa)}^*(\cdot)$ to its shift

$$\phi^{*,\kappa}(\zeta) = \phi_{\rho(\kappa)}^*(\zeta) + \ln(\varepsilon/\kappa),$$

so that for the simple test $\mathcal{T}^\kappa$ which, given observation ζ, accepts G_1 and rejects $G_2^{\rho(\kappa)}$ whenever $\phi^{*,\kappa}(\zeta) \geq 0$, and accepts $G_2^{\rho(\kappa)}$ and rejects G_1 otherwise, it holds

$$\mathrm{Risk}_1(\mathcal{T}^\kappa|G_1, G_2^{\rho(\kappa)}) \leq \frac{\kappa^2}{\varepsilon}, \mathrm{Risk}_2(\mathcal{T}^\kappa|G_1, G_2^{\rho(\kappa)}) \leq \varepsilon, \tag{2.8.66}$$

see Proposition 2.3.1 and (2.3.5).

We are nearly done. Given $\kappa \in (0, 1)$, consider the system of tests $\mathcal{T}_t^\kappa$, $t = 2, 3, ..., K$, as follows. At time $t \in \{2, 3, ..., K\}$, given observations $\omega_1, ..., \omega_t$ stemming from (2.8.63), let us form the vector

$$\zeta_t = \omega_t - \omega_1$$

and compute the quantity $\phi^{*,\kappa}(\zeta_t)$. If this quantity is negative, we claim that the change has already taken place and terminate, otherwise we claim that so far, there was no change, and proceed to time $t + 1$ (if $t < K$) or terminate (if $t = K$).

The risk analysis for the resulting system of inferences is immediate. Observe that

(!) *For every $t = 2, 3, ..., K$:*

A. *if there is no change on time horizon $1, ..., t$: $x_1 = x_2 = ... = x_t$, then the probability for $\mathcal{T}_t^\kappa$ to conclude that there was a change is at most κ^2/ε;*

B. if, on the other hand, $\|x_t - x_1\|_2^2 \geq \rho^2$, then the probability for $\mathcal{T}_t^\kappa$ to conclude that so far there was no change is at most ε.

Indeed, we clearly have

$$\zeta_t = [x_t - x_1] + \xi^t,$$

where $\xi^t = \xi_t - \xi_1 \sim \mathcal{N}(0, \sigma^2 I_d)$ with $\sigma \in [\sqrt{2}\underline{\sigma}, \sqrt{2}\overline{\sigma}]$. Our actions at time t are nothing but application of the test $\mathcal{T}^\kappa$ to the observation ζ_t. In the case of A the distribution of this observation obeys the hypothesis G_1, and the probability for $\mathcal{T}_t^\kappa$ to claim that there was a change is at most κ^2/ε by the first inequality in (2.8.66). In the case of B, the distribution of ζ_t obeys the hypothesis $G_2^{\rho(\kappa)}$, and thus the probability for $\mathcal{T}_t^\kappa$ to claim that there was no change on time horizon $1, ..., t$ is $\leq \varepsilon$ by the second inequality in (2.8.66).

In view of (!), the probability of false alarm for the system of inferences $\{\mathcal{T}_t^\kappa\}_{t=2}^K$ is at most $(K-1)\kappa^2/\varepsilon$, and specifying κ as

$$\kappa = \sqrt{\epsilon\varepsilon/(K-1)},$$

we make this probability $\leq \epsilon$. The resulting procedure, by the same (!), detects a change at time $t \in \{2, 3, ..., K\}$ with probability at least $1 - \varepsilon$, provided that the energy of this change is at least ρ_*^2, with

$$\rho_* = \rho\left(\sqrt{\epsilon\varepsilon/(K-1)}\right), \tag{2.8.67}$$

In fact we can say a bit more:

Proposition 2.8.10 *Let the deterministic sequence $x_1, ..., x_K$ underlying observations (2.8.63) be such that for some t it holds $\|x_t - x_1\|_2^2 \geq \rho_*^2$, with ρ_* given by (2.8.67). Then the probability for the system of inferences we have built to detect a change at time t or earlier is at least $1 - \varepsilon$.*

Indeed, under the premise of Proposition, the probability for $\mathcal{T}_t^\kappa$ to claim that a change already took place is at least $1 - \varepsilon$, and this probability can be only smaller than the probability to detect change on time horizon $2, 3, ..., t$.

How it works. As applied to the "movie" story we started with, the outlined procedure works as follows. The images in question are of the size 256×256, so that we are in the case of $d = 256^2 = 65536$. The images are represented by 2D arrays in gray scale, that is, as 256×256 matrices with entries in the range $[0, 255]$. In the experiment to be reported (same as in the movie) we assumed the maximal noise intensity $\overline{\sigma}$ to be 10, and used $\underline{\sigma} = \overline{\sigma}/\sqrt{2}$. The reliability tolerances ϵ, ε were set to 0.01, and K was set to 9, resulting in

$$\rho_*^2 = 7.38 \cdot 10^6,$$

which corresponds to the per pixel energy $\rho_*^2/65536 = 112.68$ – just by 12% above the allowed expected per pixel energy of noise (the latter is $\overline{\sigma}^2 = 100$). The resulting detector is

$$\phi_*(\zeta) = -2.7138\frac{\zeta^T\zeta}{10^5} + 366.9548;$$

in other words, test $\mathcal{T}_t^\kappa$ claims that the change took place when the average, over pixels, per pixel energy in the difference $\omega_t - \omega_1$ is at least 206.33, which is pretty close to the expected per pixel energy (200.0) in the noise $\xi_t - \xi_1$ affecting the difference $\omega_t - \omega_1$.

Finally, this is how the just described system of inferences worked in simulations. The underlying sequence of images was obtained from the "basic sequence"

$$\bar{x}_t = G + 0.0357(t-1)(L - G), t = 1, 2, ...^{29} \tag{2.8.68}$$

[29] The coefficient 0.0357 corresponds to 28-frame transition from G to L.

where G is the image of the gentlemen and L is the image of the lady (up to noise, these are the first and the last frames on Figure 2.5). To get the observations in a particular simulation, we augmented this sequence from the left by a random number of images G, took the first 9 images in the resulting sequence, and added to them independent across the images observation noises drawn at random from $\mathcal{N}(0, 100I_{65536})$. Augmentation was carried out in such a way that with probability 1/2, there was no change on the time horizon 1,2,...,9, and with probability 0.5 there was a change at time instant τ chosen at random according to uniform distribution on $\{2, 3, ..., 9\}$. In 3000 simulations of this type, not a *single* false alarm was observed, while the empirical probability of a miss was 0.0553. It should be added that the actual energy of a change, if any, that is, $0.0357^2\|L - G\|_F^2$, was "just" $3.37 \cdot 10^5$, that is, it was by factor ≈ 21 *less* than the energy of change ρ_*^2 which our inferences are bound to detect with probability at least 0.99. And in the series of 3000 experiments we have reported, there was no "no detection" simulations where $\max_{t\leq K} \|x_t - x_1\|_2^2$ was above ρ_*^2 (that is, no simulations where Proposition 2.8.10 ensures detectability with probability at least 0.99, and in fact the change is not detected). Thus, all misses came from simulations which are *not* covered by our risk guarantees[30]. Moreover, the change at time t, *if detected*, *never* was detected with a delay more than 1.

Finally, in the particular "movie" we started with, the change takes place at time $t = 3$, and the system of inferences we have just developed discovered the change at time 4. How this compares to the time at which you managed to detect the change?

"Numerical near-optimality." Beyond the realm of simple o.s.'s we have no theoretical guarantees of near-optimality for the inferences we are developing; this does not mean, however, that we cannot quantify conservatism of our techniques numerically. To give an example, let us forget, for the sake of simplicity, about change detection *per se* and focus on the auxiliary problem we have introduced above, the one on deciding upon hypotheses G_1 and G_2^ρ via observation (2.8.64), and let our goal be to decide on these two hypotheses from a single observation with risk $\leq \epsilon$, for a given $\epsilon \in (0, 1)$. Whether this is possible or not, it depends on ρ; let us denote by ρ^+ the smallest ρ for which we can meet the risk specification with our detector-based approach (ρ^+ is nothing but what was above called $\rho(\epsilon)$), and by $\underline{\rho}$ – the smallest ρ for which "in the nature" there exists a simple test deciding on G_1 vs. G_2^ρ with risk $\leq \epsilon$. We can look at the ratio $\rho^+/\underline{\rho}$ as at the "index of conservatism" of our approach. Now, ρ^+ is given by an efficient computation; what about $\underline{\rho}$? Well, there is a simple way to get a *lower bound* on $\underline{\rho}$, namely, as follows. Observe that if the two composite hypotheses G_1, G_2^ρ can be decided upon with risk $\leq \epsilon$, the same holds true for two simple hypotheses stating that the distribution of observation (2.8.64) is P_1, respectively, P_2, where P_1, P_2 correspond to the cases when

- (P_1): ζ is drawn from $\mathcal{N}(0, 2\overline{\sigma}^2 I_d)$
- (P_2): ζ is obtained by adding $\mathcal{N}(0, 2\underline{\sigma}^2 I_d)$-noise to a random, independent of the noise, signal u uniformly distributed on the sphere $\{\|u\|_2 = \rho\}$.

Indeed, P_1 obeys hypothesis G_1, and P_2 is a mixture of distributions obeying G_2^ρ; as a result, a simple test $\mathcal{T}$ deciding $(1-\epsilon)$-reliably on G_1 vs. G_2^ρ would induce a test deciding equally reliably on

[30] A reader can be surprised – how happens that with actual energy of change 20 times less than the "theoretical threshold" ρ_*^2, in our experiments, the empirical probability of a miss was as low as 5%, instead of being 50% or 100%. A plausible explanation is as follows: our performance guarantees are given by worst-case oriented theoretical analysis , and in random simulations we usually do not generate the "worst case" situations. For example, with model (2.8.68), the change, when happens, is of energy 20 times below the threshold; however, 3 time units after the change, the quantity $\|x_t - x_1\|_2^2$ becomes 16 times larger the energy of change, so that by Proposition 2.8.10, already worst-case analysis shows that there are good chances to detect the change when it happens "deeply inside" the observation horizon.

P_1 vs. P_2, specifically, the test which, given observation ζ, accepts P_1 if $\mathcal{T}$ on the same observation accepts G_1, and accepts P_2 otherwise.

Now, we can use two-point lower bound (Proposition 2.1.1) to lower-bound the risk of deciding on P_1 vs. P_2; since both distributions are spherically symmetric, computing this bound reduces to computing similar bound for the univariate distributions of $\zeta^T\zeta$ induced by P_1 and P_2, and these univariate distributions are easy to compute. The resulting lower risk bound depends on ρ, and we can find the smallest ρ for which the bound is ≥ 0.01, and use this ρ in the role of $\underline{\rho}$; the associated indexes of conservatism can be only larger than the true ones. Let us look what are these indexes for the data used in our change detection experiment, that is, $\epsilon = 0.01$, $d = 256^2 = 65536$, $\overline{\sigma} = 10$, $\underline{\sigma} = \overline{\sigma}/\sqrt{2}$. Computation shows that in this case we have

$$\rho^+ = 2702.4,\ \rho^+/\underline{\rho} \leq 1.04$$

– nearly no conservatism at all! When eliminating the uncertainty in the intensity of noise by increasing $\underline{\sigma}$ from $\overline{\sigma}/\sqrt{2}$ to $\overline{\sigma}$, we get

$$\rho^+ = 668.46,\ \rho^+/\underline{\rho} \leq 1.15$$

– still not that much of conservatism!

2.8.4 Proofs for Section 2.8

2.8.4.1 Proof of Proposition 2.8.1

All we need is to verify (2.8.4) and to check that the right hand side function in this relation is convex. The latter is evident, since $\phi_X(h) + \phi_X(-h) \geq 2\phi_X(0) = 0$ and $\phi_X(h) + \phi_X(-h)$ is convex. To verify (2.8.4), let us fix $P \in \mathcal{P}[X]$ and $h \in \mathbf{R}^d$ and set

$$\nu = h^T e[P],$$

so that ν is the expectation of $h^T\omega$ with $\omega \sim P$. Note that $-\phi_X(-h) \leq \nu \leq \phi_X(h)$, so that (2.8.4) definitely holds true when $\phi_X(h) + \phi_X(-h) = 0$. Now let

$$\eta := \frac{1}{2}\left[\phi_X(h) + \phi_X(-h)\right] > 0,$$

and let

$$a = \frac{1}{2}\left[\phi_X(h) - \phi_X(-h)\right],\quad \beta = (\nu - a)/\eta.$$

Denoting by P_h the distribution of $h^T\omega$ induced by the distribution P of ω and noting that this distribution is supported on $[-\phi_X(-h), \phi_X(h)] = [a - \eta, a + \eta]$ and has expectation ν, we get

$$\beta \in [-1, 1]$$

and

$$\gamma := \int \exp\{h^T\omega\}P(d\omega) = \int_{a-\eta}^{a+\eta} [\mathrm{e}^s - \lambda(s - \nu)]P_h(ds)$$

for all $\lambda \in \mathbf{R}$. Hence,

$$\begin{array}{rcl}\ln(\gamma) & \leq & \inf\limits_{\lambda} \ln\left(\max\limits_{a-\eta\leq s\leq a+\eta}[\mathrm{e}^s - \lambda(s-\nu)]\right) = a + \inf\limits_{\rho} \ln\left(\max\limits_{-\eta\leq t\leq\eta}[\mathrm{e}^t - \rho(t - [\nu - a])]\right) \\ & = & a + \inf\limits_{\rho} \ln\left(\max\limits_{-\eta\leq t\leq\eta}[\mathrm{e}^t - \rho(t - \eta\beta)]\right) \leq a + \ln\left(\max\limits_{-\eta\leq t\leq\eta}[\mathrm{e}^t - \bar{\rho}(t - \eta\beta)\right)\end{array}$$

with $\bar{\rho} = (2\eta)^{-1}(\mathrm{e}^\eta - \mathrm{e}^{-\eta})$. The function $g(t) = \mathrm{e}^t - \bar{\rho}(t - \eta\beta)$ is convex on $[-\eta, \eta]$, and

$$g(-\eta) = g(\eta) = \cosh(\eta) + \beta\sinh(\eta),$$

which combines with the above computation to yield the relation

$$\ln(\gamma) \leq a + \ln(\cosh(\eta) + \beta \sinh(\eta)), \tag{2.8.69}$$

and all we need to verify is that

$$\forall (\eta > 0, \beta \in [-1, 1]): \ \beta\eta + \frac{1}{2}\eta^2 - \ln(\cosh(\eta) + \beta \sinh(\eta)) \geq 0. \tag{2.8.70}$$

Indeed, if (2.8.70) holds true (2.8.69) implies that

$$\ln(\gamma) \leq a + \beta\eta + \frac{1}{2}\eta^2 = \nu + \frac{1}{2}\eta^2,$$

which, recalling what γ, ν and η are, is exactly what we want to prove.

Verification of (2.8.70) is as follows. The left hand side in (2.8.70) is convex in β for $\beta > -\frac{\cosh(\eta)}{\sinh(\eta)}$ containing, due to $\eta > 0$, the range of β in (2.8.70). Furthermore, the minimum of the left hand side of (2.8.70) over $\beta > -\coth(\eta)$ is attained when $\beta = \frac{\sinh(\eta) - \eta\cosh(\eta)}{\eta \sinh(\eta)}$ and is equal to

$$r(\eta) = \frac{1}{2}\eta^2 + 1 - \eta\coth(\eta) - \ln(\sinh(\eta)/\eta).$$

All we need to prove is that the latter quantity is nonnegative whenever $\eta > 0$. We have

$$r'(\eta) = \eta - \coth(\eta) - \eta(1 - \coth^2(\eta)) - \coth(\eta) + \eta^{-1} = (\eta\coth(\eta) - 1)^2\eta^{-1} \geq 0,$$

and since $r(+0) = 0$, we get $r(\eta) \geq 0$ when $\eta > 0$. □

2.8.4.2 Proof of Proposition 2.8.7

Proof of Proposition 2.8.7.i

1^0. Let $b = [0; ...; 0; 1] \in \mathbf{R}^{n+1}$, so that $B = \left[\begin{array}{c} A \\ b^T \end{array}\right]$, and let $\mathcal{A}(u) = A[u; 1]$. For any $u \in \mathbf{R}^n$, $h \in \mathbf{R}^d$, $\Theta \in \mathbf{S}^d_+$ and $H \in \mathbf{S}^d$ such that $-I \prec \Theta^{1/2}H\Theta^{1/2} \prec I$ we have

$$\begin{array}{l}
\Psi(h, H; u, \Theta) := \ln\left(\mathbf{E}_{\zeta\sim\mathcal{N}(\mathcal{A}(u),\Theta)}\left\{\exp\{h^T\zeta + \frac{1}{2}\zeta^T H\zeta\}\right\}\right) \\
\quad = \ln\left(\mathbf{E}_{\xi\sim\mathcal{N}(0,I)}\left\{\exp\{h^T[\mathcal{A}(u) + \Theta^{1/2}\xi] + \frac{1}{2}[\mathcal{A}(u) + \Theta^{1/2}\xi]^T H[\mathcal{A}(u) + \Theta^{1/2}\xi]\}\right\}\right) \\
\quad = -\frac{1}{2}\ln\mathrm{Det}(I - \Theta^{1/2}H\Theta^{1/2}) \\
+h^T\mathcal{A}(u) + \frac{1}{2}\mathcal{A}(u)^T H\mathcal{A}(u) + \frac{1}{2}[H\mathcal{A}(u) + h]^T\Theta^{1/2}[I - \Theta^{1/2}H\Theta^{1/2}]^{-1}\Theta^{1/2}[H\mathcal{A}(u) + h] \\
\quad = -\frac{1}{2}\ln\mathrm{Det}(I - \Theta^{1/2}H\Theta^{1/2}) + \frac{1}{2}[u; 1]^T\left[bh^TA + A^Thb^T + A^THA\right][u; 1] \\
\qquad\qquad\qquad + \frac{1}{2}[u; 1]^T\left[B^T[H, h]^T\Theta^{1/2}[I - \Theta^{1/2}H\Theta^{1/2}]^{-1}\Theta^{1/2}[H, h]B\right][u; 1]
\end{array} \tag{2.8.71}$$

(because $h^T\mathcal{A}(u) = [u; 1]^T bh^T A[u; 1] = [u; 1]^T A^T hb^T[u; 1]$ and $H\mathcal{A}(u) + h = [H, h]B[u; 1]$).

Observe that when $(h, H) \in \mathcal{H}^\gamma$, we have

$$\Theta^{1/2}[I - \Theta^{1/2}H\Theta^{1/2}]^{-1}\Theta^{1/2} = [\Theta^{-1} - H]^{-1} \preceq [\Theta_*^{-1} - H]^{-1},$$

so that (2.8.71) implies that for all $u \in \mathbf{R}^n$, $\Theta \in \mathcal{V}$, and $(h, H) \in \mathcal{H}^\gamma$,

$$\begin{array}{rcl}
\Psi(h, H; u, \Theta) & \leq & -\frac{1}{2}\ln\mathrm{Det}(I - \Theta^{1/2}H\Theta^{1/2}) \\
& & + \frac{1}{2}[u; 1]^T \underbrace{\left[bh^TA + A^Thb^T + A^THA + B^T[H, h]^T[\Theta_*^{-1} - H]^{-1}[H, h]B\right]}_{Q[H,h]}[u; 1] \\
& = & -\frac{1}{2}\ln\mathrm{Det}(I - \Theta^{1/2}H\Theta^{1/2}) + \frac{1}{2}\mathrm{Tr}(Q[H, h]Z(u)) \\
& \leq & -\frac{1}{2}\ln\mathrm{Det}(I - \Theta^{1/2}H\Theta^{1/2}) + \Gamma_{\mathcal{Z}}(h, H), \\
\Gamma_{\mathcal{Z}}(h, H) & = & \frac{1}{2}\phi_{\mathcal{Z}}(Q[H, h])
\end{array} \tag{2.8.72}$$

(we have taken into account that $Z(u) \in \mathcal{Z}$ when $u \in U$ (premise of the proposition) and therefore $\mathrm{Tr}(Q[H, h]Z(u)) \leq \phi_{\mathcal{Z}}(Q[H, h])$).

2^0. We need the following

Lemma 2.8.1 *Let Θ_* be a $d\times d$ symmetric positive definite matrix, let $\delta\in[0,2]$, and let $\mathcal{V}$ be a closed convex subset of $\mathbf{S}^d_+$ such that*

$$\Theta\in\mathcal{V}\Rightarrow\{\Theta\preceq\Theta_*\}\ \&\ \{\|\Theta^{1/2}\Theta_*^{-1/2}-I\|\leq\delta\} \tag{2.8.73}$$

(cf. (2.8.35)). Let also $\mathcal{H}^o:=\{H\in\mathbf{S}^d:-\Theta_*^{-1}\prec H\prec\Theta_*^{-1}\}$. Then

$$\begin{array}{l}\forall(H,\Theta)\in\mathcal{H}^o\times\mathcal{V}:\\ G(H;\Theta):=-\frac{1}{2}\ln\mathrm{Det}(I-\Theta^{1/2}H\Theta^{1/2})\\ \quad\leq G^+(H;\Theta):=-\frac{1}{2}\ln\mathrm{Det}(I-\Theta_*^{1/2}H\Theta_*^{1/2})+\frac{1}{2}\mathrm{Tr}([\Theta-\Theta_*]H)+\frac{\delta(2+\delta)}{2(1-\|\Theta_*^{1/2}H\Theta_*^{1/2}\|)}\|\Theta_*^{1/2}H\Theta_*^{1/2}\|_F^2,\end{array} \tag{2.8.74}$$

where $\|\cdot\|$ is the spectral, and $\|\cdot\|_F$ - the Frobenius norm of a matrix. In addition, $G^+(H,\Theta)$ is continuous function on $\mathcal{H}^o\times\mathcal{V}$ which is convex in $H\in H^o$ and concave (in fact, affine) in $\Theta\in\mathcal{V}$

Proof. Let us set

$$d(H)=\|\Theta_*^{1/2}H\Theta_*^{1/2}\|,$$

so that $d(H)<1$ for $H\in\mathcal{H}^o$. For $H\in\mathcal{H}^o$ and $\Theta\in\mathcal{V}$ fixed we have

$$\begin{array}{rcl}\|\Theta^{1/2}H\Theta^{1/2}\| & = & \|[\Theta^{1/2}\Theta_*^{-1/2}][\Theta_*^{1/2}H\Theta_*^{1/2}][\Theta^{1/2}\Theta_*^{-1/2}]^T\|\\ & \leq & \|\Theta^{1/2}\Theta_*^{-1/2}\|^2\|\Theta_*^{1/2}H\Theta_*^{1/2}\|\leq\|\Theta_*^{1/2}H\Theta_*^{1/2}\|=d(H)\end{array} \tag{2.8.75}$$

(we have used the fact that $0\preceq\Theta\preceq\Theta_*$ implies $\|\Theta^{1/2}\Theta_*^{-1/2}\|\leq1$). Noting that $\|AB\|_F\leq\|A\|\|B\|_F$, computation completely similar to the one in (2.8.75) yields

$$\|\Theta^{1/2}H\Theta^{1/2}\|_F\leq\|\Theta_*^{1/2}H\Theta_*^{1/2}\|_F=:D(H) \tag{2.8.76}$$

Besides this, setting $F(X)=-\ln\mathrm{Det}(X):\mathrm{int}\,\mathbf{S}^d_+\to\mathbf{R}$ and equipping $\mathbf{S}^d$ with the Frobenius inner product, we have $\nabla F(X)=-X^{-1}$, so that with $R_0=\Theta_*^{1/2}H\Theta_*^{1/2}$, $R_1=\Theta^{1/2}H\Theta^{1/2}$, and $\Delta=R_1-R_0$, we have for properly selected $\lambda\in(0,1)$ and $R_\lambda=\lambda R_0+(1-\lambda)R_1$:

$$\begin{array}{rcl}F(I-R_1) & = & F(I-R_0-\Delta)=F(I-R_0)+\langle\nabla F(I-R_\lambda),-\Delta\rangle=F(I-R_0)+\langle(I-R_\lambda)^{-1},\Delta\rangle\\ & = & F(I-R_0)+\langle I,\Delta\rangle+\langle(I-R_\lambda)^{-1}-I,\Delta\rangle.\end{array}$$

We conclude that

$$F(I-R_1)\leq F(I-R_0)+\mathrm{Tr}(\Delta)+\|I-(I-R_\lambda)^{-1}\|_F\|\Delta\|_F. \tag{2.8.77}$$

Denoting by μ_i the eigenvalues of R_λ and noting that $\|R_\lambda\|\leq\max[\|R_0\|,\|R_1\|]=d(H)$ (see (2.8.75)), we have $|\mu_i|\leq d(H)$, and therefore eigenvalues $\nu_i=1-\frac{1}{1-\mu_i}=-\frac{\mu_i}{1-\mu_i}$ of $I-(I-R_\lambda)^{-1}$ satisfy $|\nu_i|\leq|\mu_i|/(1-\mu_i)\leq|\mu_i|/(1-d(H))$, whence

$$\|I-(I-R_\lambda)^{-1}\|_F\leq\|R_\lambda\|_F/(1-d(H)).$$

Noting that $\|R_\lambda\|_F\leq\max[\|R_0\|_F,\|R_1\|_F]\leq D(H)$, see (2.8.76), we conclude that $\|I-(I-R_\lambda)^{-1}\|_F\leq D(H)/(1-d(H))$, so that (2.8.77) yields

$$F(I-R_1)\leq F(I-R_0)+\mathrm{Tr}(\Delta)+D(H)\|\Delta\|_F/(1-d(H)). \tag{2.8.78}$$

Further, by (2.8.35) the matrix $D=\Theta^{1/2}\Theta_*^{-1/2}-I$ satisfies $\|D\|\leq\delta$, whence

$$\Delta=\underbrace{\Theta^{1/2}H\Theta^{1/2}}_{R_1}-\underbrace{\Theta_*^{1/2}H\Theta_*^{1/2}}_{R_0}=(I+D)R_0(I+D^T)-R_0=DR_0+R_0D^T+DR_0D^T.$$

Consequently,

$$\|\Delta\|_F\leq\|DR_0\|_F+\|R_0D^T\|_F+\|DR_0D^T\|_F\leq[2\|D\|+\|D\|^2]\|R_0\|_F\leq\delta(2+\delta)\|R_0\|_F=\delta(2+\delta)D(H).$$

This combines with (2.8.78) and the relation

$$\mathrm{Tr}(\Delta)=\mathrm{Tr}(\Theta^{1/2}H\Theta^{1/2}-\Theta_*^{1/2}H\Theta_*^{1/2})=\mathrm{Tr}([\Theta-\Theta_*]H)$$

to yield

$$F(I-R_1)\leq F(I-R_0)+\mathrm{Tr}([\Theta-\Theta_*]H)+\frac{\delta(2+\delta)}{1-d(H)}\|\Theta_*^{1/2}H\Theta_*^{1/2}\|_F^2,$$

and we arrive at (2.8.74). It remains to prove that $G^+(H;\Theta)$ is convex-concave and continuous on $\mathcal{H}^o\times\mathcal{V}$. The only component of this claim which is not completely evident is convexity of the function in $H\in\mathcal{H}^o$; to see that it is the case, note that $\ln\mathrm{Det}(S)$ is concave on the interior of the semidefinite cone, the function $f(u,v)=\frac{u^2}{1-v}$ is convex and nondecreasing in u,v in the convex domain $\Pi=\{(u,v):u\geq 0,v<1\}$, and the function $\frac{\|\Theta_*^{1/2}H\Theta_*^{1/2}\|_F^2}{1-\|\Theta_*^{1/2}H\Theta_*^{1/2}\|}$ is obtained from f by convex substitution of variables $H\mapsto(\|\Theta_*^{1/2}H\Theta_*^{1/2}\|_F,\|\Theta_*^{1/2}H\Theta_*^{1/2}\|)$ mapping $\mathcal{H}^o$ into Π. □

.

3^0. Combining (2.8.74), (2.8.72), (2.8.37) and the origin of Ψ, see (2.8.71), we arrive at

$$\forall((u,\Theta)\in U\times\mathcal{V},(h,H)\in\mathcal{H}^\gamma=\mathcal{H}):\ \ln\left(\mathbf{E}_{\zeta\sim\mathcal{N}(A[u;1],\Theta)}\left\{\exp\{h^T\zeta+\frac{1}{2}\zeta^TH\zeta\}\right\}\right)\leq\Phi_{A,\mathcal{Z}}(h,H;\Theta),$$

as claimed in (2.8.40).

4^0. Now let us check that $\Phi_{A,\mathcal{Z}}(h,H;\Theta):\mathcal{H}\times\mathcal{V}\to\mathbf{R}$ is continuous and convex-concave. Recalling that the function $G^+(H;\Theta)$ from (2.8.74) is convex-concave and continuous on $\mathcal{H}^o\times\mathcal{V}$, all we need to verify is that $\Gamma_{\mathcal{Z}}(h,H)$ is convex and continuous on $\mathcal{H}$. Recalling that $\mathcal{Z}$ is nonempty compact set, the function $\phi_{\mathcal{Z}}(\cdot):\mathbf{S}^{d+1}\to\mathbf{R}$ is continuous, implying the continuity of $\Gamma_{\mathcal{Z}}(h,H)=\frac{1}{2}\phi_{\mathcal{Z}}(Q[H,h])$ on $\mathcal{H}=\mathcal{H}^\gamma$ ($Q[H,h]$ is defined in (2.8.72)). To prove convexity of $\Gamma_{\mathcal{Z}}$, note that $\mathcal{Z}$ is contained in $\mathbf{S}^{n+1}_+$, implying that $\phi_{\mathcal{Z}}(\cdot)$ is convex and $\succeq$-monotone. On the other hand, by Schur Complement Lemma, we have

$$\begin{array}{rcl}S&:=&\{(h,H,G):G\succeq Q[H,h],(h,H)\in\mathcal{H}^\gamma\}\\&=&\left\{(h,H,G):\left[\begin{array}{c|c}G-[bh^TA+A^Thb^T+A^THA]&B^T[H,h]^T\\\hline [H,h]B&\Theta_*^{-1}-H\end{array}\right]\succeq 0,(h,H)\in\mathcal{H}^\gamma\right\},\end{array}$$

implying that S is convex. Since $\phi_{\mathcal{Z}}(\cdot)$ is $\succeq$-monotone, we have

$$\{(h,H,\tau):(h,H)\in\mathcal{H}^\gamma,\ \tau\geq\Gamma)_{\mathcal{Z}}(h,H)\}=\{(h,H,\tau):\ \exists G:G\succeq Q[H,h],\ 2\tau\geq\phi_{\mathcal{Z}}(G),\ (h,H)\in\mathcal{H}^\gamma\},$$

and we see that the epigraph of $\Gamma_{\mathcal{Z}}$ is convex (since the set S and the epigraph of $\phi_{\mathcal{Z}}$ are so), as claimed.

5^0. It remains to prove that $\Phi_{\mathcal{A},\mathcal{Z}}$ is coercive in H,h. Let $\Theta\in\mathcal{V}$ and $(h_i,H_i)\in\mathcal{H}^\gamma$ with $\|(h_i,H_i)\|\to\infty$ as $i\to\infty$, and let us prove that $\Phi_{\mathcal{A},\mathcal{Z}}(h_i,H_i;\Theta)\to\infty$. Looking at the expression for $\Phi_{\mathcal{A},\mathcal{Z}}(h_i,H_i;\Theta)$, it is immediately seen that all terms in this expression, except for the terms coming from $\phi_{\mathcal{Z}}(\cdot)$, remain bounded as i grows, so that all we need to verify is that the $\phi_{\mathcal{Z}}(\cdot)$-term goes to ∞ as $i\to\infty$. Observe that H_i are uniformly bounded due to $(h_i,H_i)\in\mathcal{H}^\gamma$, implying that $\|h_i\|_2\to\infty$ as $i\to\infty$. Denoting $e=[0;...;0;1]\in\mathbf{R}^{d+1}$ and, as before, $b=[0;...;0;1]\in\mathbf{R}^{n+1}$, note that, by construction, $B^Te=b$. Now let $W\in\mathcal{Z}$, so that $W_{n+1,n+1}=1$. Taking into account that the matrices $[\Theta_*^{-1}-H_i]^{-1}$ satisfy $\alpha I_d\preceq[\Theta_*^{-1}-H_i]^{-1}\preceq\beta I_d$ for some positive α,β due to $H_i\in\mathcal{H}^\gamma$, observe that

$$\underbrace{\left[\left[\begin{array}{c|c}H_i&h_i\\\hline h_i^T&\end{array}\right]+[H_i,h_i]^T\left[\Theta_*^{-1}-H_i\right]^{-1}[H_i,h_i]\right]}_{Q_i}=\underbrace{\left[h_i^T[\Theta_*^{-1}-H_i]^{-1}h_i\right]}_{\alpha_i\|h_i\|_2^2}ee^T+R_i,$$

where $\alpha_i\geq\alpha>0$ and $\|R_i\|_F\leq C(1+\|h_i\|_2)$. As a result,

$$\begin{array}{rcl}\phi_{\mathcal{Z}}(B^TQ_iB)&\geq&\mathrm{Tr}(WB^TQ_iB)=\mathrm{Tr}(WB^T[\alpha_i\|h_i\|_2^2ee^T+R_i]B)\\&=&\alpha_i\|h_i\|_2^2\underbrace{\mathrm{Tr}(Wbb^T)}_{=W_{n+1,n+1}=1}-\|BWB^T\|_F\|R_i\|_F\geq\alpha\|h_i\|_2^2-C(1+\|h_i\|_2)\|BWB^T\|_F,\end{array}$$

and the concluding quantity tends to ∞ as $i \to \infty$ due to $\|h_i\|_2 \to \infty$, $i \to \infty$. Part (i) is proved.

Proof of Proposition 2.8.7.ii

By (i) the function $\Phi(h, H; \Theta_1, \Theta_2)$ is continuous and convex-concave on the domain $\underbrace{(\mathcal{H}_1 \cap \mathcal{H}_2)}_{\mathcal{H}} \times \underbrace{(\mathcal{V}_1 \times \mathcal{V}_2)}_{\mathcal{V}}$ and are coercive in (h, H), while $\mathcal{H}$ and $\mathcal{V}$ are closed and convex, and $\mathcal{V}$ in addition is compact, saddle point problem (2.8.42) is solvable (Sion-Kakutani Theorem, a.k.a. Theorem 2.4.1). Now let $(h_*, H_*; \Theta_1^*, \Theta_2^*)$ be a saddle point. To prove (2.8.44), let $P \in \mathcal{G}_1$, that is, $P = \mathcal{N}(A_1[u; 1], \Theta_1)$ for some $\Theta_1 \in \mathcal{V}_1$ and some u with $[u; 1][u; 1]^T \in \mathcal{Z}_1$. Applying (2.8.40) to the first collection of data, with a given by (2.8.43), we get the first $\leq$ in the following chain:

$$\ln\left(\int \mathrm{e}^{-\frac{1}{2}\omega^T H_* \omega - \omega^T h_* - a} P(d\omega)\right) \leq \Phi_{A_1, \mathcal{Z}_1}(-h_*, -H_*; \Theta_1) - a \underbrace{\leq}_{(a)} \Phi_{A_1, \mathcal{Z}_1}(-h_*, -H_*; \Theta_1^*) - a \underbrace{=}_{(b)} \mathcal{SV},$$

where (a) is due to the fact that $\Phi_{A_1, \mathcal{Z}_1}(-h_*, -H_*; \Theta_1) + \Phi_{A_2, \mathcal{Z}_2}(h_*, H_*; \Theta_2)$ attains its maximum over $(\Theta_1, \Theta_2) \in \mathcal{V}_1 \times \mathcal{V}_2$ at the point (Θ_1^*, Θ_2^*), and (b) is due to the origin of a and the relation $\mathcal{SV} = \frac{1}{2}[\Phi_{A_1, \mathcal{Z}_1}(-h_*, -H_*; \Theta_1^*) + \Phi_{A_2, \mathcal{Z}_2}(h_*, H_*; \Theta_2^*)]$. The bound in (2.8.44.a) is proved. Similarly, let $P \in \mathcal{G}_2$, that is, $P = \mathcal{N}(A_2[u; 1], \Theta_2)$ for some $\Theta_2 \in \mathcal{V}_2$ and some u with $[u; 1][u; 1]^T \in \mathcal{Z}_2$. Applying (2.8.40) to the second collection of data, with the same a as above, we get the first $\leq$ in the following chain:

$$\ln\left(\int \mathrm{e}^{\frac{1}{2}\omega^T H_* \omega + \omega^T h_* + a} P(d\omega)\right) \leq \Phi_{A_2, \mathcal{Z}_2}(h_*, H_*; \Theta_2) + a \underbrace{\leq}_{(a)} \Phi_{A_2, \mathcal{Z}_2}(h_*, H_*; \Theta_2^*) + a \underbrace{=}_{(b)} \mathcal{SV},$$

with exactly the same as above justification of (a) and (b). The bound in (2.8.44.b) is proved. □

2.8.4.3 Proof of Proposition 2.8.9

2.8.4.4 Preliminaries

We start with the following result:

Lemma 2.8.2 *Let $\bar{\Theta}$ be a positive definite $d \times d$ matrix, and let*

$$u \mapsto \mathcal{C}(u) = A[u; 1]$$

be an affine mapping from $\mathbf{R}^n$ into $\mathbf{R}^d$. Finally, let $h \in \mathbf{R}^d$, $H \in \mathbf{S}^d$ and $P \in \mathbf{S}^d$ satisfy the relations

$$0 \preceq P \prec I_d \ \&\ P \succeq \bar{\Theta}^{1/2} H \bar{\Theta}^{1/2}. \tag{2.8.79}$$

Then, setting $B = \left[\begin{array}{c} A \\ 0, ..., 0, 1 \end{array}\right]$, for every $u \in \mathbf{R}^n$ it holds

$$\begin{array}{l} \zeta \sim \mathcal{SG}(\mathcal{C}(u), \bar{\Theta}) \Rightarrow \ln\left(\mathbf{E}_\zeta\left\{\mathrm{e}^{h^T\zeta + \frac{1}{2}\zeta^T H \zeta}\right\}\right) \leq -\frac{1}{2}\ln \mathrm{Det}(I - P) \\ \quad + \frac{1}{2}[u; 1]^T B^T \left[\left[\begin{array}{c|c} H & h \\ \hline h^T & \end{array}\right] + [H, h]^T \bar{\Theta}^{1/2}[I - P]^{-1}\bar{\Theta}^{1/2} [H, h]\right] B[u; 1] \end{array} \tag{2.8.80}$$

Equivalently (set $G = \bar{\Theta}^{-1/2} P \bar{\Theta}^{-1/2}$): Whenever $h \in \mathbf{R}^d$, $H \in \mathbf{S}^d$ and $G \in \mathbf{S}^d$ satisfy the relations

$$0 \preceq G \prec \bar{\Theta}^{-1} \ \&\ G \succeq H, \tag{2.8.81}$$

one has for every for every $u \in \mathbf{R}^n$:

$$\begin{array}{l} \zeta \sim \mathcal{SG}(\mathcal{C}(u), \bar{\Theta}) \Rightarrow \ln\left(\mathbf{E}_\zeta\left\{\mathrm{e}^{h^T\zeta + \frac{1}{2}\zeta^T H \zeta}\right\}\right) \leq -\frac{1}{2}\ln \mathrm{Det}(I - \bar{\Theta}^{1/2} G \bar{\Theta}^{1/2}) \\ \quad + \frac{1}{2}[u; 1]^T B^T \left[\left[\begin{array}{c|c} H & h \\ \hline h^T & \end{array}\right] + [H, h]^T \left[\bar{\Theta}^{-1} - G\right]^{-1} [H, h]\right] B[u; 1] \end{array} \tag{2.8.82}$$

Proof. 1^0. Let us start with the following observation:

Lemma 2.8.3 *Let $\Theta\in\mathbf{S}^d_+$ and $S\in\mathbf{R}^{d\times d}$ be such that $S\Theta S^T\prec I_d$. Then for every $\nu\in\mathbf{R}^d$ one has*

$$\begin{array}{rcl}\ln\left(\mathbf{E}_{\xi\sim\mathcal{SG}(0,\Theta)}\left\{\mathrm{e}^{\nu^TS\xi+\frac{1}{2}\xi^TS^TS\xi}\right\}\right)&\leq&\ln\left(\mathbf{E}_{x\sim\mathcal{N}(\nu,I_d)}\left\{\mathrm{e}^{\frac{1}{2}x^TS\Theta S^Tx}\right\}\right)\\ &=&-\frac{1}{2}\ln\mathrm{Det}(I_d-S\Theta S^T)+\frac{1}{2}\nu^T\left[S\Theta S^T(I_d-S\Theta S^T)^{-1}\right]\nu.\end{array}\tag{2.8.83}$$

Indeed, let $\xi\sim\mathcal{SG}(0,\Theta)$ and $x\sim\mathcal{N}(\nu,I_d)$ be independent. We have

$$\mathbf{E}_\xi\left\{\mathrm{e}^{\nu^TS\xi+\frac{1}{2}\xi^TS^TS\xi}\right\}\underbrace{=}_{a}\mathbf{E}_\xi\left\{\mathbf{E}_x\left\{\mathrm{e}^{[S\xi]^Tx}\right\}\right\}=\mathbf{E}_x\left\{\mathbf{E}_\xi\left\{\mathrm{e}^{[S^Tx]^T\xi}\right\}\right\}\underbrace{\leq}_{b}\mathbf{E}_x\left\{\mathrm{e}^{\frac{1}{2}x^TS\Theta S^Tx}\right\},$$

where a is due to $x\sim\mathcal{N}(\nu,I_d)$ and b is due to $\xi\sim\mathcal{SG}(0,\Theta)$. We have verified the inequality in (2.8.83); the equality in (2.8.83) is given by direct computation. □

2^0. Now, in the situation described in Lemma 2.8.2, by continuity it suffices to prove (2.8.80) in the case when $P\succeq0$ in (2.8.79) is replaced with $P\succ0$. Under the premise of Lemma, given $u\in\mathbf{R}^n$ and assuming $P\succ0$, let us set $\mu=\mathcal{C}(u)=A[u;1]$, $\nu=P^{-1/2}\bar\Theta^{1/2}[H\mu+h]$, $S=P^{1/2}\bar\Theta^{-1/2}$, so that $S\bar\Theta S^T=P\prec I_d$, and let $G=\bar\Theta^{-1/2}P\bar\Theta^{-1/2}$, so that $G\succeq H$. Let $\zeta\sim\mathcal{SG}(\mu,\bar\Theta)$. Representing ζ as $\zeta=\mu+\xi$ with $\xi\sim\mathcal{SG}(0,\bar\Theta)$, we have

$$\begin{array}{l}\ln\left(\mathbf{E}_\zeta\left\{\mathrm{e}^{h^T\zeta+\frac{1}{2}\zeta^TH\zeta}\right\}\right)=h^T\mu+\frac{1}{2}\mu^TH\mu+\ln\left(\mathbf{E}_\xi\left\{\mathrm{e}^{[h+H\mu]^T\xi+\frac{1}{2}\xi^TH\xi}\right\}\right)\\ \leq h^T\mu+\frac{1}{2}\mu^TH\mu+\ln\left(\mathbf{E}_\xi\left\{\mathrm{e}^{[h+H\mu]^T\xi+\frac{1}{2}\xi^TG\xi}\right\}\right)\ \text{[since } G\succeq H\text{]}\\ =h^T\mu+\frac{1}{2}\mu^TH\mu+\ln\left(\mathbf{E}_\xi\left\{\mathrm{e}^{\nu^TS\xi+\frac{1}{2}\xi^TS^TS\xi}\right\}\right)\ \text{[since } S^T\nu=h+H\mu \text{ and } G=S^TS\text{]}\\ \leq h^T\mu+\frac{1}{2}\mu^TH\mu-\frac{1}{2}\ln\mathrm{Det}(I_d-S\bar\Theta S^T)+\frac{1}{2}\nu^T\left[S\bar\Theta S^T(I_d-S\bar\Theta S^T)^{-1}\right]\nu\ \text{[by Lemma 2.8.3 with } \Theta=\bar\Theta\text{]}\\ =h^T\mu+\frac{1}{2}\mu^TH\mu-\frac{1}{2}\ln\mathrm{Det}(I_d-P)+\frac{1}{2}[H\mu+h]^T\bar\Theta^{1/2}(I_d-P)^{-1}\bar\Theta^{1/2}[H\mu+h]\ \text{[plugging in } S \text{ and } \nu\text{]}\end{array}$$

It is immediately seen that the concluding quantity in this chain is nothing but the right hand side quantity in (2.8.80). □

2.8.4.5 Proof of Proposition 2.8.9.i

1^0. Let us prove (2.8.54.a). By Lemma 2.8.2 (see (2.8.82)) applied with $\bar\Theta=\Theta_*$, setting $\mathcal{C}(u)=A[u;1]$, we have

$$\begin{array}{l}\forall\left((h,H)\in\mathcal{H},G:0\preceq G\preceq\gamma^+\Theta_*^{-1},G\succeq H,u\in\mathbf{R}^n:[u;1][u;1]^T\in\mathcal{Z}\right):\\ \ln\left(\mathbf{E}_{\zeta\sim\mathcal{SG}(\mathcal{C}(u),\Theta_*)}\left\{\mathrm{e}^{h^T\zeta+\frac{1}{2}\zeta^TH\zeta}\right\}\right)\leq-\frac{1}{2}\ln\mathrm{Det}(I-\Theta_*^{1/2}G\Theta_*^{1/2})\\ \qquad\qquad+\frac{1}{2}[u;1]^TB^T\left[\left[\begin{array}{c|c}H&h\\ \hline h^T&\end{array}\right]+[H,h]^T\left[\Theta_*^{-1}-G\right]^{-1}[H,h]\right]B[u;1]\\ \leq-\frac{1}{2}\ln\mathrm{Det}(I-\Theta_*^{1/2}G\Theta_*^{1/2})+\frac{1}{2}\phi_{\mathcal{Z}}\left(B^T\left[\left[\begin{array}{c|c}H&h\\ \hline h^T&\end{array}\right]+[H,h]^T\left[\Theta_*^{-1}-G\right]^{-1}[H,h]\right]B\right)\\ =\Psi_{A,\mathcal{Z}}(h,H,G),\end{array}\tag{2.8.84}$$

implying, due to the origin of $\Phi_{A,\mathcal{Z}}$, that under the premise of (2.8.84) we have

$$\ln\left(\mathbf{E}_{\zeta\sim\mathcal{SG}(\mathcal{C}(u),\Theta_*)}\left\{\mathrm{e}^{h^T\zeta+\frac{1}{2}\zeta^TH\zeta}\right\}\right)\leq\Phi_{A,\mathcal{Z}}(h,H),\ \forall(h,H)\in\mathcal{H}.$$

Taking into account that when $\zeta\sim\mathcal{SG}(\mathcal{C}(u),\Theta)$ with $\Theta\in\mathcal{V}$, we have also $\zeta\sim\mathcal{SG}(\mathcal{C}(u),\Theta_*)$, (2.8.54.$a$) follows.

2^0. Now let us prove (2.8.54.b). All we need is to verify the relation

$$\begin{array}{c}\forall\left((h,H)\in\mathcal{H},G:0\preceq G\preceq\gamma^+\Theta_*^{-1},G\succeq H,u\in\mathbf{R}^n:[u;1][u;1]^T\in\mathcal{Z},\Theta\in\mathcal{V}\right):\\ \ln\left(\mathbf{E}_{\zeta\sim\mathcal{SG}(\mathcal{C}(u),\Theta)}\left\{\mathrm{e}^{h^T\zeta+\frac{1}{2}\zeta^TH\zeta}\right\}\right)\leq\Psi^\delta_{A,\mathcal{Z}}(h,H,G;\Theta);\end{array}\tag{2.8.85}$$

with this relation at our disposal (2.8.54.b) can be obtained by the same argument as the one we used in item 1^0 to derive (2.8.54.a).

To establish (2.8.85), let us fix h, H, G, u, Θ satisfying the premise of (2.8.85); note that under the premise of Proposition 2.8.9.i, we have $0 \preceq \Theta \preceq \Theta_*$. Now let $\lambda \in (0,1)$, and let $\Theta_\lambda = \Theta + \lambda(\Theta_* - \Theta)$, so that $0 \prec \Theta_\lambda \preceq \Theta_*$, and let $\delta_\lambda = \|\Theta_\lambda^{1/2}\Theta_*^{-1/2} - I\|$, so that $\delta_\lambda \in [0,2]$. We have $0 \preceq G \preceq \gamma^+\Theta_*^{-1} \preceq \gamma^+\Theta_\lambda^{-1}$ that is, H, G satisfy (2.8.81) w.r.t. $\bar{\Theta} = \Theta_\lambda$. As a result, for our h, G, H, u and the just defined $\bar{\Theta}$ relation (2.8.82) holds true:

$$\begin{array}{l}
\zeta \sim \mathcal{SG}(\mathcal{C}(u), \Theta_\lambda) \Rightarrow\\
\ln\left(\mathbf{E}_\zeta\left\{\mathrm{e}^{h^T\zeta + \frac{1}{2}\zeta^T H\zeta}\right\}\right) \leq -\frac{1}{2}\ln\mathrm{Det}(I - \Theta_\lambda^{1/2}G\Theta_\lambda^{1/2})\\
\qquad\qquad + \frac{1}{2}[u;1]^T B^T\left[\left[\begin{array}{c|c} H & h\\ \hline h^T & \end{array}\right] + [H,h]^T\left[\Theta_\lambda^{-1} - G\right]^{-1}[H,h]\right]B[u;1]\\
\leq -\frac{1}{2}\ln\mathrm{Det}(I - \Theta_\lambda^{1/2}G\Theta_\lambda^{1/2}) + \frac{1}{2}\phi_{\mathcal{Z}}\left(B^T\left[\left[\begin{array}{c|c} H & h\\ \hline h^T & \end{array}\right] + [H,h]^T\left[\Theta_\lambda^{-1} - G\right]^{-1}[H,h]\right]B\right)
\end{array} \tag{2.8.86}$$

(recall that $[u;1][u;1]^T \in \mathcal{Z}$). As a result,

$$\begin{array}{l}
\zeta \sim \mathcal{SG}(\mathcal{C}(u), \Theta) \Rightarrow \ln\left(\mathbf{E}_\zeta\left\{\mathrm{e}^{h^T\zeta + \frac{1}{2}\zeta^T H\zeta}\right\}\right) \leq -\frac{1}{2}\ln\mathrm{Det}(I - \Theta_\lambda^{1/2}G\Theta_\lambda^{1/2})\\
\qquad\qquad + \frac{1}{2}\phi_{\mathcal{Z}}\left(B^T\left[\left[\begin{array}{c|c} H & h\\ \hline h^T & \end{array}\right] + [H,h]^T\left[\Theta_*^{-1} - G\right]^{-1}[H,h]\right]B\right)
\end{array} \tag{2.8.87}$$

When deriving (2.8.87) from (2.8.86), we have used that
— $\Theta \preceq \Theta_\lambda$, so that when $\zeta \sim \mathcal{SG}(\mathcal{C}(u), \Theta)$, we have also $\zeta \sim \mathcal{SG}(\mathcal{C}(u), \Theta_\lambda)$,
— $0 \preceq \Theta_\lambda \preceq \Theta_*$ and $G \prec \Theta_*^{-1}$, whence $[\Theta_\lambda^{-1} - G]^{-1} \preceq [\Theta_*^{-1} - G]^{-1}$,
— $\mathcal{Z} \subset \mathbf{S}^{n+1}_+$, whence $\phi_{\mathcal{Z}}$ is $\succeq$-monotone: $\phi_{\mathcal{Z}}(M) \leq \phi_{\mathcal{Z}}(N)$ whenever $M \preceq N$.

By Lemma 2.8.1 applied with Θ_λ in the role of Θ and δ_λ in the role of δ, we have

$$\begin{array}{l}
-\frac{1}{2}\ln\mathrm{Det}(I - \Theta_\lambda^{1/2}G\Theta_\lambda^{1/2})\\
\qquad \leq -\frac{1}{2}\ln\mathrm{Det}(I - \Theta_*^{1/2}G\Theta_*^{1/2}) + \frac{1}{2}\mathrm{Tr}([\Theta_\lambda - \Theta_*]G) + \frac{\delta_\lambda(2+\delta_\lambda)}{2(1-\|\Theta_*^{1/2}G\Theta_*^{1/2}\|)}\|\Theta_*^{1/2}G\Theta_*^{1/2}\|_F^2.
\end{array}$$

Consequently, (2.8.87) implies that

$$\begin{array}{l}
\zeta \sim \mathcal{SG}(\mathcal{C}(u), \Theta) \Rightarrow\\
\ln\left(\mathbf{E}_\zeta\left\{\mathrm{e}^{h^T\zeta + \frac{1}{2}\zeta^T H\zeta}\right\}\right)\\
\qquad \leq -\frac{1}{2}\ln\mathrm{Det}(I - \Theta_*^{1/2}G\Theta_*^{1/2}) + \frac{1}{2}\mathrm{Tr}([\Theta_\lambda - \Theta_*]G) + \frac{\delta_\lambda(2+\delta_\lambda)}{2(1-\|\Theta_*^{1/2}G\Theta_*^{1/2}\|)}\|\Theta_*^{1/2}G\Theta_*^{1/2}\|_F^2\\
\qquad\qquad + \frac{1}{2}\phi_{\mathcal{Z}}\left(B^T\left[\left[\begin{array}{c|c} H & h\\ \hline h^T & \end{array}\right] + [H,h]^T\left[\Theta_*^{-1} - G\right]^{-1}[H,h]\right]B\right).
\end{array}$$

The resulting inequality holds true for all small positive λ; taking $\liminf$ of the right hand side as $\lambda \to +0$, and recalling that $\Theta_0 = \Theta$, we get

$$\begin{array}{l}
\zeta \sim \mathcal{SG}(\mathcal{C}(u), \Theta) \Rightarrow\\
\ln\left(\mathbf{E}_\zeta\left\{\mathrm{e}^{h^T\zeta + \frac{1}{2}\zeta^T H\zeta}\right\}\right)\\
\qquad \leq -\frac{1}{2}\ln\mathrm{Det}(I - \Theta_*^{1/2}G\Theta_*^{1/2}) + \frac{1}{2}\mathrm{Tr}([\Theta - \Theta_*]G) + \frac{\delta(2+\delta)}{2(1-\|\Theta_*^{1/2}G\Theta_*^{1/2}\|)}\|\Theta_*^{1/2}G\Theta_*^{1/2}\|_F^2\\
\qquad\qquad + \frac{1}{2}\phi_{\mathcal{Z}}\left(B^T\left[\left[\begin{array}{c|c} H & h\\ \hline h^T & \end{array}\right] + [H,h]^T\left[\Theta_*^{-1} - G\right]^{-1}[H,h]\right]B\right)
\end{array}$$

(note that under the premise of Proposition 2.8.9.i we clearly have $\liminf_{\lambda\to+0}\delta_\lambda \leq \delta$). The right hand side of the resulting inequality is nothing but $\Psi^\delta_{A,\mathcal{Z}}(h, H, G; \Theta)$, see (2.8.53), and we arrive at the inequality required in the conclusion of (2.8.85).

3^0. To complete the proof of Proposition 2.8.9.i, it remains to prove that the functions $\Phi_{A,\mathcal{Z}}$, $\Phi^\delta_{A,\mathcal{Z}}$ possess the announced in Proposition continuity, convexity-concavity, and coerciveness properties. Let us verify that this indeed is so for $\Phi^\delta_{A,\mathcal{Z}}$; reasoning to follow, with evident simplifications, is applicable to $\Phi_{A,\mathcal{Z}}$ as well.

Observe, first, that by exactly the same reasons as in item **4^0** of the proof of Proposition 2.8.7, the function $\Psi^\delta_{A,\mathcal{Z}}(h, H, G; \Theta)$ is real valued, continuous and convex-concave on the domain

$$\widehat{\mathcal{H}} \times \mathcal{V} = \{(h, H, G) : -\gamma^+\Theta_*^{-1} \preceq H \preceq \gamma^+\Theta_*^{-1}, 0 \preceq G \preceq \gamma^+\Theta_*^{-1}, H \preceq G\} \times \mathcal{V}.$$

The function $\Phi^\delta_{A,\mathcal{Z}}(h,H;\Theta):\mathcal{H}\times\mathcal{V}\to\mathbf{R}$ is obtained from $\Psi^\delta(h,H,G;\Theta)$ by the following two operations: we first minimize $\Psi^\delta_{A,\mathcal{Z}}(h,H,G;\Theta)$ over G linked to (h,H) by the convex constraints $0\preceq G\preceq\gamma^+\Theta_*^{-1}$ and $G\succeq H$, thus obtaining a function

$$\bar\Phi(h,H;\Theta):\underbrace{\{(h,H):-\gamma^+\Theta_*^{-1}\preceq H\preceq\gamma^+\Theta_*^{-1}\}}_{\bar{\mathcal{H}}}\times\mathcal{V}\to\mathbf{R}\cup\{+\infty\}\cup\{-\infty\}.$$

Second, we restrict the function $\bar\Phi(h,H;\Theta)$ from $\bar{\mathcal{H}}\times\mathcal{V}$ onto $\mathcal{H}\times\mathcal{V}$. For $(h,H)\in\bar{\mathcal{H}}$, the set of G's linked to (h,H) by the above convex constraints clearly is a nonempty compact set; as a result, $\bar\Phi$ is real-valued convex-concave function on $\bar{\mathcal{H}}\times\mathcal{V}$. From continuity of $\Psi^\delta_{A,\mathcal{Z}}$ on its domain it immediately follows that $\Psi^\delta_{A,\mathcal{Z}}$ is bounded and uniformly continuous on every bounded subset of this domain, implying by evident reasons that $\bar\Phi(h,H;\Theta)$ is bounded in every domain of the form $\bar B\times\mathcal{V}$, where $\bar B$ is a bounded subset of $\bar{\mathcal{H}}$, and is continuous on $\bar B\times\mathcal{V}$ in $\Theta\in\mathcal{V}$ with properly selected modulus of continuity independent of $(h,H)\in\bar B$. Besides this, by construction, $\mathcal{H}\subset\operatorname{int}\bar{\mathcal{H}}$, implying that if B is a convex compact subset of $\mathcal{H}$, it belongs to the interior of a properly selected convex compact subset $\bar B$ of $\bar{\mathcal{H}}$. Since $\bar\Phi$ is bounded on $\bar B\times\mathcal{V}$ and is convex in (h,H), the function $\bar\Phi$ is Lipschitz continuous in $(h,H)\in B$ with Lipschitz constant which can be selected to be independent of $\Theta\in\mathcal{V}$. Taking into account that $\mathcal{H}$ is convex and closed, the bottom line is that $\Phi^\delta_{A,\mathcal{Z}}$ is not just real-valued convex-concave function on the domain $\mathcal{H}\times\mathcal{V}$, it is also continuous on this domain.

Coerciveness of $\Phi^\delta_{A,\mathcal{Z}}(h,H;\Theta)$ in (h,H) is proved in exactly the same fashion as the similar property of function (2.8.37), see item 5^0 in the proof of Proposition 2.8.7. The proof of item (i) of Proposition 2.8.9 is complete.

Item (ii) of Proposition 2.8.9 can be derived from item (i) of Proposition in exactly the same fashion as when proving Proposition 2.8.7. □

2.9 Exercises for Lecture 2

† marks more difficult exercises.

2.9.1 Two-point lower risk bound

Exercise 2.1 *Let p and q be two distinct from each other probability distributions on d-element observation space $\Omega=\{1,...,d\}$, and consider two simple hypotheses on the distribution of observation $\omega\in\Omega$, $H_1:\omega\sim p$, and $H_2:\omega\sim q$.*

1. *Is it true that there always exists a simple deterministic test deciding on H_1,H_2 with risk $<1/2$?*

2. *Is it true that there always exists a simple randomized test deciding on H_1,H_2 with risk $<1/2$?*

3. *Is it true that when quasi-stationary K-repeated observations are allowed, one can decide on H_1, H_2 with a whatever small risk, provided K is large enough?*

2.9.2 Hypothesis testing via ℓ_1-separation

Let d be a positive integer, and the observation space Ω be the finite set $\{1,...,d\}$ equipped with the counting reference measure[31]. Probability distributions on Ω can be identified with points p of d-dimensional *probabilistic simplex*

$$\mathbf{\Delta}_d=\{p\in\mathbf{R}^d:p\geq0,\sum_i p_i=1\};$$

[31] Counting measure is the measure on a discrete (finite or countable) set Ω which assigns every point of Ω with mass 1, so that the measure of a subset of Ω is the cardinality of the subset when it is finite and is $+\infty$ otherwise.

i-th entry p_i in $p \in \mathbf{\Delta}_d$ is the probability for the distributed according to p random variable to take value $i \in \{1, ..., d\}$. With this interpretation, p is the probability density taken w.r.t. the counting measure on Ω.

Assume B and W are two nonintersecting nonempty closed convex subsets of $\mathbf{\Delta}_d$; we interpret B and W as black and white probability distributions on Ω, and our goal is to find optimal, in terms of its total risk, test deciding on the hypotheses

$$H_1 : p \in B,\ H_2 : p \in W$$

via a single observation $\omega \sim p$.
Warning: Everywhere in this Section, "test" means "simple test."

Exercise 2.2 *Consider the convex optimization problem*

$$\mathrm{Opt} = \min_{p\in B, q\in W} \left[f(p,q) := \sum_{i=1}^{d} |p_i - q_i| \right] \tag{2.9.1}$$

and let (p^*, q^*) be an optimal solution to this problem (it clearly exists).

1. *Extract from optimality conditions that there exist reals $\rho_i \in [-1, 1]$, $1 \le i \le n$, such that*

$$\rho_i = \left\{ \begin{array}{rl} 1, & p_i^* > q_i^* \\ -1, & p_i^* < q_i^* \end{array} \right. \tag{2.9.2}$$

and

$$\rho^T(p - p^*) \ge 0 \forall p \in B \;\&\; \rho^T(q - q^*) \le 0\ \forall q \in W. \tag{2.9.3}$$

2. *Extract from the previous item that the test $\mathcal{T}$ which, given an observation $\omega \in \{1, ..., d\}$, accepts H_1 with probability $\pi_\omega = (1 + \rho_\omega)/2$ and accepts H_2 with complementary probability, has total risk equal to*

$$\sum_{\omega\in\Omega} \min[p_\omega^*, q_\omega^*] \tag{2.9.4}$$

and thus is minimax optimal in terms of the total risk.

Comments. Exercise 2.2 describes an efficiently computable and *optimal in terms of worst-case total risk* simple test deciding on a pair of "convex" composite hypotheses on the distribution of a discrete random variable. While it seems an attractive result, we believe *by itself* this result is useless, since usually in the testing problem in question a *single* observation by far is not enough for a reasonable inference; such an inference requires observing *several* independent realizations $\omega_1, ..., \omega_K$ of the random variable in question. And construction presented in Exercise 2.2 says nothing on how to adjust the test to the case of repeated observation. Of course, when $\omega^K = (\omega_1, ..., \omega_K)$ is K-element i.i.d. sample drawn from a probability distribution p on $\Omega = \{1, ..., d\}$, ω^K can be thought of as a single observation of discrete random variable taking value in the set $\Omega^K = \underbrace{\Omega \times ... \times \Omega}_{K}$, the probability distribution p^K of ω^K being readily given by p; so why not to apply the construction from Exercise 2.2 to ω^K in the role of ω? On a close inspection, this idea fails. One serious reason for this failure is that the cardinality of Ω^K (which, among other factors, is responsible for the computational complexity of building the test in Exercise 2.2) blows up exponentially as K grows. Another, even more serious, complication is that p^K depends on p nonlinearly, so that the family of distributions p^K of ω^K induced by a convex family of distributions p of ω, convexity meaning that p's in question fill a *convex* subset of the probabilistic simplex, is not convex; and convexity of the sets B, W in the context of Exercise 2.2 is crucial. Thus, passing from single realization of discrete random variable to the sample of $K > 1$ independent realizations of the variable results in severe structural and quantitative complications "killing," at least at the first glance, the approach undertaken in Exercise 2.2.

In spite of the above pessimistic conclusions, the single-observation test from Exercise 2.2 admits a meaningful multi-observation modification, which is the subject of our next Exercise.

Exercise 2.3 There is a straightforward way to use the optimal, in terms of its total risk, single-observation test built in Exercise 2.2 in the "multi-observation" environment. Specifically, following the notation from Exercise 2.2, let $\rho \in \mathbf{R}^d, p^*, q^*$ be the entities built in this Exercise, so that $p^* \in B$, $q^* \in W$, all entries in ρ belong to $[-1, 1]$, and

$$\{\rho^T p \geq \alpha := \rho^T p^* \;\; \forall p \in B\} \;\&\; \{\rho^T q \leq \beta := \rho^T q^* \,\forall q \in W\} \;\&\; \alpha - \beta = \rho^T[p^* - q^*] = \|p^* - q^*\|_1.$$

Given an i.i.d. sample $\omega^K = (\omega_1, ..., \omega_K)$ with $\omega_t \sim p$, where $p \in B \cup W$, we could try to decide on the hypotheses $H_1 : p \in B$, $H_2 : p \in W$ as follows. Let us set $\zeta_t = \rho_{\omega_t}$. For large K the observable, given ω^K, quantity $\zeta^K := \frac{1}{K}\sum_{t=1}^K \zeta_t$, by the Law of Large Numbers, will be with overwhelming probability close to $\mathbf{E}_{\omega\sim p}\{\rho_\omega\} = \rho^T p$, and the latter quantity is $\geq \alpha$ when $p \in B$ and is $\leq \beta < \alpha$ when $p \in W$. Consequently, selecting a "comparison level" $\ell \in (\beta, \alpha)$, we can decide on the hypotheses $p \in B$ vs. $p \in W$ by computing ζ^K, comparing the result with ℓ, and accepting the hypothesis $p \in B$ when $\zeta^K \geq \ell$, otherwise accepting the alternative $p \in W$. The goal of this Exercise is to quantify the above qualitative considerations. To this end let us fix $\ell \in (\beta, \alpha)$ and K and ask ourselves the following questions:

A. For $p \in B$, how to upper-bound the probability $\mathrm{Prob}_{p_K}\{\zeta^K \leq \ell\}$?

B. For $p \in W$, how to upper-bound the probability $\mathrm{Prob}_{p_K}\{\zeta^K \geq \ell\}$?

Here p_K is the probability distribution of the i.i.d. sample $\omega^K = (\omega_1, ..., \omega_K)$ with $\omega_t \sim p$.

The simplest way to answer these questions is to use Bernstein's bounding scheme. Specifically, to answer question A, let us select $\gamma \geq 0$ and observe that for every probability distribution p on $\{1, 2, ..., d\}$ it holds

$$\underbrace{\mathrm{Prob}_{p_K}\left\{\zeta^K \leq \ell\right\}}_{\pi_{K,-}[p]} \exp\{-\gamma\ell\} \leq \mathbf{E}_{p_K}\left\{\exp\{-\gamma\zeta^K\}\right\} = \left[\sum_{i=1}^d p_i \exp\{-\frac{1}{K}\gamma\rho_i\}\right]^K,$$

whence

$$\ln(\pi_{K,-}[p]) \leq K \ln\left(\sum_{i=1}^d p_i \exp\{-\frac{1}{K}\gamma\rho_i\}\right) + \gamma\ell,$$

implying, via substitution $\gamma = \mu K$, that

$$\forall \mu \geq 0 : \ln(\pi_{K,-}[p]) \leq K\psi_-(\mu, p),\;\; \psi_-(\mu, p) = \ln\left(\sum_{i=1}^d p_i \exp\{-\mu\rho_i\}\right) + \mu\ell.$$

Similarly, setting $\pi_{K,+}[p] = \mathrm{Prob}_{p_K}\left\{\zeta^K \geq \ell\right\}$, we get

$$\forall \nu \geq 0 : \ln(\pi_{K,+}[p]) \leq K\psi_+(\nu, p),\;\; \psi_+(\nu, p) = \ln\left(\sum_{i=1}^d p_i \exp\{\nu\rho_i\}\right) - \nu\ell.$$

Now goes the exercise:

1. *Extract from the above observations that*

$$\mathrm{Risk}(\mathcal{T}^{K,\ell}|H_1,H_2) \leq \exp\{K\varkappa\},\ \varkappa = \max\left[\max_{p\in B}\inf_{\mu\geq 0}\psi_-(\mu,p), \max_{q\in W}\inf_{\nu\geq 0}\psi_+(\nu,q)\right],$$

where $\mathcal{T}^{K,\ell}$ is the K-observation test which accepts the hypothesis $H_1 : p \in B$ when $\zeta^K \geq \ell$ and accepts the hypothesis $H_2 : p \in W$ otherwise.

2. *Verify that $\psi_-(\mu,p)$ is convex in μ and concave in p, and similarly for $\psi_+(\nu,q)$, so that*

$$\max_{p\in B}\inf_{\mu\geq 0}\psi_-(\mu,p) = \inf_{\mu\geq 0}\max_{p\in B}\psi_-(\mu,p),\ \max_{q\in W}\inf_{\nu\geq 0}\psi_+(\nu,q) = \inf_{\nu\geq 0}\max_{q\in W}\psi_+(\nu,q)$$

Thus, computing $\varkappa$ reduces to minimizing on the nonnegative ray the convex functions $\phi_-(\mu) = \max_{p\in B}\psi_+(\mu,p)$ and $\phi_+(\nu) = \max_{q\in W}\psi_+(\nu,q)$.

3. *Prove that when $\ell = \frac{1}{2}[\alpha+\beta]$, one has*

$$\varkappa \leq -\frac{1}{12}\Delta^2,\ \Delta = \alpha-\beta = \|p^*-q^*\|_1. \tag{2.9.5}$$

Note that the above test and the quantity $\varkappa$ responsible for the upper bound on its risk depend, as on a parameter, on the "acceptance level" $\ell \in (\beta,\alpha)$. The simplest way to select a reasonable value of ℓ is to minimize $\varkappa$ over an equidistant grid $\Gamma \subset (\beta,\alpha)$, of small cardinality, of values of ℓ.

Now let us consider an alternative way to pass from a "good" single-observation test to its multi-observation version. Our "building block" now is the minimum rrisk randomized single-observation test[32]*, and its multi-observation modification is just the majority version of this building block. Our first observation is that building the minimum risk single-observation test reduces to solving a convex optimization problem:*

Exercise 2.4 *Let, as above, B and W be nonempty nonintersecting closed convex subsets of probabilistic simplex $\mathbf{\Delta}_d$. Demonstrate that the problem of finding the best, in terms of its risk, randomized single-observation test deciding on $H_1 : p \in B$ vs. $H_2 : p \in W$ via observation $\omega \sim p$ reduces to solving a convex optimization problem. Write down this problem as an explicit LO program when B and W are polyhedral sets given by polyhedral representations:*

$$\begin{array}{rcl} B &=& \{p : \exists u : P_Bp + Q_Bu \leq a_B\}, \\ W &=& \{p : \exists u : P_Wp + Q_Wu \leq a_W\}. \end{array}$$

We see that the "ideal building block" – the minimum-risk single-observation test – can be built efficiently. What is at this point unclear, is whether this block is of any use for majority modifications, that is, whether it is true that the risk of this test is $< 1/2$ – this is what we need to get the possibility for low-risk testing from repeated observations via majority version of the minimum-risk single-observation test.

Exercise 2.5 *Extract from Exercise 2.2 that in the situation of this Section, denoting by Δ the optimal value in the optimization problem* (2.9.1)*, one has*

1. *The risk of any single-observation test, deterministic or randomized alike, is $\geq \frac{1}{2} - \frac{\Delta}{4}$*

2. *There exists a single-observation randomized test with risk $\leq \frac{1}{2} - \frac{\Delta}{8}$, and thus the risk of the minimum risk single-observation test given by Exercise 2.4 does not exceed $\frac{1}{2} - \frac{\Delta}{8} < 1/2$ as well.*

[32] this test can differ from the one built in Exercise 2.2 – the latter test is optimal in terms of the sum, rather than the maximum, of its partial risks.

Pay attention to the fact that $\Delta > 0$ (since, by assumption,. B and W do not intersect).

The bottom line is that in the situation of this Section, given a target value ϵ of risk and assuming stationary repeated observations are allowed, we have (at least) three options to meet the risk specifications:

1. To start with the optimal, in terms of its total risk, single-observation detector as explained in Exercise 2.2, and the to pass to its multi-observation version built in Exercise 2.3;

2. To use the majority version of the minimum-risk randomized single-observation test built in Exercise 2.4;

3. To use the test based on the minimum risk detector for B, W, as explained in the main body of Lecture 2.

In all cases, the number K of observations should be specified as "presumably the smallest" K ensuring that the risk of the resulting multi-observation test is at most a given target ϵ; this K can be easily specified by utilizing the results on the risk of a detector-based test in a Discrete o.s. from the main body of Lecture 2 along with risk-related results of Exercises 2.3, 2.4.

Exercise 2.6 *Run numerical experimentation to get an idea whether one of the three options above always dominates other options (that is, requires smaller sample of observations to ensure the same risk).*

Now let us focus on theoretical comparison of the detector-based test and the majority version of the minimum-risk single-observation test (options 1 and 2 above) in the general situation described in the beginning of Section 2.9.2. Given $\epsilon \in (0, 1)$, the corresponding sample sizes, $K_{\rm d}$ and $K_{\rm m}$, are completely specified each by its own "measure of closeness" between B and W. Specifically,

- For $K_{\rm d}$, the closeness measure is

$$\rho_{\rm d}(B, W) = 1 - \max_{p \in B, q \in W} \sum_{\omega} \sqrt{p_\omega q_\omega}; \tag{2.9.6}$$

  $1 - \rho_{\rm d}(B, W)$ is the minimal risk of a detector for B, W, and for $\rho_{\rm d}(B, W)$ and ϵ small, we have $K_{\rm d} \approx \ln(1/\epsilon)/\rho_{\rm d}(B, W)$ (why?).

- Given ϵ, $K_{\rm m}$ is fully specified by the minimal risk ρ of simple randomized single-observation test $\mathcal{T}$ deciding on the associated with B, W hypotheses; by Exercise 2.5, we have $\rho = \frac{1}{2} - \delta$, where δ is within absolute constant factor of the optimal value $\Delta = \min_{p \in B, q \in W} \|p - q\|_1$ of (2.9.1). The risk bound for the K-observation majority version of $\mathcal{T}$ is the probability to get at least $K/2$ heads in K independent tosses of coin with probability to get head in a single toss equal to $\rho = 1/2 - \delta$. When ρ is not close to 0 and ϵ is small, the $(1 - \epsilon)$-quantile of the number of heads in our K coin tosses is $K\rho + O(1)\sqrt{K \ln(1/\epsilon)} = K/2 - \delta K + O(1)\sqrt{K \ln(1/\epsilon)}$ (why?). $K_{\rm m}$ is the smallest K for which this quantile is $< K/2$, so that $K_{\rm m}$ is of order of $\ln(1/\epsilon)/\delta^2$, or, which is the same, of order of $\ln(1/\epsilon)/\Delta^2$. We see that the "responsible for $K_{\rm m}$" closeness between B and W is

$$\rho_{\rm m}(B, W) = \Delta^2 = \left[\min_{p \in B, q \in W} \|p - q\|_1\right]^2, \tag{2.9.7}$$

  and $K_{\rm m}$ is of order of $\ln(1/\epsilon)/\rho_{\rm m}(B, W)$.

The goal of the next exercise is to compare $\rho_{\rm b}$ and $\rho_{\rm m}$.

Exercise 2.7 . *Prove that in the situation of this Section one has*

$$\frac{1}{8}\rho_{\rm m}(B,W) \leq \rho_{\rm d}(B,W) \leq \frac{1}{2}\sqrt{\rho_{\rm m}(B,W)}. \tag{2.9.8}$$

Relation (2.9.8) suggests that while $K_{\rm d}$ never is "much larger" than $K_{\rm m}$ (this we know in advance: in repeated version of Discrete o.s., properly built detector-based test provably is nearly optimal), but $K_{\rm m}$ could be much larger than $K_{\rm d}$. This indeed is the case:

Exercise 2.8 *Given $\delta \in (0,1/2)$, let $B = \{[\delta;0;1-\delta]\}$ and $W = \{[0;\delta;1-\delta]\}$. Verify that in this case, the numbers of observations $K_{\rm d}$ and $K_{\rm m}$ resulting in a given risk $\epsilon \ll 1$ of multi-observation tests, as functions of δ are proportional to $1/\delta$ and $1/\delta^2$, respectively. Compare the numbers when $\epsilon = 0.01$ and $\delta \in \{0.01;0.05;0.1\}$.*

2.9.3 Miscellaneous exercises

Exercise 2.9 *Prove that the conclusion in Proposition 2.3.5 remains true when the test $\mathcal{T}$ in the premise of Proposition is randomized.*

Exercise 2.10 *Let $p_1(\omega), p_2(\omega)$ be two positive probability densities, taken w.r.t. a reference measure Π, on an observation space Ω, and let $\mathcal{P}_\chi = \{p_\chi\}$, $\chi = 1,2$. Find the optimal, in terms of its risk, balanced detector for $\mathcal{P}_\chi$, $\chi = 1,2$.*

Exercise 2.11 *Recall that the exponential, with parameter $\mu > 0$, distribution on $\Omega = \mathbf{R}_+$ is the distribution with the density $p_\mu(\omega) = \mu e^{-\mu\omega}$, $\omega \geq 0$. Given positive reals $\alpha < \beta$, consider two families of exponential distributions, $\mathcal{P}_1 = \{p_\mu : 0 < \mu \leq \alpha\}$, and $\mathcal{P}_2 = \{p_\mu : \mu \geq \beta\}$. Build the optimal, in terms of its risk, balanced detector for $\mathcal{P}_1$, $\mathcal{P}_2$. What happens with the risk of the detector you have built when the families $\mathcal{P}_\chi$, $\chi = 1,2$, are replaced with their convex hulls?*

Exercise 2.12 [Follow-up to Exercise 2.11] *Assume that the "lifetime" ζ of a lightbulb is a realization of random variable with exponential distribution (i.e., the density $p_\mu(\zeta) = \mu e^{-\mu\zeta}$, $\zeta \geq 0$; in particular, the expected lifetime of a lightbulb in this model is $1/\mu$)* [33]. *Given a lot of lightbulbs, you should decide whether they were produced under normal conditions (resulting in $\mu \leq \alpha = 1$) or under abnormal ones (resulting in $\mu \geq \beta = 1.5$). To this end, you can select at random K lightbulbs and test them. How many lightbulbs should you test in order to make a 0.99-reliable conclusion? Answer this question in the situations when the observation ω in a test is*

1. *the lifetime of a lightbulb (i.e., $\omega \sim p_\mu(\cdot)$)*

2. *the minimum $\omega = \min[\zeta, \delta]$ of the lifetime $\zeta \sim p_\mu(\cdot)$ of a lightbulb and the allowed duration $\delta > 0$ of your test (i.e., if the lightbulb you are testing does not "die" on time horizon δ, you terminate the test)*

3. *$\omega = \chi_{\zeta<\delta}$, that is, $\omega = 1$ when $\zeta < \delta$, and $\omega = 0$ otherwise; here, as above, $\zeta \sim p_\mu(\cdot)$ is the random lifetime of a lightbulb, and $\delta > 0$ is the allowed test duration (i.e., you observe whether or not a lightbulb "dies" on time horizon δ, but do not register the lifetime when it is $< \delta$).*

[33] In Reliability, probability distribution of the lifetime ζ of an organism or a technical device is characterized by the *failure rate* $\lambda(t) = \lim_{dt\to+0} \frac{\mathrm{Prob}\{t\leq\zeta\leq t+dt\}}{dt\cdot\mathrm{Prob}\{\zeta\geq t\}}$ (so that for small dt, $\lambda(t)dt$ is the conditional probability to "die" in the time interval $[t, t+dt]$ provided the lifetime is at least t). The exponential distribution corresponds to the case of failure rate independent of t; usually, this indeed is nearly so except for "very small" and "very large" values of t.

Consider the values $0.25, 0.5, 1, 2, 4$ *of* δ.

Exercise 2.13 [Follow-up to Exercise 2.12] *In the situation of Exercise 2.12, build a sequential test for deciding on Null hypothesis "the lifetime of a lightbulb from a given lot is* $\zeta \sim p_\mu(\cdot)$ *with* $\mu \leq 1$*" (recall that* $p_\mu(z)$ *is the exponential density* $\mu e^{-\mu z}$ *on the ray* $\{z \geq 0\}$*) vs. the alternative "the lifetime is* $\zeta \sim p_\mu(\cdot)$ *with* $\mu > 1$.*" In this test, you can select a number* K *of lightbulbs from the lot, switch them on at time 0 and record the actual lifetimes of the lightbulbs you are testing. As a result at the end of (any) observation interval* $\Delta = [0, \delta]$*, you observe* K *independent realizations of r.v.* $\min[\zeta, \delta]$*, where* $\zeta \sim p_\mu(\cdot)$ *with some unknown* μ*. In your sequential test, you are welcome to make conclusions at the endpoints* $\delta_1 < \delta_2 < ... < \delta_S$ *of several observation intervals.*

Note: We deliberately skip details of problem's setting; how you decide on these missing details, is part of your solution to Exercise.

Exercise 2.14 *In Section 2.6, we considered a model of elections where every member of population was supposed to cast a vote. Enrich the model by incorporating the option for a voter not to participate in the elections at all. Implement Sequential test for the resulting model and run simulations.*

Exercise 2.15 *Work out the following extension of the DOP problem. You are given two finite sets,* $\Omega_1 = \{1, ..., I\}$ *and* $\Omega_2 = \{1, ..., M\}$*, along with* L *nonempty closed convex subsets* Y_ℓ *of the set*

$$\mathbf{\Delta}_{IM} = \{[y_{im} > 0]_{i,m} : \sum_{i=1}^{I} \sum_{m=1}^{M} y_{im} = 1\}$$

of all non-vanishing probability distributions on $\Omega = \Omega_1 \times \Omega_2 = \{(i, m) : 1 \leq i \leq I, 1 \leq m \leq M\}$*. The sets* Y_ℓ *are such that all distributions from* Y_ℓ *have a common marginal distribution* $\theta^\ell > 0$ *of* i*:*

$$\sum_{m=1}^{M} y_{im} = \theta_i^\ell, \; 1 \leq i \leq I, \; \forall y \in Y_\ell, \; 1 \leq \ell \leq L.$$

Your observations $\omega_1, \omega_2, ...$ *are sampled, independently of each other, from a distribution partly selected "by nature," and partly – by you. Specifically, the nature selects* $\ell \leq L$ *and a distribution* $y \in Y_\ell$*, and you select a positive* I*-dimensional probabilistic vector* q *from a given convex compact subset* $\mathcal{Q}$ *of the positive part of* I*-dimensional probabilistic simplex. Let* $y_{|i}$ *be the conditional,* i *being given, distribution of* $m \in \Omega_2$ *induced by* y*, so that* $y_{|i}$ *is the* M*-dimensional probabilistic vector with entries*

$$[y_{|i}]_m = \frac{y_{im}}{\sum_{\mu \leq M} y_{i\mu}} = \frac{y_{im}}{\theta_i^\ell}.$$

In order to generate $\omega_t = (i_t, m_t) \in \Omega$*, you draw* i_t *at random from the distribution* q*, and then the nature draws* m_t *at random from the distribution* $y_{|i_t}$*.*

Given closeness relation $\mathcal{C}$*, your goal is to decide, up to closeness* $\mathcal{C}$*, on the hypotheses* $H_1, ..., H_L$*, with* H_ℓ *stating that the distribution* y *selected by the nature belongs to* Y_ℓ*. Given "observation budget" (a number* K *of observations* ω_k *you can use), you want to find a probabilistic vector* q *which results in the test with as small* $\mathcal{C}$*-risk as possible. Pose this Measurement Design problem as an efficiently solvable convex optimization problem.*

Exercise 2.16 [probabilities of deviations from the mean]

The goal of what follows is to present the most straightforward application of simple families of distributions – bounds on probabilities of deviations of random vectors from their means.

Let $\mathcal{H} \subset \Omega = \mathbf{R}^d, \mathcal{M}, \Phi$ be a regular data such that $0 \in \operatorname{int} \mathcal{H}$, $\mathcal{M}$ is compact, $\Phi(0;\mu) = 0\,\forall \mu \in \mathcal{M}$, and $\Phi(h;\mu)$ is differentiable at $h = 0$ for every $\mu \in \mathcal{M}$. Let, further, $\bar{P} \in \mathcal{S}[\mathcal{H}, \mathcal{M}, \Phi]$ and let $\bar{\mu} \in \mathcal{M}$ be a parameter of $\bar{P}$. Prove that

1. *$\bar{P}$ possesses expectation $e[\bar{P}]$, and*

$$e[\bar{P}] = \nabla_h \Phi(0; \bar{\mu})$$

2. *For every linear form $e^T\omega$ on Ω it holds*

$$\pi := \bar{P}\{\omega : e^T(\omega - e[\bar{P}]) \geq 1\} \leq \exp\left\{ \sup_{\mu \in \mathcal{M}} \inf_{t \geq 0: te \in \mathcal{H}} \left[\Phi(te; \bar{\mu}) - te^T \nabla_h \Phi(0; \bar{\mu}) - t\right] \right\}. \quad (2.9.9)$$

Exercise 2.17 [testing convex hypotheses on mixtures] *Consider the situation as follows. For given positive integers K, L and for $\chi = 1, 2$, given are*

- *nonempty convex compact signal sets $U_\chi \subset \mathbf{R}^{n_\chi}$*
- *regular data $\mathcal{H}^\chi_{k\ell} \subset \mathbf{R}^{d_k}, \mathcal{M}^\chi_{k\ell}, \Phi^\chi_{k\ell}$, and affine mappings*

$$u_\chi \mapsto A^\chi_{k\ell}[u_\chi; 1] : \mathbf{R}^{n_\chi} \to \mathbf{R}^{d_k}$$

  such that

$$u_\chi \in U_\chi \Rightarrow A^\chi_{k\ell}[u_\chi; 1] \in \mathcal{M}^\chi_{k\ell},$$

  $1 \leq k \leq K$, $1 \leq \ell \leq L$,

- *probability vectors $\mu^k = [\mu^k_1; ...; \mu^k_L]$, $1 \leq k \leq K$.*

We can associate with the outlined data families of probability distributions $\mathcal{P}_\chi$ on the observation space $\Omega = \mathbf{R}^{d_1} \times ... \times \mathbf{R}^{d_K}$ as follows. For $\chi = 1, 2$, $\mathcal{P}_\chi$ is comprised of all probability distributions P of random vectors $\omega^K = [\omega_1; ...; \omega_K] \in \Omega$ generated as follows:
We select

- *a signal $u \in U_\chi$,*
- *a collection of probability distributions $P_{k\ell} \in \mathcal{S}[\mathcal{H}^\chi_{k\ell}, \mathcal{M}^\chi_{k\ell}, \Phi^\chi_{k\ell}]$, $1 \leq k \leq K$, $1 \leq \ell \leq L$, in such a way that $A^\chi_{k\ell}[u_\chi; 1]$ is a parameter of $P_{k\ell}$:*

$$\forall h \in \mathcal{H}^\chi_{k\ell} : \ln\left(\mathbf{E}_{\omega_k \sim P_{k\ell}} e^{h^T \omega_k}\right) \leq \Phi^\chi_{k\ell}(h_k; A^\chi_{k\ell}[u_\chi; 1]);$$

- *we generate the components ω_k, $k = 1, ..., K$, independently across k, from μ^k-mixture $\Pi[\{P_{k\ell}\}_{\ell=1}^L, \mu]$ of distributions $P_{k\ell}$, $\ell = 1, ..., L$, that is, draw at random, from distribution μ^k on $\{1, ..., L\}$, index ℓ, and then draw ω_k from the distribution $P_{k\ell}$.*

Prove that setting

$$\begin{array}{rcl}
\mathcal{H}_\chi &=& \{h = [h_1; ...; h_K] \in \mathbf{R}^{d = d_1 + ... + d_K} : h_k \in \bigcap\limits_{\ell=1}^{L} \mathcal{H}^\chi_{k\ell}, 1 \leq k \leq K\}, \\
\mathcal{M}_\chi &=& \{0\} \subset \mathbf{R}, \\
\Phi_\chi(h; \mu) &=& \sum_{k=1}^K \ln\left(\sum_{\ell=1}^L \mu^k_\ell \max\limits_{u_\chi \in U_\chi} \Phi^\chi_{k\ell}(h_k; A^\chi_{k\ell}[u_\chi; 1])\right) : \mathcal{H}_\chi \times \mathcal{M}_\chi \to \mathbf{R},
\end{array}$$

we get regular data such that

$$\mathcal{P}_\chi \subset \mathcal{S}[\mathcal{H}_\chi, \mathcal{M}_\chi, \Phi_\chi].$$

Explain how to use this observation to compute via Convex Programming affine detector, along with its risk, for the families of distributions $\mathcal{P}_1, \mathcal{P}_2$.

Exercise 2.18 [mixture of sub-Gaussian distributions] *Let P_ℓ be sub-Gaussian distributions on $\mathbf{R}^d$ with sub-Gaussianity parameters θ_ℓ, Θ, $1 \leq \ell \leq L$, with a common Θ-parameter, and let $\nu = [\nu_1; ...; \nu_L]$ be a probabilistic vector. Consider the ν-mixture $P = \Pi[P^L, \nu]$ of distributions P_ℓ, so that $\omega \sim P$ is generated as follows: we draw at random from distribution ν index ℓ and then draw ω at random from distribution P_ℓ. Prove that P is sub-Gaussian with sub-Gaussianity parameters $\bar\theta = \sum_\ell \nu_\ell \theta_\ell$ and $\bar\Theta$, with (any) $\bar\Theta$ chosen to satisfy*

$$\bar\Theta \succeq \Theta + \frac{6}{5}[\theta_\ell - \bar\theta][\theta_\ell - \bar\theta]^T \,\forall \ell,$$

in particular, according to any one of the following rules:

1. $\bar\Theta = \Theta + \left(\frac{6}{5}\max_\ell \|\theta_\ell - \bar\theta\|_2^2\right) I_d$,

2. $\bar\Theta = \Theta + \frac{6}{5}\sum_\ell (\theta_\ell - \bar\theta)(\theta_\ell - \bar\theta)^T$,

3. $\bar\Theta = \Theta + \frac{6}{5}\sum_\ell \theta_\ell \theta_\ell^T$, *provided that $\nu_1 = ... = \nu_L = 1/L$,*

Exercise 2.19 The goal of this Exercise is to give a simple sufficient condition for quadratic lift "to work" in the Gaussian case. Specifically, let $\mathcal{A}_\chi$, U_χ, $\mathcal{V}_\chi$, $\mathcal{G}_\chi$, $\chi = 1, 2$, be as in Section 2.8.3.3, with the only difference that now we do *not* assume the compact sets $\mathcal{U}_\chi$ to be convex, and let $\mathcal{Z}_\chi$ be convex compact subsets of the sets $\mathcal{Z}^{n_\chi}$, see (2.8.34), such that

$$[u_\chi; 1][u_\chi; 1]^T \in \mathcal{Z}_\chi \;\forall u_\chi \in U_\chi,\; \chi = 1, 2.$$

Augmenting the above data with $\Theta_\chi^{(*)}$, δ_χ such that $\mathcal{V} = \mathcal{V}_\chi$, $\Theta_* = \Theta_*^{(\chi)}$, $\delta = \delta_\chi$ satisfy (2.8.35), $\chi = 1, 2$, and invoking Proposition 2.8.7.ii, we get at our disposal a quadratic detector $\phi_{\rm lift}$ such that

$$\mathrm{Risk}[\phi_{\rm lift}|\mathcal{G}_1, \mathcal{G}_2] \leq \exp\{-\mathrm{SadVal}_{\rm lift}\},$$

with $\mathrm{SadVal}_{\rm lift}$ given by (2.8.42). A natural question is, when $\mathrm{SadVal}_{\rm lift}$ is negative, meaning that our quadratic detector indeed "is working" – its risk is < 1, implying that when repeated observations are allowed, tests based upon this detector allow to decide on the hypotheses $H_\chi : P \in \mathcal{G}_\chi$, $\chi = 1, 2$, on the distribution of observation $\zeta \sim P$ with a whatever small desired risk $\epsilon \in (0, 1)$. With our computation-oriented ideology, this is not too important question, since we can answer it via efficient computation. This being said, there is no harm in a "theoretical" answer which could provide us with an additional insight. The goal of the Exercise is to justify a simple result on the subject. Here is the Exercise:

In the situation in question, assume that $\mathcal{V}_1 = \mathcal{V}_2 = \{\Theta_*\}$, which allows to set $\Theta_*^{(\chi)} = \Theta_*$, $\delta_\chi = 0$, $\chi = 1, 2$. Prove that in the case in question a necessary and sufficient condition for $\mathrm{SadVal}_{\rm lift}$ *to be negative is that the convex compact sets*

$$\mathcal{U}_\chi = \{B_\chi Z B_\chi^T : Z \in \mathcal{Z}_\chi\} \subset \mathbf{S}_+^{d+1},\; \chi = 1, 2$$

do not intersect with each other.

Exercise 2.20 *Prove if X is a nonempty convex compact set in $\mathbf{R}^d$, then the function $\widehat\Phi(h; \mu)$ given by* (2.8.14) *is real-valued and continuous on $\mathbf{R}^d \times X$ and is convex in h and concave in μ.*

Lecture 3

Estimating Functions via Hypothesis Testing

In this Lecture we apply the hypothesis testing techniques developed in Lecture 2 to estimating properly structured scalar functionals in simple o.s.'s (Section 3.2) and beyond (Section 3.4).

3.1 Estimating linear forms on unions of convex sets

3.1.1 The problem

Let $\mathcal{O} = ((\Omega, \Pi), \{p_\mu(\cdot) : \mu \in \mathcal{M}\}, \mathcal{F})$ be a simple observation scheme. The problem we are interested in this section is as follows:

> We are given a positive integer K and I nonempty convex compact sets $X_j \subset \mathbf{R}^n$, along with affine mappings $A_j(\cdot) : \mathbf{R}^n \to \mathbf{R}^M$ such that $A_j(x) \in \mathcal{M}$ whenever $x \in X_j$, $1 \leq j \leq I$. In addition, we are given a linear function $g^T x$ on $\mathbf{R}^n$.
>
> Given random observation
> $$\omega^K = (\omega_1, ..., \omega_K)$$
> with ω_k drawn, independently across k, from $p_{A_j(x)}$ with $j \leq I$ and $x \in X_j$, we want to recover $g^T x$. It should be stressed that we do *not* know neither j nor x underlying our observation.

Given reliability tolerance $\epsilon \in (0, 1)$, we quantify the performance of a candidate estimate – a Borel function $\widehat{g}(\cdot) : \Omega \to \mathbf{R}$ – by the worst case, over j and x, width of $(1-\epsilon)$-confidence interval, Specifically, we say that $\widehat{g}(\cdot)$ is (ρ, ϵ)-reliable, if

$$\forall (j \leq I, x \in X_j) : \mathrm{Prob}_{\omega \sim p_{A_j(x)}}\{|\widehat{g}(\omega) - g^T x| > \rho\} \leq \epsilon.$$

We define ϵ-risk of the estimate as

$$\mathrm{Risk}_\epsilon[\widehat{g}] = \inf\{\rho : \widehat{g} \text{ is } (\rho, \epsilon)\text{-reliable}\};$$

note that $\widehat{g}$ is the smallest ρ such that $\widehat{g}$ is (ρ, ϵ)-reliable.

We remark that the technique we are about to use originates from [92] where recovery, in a simple o.s., of a linear form on a convex compact set (i.e., the case $I = 1$ of the estimation problem at hand) was considered; it was proved that in this situation the estimate

$$\widehat{g}(\omega^K) = \sum_k \phi(\omega_k) + \varkappa$$

with properly selected $\phi \in \mathcal{F}$ and $\kappa \in \mathbf{R}$ is near-optimal; for Gaussian o.s. similar fact was discovered, by different technique, by D. Donoho [51] as early as in 1994.

3.1.2 The estimate

In the sequel, we associate with the simple o.s. $\mathcal{O} = ((\Omega,\Pi),\{p_\mu(\cdot):\mu\in\mathcal{M}\},\mathcal{F})$ in question the function

$$\Phi_{\mathcal{O}}(\phi;\mu)=\ln\left(\int \mathrm{e}^{\phi(\omega)}p_\mu(\omega)\Pi(d\omega)\right):\mathcal{F}\times\mathcal{M}\to\mathbf{R}.$$

Recall that by definition of a simple o.s., this function is real-valued on its domain and is concave in $\mu\in\mathcal{M}$, convex in $\phi\in\mathcal{F}$, and continuous on $\mathcal{F}\times\mathcal{M}$ (the latter follows from convexity-concavity and relative openness of $\mathcal{M}$ and $\mathcal{F}$).

Let us associate with a pair (i,j), $1\le i,j\le I$, the functions

$$\begin{array}{rcl}
\Phi_{ij}(\alpha,\phi;x,y) &=& \frac{1}{2}\left[K\alpha\Phi_{\mathcal{O}}(\phi/\alpha;A_i(x))+K\alpha\Phi_{\mathcal{O}}(-\phi/\alpha;A_j(y))+g^T(y-x)+2\alpha\ln(2I/\epsilon)\right]:\\
&&\qquad\qquad\{\alpha>0,\phi\in\mathcal{F}\}\times[X_i\times X_j]\to\mathbf{R},\\
\Psi_{ij}(\alpha,\phi) &=& \max\limits_{x\in X_i,y\in X_j}\Phi_{ij}(\alpha,\phi;x,y)=\frac{1}{2}\left[\Psi_{i,+}(\alpha,\phi)+\Psi_{j,-}(\alpha,\phi)\right]:\{\alpha>0\}\times\mathcal{F}\to\mathbf{R},\\
\Psi_{\ell,+}(\beta,\psi) &=& \max_{x\in X_\ell}\left[K\beta\Phi_{\mathcal{O}}(\psi/\beta;A_\ell(x))-g^Tx+\beta\ln(2I/\epsilon)\right]:\{\beta>0,\psi\in\mathcal{F}\}\to\mathbf{R},\\
\Psi_{\ell,-}(\beta,\psi) &=& \max_{x\in X_\ell}\left[K\beta\Phi_{\mathcal{O}}(-\psi/\beta;A_\ell(x))+g^Tx+\beta\ln(2I/\epsilon)\right]:\{\beta>0,\psi\in\mathcal{F}\}\to\mathbf{R}.
\end{array}\tag{3.1.1}$$

Note that the function $\alpha\Phi_{\mathcal{O}}(\phi/\alpha;A_i(x))$ is obtained from continuous convex-concave function $\Phi_{\mathcal{O}}(\cdot,\cdot)$ by projective transformation in the convex argument, and affine substitution in the concave argument, so that the former function is convex-concave and continuous on the domain $\{\alpha>0,\phi\in\mathcal{X}\}\times X_i$. By similar argument, the function $\alpha\Phi_{\mathcal{O}}(-\phi/\alpha;A_j(y))$ is convex-concave and continuous on the domain $\{\alpha>0,\phi\in\mathcal{F}\}\times X_j$. These observations combine with compactness of X_i,X_j to imply that $\Psi_{ij}(\alpha,\phi)$ is real-valued continuous convex function on the domain

$$\mathcal{F}^+=\{\alpha>0\}\times\mathcal{F}.$$

Observe that functions $\Psi_{ii}(\alpha,\phi)$ are nonnegative on $\mathcal{F}^+$. Indeed, selecting somehow $\bar{x}\in X_i$, and setting $\mu=A_i(\bar{x})$, we have

$$\begin{array}{l}
\Psi_{ii}(\alpha,\phi)\ge\Phi_{ii}(\alpha,\phi;\bar{x},\bar{x})=\frac{\alpha}{2}\left[K[\Phi_{\mathcal{O}}(\phi/\alpha;\mu)+\Phi_{\mathcal{O}}(-\phi/\alpha;\mu)]+2\ln(2I/\epsilon)\right]\\
=\frac{\alpha}{2}\left[K\ln\left(\underbrace{\left[\int\exp\{\phi(\omega)/\alpha\}p_\mu(\omega)\Pi(d\omega)\right]\left[\int\exp\{-\phi(\omega)/\alpha\}p_\mu(\omega)\Pi(d\omega)\right]}_{\ge[\int\exp\{\frac{1}{2}\phi(\omega)/\alpha\}\exp\{-\frac{1}{2}\phi(\omega)/\alpha\}p_\mu(\omega)\Pi(d\omega)]^2=1}\right)+2\ln(2I/\epsilon)\right]\\
\ge\alpha\ln(2I/\epsilon)>0
\end{array}$$

(we have used Cauchy inequality).

Functions Ψ_{ij} give rise to convex and feasible optimization problems

$$\mathrm{Opt}_{ij}=\mathrm{Opt}_{ij}(K)=\min_{(\alpha,\phi)\in\mathcal{F}^+}\Psi_{ij}(\alpha,\phi).\tag{3.1.2}$$

By its origin, Opt_{ij} is either a real, or $-\infty$; by the observation above, Opt_{ii} are nonnegative. Our estimate is as follows.

1. For $1\le i,j\le I$, we select somehow feasible solutions α_{ij},ϕ_{ij} to problems (3.1.2) (the less the values of the corresponding objectives, the better) and set

$$\begin{array}{rcl}
\rho_{ij} &=& \Psi_{ij}(\alpha_{ij},\phi_{ij})=\frac{1}{2}\left[\Psi_{i,+}(\alpha_{ij},\phi_{ij})+\Psi_{j,-}(\alpha_{ij},\phi_{ij})\right]\\
\varkappa_{ij} &=& \frac{1}{2}\left[\Psi_{j,-}(\alpha_{ij},\phi_{ij})-\Psi_{i,+}(\alpha_{ij},\phi_{ij})\right]\\
g_{ij}(\omega^K) &=& \sum_{k=1}^K\phi_{ij}(\omega_k)+\varkappa_{ij}\\
\rho &=& \max_{1\le i,j\le I}\rho_{ij}
\end{array}\tag{3.1.3}$$

2. Given observation ω^K, we specify the estimate $\widehat{g}(\omega^K)$ as follows:

$$\begin{array}{rcl}
r_i &=& \max_{j\le I}g_{ij}(\omega^K)\\
c_j &=& \min_{i\le I}g_{ij}(\omega^K)\\
\widehat{g}(\omega^K) &=& \frac{1}{2}\left[\min_{i\le I}r_i+\max_{j\le I}c_j\right].
\end{array}\tag{3.1.4}$$

3.1.3 Main result

Proposition 3.1.1 *The ϵ-risk of the estimate we have built can be upper-bounded as follows:*

$$\mathrm{Risk}_\epsilon[\widehat{g}] \leq \rho. \tag{3.1.5}$$

Proof. Let the common distribution p of independent across k components ω_k in observation ω^K be $p_{A_\ell(u)}$ for some $\ell \leq I$ and $u \in X_\ell$. Let us fix these ℓ and u, let $\mu = A_\ell(u)$, and let p^K stand for the distribution of ω^K.

1^0. We have

$$\begin{array}{l}
\Psi_{\ell,+}(\alpha_{\ell j}, \phi_{\ell j}) = \max_{x \in X_\ell} \left[K\alpha_{\ell j}\Phi_{\mathcal{O}}(\phi_{\ell j}/\alpha_{\ell j}, A_\ell(x)) - g^T x\right] + \alpha_{\ell j}\ln(2I/\epsilon) \\
\geq K\alpha_{\ell j}\Phi_{\mathcal{O}}(\phi_{\ell j}/\alpha_{\ell j}, \mu) - g^T u + \alpha_{\ell j}\ln(2I/\epsilon) \text{ [since } u \in X_\ell \text{ and } \mu = A_\ell(u)] \\
= K\alpha_{\ell j}\ln\left(\int \exp\{\phi_{\ell j}(\omega)/\alpha_{\ell j}\}p_\mu(\omega)\Pi(d\omega)\right) - g^T u + \alpha_{\ell j}\ln(2I/\epsilon) \text{ [by definition of } \Phi_{\mathcal{O}}] \\
= \alpha_{\ell j}\ln\left(\mathbf{E}_{\omega^K \sim p^K}\left\{\exp\{\alpha_{\ell j}^{-1}\sum_k \phi_{\ell j}(\omega_k)\}\right\}\right) - g^T u + \alpha_{\ell j}\ln(2I/\epsilon) \\
= \alpha_{\ell j}\ln\left(\mathbf{E}_{\omega^K \sim p^K}\left\{\exp\{\alpha_{\ell j}^{-1}[g_{\ell j}(\omega^K) - \varkappa_{\ell j}]\}\right\}\right) - g^T u + \alpha_{\ell j}\ln(2I/\epsilon) \\
= \alpha_{\ell j}\ln\left(\mathbf{E}_{\omega^K \sim p^K}\left\{\exp\{\alpha_{\ell j}^{-1}[g_{\ell j}(\omega^K) - g^T u - \rho_{\ell j}]\}\right\}\right) + \rho_{\ell j} - \varkappa_{\ell j} + \alpha_{\ell j}\ln(2I/\epsilon) \\
\geq \alpha_{\ell j}\ln\left(\mathrm{Prob}_{\omega^K \sim p^K}\left\{g_{\ell j}(\omega^K) > g^T u + \rho_{\ell j}\right\}\right) + \rho_{\ell j} - \varkappa_{\ell j} + \alpha_{\ell j}\ln(2I/\epsilon) \\
\Rightarrow \\
\alpha_{\ell j}\ln\left(\mathrm{Prob}_{\omega^K \sim p^K}\left\{g_{\ell j}(\omega^K) > g^T u + \rho_{\ell j}\right\}\right) \leq \Psi_{\ell,+}(\alpha_{\ell j}, \phi_{\ell j}) + \varkappa_{\ell j} - \rho_{\ell j} + \alpha_{\ell j}\ln(\frac{\epsilon}{2I}) \\
= \alpha_{\ell j}\ln(\frac{\epsilon}{2I}) \text{ [by (3.1.3)]}
\end{array}$$

and we arrive at

$$\mathrm{Prob}_{\omega^K \sim p^K}\left\{g_{\ell j}(\omega^K) > g^T u = \rho_{\ell j}\right\} \leq \frac{\epsilon}{2I}. \tag{3.1.6}$$

Similarly,

$$\begin{array}{l}
\Psi_{\ell,-}(\alpha_{i\ell}, \phi_{i\ell}) = \max_{y \in X_\ell} \left[K\alpha_{i\ell}\Phi_{\mathcal{O}}(-\phi_{i\ell}/\alpha_{i\ell}, A_\ell(y)) + g^T y\right] + \alpha_{i\ell}\ln(2I/\epsilon) \\
\geq K\alpha_{i\ell}\Phi_{\mathcal{O}}(-\phi_{i\ell}/\alpha_{i\ell}, \mu) + g^T u + \alpha_{i\ell}\ln(2I/\epsilon) \text{ [since } u \in X_\ell \text{ and } \mu = A_\ell(u)] \\
= K\alpha_{i\ell}\ln\left(\int \exp\{-\phi_{i\ell}(\omega)/\alpha_{i\ell}\}p_\mu(\omega)\Pi(d\omega)\right) + g^T u + \alpha_{i\ell}\ln(2I/\epsilon) \text{ [by definition of } \Phi_{\mathcal{O}}] \\
= \alpha_{i\ell}\ln\left(\mathbf{E}_{\omega^K \sim p^K}\left\{\exp\{-\alpha_{i\ell}^{-1}\sum_k \phi_{i\ell}(\omega_k)\}\right\}\right) + g^T u + \alpha_{i\ell}\ln(2I/\epsilon) \\
= \alpha_{i\ell}\ln\left(\mathbf{E}_{\omega^K \sim p^K}\left\{\exp\{\alpha_{i\ell}^{-1}[-g_{i\ell}(\omega^K) + \varkappa_{i\ell}]\}\right\}\right) + g^T u + \alpha_{i\ell}\ln(2I/\epsilon) \\
= \alpha_{i\ell}\ln\left(\mathbf{E}_{\omega^K \sim p^K}\left\{\exp\{\alpha_{i\ell}^{-1}[-g_{i\ell}(\omega^K) + g^T u - \rho_{i\ell}]\}\right\}\right) + \rho_{i\ell} + \varkappa_{i\ell} + \alpha_{i\ell}\ln(2I/\epsilon) \\
\geq \alpha_{i\ell}\ln\left(\mathrm{Prob}_{\omega^K \sim p^K}\left\{g_{i\ell}(\omega^K) < g^T u - \rho_{i\ell}\right\}\right) + \rho_{i\ell} + \varkappa_{i\ell} + \alpha_{i\ell}\ln(2I/\epsilon) \\
\Rightarrow \\
\alpha_{i\ell}\ln\left(\mathrm{Prob}_{\omega^K \sim p^K}\left\{g_{i\ell}(\omega^K) < g^T u - \rho_{i\ell}\right\}\right) \leq \Psi_{\ell,-}(\alpha_{i\ell}, \phi_{i\ell}) - \varkappa_{i\ell} - \rho_{i\ell} + \alpha_{i\ell}\ln(\frac{\epsilon}{2I}) \\
= \alpha_{i\ell}\ln(\frac{\epsilon}{2I}) \text{ [by (3.1.3)]}
\end{array}$$

and we arrive at

$$\mathrm{Prob}_{\omega^K \sim p^K}\left\{g_{i\ell}(\omega^K) < g^T u - \rho_{i\ell}\right\} \leq \frac{\epsilon}{2I}. \tag{3.1.7}$$

2^0. Let

$$\mathcal{E} = \{\omega^K : g_{\ell j}(\omega^K) \leq g^T u + \rho_{\ell j}, g_{i\ell}(\omega^K) \geq g^T u - \rho_{i\ell}, 1 \leq i, j \leq I\}.$$

From (3.1.6), (3.1.7) and the union bound it follows that p^K-probability of the event $\mathcal{E}$ is $\geq 1 - \epsilon$. As a result, all we need to complete the proof of Proposition is to verify that

$$\omega^K \in \mathcal{E} \Rightarrow |\widehat{g}(\omega^K) - g^T u| \leq \rho. \tag{3.1.8}$$

Indeed, let us fix $\omega^K \in \mathcal{E}$, and let E be the $I \times I$ matrix with entries $E_{ij} = g_{ij}(\omega^K)$, $1 \leq i, j \leq I$. The quantity r_i, see (3.1.4), is the maximum of entries in i-th row of E, and the quantity c_j is the minimum of entries in j-th column of E; in particular, $r_i \geq E_{ij} \geq c_j$ for all i,j, implying that $r_i \geq c_j$ for all i,j. Now, since $\omega^K \in \mathcal{E}$, we have $E_{\ell\ell} = g_{\ell\ell}(\omega^K) \geq g^T u - \rho_{\ell\ell} \geq g^T u - \rho$ and $E_{\ell j} = g_{\ell j}(\omega^K) \leq g^T u + \rho_{\ell j} \leq g^T u + \rho$ for all j, implying that $r_\ell = \max_j E_{\ell j} \in \Delta = [g^T u - \rho, g^T u + \rho]$. Similarly, $\omega \in \mathcal{E}$ implies that $E_{\ell\ell} = g_{\ell\ell}(\omega^K) \leq g^T u + \rho$ and $E_{i\ell} = g_{i\ell}(\omega^K) \geq g^T u - \rho_{i\ell} \geq g^T u - \rho$ for all i, implying that $c_\ell = \min_i E_{i\ell} \in \Delta$. We see that both r_ℓ and c_ℓ belong to Δ; since $r_* := \min_i r_i \leq r_\ell$ and, as have already seen, $r_i \geq c_\ell$ for all i, we conclude that $r_* \in \Delta$. By similar argument, $c_* := \max_j c_j \in \Delta$ as well. By construction, $\widehat{g}(\omega^K) = \frac{1}{2}[r_* + c_*]$, that is, $\widehat{g}(\omega^K) \in \Delta$, and the conclusion in (3.1.8) indeed takes place. □

3.1.4 Near-optimality

Observe that properly selecting ϕ_{ij} and α_{ij} we can make, in a computationally efficient manner, the upper bound ρ on the ϵ-risk of the above estimate arbitrarily close to

$$\mathrm{Opt}(K) = \max_{1 \leq i,j \leq I} \mathrm{Opt}_{ij}(K).$$

We are about to demonstrate that the quantity $\mathrm{Opt}(K)$ "nearly lower-bounds" the minimax optimal ϵ-risk

$$\mathrm{Risk}^*_\epsilon(K) = \inf_{\widehat{g}(\cdot)} \mathrm{Risk}_\epsilon[\widehat{g}],$$

the infimum being taken over all K-observation Borel estimates. The precise statement is as follows:

Proposition 3.1.2 *In the situation of this Section, let $\epsilon \in (0, 1/2)$ and $\bar{K}$ be a positive integer. Then for every integer K satisfying*

$$K/\bar{K} > \frac{2\ln(2I/\epsilon)}{\ln(\frac{1}{4\epsilon(1-\epsilon)})}$$

one has

$$\mathrm{Opt}(K) \leq \mathrm{Risk}^*_\epsilon(\bar{K}). \tag{3.1.9}$$

In addition, in the special case where for every i,j there exists $x_{ij} \in X_i \cap X_j$ such that $A_i(x_{ij}) = A_j(x_{ij})$ one has

$$K \geq \bar{K} \Rightarrow \mathrm{Opt}(K) \leq \frac{2\ln(2I/\epsilon)}{\ln(\frac{1}{4\epsilon(1-\epsilon)})} \mathrm{Risk}^*_\epsilon(\bar{K}). \tag{3.1.10}$$

Proof. 1^0. Observe that $\mathrm{Opt}_{ij}(K)$ is the saddle point value in the convex-concave saddle point problem:

$$\mathrm{Opt}_{ij}(K) = \inf_{\alpha>0,\phi\in\mathcal{F}} \max_{x\in X_i, y\in X_j} \left[\frac{1}{2}K\alpha\left\{\Phi_{\mathcal{O}}(\phi/\alpha; A_i(x)) + \Phi_{\mathcal{O}}(-\phi/\alpha; A_j(y))\right\} + \frac{1}{2}g^T[y-x] + \alpha\ln(2I/\epsilon)\right].$$

The domain of the maximization variable is compact and the cost function is continuous on its domain, whence, by Sion-Kakutani Theorem, we have also

$$\begin{array}{rcl}\mathrm{Opt}_{ij}(K) & = & \max\limits_{x\in X_i, y\in X_j} \Theta_{ij}(x,y),\\ \Theta_{ij}(x,y) & = & \inf\limits_{\alpha>0,\phi\in\mathcal{F}} \left[\frac{1}{2}K\alpha\left\{\Phi_{\mathcal{O}}(\phi/\alpha; A_i(x)) + \Phi_{\mathcal{O}}(-\phi/\alpha; A_j(y))\right\} + \alpha\ln(2I/\epsilon)\right]\\ & & +\frac{1}{2}g^T[y-x].\end{array} \tag{3.1.11}$$

We have

$$\begin{array}{rcl}\Theta_{ij}(x,y) & = & \inf\limits_{\alpha>0,\psi\in\mathcal{F}} \left[\frac{1}{2}K\alpha\left\{\Phi_{\mathcal{O}}(\psi; A_i(x)) + \Phi_{\mathcal{O}}(-\psi; A_j(y))\right\} + \frac{1}{2}g^T[y-x] + \alpha\ln(2I/\epsilon)\right]\\ & = & \inf\limits_{\alpha>0} \left[\frac{1}{2}\alpha K \inf\limits_{\psi\in\mathcal{F}} \left\{\Phi_{\mathcal{O}}(\psi; A_i(x)) + \Phi_{\mathcal{O}}(-\psi; A_j(y))\right\} + \alpha\ln(2I/\epsilon)\right] + \frac{1}{2}g^T[y-x]\end{array}$$

Given $x \in X_i$, $y \in X_j$ and setting $\mu = A_i(x)$, $\nu = A_j(y)$, we obtain

$$\begin{array}{rcl}\inf\limits_{\psi\in\mathcal{F}}[\Phi_{\mathcal{O}}(\psi;A_i(x))+\Phi_{\mathcal{O}}(-\psi;A_j(y))] &=& \inf\limits_{\psi\in\mathcal{F}}\left[\ln\left(\int\exp\{\psi(\omega)\}p_\mu(\omega)P(d\omega)\right)\right.\\ && \left.+\ln\left(\int\exp\{-\psi(\omega)\}p_\nu(\omega)P(d\omega)\right)\right].\end{array}$$

Since $\mathcal{O}$ is a good o.s., the function $\bar\psi(\omega)=\frac{1}{2}\ln(p_\nu(\omega)/p_\mu(\omega))$ belongs to $\mathcal{F}$, and

$$\begin{array}{l}\inf\limits_{\psi\in\mathcal{F}}\left[\ln\left(\int\exp\{\psi(\omega)\}p_\mu(\omega)P(d\omega)\right)+\ln\left(\int\exp\{-\psi(\omega)\}p_\nu(\omega)P(d\omega)\right)\right]\\ \quad=\inf\limits_{\delta\in\mathcal{F}}\left[\ln\left(\int\exp\{\bar\psi(\omega)+\delta(\omega)\}p_\mu(\omega)P(d\omega)\right)+\ln\left(\int\exp\{-\bar\psi(\omega)-\delta(\omega)\}p_\nu(\omega)P(d\omega)\right)\right]\\ \quad=\inf\limits_{\delta\in\mathcal{F}}\underbrace{\left[\ln\left(\int\exp\{\delta(\omega)\}\sqrt{p_\mu(\omega)p_\nu(\omega)}P(d\omega)\right)+\ln\left(\int\exp\{-\delta(\omega)\}\sqrt{p_\mu(\omega)p_\nu(\omega)}P(d\omega)\right)\right]}_{f(\delta)}.\end{array}$$

Observe that $f(\delta)$ clearly is a convex and even function of $\delta\in\mathcal{F}$; as such, it attains its minimum over $\delta\in\mathcal{F}$ when $\delta=0$. The bottom line is that

$$\inf_{\psi\in\mathcal{F}}[\Phi_{\mathcal{O}}(\psi;A_i(x))+\Phi_{\mathcal{O}}(-\psi;A_j(y))]=2\ln\left(\int\sqrt{p_{A_i(x)}(\omega)p_{A_j(y)}(\omega)}P(d\omega)\right), \tag{3.1.12}$$

and

$$\begin{array}{rcl}\Theta_{ij}(x,y) &=& \inf\limits_{\alpha>0}\alpha\left[K\ln\left(\int\sqrt{p_{A_i(x)}(\omega)p_{A_j(y)}(\omega)}P(d\omega)\right)+\ln(2I/\epsilon)\right]+\frac{1}{2}g^T[y-x]\\ &=& \left\{\begin{array}{ll}\frac{1}{2}g^T[y-x] & ,K\ln\left(\int\sqrt{p_{A_i(x)}(\omega)p_{A_j(y)}(\omega)}P(d\omega)\right)+\ln(2I/\epsilon)\geq 0,\\ -\infty, & \text{otherwise.}\end{array}\right.\end{array}$$

This combines with (3.1.11) to imply that

$$\mathrm{Opt}_{ij}(K)=\max_{x,y}\left\{\frac{1}{2}g^T[y-x]:x\in X_i,y\in X_j,\left[\int\sqrt{p_{A_i(x)}(\omega)p_{A_j(y)}(\omega)}P(d\omega)\right]^K\geq\frac{\epsilon}{2I}\right\}. \tag{3.1.13}$$

2^0. We claim that under the premise of Proposition, for all i,j, $1\le i,j\le I$, one has

$$\mathrm{Opt}_{ij}(K)\le\mathrm{Risk}^*_\epsilon(\bar K),$$

implying the validity of (3.1.9). Indeed, assume that for some pair i,j the opposite inequality holds true:

$$\mathrm{Opt}_{ij}(K)>\mathrm{Risk}^*_\epsilon(\bar K),$$

and let us lead this assumption to a contradiction. Under our assumption optimization problem in (3.1.13) has a feasible solution $(\bar x,\bar y)$ such that

$$r:=\frac{1}{2}g^T[\bar y-\bar x]>\mathrm{Risk}^*_\epsilon(\bar K), \tag{3.1.14}$$

implying, due to the origin of $\mathrm{Risk}^*_\epsilon(\bar K)$, that there exists an estimate $\widehat g(\omega^{\bar K})$ such that for $\mu=A_i(\bar x)$, $\nu=A_j(\bar y)$ it holds

$$\begin{array}{rcl}\mathrm{Prob}_{\omega^{\bar K}\sim p_\nu^{\bar K}}\left\{\widehat g(\omega^{\bar K})\le\frac{1}{2}g^T[\bar x+\bar y]\right\} &\le& \mathrm{Prob}_{\omega^{\bar K}\sim p_\nu^{\bar K}}\left\{|\widehat g(\omega^{\bar K})-g^T\bar y|\ge r\right\}\le\epsilon\\ \mathrm{Prob}_{\omega^{\bar K}\sim p_\mu^{\bar K}}\left\{\widehat g(\omega^{\bar K})\ge\frac{1}{2}g^T[\bar x+\bar y]\right\} &\le& \mathrm{Prob}_{\omega^{\bar K}\sim p_\mu^{\bar K}}\left\{|\widehat g(\omega^{\bar K})-g^T\bar x|\ge r\right\}\le\epsilon,\end{array}$$

so that we can decide on two simple hypotheses stating that observation $\omega^{\bar{K}}$ obeys distribution $p_\mu^{\bar{K}}$, resp., $p_\nu^{\bar{K}}$, with risk $\leq \epsilon$. Therefore,

$$\int \min\left[p_\mu^{\bar{K}}(\omega^{\bar{K}}), p_\nu^{\bar{K}}(\omega^{\bar{K}})\right] P^{\bar{K}}(d\omega^{\bar{K}}) \leq 2\epsilon.$$

Hence, when setting $p_\theta^{\bar{K}}(\omega^{\bar{K}}) = \prod_k p_\theta(\omega_k)$ and $P^{\bar{K}} = \underbrace{P \times ... \times P}_{\bar{K}}$, we have

$$\begin{array}{l}
\left[\int \sqrt{p_\mu(\omega)p_\nu(\omega)}P(d\omega)\right]^{\bar{K}} = \int \sqrt{p_\mu^{\bar{K}}(\omega^{\bar{K}})p_\nu^{\bar{K}}(\omega^{\bar{K}})}P^{\bar{K}}(d\omega^{\bar{K}}) \\
\quad = \ \int \sqrt{\min\left[p_\mu^{\bar{K}}(\omega^{\bar{K}}), p_\nu^{\bar{K}}(\omega^{\bar{K}})\right]}\sqrt{\max\left[p_\mu^{\bar{K}}(\omega^{\bar{K}}), p_\nu^{\bar{K}}(\omega^{\bar{K}})\right]}P^{\bar{K}}(d\omega^{\bar{K}}) \\
\quad \leq \ \left[\int \min\left[p_\mu^{\bar{K}}(\omega^{\bar{K}}), p_\nu^{\bar{K}}(\omega^{\bar{K}})\right] P^{\bar{K}}(d\omega^{\bar{K}})\right]^{1/2} \left[\int \max\left[p_\mu^{\bar{K}}(\omega^{\bar{K}}), p_\nu^{\bar{K}}(\omega^{\bar{K}})\right] P^{\bar{K}}(d\omega^{\bar{K}})\right]^{1/2} \\
\quad = \ \left[\int \min\left[p_\mu^{\bar{K}}(\omega^{\bar{K}}), p_\nu^{\bar{K}}(\omega^{\bar{K}})\right] P^{\bar{K}}(d\omega^{\bar{K}})\right]^{1/2} \\
\qquad \times \left[\int \left[p_\mu^{\bar{K}}(\omega^{\bar{K}}) + p_\nu^{\bar{K}}(\omega^{\bar{K}}) - \min\left[p_\mu^{\bar{K}}(\omega^{\bar{K}}), p_\nu^{\bar{K}}(\omega^{\bar{K}})\right]\right] P^{\bar{K}}(d\omega^{\bar{K}})\right]^{1/2} \\
\quad = \ \left[\int \min\left[p_\mu^{\bar{K}}(\omega^{\bar{K}}), p_\nu^{\bar{K}}(\omega^{\bar{K}})\right] P^{\bar{K}}(d\omega^{\bar{K}})\right]^{1/2} \left[2 - \int \min\left[p_\mu^{\bar{K}}(\omega^{\bar{K}}), p_\nu^{\bar{K}}(\omega^{\bar{K}})\right] P^{\bar{K}}(d\omega^{\bar{K}})\right]^{1/2} \\
\quad \leq \ 2\sqrt{\epsilon(1-\epsilon)}.
\end{array}$$

Consequently,

$$\left[\int \sqrt{p_\mu(\omega)p_\nu(\omega)}P(d\omega)\right]^K \leq [2\sqrt{\epsilon(1-\epsilon)}]^{K/\bar{K}} < \frac{\epsilon}{2I},$$

which is the desired contradiction (recall that $\mu = A_i(\bar{x})$, $\nu = A_j(\bar{y})$ and $(\bar{x}, \bar{y})$ is feasible for (3.1.13)).

3^0. Now let us prove that under the premise of Proposition, (3.1.10) takes place. To this end let us set

$$w_{ij}(s) = \max_{x \in X_j, y \in X_j} \left\{\frac{1}{2}g^T[y - x] : \underbrace{\bar{K}\ln\left(\int \sqrt{p_{A_i(x)}(\omega)p_{A_j(y)}(\omega)}P(d\omega)\right)}_{H(x,y)} + s \geq 0\right\}. \qquad (3.1.15)$$

As we have seen in item 1^0, see (3.1.12), one has

$$H(x, y) = \inf_{\psi \in \mathcal{F}} \frac{1}{2}\left[\Phi_{\mathcal{O}}(\psi; A_i(x)) + \Phi_{\mathcal{O}}(-\psi, A_j(y))\right],$$

that is, $H(x, y)$ is the infimum of a parametric family of concave functions of $(x, y) \in X_i \times X_j$ and as such is concave. Besides this, the optimization problem in (3.1.15) is feasible whenever $s \geq 0$, a feasible solution being $y = x = x_{ij}$. At this feasible solution we have $g^T[y - x] = 0$, implying that $w_{ij}(s) \geq 0$ for $s \geq 0$. Observe also that from concavity of $H(x, y)$ it follows that $w_{ij}(s)$ is concave on the ray $\{s \geq 0\}$. Finally, we claim that

$$w_{ij}(\bar{s}) \leq \mathrm{Risk}_\epsilon^*(\bar{K}), \ \bar{s} = -\ln(2\sqrt{\epsilon(1-\epsilon)}). \qquad (3.1.16)$$

Indeed, $w_{ij}(s)$ is nonnegative, concave and bounded (since X_i, X_j are compact) on $\mathbf{R}_+$, implying that $w_{ij}(s)$ is continuous on $\{s > 0\}$. Assuming, on the contrary to our claim, that $w_{ij}(\bar{s}) > \mathrm{Risk}_\epsilon^*(\bar{K})$, there exists $s' \in (0, \bar{s})$ such that $w_{ij}(s') > \mathrm{Risk}_\epsilon^*(\bar{K})$ and thus there exist $\bar{x} \in X_i$, $\bar{y} \in X_j$ such that $(\bar{x}, \bar{y})$ is feasible for the optimization problem specifying $w_{ij}(s')$ and (3.1.14) takes place. We have seen in item 2^0 that the latter relation implies that for $\mu = A_i(\bar{x})$, $\nu = A_j(\bar{y})$ it holds

$$\left[\int \sqrt{p_\mu(\omega)p_\nu(\omega)}P(d\omega)\right]^{\bar{K}} \leq 2\sqrt{\epsilon(1-\epsilon)},$$

that is,

$$\bar{K}\ln\left(\int\sqrt{p_\mu(\omega)p_\nu(\omega)}P(d\omega)\right)+\bar{s}\leq 0,$$

whence

$$\bar{K}\ln\left(\int\sqrt{p_\mu(\omega)p_\nu(\omega)}P(d\omega)\right)+s'<0,$$

contradicting the fact that $(\bar{x},\bar{y})$ is feasible for the optimization problem specifying $w_{ij}(s')$.

It remains to note that (3.1.16) combines with concavity of $w_{ij}(\cdot)$ and the relation $w_{ij}(0)\geq 0$ to imply that

$$w_{ij}(\ln(2I/\epsilon))\leq\vartheta w_{ij}(\bar{s})\leq\vartheta\mathrm{Risk}^*_\epsilon(\bar{K}),\;\;\vartheta=\ln(2I/\epsilon)/\bar{s}=\frac{2\ln(2I/\epsilon)}{\ln([4\epsilon(1-\epsilon)]^{-1})}.$$

Invoking (3.1.13), we conclude that

$$\mathrm{Opt}_{ij}(\bar{K})=w_{ij}(\ln(2I/\epsilon))\leq\vartheta\mathrm{Risk}^*_\epsilon(\bar{K})\,\forall i,j.$$

Finally, from (3.1.13) it immediately follows that $\mathrm{Opt}_{ij}(K)$ is nonincreasing in K (since as K grows, the feasible set of the right hand side optimization problem in (3.1.13) shrinks), that is,

$$K\geq\bar{K}\Rightarrow\mathrm{Opt}(K)\leq\mathrm{Opt}(\bar{K})=\max_{i,j}\mathrm{Opt}_{ij}(\bar{K})\leq\vartheta\mathrm{Risk}^*_\epsilon(\bar{K}),$$

and (3.1.10) follows. □

3.1.5 Illustration

We illustrate our construction on the simplest possible example – one where $X_i=\{x_i\}$ are singletons in $\mathbf{R}^n$, $i=1,...,I$, the observation scheme is Gaussian. Thus, setting $y_i=A_i(x_i)\in\mathbf{R}^m$, the observation's components ω_k, $1\leq k\leq K$, stemming from signal x_i, are drawn, independently of each other, from the normal distribution $\mathcal{N}(y_i,I_m)$. The family $\mathcal{F}$ of functions ϕ associated with Gaussian o.s. is the family of all affine functions $\phi(\omega)=\phi_0+\varphi^T\omega$ on the observation space (which at present is $\mathbf{R}^m$); we identify $\phi\in\mathcal{F}$ with the pair (ϕ_0,φ). The function $\Psi_{\mathcal{O}}$ associated with the Gaussian observation scheme with m-dimensional observations is

$$\Phi_{\mathcal{O}}(\phi;\mu)=\phi_0+\varphi^T\mu+\frac{1}{2}\varphi^T\varphi:(\mathbf{R}\times\mathbf{R}^m)\times\mathbf{R}^m\to\mathbf{R},$$

a straightforward computation shows that in the case in question, setting

$$\theta=\ln(2I/\epsilon),$$

we have

$$\begin{array}{rcl}\Psi_{i,+}(\alpha,\phi)&=&K\alpha\left[\phi_0+\varphi^Ty_i/\alpha+\frac{1}{2}\varphi^T\varphi/\alpha^2\right]+\alpha\theta-g^Tx_i\\&=&K\phi_0+K\varphi^Ty_i-g^Tx_i+\frac{K}{2\alpha}\varphi^T\varphi+\alpha\theta\\\Psi_{j,-}(\alpha,\phi)&=&-K\phi_0-K\varphi^Ty_j+g^Tx_j+\frac{K}{2\alpha}\varphi^T\varphi+\alpha\theta\\\mathrm{Opt}_{ij}&=&\inf_{\alpha>0,\phi}\frac{1}{2}\left[\Psi_{i,+}(\alpha,\phi)+\Psi_{j,-}(\alpha,\phi)\right]\\&=&\frac{1}{2}g^T[x_j-x_i]+\inf_\varphi\left[\frac{K}{2}\varphi^T[y_i-y_j]+\inf_{\alpha>0}\left[\frac{K}{2\alpha}\varphi^T\varphi+\alpha\theta\right]\right]\\&=&\frac{1}{2}g^T[x_j-x_i]+\inf_\varphi\left[\frac{K}{2}\varphi^T[y_i-y_j]+\sqrt{2K\theta}\|\varphi\|_2\right]\\&=&\left\{\begin{array}{ll}\frac{1}{2}g^T[x_j-x_i],&\|y_i-y_j\|_2\leq2\sqrt{2\theta/K}\\-\infty,&\|y_i-y_j\|_2>2\sqrt{2\theta/K}.\end{array}\right.\end{array}\qquad(3.1.17)$$

We see that we can safely set $\phi_0=0$ and that setting

$$\mathcal{I}=\{(i,j):\|y_i-y_j\|_2\leq2\sqrt{2\theta/K}\},$$

$\mathrm{Opt}_{ij}(K)$ is finite iff $(i,j)\in\mathcal{I}$ and is $-\infty$ otherwise; in both cases, the optimization problem specifying Opt_{ij} has no optimal solution. Indeed, this clearly is the case when $(i,j)\not\in\mathcal{I}$; when $(i,j)\in\mathcal{I}$, a minimizing sequence is, e.g., $\phi_0\equiv 0$, $\varphi\equiv 0$, $\alpha_i\to 0$, but its limits is not in the minimization domain (on this domain, α should be positive). Coping with this case was exactly the reason why in our construction we required from ϕ_{ij},α_{ij} to be feasible, and not necessary optimal, solutions to the optimization problems in question). In the illustration under consideration, the simplest way to overcome the difficulty is to restrict the optimization domain $\mathcal{F}^+$ in (3.1.2) with its compact subset $\{\alpha\geq 1/R,\phi_0=0,\|\varphi\|_2\leq R\}$ with large R, like $R=10^{10}$ or 10^{20}. With this approach, we specify the entities participating in (3.1.3) as

$$\begin{array}{rcl}
\phi_{ij}(\omega) &=& \varphi_{ij}^T\omega,\ \varphi_{ij}=\left\{\begin{array}{ll}0, & (i,j)\in\mathcal{I}\\ -R[y_i-y_j]/\|y_i-y_j\|_2, & (i,j)\not\in\mathcal{I}\end{array}\right.\\
\alpha_{ij} &=& \left\{\begin{array}{ll}1/R, & (i,j)\in\mathcal{I}\\ \sqrt{\frac{K}{2\theta}}R, & (i,j)\not\in\mathcal{I}\end{array}\right.
\end{array}\tag{3.1.18}$$

resulting in

$$\begin{array}{rcl}
\varkappa_{ij} &=& \frac{1}{2}\left[\Psi_{j,-}(\alpha_{ij},\phi_{ij})-\Psi_{i,+}(\alpha_{ij},\phi_{ij})\right]\\
&=& \frac{1}{2}\left[-K\varphi_{ij}^Ty_j+g^Tx_j+\frac{K}{2\alpha_{ij}}\varphi_{ij}^T\varphi_{ij}+\alpha_{ij}\theta-K\varphi_{ij}^Ty_i+g^Tx_i-\frac{K}{2\alpha_{ij}}\varphi_{ij}^T\varphi_{ij}-\alpha_{ij}\theta\right]\\
&=& \frac{1}{2}g^T[x_i+x_j]-\frac{K}{2}\varphi_{ij}^T[y_i+y_j]\\
\rho_{ij} &=& \frac{1}{2}\left[\Psi_{i,+}(\alpha_{ij},\phi_{ij})+\Psi_{j,-}(\alpha_{ij},\phi_{ij})\right]\\
&=& \frac{1}{2}\left[K\varphi_{ij}^Ty_i-g^Tx_i+\frac{K}{2\alpha_{ij}}\varphi_{ij}^T\varphi_{ij}+\alpha_{ij}\theta-K\varphi_{ij}^Ty_j+g^Tx_j+\frac{K}{2\alpha_{ij}}\varphi_{ij}^T\varphi_{ij}+\alpha_{ij}\theta\right]\\
&=& \frac{K}{2\alpha_{ij}}\varphi_{ij}^T\phi_{ij}+\alpha_{ij}\theta+\frac{1}{2}g^T[x_j-x_i]+\frac{K}{2}\varphi_{ij}^T[y_i-y_j]\\
&=& \left\{\begin{array}{ll}\frac{1}{2}g^T[x_j-x_i]+R^{-1}\theta, & (i,j)\in\mathcal{I}\\ \frac{1}{2}g^T[x_j-x_i]+[\sqrt{2K\theta}-\frac{K}{2}\|y_i-y_j\|_2]R, & (i,j)\not\in\mathcal{I}\end{array}\right.
\end{array}\tag{3.1.19}$$

In the numerical experiments we are about to report we used $n=20$, $m=10$, and $I=100$, with x_i, $i\leq I$, drawn independently of each other from $\mathcal{N}(0,I_n)$, and $y_i=Ax_i$ with randomly generated matrix A (specifically, matrix with independent $\mathcal{N}(0,1)$ entries normalized to have unit spectral norm), and used $R=10^{20}$; the linear form to be recovered was just the first coordinate of x. The results of typical experiment are as follows:

K	$\max_{i,j}\rho_{ij}$	Empirical recovery error [mean/median/max]
2	2.541	0.9243/0.8292/2.541
4	2541	0.9859/0.9066/2.541
8	2.541	0.8057/0.7316/2.541
16	2.541	0.6807/0.6567/2.115
32	1.758	0.3630/0.2845/1.758
64	0.954	0.0860/0.0000/0.954
128	0.000	0.0000/0.0000/0.000
256	0.000	0.0000/0.0000/0.000

For every K, the empirical recovery errors shown in the table stem from 20 experiments, with the signal underlying an experiment selected at random among $x_1,...,x_{100}$.

3.2 Estimating N-convex functions on unions of convex sets

In this Section, we apply our testing machinery to the estimation problem as follows.

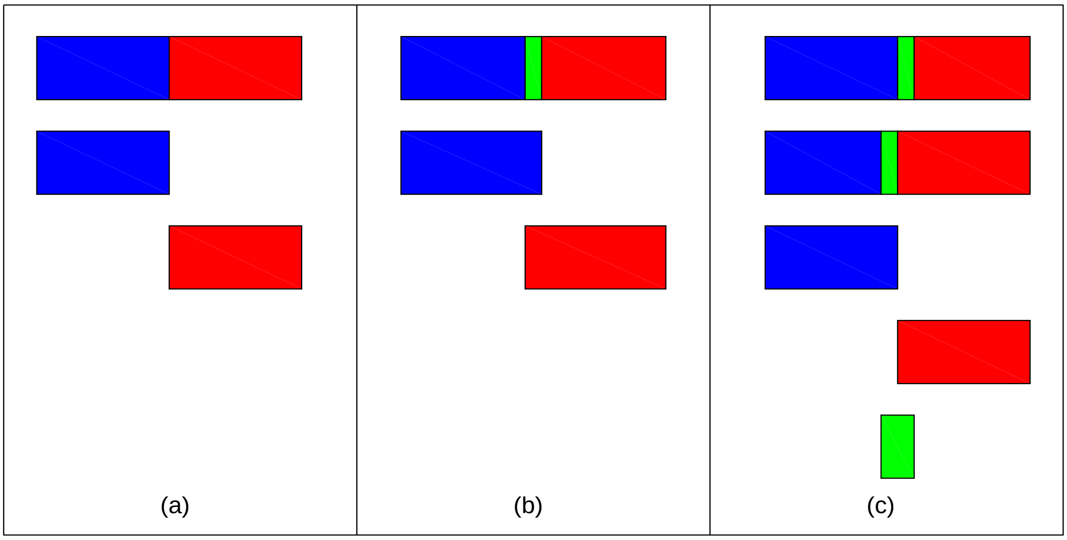


Figure 3.1: Bisection via Hypothesis Testing

Given are:

- a simple o.s. $\mathcal{O} = (\Omega, \Pi; \{p_\mu : \mu \in \mathcal{M}\}; \mathcal{F})$,
- a *signal space* $X \subset \mathbf{R}^n$ along with affine *encoding* $x \mapsto A(x) : X \to \mathcal{M}$,
- a real-valued function f on X.

Given observation $\omega \sim p_{A(x_*)}$ stemming from unknown signal x_* known to belong to X, we want to recover $f(x_*)$.

Our approach imposes severe restrictions on f (satisfied, e.g., when f is linear, or linear-fractional, or is the maximum of several linear functions); as a compensation, we allow for rather "complex" X – finite unions of convex sets.

3.2.1 Outline

The approach we intend to develop is, in nutshell, extremely simple; its formal description, however, turns to be lengthy and obscures, to some extent, the simple ideas underlying the construction. By this reason, it makes sense to start with informal outline of the strategy underlying the forthcoming developments. Consider the situation where the signal space X is the 2D rectangle depicted on the top of Figure 3.1.(a), and let the function to be recovered be $f(u) = u_1$. Thus, "the nature" has somehow selected x in the rectangle, and we observe, say, Gaussian random variable with the mean $A(x)$ and known covariance matrix, where $A(\cdot)$ is a given affine mapping. Note that hypotheses $f(x) \geq b$ and $f(x) \leq a$ translate into convex hypotheses on the expectation of the observed Gaussian r.v., so that we can use out hypothesis testing machinery to decide on hypotheses of this type and to localize $f(x)$ in a (hopefully, small) segment by a Bisection-type process. Before describing the process, let us make a terminological agreement. In the sequel we shall use pairwise hypothesis testing in the situation where it may happen the *neither one* of the hypotheses H_1, H_2 we are deciding upon is true. In this case, we will say that the outcome of a test is correct, if the rejected hypothesis indeed is wrong (the accepted hypothesis can be wrong as well, but the latter can happen only in the case when both our hypotheses are wrong).

This is how our Bisection could look like.

1. Were we able to decide reliably on the Blue and Red hypotheses on Figure 3.1.(a), that is, to understand via observations whether x belongs to the left or to the right half of the original rectangle, our course of actions would be clear: depending on this decision, we would replace our original rectangle with a smaller rectangle localizing x, as shown on Figure 3.1.(a), and then iterate this process. The difficulty, of course, is that our Red and Blue hypotheses intersect, so that is impossible to decide on them reliably.

2. In order to make Red and Blue hypotheses distinguishable from each other, we could act as shown on Figure 3.1.(b), by shrinking a little bit the blue and the red rectangles and inserting between the resulting rectangles the green "no-man land." Assuming that the width of the green rectangle allows to decide reliably on our new Blue and Red hypotheses and utilizing available observation, we can localize x either in the blue, or in the red rectangles as shown on Figure 3.1.(b). Specifically, assume that our "Red vs. Blue" test rejected correctly the red hypothesis. Then x can be located either in blue, or in green rectangles shown on the top of the figure, and thus x is in the new blue localizer which is the union of the blue and the green original rectangles. Similarly, if our test rejects correctly the blue hypothesis, then we can take, as the new localizer of x, the union of the original red and green rectangles, as shown on Figure 3.1.(b). Note that our localization is as reliable as our test is, and that it reduces the width of localizer by factor close to 2, provided the width of the green rectangle is small as compared to the width of the original "tricolor" localizer of x. We can iterate this process, with the new – smaller – localizer in the role of the old till arriving at a localizer so narrow that "no-man land" part of it (this part cannot be too narrow, since it should allow for reliable decision on the current blue and red hypotheses) becomes too large to allow for significant progress in localizer's width.

The bottleneck of this approach is where to take observations to be used in our subsequent tests. In principle, we could use in all of them the initial observation; the difficulty with this approach is, that the hypotheses we need to decide upon depend on the observations (e.g., when x belongs to the green part of the "tricolor" rectangle on Figure 3.1, deciding on Blue vs. Red can, depending on observation, lead to accepting either red or blue hypothesis, thus leading to different updated localizers), and we arrive at the situation when we should decide on *random* hypotheses via observation statistically depending on these hypotheses – a mess we have no idea how to analyze. To circumvent this difficulty, we could use in every one of the tests its own observation drawn, independently of the previous observations, from the distribution $p_{A(x)}$. However, to do this, we need repeated observations to be allowed, and the number of observations we will use will be proportional to the number of tests we intend to run.

3. Finally, there is a theoretically sound way to implement Bisection based on a *single* observation, and this is what we intend to do. The policy we use now is as follows: given current localizer for x (at the first step - our initial rectangle), we consider two "tricolor" partitions of it depicted at the top of Figure 3.1.(c). In the first partition, the blue rectangle is the left half of the original rectangle, in the second the red rectangle is the right half of the original rectangle. We then run *two* Blue vs. Red tests, the first on the pair of Blue and Red hypotheses stemming from the first partition, and the second on the pair of Blue and Red hypotheses stemming from the second partition. Assuming that in both tests the rejected hypotheses indeed were wrong, the results of these tests allow us to make conclusions as follows:

- when both tests reject red hypotheses from the corresponding pairs, x is located in the left half of the initial rectangle (since otherwise in the second test the rejected hypothesis were in fact true, contradicting to the assumption that both tests make no wrong rejections);
- when both tests reject blue hypotheses from the corresponding pairs, x is located in the right half of the original rectangle (by the same reasons as in the previous case);

- when the tests "disagree," rejecting hypotheses of different colors, x is located in the union of the two green rectangles we deal with. Indeed, otherwise x should be either in the blue rectangles of both our "tricolors," or in the red rectangles of both of them. Since we have assumed that in both tests no wrong rejections took place, in the first case both tests must reject red hypotheses, and in the second both should reject blue ones, while in fact neither one of these two options took place.

Now, in the first two cases we can safely say to which one of "halves" – left or right – of the initial rectangle x belongs, and take this half as our new localizer. In the third case, we take as a new localizer for x the green rectangle shown on the bottom of Figure 3.1 *and terminate our estimation process* – the new localizer already is narrow! Now, in the proposed algorithm, unless we terminate at the very first step, we carry out the second step exactly in the same fashion as the first one, with the localizer of x yielded by the first step in the role of the initial localizer, then carry out, in the same fashion, the third step, etc., until termination either due to running into a disagreement, or due to reaching a prescribed number of steps. Upon termination, we return the last localizer for x which we have built, and claim that $f(x) = x_1$ belongs to the projection of this localizer onto the x_1-axis. *In all tests from the above process, we use the same observation.* Note that in our current situation, in contrast to the one we have discussed earlier, re-utilizing a single observation creates no difficulties, since *with no wrong rejections in the pairwise tests we use, the pairs of hypotheses participating in the tests are not random at all – they are uniquely defined by $f(x) = x_1$!* Indeed, with no wrong rejections, prior to termination everything is *as if* we were running perfect Bisection, that is, were updating subsequent rectangles Δ_t containing x according to the rules

- Δ_1 is a given in advance rectangle containing x,
- Δ_{t+1} is either the left, or the right half of Δ_t, depending on which one of these two halves contains x.

Thus, given x and with no wrong rejections, the situation is as if a single observation were used in a number L of tests "in parallel" rather than sequentially, and the only elaboration caused by the sequential nature of our process is in "risk accumulation" – we want the probability of error *in one or more of our L tests* to be less than the desired risk ϵ of wrong "bracketing" of $f(x)$, implying, for absence of something better, that the risks of the individual tests should be at most ϵ/L. These risks, in turn, define the allowed width of "no man land" zones, and thus – the accuracy to which $f(x)$ can be estimated. It should be noted that the number L of steps of Bisection always is a moderate integer (since otherwise the width of "no-man land" zone, which at the concluding Bisection steps is of order of 2^{-L}, will be by far too small to allow for deciding on the concluding pairs of our hypotheses with risk ϵ/L, at least when our observations possess non-negligible volatility). As a result, "the price" of Bisection turns out to be low as compared to the case where every test uses its own observation.

We have outlined the strategy we are about to implement. From the outline it is clear that all what matters is our ability to decide on the pairs of hypotheses $\{x \in X : f(x) \leq a\}$ and $\{x \in X : f(X) \geq b\}$, with a and b given, via observation drawn from $p_{A(x)}$. In our outline, these were convex hypotheses in Gaussian o.s., and in this case we can use detector-based pairwise tests yielded by Theorem 2.4.2. Applying the machinery developed in Section 2.5.1, we could also handle the case when the sets $\{x \in X : f(x) \leq a\}$ and $\{x \in X : f(X) \geq b\}$ are unions of a moderate number of convex sets (e.g., f is affine, and X is the union of a number of convex sets), the o.s. in question still being simple, and this is the situation we intend to consider.

3.2.2 Estimating N-convex functions: problem's setting

In the rest of this Section, we consider the situation as follows. Given are:

1. simple o.s. $\mathcal{O} = ((\Omega, P), \{p_\mu(\cdot) : \mu \in \mathcal{M}\}, \mathcal{F})$,
2. convex compact set $\mathcal{X} \subset \mathbf{R}^n$ along with a collection of I convex compact sets $X_i \subset \mathcal{X}$,
3. affine "encoding" $x \mapsto A(x) : \mathcal{X} \to \mathcal{M}$,
4. a continuous function $f(x) : \mathcal{X} \to \mathbf{R}$ which is N-*convex*, meaning that for every $a \in \mathbf{R}$ the sets $\mathcal{X}^{a,\geq} = \{x \in \mathcal{X} : f(x) \geq a\}$ and $\mathcal{X}^{a,\leq} = \{x \in \mathcal{X} : f(x) \leq a\}$ can be represented as the unions of at most N closed convex sets $\mathcal{X}_\nu^{a,\geq}$, $\mathcal{X}_\nu^{a,\leq}$:

$$\mathcal{X}^{a,\geq} = \bigcup_{\nu=1}^N \mathcal{X}_\nu^{a,\geq},\ \mathcal{X}^{a,\leq} = \bigcup_{\nu=1}^N \mathcal{X}_\nu^{a,\leq}. \tag{3.2.1}$$

For some *unknown* x known to belong to $X = \bigcup\limits_{i=1}^I X_i$, we have at our disposal observation $\omega^K = (\omega_1, ..., \omega_K)$ with i.i.d. $\omega_t \sim p_{A(x)}(\cdot)$, and our goal is to estimate from this observation the quantity $f(x)$.

Given tolerances $\rho > 0$, $\epsilon \in (0,1)$, let us call a candidate estimate $\widehat{f}(\omega^K)$ (ρ, ϵ)-*reliable*, if for every $x \in X$, with the $p_{A(x)}$-probability at least $1 - \epsilon$ it holds $|\widehat{f}(\omega^K) - f(x)| \leq \rho$ or, which is the same, if

$$\forall (x \in X) : \mathrm{Prob}_{\omega^K \sim p_{A(x)} \times ... \times p_{A(x)}} \left\{ |\widehat{f}(\omega^K) - f(x)| > \rho \right\} \leq \epsilon. \tag{3.2.2}$$

3.2.2.1 Examples of N-convex functions

Example 3.1 [Minima and Maxima of linear-fractional functions] *Every function which can be obtained from linear-fractional functions $\frac{g_\nu(x)}{h_\nu(x)}$ (g_ν, h_ν are affine functions on $\mathcal{X}$ and h_ν are positive on $\mathcal{X}$) by taking maxima and minima is N-convex for appropriately selected N due to the following immediate observations:*

- *linear-fractional function $\frac{g(x)}{h(x)}$ with positive on $\mathcal{X}$ denominator is 1-convex on $\mathcal{X}$;*
- *if $f(x)$ is N-convex, so is $-f(x)$;*
- *if $f_i(x)$ is N_i-convex, $i = 1, 2, ..., I$, then $f(x) = \max_i f_i(x)$ is ($N = \max[\prod_i N_i, \sum_i N_i]$)-convex, due to*

$$\{x \in \mathcal{X} : f(x) \leq a\} = \bigcap_{i=1}^I \{x : f_i(x) \leq a\},\ \{x \in \mathcal{X} : f(x) \geq a\} = \bigcup_{i=1}^I \{x : f_i(x) \geq a\}.$$

The first right hand side set is the intersection of I unions of convex sets with N_i components in i-th union, and thus is the union of $\prod_i N_i$ convex sets; the second right hand side set is the union of I unions, N_i components in i-th of them, of convex sets, and thus is the union of $\sum_i N_i$ convex sets.

Example 3.2 [Conditional quantile]. *Let $S = \{s_1 < s_2 < ... < s_N\} \subset \mathbf{R}$ and T be a finite set, and let $\mathcal{X}$ be a convex compact set in the space of nonvanishing probability distributions on $S \times T$. Given $\tau \in T$, consider the conditional, by the condition $t = \tau$, distribution $q_\tau[p]$ of $s \in S$ induced by a distribution $p(\cdot, \cdot) \in \mathcal{X}$:*

$$(q_\tau[p])_\mu = \frac{p(\mu, \tau)}{\sum_{\nu=1}^N p(\nu, \tau)}.$$

For a nonvanishing probability distribution q on S and $\alpha \in (0,1)$, let $\chi_\alpha(q)$ be the regularized α-quantile of q defined as follows: we pass from q to the distribution on $[s_1, s_N]$ by spreading uniformly the mass q_ν, $1 \leq \nu < N$, over $[s_\nu, s_{\nu+1}]$, and assigning mass q_N to the point s_N; $\chi_\alpha(q)$ is the usual α-quantile of the resulting distribution $\bar{q}$:

$$\chi_\alpha(q) = \min\{s \in [s_1, s_N] : \bar{q}\{(s, s_N]\} \leq \alpha\}$$

The function $\chi_\alpha(q_\mathcal{T}[p]) : \mathcal{X} \to \mathbf{R}$ turns out to be 2-convex, see Section 3.2.10.

3.2.3 Bisection Estimate: Construction

While the construction to be presented admits numerous refinements, we focus here on its simplest version as follows.

3.2.3.1 Preliminaries

Upper an lower feasibility/infeasibility, sets $Z_i^{a,\geq}$ and $Z_i^{a,\leq}$. Let a be a real. We associate with a the collection of *upper a-sets* defined as follows: we look at the sets $X_i \cap \mathcal{X}_\nu^{a,\geq}$, $1 \leq i \leq I$, $1 \leq \nu \leq N$, and arrange the nonempty sets from this family into a sequence $Z_i^{a,\geq}$, $1 \leq i \leq I_{a,\geq}$, where $I_{a,\geq} = 0$ if all sets in the family are empty; in the latter case, we call *a upper-infeasible*, otherwise *upper-feasible*. Similarly, we associate with a the collection of *lower a-sets* $Z_i^{a,\leq}$, $1 \leq i \leq I_{a,\leq}$ by arranging into a sequence all nonempty sets from the family $X_i \cap \mathcal{X}_\nu^{a,\leq}$, and call a lower-feasible or lower-infeasible depending on whether $I_{a,\leq}$ is positive or zero. Note that upper and lower a-sets are nonempty convex compact sets, and

$$X^{a,\geq} := \{x \in X : f(x) \geq a\} = \bigcup_{1 \leq i \leq I_{a,\geq}} Z_i^{a,\geq},\ X^{a,\leq} := \{x \in X : f(x) \leq a\} = \bigcup_{1 \leq i \leq I_{a,\leq}} Z_i^{a,\leq}. \tag{3.2.3}$$

Right-side tests. Given a segment $\Delta = [a, b]$ of positive length with lower-feasible a, we associate with this segment *right-side test* – a function $\mathcal{T}_{\Delta,\mathrm{r}}^K(\omega^K)$ taking values `red` and `blue`, and risk $\sigma_{\Delta,\mathrm{r}} \geq 0$ – as follows:

1. if b is upper-infeasible, $\mathcal{T}_{\Delta,\mathrm{r}}^K(\cdot) \equiv$ `blue` and $\sigma_{\Delta,\mathrm{r}} = 0$.

2. if b is upper-feasible, the collections $\{A(Z_i^{b,\geq})\}_{i \leq I_{b,\geq}}$ ("red sets"), $\{A(Z_j^{a,\leq})\}_{j \leq I_{a,\leq}}$ ("blue sets"), are nonempty, and the test is given by the construction from Section 2.5.1 *as applied to these sets and the stationary K-repeated version of $\mathcal{O}$ in the role of $\mathcal{O}$*, specifically,

   - for $1 \leq i \leq I_{b,\geq}$, $1 \leq j \leq I_{a,\leq}$, we build the detectors $\phi_{ij\Delta}^K(\omega^K) = \sum_{t=1}^K \phi_{ij\Delta}(\omega_t)$, with $\phi_{ij\Delta}(\omega)$ given by

$$\begin{array}{rcl} (r_{ij\Delta}, s_{ij\Delta}) & \in & \mathrm{Argmin}_{r \in Z_i^{b,\geq}, s \in Z_j^{a,\leq}} \ln\left(\int_\Omega \sqrt{p_{A(r)}(\omega) p_{A(s)}(\omega)} \Pi(d\omega)\right), \\ \phi_{ij\Delta}(\omega) & = & \frac{1}{2}\ln\left(p_{A(r_{ij\Delta})}(\omega)/p_{A(s_{ij\Delta})}(\omega)\right) \end{array} \tag{3.2.4}$$

   set

$$\epsilon_{ij\Delta} = \int_\Omega \sqrt{p_{A(r_{ij\Delta})}(\omega) p_{A(s_{ij\Delta})}(\omega)} \Pi(d\omega) \tag{3.2.5}$$

   and build the $I_{b,\geq} \times I_{a,\leq}$ matrix $E_{\Delta,\mathrm{r}} = [\epsilon_{ij\Delta}^K]_{\substack{1 \leq i \leq I_{b,\geq} \\ 1 \leq j \leq I_{a,\leq}}}$;

- $\sigma_{\Delta,\mathrm{r}}$ is defined as the spectral norm of $E_{\Delta,\mathrm{r}}$. We compute the Perron-Frobenius eigenvector $[g^{\Delta,\mathrm{r}}; h^{\Delta,\mathrm{r}}]$ of the matrix $\left[\begin{array}{c|c} & E_{\Delta,\mathrm{r}} \\ \hline E^T_{\Delta,\mathrm{r}} & \end{array}\right]$, so that (see Section 2.5.1.2)

$$g^{\Delta,\mathrm{r}} > 0,\, h^{\Delta,\mathrm{r}} > 0, \sigma_{\Delta,\mathrm{r}} g^{\Delta,\mathrm{r}} = E_{\Delta,\mathrm{r}} h^{\Delta,\mathrm{r}},\ \sigma_{\Delta,\mathrm{r}} h^{\Delta,\mathrm{r}} = E^T_{\Delta,\mathrm{r}} g^{\Delta,\mathrm{r}}.$$

Finally, we define the matrix-valued function

$$D_{\Delta,\mathrm{r}}(\omega^K) = [\phi^K_{ij\Delta}(\omega^K) + \ln(h^{\Delta,\mathrm{r}}_j) - \ln(g^{\Delta,\mathrm{r}}_i)]_{\substack{1\le i\le I_{b,\ge} \\ 1\le j\le I_{a,\le}}}.$$

Test $\mathcal{T}^K_{\Delta,\mathrm{r}}(\omega^K)$ takes value `red` iff the matrix $D_{\Delta,\mathrm{r}}(\omega^K)$ has a nonnegative row, and takes value `blue` otherwise.

Given $\delta > 0$, $\varkappa > 0$, we call segment $\Delta = [a, b]$ *δ-good (right)*, if a is lower-feasible, $b > a$, and $\sigma_{\Delta,\mathrm{r}} \le \delta$. We call a δ-good (right) segment $\Delta = [a, b]$ *$\varkappa$-maximal*, if the segment $[a, b - \varkappa]$ is not δ-good (right).

Left-side tests. The "mirror" version of the above is as follows. Given a segment $\Delta = [a, b]$ of positive length with upper-feasible b, we associate with this segment *left-side test* – a function $\mathcal{T}^K_{\Delta,\mathrm{l}}(\omega^K)$ taking values `red` and `blue`, and risk $\sigma_{\Delta,\mathrm{l}} \ge 0$ – as follows:

1. if a is lower-infeasible, $\mathcal{T}^K_{\Delta,\mathrm{l}}(\cdot) \equiv$ `red` and $\sigma_{\Delta,\mathrm{l}} = 0$.
2. if a is lower-feasible, we set $\mathcal{T}^K_{\Delta,\mathrm{l}} \equiv \mathcal{T}^K_{\Delta,\mathrm{r}}$, $\sigma_{\Delta,\mathrm{l}} = \sigma_{\Delta,\mathrm{r}}$.

Given $\delta > 0$, $\varkappa > 0$, we call segment $\Delta = [a, b]$ *δ-good (left)*, if b is upper-feasible, $b > a$, and $\sigma_{\Delta,\mathrm{l}} \le \delta$. We call a δ-good (left) segment $\Delta = [a, b]$ *$\varkappa$-maximal*, if the segment $[a + \varkappa, b]$ is not δ-good (left).

Explanation: When $a < b$ and a is lower-feasible, b is upper-feasible, so that the sets

$$X^{a,\le} = \{x \in X : f(x) \le a\},\ X^{b,\ge} = \{x \in X : f(x) \ge b\}$$

are nonempty, the right-side and the left-side tests $\mathcal{T}^K_{\Delta,\mathrm{l}}$, $\mathcal{T}^K_{\Delta,\mathrm{r}}$ are identical to each other and coincide with the minimal risk test, built as explained in Section 2.5.1, deciding, via stationary K-repeated observations, on the "color" of the distribution $p_{A(x)}$ underlying observations – whether this color is blue ("blue" hypothesis stating that $x \in X$ and $f(x) \le a$, whence $A(x) \in \bigcup\limits_{1\le i\le I_{a,\le}} A(Z^{a,\le}_i)$), or red ("red" hypothesis, stating that $x \in X$ and $f(x) \ge b$, whence $A(x) \in \bigcup\limits_{1\le i\le I_{b,\ge}} A(Z^{b,\ge}_i)$). When a is lower-feasible and b is *not* upper-feasible, the red one of the above two hypotheses is empty, and the left-side test associated with $[a, b]$, naturally, always accepts the blue hypothesis; similarly, when a is lower-infeasible and b is upper-feasible, the right-side test associated with $[a, b]$ always accepts the red hypothesis.

A segment $[a, b]$ with $a < b$ is δ-good (left), if the corresponding to the segment "red" hypothesis is nonempty, and the left-hand side test $\mathcal{T}^K_{\Delta\ell}$ associated with $[a, b]$ decides on the "red" and the "blue" hypotheses with risk $\le \delta$, and similarly for δ-good (right) segment $[a, b]$.

3.2.4 Building the Bisection estimate

3.2.4.1 Control parameters

The control parameters of our would-be Bisection estimate are

1. positive integer L – the maximum allowed number of bisection steps,
2. tolerances $\delta \in (0, 1)$ and $\varkappa > 0$.

3.2.4.2 Bisection estimate: construction

The estimate of $f(x)$ (x is the signal underlying our observations: $\omega_t \sim p_{A(x)}$) is given by the following recurrence run on the observation $\bar{\omega}^K = (\bar{\omega}_1, ..., \bar{\omega}_K)$ which we have at our disposal:

1. **Initialization.** We find a valid upper bound b_0 on $\max_{u \in X} f(u)$ and valid lower bound a_0 on $\min_{u \in X} f(u)$ and set $\Delta_0 = [a_0, b_0]$. We assume w.l.o.g. that $a_0 < b_0$, otherwise the estimation is trivial.
   Note: $f(a) \in \Delta_0$.

2. **Bisection Step ℓ,** $1 \le \ell \le L$. Given *localizer* $\Delta_{\ell-1} = [a_{\ell-1}, b_{\ell-1}]$ with $a_{\ell-1} < b_{\ell-1}$, we act as follows:

   (a) We set $c_\ell = \frac{1}{2}[a_{\ell-1} + b_{\ell-1}]$. If c_ℓ is not upper-feasible, we set $\Delta_\ell = [a_{\ell-1}, c_\ell]$ and pass to 2e, and if c_ℓ is not lower-feasible, we set $\Delta_\ell = [c_\ell, b_{\ell-1}]$ and pass to 2e.
   Note: In the latter two cases, $\Delta_\ell \backslash \Delta_{\ell-1}$ does not intersect with $f(X)$; in particular, in these cases $f(x) \in \Delta_\ell$ provided that $f(x) \in \Delta_{\ell-1}$.

   (b) When c_ℓ is both upper- and lower-feasible, we check whether the segment $[c_\ell, b_{\ell-1}]$ is δ-good (right). If it is not the case, we terminate and claim that $f(x) \in \bar{\Delta} := \Delta_{\ell-1}$, otherwise find v_ℓ, $c_\ell < v_\ell \le b_{\ell-1}$, such that the segment $\Delta_\ell^{,\mathrm{rg}} = [c_\ell, v_\ell]$ is δ-good (right) $\varkappa$-maximal.
   Note: In terms of the outline of our strategy presented in Section 3.2.1, termination when the segment $[c_\ell, b_{\ell-1}]$ is not δ-good (right) corresponds to the case when the current localizer is too small to allow for "no-man land" wide enough to ensure low-risk decision on the blue and the red hypotheses.
   Note: To find v_ℓ, we look one by one at the candidates with $v_\ell^k = b_{\ell-1} - k\varkappa$, $k = 0, 1, ...$ until arriving for the first time at segment $[c_\ell, v_\ell^k]$ which is not δ-good (right), and take, as v_ℓ, the quantity v^{k-1} (when v_ℓ indeed is sought, we clearly have $k \ge 1$, so that our recipe for building v_ℓ is well-defined and clearly meets the above requirements on v_ℓ).

   (c) Similarly, we check whether the segment $[a_{\ell-1}, c_\ell]$ is δ-good (left). If it is not the case, we terminate and claim that $f(x) \in \bar{\Delta} := \Delta_{\ell-1}$, otherwise find u_ℓ, $a_{\ell-1} \le u_\ell < c_\ell$, such that the segment $\Delta_{\ell,\mathrm{lf}} = [u_\ell, c_\ell]$ is δ-good (left) $\varkappa$-maximal.
   Note: The rules for building u_ℓ are completely similar to those for v_ℓ.

   (d) We compute $\mathcal{T}^K_{\Delta_{\ell,\mathrm{rg}},\mathrm{r}}(\bar{\omega}^K)$ and $\mathcal{T}^K_{\Delta_{\ell,\mathrm{lf}},\mathrm{l}}(\bar{\omega}^K)$. If $\mathcal{T}^K_{\Delta_{\ell,\mathrm{rg}},\mathrm{r}}(\bar{\omega}^K) = \mathcal{T}^K_{\Delta_{\ell,\mathrm{lf}},\mathrm{l}}(\bar{\omega}^K)$ ("consensus"), we set
   $$\Delta_\ell = [a_\ell, b_\ell] = \left\{ \begin{array}{ll} [c_\ell, b_{\ell-1}], & \mathcal{T}^K_{\Delta_{\ell,\mathrm{rg}},\mathrm{r}}(\bar{\omega}^K) = \texttt{red}, \\ [a_{\ell-1}, c_\ell], & \mathcal{T}^K_{\Delta_{\ell,\mathrm{rg}},\mathrm{r}}(\bar{\omega}^K) = \texttt{blue} \end{array} \right. \tag{3.2.6}$$
   and pass to 2e. Otherwise ("disagreement") we terminate and claim that $f(x) \in \bar{\Delta} = [u_\ell, v_\ell]$.

   (e) When arriving at this rule, Δ_ℓ is already built. When $\ell < L$, we pass to step $\ell + 1$, otherwise we terminate with the claim that $f(x) \in \bar{\Delta} := \Delta_L$.

3. **Output of the estimation procedure** is the segment $\bar{\Delta}$ built upon termination and claimed to contain $f(x)$, see rules 2b–2e; the midpoint of this segment is the estimate of $f(x)$ yielded by our procedure.

3.2.5 Bisection estimate: Main result

Our main result on Bisection is as follows:

Proposition 3.2.1 *Consider the situation described in the beginning of Section 3.2.2, and let $\epsilon \in (0,1/2)$ be given. Then*

(i) [reliability of Bisection] *For every positive integer L and every $\kappa > 0$, Bisection with control parameters L,*

$$\delta = \frac{\epsilon}{2L},$$

κ is $(1-\epsilon)$-reliable: for every $x \in X$, the $p_{A(x)}$-probability of the event

$$f(x) \in \bar{\Delta}$$

($\bar{\Delta}$ is the output of Bisection as defined above) is at least $1-\epsilon$.

(ii) [near-optimality] *Let $\rho > 0$ and positive integer $\bar{K}$ be such that in the nature there exists a (ρ,ϵ)-reliable estimate $\widehat{f}(\cdot)$ of $f(x)$, $x \in X := \bigcup_{i \leq I} X_i$, via stationary $\bar{K}$-repeated observation $\omega^{\bar{K}}$ with $\omega_k \sim p_{A(x)}$, $1 \leq k \leq \bar{K}$. Given $\widehat{\rho} > 2\rho$, the Bisection estimate utilizing stationary K-repeated observations, with*

$$K = \rfloor \frac{2\ln(2LNI/\epsilon)}{\ln(1/\epsilon) - \ln(4(1-\epsilon))} \bar{K} \lfloor, \tag{3.2.7}$$

the control parameters of the estimate being

$$L = \rfloor \log_2\left(\frac{b_0 - a_0}{2\widehat{\rho}}\right) \lfloor, \;\; \delta = \frac{\epsilon}{2L}, \;\; \varkappa = \widehat{\rho} - 2\rho, \tag{3.2.8}$$

is $(\widehat{\rho},\epsilon)$-reliable. Not that K is only "slightly larger" than $\bar{K}$.

For proof, see Section 3.2.8.

Note that the running time K of Bisection estimate as given by (3.2.7) is just by (at most) logarithmic in N, I, L, $1/\epsilon$ factor larger than $\bar{K}$; note also that L is just logarithmic in $1/\rho$. Assume, e.g., that for some $\gamma > 0$ "in the nature" there exist $(\epsilon^\gamma, \epsilon)$ reliable estimates, parameterized by $\epsilon \in (0,1/2)$, with $\bar{K} = \bar{K}(\epsilon)$. Bisection with the volume of observation and control parameters given by (3.2.7) (3.2.8), where $\bar{\rho} = 3\rho = 3\epsilon^\gamma$ and $\bar{K} = \bar{K}(\epsilon)$, is $(3\epsilon^\gamma, \epsilon)$-reliable and requires $K = K(\epsilon)$-repeated observations with $\overline{\lim}_{\epsilon \to +0} K(\epsilon)/\bar{K}(\epsilon) \leq 2$.

3.2.6 Illustration

To illustrate bisection-based estimation of N-convex functional, consider the situation as follows[1]. There are M devices ("receivers") recording a signal u known to belong to a given convex compact and nonempty set $U \subset \mathbf{R}^n$; the output of i-th receiver is the vector

$$y_i = A_i u + \sigma\xi \in \mathbf{R}^m \qquad [\xi \sim \mathcal{N}(0, I_m)]$$

where A_i are given $m \times n$ matrices; you may think about M allowed positions of a single receiver, and on y_i – as on the output of receiver when the latter is in position i. Our observation ω is one of the vectors y_i, $1 \leq i \leq M$ *with unknown to us index i* ("we observe a noisy record of signal, but do not know the position in which this record was taken"). Given ω, we want to recover a given linear function $g(x) = e^T u$ of the signal.

The problem can be modeled as follows. Consider the sets

$$X_i = \{x = [x^1; ...; x^M] \in \mathbf{R}^{Mn} = \underbrace{\mathbf{R}^n \times ... \times \mathbf{R}^n}_{M} : x^j = 0, j \neq i; x^i \in U\}$$

[1] Our local goal is to illustrate a mathematical construction rather than to work out a particular application; the reader is welcome to invent a plausible "covering story" for this construction.

along with the linear mapping

$$A[x^1;...;x^M] = \sum_{i=1}^M A_i x^i : \mathbf{R}^{Mn} \to \mathbf{R}^m$$

and linear function

$$f([x^1;...;x^M]) = e^T \sum_i x^i : \mathbf{R}^{Mn} \to \mathbf{R},$$

and let $\mathcal{X}$ be a convex compact set in $\mathbf{R}^{Mn}$ containing all the sets X_i, $1 \le i \le m$. Observe that the problem we are interested in is nothing but the problem of recovering $f(x)$ via observation

$$\omega = Ax + \sigma\xi,\ \ \xi \sim \mathcal{N}(0, I_m), \tag{3.2.9}$$

where the unknown signal x is known to belong to the union $\bigcup_{i=1}^M X_i$ of known convex compact sets X_i. As a result,our problem can be solved via the machinery we have developed.

Numerical illustration. In the numerical results to be reported, we used $n = 128$, $m = 64$ and $M = 2$. The data was generated as follows:

- The set $U \subset \mathbf{R}^{128}$ of candidate signals was comprised by restrictions onto equidistant ($n = 128$)-point grid in $[0, 1]$ of twice differentiable functions $h(t)$ of continuous argument $t \in [0, 1]$ satisfying the relations $|h(0)| \le 1$, $|h'(0)| \le 1$, $|h''(t)| \le 1$, $0 \le t \le 1$, which for the discretized signal $u = [h(0); h(1/n); h(2/n); ...; h(1 - 1/n)]$ translates to the system of convex constraints

$$|u_1| \le 1, n|u_2 - u_1| \le 1,\ n^2|u_{i+1} - 2u_i + u_{i-1}| \le 1,\ 2 \le i \le n - 1.$$

- We were interested to recover the discretized counterpart of the integral $\int_0^1 h(t)dt$, specifically, dealt with $e = \bar{e}$, $\bar{e}^T u = \alpha \sum_{i=1}^n u_i$. The normalizing constant α was selected to ensure $\max_{u\in U} \bar{e}^T u = 1$, $\min_{u\in U} \bar{e}^T u = -1$, allowing to run Bisection with $\Delta_0 = [-1; 1]$.

- We generated A_1 as $(m = 64) \times (n = 128)$ matrix with singular values $\sigma_i = \theta^{i-1}$, $1 \le i \le m$, with θ selected from the requirement $\sigma_m = 0.1$. The system of left singular vectors of A_1 was obtained from the system of basic orths in $\mathbf{R}^n$ by random rotation.

  Matrix A_2 was selected as $A_2 = A_1 S$, where S was "reflection w.r.t. the axis $\bar{e}$", that is,

$$S\bar{e} = \bar{e} \ \&\ Sh = -h \text{ whenever } h \text{ is orthogonal to } \bar{e}. \tag{3.2.10}$$

  Signals u underlying the observations were selected in U at random.

- The reliability $1 - \epsilon$ of our estimate was set to 0.99, and the maximal allowed number L of Bisection steps was set to 8. We used single observation (3.2.9) (i.e., used $K = 1$ in our general scheme) with σ set to 0.01.

The results of our experiments are presented in Table 3.1. Note that in the problem we are considering, there exists an intrinsic obstacle for high accuracy estimation even in the case of noiseless observations and invertible matrices A_i, $i = 1, 2$ (recall that we are in the case of $M = 2$). Indeed, assume that there exist $u \in U$, $u' \in U$ such that $A_1 u = A_2 u'$ and $e^T u \neq e^T u'$. In this case, when the signal is u and the (noiseless) observation is $A_1 u$, the true quantity to be estimated is $e^T u$, and when the signal is u' and the observation is $A_2 u'$, the true quantity to be estimated is $e^T u' \neq e^T u$. Since we do not know which of the matrices, A_1 or A_2, underlies the observation and $A_1 u = A_2 u'$, there is no way to distinguish between the two cases we have described, implying that the quantity

$$\rho = \max_{u,u'\in U} \left\{ \frac{1}{2} |e^T(u - u')| : A_1 u = A_2 u' \right\} \tag{3.2.11}$$

Characteristic	min	median	mean	max
error bound	0.008	0.015	0.014	0.015
actual error	0.001	0.002	0.002	0.005
# of Bisection steps	5	7.00	6.60	8

Table 3.1: Experiments with Bisection, data over 10 experiments, $\sigma = 0.01$. In the table, "error bound" is half-length of final localizer, which is an 0.99-reliable upper bound on the estimation error, and "actual error" is the actual estimation error.

is a lower bound on the worst-case, over signals from U, error of a reliable recovery of $e^T u$, independently of how small is the noise. In the reported experiments, we used $A_2 = A_1 S$ with S linked to $e = \bar{e}$, see (3.2.10); with this selection of S, $e = \bar{e}$ and A_2, *were* A_1 *invertible*, the lower bound ρ would be just trivial – zero. In fact, our A_1 was not invertible, resulting in a positive ρ; computation shows, however, that with our data, this positive ρ is negligibly small (about 2.0e-5). When we destroy the link between e and S, the estimation problem can become intrinsically more difficult, and the performance of our estimation procedure can deteriorate. Let us look what happens when we keep A_1 and $A_2 = A_1 S$ exactly as they are, but replace the linear form $\bar{e}^T u$ to be estimated with $e^T u$, e being randomly selected e [2]. The corresponding data are presented in Table 3.2. The data in the top part of Table relate to the case of "difficult" signals u – those participating in forming the lower bound (3.2.11) on the recovery error, while the data in the bottom part of Table relate to randomly selected signals [3]. We see that when recovering the value of a randomly selected linear form, the error bounds indeed deteriorate, as compared to those in Table 3.1. We see also that the resulting error bounds are in reasonably good agreement with the lower bound ρ, illustrating the basic property of nearly optimal estimates: the guaranteed performance of an estimate can be bad or good, but it always is nearly as good as is possible under the circumstances. As about actual estimation errors, they in some experiments were essentially less than the error bounds, especially when random signals were used. This phenomenon, of course, should not be overestimated; remember that even a broken clock twice a day shows the correct time.

3.2.7 Estimating N-convex functions: an alternative

Observe that the problem of estimating an N-convex function on the union of convex sets posed in Section 3.2.2 can be processed not only by Bisection. An alternative is as follows. In the notation from Section 3.2.2, we start with computing the range Δ of function f on the set $X = \bigcup\limits_{i \le I} X_i$, that is, we compute the quantities

$$\underline{f} = \min_{x\in X} f(x), \ \overline{f} = \max_{x \in X} f(x)$$

and set $\Delta = [\underline{f}, \overline{f}]$. We assume that this segment is not a singleton, otherwise estimating f is trivial. Further, we split Δ in a number L of consecutive bins – segments Δ_ℓ of equal length $\delta_L = (\overline{f} - \underline{f})/L$. δ_L will be the accuracy of our estimate; given a desired accuracy, we can select L accordingly. We now consider the sets

$$X_{i\ell} = \{x \in X_i : f(x) \in \Delta_\ell\}, \, 1 \le i \le I, 1 \le \ell \le L.$$

[2] in the experiments to be reported, e was selected as follows: we start with a random unit vector drawn from the uniform distribution on the unit sphere in $\mathbf{R}^n$ and then normalize it to make $\max_{u\in U} e^T u - \min_{u \in U} e^T u = 2$.

[3] specifically, to generate a signal u, we drew a point $\bar{u}$ at random, from the uniform distribution on the sphere of radius $10\sqrt{n}$, and took as u the $\|\cdot\|_2$-closest to $\bar{u}$ point of U.

Characteristic	min	median	mean	max
error bound	0.057	0.457	0.441	1.000
actual error	0.001	0.297	0.350	1.000
# of Bisection steps	1	1.00	2.20	5

“Difficult” signals, data over 10 experiments

ρ	0.0223	0.0281	0.1542	0.1701	0.2130	0.2482	0.2503	0.4999	0.6046	0.9238
error bound	0.0569	0.0625	0.2188	0.2393	0.4063	0.5078	0.5156	0.6250	0.7734	1.0000

Error bound vs. ρ, experiments sorted according to the values of ρ

Characteristic	min	median	mean	max
error bound	0.016	0.274	0.348	1.000
actual error	0.005	0.066	0.127	0.556
# of Bisection steps	1	2.00	2.80	7

Random signals, data over 10 experiments

ρ	0.0100	0.0853	0.1768	0.2431	0.2940	0.3336	0.3365	0.5535	0.6300	0.7616
error bound	0.0156	0.1816	0.3762	0.4375	0.6016	0.0293	0.0313	0.6875	0.1250	1.0000

Error bound vs. ρ, experiments sorted according to the values of ρ

Table 3.2: Experiments with randomly selected linear form, $\sigma = 0.01$

Since f is N-convex, every one of these sets is the union of $M_{i\ell} \leq N^2$ convex compact sets $X_{i\ell j}$, $1 \leq j \leq M_{i\ell}$. Thus, we get at our disposal a collection of at most ILN^2 convex compact sets; let us eliminate from this collection empty sets and arrange the nonempty ones into a sequence $Y_1, ..., Y_M$, $M \leq ILN^2$. Note that $\bigcup_{s\leq M} Y_s = X$, so that the goal posed in Section 3.2.2 can be reformulated as follows:

> For some *unknown* x known to belong to $X = \bigcup\limits_{s=1}^{M} Y_s$, we have at our disposal observation $\omega^K = (\omega_1, ..., \omega_K)$ with i.i.d. $\omega_t \sim p_{A(x)}(\cdot)$; our goal is to estimate from this observation the quantity $f(x)$.

The sets Y_s give rise to M hypotheses $H_1, ..., H_M$ on the distribution of our observations ω_t, $1 \leq t \leq K$; according to H_s, $\omega_t \sim p_{A(x)}(\cdot)$ with some $x \in Y_s$.

Let us define a closeness $\mathcal{C}$ on the set of our M hypotheses as follows. Given $s \leq M$, the set Y_s is some $X_{i(s)\ell(s)j(s)}$; we say that two hypotheses, H_s and $H_{s'}$, are $\mathcal{C}$-close, if the segments $\Delta_{\ell(s)}$ and $\Delta_{\ell(s')}$ intersect. Observe that when H_s and $H_{s'}$ are *not* $\mathcal{C}$-close, the convex compact sets Y_s, Y_s' do not intersect, since the values of f on Y_s belong to $\Delta_{\ell(s)}$, the values of f on $Y_{s'}$ belong to $\Delta_{\ell(s')}$, and the segments $\Delta_{\ell(s)}$ and $\Delta_{\ell(s')}$ do not intersect.

Now let us apply to the hypotheses $H_1, ..., H_M$ our machinery for testing up to closeness $\mathcal{C}$, see Section 2.5.2. Assuming that whenever H_s and $H_{s'}$ are not $\mathcal{C}$-close, the risks $\epsilon_{ss'}$ defined in Section 2.5.2.2 are < 1 [4], we, given tolerance $\epsilon \in (0,1)$, can find $K = K(\epsilon)$ such that stationary K-repeated observation ω^K allows to decide $(1-\epsilon)$-reliably on $H_1, ..., H_M$ up to closeness $\mathcal{C}$. As applied to ω^K, the corresponding test $\mathcal{T}^K$ will accept some (perhaps, none) of the hypotheses, let the indexes of the accepted hypotheses form set $S = S(\omega^K)$. We convert S into an estimate $\widehat{f}(\omega^K)$ of $f(x)$, $x \in X = \bigcup_{s\leq M} Y_s$ being the signal underlying our observation, as follows:

- when S is empty, the estimate is, say $(\overline{f} + \underline{f})/2$;
- when S is nonempty, we take the union $\Delta(S)$ of the segments $\Delta_{\ell(s)}$, $s \in S$, and our estimate is the average of the largest and the smallest elements of $\Delta(S)$.

It is immediately seen (check it!) that if the signal x underlying our stationary K-repeated observation ω^K belongs to some Y_{s_*}, so that the hypothesis H_{s_*} is true, and the outcome S of $\mathcal{T}^K$ contains s_* and is such that for all $s \in S$ H_s and H_{s_*} are $\mathcal{C}$-close to each other, we have $|f(x) - \widehat{f}(\omega^K)| \leq \delta_L$. Note that since $\mathcal{C}$-risk of $\mathcal{T}^K$ is $\leq \epsilon$, the $p_{A(x)}$-probability to get $|f(x) - \widehat{f}(\omega^K)| \leq \delta_L$ is at least $1 - \epsilon$.

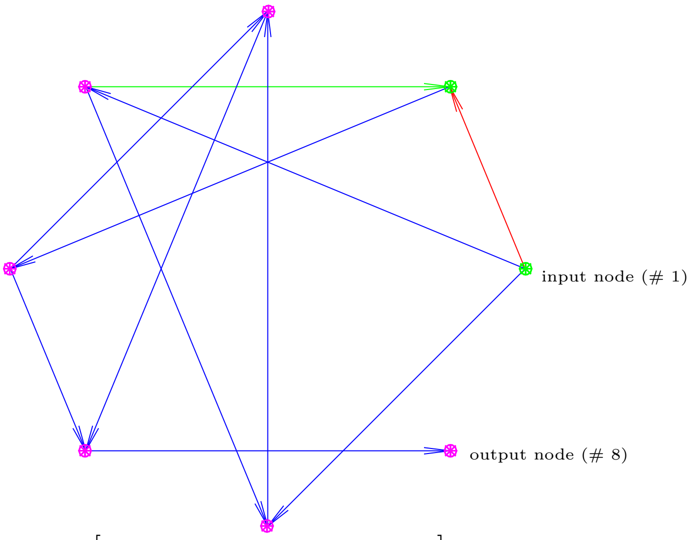
3.2.7.1 Numerical illustration

Our illustration deals with the situation when $I = 1$, $X = X_1$ is a convex compact set, and $f(x)$ is fractional-linear: $f(x) = a^T x / c^T x$ with positive on X denominator. Specifically, assume we are given noisy measurements of voltages V_i at *some* nodes i and currents I_{ij} in *some* arcs (i,j) of an electric circuit, and want to recover the resistance of a particular arc $(\hat{i}, \hat{j})$:

$$r_{\hat{i}\hat{j}} = \frac{V_{\hat{j}} - V_{\hat{i}}}{I_{\hat{i}\hat{j}}}.$$

[4] In our standard simple o.s.'s, this is the case whenever for s, s' in question the images of Y_s and $Y_{s'}$ under the mapping $x \mapsto A(x)$ do not intersect; this definitely is the case when $A(\cdot)$ is an embedding, since for our s, s', Y_s and $Y_{s'}$ do not intersect.

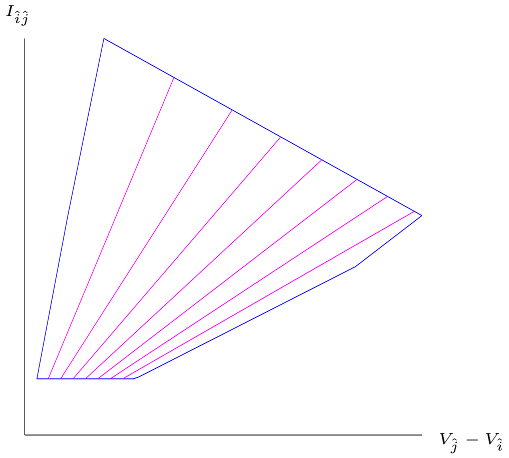
In our experiment, we work with the data as follows:



$$x = \big[\text{voltages at nodes; currents in arcs}\big]$$
$$Ax = \big[\text{observable voltages; observable currents}\big]$$

- Currents are measured in blue arcs only
- Voltages are measured in magenta nodes only
- We want to recover resistance of red arc
- $X: \left\{\begin{array}{l}\textit{conservation of current, except for nodes ##1,8}\\ \textit{zero voltage at node #1, nonnegative currents}\\ \textit{current in red arc at least 1, total of currents at most 33}\\ \textit{Ohm Law, resistances of arcs between 1 and 10}\end{array}\right.$

We are in the situation $N = 1$, $I = 1$, implying $M = L$. When using $L = 8$, the projections of the sets Y_s, $1 \leq s \leq L = 8$ onto the 2D plane of variables $(V_{\hat{j}} - V_{\hat{i}}, I_{\hat{i}\hat{j}})$ are the "stripes" shown below:



The range of the unknown resistance turns out to be $\Delta = [1, 10]$.

In our experiment we worked with $\epsilon = 0.01$. Instead of looking for K such that K-repeated observation allows to recover 0.99-reliably the resistance in the arc of interest within accuracy $|\Delta|/L$, we looked for the largest observation noise σ allowing to achieve the desired recovery with single observation. The results for $L = 8, 16, 32$ are as follows

L	8	16	32
δ_L	$9/8 \approx 1.13$	$9/16 \approx 0.56$	$9/32 \approx 0.28$
σ	0.024	0.010	0.005
$\sigma_{\rm opt}/\sigma \leq$	1.31	1.31	1.33
σ	0.031	0.013	0.006
$\sigma_{\rm opt}/\sigma \leq$	1.01	1.06	1.08

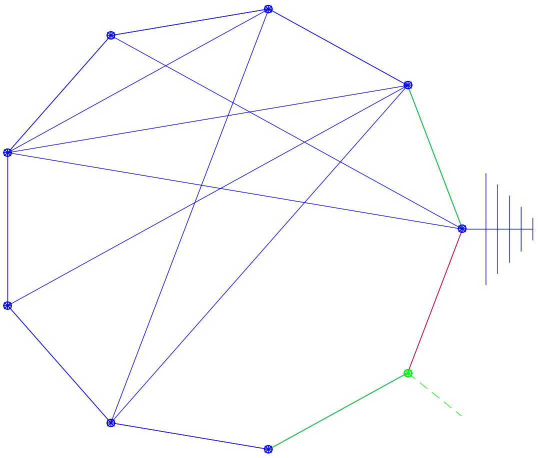


Figure 3.2: A circuit (9 nodes, 16 arcs). Red: arc of interest; Green: arcs with measured currents and nodes with measured voltages.

In the table:

- $\sigma_{\rm opt}$ is the largest σ for which "in the nature" there exists a test deciding on $H_1,...,H_L$ with $\mathcal{C}$-risk ≤ 0.01;
- Red data: Risks $\epsilon_{ss'}$ of pairwise tests are bounded via risks of optimal detectors, $\mathcal{C}$-risk of $\mathcal{T}$ is bounded by
$$\left\|[\epsilon_{ss'}\chi_{ss'}]_{s,s'=1}^L\right\|_{2,2}, \chi_{ss'}=\left\{\begin{array}{ll}1, & (s,s')\not\in\mathcal{C}\\ 0, & (s,s')\in\mathcal{C}\end{array}\right.,$$
see Proposition 2.5.4;
- Brown data: Risks $\epsilon_{ss'}$ of pairwise tests are bounded via error function, $\mathcal{C}$-risk of $\mathcal{T}$ is bounded by
$$\max_s \sum_{s':(s,s')\not\in\mathcal{C}} \epsilon_{ss'}$$
(check that in the case of Gaussian o.s., this indeed is a legitimate risk bound).

3.2.7.2 Estimating dissipated power

The alternative approach to estimating N-convex functions proposed in Section 3.2.7 can be combined with quadratic lifting from Section 2.8.3 to yield, under favorable circumstances, estimates of quadratic and quadratic fractional functions. We are about to consider an instructive example of this sort. Figure 3.2 represent a DC electrical circuit. We have access to repeated noisy measurements of currents in green arcs and voltages at green nodes, with the voltage of the ground node equal to 0. The arcs are somehow oriented; this orientation, however, is of no relevance in our context and therefore is not displayed. Our goal is to use these observations to estimate the power dissipated in a given "arc of interest." Our a priori information is as follows:

- the (unknown) resistances of arcs are known to belong to a given range $[r,R]$, with $0<r<R<\infty$;
- the currents and the voltages are linked by Kirchhoff Laws:

- at every node, the sum of currents in the outgoing arcs is equal to the sum of currents in the incoming arcs plus the external current at the node.
  In our circuit, there are just two external currents, one at the ground node and one at the input node marked by dashed line.
- the voltages and the currents are linked by Ohm's Law: for every (inner) arc γ, we have

$$I_\gamma r_\gamma = V_{j(\gamma)} - V_{i(\gamma)}$$

where I_γ is the current in the arc, r_γ is the arc's resistance, V_s is the voltage at node s, and $i(\gamma)$, $j(\gamma)$ are the initial and the final nodes linked by arc γ;

- magnitudes of all currents and voltages are bounded by 1.

We assume that the measurements of observable currents and voltages are affected by zero mean Gaussian noise with scalar covariance matrix $\theta^2 I$, with unknown θ from a given range $[\underline{\sigma}, \overline{\sigma}]$.

Processing the problem. We specify the "signal" underlying our observation as the collection u of the voltages at our 9 nodes and currents I_γ in our 16 (inner) arcs γ, augmented by external current I_o at the input node (so that $-I_o$ is the external current at the ground node), so that our single-time observation is

$$\zeta = Au + \theta\xi, \tag{3.2.12}$$

where A extracts from u four entries, $\xi \sim \mathcal{N}(0, I_4)$, and $\theta \in [\underline{\sigma}, \overline{\sigma}]$. Our a priori information on u states that u belongs to the compact set U given by the quadratic constraints, namely, as follows:

$$U = \left\{ u = \{I_\gamma, I_o, V_i\} : \begin{array}{c} I_\gamma^2 \le 1, V_i^2 \le 1\,\forall\gamma, i; u^T J^T J u = 0 \\ \left.\begin{array}{c}[V_{j(\gamma)} - V_{i(\gamma)}]^2/R - I_\gamma[V_{j(\gamma)} - V_{i(\gamma)}] \le 0 \\ I_\gamma[V_{j(\gamma)} - V_{i(\gamma)}] - [V_{j(\gamma)} - V_{i(\gamma)}]^2/r \le 0\end{array}\right\} \forall\gamma \quad (a) \\ \left.\begin{array}{c} rI_\gamma^2 - I_\gamma[V_{j(\gamma)} - V_{i(\gamma)}] \le 0 \\ I_\gamma[V_{j(\gamma)} - V_{i(\gamma)}] - RI_\gamma^2 \le 0\end{array}\right\} \forall\gamma \quad (b) \end{array} \right\} \tag{3.2.13}$$

where $Ju = 0$ expresses the first Kirchhoff's Law, and quadratic constraints (a), (b) account for the Ohm's Law in the situation when we do not know the resistances, just the range $[r, R]$ of them. Note that the groups (a), (b) of constraints in (3.2.13) are "logical consequences" of each other, and thus one of groups seems to be redundant. However, on a closest inspection, valid on U quadratic inequalities are indeed redundant in our context, that is, do not tighten the outer approximation $\mathcal{Z}$ of $\mathcal{Z}[U]$, only when these inequalities can be obtained from the inequalities we do include into the description of $\mathcal{Z}$ "in a linear fashion" – by taking weighted sum with nonnegative coefficients; this is *not* how (b) is obtained from (a). As a result, to get a smaller $\mathcal{Z}$, it makes sense to keep both (a) and (b).

The dissipated power we are interested to estimate is the quadratic function

$$f(u) = I_{\gamma_*}[V_{j_*} - V_{i_*}] = [u; 1]^T G [u; 1]$$

where $\gamma_* = (i_*, j_*)$ is the arc of interest, and $G \in \mathbf{S}^{n+1}$, $n = \dim u$, is a properly built matrix.

In order to build an estimate, we "lift quadratically" the observations:

$$\zeta \mapsto \omega = (\zeta, \zeta\zeta^T)$$

and pass from the domain U of actual signals to the outer approximation $\mathcal{Z}$ of the quadratic lifting of U:

$$\mathcal{Z} := \{Z \in \mathbf{S}^{n+1} : Z \succeq 0, Z_{n+1,n+1} = 1, \mathrm{Tr}(Q_s Z) \le c_s, 1 \le s \le S\} \supset \left\{[u; 1][u; 1]^T : u \in \mathcal{V}\right\},$$

where the matrix $Q_s \in \mathbf{S}^{n+1}$ represents the left hand side $F_s(u)$ of s-th quadratic constraint participating in the description (3.2.13) of U: $F_s(u) \equiv [u;1]^T Q_s [u;1]$, and c_s is the right hand side of s-th constraint.

We process the problem similarly to what was done in Section 3.2.7.1, where our goal was to estimate a fractional-linear function. Specifically,

1. We compute the range of f on U; the smallest value $\underline{f}$ of f on U clearly is zero, and an upper bound on the maximum of $f(u)$ over $u \in U$, is the optimal value in the convex optimization problem
$$\overline{f} = \max_{Z\in\mathcal{Z}} \mathrm{Tr}(GZ)$$

2. Given a positive integer L, we split the range $[\underline{f},\overline{f}]$ into L segments $\Delta_\ell = [a_{\ell-1}, a_\ell]$ of equal length $\delta_L = (\overline{f} - \underline{f})/L$ and define convex compact sets
$$\mathcal{Z}_\ell = \{Z \in \mathcal{Z} : a_{\ell-1} \leq \mathrm{Tr}(GZ) \leq a_\ell\},\ 1 \leq \ell \leq L,$$
so that
$$u \in U, f(u) \in \Delta_\ell \Rightarrow [u;1][u;1]^T \in \mathcal{Z}_\ell,\ 1 \leq \ell \leq L;$$

3. We specify L quadratically constrained hypotheses $H_1, ..., H_L$ on the distribution of observation (3.2.12), with H_ℓ stating that $\zeta \sim \mathcal{N}(Au, \theta^2 I_4)$ with some $u \in U$ satisfying $f(u) \in \Delta_\ell$ (so that $[u;1][u;1]^T \in \mathcal{Z}_\ell$), and θ belongs to the above segment $[\underline{\sigma}, \overline{\sigma}]$].

   We equip our hypotheses with closeness relation $\mathcal{C}$, specifically, say that H_ℓ, $H_{\ell'}$ are $\mathcal{C}$-close if and only if the segments Δ_ℓ and $\Delta_{\ell'}$ intersect.

4. We use Proposition 2.8.7.ii to build quadratic in ζ detectors $\phi_{\ell\ell'}$ for the families of distributions obeying H_ℓ and $H_{\ell'}$, respectively, along with upper bounds $\epsilon_{\ell\ell'}$ on the risks of these detectors, and then use the machinery from Section 2.5.2 to find the smallest K and a test $\mathcal{T}_{\mathcal{C}}^K$, based on stationary K-repeated version of observation (3.2.12), capable to decide on $H_1, ..., H_L$ with $\mathcal{C}$-risk $\leq \epsilon$, where $\epsilon \in (0,1)$ is a given tolerance.

Finally, given stationary K-repeated observation (3.2.12), we apply to it test $\mathcal{T}_{\mathcal{C}}^K$, look at the hypotheses, if any, accepted by the test, and build the union Δ of the corresponding segments Δ_ℓ. If $\Delta = \emptyset$, we estimate $f(u)$ by the midpoint of the range $[\underline{f},\overline{f}]$ of power, otherwise the estimate is the mean of the largest and the smallest points in Δ. It is easily seen (check it!) that for this estimate, the probability for the estimation error to be $> \delta_\ell$ is $\leq \epsilon$.

Numerical results we are about to report deal with the circuit presented on Figure 3.2; we used $\overline{\sigma} = 0.01$, $\underline{\sigma} = \overline{\sigma}/\sqrt{2}$, $[r, R] = [1, 2]$, $\epsilon = 0.01$, and $L = 8$. The numerical results are as follows. The range $[\underline{f},\overline{f}]$ of the dissipated power turned out to be $[0, 0.821]$, so that the estimate we have built with reliability 0.99 recovers the dissipated power within accuracy 0.103. The resulting value of K was $K = 95$.

In a series of 500 simulations, the actual recovery error *all the time* was less than the bound 0.103, and the average error was as small as 0.041.

3.2.8 Proof of Proposition 3.2.1

3.2.8.1 Proof of Proposition 3.2.1.i

We call step ℓ *essential*, if at this step rule 2d is invoked.

1^0. Let $x \in X$ be the true signal underlying our observation $\bar{\omega}^K$, so that $\bar{\omega}_1, ..., \bar{\omega}_K$ are independently of each other drawn from the distribution $p_{A(x)}$. Consider the "ideal" estimate given by exactly the same rules as the procedure above (in the sequel, we call the latter the "true" one), up to the fact that the role of the tests $\mathcal{T}^K_{\Delta_{\ell,\mathrm{rg}},\mathrm{r}}(\cdot)$, $\mathcal{T}^K_{\Delta_{\ell,\mathrm{lf}},\mathrm{l}}(\cdot)$ in rule 2d is played by the "tests"

$$\widehat{T}_{\Delta_{\ell,\mathrm{rg}},\mathrm{r}} = \widehat{T}_{\Delta_{\ell,\mathrm{lf}},\mathrm{l}} = \left\{ \begin{array}{ll} \texttt{red}, & f(x) > c_\ell \\ \texttt{blue}, & f(x) \leq c_\ell \end{array} \right.$$

Marking by * the entities produced by the resulting *fully deterministic* procedure, we arrive at nested sequence of segments $\Delta^*_\ell = [a^*_\ell, b^*_\ell]$, $0 \leq \ell \leq L^* \leq L$, along with subsegments $\Delta^*_{\ell,\mathrm{rg}} = [c^*_\ell, v^*_\ell]$, $\Delta^*_{\ell,\mathrm{lf}} = [u^*_\ell, c^*_\ell]$ of $\Delta^*_{\ell-1}$, defined for all *-essential values of ℓ, and the output segment $\bar{\Delta}^*$ claimed to contain $f(x)$. Note that the ideal procedure cannot terminate due to arriving at a disagreement, and that $f(x)$, as is immediately seen, is contained in all segments Δ^*_ℓ, $0 \leq \ell \leq L^*$, same as $f(x) \in \bar{\Delta}^*$.

Let $\mathcal{L}^*$ be the set of all *-essential values of ℓ. For $\ell \in \mathcal{L}^*$, let the event $\mathcal{E}_\ell[x]$ parameterized by x be defined as follows:

$$\mathcal{E}_\ell[x] = \left\{ \begin{array}{ll} \{\omega^K : \mathcal{T}^K_{\Delta^*_{\ell,\mathrm{rg}},\mathrm{r}}(\omega^K) = \texttt{red} \text{ or } \mathcal{T}^K_{\Delta^*_{\ell,\mathrm{lf}},\mathrm{l}}(\omega^K) = \texttt{red}\}, & f(x) \leq u^*_\ell \\ \{\omega^K : \mathcal{T}^K_{\Delta^*_{\ell,\mathrm{rg}},\mathrm{r}}(\omega^K) = \texttt{red}\}, & u^*_\ell < f(x) \leq c^*_\ell \\ \{\omega^K : \mathcal{T}^K_{\Delta^*_{\ell,\mathrm{lf}},\mathrm{l}}(\omega^K) = \texttt{blue}\}, & c^*_\ell < f(x) < v^*_\ell \\ \{\omega^K : \mathcal{T}^K_{\Delta^*_{\ell,\mathrm{rg}},\mathrm{r}}(\omega^K) = \texttt{blue} \text{ or } \mathcal{T}^K_{\Delta^*_{\ell,\mathrm{lf}},\mathrm{l}}(\omega^K) = \texttt{blue}\}, & f(x) \geq v^*_\ell \end{array} \right. \tag{3.2.14}$$

2^0. Observe that by construction and in view of Proposition 2.5.2 we have

$$\forall \ell \in \mathcal{L}^* : \mathrm{Prob}_{\omega^K \sim p_{A(x)} \times ... \times p_{A(x)}} \{\mathcal{E}_\ell[x]\} \leq 2\delta. \tag{3.2.15}$$

Indeed, let $\ell \in \mathcal{L}^*$.

- When $f(x) \leq u^*_\ell$, we have $x \in X$ and $f(x) \leq u^*_\ell \leq c^*_\ell$, implying that $\mathcal{E}_\ell[x]$ takes place only when either the left-side test $\mathcal{T}^K_{\Delta^*_{\ell,\mathrm{lf}},\mathrm{l}}$, or the right side test $\mathcal{T}^K_{\Delta^*_{\ell,\mathrm{rg}},\mathrm{r}}$, or both, accepted wrong – red – hypotheses from the pairs of red and blue hypotheses the tests were applied to. Since the corresponding intervals ($[u^*_\ell, c^*_\ell]$ for the left side test, $[c^*_\ell, v^*_\ell]$ for the right side one) are δ-good left/right, respectively, the risks of the tests do not exceed δ, and the $p_{A(x)}$-probability of the event $\mathcal{E}_\ell[x]$ is at most 2δ;
- when $u^*_\ell < f(x) \leq c^*_\ell$, the event $\mathcal{E}_\ell[x]$ takes place only when the right side test $\mathcal{T}^K_{\Delta^*_{\ell,\mathrm{rg}},\mathrm{r}}$ accepts wrong – red – of the hypotheses from the pair it is applied to; similarly to the above, this can happen with $p_{A(x)}$-probability at most δ;
- when $c_\ell < f(x) \leq v_\ell$, the event $\mathcal{E}_\ell[x]$ takes place only when the left-side test $\mathcal{T}^K_{\Delta^*_{\ell,\mathrm{lf}},\mathrm{l}}$ accepted wrong – blue – hypothesis from the pair it was applied to, which again happens with $p_{A(x)}$-probability $\leq \delta$;
- finally, when $f(x) > v_\ell$, the event $\mathcal{E}_\ell[x]$ takes place only when either the left-side test $\mathcal{T}^K_{\Delta^*_{\ell,\mathrm{lf}},\mathrm{l}}$, or the right side test $\mathcal{T}^K_{\Delta^*_{\ell,\mathrm{rg}},\mathrm{r}}$, or both, accepted wrong – blue – hypotheses from the pairs of red and blue hypotheses the tests were applied to; same as above, this can happen with $p_{A(x)}$-probability at most 2δ.

3^0. Let $\bar{L} = \bar{L}(\bar{\omega}^K)$ be the last step of the "true" estimating procedure as run on the observation $\bar{\omega}^K$. We claim that the following holds true:

> **(!)** *Let $\mathcal{E} := \bigcup_{\ell \in \mathcal{L}^*} \mathcal{E}_\ell[x]$, so that the $p_{A(x)}$-probability of the event $\mathcal{E}$, the observations stemming from x, is at most*
>
> $$2\delta L = \epsilon$$

by (3.2.15). *Assume that* $\bar{\omega}^K \not\in \mathcal{E}$. *Then* $\bar{L}(\bar{\omega}^K) \leq L^*$, *and just two cases are possible:*

(!.A) *The true estimating procedure was not terminated due to arriving at disagreement. In this case* $L^* = \bar{L}(\bar{\omega}^K)$ *and the trajectories of the ideal and the true procedures are identical* (the same localizers, essential steps, the same output segments, etc.), and in particular $f(x) \in \bar{\Delta}$, *or*

(!.B) *The true estimating procedure was terminated due to arriving at a disagreement. Then* $\Delta_\ell = \Delta_\ell^*$ *for* $\ell < \bar{L}$, *and* $f(x) \in \bar{\Delta}$.

In view of **A**, **B** *the* $p_{A(x)}$*-probability of the event* $f(x) \in \bar{\Delta}$ *is at least* $1 - \epsilon$, *as claimed in Proposition 3.2.1.*

To prove **(!)**, note that the actions at step ℓ in the ideal and the true procedures depend solely on $\Delta_{\ell-1}$ and on the outcome of rule 2d. Taking into account that $\Delta_0 = \Delta_0^*$, all we need to verify is the following claim:

(!!) *Let* $\bar{\omega}^K \not\in \mathcal{E}$, *and let* $\ell \leq L^*$ *be such that* $\Delta_{\ell-1} = \Delta_{\ell-1}^*$, *whence also* $u_\ell = u_\ell^*, c_\ell = c_\ell^*, v_\ell = v_\ell^*$. *Assume that* ℓ *is essential (given that* $\Delta_{\ell-1} = \Delta_{\ell-1}^*$, *this may happen if and only if* ℓ *is* **-essential as well). Then either*

C. *At step* ℓ *the true procedure is terminated due to disagreement, in which case* $f(x) \in \bar{\Delta}$, *or*

D. *At step* ℓ *there was no disagreement, in which case* Δ_ℓ *as given by* (3.2.6) *is identical to* Δ_ℓ^* *as given by the ideal counterpart of* (3.2.6) *in the case of* $\Delta_{\ell-1}^* = \Delta_{\ell-1}$, *that is, by the rule*

$$\Delta_\ell^* = \left\{ \begin{array}{ll} [c_\ell, b_{\ell-1}], & f(x) > c_\ell, \\ [a_{\ell-1}, c_\ell], & f(x) \leq c_\ell \end{array} \right. \tag{3.2.16}$$

To verify **(!!)**, let $\bar{\omega}^K$ and ℓ satisfy the premise of **(!!)**. Note that due to $\Delta_{\ell-1} = \Delta_{\ell-1}^*$ we have $u_\ell = u_\ell^*$, $c_\ell = c_\ell^*$, and $v_\ell = v_\ell^*$, and thus also $\Delta_{\ell,\mathrm{lf}}^* = \Delta_{\ell,\mathrm{lf}}$, $\Delta_{\ell,\mathrm{rg}}^* = \Delta_{\ell,\mathrm{rg}}$. Consider first the case when at the step ℓ the true estimation procedure is terminated due to disagreement, so that $\mathcal{T}^K_{\Delta^*_{\ell,\mathrm{lf}},\mathrm{l}}(\bar{\omega}^K) \neq \mathcal{T}^K_{\Delta^*_{\ell,\mathrm{rg}},\mathrm{r}}(\bar{\omega}^K)$. Assuming for a moment that $f(x) < u_\ell = u_\ell^*$, the relation $\bar{\omega}^K \not\in \mathcal{E}_\ell[x]$ combines with (3.2.14) to imply that $\mathcal{T}^K_{\Delta^*_{\ell,\mathrm{rg}},\mathrm{r}}(\bar{\omega}^K) = \mathcal{T}^K_{\Delta^*_{\ell,\mathrm{lf}},\mathrm{l}}(\bar{\omega}^K) = \texttt{blue}$, which under disagreement is impossible. Assuming $f(x) > v_\ell = v_\ell^*$, the same argument results in $\mathcal{T}^K_{\Delta^*_{\ell,\mathrm{rg}},\mathrm{r}}(\bar{\omega}^K) = \mathcal{T}^K_{\Delta^*_{\ell,\mathrm{lf}},\mathrm{l}}(\bar{\omega}^K) = \texttt{red}$, which again is impossible. We conclude that in the case in question $u_\ell \leq f(x) \leq v_\ell$, i.e., $f(x) \in \bar{\Delta}$, as claimed in **C**. **C** is proved.

Now let in the true estimating procedure there was a consensus at the step ℓ. The relation $\bar{\omega}^K \not\in \mathcal{E}_\ell[x]$ implies that one of the following four options takes place:

1. $\mathcal{T}^K_{\Delta^*_{\ell,\mathrm{rg}},\mathrm{r}}(\bar{\omega}^K) = \texttt{blue}$ and $f(x) \leq u_\ell = u_\ell^*$,

2. $\mathcal{T}^K_{\Delta^*_{\ell,\mathrm{rg}},\mathrm{r}}(\bar{\omega}^K) = \texttt{blue}$ and $u_\ell < f(x) \leq c_\ell = c_\ell^*$,

3. $\mathcal{T}^K_{\Delta^*_{\ell,\mathrm{lf}},\mathrm{l}}(\bar{\omega}^K) = \texttt{red}$ and $c_\ell < f(x) < v_\ell = v_\ell^*$,

4. $\mathcal{T}^K_{\Delta^*_{\ell,\mathrm{lf}},\mathrm{l}}(\bar{\omega}^K) = \texttt{red}$ and $v_\ell \leq f(x)$,

In situations 1-2 and due to consensus at the step ℓ, (3.2.6) says that $\Delta_\ell = [a_{\ell-1}, c_\ell]$, which combines with (3.2.16) and $v_\ell = v_\ell^*$ to imply that $\Delta_\ell = \Delta_\ell^*$. Similarly, in situations 3-4 and due to consensus at the step ℓ, (3.2.6) says that $\Delta_\ell = [c_\ell, b_{\ell-1}]$, which combines with $u_\ell = u_\ell^*$ and (3.2.16) to imply that $\Delta_\ell = \Delta_\ell^*$. **B** is proved. □

3.2.9 Proof of Proposition 3.2.1.ii

There is nothing to prove when $\frac{b_0-a_0}{2} \leq \widehat{\rho}$, since in this case the estimate $\frac{a_0+b_0}{2}$ which does not use observations at all is $(\widehat{\rho}, 0)$-reliable. From now on we assume that $b_0 - a_0 > 2\widehat{\rho}$, implying that L is positive integer.

1^0. Observe, first, that if a, b are such that a is lower-feasible, b is upper-feasible, and $b - a > 2\rho$, then for every $i \leq I_{b,\geq}$ and $j \leq I_{a,\leq}$ there exists a test, based on $\bar{K}$ observations, which decides upon the hypotheses H_1, H_2, stating that the observations are drawn from $p_{A(x)}$ with $x \in Z_i^{b,\geq}$ (H_1) and with $x \in Z_j^{a,\leq}$ (H_2) with risk at most ϵ. Indeed, it suffices to consider the test which accepts H_1 and rejects H_2 when $\widehat{f}(\omega^{\bar{K}}) \geq \frac{a+b}{2}$ and accepts H_2 and rejects H_1 otherwise.

2^0. With parameters of Bisection chosen according to (3.2.8), by already proved Proposition 3.2.1.i, we have

> **E.1:** *For every $x \in X$, the $p_{A(x)}$-probability of the event $f(x) \in \bar{\Delta}$, $\bar{\Delta}$ being the output segment of our Bisection, is at least $1 - \epsilon$.*

3^0. We claim also that

F.1. *Every segment $\Delta = [a, b]$ with $b - a > 2\rho$ and lower-feasible a is δ-good (right),*

F.2. *Every segment $\Delta = [a, b]$ with $b - a > 2\rho$ and upper-feasible b is δ-good (left),*

F.3. *Every $\varkappa$-maximal δ-good (left or right) segment has length at most $2\rho + \varkappa = \widehat{\rho}$. As a result, for every essential step ℓ, the lengths of the segments $\Delta_{\ell,\mathrm{rg}}$ and $\Delta_{\ell,\mathrm{lf}}$ do not exceed $\widehat{\rho}$.*

Let us verify F.1 (verification of F.2 is completely similar, and F.3 is an immediate consequence of the definitions and F.1-2). Let $[a, b]$ satisfy the premise of F.1. It may happen that b is upper-infeasible, whence $\Delta = [a, b]$ is 0-good (right), and we are done. Now let b be upper-feasible. As we have already seen, whenever $i \leq I_{b,\geq}$ and $j \leq I_{a,\leq}$, the hypotheses stating that ω_k are sampled from $p_{A(x)}$ for some $x \in Z_i^{b,\geq}$, resp., from some $x \in Z_j^{a,\leq}$, can be decided upon with risk $\leq \epsilon$, implying, same as in the proof of Proposition 2.4.2, that

$$\epsilon_{ij\Delta} \leq [2\sqrt{\epsilon(1-\epsilon)}]^{1/\bar{K}}$$

whence, taking into account that the column and the row sizes of $E_{\Delta,\mathrm{r}}$ do not exceed NI,

$$\sigma_{\Delta,\mathrm{r}} \leq NI \max_{i,j} \epsilon_{ij\Delta}^K \leq NI[2\sqrt{\epsilon(1-\epsilon)}]^{K/\bar{K}} \leq \frac{\epsilon}{2L} = \delta$$

(we have used (3.2.8)), that is, Δ indeed is δ-good (right).

4^0. Let us fix $x \in X$ and consider a trajectory of Bisection, the observation being drawn from $p_{A(x)}$. The output $\bar{\Delta}$ of the procedure is given by one of the following options:

1. At some step ℓ of Bisection, the process was terminated according to rules in 2b or 2c. In the first case, the segment $[c_\ell, b_{\ell-1}]$ has lower-feasible left endpoint and is not δ-good (right), implying by F.1 that the length of this segment (which is 1/2 of the length of $\bar{\Delta} = \Delta_{\ell-1}$) is $\leq 2\rho$, so that the length $|\bar{\Delta}|$ of $\bar{\Delta}$ is at most $4\rho \leq 2\widehat{\rho}$. The same conclusion, by completely similar argument, holds true when the process was terminated at step ℓ according to rule 2c.

2. At some step ℓ of Bisection, the process was terminated due to disagreement. In this case, by F.3, we have $|\bar{\Delta}| \leq 2\widehat{\rho}$.

3. Bisection was terminated at step L, and $\bar{\Delta} = \Delta_L$. In this case, termination clauses in rules 2b, 2c and 2d were never invoked, clearly implying that $|\Delta_s| \leq \frac{1}{2}|\Delta_{s-1}|$, $1 \leq s \leq L$, and thus $|\bar{\Delta}| = |\Delta_L| \leq \frac{1}{2^L}|\Delta_0| \leq 2\widehat{\rho}$ (see (3.2.8)).

Thus, along with E.1 we have

E.2: *It always holds* $|\bar{\Delta}| \leq 2\widehat{\rho}$,

implying that whenever the signal $x \in X$ underlying observations and the output segment $\bar{\Delta}$ are such that $f(x) \in \bar{\Delta}$, the error of the Bisection estimate (which is the midpoint of $\bar{\Delta}$) is at most $\widehat{\rho}$. Invoking E.1, we conclude that the Bisection estimate is $(\widehat{\rho}, \epsilon)$-reliable. □

3.2.10 Appendix: 2-convexity of conditional quantile

A. Let $\mathcal{Q}$ be the family of non-vanishing probability distributions on $S = \{s_1 < s_2 < ... < s_M\} \subset \mathbf{R}$. For $q \in \mathcal{Q}$, let

$$\ell_m(q) = \sum_{i=m}^{M} q_i,\ 1 \leq m \leq M,$$

so that $1 = \ell_1(q) > \ell_2(q) > ... > \ell_M(q) > 0$.

Given $\alpha \in [0,1]$, let us define (regularized) α-quantile of $q \in \mathcal{Q}$, $\chi_\alpha(q)$, as follows:

- if $\ell_M(q) > \alpha$, we set $\chi_\alpha(q) = s_M$;
- otherwise, there exists $m \in \{1, ..., M-1\}$ such that $\ell_m(q) \geq \alpha \geq \ell_{m+1}(q)$. We select an m with this property, set
$$\beta = \frac{\alpha - \ell_{m+1}(q)}{\ell_m(q) - \ell_{m+1}(q)},$$
so that $\beta \in [0,1]$ and $\beta\ell_m(q) + (1-\beta)\ell_{m+1}(q) = \alpha$, and set
$$\chi_\alpha(q) = \beta s_m + (1-\beta)s_{m+1}.$$

Note that for some q, the above m is not uniquely defined; this happens if and only if $\ell_\mu(q) = \alpha$ for some μ, $1 < \mu < M$, in which case there are exactly two options for selecting m, one $m = \mu$, and another $m = \mu - 1$. The first option results in

$$\beta = \frac{\alpha - \ell_{\mu+1}(q)}{\ell_\mu(q) - \ell_{\mu+1}(q)} = \frac{\ell_\mu(q) - \ell_{\mu+1}(q)}{\ell_\mu(q) - \ell_{\mu+1}(q)} = 1 \Rightarrow \beta s_m + (1-\beta)s_{m+1} = s_\mu,$$

and the second option results in

$$\beta = \frac{\alpha - \ell_\mu(q)}{\ell_{\mu-1}(q) - \ell_\mu(q)} = \frac{\ell_\mu(q) - \ell_\mu(q)}{\ell_{\mu-1}(q) - \ell_\mu(q)} = 0 \Rightarrow \beta s_m + (1-\beta)s_{m+1} = s_\mu;$$

Thus, in spite of the fact that m above is not always uniquely defined by α, q, $\chi_\alpha(q)$ is well defined – the value we assign to $\chi_\alpha(q)$ according to the above construction is independent of how the required m is selected.

B. From what was said so far, it is immediately seen that for $s_1 \leq s < s_M$, the relation $\chi_\alpha(q) = s$ is exactly equivalent to the relation

(!) For some $m \in \{1, ..., M-1\}$, we have $\ell_m(q) \geq \alpha \geq \ell_{m+1}(q)$ and $\frac{[\alpha - \ell_{m+1}(q)]s_m + [\ell_m(q) - \alpha]s_{m+1}}{\ell_m(q) - \ell_{m+1}(q)} = s$.

C. Let $\theta_q(s)$, $s_1 \leq s \leq s_M$, be the piecewise linear version to the cumulative distribution of q, that is, the piecewise linear function on $\Delta = [s_1, s_M]$ with breakpoints at $s_1, ..., s_M$ and such that $\theta_q(s_m) = \ell_m(q)$, $1 \leq m \leq M$; this is a strictly decreasing function mapping Δ onto $\Delta^+ := [\ell_M(q), 1]$. For given q, $\chi_\alpha(q)$, as a function of $\alpha \in [0,1]$, is obtained from the inverse of $\theta_q(\cdot)$ by extending this inverse from its natural domain $\Delta^+ \subset [0,1]$ to the entire $[0,1]$ by the value S_M to the left of the left endpoint, $\ell_M(q)$, of Δ^+. As a consequence of this representation, it is immediately seen that $\chi_\alpha(q)$ is continuous in $(\alpha, q) \in [0,1] \times \mathcal{Q}$. Note that $\chi_\alpha(q)$ takes all its values in $\Delta = [s_1, s_M]$.

Note that we have demonstrated the equivalence between the definition of $\chi_\alpha(\cdot)$ via "spreading masses" used in Section 3.2.2.1 and the definition we started with in this Section.

D. Let us fix $\alpha \in (0,1)$. Given $s \in \Delta$, let us look at the set $Q_s^- := \{q \in \mathcal{Q} : \chi_\alpha(q) \leq s\}$. This set is as follows:

1. When $s = s_M$, we have $Q_s^- = \mathcal{Q}$.

2. Now let $s \in \Delta$ be $< s_M$, so that $s \in [s_1, s_M)$. Then for some $\mu = \mu(s) \in \{1, ..., M-1\}$ we have $s_\mu \leq s < s_{\mu+1}$. We claim that now the set Q_s^- is the union of two convex sets:

$$\begin{array}{l} Q_s^- = \underbrace{\{q \in \mathcal{Q} : \ell_\mu(q) \leq \alpha\}}_{A} \cup \\ \underbrace{\{q \in \mathcal{Q} : \ell_\mu(q) \geq \alpha \;\&\; \ell_{\mu+1}(q) \leq \alpha \;\&\; [\alpha - \ell_{\mu+1}(q)]s_\mu + [\ell_\mu(q) - \alpha]s_{\mu+1} \leq s[\ell_\mu(q) - \ell_{\mu+1}(q)]\}}_{B} \end{array} \tag{3.2.17}$$

Indeed, when $q \in A$, then $\mu > 1$, since $\ell_1(q) = 1 > \alpha$; thus, $\ell_1(q) > \alpha$ and $\ell_\mu(q) \leq \alpha$, so that we can find $m \in \{1, ..., \mu - 1\}$ such that $\ell_m(q) \geq \alpha \geq \ell_{m+1}(q)$. By **A**, it implies that $\chi_\alpha(q)$ is a convex combination of s_m and s_{m+1}, and both these quantities are $\leq s_\mu \leq s$, so that $\chi_\alpha(q) \leq s$ as well, i.e., $q \in Q_s^-$; thus, $A \subset Q_s^-$. Now let $q \in B$. In this case we have $\ell_\mu(q) \geq \alpha$, $\ell_{\mu+1}(q) \leq \alpha$, implying that when specifying $\chi_\alpha(q)$ by **A**, we can take $m = \mu$, resulting, by (!), in $\chi_\alpha(q) = \frac{[\alpha - \ell_{\mu+1}(q)]s_\mu + [\ell_\mu(q) - \alpha]s_{\mu+1}}{\ell_\mu(q) - \ell_{\mu+1}(q)}$. The latter expression, by the third inequality in the description of B, is $\leq s$, and we end up with $\chi_\alpha(q) \leq s$ and thus $q \in Q_s^-$. Thus, $A \cup B \subset Q_s^-$. To verify the inverse inclusion, let $q \in \mathcal{Q}$ be such that $\chi_\alpha(q) \leq s$, and let us verify that $q \in A \cup B$. It may happen that $\ell_\mu(q) \leq \alpha$; then $q \in A$, and we are done. Now assume that $\ell_\mu(q) > \alpha$. Observe that in this case $\ell_{\mu+1}(q) \leq \alpha$, since otherwise, by **A**, we would have $\chi_\alpha(q) \geq s_{\mu+1}$, while we are in the case $\chi_\alpha(q) \leq s < s_{\mu+1}$. Thus, $\ell_\mu(q) \geq \alpha \geq \ell_{\mu+1}(q)$, i.e., q satisfies the first two inequalities from the description of B. As a result, by (!), it holds $\chi_\alpha(q) = \frac{[\alpha - \ell_{\mu+1}(q)]s_\mu + [\ell_\mu(q) - \alpha]s_{\mu+1}}{\ell_\mu(q) - \ell_{\mu+1}(q)}$, which combines with $\chi_\alpha(q) \leq s$ to imply the validity at q of the third inequality in the description of B, and we are done.

The bottom line is that Q_s^- is the union of two closed in $\mathcal{Q}$ convex sets, A and B.

Now let us look at the set $Q_s^+ = \{q \in \mathcal{Q} : \chi_\alpha(q) \geq s\}$, where $s \in \Delta$. This set is as follows:

1. When $s = s_M$, Q_s^+, by **A**, is exactly the set $\{q \in \mathcal{Q} : \ell_M(q) \geq \alpha\}$.

2. Now let $s \in \Delta$ be $< s_M$, so that for some $\mu = \mu(s) \in \{1, ..., M-1\}$ we have $s_\mu \leq s < s_{\mu+1}$. We claim that now the set Q_s^+ is the union of two convex sets:

$$\begin{array}{l} Q_s^+ = \underbrace{\{q \in \mathcal{Q} : \ell_{\mu+1}(q) \geq \alpha\}}_{A'} \cup \\ \underbrace{\{q \in \mathcal{Q} : \ell_\mu(q) \geq \alpha \;\&\; \ell_{\mu+1}(q) \leq \alpha \;\&\; [\alpha - \ell_{\mu+1}(q)]s_\mu + [\ell_\mu(q) - \alpha]s_{\mu+1} \geq s[\ell_\mu(q) - \ell_{\mu+1}(q)]\}}_{B'} \end{array} \tag{3.2.18}$$

Indeed, if $q \in A'$, then, by **A**, we either have $\chi_\alpha(q) = s_M > s$, or m in **A** can be chosen to be $\geq \mu + 1$ implying by **A** that $\chi_\alpha(q) \geq s_{\mu+1} > s$; thus, $A' \subset Q_s^+$. Now let $q \in B'$.

From the first two inequalities in the description of B', by (!), we conclude that $\chi_\alpha(q) = \frac{[\alpha-\ell_{\mu+1}(q)]s_\mu+[\ell_\mu(q)-\alpha]s_{\mu+1}}{\ell_\mu(q)-\ell_{\mu+1}(q)}$, and the latter quantity, by the third inequality in the description of B', is $\geq s$, implying that $q \in Q_s^+$. Thus, $B' \subset Q_s^+$, and we see that $A' \cup B' \subset Q_s^+$. To verify the inverse inclusion, let $q \in \mathcal{Q}$ be such that $\chi_\alpha(q) \geq s$, and let us prove that $\chi \in A' \cup B'$. It may happen that $\ell_{\mu+1}(q) \geq \alpha$, in which case $q \in A'$, and we are done. Now let $\ell_{\mu+1}(q) < \alpha$. We claim that $\ell_\mu(q) \geq \alpha$. Indeed, otherwise m in **A** is $< \mu$, implying by **A** $\chi_\alpha(q) \leq s_{m+1} \leq s_\mu$; the equality $\chi_\alpha(q) = s_\mu$ would be possible only when $m = \mu - 1$ and β, as defined in **A**, is equal to 0, that is, $\alpha = \ell_{m+1}(q) = \ell_\mu(q)$, which is not the case. Thus, under assumption $\ell_\mu(q) < \alpha$ it holds $\chi_\alpha(q) < s_\mu$, which is impossible due to $s_\mu \leq s$ and $\chi_\alpha(q) \geq s$. Thus, we are in the case when $\ell_\mu(q) \geq \alpha > \ell_{\mu+1}(q)$, that is, the first two inequalities in the description of B' hold true, which, by (!), implies that $\chi_\alpha(q) = \frac{[\alpha-\ell_{\mu+1}(q)]s_\mu+[\ell_\mu(q)-\alpha]s_{\mu+1}}{\ell_\mu(q)-\ell_{\mu+1}(q)}$; the latter combines with $\chi_\alpha(q) \geq s$ to imply that q satisfies the third inequality in the description of B', that is, $q \in B'$, and we are done.

The bottom line is that Q_s^+ is the union of two closed in $\mathcal{Q}$ convex sets, A' and B'.

We have arrived at the following

Proposition 3.2.2 *Let $S = \{s_1 < s_2 < ... < s_M\}$ be a finite subset of $\mathbf{R}$, T be a finite set, and $\mathcal{P}$ be the set of non-vanishing probability distributions on $\Omega = S \times T$. Given $\tau \in T$ and $\alpha \in (0,1)$, let $\zeta_{\tau,\alpha}(p) : \mathcal{P} \to [s_1, s_M]$ be the α-quantile of the conditional distribution on S induced by a distribution $p \in \mathcal{P}$ and the condition $t = \tau$:*

$$\zeta_{\tau,\alpha}(p) = \chi_\alpha(q_\tau[p]),\ \ (q_\tau[p])_m = \frac{p(m,\tau)}{\sum_{\mu=1}^M p(\mu,\tau)},\ 1 \leq m \leq M.$$

The function $\zeta_{\tau,\alpha}(\cdot)$ is 2-convex on $\mathcal{P}$: for every $s \in [s_1, s_M)$, selecting $\mu \in \{1, ..., M-1\}$ in such a way that $s_\mu \leq s < s_{\mu+1}$, we have

$$\begin{array}{rcl}
\{p \in \mathcal{P} : \zeta_{\tau,\alpha}(p) \leq s\} & = & \{p \in \mathcal{P} : \ell_\mu(p) - \alpha\ell(p) \leq 0\} \\
 & & \bigcup \left\{ p \in \mathcal{P} : \begin{array}{l} \ell_\mu(p) \geq \alpha\ell(p) \geq \ell_{\mu+1}(p), \\ {[\alpha\ell(p) - \ell_{\mu+1}(p)]s_\mu + [\ell_\mu(p) - \alpha\ell(p)]s_{\mu+1} \leq s[\ell_\mu(p) - \ell_{\mu+1}(p)]} \end{array} \right\}, \\
\{p \in \mathcal{P} : \zeta_{\tau,\alpha}(p) \geq s\} & = & \{p \in \mathcal{P} : \ell_{\mu+1}(p) - \alpha\ell(p) \geq 0\} \\
 & & \bigcup \left\{ p \in \mathcal{P} : \begin{array}{l} \ell_\mu(p) \geq \alpha\ell(p) \geq \ell_{\mu+1}(p), \\ {[\alpha\ell(p) - \ell_{\mu+1}(p)]s_\mu + [\ell_\mu(p) - \alpha\ell(p)]s_{\mu+1} \geq s[\ell_\mu(p) - \ell_{\mu+1}(p)]} \end{array} \right\},
\end{array}$$

where

$$\ell_m(p) = \sum_{i=m}^M p(i,\tau), \quad \ell(p) = \sum_{m=1}^M p(m,\tau),$$

and

$$\begin{array}{rcl}
s < s_1 & \Rightarrow & \{p \in \mathcal{P} : \zeta_{\tau,\alpha}(p) \leq s\} = \emptyset,\ \{p \in \mathcal{P} : \zeta_{\tau,\alpha}(p) \geq s\} = \mathcal{P}, \\
s = s_M & \Rightarrow & \{p \in \mathcal{P} : \zeta_{\tau,\alpha}(p) \leq s\} = \mathcal{P},\ \{p \in \mathcal{P} : \zeta_{\tau,\alpha}(p) \geq s\} = \{p \in P : \ell_M(p) \geq \alpha\ell(p)\}. \\
s > s_M & \Rightarrow & \{p \in \mathcal{P} : \zeta_{\tau,\alpha}(p) \leq s\} = \mathcal{P},\ \{p \in \mathcal{P} : \zeta_{\tau,\alpha}(p) \geq s\} = \emptyset.
\end{array}$$

Indeed, it suffices to apply (3.2.18) and (3.2.17) to $q = q_\tau[p]$. □

3.3 Estimating linear forms

We are about to demonstrate that the techniques developed in Section 2.8 can be applied to building estimates of linear and quadratic forms of the parameters of observed distributions. As compared to the machinery of Section 3.2, our new approach has somehow restricted scope: we cannot estimate anymore general N-convex functions and/or handle domains which are unions of convex sets; now we need the function to be linear (perhaps, after quadratic lifting of observations) and the domain to be convex. As a compensation, the new approach, when applicable, seems to be cheaper computationally: the estimate is yielded by solving a single convex problem, while the techniques developed so far require solving several (perhaps even few tens) of problems of similar structure and complexity. In this Section, we focus on estimating linear forms; estimating quadratic forms will be our subject in Section 3.4.

3.3.1 Situation and goal

Consider the situation as follows: given are Euclidean spaces $\Omega = \mathcal{E}_H$, $\mathcal{E}_M$, $\mathcal{E}_X$ along with

- regular data $\mathcal{H} \subset \mathcal{E}_H, \mathcal{M} \subset \mathcal{E}_M, \Phi(\cdot;\cdot) : \mathcal{H} \times \mathcal{M} \to \mathbf{R}$, with $0 \in \operatorname{int} \mathcal{H}$,
- a nonempty convex compact set $\mathcal{X} \subset \mathcal{E}_X$,
- an affine mapping $x \mapsto \mathcal{A}(x) : \mathcal{E}_X \to \mathcal{E}_M$ such that $\mathcal{A}(\mathcal{X}) \subset \mathcal{M}$,
- a continuous convex *calibrating function* $\upsilon(x) : \mathcal{X} \to \mathbf{R}$
- a vector $g \in \mathcal{E}_X$ and a constant c specifying the linear form $G(x) = \langle g, x\rangle + c : \mathcal{E}_X \to \mathbf{R}$ [5],
- a tolerance $\epsilon \in (0, 1)$.

These data specify, in particular, the family

$$\mathcal{P} = \mathcal{S}[\mathcal{H}, \mathcal{M}, \Phi]$$

of probability distributions on $\Omega = \mathcal{E}_H$, see Section 2.8.1.1. Given random observation

$$\omega \sim P(\cdot) \tag{3.3.1}$$

where $P \in \mathcal{P}$ is such that

$$\forall h \in \mathcal{H} : \ln\left(\int_{\mathcal{E}_H} \mathrm{e}^{\langle h,\omega\rangle} P(d\omega)\right) \leq \Phi(h; \mathcal{A}(x)) \tag{3.3.2}$$

for some $x \in \mathcal{X}$ (that is, $\mathcal{A}(x)$ is a parameter, as defined in Section 2.8.1.1, of distribution P), we want to recover the quantity $G(x)$.

ϵ-risk. Given $\rho > 0$, we call an estimate $\widehat{g}(\cdot) : \mathcal{E}_H \to \mathbf{R}$ $(\rho, \epsilon, \upsilon(\cdot))$*-accurate*, if for all pairs $x \in \mathcal{X}$, $P \in \mathcal{P}$ satisfying (3.3.2) it holds

$$\mathrm{Prob}_{\omega\sim P}\left\{|\widehat{g}(\omega) - G(x)| > \rho + \upsilon(x)\right\} \leq \epsilon. \tag{3.3.3}$$

If ρ_* is the infimum of those ρ for which estimate $\widehat{g}$ is $(\rho, \epsilon, \upsilon(\cdot))$-accurate, then clearly $\widehat{g}$ is $(\rho_*, \epsilon, \upsilon(\cdot))$-accurate; we shall call ρ_* the *ϵ-risk* of the estimate $\widehat{g}$ taken w.r.t. the data $G(\cdot)$, $\mathcal{X}$, $\upsilon(\cdot)$ and $(\mathcal{A}, \mathcal{H}, \mathcal{M}, \Phi)$:

$$\begin{array}{l}\mathrm{Risk}_\epsilon(\widehat{g}(\cdot)|G, \mathcal{X}, \upsilon, \mathcal{A}, \mathcal{H}, \mathcal{M}, \Phi) = \min\Bigg\{\rho : \mathrm{Prob}_{\omega\sim P}\{\omega : |\widehat{g}(\omega) - G(x)| > \rho + \upsilon(x)\} \leq \epsilon \\ \qquad\qquad\qquad\qquad \forall (x, P) : \left\{\begin{array}{l} P \in \mathcal{P}, x \in X \\ \ln\left(\int \mathrm{e}^{h^T\omega} P(d\omega)\right) \leq \Phi(h; \mathcal{A}(x)) \forall h \in \mathcal{H}\end{array}\right\}\Bigg\}.\end{array} \tag{3.3.4}$$

When $G, X, \upsilon, \mathcal{A}, \mathcal{H}, \mathcal{M}, \Phi$ are clear from the context, we shorten $\mathrm{Risk}_\epsilon(\widehat{g}(\cdot)|G, X, \upsilon, \mathcal{A}, \mathcal{H}, \mathcal{M}, \Phi)$ to $\mathrm{Risk}_\epsilon(\widehat{g}(\cdot))$.

Given the data listed in the beginning of this section, we are about to build, in a computationally efficient fashion, an affine estimate $\widehat{g}(\omega) = \langle h_*, \omega\rangle + \varkappa$ along with ρ_* such that the estimate is $(\rho_*, \epsilon, \upsilon(\cdot))$-accurate.

[5] from now on, $\langle u, v\rangle$ denotes the inner product of vectors u, v belonging to a Euclidean space; what is this space, it always will be clear from the context.

3.3.2 Construction & Main results

Let us set

$$\mathcal{H}^+ = \{(h,\alpha) : h \in \mathcal{E}_H, \alpha > 0, h/\alpha \in \mathcal{H}\}$$

so that $\mathcal{H}^+$ is a nonempty convex set in $\mathcal{E}_H \times \mathbf{R}_+$, and let

$$\begin{array}{lrcl}
(a) & \Psi_+(h,\alpha) & = & \sup\limits_{x\in\mathcal{X}} \left[\alpha\Phi(h/\alpha, \mathcal{A}(x)) - G(x) - \upsilon(x)\right] : \mathcal{H}^+ \to \mathbf{R}, \\
(b) & \Psi_-(h,\beta) & = & \sup\limits_{x\in\mathcal{X}} \left[\beta\Phi(-h/\beta, \mathcal{A}(x)) + G(x) - \upsilon(x)\right] : \mathcal{H}^+ \to \mathbf{R},
\end{array} \tag{3.3.5}$$

so that $\Psi_\pm$ are convex real-valued functions on $\mathcal{H}^+$ (recall that Φ is convex-concave and continuous on $\mathcal{H} \times \mathcal{M}$, while $\mathcal{A}(\mathcal{X})$ is a compact subset of $\mathcal{M}$).

Our starting point is pretty simple:

Proposition 3.3.1 *Given $\epsilon \in (0,1)$, let $\bar{h}$, $\bar{\alpha}$, $\bar{\beta}$, $\bar{\varkappa}$, $\bar{\rho}$ be a feasible solution to the system of convex constraints*

$$\begin{array}{lrcl}
(a_1) & (h,\alpha) & \in & \mathcal{H}^+ \\
(a_2) & (h,\beta) & \in & \mathcal{H}^+ \\
(b_1) & \alpha\ln(\epsilon/2) & \geq & \Psi_+(h,\alpha) - \rho + \varkappa \\
(b_2) & \beta\ln(\epsilon/2) & \geq & \Psi_-(h,\beta) - \rho - \varkappa
\end{array} \tag{3.3.6}$$

in variables h, α, β, ρ, $\varkappa$. Setting

$$\widehat{g}(\omega) = \langle \bar{h}, \omega\rangle + \bar{\varkappa},$$

we get an estimate with ϵ-risk at most $\bar{\rho}$.

Proof. Let $\epsilon \in (0,1)$, $\bar{h}$, $\bar{\alpha}$, $\bar{\beta}$, $\bar{\varkappa}$, $\bar{\rho}$ satisfy the premise of Proposition, and let $x \in X, P$ satisfy (3.3.2). We have

$$\begin{array}{rl}
& \mathrm{Prob}_{\omega\sim P}\{\widehat{g}(\omega) > G(x) + \bar{\rho} + \upsilon(x)\} = \mathrm{Prob}_{\omega\sim P}\left\{\frac{\langle\bar{h},\omega\rangle}{\bar{\alpha}} > \frac{G(x)+\bar{\rho}-\bar{\varkappa}+\upsilon(x)}{\bar{\alpha}}\right\} \\
\Rightarrow & \mathrm{Prob}_{\omega\sim P}\{\widehat{g}(\omega) > G(x) + \bar{\rho} + \upsilon(x)\} \leq \left[\int \mathrm{e}^{\langle\bar{h},\omega\rangle/\bar{\alpha}} P(d\omega)\right] \mathrm{e}^{-\frac{G(x)+\bar{\rho}-\bar{\varkappa}+\upsilon(x)}{\bar{\alpha}}} \\
& \leq \mathrm{e}^{\Phi(\bar{h}/\bar{\alpha}, \mathcal{A}(x))} \mathrm{e}^{-\frac{G(x)+\bar{\rho}-\bar{\varkappa}+\upsilon(x)}{\bar{\alpha}}} \\
\Rightarrow & \bar{\alpha}\ln\left(\mathrm{Prob}_{\omega\sim P}\{\widehat{g}(\omega) > G(x) + \bar{\rho} + \upsilon(x)\}\right) \leq \bar{\alpha}\Phi(\bar{h}/\bar{\alpha}, \mathcal{A}(x)) - G(x) - \bar{\rho} - \upsilon(x) + \bar{\varkappa} \\
& \leq \Psi_+(\bar{h},\bar{\alpha}) - \bar{\rho} + \bar{\varkappa} \text{ [by definition of } \Psi_+ \text{ and due to } x \in X] \\
& \leq \bar{\alpha}\ln(\epsilon/2) \text{ [by } (b_1)] \\
\Rightarrow & \mathrm{Prob}_{\omega\sim P}\{\widehat{g}(\omega) > G(x) + \bar{\rho} + \upsilon(x)\} \leq \epsilon/2,
\end{array}$$

and similarly

$$\begin{array}{rl}
& \mathrm{Prob}_{\omega\sim P}\{\widehat{g}(\omega) < G(x) - \bar{\rho} - \upsilon(x)\} = \mathrm{Prob}_{\omega\sim P}\left\{\frac{-\langle\bar{h},\omega\rangle}{\bar{\beta}} > \frac{-G(x)+\bar{\rho}+\bar{\varkappa}+\upsilon(x)}{\bar{\beta}}\right\} \\
\Rightarrow & \mathrm{Prob}_{\omega\sim P}\{\widehat{g}(\omega) < G(x) - \bar{\rho} - \upsilon(x)\} \leq \left[\int \mathrm{e}^{-\langle\bar{h},\omega\rangle/\bar{\beta}} P(d\omega)\right] \mathrm{e}^{-\frac{-G(x)+\bar{\rho}+\bar{\varkappa}+\upsilon(x)}{\bar{\beta}}} \\
& \leq \mathrm{e}^{\Phi(-\bar{h}/\bar{\beta}, \mathcal{A}(x))} \mathrm{e}^{\frac{G(x)-\bar{\rho}-\bar{\varkappa}-\upsilon(x)}{\bar{\beta}}} \\
\Rightarrow & \bar{\beta}\ln\left(\mathrm{Prob}_{\omega\sim P}\{\widehat{g}(\omega) < G(x) - \bar{\rho} - \upsilon(x)\}\right) \leq \bar{\beta}\Phi(-\bar{h}/\bar{\beta}, \mathcal{A}(x)) + G(x) - \bar{\rho} - \bar{\varkappa} - \upsilon(x) \\
& \leq \Psi_-(\bar{h},\bar{\beta}) - \bar{\rho} - \bar{\varkappa} \text{ [by definition of } \Psi_- \text{ and due to } x \in X] \\
& \leq \bar{\beta}\ln(\epsilon/2) \text{ [by } (b_2)] \\
\Rightarrow & \mathrm{Prob}_{\omega\sim P}\{\widehat{g}(\omega) < G(x) - \bar{\rho} - \upsilon(x)\} \leq \epsilon/2.
\end{array}$$

□

Corollary 3.3.1 *In the situation described in Section 3.3.1, let Φ satisfy the relation*

$$\Phi(0;\mu) \geq 0 \;\forall \mu \in \mathcal{M}. \tag{3.3.7}$$

Then

$$\begin{array}{rcll}\widehat{\Psi}_+(h) & := & \inf_\alpha\left\{\Psi_+(h,\alpha)+\alpha\ln(2/\epsilon):\alpha>0,(h,\alpha)\in\mathcal{H}^+\right\} & \\ & = & \sup_{x\in\mathcal{X}}\inf_{\alpha>0,(h,\alpha)\in\mathcal{H}^+}\left[\alpha\Phi(h/\alpha,\mathcal{A}(x))-G(x)-\upsilon(x)+\alpha\ln(2/\epsilon)\right], & (a)\\ \widehat{\Psi}_-(h) & := & \inf_\alpha\left\{\Psi_-(h,\alpha)+\alpha\ln(2/\epsilon):\alpha>0,(h,\alpha)\in\mathcal{H}^+\right\} & \\ & = & \sup_{x\in\mathcal{X}}\inf_{\alpha>0,(h,\alpha)\in\mathcal{H}^+}\left[\alpha\Phi(-h/\alpha,\mathcal{A}(x))+G(x)-\upsilon(x)+\alpha\ln(2/\epsilon)\right]. & (b)\end{array}\qquad(3.3.8)$$

and functions $\widehat{\Psi}_\pm:\mathcal{E}_H\to\mathbf{R}$ *are convex. Furthermore, let* $\bar{h}$, $\bar{\varkappa}$, $\widetilde{\rho}$ *be a feasible solution to the system of convex constraints*

$$\widehat{\Psi}_+(h)\leq\rho-\varkappa,\ \widehat{\Psi}_-(h)\leq\rho+\varkappa\qquad(3.3.9)$$

in variables h, ρ, $\varkappa$. *Then, setting*

$$\widehat{g}(\omega)=\langle\bar{h},\omega\rangle+\bar{\varkappa},$$

we get an estimate of $G(x)$, $x\in X$, *with* ϵ*-risk at most* $\widehat{\Psi}(\bar{h})$*:*

$$\mathrm{Risk}_\epsilon(\widehat{g}(\cdot)|G,X,\upsilon,\mathcal{A},\mathcal{H},\mathcal{M},\Phi)\leq\widehat{\Psi}(\bar{h}).\qquad(3.3.10)$$

Relation (3.3.9) *(and thus – the risk bound* (3.3.10)*) clearly holds true when* $\bar{h}$ *is a candidate solution to the convex optimization problem*

$$\mathrm{Opt}=\min_h\left\{\widehat{\Psi}(h):=\frac{1}{2}\left[\widehat{\Psi}_+(h)+\widehat{\Psi}_-(h)\right]\right\},\qquad(3.3.11)$$

$\bar{\rho}=\widehat{\Psi}(\bar{h})$, *and*

$$\bar{\varkappa}=\frac{\widehat{\Psi}_-(\bar{h})-\widehat{\Psi}_+(\bar{h})}{2}.$$

As a result, properly selecting $\bar{h}$, *we can make (an upper bound on) the* ϵ*-risk of estimate* $\widehat{g}(\cdot)$ *arbitrarily close to* Opt, *and equal to* Opt *when optimization problem* (3.3.11) *is solvable.*

Proof. Let us first verify the equalities in (3.3.8). The function

$$\Theta_+(h,\alpha;x)=\alpha\Phi(h/\alpha,\mathcal{A}(x))-G(x)-\upsilon(x)+\alpha\ln(2/\epsilon):\mathcal{H}^+\times\mathcal{X}\to\mathbf{R}$$

is convex-concave and continuous, and $\mathcal{X}$ is compact, whence by Sion-Kakutani Theorem

$$\begin{array}{rcl}\widehat{\Psi}_+(h) & := & \inf_\alpha\left\{\Psi_+(h,\alpha)+\alpha\ln(2/\epsilon):\alpha>0,(h,\alpha)\in\mathcal{H}^+\right\}\\ & = & \inf_{\alpha>0,(h,\alpha)\in\mathcal{H}^+}\max_{x\in\mathcal{X}}\Theta_+(h,\alpha;x)=\sup_{x\in\mathcal{X}}\inf_{\alpha>0,(h,\alpha)\in\mathcal{H}^+}\Theta_+(h,\alpha;x)\\ & = & \sup_{x\in\mathcal{X}}\inf_{\alpha>0,(h,\alpha)\in\mathcal{H}^+}\left[\alpha\Phi(h/\alpha,\mathcal{A}(x))-G(x)-\upsilon(x)+\alpha\ln(2/\epsilon)\right],\end{array}$$

as required in (3.3.8.a). As we know, $\Psi_+(h,\alpha)$ is real-valued continuous function on $\mathcal{H}^+$, so that $\widehat{\Psi}_+$ is convex on $\mathcal{E}_H$, provided that the function is real-valued. Now, let $\bar{x}\in\mathcal{X}$, and let e be a subgradient of $\phi(h)=\Phi(h;\mathcal{A}(x))$ taken at $h=0$. For $h\in\mathcal{E}_H$ and all $\alpha>0$ such that $(h,\alpha)\in\mathcal{H}^+$ we have

$$\begin{array}{rcl}\Psi_+(h,\alpha) & \geq & \alpha\Phi(h/\alpha;\mathcal{A}(\bar{x}))-G(\bar{x})-\upsilon(\bar{x})+\alpha\ln(2/\epsilon)\\ & \geq & \alpha[\Phi(0;\mathcal{A}(\bar{x}))+\langle e,h/\alpha\rangle]-G(\bar{x})-\upsilon(\bar{x})+\alpha\ln(2/\epsilon)\geq\langle e,h\rangle-G(\bar{x})-\upsilon(\bar{x})\end{array}$$

(we have used (3.3.7)), and therefore $\Psi_+(h,\alpha)$ is bounded below on the set $\{\alpha>0:h/\alpha\in\mathcal{H}\}$; in addition, this set is nonempty, since $\mathcal{H}$ contains a neighbourhood of the origin. Thus, $\widehat{\Psi}_+$ is real-valued and convex on $\mathcal{E}_H$. Verification of (3.3.8.b) and of the fact that $\widehat{\Psi}_-(h)$ is real-valued convex function on $\mathcal{E}_H$ is completely similar.

Now, given a feasible solution $(\bar{h}, \bar{\varkappa}, \widetilde{\rho})$ to (3.3.9), let us select somehow $\bar{\rho} > \widetilde{\rho}$. Taking into account the definition of $\widehat{\Psi}_{\pm}$, we can find $\bar{\alpha}$ and $\bar{\beta}$ such that

$$\begin{array}{l}(\bar{h}, \bar{\alpha}) \in \mathcal{H}^+ \ \& \ \Psi_+(\bar{h}, \bar{\alpha}) + \bar{\alpha}\ln(2/\epsilon) \leq \bar{\rho} - \bar{\varkappa},\\ (\bar{h}, \bar{\beta}) \in \mathcal{H}^+ \ \& \ \Psi_-(\bar{h}, \bar{\beta}) + \bar{\beta}\ln(2/\epsilon) \leq \bar{\rho} + \bar{\varkappa},\end{array}$$

implying that the collection $(\bar{h}, \bar{\alpha}, \bar{\beta}, \bar{\varkappa}, \bar{\rho})$ is a feasible solution to (3.3.6). Invoking Proposition 3.3.1, we get

$$\mathrm{Prob}_{\omega\sim P}\left\{\omega : |\widehat{g}(\omega) - G(x)| > \bar{\rho} + \upsilon(x)\right\} \leq \epsilon$$

for all $(x \in X, P \in \mathcal{P})$ satisfying (3.3.2). Since $\bar{\rho}$ can be selected arbitrarily close to $\widetilde{\rho}$, $\widehat{g}(\cdot)$ indeed is a $(\widetilde{\rho}, \epsilon, \upsilon(\cdot))$-accurate estimate. □

3.3.3 Estimation from repeated observations

Assume that in the situation described in section 3.3.1 we have access to K observations $\omega_1, ..., \omega_K$ sampled, independently of each other, from a probability distribution P, and are allowed to build our estimate based on these K observations rather than on a single observation. We can immediately reduce this new situation to the previous one, just by redefining the data. Specifically, given initial data $\mathcal{H} \subset \mathcal{E}_H$, $\mathcal{M} \subset \mathcal{E}_M$, $\Phi(\cdot;\cdot) : \mathcal{H} \times \mathcal{M} \to \mathbf{R}$, $X \subset \mathcal{X} \subset \mathcal{E}_X$, $\mathcal{A}(\cdot)$, $G(x) = g^Tx + c$, see section 3.3.1 and a positive integer K, let us update part of the data, specifically, replace $\mathcal{H} \subset \mathcal{E}_{\mathcal{H}}$ with $\mathcal{H}^K := \underbrace{\mathcal{H} \times ... \times \mathcal{H}}_{K} \subset \mathcal{E}_H^K := \underbrace{\mathcal{E}_H \times ... \times \mathcal{E}_H}_{K}$ and replace $\Phi(\cdot, \cdot) : \mathcal{H} \times \mathcal{M} \to \mathbf{R}$ with $\Phi^K(h^K = (h_1, ..., h_K); \mu) = \sum_{i=1}^K \Phi(h_i; \mu) : \mathcal{H}^K \times \mathcal{M} \to \mathbf{R}$. It is immediately seen that the updated data satisfy all requirements imposed on the data in section 3.3.1, and that whenever a Borel probability distribution $\mathcal{P}$ on $\mathcal{E}_{\mathcal{H}}$ and $x \in X$ are linked by (3.3.2), the distribution P^K of K-element i.i.d. sample $\omega^K = (\omega_1, ..., \omega_K)$ drawn from P and x are linked by the relation

$$\begin{array}{rcl}\forall h^K = (h_1, ..., h_K) \in \mathcal{H}^K : \ln\left(\int_{\mathcal{E}_H^K} \mathrm{e}^{\langle h^K, \omega^K\rangle} P^K(d\omega^K)\right) & = & \sum_i \ln\left(\int_{\mathcal{E}_H} \mathrm{e}^{\langle h_i, \omega_i\rangle} P(d\omega_i)\right)\\ & \leq & \Phi^K(h^K; \mathcal{A}(x)).\end{array} \tag{3.3.12}$$

Applying to our new data the construction from section 3.3.2, we arrive at "repeated observations" versions of Proposition 3.3.1 and Corollary 3.3.1. Note that the resulting convex constraints/objectives are symmetric w.r.t. permutations functions of the components $h_1, ..., h_K$ of h^K, implying that we lose nothing when restricting ourselves with collections h^K with equal to each other components; it is convenient to denote the common value of these components h/K. With these observations, Proposition 3.3.1 and Corollary 3.3.1 become the statements as follows (we use the assumptions and the notation from the previous sections):

Proposition 3.3.2 *Given $\epsilon \in (0, 1)$ and positive integer K, let*

$$\begin{array}{lrcl}(a) & \Psi_+(h, \alpha) & = & \sup\limits_{x\in\mathcal{X}} \left[\alpha\Phi(h/\alpha, \mathcal{A}(x)) - G(x) - \upsilon(x)\right] : \mathcal{H}^+ \to \mathbf{R},\\ (b) & \Psi_-(h, \beta) & = & \sup\limits_{x\in\mathcal{X}} \left[\beta\Phi(-h/\beta, \mathcal{A}(x)) + G(x) - \upsilon(x)\right] : \mathcal{H}^+ \to \mathbf{R},\end{array}$$

and let $\bar{h}$, $\bar{\alpha}$, $\bar{\beta}$, $\bar{\varkappa}$, $\bar{\rho}$ be a feasible solution to the system of convex constraints

$$\begin{array}{lrcl}(a_1) & (h, \alpha) & \in & \mathcal{H}^+\\ (a_2) & (h, \beta) & \in & \mathcal{H}^+\\ (b_1) & \alpha K^{-1}\ln(\epsilon/2) & \geq & \Psi_+(h, \alpha) - \rho + \varkappa\\ (b_2) & \beta K^{-1}\ln(\epsilon/2) & \geq & \Psi_-(h, \beta) - \rho - \varkappa\end{array} \tag{3.3.13}$$

in variables h, α, β, ρ, $\varkappa$. Setting

$$\widehat{g}(\omega^K) = \langle \bar{h}, \frac{1}{K}\sum_{i=1}^K \omega_i\rangle + \bar{\varkappa},$$

we get an estimate of $G(x)$ via independent K-repeated observations

$$\omega_i \sim P, \, i = 1, ..., K$$

with ϵ-risk on X not exceeding $\bar{\rho}$, meaning that whenever $x \in X$ and a Borel probability distribution P on $\mathcal{E}_\mathcal{H}$ are linked by (3.3.2), one has

$$\mathrm{Prob}_{\omega^K\sim P^K}\left\{\omega^K : |\widehat{g}(\omega^K) - G(x)| > \bar{\rho} + \upsilon(x)\right\} \leq \epsilon. \tag{3.3.14}$$

Corollary 3.3.2 *In the situation described in the beginning of section 3.3.1, let Φ satisfy the relation (3.3.7), and let a positive integer K be given. Then*

$$\begin{array}{rcll}
\widehat{\Psi}_+(h) & := & \inf_\alpha \left\{\Psi_+(h,\alpha) + K^{-1}\alpha\ln(2/\epsilon) : \alpha > 0, (h,\alpha) \in \mathcal{H}^+\right\} & \\
& = & \sup_{x\in\mathcal{X}} \inf_{\alpha>0,(h,\alpha)\in\mathcal{H}^+} \left[\alpha\Phi(h/\alpha, \mathcal{A}(x)) - G(x) - \upsilon(x) + K^{-1}\alpha\ln(2/\epsilon)\right], & (a)\\
\widehat{\Psi}_-(h) & := & \inf_\alpha \left\{\Psi_-(h,\alpha) + K^{-1}\alpha\ln(2/\epsilon) : \alpha > 0, (h,\alpha) \in \mathcal{H}^+\right\} & \\
& = & \sup_{x\in\mathcal{X}} \inf_{\alpha>0,(h,\alpha)\in\mathcal{H}^+} \left[\alpha\Phi(-h/\alpha, \mathcal{A}(x)) + G(x) - \upsilon(x) + K^{-1}\alpha\ln(2/\epsilon)\right]. & (b)
\end{array} \tag{3.3.15}$$

and functions $\widehat{\Psi}_\pm : \mathcal{E}_H \to \mathbf{R}$ are convex. Furthermore, let $\bar{h}$, $\bar{\varkappa}$, $\widetilde{\rho}$ be a feasible solution to the system of convex constraints

$$\widehat{\Psi}_+(h) \leq \rho - \varkappa, \; \widehat{\Psi}_-(h) \leq \rho + \varkappa \tag{3.3.16}$$

in variables h, ρ, $\varkappa$. Then, setting

$$\widehat{g}(\omega^K) = \langle \bar{h}, \frac{1}{K}\sum_{i=1}^K \omega_i\rangle + \bar{\varkappa},$$

we get an estimate of $G(x)$, $x \in X$, with ϵ-risk at most $\widehat{\Psi}(\bar{h})$, meaning that whenever $x \in X$ and a Borel probability distribution P on $\mathcal{E}_\mathcal{H}$ are linked by (3.3.2), relation (3.3.14) holds true.

Relation (3.3.16) clearly holds true when $\bar{h}$ is a candidate solution to the convex optimization problem

$$\mathrm{Opt} = \min_h \left\{\widehat{\Psi}(h) := \frac{1}{2}\left[\widehat{\Psi}_+(h) + \widehat{\Psi}_-(h)\right]\right\}, \tag{3.3.17}$$

$\bar{\rho} = \widehat{\Psi}(\bar{h})$ and

$$\bar{\varkappa} = \frac{\widehat{\Psi}_-(\bar{h}) - \widehat{\Psi}_+(\bar{h})}{2}.$$

As a result, properly selecting $\bar{h}$, we can make (an upper bound on) the ϵ-risk of estimate $\widehat{g}(\cdot)$ arbitrarily close to Opt*, and equal to* Opt *when optimization problem (3.3.17) is solvable.*

From now on, if otherwise is not explicitly stated, we deal with K-repeated observations; to get back to single-observation case, it suffices to set $K = 1$.

3.3.4 Application: Estimating linear form of sub-Gaussianity parameters

Consider the simplest case of the situation from sections 3.3.1, 3.3.3, where

- $\mathcal{H} = \mathcal{E}_H = \mathbf{R}^d$, $\mathcal{M} = \mathcal{E}_M = \mathbf{R}^d \times \mathbf{S}^d_+$, $\Phi(h;\mu,M) = h^T\mu + \frac{1}{2}h^TMh : \mathbf{R}^d \times (\mathbf{R}^d \times \mathbf{S}^d_+) \to \mathbf{R}$, so that $\mathcal{S}[\mathcal{H},\mathcal{M},\Phi]$ is the family of all sub-Gaussian distributions on $\mathbf{R}^d$;
- $X = \mathcal{X} \subset \mathcal{E}_X = \mathbf{R}^{n_x}$ is a nonempty convex compact set, and
- $\mathcal{A}(x) = (Ax + a, M(x))$, where A is $d \times n_x$ matrix, and $M(x)$ is affinely depending on x symmetric $d \times d$ matrix such that $M(x)$ is $\succeq 0$ when $x \in X$,
- $\upsilon(x)$ is a convex continuous function on $\mathcal{X}$,
- $G(x)$ is an affine function on $\mathcal{E}_X$.

In the case in question (3.3.7) clearly takes place, and the left hand sides in the constraints (3.3.16) are

$$\begin{array}{rcl}
\widehat{\Psi}_+(h) &=& \sup_{x\in X}\inf_{\alpha>0}\left\{h^T[Ax+a] + \frac{1}{2\alpha}h^TM(x)h + K^{-1}\alpha\ln(2/\epsilon) - G(x) - \upsilon(x)\right\}\\
&=& \max_{x\in X}\left\{\sqrt{2K^{-1}\ln(2/\epsilon)[h^TM(x)h]} + h^T[Ax+a] - G(x) - \upsilon(x)\right\},\\
\widehat{\Psi}_-(h) &=& \sup_{x\in X}\inf_{\alpha>0}\left\{-h^T[Ax+a] + \frac{1}{2\alpha}h^TM(x)h + K^{-1}\alpha\ln(2/\epsilon) + G(x) - \upsilon(x)\right\}\\
&=& \max_{x\in X}\left\{\sqrt{2K^{-1}\ln(2/\epsilon)[h^TM(x)h]} - h^T[Ax+a] + G(x) - \upsilon(x)\right\}.
\end{array}$$

Thus, system (3.3.16) reads

$$\begin{array}{rcl}
a^Th + \max\limits_{x\in X}\left[\sqrt{2K^{-1}\ln(2/\epsilon)[h^TM(x)h]} + h^TAx - G(x) - \upsilon(x)\right] &\leq& \rho - \varkappa,\\
-a^Th + \max\limits_{x\in X}\left[\sqrt{2K^{-1}\ln(2/\epsilon)[h^TM(x)h]} - h^TAx + G(x) - \upsilon(x)\right] &\leq& \rho + \varkappa.
\end{array}$$

We arrive at the following version of Corollary 3.3.2:

Proposition 3.3.3 *In the situation described in the beginning of section 3.3.4, given $\epsilon \in (0,1)$, let $\bar{h}$ be a feasible solution to the convex optimization problem*

$$\mathrm{Opt} = \min_{h\in\mathbf{R}^d}\left\{\widehat{\Psi}(h) := \frac{1}{2}\left[\begin{array}{l}\overbrace{\max\limits_{x\in X}\left[\sqrt{2K^{-1}\ln(2/\epsilon)[h^TM(x)h]} + h^TAx - G(x) - \upsilon(x)\right] + a^Th}^{\widehat{\Psi}_+(h)}\\ \underbrace{+\max\limits_{y\in X}\left[\sqrt{2K^{-1}\ln(2/\epsilon)[h^TM(y)h]} - h^TAy + G(y) - \upsilon(y)\right] - a^Th}_{\widehat{\Psi}_-(h)}\end{array}\right]\right\}. \tag{3.3.18}$$

Then, setting

$$\bar{\varkappa} = \frac{1}{2}\left[\widehat{\Psi}_-(\bar{h}) - \widehat{\Psi}_+(\bar{h})\right],\ \bar{\rho} = \widehat{\Psi}(\bar{h}), \tag{3.3.19}$$

the affine estimate

$$\widehat{g}(\omega^K) = \frac{1}{K}\sum_{i=1}^K \bar{h}^T\omega_i + \bar{\varkappa}$$

has ϵ-risk, taken w.r.t. the data listed in the beginning of this section, at most $\bar{\rho}$.

It is immediately seen that optimization problem (3.3.18) is solvable, provided that

$$\bigcap_{x\in X}\mathrm{Ker}(M(x))=\{0\},$$

and an optimal solution h_* to the problem, taken along with

$$\varkappa_*=\frac{1}{2}\left[\widehat{\Psi}_-(h_*)-\widehat{\Psi}_+(h_*)\right], \tag{3.3.20}$$

yields the affine estimate

$$\widehat{g}_*(\omega)=\frac{1}{K}\sum_{i=1}^{K}h_*^T\omega_i+\varkappa_*$$

with ϵ-risk, taken w.r.t. the data listed in the beginning of this section, at most Opt.

3.3.4.1 Consistency

Assuming $\upsilon(x)\equiv 0$, we can easily answer the natural question "when the proposed estimation scheme is consistent", meaning that for every $\epsilon\in(0,1)$, it allows to achieve arbitrarily small ϵ-risk, provided that K is large enough. Specifically, denoting by g^Tx the linear part of $G(x)$: $G(x)=g^Tx+c$, from Proposition 3.3.3 it is immediately seen that a sufficient condition for consistency is the existence of $\bar{h}\in\mathbf{R}^d$ such that $\bar{h}^TAx=g^Tx$ for all $x\in\mathcal{X}-\mathcal{X}$, or, equivalently, the condition that g is orthogonal to the intersection of the kernel of A with the linear span of $\mathcal{X}-\mathcal{X}$. Indeed, under this assumption, for every fixed $\epsilon\in(0,1)$ we clearly have $\lim_{K\to\infty}\widehat{\Phi}(\bar{h})=0$, implying that $\lim_{K\to\infty}\mathrm{Opt}=0$, with $\widehat{\Psi}$ and Opt given by (3.3.18). Still assuming $\upsilon(x)\equiv 0$, the condition in question is necessary for consistency as well, since when the condition is violated, we have $Ax'=Ax''$ for properly selected $x',x''\in\mathcal{X}$ with $G(x')\neq G(x'')$, making low risk recovery of $G(x)$, $x\in\mathcal{X}$, impossible already in the case of zero noisy component in observations[6].

3.3.4.2 Direct product case

Further simplifications are possible in the *direct product case*, where, in addition to what was assumed in the beginning of section 3.3.4,

- $\mathcal{E}_X=\mathcal{E}_U\times\mathcal{E}_V$ and $X=U\times V$, with convex compact sets $U\subset\mathcal{E}_U=\mathbf{R}^{n_u}$ and $V\subset E_V=\mathbf{R}^{n_v}$,
- $\mathcal{A}(x=(u,v))=[Au+a,M(v)]:U\times V\to\mathbf{R}^d\times\mathbf{S}^d$, with $M(v)\succeq 0$ for $v\in V$,
- $G(x=(u,v))=g^Tu+c$ depends solely on u, and
- $\upsilon(x=(u,v))=\varrho(u)$ depends solely on u.

It is immediately seen that in the direct product case problem (3.3.18) reads

$$\mathrm{Opt}=\min_{h\in\mathbf{R}^d}\left\{\frac{\phi_U(A^Th-g)+\phi_U(-A^Th+g)}{2}+\max_{v\in V}\sqrt{2K^{-1}\ln(2/\epsilon)h^TM(v)h}\right\}, \tag{3.3.21}$$

where

$$\phi_U(f)=\max_{u\in U}\left[u^Tf-\varrho(u)\right]. \tag{3.3.22}$$

[6] Note that in Gaussian case with $M(x)$ depending on x the above condition is, in general, not necessary for consistency, since a nontrivial information on x (and thus on $G(x)$) can, in principle, be extracted from the covariance matrix $M(x)$ which can be estimated from observations.

Assuming $\bigcap_{v\in V}\mathrm{Ker}(M(v))=\{0\}$, the problem is solvable, and its optimal solution h_* produces affine estimate

$$\widehat{g}_*(\omega^K)=\frac{1}{K}\sum_i h_*^T\omega_i+\varkappa_*,\ \varkappa_*=\frac{1}{2}[\phi_U(-A^Th+g)-\phi_U(A^Th-g)]-a^Th_*+c$$

with ϵ-risk $\leq$ Opt.

Near-optimality In addition to the assumption that we are in the direct product case, assume that $v(\cdot)\equiv 0$ and, for the sake of simplicity, that $M(v)\succ 0$ whenever $v\in V$. In this case (3.3.18) reads

$$\mathrm{Opt}=\min_h\max_{v\in V}\left\{\Theta(h,v):=\frac{1}{2}[\phi_U(A^Th-g)+\phi_U(-A^Th+g)]+\sqrt{2K^{-1}\ln(2/\epsilon)h^TM(v)h}\right\},$$

whence, taking into account that $\Theta(h,v)$ clearly is convex in h and concave in v, while V is a convex compact set, by Sion-Kakutani Theorem we get also

$$\mathrm{Opt}=\max_{v\in V}\left[\mathrm{Opt}(v)=\min_h\frac{1}{2}[\phi_U(A^Th-g)+\phi_U(-A^Th+g)]+\sqrt{2K^{-1}\ln(2/\epsilon)h^TM(v)h}\right].\tag{3.3.23}$$

Now consider the problem of recovering g^Tu from observation ω_i, $i\leq K$, independently of each other sampled from $\mathcal{N}(Au+a,M(v))$, where unknown u is known to belong to U and $v\in V$ is known. Let $\rho_\epsilon(v)$ be the minimax ϵ-risk of the recovery:

$$\rho_\epsilon(v)=\inf_{\widehat{g}(\cdot)}\left\{\rho:\mathrm{Prob}_{\omega^K\sim[\mathcal{N}(Au+a,M(v))]^K}\{\omega^K:|\widehat{g}(\omega^K)-g^Tu|>\rho\}\leq\epsilon\ \forall u\in U\right\},$$

where inf is taken over all Borel functions $\widehat{g}(\cdot):\mathbf{R}^{Kd}\to\mathbf{R}$. Invoking [92, Theorem 3.1], it is immediately seen that whenever $\epsilon<1/4$, one has

$$\rho_\epsilon(v)\geq\left[\frac{2\ln(2/\epsilon)}{\ln\left(\frac{1}{4\epsilon}\right)}\right]^{-1}\mathrm{Opt}(v).$$

Since the family $\mathcal{SG}(\mathcal{U},\mathcal{V})$ of all sub-Gaussian, with parameters $(Au+a,M(v))$, $u\in U$, $v\in V$, distributions on $\mathbf{R}^d$ contains all Gaussian distributions $\mathcal{N}(Au+a,M(v))$ induced by $(u,v)\in U\times V$, we arrive at the following conclusion:

Proposition 3.3.4 *In the just described situation, the minimax optimal ϵ-risk*

$$\mathrm{Risk}_\epsilon^{\mathrm{opt}}(K)=\inf_{\widehat{g}(\cdot)}\mathrm{Risk}_\epsilon(\widehat{g}(\cdot)),$$

of recovering g^Tu from K-repeated i.i.d. sub-Gaussian, with parameters $(Au+a,M(v))$, $(u,v)\in U\times V$, random observations is within a moderate factor of the upper bound Opt *on the ϵ-risk, taken w.r.t. the same data, of the affine estimate $\widehat{g}_*(\cdot)$ yielded by an optimal solution to* (3.3.21)*, namely,*

$$\mathrm{Opt}\leq\frac{2\ln(2/\epsilon)}{\ln\left(\frac{1}{4\epsilon}\right)}\mathrm{Risk}_\epsilon^{\mathrm{opt}}.$$

3.3.4.3 Numerical illustration

The numerical illustration we are about to discuss models the situation when we want to recover a linear form of a signal x known to belong to a given convex compact subset X via indirect observations Ax affected by sub-Gaussian "relative noise," meaning that the variance of observation is the larger the larger is the signal. Specifically, our observation is

$$\omega \sim \mathcal{SG}(Ax, M(x)),$$

where

$$x \in X = \left\{x \in \mathbf{R}^n : 0 \leq x_j \leq j^{-\alpha}, 1 \leq j \leq n\right\},\ M(x) = \sigma^2 \sum_{j=1}^n x_j \Theta_j \tag{3.3.24}$$

where $A \in \mathbf{R}^{d\times n}$ and $\Theta_j \in \mathbf{S}^d_+$, $j = 1, ..., n$, are given matrices; the linear form to be recovered from observation ω is $G(x) = g^T x$. The entities $g, A, \{\Theta_j\}_{j=1}^n$ and reals $\alpha \geq 0$ ("degree of smoothness"), $\sigma > 0$ ("noise intensity") are parameters of the estimation problem we intend to process. The parameters g, A, Θ_j were generated as follows:

- $g \geq 0$ was selected at random and then normalized to have $\max\limits_{x\in X} g^T x = \max\limits_{x,y\in X} g^T[x-y] = 2$;
- we dealt with $n > d$ ("deficient observations"); the d nonzero singular values of A were set to $\theta^{-\frac{i-1}{d-1}}$, where "condition number" $\theta \geq 1$ is a parameter; the orthonormal systems U and V of the first d left, respectively, right singular vectors of A were drawn at random from rotationally invariant distributions;
- the positive semidefinite $d \times d$ matrices Θ_j were orthogonal projectors on randomly selected subspaces in $\mathbf{R}^d$ of dimension $\lfloor d/2 \rfloor$;
- in all our experiments, we dealt with single-observation case $K = 1$, and used $\upsilon(\cdot) \equiv 0$.

Note that X possesses $\geq$-largest point $\bar{x}$, whence $M(x) \preceq M(\bar{x})$ whenever $x \in X$; as a result, sub-Gaussian distributions with matrix parameter $M(x)$, $x \in X$, can be thought also to have matrix parameter $M(\bar{x})$. One of the goals of experiment to be reported was to understand how much would be lost were we replacing $M(\cdot)$ with $\widehat{M}(x) \equiv M(\bar{x})$, that is, were we ignoring the fact that small signals result in low-noise observations.

In the experiment to be reported, we use $d = 32$, $m = 48$, $\alpha = 2$, $\theta = 2$, and $\sigma = 0.01$. Utilizing these parameters, we generated at random, as described above, 10 collections $\{g, A, \Theta_j, j \leq d\}$, thus arriving at 10 estimation problems. For every one of these problems, we used the outlined machinery to build affine in ω estimate of $g^T x$ as yielded by optimal solution to (3.3.18), and computed upper bound Opt on ($\epsilon = 0.01$)-risk of this estimate. In fact, for every one of the 10 generated estimation problems, we build two estimates and two risk bounds: the first – for the problem "as is," and the second – for the aforementioned "direct product envelope" of the problem, where the mapping $x \mapsto M(x)$ is replaced with $x \mapsto \widehat{M}(x) := M(\bar{x})$. The results are as follows:

min	median	mean	max
0.138	0.190	0.212	0.299
0.150	0.210	0.227	0.320

0.01-Risk, data over 10 estimation problems $[d = 32, m = 48, \alpha = 2, \theta = 2, \sigma = 0.1]$
First row: $\omega \sim \mathcal{SG}(Ax, M(x))$. Second row: $\omega \sim \mathcal{SG}(Ax, M(\bar{x}))$

Pay attention to "amplification of noise" in the estimate (about 20 times the level σ of observation noise) and significant variability of risk across the experiments; seemingly, both these phenomena stem from the fact that we have highly deficient observations ($n/d = 1.5$) combined with "random interplay" between the directions of coordinate axes in $\mathbf{R}^m$ (along these directions, X becomes more and more thin) and the orientation of the 16-dimensional kernel of A.

3.4 Estimating quadratic forms via quadratic lifting

In the situation of Section 3.3.1, passing from "original" observations (3.3.1) to their quadratic lifting: we can use the just developed machinery to estimate quadratic forms of the underlying parameters rather than linear ones. We are about to investigate the related possibilities in the cases of Gaussian and sub-Gaussian observations.

3.4.1 Estimating quadratic forms, Gaussian case

3.4.1.1 Preliminaries

Consider the situation where we are given

- a nonempty bounded set U in $\mathbf{R}^m$;
- a nonempty convex compact subset $\mathcal{V}$ of the positive semidefinite cone $\mathbf{S}^d_+$;
- a matrix $\Theta_* \succ 0$ such that $\Theta_* \succeq \Theta$ for all $\Theta \in \mathcal{V}$;
- an affine mapping $u \mapsto A[u;1] : \mathbf{R}^m \to \Omega = \mathbf{R}^d$, where A is a given $d \times (m+1)$ matrix,
- a convex continuous function $\varrho(\cdot)$ on $\mathbf{S}^{m+1}_+$.

A pair $(u \in U, \Theta \in \mathcal{V})$ specifies Gaussian random vector $\zeta \sim \mathcal{N}(A[u;1], \Theta)$ and thus specifies probability distribution $P[u,\Theta]$ of $(\zeta, \zeta\zeta^T)$. Let $\mathcal{Q}(U,\mathcal{V})$ be the family of probability distributions on $\Omega = \mathbf{R}^d \times \mathbf{S}^d$ stemming in this fashion from Gaussian distributions with parameters from $U \times \mathcal{V}$. Our goal is to cover the family $\mathcal{Q}(U,\mathcal{V})$ by a family of the type $\mathcal{S}[N, \mathcal{M}, \Phi]$.

It is convenient to represent a linear form on $\Omega = \mathbf{R}^d \times \mathbf{S}^d$ as

$$h^Tz + \frac{1}{2}\mathrm{Tr}(HZ),$$

where $(h,H) \in \mathbf{R}^d \times \mathbf{S}^d$ is the "vector of coefficients" of the form, and $(z,Z) \in \mathbf{R}^d \times \mathbf{S}^d$ is the argument of the form.

We assume that for some $\delta \in [0,2]$ it holds

$$\|\Theta^{1/2}\Theta_*^{-1/2} - I\| \leq \delta \;\forall \Theta \in \mathcal{V}, \tag{3.4.1}$$

where $\|\cdot\|$ is the spectral norm (cf. (2.8.35). Finally, we set $B = \left[\begin{array}{c} A \\ b^T \end{array}\right]$ and

$$\mathcal{Z}^+ = \{W \in \mathbf{S}^{m+1}_+ : W_{m+1,m+1} = 1\}. \tag{3.4.2}$$

The statement below is a straightforward reformulation of Proposition 2.8.7.i:

Proposition 3.4.1 *In the just described situation, let us select $\gamma \in (0,1)$ and set*

$$\begin{array}{rcl}
\mathcal{H} & = & \mathcal{H}_\gamma := \{(h,H) \in \mathbf{R}^d \times \mathbf{S}^d : -\gamma\Theta_*^{-1} \preceq H \preceq \gamma\Theta_*^{-1}\},\\
\mathcal{M}^+ & = & \mathcal{V} \times \mathcal{Z}^+,\\
\Phi(h,H;\Theta,Z) & = & -\frac{1}{2}\ln\mathrm{Det}(I - \Theta_*^{1/2}H\Theta_*^{1/2}) + \frac{1}{2}\mathrm{Tr}([\Theta - \Theta_*]H) + \frac{\delta(2+\delta)}{2(1-\|\Theta_*^{1/2}H\Theta_*^{1/2}\|)}\|\Theta_*^{1/2}H\Theta_*^{1/2}\|_H^2\\
& & +\Gamma(h,H;Z) : \mathcal{H} \times \mathcal{M}^+ \to \mathbf{R}[\|\cdot\| \text{ is the spectral, and } \|\cdot\|_H \text{ is the Frobenius norm}],\\
\Gamma(h,H;Z) & = & \frac{1}{2}\mathrm{Tr}\left(Z[bh^TA + A^Thb^T + A^THA + B^T[H,h]^T[\Theta_*^{-1} - H]^{-1}[H,h]B]\right)\\
& = & \frac{1}{2}\mathrm{Tr}\left(Z\left(B^T\left[\left[\begin{array}{c|c} H & h \\ \hline h^T & \end{array}\right] + [H,h]^T\left[\Theta_*^{-1} - H\right]^{-1}[H,h]\right]B\right)\right).
\end{array} \tag{3.4.3}$$

Then $\mathcal{H}, \mathcal{M}^+, \Phi$ form a regular data, and for every $(u,\Theta) \in \mathbf{R}^m \times \mathcal{V}$ it holds

$$\forall (h,H) \in \mathcal{H} : \ln\left(\mathbf{E}_{\zeta\sim\mathcal{N}(\mathcal{C}(u),\Theta)}\left\{e^{h^T\zeta + \frac{1}{2}\zeta^T H\zeta}\right\}\right) \leq \Phi(h,H;\Theta,[u;1][u;1]^T). \tag{3.4.4}$$

Besides this, function $\Phi(h,H;\Theta,Z)$ is coercive in the convex argument: whenever $(\Theta,Z) \in \mathcal{M}$ and $(h_i,H_i) \in \mathcal{H}$ and $\|(h_i,H_i)\| \to \infty$ as $i \to \infty$, we have $\Phi(h_i,H_i;\Theta,Z) \to \infty$, $i \to \infty$.

3.4.1.2 Estimating quadratic form: Situation & goal

We are interested in the situation as follows: we are given a sample $\zeta^K = (\zeta_1, ..., \zeta_K)$ of independent across i and identically distributed random observations

$$\zeta_i \sim \mathcal{N}(A[u;1], M(v)),\ 1 \leq i \leq K, \tag{3.4.5}$$

where

- (u,v) is unknown “signal” known to belong to a given set $U \times V$, where
  - $U \subset \mathbf{R}^m$ is a compact set, and
  - $V \subset \mathbf{R}^k$ is a compact convex set;
- A is a given $d \times (m+1)$ matrix, and $v \mapsto M(v) : \mathbf{R}^k \to \mathbf{S}^d$ is affine mapping such that $M(v) \succeq 0$ whenever $v \in V$.

We are also given a convex calibrating function $\varrho(Z) : \mathbf{S}^{m+1}_+ \to \mathbf{R}$ and “functional of interest”

$$F(u,v) = [u;1]^T Q[u;1] + q^T v, \tag{3.4.6}$$

where Q and q are known $(m+1)\times(m+1)$ symmetric matrix and k-dimensional vector, respectively. Our goal is to recover $F(u,v)$, for unknown (u,v) known to belong to $U \times V$, via observation (3.4.5). Given a tolerance $\epsilon \in (0,1)$, we quantify the quality of a candidate estimate $\widehat{g}(\zeta^K)$ of $F(u,v)$ by the smallest ρ such that for all $(u,v) \in U \times V$ it holds

$$\mathrm{Prob}_{\zeta^K\sim\mathcal{N}(A[u;1],M(v))\times\ldots\times\mathcal{N}(A[u;1],M(v))}\left\{|\widehat{g}(\zeta^K) - F(u,v)\,\rho + \varrho([u;1][u;1]^T)\right\} \leq \epsilon. \tag{3.4.7}$$

3.4.1.3 Construction & Result

Let

$$\mathcal{V} = \{M(v) : v \in V\},$$

so that $\mathcal{V}$ is a convex compact subset of the positive semidefinite cone $\mathbf{S}^{d+1}_+$. Let us select somehow

1. a matrix $\Theta_* \succ 0$ such that $\Theta_* \succeq \Theta$, for all $\Theta \in \mathcal{V}$;
2. a convex compact subset $\mathcal{Z}$ of the set $\mathcal{Z}^+ = \{Z \in \mathbf{S}^{m+1}_+ : Z_{m+1,m+1} = 1\}$ such that $[u;1][u;1]^T \in \mathcal{Z}$ for all $u \in U$;
3. a real $\gamma \in (0,1)$ and a nonnegative real δ such that (3.4.1) takes place.

We further set (cf. Proposition 3.4.1)

$$
\begin{array}{rcl}
B &=& \left[\begin{array}{c} A \\ {[0,...,0,1]} \end{array}\right] \in \mathbf{R}^{(d+1)\times(m+1)},\\
\mathcal{H} &=& \mathcal{H}_\gamma := \{(h,H)\in\mathbf{R}^d\times\mathbf{S}^d : -\gamma\Theta_*^{-1}\preceq H\preceq\gamma\Theta_*^{-1}\},\\
\mathcal{M} &=& \mathcal{V}\times\mathcal{Z},\\
\Phi(h,H;\Theta,Z) &=& -\frac{1}{2}\ln\mathrm{Det}(I-\Theta_*^{1/2}H\Theta_*^{1/2})+\frac{1}{2}\mathrm{Tr}([\Theta-\Theta_*]H)+\frac{\delta(2+\delta)}{2(1-\|\Theta_*^{1/2}H\Theta_*^{1/2}\|)}\|\Theta_*^{1/2}H\Theta_*^{1/2}\|_H^2\\
&&+\Gamma(h,H;Z) : \mathcal{H}\times\mathcal{M}\to\mathbf{R}[\|\cdot\| \text{ is the spectral, and } \|\cdot\|_H \text{ is the Frobenius norm}],\\
\Gamma(h,H;Z) &=& \frac{1}{2}\mathrm{Tr}\left(Z[bh^TA+A^Thb^T+A^THA+B^T[H,h]^T[\Theta_*^{-1}-H]^{-1}[H,h]B]\right)\\
&=& \frac{1}{2}\mathrm{Tr}\left(Z\left(B^T\left[\left[\begin{array}{c|c} H & h\\ \hline h^T & \end{array}\right]+[H,h]^T\left[\Theta_*^{-1}-H\right]^{-1}[H,h]\right]B\right)\right)
\end{array}
\tag{3.4.8}
$$

and treat, as our observation, the quadratic lift of observation (3.4.5), that is, our observation is

$$
\omega^K=\{\omega_i=(\zeta_i,\zeta_i\zeta_i^T)\}_{i=1}^K, \zeta_i\sim\mathcal{N}(A[u;1],M(v)) \text{ are independent across } i. \tag{3.4.9}
$$

Note that by Proposition 3.4.1, function $\Phi(h,H;\Theta,Z):\mathcal{H}\times\mathcal{M}\to\mathbf{R}$ is continuous convex-concave function which is coercive in convex argument and is such that

$$
\forall (u\in U, v\in V, (h,H)\in\mathcal{H}) : \ln\left(\mathbf{E}_{\zeta\sim\mathcal{N}(A[u;1],M(v))}\left\{\mathrm{e}^{\frac{1}{2}\zeta^TH\zeta+h^T\zeta}\right\}\right)\le\Phi(h,H;M(v),[u;1][u;1]^T). \tag{3.4.10}
$$

We are about to demonstrate that as far as estimating the functional of interest (3.4.6) at a point $(u,v)\in U\times V$ via observation (3.4.9) is concerned, we are in the situation considered in Section 3.3 and can use the machinery developed there. Indeed, let us specify the data introduced in section 3.3.1 and participating in the constructions of section 3.3 as follows:

- $\mathcal{H}=\{f=(h,H)\in\mathcal{H}\}\subset\mathcal{E}_H=\mathbf{R}^d\times\mathbf{S}^d$, with $\mathcal{H}$ defined in (3.4.8), and the inner product on $\mathcal{E}_H$ defined as
$$
\langle (h,H),(h',H')\rangle = h^Th'+\frac{1}{2}\mathrm{Tr}(HH'),
$$
$\mathcal{E}_M=\mathbf{S}^{d+1}\times\mathbf{S}^{m+1}$, and $\mathcal{M}$, Φ are as defined in (3.4.8);

- $\mathcal{E}_X=\mathbf{R}^k\times\mathbf{S}^{d+1}$, $X:=\{x=(v,Z): v\in V, Z=[u;1][u;1]^T, u\in U\}\subset\mathcal{X}:=\{x=(v,Z)\in\mathcal{V}\times\mathcal{Z}\}$;

- $\mathcal{A}(x)=\mathcal{A}(v,Z)=(M(v),Z)$; note that $\mathcal{A}$ is affine mapping from $\mathcal{E}_X$ into $\mathcal{E}_M$ mapping $\mathcal{X}$ into $\mathcal{M}$, as required in section 3.3. Observe that when $u\in U$ and $v\in V$, the distribution $P=P_{u,v}$ of observation ω defined by (3.4.9) satisfies the relation
$$
\begin{array}{l}
\forall (f=(h,H)\in\mathcal{H}) :\\
\ln\left(\mathbf{E}_{\omega\sim P}\left\{\mathrm{e}^{\langle f,\omega\rangle}\right\}\right)=\ln\left(\mathbf{E}_{\zeta\sim\mathcal{N}(A[u;1],M(v))}\left\{\mathrm{e}^{h^T\zeta+\frac{1}{2}\zeta^TH\zeta}\right\}\right)\le\Phi(h,H;M(v),[u;1][u;1]^T),
\end{array}
\tag{3.4.11}
$$
see (3.4.10);

- $v(x=(v,Z))=\varrho(Z):\mathcal{X}\to\mathbf{R}$,

- we define affine functional $G(x)$ on $\mathcal{E}_X$ by the relation
$$
\langle g, x:=(v,Z)\rangle = q^Tv+\mathrm{Tr}(QZ),
$$
see (3.4.6). As a result, for $x\in X$, that is, for $x=(v,[u;1][u;1]^T)$ with $v\in V$ and $u\in U$ we have
$$
F(u,v)=G(x). \tag{3.4.12}
$$

Applying to the just specified data Corollary 3.3.2 (which is legitimate, since our Φ clearly satisfies (3.3.7)), we arrive at the result as follows:

Proposition 3.4.2 *In the just described situation, let us set*

$$\begin{array}{l}
\widehat{\Psi}_+(h,H)\\
:= \inf\limits_{\alpha}\left\{\max\limits_{(v,Z)\in V\times\mathcal{Z}}\left[\alpha\Phi(h/\alpha,H/\alpha;M(v),Z)-G(v,Z)-\varrho(Z)+K^{-1}\alpha\ln(2/\epsilon)\right]:\alpha>0,-\gamma\alpha\Theta_*^{-1}\preceq H\preceq\gamma\alpha\Theta_*^{-1}\right\}\\
=\max\limits_{(v,Z)\in V\times\mathcal{Z}}\inf\limits_{\substack{\alpha>0,\\ -\gamma\alpha\Theta_*^{-1}\preceq H\preceq\gamma\alpha\Theta_*^{-1}}}\left[\alpha\Phi(h/\alpha,H/\alpha;M(v),Z)-G(v,Z)-\varrho(Z)+K^{-1}\alpha\ln(2/\epsilon)\right],\\
\widehat{\Psi}_-(h,H)\\
:= \inf\limits_{\alpha}\left\{\max\limits_{(v,Z)\in V\times\mathcal{Z}}\left[\alpha\Phi(-h/\alpha,-H/\alpha;M(v),Z)+G(v,Z)-\varrho(Z)+K^{-1}\alpha\ln(2/\epsilon)\right]:\alpha>0,-\gamma\alpha\Theta_*^{-1}\preceq H\preceq\gamma\alpha\Theta_*^{-1}\right\}\\
=\max\limits_{(v,Z)\in V\times\mathcal{Z}}\inf\limits_{\substack{\alpha>0,\\ -\gamma\alpha\Theta_*^{-1}\preceq H\preceq\gamma\alpha\Theta_*^{-1}}}\left[\alpha\Phi(-h/\alpha,-H/\alpha;M(v),Z)+G(v,Z)-\varrho(Z)+K^{-1}\alpha\ln(2/\epsilon)\right].
\end{array}\tag{3.4.13}$$

so that the functions $\widehat{\Psi}_\pm(h,H):\mathbf{R}^d\times\mathbf{S}^d\to\mathbf{R}$ *are convex. Furthermore, whenever* $\bar{h},\bar{H},\bar{\rho},\bar{\varkappa}$ *form a feasible solution to the system of convex constraints*

$$\widehat{\Psi}_+(h,H)\leq\rho-\varkappa,\ \ \widehat{\Psi}_-(h,H)\leq\rho+\varkappa\tag{3.4.14}$$

in variables $(h,H)\in\mathbf{R}^d\times\mathbf{S}^d$, $\rho\in\mathbf{R}$, $\varkappa\in\mathbf{R}$, *setting*

$$\widehat{g}(\zeta^K:=(\zeta_1,...,\zeta_K))=\frac{1}{K}\sum_{i=1}^K\left[h^T\zeta_i+\frac{1}{2}\zeta_i^TH\zeta_i\right]+\bar{\varkappa},\tag{3.4.15}$$

we get an estimate of the functional of interest $F(u,v)=[u;1]^TQ[u;1]+q^Tv$ *via* K *independent observations*

$$\zeta_i\sim\mathcal{N}(A[u;1],M(v)),\ i=1,...,K,$$

with the following property:

$$\forall(u,v)\in U\times V:\mathrm{Prob}_{\zeta^K\sim[\mathcal{N}(A[u;1],M(v))]^K}\left\{|F(u,v)-\widehat{g}(\zeta^K)|>\bar{\rho}+\varrho([u;1][u;1]^T)\right\}\leq\epsilon.\tag{3.4.16}$$

Proof. Under the premise of Proposition, let us fix $u\in U$, $v\in V$, so that $x:=(v,Z:=[u;1][u;1]^T)\in X$. Denoting, as above, by $P=P_{u,v}$ the distribution of $\omega:=(\zeta,\zeta\zeta^T)$ with $\zeta\sim\mathcal{N}(A[u;1],M(v))$, and invoking (3.4.11), we see that for just defined (x,P), relation (3.3.2) takes place. Applying Corollary 3.3.2, we conclude that

$$\mathrm{Prob}_{\zeta^K\sim[\mathcal{N}(A[u;1],M(v))]^K}\left\{|\widehat{g}(\zeta^K)-G(x)|>\bar{\rho}+\varrho([u;1][u;1]^T)\right\}\leq\epsilon.$$

It remains to note that by construction for $x=(v,Z)$ in question it holds

$$G(x)=q^Tv+\mathrm{Tr}(QZ)=q^Tv+\mathrm{Tr}(Q[u;1][u;1]^T)=q^Tv+[u;1]^TQ[u,1]=F(u,v).\qquad\square$$

An immediate consequence of Proposition 3.4.2 is as follows:

Corollary 3.4.1 *Under the premise and in the notation of Proposition 3.4.2, let* $(h,H)\in\mathbf{R}^d\times\mathbf{S}^d$. *Setting*

$$\begin{array}{rcl}\rho&=&\frac{1}{2}\left[\widehat{\Psi}_+(h,H)+\widehat{\Psi}_-(h,H)\right],\\ \varkappa&=&\frac{1}{2}\left[\widehat{\Psi}_-(h,H)-\widehat{\Psi}_+(h,H)\right],\end{array}\tag{3.4.17}$$

the ϵ*-risk of estimate* (3.4.15) *does not exceed* ρ.

Indeed, with ρ and $\varkappa$ given by (3.4.17), $h,H,\rho,\varkappa$ satisfy (3.4.14).

3.4.1.4 Consistency

We are about to present a simple sufficient condition for the estimator suggested by Proposition 3.4.2 to be consistent, in the sense of Section 3.3.4.1. Specifically, in the situation and with the notation from Sections 3.4.1.1, 3.4.1.3 assume that

A.1. $\varrho(\cdot) \equiv 0$,

A.2. $V = \{\bar{v}\}$ is a singleton, which allows to set $\Theta_* = M(\bar{v})$, to satisfy (3.4.1) with $\delta = 0$, and to assume w.l.o.g. that

$$F(u,v) = [u;1]^T Q[u;1],\ G(Z) = \mathrm{Tr}(QZ);$$

A.3. the first m columns of the $d \times (m+1)$ matrix A are linearly independent.

By A.3, the columns of $(d+1)\times(m+1)$ matrix B, see (3.4.8), are linearly independent, so that we can find $(m+1)\times(d+1)$ matrix C such that $CB = I_{m+1}$. Let us define $(\bar{h}, \bar{H}) \in \mathbf{R}^d \times \mathbf{S}^d$ from the relation

$$\left[\begin{array}{c|c} \bar{H} & \bar{h} \\ \hline \bar{h}^T & \end{array}\right] = 2(C^T QC)^o, \tag{3.4.18}$$

where for $(d+1)\times(d+1)$ matrix S, S^o is the matrix obtained from S by zeroing our the entry in the cell $(d+1, d+1)$.

The consistency of our estimation machinery is given by the following simple statement:

Proposition 3.4.3 *In the just described situation and under assumptions A.1-3, given $\epsilon \in (0,1)$, consider the estimate*

$$\widehat{g}_{K,\epsilon}(\zeta^K) = \frac{1}{K}\sum_{k=1}^K [\bar{h}^T\zeta_k + \frac{1}{2}\zeta^T \bar{H}\zeta_k] + \varkappa_{K,\epsilon},$$

where

$$\varkappa_{K,\epsilon} = \frac{1}{2}\left[\widehat{\Psi}_-(\bar{h},\bar{H}) - \widehat{\Psi}_+(\bar{h},\bar{H})\right]$$

and $\widehat{\Psi}_\pm = \widehat{\Psi}_\pm^{K,\epsilon}$ are given by (3.4.13). Then the ϵ-risk of $\widehat{g}_{K,\epsilon}(\cdot)$ goes to 0 as $K \to \infty$.

For proof, see Section 3.4.3.

3.4.1.5 A modification

In the situation described in the beginning of this Section, let a set $W \subset U \times V$ be given, and assume we are interested in recovering functional of interest (3.4.6) at points $(u,v) \in W$ only. When reducing the "domain of interest" $U \times V$ to W, we hopefully can reduce the achievable ϵ-risk of recovery. To utilize for this purpose the machinery we have developed, assume that we can point our a convex compact set $\mathcal{W} \subset V \times \mathcal{Z}$ such that

$$(u,v) \in W \Rightarrow (v, [u;1][u;1]^T) \in \mathcal{W}$$

A straightforward inspection justifies the following

Remark 3.4.1 *In the just described situation, the conclusion of Proposition 3.4.2 remains valid when the set $U \times V$ participating in (3.4.16) and in relations (3.4.13) is reduced to $\mathcal{W}$. This modification enlarges the feasible set of (3.4.14) and thus reduces the achievable values of risk bound $\bar{\rho}$.*

3.4.2 Estimating quadratic form, sub-Gaussian case

3.4.2.1 Situation

In the rest of this Section we are interested in the situation is as follows: we are given i.i.d. random observations

$$\zeta_i \sim \mathcal{SG}(A[u;1], M(v)),\ i = 1, ..., K, \tag{3.4.19}$$

where $\zeta \sim \mathcal{SG}(\mu, \Theta)$ means that ζ is sub-Gaussian with parameters $\mu \in \mathbf{R}^d$, $\Theta \in \mathcal{S}^d_+$, and

- (u, v) is unknown "signal" known to belong to a given set $U \times V$, where
  - $U \subset \mathbf{R}^m$ is a compact set, and
  - $V \subset \mathbf{R}^k$ is a compact convex set;
- A is a given $d \times (m + 1)$ matrix, and $v \mapsto M(v) : \mathbf{R}^k \to \mathbf{S}^{d+1}$ is affine mapping such that $M(v) \succeq 0$ whenever $v \in V$.

We are also given a convex calibrating function $\varrho(Z) : \mathbf{S}^{m+1}_+ \to \mathbf{R}$ and "functional of interest"

$$F(u, v) = [u; 1]^T Q[u; 1] + q^T v, \tag{3.4.20}$$

where Q and q are known $(m+1)\times(m+1)$ symmetric matrix and k-dimensional vector, respectively. Our goal is to recover $F(u, v)$, for unknown (u, v) known to belong to $U\times V$, via observation (3.4.19).

Note that the only difference of our present situation with the one considered in Section 3.4.1.1 is that now we allow for sub-Gaussian, and not necessary Gaussian, observations.

3.4.2.2 Construction & Result

Let

$$\mathcal{V} = \{M(v) : v \in V\},$$

so that $\mathcal{V}$ is a convex compact subset of the positive semidefinite cone $\mathbf{S}^d_+$. Let us select somehow

1. a matrix $\Theta_* \succ 0$ such that $\Theta_* \succeq \Theta$, for all $\Theta \in \mathcal{V}$;
2. a convex compact subset $\mathcal{Z}$ of the set $\mathcal{Z}^+ = \{Z \in \mathbf{S}^{m+1}_+ : Z_{m+1,m+1} = 1\}$ such that $[u; 1][u; 1]^T \in \mathcal{Z}$ for all $u \in U$;
3. reals $\gamma, \gamma^+ \in (0, 1)$ with $\gamma < \gamma^+$ (say, $\gamma = 0.99, \gamma^+ = 0.999$).

Preliminaries Given the data of the above description and $\delta \in [0, 2]$, we set (cf. Proposition 3.4.1)

$$\begin{array}{l}
\mathcal{H} = \mathcal{H}_\gamma := \{(h, H) \in \mathbf{R}^d \times \mathbf{S}^d : -\gamma\Theta_*^{-1} \preceq H \preceq \gamma\Theta_*^{-1}\},\\
B = \left[\begin{array}{c} A \\ [0, ..., 0, 1] \end{array}\right] \in \mathbf{R}^{(d+1)\times(m+1)},\\
\mathcal{M} = \mathcal{V} \times \mathcal{Z},\\
\Psi(h, H, G; Z)\\
\quad = -\frac{1}{2}\ln\mathrm{Det}(I - \Theta_*^{1/2} G \Theta_*^{1/2}) + \frac{1}{2}\mathrm{Tr}\left(ZB^T\left[\left[\begin{array}{c|c} H & h \\ \hline h^T & \end{array}\right] + [H, h]^T[\Theta_*^{-1} - G]^{-1}[H, h]\right]B\right) :\\
\hfill \left(\mathcal{H} \times \{G : 0 \preceq G \preceq \gamma^+\Theta_*^{-1}\}\right) \times \mathcal{Z} \to \mathbf{R},\\
\Psi_\delta(h, H, G; \Theta, Z) = -\frac{1}{2}\ln\mathrm{Det}(I - \Theta_*^{1/2} G \Theta_*^{1/2}) + \frac{1}{2}\mathrm{Tr}([\Theta - \Theta_*]G) + \frac{\delta(2+\delta)}{2(1-\|\Theta_*^{1/2} G \Theta_*^{1/2}\|)}\|\Theta_*^{1/2} G \Theta_*^{1/2}\|_H^2\\
\quad +\frac{1}{2}\mathrm{Tr}\left(ZB^T\left[\left[\begin{array}{c|c} H & h \\ \hline h^T & \end{array}\right] + [H, h]^T[\Theta_*^{-1} - G]^{-1}[H, h]\right]B\right) :\\
\hfill \left(\mathcal{H} \times \{G : 0 \preceq G \preceq \gamma^+\Theta_*^{-1}\}\right) \times (\{0 \preceq \Theta \preceq \Theta_*\} \times \mathcal{Z}) \to \mathbf{R},\\
\Phi(h, H; Z) = \min\limits_G \left\{\Psi(h, H, G; Z) : 0 \preceq G \preceq \gamma^+\Theta_*^{-1}, G \succeq H\right\} : \mathcal{H} \times \mathcal{Z} \to \mathbf{R},\\
\Phi_\delta(h, H; \Theta, Z) = \min\limits_G \left\{\Psi_\delta(h, H, G; \Theta, Z) : 0 \preceq G \preceq \gamma^+\Theta_*^{-1}, G \succeq H\right\} : \mathcal{H} \times (\{0 \preceq \Theta \preceq \Theta_*\} \times \mathcal{Z}) \to \mathbf{R}.
\end{array} \tag{3.4.21}$$

The following statement is straightforward reformulation of Proposition 2.8.9.i:

Proposition 3.4.4 *In the situation described in Section 3.4.2.1, we have*

(i) *Φ is well-defined real-valued continuous function on the domain $\mathcal{H}\times\mathcal{Z}$; the function is convex in $(h,H)\in\mathcal{H}$, concave in $Z\in\mathcal{Z}$, and $\Phi(0;Z)\geq 0$. Furthermore, let $(h,H)\in\mathcal{H}$, $u\in U$, $v\in V$, and let $\zeta\sim\mathcal{SG}(A[u;1],M(v))$. Then*

$$\ln\left(\mathbf{E}_\zeta\left\{\exp\{h^T\zeta+\frac{1}{2}\zeta^TH\zeta\}\right\}\right)\leq\Phi(h,H;[u;1][u;1]^T). \tag{3.4.22}$$

(ii) *Let $\mathcal{V}$ be a convex compact subset of $\mathbf{S}^d_+$ such that $M(v)\in\mathcal{V}$ for all $v\in V$, and let $\delta\in[0,2]$ be such that*

$$\Theta\in\mathcal{V}\Rightarrow\{\Theta\preceq\Theta_*\}\ \&\ \{\|\Theta^{1/2}\Theta_*^{-1/2}-I\|\leq\delta\}. \tag{3.4.23}$$

Then $\Phi_\delta(h,H;\Theta,Z)$ is well-defined real-valued continuous function on the domain $\mathcal{H}\times(\mathcal{V}\times\mathcal{Z})$; the function is convex in $(h,H)\in\mathcal{H}$, concave in $(\Theta,Z)\in\mathcal{V}\times\mathcal{Z}$, and $\Phi_\delta(0;\Theta,Z)\geq 0$. Furthermore, let $(h,H)\in\mathcal{H}$, $u\in U$, $v\in V$, and let $\zeta\sim\mathcal{SG}(A[u;1],M(v))$. Then

$$\ln\left(\mathbf{E}_\zeta\left\{\exp\{h^T\zeta+\frac{1}{2}\zeta^TH\zeta\}\right\}\right)\leq\Phi_\delta(h,H;M(v),[u;1][u;1]^T). \tag{3.4.24}$$

The estimate. We estimate the functional of interest similarly to the case of Gaussian observations. Specifically, let us pass from observations (3.4.19) to their quadratic lifts, so that our observations become

$$\omega_i=(\zeta_i,\zeta_i\zeta_i^T),\ 1\leq i\leq K,\zeta_i\sim\mathcal{SG}(A[u;1],M(v))\text{ are i.i.d.} \tag{3.4.25}$$

As in the Gaussian case, we find ourselves in the situation considered in section 3.3.3 and can use the machinery developed there. Indeed, let us specify the data introduced in section 3.3.1 and participating in the constructions of section 3.3 as follows:

- $\mathcal{H}=\{f=(h,H)\in\mathcal{H}\}\subset\mathcal{E}_H=\mathbf{R}^d\times\mathbf{S}^d$, with $\mathcal{H}$ defined in (3.4.21), and the inner product on $\mathcal{E}_H$ defined as
$$\langle(h,H),(h',H')\rangle=h^Th'+\frac{1}{2}\mathrm{Tr}(HH'),$$
$\mathcal{E}_M=\mathbf{S}^{m+1}$, and $\mathcal{M}$, Φ are as defined in (3.4.21);
- $\mathcal{E}_X=\mathbf{R}^k\times\mathbf{S}^{d+1}$, $X:=\{(v,Z):v\in V,Z=[u;1][u;1]^T,u\in U\}\subset\mathcal{X}:=\{(v,Z):v\in V,Z\in\mathcal{Z}\}$;
- $\mathcal{A}(x)=\mathcal{A}(v,Z)=(M(v),Z)$; note that $\mathcal{A}$ is affine mapping from $\mathcal{E}_X$ into $\mathcal{E}_M$ mapping $\mathcal{X}$ into $\mathcal{M}$, as required in section 3.3. Observe that when $u\in U$ and $v\in V$, the distribution $P=P$ of observation ω_i defined by (3.4.25) satisfies the relation
$$\begin{array}{l}\forall(f=(h,H)\in\mathcal{H}):\\ \ln\left(\mathbf{E}_{\omega\sim P}\left\{\mathrm{e}^{\langle f,\omega\rangle}\right\}\right)=\ln\left(\mathbf{E}_{\zeta\sim\mathcal{SG}(A[u;1],M(v))}\left\{\mathrm{e}^{h^T\zeta+\frac{1}{2}\zeta^TH\zeta}\right\}\right)\leq\Phi(h,H;[u;1][u;1]^T),\end{array} \tag{3.4.26}$$
see (3.4.22). Moreover, in the case of (3.4.23), we have also
$$\begin{array}{l}\forall(f=(h,H)\in\mathcal{H}):\\ \ln\left(\mathbf{E}_{\omega\sim P}\left\{\mathrm{e}^{\langle f,\omega\rangle}\right\}\right)=\ln\left(\mathbf{E}_{\zeta\sim\mathcal{SG}(A[u;1],M(v))}\left\{\mathrm{e}^{h^T\zeta+\frac{1}{2}\zeta^TH\zeta}\right\}\right)\leq\Phi_\delta(h,H;M(v),[u;1][u;1]^T),\end{array} \tag{3.4.27}$$
see (3.4.24);

- we set $\upsilon(x = (v, Z)) = \varrho(Z)$,
- we define affine functional $G(x)$ on $\mathcal{E}_X$ by the relation

$$G(x := (v, Z)) = q^T v + \mathrm{Tr}(QZ),$$

see (3.4.20). As a result, for $x \in X$, that is, for $x = (v, [u; 1][u; 1]^T)$ with $v \in V$ and $u \in U$ we have

$$F(u, v) = G(x). \tag{3.4.28}$$

The result. Applying to the just specified data Corollary 3.3.2 (which is legitimate, since our Φ clearly satisfies (3.3.7)), we arrive at the result as follows:

Proposition 3.4.5 *In the situation described in Section 3.4.2.1, let us set*

$$\begin{array}{l}
\widehat{\Psi}_+(h,H)\\
\quad := \inf\limits_{\alpha}\left\{\max\limits_{(v,Z)\in V\times\mathcal{Z}}\left[\alpha\Phi(h/\alpha, H/\alpha; Z) - G(v,Z) - \varrho(Z) + \alpha K^{-1}\ln(2/\epsilon)\right] : \alpha > 0, -\gamma\alpha\Theta_*^{-1} \preceq H \preceq \gamma\alpha\Theta_*^{-1}\right\}\\
\quad = \max\limits_{(v,Z)\in V\times\mathcal{Z}} \inf\limits_{\substack{\alpha>0,\\ -\gamma\alpha\Theta_*^{-1}\preceq H\preceq\gamma\alpha\Theta_*^{-1}}} \left[\alpha\Phi(h/\alpha, H/\alpha; Z) - G(v,Z) - \varrho(Z) + \alpha K^{-1}\ln(2/\epsilon)\right],\\
\widehat{\Psi}_-(h,H)\\
\quad := \inf\limits_{\alpha}\left\{\max\limits_{(v,Z)\in V\times\mathcal{Z}}\left[\alpha\Phi(-h/\alpha, -H/\alpha; Z) + G(v,Z) - \varrho(Z) + \alpha K^{-1}\ln(2/\epsilon)\right] : \alpha > 0, -\gamma\alpha\Theta_*^{-1} \preceq H \preceq \gamma\alpha\Theta_*^{-1}\right\}\\
\quad = \max\limits_{(v,Z)\in V\times\mathcal{Z}} \inf\limits_{\substack{\alpha>0,\\ -\widehat{\gamma}_d\alpha\Theta_*^{-1}\preceq H\preceq\widehat{\gamma}_d\alpha\Theta_*^{-1}}} \left[\alpha\Phi(-h/\alpha, -H/\alpha; Z) + G(v,Z) - \varrho(Z) + \alpha K^{-1}\ln(2/\epsilon)\right].
\end{array} \tag{3.4.29}$$

so that the functions $\widehat{\Psi}_\pm(h,H) : \mathbf{R}^d \times \mathbf{S}^d \to \mathbf{R}$ are convex. Furthermore, whenever $\bar{h}, \bar{H}, \bar{\rho}, \bar{\varkappa}$ form a feasible solution to the system of convex constraints

$$\widehat{\Psi}_+(h,H) \le \rho - \varkappa, \ \widehat{\Psi}_-(h,H) \le \rho + \varkappa \tag{3.4.30}$$

in variables $(h,H) \in \mathbf{R}^d \times \mathbf{S}^d$, $\rho \in \mathbf{R}$, $\varkappa \in \mathbf{R}$, setting

$$\widehat{g}(\zeta^K) = \frac{1}{K}\sum_{i=1}^K \left[h^T\zeta_i + \frac{1}{2}\zeta_i^T H\zeta_i\right] + \bar{\varkappa},$$

we get an estimate of the functional of interest $F(u,v) = [u;1]^T Q[u;1] + q^T v$ via i.i.d. observations

$$\zeta_i \sim \mathcal{SG}(A[u;1], M(v)),\ 1 \le i \le K,$$

with the following property:

$$\forall (u,v) \in U \times V : \mathrm{Prob}_{\zeta^K\sim[\mathcal{SG}(A[u;1],M(v))]^K}\left\{|F(u,v) - \widehat{g}(\zeta^K)| > \bar{\rho} + \varrho([u;1][u;1]^T)\right\} \le \epsilon. \tag{3.4.31}$$

Proof. Under the premise of Proposition, let us fix $u \in U$, $v \in V$, so that $x := (v, Z := [u;1][u;1]^T) \in X$. Denoting by P the distribution of $\omega := (\zeta, \zeta\zeta^T)$ with $\zeta \sim \mathcal{SG}(A[u;1], M(v))$, and invoking (3.4.26), we see that for just defined (x, P), relation (3.3.2) takes place. Applying Corollary 3.3.2, we conclude that

$$\mathrm{Prob}_{\zeta^K\sim[\mathcal{N}(A[u;1],M(v))]^K}\left\{|\widehat{g}(\zeta^K) - G(x)| > \bar{\rho} + \varrho([u;1][u;1]^T)\right\} \le \epsilon.$$

It remains to note that by construction for $x = (v, Z)$ in question it holds

$$G(x) = q^T v + \mathrm{Tr}(QZ) = q^T v + \mathrm{Tr}(Q[u;1][u;1]^T) = q^T v + [u;1]^T Q[u,1] = F(u,v). \qquad \square$$

Remark 3.4.2 *In the situation described in section 3.4.2.1, let $\delta \in [0,2]$ be such that*

$$\|\Theta^{1/2}\Theta_*^{-1/2} - I\| \leq \delta \;\forall \Theta \in \mathcal{V}.$$

Then the conclusion of Proposition 3.4.5 remains valid when the function Φ in (3.4.29) is replaced with the function Φ_δ, that is, when $\widehat{\Psi}_\pm$ are defined as

$$\begin{array}{l} \widehat{\Psi}_+(h,H) \\ \quad := \inf\limits_{\alpha} \left\{ \max\limits_{(v,Z)\in V\times\mathcal{Z}} \left[\alpha\Phi_\delta(h/\alpha, H/\alpha; M(v), Z) - G(v,Z) - \varrho(Z) + \alpha K^{-1}\ln(2/\epsilon)\right] : \right. \\ \hfill \left. \alpha > 0, -\gamma\alpha\Theta_*^{-1} \preceq H \preceq \gamma\alpha\Theta_*^{-1} \right\} \\ \quad = \max\limits_{(v,Z)\in V\times\mathcal{Z}} \inf\limits_{\substack{\alpha>0,\\ -\gamma\alpha\Theta_*^{-1}\preceq H\preceq\gamma\alpha\Theta_*^{-1}}} \left[\alpha\Phi_\delta(-h/\alpha, H/\alpha; M(v), Z) - G(v,Z) - \varrho(Z) + \alpha K^{-1}\ln(2/\epsilon)\right], \\ \widehat{\Psi}_-(h,H) \\ \quad := \inf\limits_{\alpha} \left\{ \max\limits_{(v,Z)\in V\times\mathcal{Z}} \left[\alpha\Phi_\delta(-h/\alpha, -H/\alpha; M(v), Z) + G(v,Z) - \varrho(Z) + \alpha K^{-1}\ln(2/\epsilon)\right] : \right. \\ \hfill \left. \alpha > 0, -\gamma\alpha\Theta_*^{-1} \preceq H \preceq \gamma\alpha\Theta_*^{-1} \right\} \\ \quad = \max\limits_{(v,Z)\in V\times\mathcal{Z}} \inf\limits_{\substack{\alpha>0,\\ -\widehat{\gamma}_d\alpha\Theta_*^{-1}\preceq H\preceq\widehat{\gamma}_d\alpha\Theta_*^{-1}}} \left[\alpha\Phi_\delta(-h/\alpha, -H/\alpha; M(v), Z) + G(v,Z) - \varrho(Z) + \alpha K^{-1}\ln(2/\epsilon)\right]. \end{array} \tag{3.4.32}$$

To justify Remark 3.4.2, it suffices to use in the proof of Proposition 3.4.5 relation (3.4.27) in the role of (3.4.26). Note that what is better in terms of the risk of the resulting estimate – Proposition 3.4.5 "as is" or its modification presented in Remark 3.4.2 – depends on the situation, so that it makes sense to keep in mind both options.

3.4.2.3 Numerical illustration, direct observations

The problem. Our initial illustration is deliberately selected to be extremely simple: given direct noisy observations

$$\zeta = u + \xi$$

of unknown signal $u \in \mathbf{R}^m$ known to belong to a given set U, we want to recover the "energy" u^Tu of u; what we are interested in, is the quadratic in ζ estimate with as small ϵ-risk on U as possible; here $\epsilon \in (0,1)$ is a given design parameter. The details of our setup are as follows:

- U is the "spherical layer" $U = \{u \in \mathbf{R}^m : r^2 \leq u^Tu \leq R^2\}$, where r, R, $0 \leq r < R < \infty$ are given. As a result, the "main ingredient" in constructions from sections 3.4.1.3, 3.4.2.2 – the convex compact subset $\mathcal{Z}$ of the set $\{Z \in \mathbf{S}_+^{m+1} : Z_{m+1,m+1} = 1\}$ containing all matrices $[u;1][u;1]^T$, $u \in U$, can be specified as

$$\mathcal{Z} = \{Z \in \mathbf{S}_+^{m+1} : Z_{m+1,m+1} = 1, 1 + r^2 \leq \mathrm{Tr}(Z) \leq 1 + R^2\};$$

- ξ is either $\sim \mathcal{N}(0,\Theta)$ (Gaussian case), or $\sim \mathcal{SG}(0,\Theta)$ (sub-Gaussian case), with matrix Θ known to be diagonal with diagonal entries satisfying $\theta\sigma^2 \leq \Theta_{ii} \leq \sigma^2$, $1 \leq i \leq d = m$, with known $\theta \in [0,1]$ and $\sigma^2 > 0$;

- the calibrating function $\varrho(Z)$ is $\varrho(Z) = \varsigma(\sum_{i=1}^m Z_{ii})$, where ς is a convex continuous real-valued function on $\mathbf{R}_+$. Note that with this selection, the claim that ϵ-risk of an estimate $\widehat{g}(\cdot)$ is $\leq \rho$ means that whenever $u \in U$, one has

$$\mathrm{Prob}\{|\widehat{g}(u+\xi) - u^Tu| > \rho + \varsigma(u^Tu)\} \leq \epsilon. \tag{3.4.33}$$

Processing the problem. It is easily seen that in the situation in question the machinery of sections 3.4.1, 3.4.2 boils down to the following:

1. We lose nothing when restricting ourselves with estimates of the form
$$\widehat{g}(\zeta) = \frac{1}{2}\eta\zeta^T\zeta + \varkappa, \tag{3.4.34}$$
with properly selected scalars η and $\varkappa$;

2. In Gaussian case, η and $\varkappa$ are yielded by the convex optimization problem with just 3 variables α_+, α_-, η, namely the problem
$$\begin{array}{rcl}\multicolumn{3}{l}{\min\limits_{\alpha_\pm,\eta}\left\{\widehat{\Psi}(\alpha_+,\alpha_-,\eta) = \frac{1}{2}\left[\widehat{\Psi}_+(\alpha_+,\eta) + \widehat{\Psi}_-(\alpha_-,\eta)\right] : \sigma^2|\eta| < \alpha_\pm\right\},}\\ \widehat{\Psi}_+(\alpha_+,\eta) &=& -\frac{d\alpha_+}{2}\ln(1-\sigma^2\eta/\alpha_+) + \frac{d}{2}\sigma^2(1-\theta)\max[-\eta,0] + \frac{d\delta(2+\delta)\sigma^4\eta^2}{2(\alpha_+-\sigma^2|\eta|)}\\ && +\max\limits_{r^2\leq t\leq R^2}\left[\left[\frac{\alpha_+\eta}{2(\alpha_+-\sigma^2\eta)}-1\right]t - \varsigma(t)\right] + \alpha_+\ln(2/\epsilon)\\ \widehat{\Psi}_-(\alpha_+,\eta) &=& -\frac{d\alpha_-}{2}\ln(1+\sigma^2\eta/\alpha_-) + \frac{d}{2}\sigma^2(1-\theta)\max[\eta,0] + \frac{d\delta(2+\delta)\sigma^4\eta^2}{2(\alpha_--\sigma^2|\eta|)}\\ && +\max\limits_{r^2\leq t\leq R^2}\left[\left[-\frac{\alpha_-\eta}{2(\alpha_-+\sigma^2\eta)}+1\right]t - \varsigma(t)\right] + \alpha_-\ln(2/\epsilon)\ ,\end{array} \tag{3.4.35}$$
where $\delta = 1-\sqrt{\theta}$. Specifically, the η-component of a feasible solution to (3.4.35) augmented by the quantity
$$\varkappa = \frac{1}{2}\left[\widehat{\Psi}_-(\alpha_-,\eta) - \widehat{\Psi}_+(\alpha_+,\eta)\right]$$
yields estimate (3.4.34) with ϵ-risk on U not exceeding $\widehat{\Psi}(\alpha_+,\alpha_-,\eta)$;

3. In sub-Gaussian case, η and $\varkappa$ are yielded by convex optimization problem with just 5 variables, $\alpha_\pm, g_\pm, \eta$, namely, the problem
$$\begin{array}{rcl}\multicolumn{3}{l}{\min\limits_{\alpha_\pm,g_\pm,\eta}\left\{\widehat{\Psi}(\alpha_\pm,g_\pm,\eta) = \frac{1}{2}\left[\widehat{\Psi}_+(\alpha_+,\lambda_+,g_+\eta) + \widehat{\Psi}_-(\alpha_-,\lambda_-,g_-,\eta)\right] :\right.}\\ \multicolumn{3}{r}{\left. 0\leq\sigma^2 g_\pm\leq\gamma\alpha_\pm,\ \sigma^2\eta\geq-\alpha_+, \sigma^2\eta\leq\alpha_-,\ \eta\leq g_+,\ -\eta\leq g_-\right\},}\\ \widehat{\Psi}_+(\alpha_+,g_+,\eta) &=& -\frac{d\alpha_+}{2}\ln(1-\sigma^2 g_+) + \alpha_+\ln(2/\epsilon) + \max\limits_{r^2\leq t\leq R^2}\left[\left[\frac{\sigma^2\eta^2}{2(\alpha_+-g_+)}+\frac{1}{2}\eta-1\right]r - \varsigma(t)\right]\\ \widehat{\Psi}_-(\alpha_-,g_-,\eta) &=& -\frac{d\alpha_-}{2}\ln(1-\sigma^2 g_-) + \alpha_-\ln(2/\epsilon) + \max\limits_{r^2\leq t\leq R^2}\left[\left[\frac{\sigma^2\eta^2}{2(\alpha_--g_-)}-\frac{1}{2}\eta+1\right]r - \varsigma(t)\right]\end{array} \tag{3.4.36}$$
where $\gamma\in(0,1)$ is construction's parameter (we used $\gamma = 0.99$). Specifically, the η-component of a feasible solution to (3.4.36) augmented by the quantity
$$\varkappa = \frac{1}{2}\left[\widehat{\Psi}_-(\alpha_-,g_-,\eta) - \widehat{\Psi}_+(\alpha_+,g_+,\eta)\right]$$
yields estimate (3.4.34) with ϵ-risk on U not exceeding $\widehat{\Psi}(\alpha_\pm,g_\pm,\eta)$.

Note that the Gaussian case of our "energy estimation" problem is well studied in the literature, mainly in the case $\xi\sim\mathcal{N}(0,\sigma^2 I_m)$ of white Gaussian noise with exactly known variance σ^2; available results investigate analytically the interplay between the dimension m of signal, noise intensity σ^2 and the parameters R, r and offer provably optimal, up to absolute constant factors, estimates. A nice property of the proposed approach is that (3.4.35) automatically takes care of the parameters and results in estimates with seemingly near-optimal performance, as is witnessed by the numerical results we are about to present.

d	r	R	θ	Relative 0.01-risk, Gaussian case	Relative 0.01-risk, sub-Gaussian case	Optimality ratio
64	0	16	1	0.34808	0.44469	1.22
64	0	16	0.5	0.43313	0.44469	1.48
64	0	128	1	0.04962	0.05181	1.28
64	0	128	0.5	0.05064	0.05181	1.34
64	8	80	1	0.07827	0.08376	1.28
64	8	80	0.5	0.08095	0.08376	1.34
256	0	32	1	0.19503	0.30457	1.28
256	0	32	0.5	0.26813	0.30457	1.41
256	0	512	1	0.01264	0.01314	1.28
256	0	512	0.5	0.01289	0.01314	1.34
256	16	160	1	0.03996	0.04501	1.28
256	16	160	0.5	0.04255	0.04501	1.34
1024	0	64	1	0.10272	0.21923	1.28
1024	0	64	0.5	0.17032	0.21923	1.34
1024	0	2048	1	0.00317	0.00330	1.28
1024	0	2048	0.5	0.00324	0.00330	1.34
1024	32	320	1	0.02019	0.02516	1.28
1024	32	320	0.5	0.02273	0.02516	1.41

Table 3.3: Recovering signal's energy from direct observations

Numerical results. In the first series of experiments, we used the trivial calibrating function: $\varsigma(\cdot) \equiv 0$.

A typical sample of numerical results is presented in Table 3.3. To avoid large numbers, we display in the table *relative* 0.01-risk achievable with our machinery, that is, the plain risk divided by R^2; keeping this in mind, one should not be surprised that when extending the range $[r, R]$ of allow norms of the observed signal, all other components of the setup being fixed, the relative risk can decrease (the actual risk, of course, can only increase). Note that in all our experiments σ was set to 1.

Along with the values of the relative 0.01-risk, we present also the values of "optimality ratios" – the ratios of the relative risks achievable with our machinery in the Gaussian case to the (lower bounds on the) the best possible under circumstances relative 0.01-risks. These lower bounds are obtained as follows. Let us select somehow values $r_1 < r_2$ in the allowed under the circumstances range $[r, R]$ of $\|u\|_2$, and two values, σ_1, σ_2, in the allowed range $[\theta\sigma, \sigma] = [\theta, 1]$ of values of diagonal entries in diagonal matrices Θ, and consider two distributions of observations P_1 and P_2 as follows: P_χ is the distribution of random vector $x + \zeta$, where x and ξ are independent, x is uniformly distributed on the sphere $\|x\|_2 = r_\chi$ and $\zeta \sim \mathcal{N}(0, \sigma_\chi^2 I_d)$. It is immediately seen that whenever the two simple hypotheses $\omega \sim P_1$, and $\omega \sim P_2$, can*not* be decided upon via a single observation by a test with total risk $\leq 2\epsilon$ (with the total risk of a test defined as the sum, over the two hypotheses in question, of probabilities for the test to reject the hypothesis when it is true), the quantity $\delta = \frac{r_2^2 - r_1^2}{2}$ is a lower bound on the optimal ϵ-risk, Risk_ϵ^*, defined as the infimum, over all estimates recovering $\|u\|_2^2$ via single observation $\omega = u + \zeta$, of the ϵ-risk of the estimate, where the ϵ-risk is taken w.r.t. u running through the spherical layer $U = \{u : r^2 \leq u^T u \leq R^2\}$, and the covariance matrices Θ of Gaussian zero mean noise running through the set of scalar matrices with diagonal entries varying in $[\theta,1]$. In other words, denoting by $p_\chi(\cdot)$ the density of P_χ, we have

$$0.02 < \int_{\mathbf{R}^d} \min[p_1(\omega), p_2(\omega)]d\omega \Rightarrow \mathrm{Risk}^*_{0.01} \geq \frac{r_2^2 - r_1^2}{2}.$$

Now, the densities p_χ are spherically symmetric, whence, denoting by $q_\chi(\cdot)$ the univariate density of the energy $\omega^T\omega$ of observation $\omega \sim P_\chi$, we have

$$\int_{\mathbf{R}^d} \min[p_1(\omega), p_2(\omega)]d\omega = \int_0^\infty \min[q_1(s), q_2(s)]ds,$$

so that

$$0.02 < \int_0^\infty \min[q_1(s), q_2(s)]ds \Rightarrow \mathrm{Risk}^*_{0.01} \geq \frac{r_2^2 - r_1^2}{2}. \tag{3.4.37}$$

Now, on a closest inspection, q_χ is the convolution of two univariate densities representable by explicit computation-friendly formulas, implying that given $r_1, r_2, \sigma_1, \sigma_2$, we can check numerically whether the premise in (3.4.37) indeed takes place, and whenever the latter is the case, the quantity $\frac{r_2^2-r_1^2}{2}$ is a lower bound on $\mathrm{Risk}^*_{0.01}$. In our experiments, we used a simple search strategy (not described here) aimed at crude maximizing this bound in $r_1, r_2, \sigma_1, \sigma_2$ and used the resulting lower bounds on $\mathrm{Risk}^*_{0.01}$ to compute the optimality ratios presented in the table[7].

We believe that quite moderate values of the optimality ratios presented in the table (these results are typical for a much larger series of experiments we have conducted) witness quite good performance of our machinery.

Optimizing the relative risk. The "relative risk" displayed in Table 3.3 is the corresponding to the trivial calibrating function 0.01-risk in recovery u^Tu divided by the largest value R^2 of this risk allowed by the inclusion $u \in U$. When R is large, low relative risk can correspond to pretty high "actual" risk. For example, with $d := \dim u = 1024$, $\theta = 1$, and $U = \{u \in \mathbf{R}^d : \|u\|_2 \leq 1.e6\}$, the 0.01-risk becomes as large as $\rho \approx 6.5e6$; for "relatively small" signals, like $u^Tu \approx 10^4$, recovering u^Tu within accuracy ρ does not make much sense. In order to allow for "large" domains U, it makes sense to pass from the trivial calibrating function to a nontrivial one, like $\varsigma(t) = \alpha t$, with small positive α. With this calibrating function, (3.4.33) reads

$$\mathrm{Prob}\left\{|\widehat{g}(u+\xi) - u^Tu| > \rho + \alpha u^T u\right\} \leq \epsilon.$$

It turns out that (quite reasonable when U is large) "relative" characterization of risk results in much smaller values of ρ as compared to the case $\alpha = 0$ of "plain" risk. Here is instructive numerical data:

r	R	0.01-Risk, $\alpha = 0$	0.01-Risk, $\alpha = 0.01$	0.01-Risk, $\alpha = 0.1$
0	1.e7	6.51e7/6.51e7	1.33e3/1.58e3	474/642
1.e2	1.e7	6.51e7/6.51e7	1.33e3/1.58e3	−123/92.3
1.1e3	1.e7	6.51e7/6.51e7	−4.73e3/−4.48e3	−1.14e5/−1.14e5

$U = \{u \in \mathbf{R}^{1024} : r \leq \|u\|_2 \leq R\}$, $\theta = 1/2$
Left/Right: risks in Gaussian/sub-Gaussian cases

3.4.2.4 Numerical illustration, indirect observations

The problem. The estimation problem we are about to process numerically is as follows. Our observations are

$$\zeta = Au + \xi, \tag{3.4.38}$$

where

[7] The reader should not be surprised by "narrow numerical spectrum" of optimality ratios displayed in Table 3.3: our lower bounding scheme was restricted to identify actual optimality ratios among the candidate values 1.05^i, $i = 1, 2, ...$

- A is a given $d \times m$ matrix, with $m > d$ ("under-determined observations"),

- $u \in \mathbf{R}^m$ is a signal known to belong to a compact set U,

- $\xi \sim \mathcal{N}(0, \Theta)$ (Gaussian case) of $\xi \sim \mathcal{SG}(0, \Theta)$ (sub-Gaussian case) is the observation noise; Θ is positive semidefinite $d \times d$ matrix known to belong to a given convex compact set $\mathcal{V} \subset \mathbf{S}^d_+$.

Our goal is to recover the energy

$$F(u) = \frac{1}{m}\|u\|_2^2$$

of the signal from a single observation (3.4.38).

In our experiment, the data is specified as follows:

1. We think of $u \in \mathbf{R}^m$ as of discretization of a smooth function $x(t)$ of continuous argument $t \in [0; 1]$: $u_i = x(\frac{i}{m})$, $1 \leq i \leq m$. We set $U = \{u : \|Su\|_2 \leq 1\}$, where $u \mapsto Su$ is the finite-difference approximation of the mapping $x(\cdot) \mapsto (x(0), x'(0), x''(\cdot))$, so that U is a natural discrete-time analogy of the Sobolev-type ball $\{x : [x(0)]^2 + [x'(0)]^2 + \int_0^1 [x''(t)]^2 dt \leq 1\}$.

2. $d \times m$ matrix A is of the form UDV^T, where U and V are randomly selected $d \times d$ and $m \times m$ orthogonal matrices, and the d diagonal entries in diagonal $d \times m$ matrix D are of the form $\theta^{-\frac{i-1}{d-1}}$, $1 \leq i \leq d$; the "condition number" θ of A is design parameter.

3. The set $\mathcal{V}$ of allowed matrices Θ is the set of all diagonal $d \times d$ matrices with diagonal entries varying from 0 to σ^2, where the "noise intensity" σ is design parameter.

Processing the problem. Our estimating problem clearly is covered by the setups considered in sections 3.4.1 (Gaussian case) and 3.4.2 (sub-Gaussian case); in terms of these setups, it suffices to specify Θ_* as $\sigma^2 I_d$, $M(v)$ as the identity mapping of $\mathcal{V}$ onto itself, the mapping $u \mapsto A[u; 1]$ as the mapping $u \mapsto Pu$, and the set $\mathcal{Z}$ (which should be a convex compact subset of the set $\{Z \in \mathbf{S}^{d+1}_+ : Z_{d+1,d+1} = 0\}$ containing all matrices of the form $[u; 1][u; 1]^T$, $u \in U$) as the set

$$\mathcal{Z} = \{Z \in \mathbf{S}^{d+1}_+ : Z_{d+1,d+1} = 1, \mathrm{Tr}\left(Z\mathrm{Diag}\{S^T S, 0\}\right) \leq 1\}.$$

As suggested by Propositions 3.4.2 (Gaussian case) and 3.4.5 (sub-Gaussian case), the linear in "lifted observation" $\omega = (\zeta, \zeta\zeta^T)$ estimates of $F(u) = \frac{1}{m}\|u\|_2^2$ stem from the optimal solution (h_*, H_*) to the convex optimization problem

$$\mathrm{Opt} = \min_{h,H} \frac{1}{2}\left[\widehat{\Psi}_+(h, H) + \widehat{\Psi}_-(h, H)\right], \tag{3.4.39}$$

with $\widehat{\Psi}_\pm(\cdot)$ given by (3.4.13) in the Gaussian, and by (3.4.29) in the sub-Gaussian cases, with the number K of observations in (3.4.13), (3.4.29) set to 1. The resulting estimate is

$$\zeta \mapsto h_*^T\zeta + \frac{1}{2}\zeta^T H_*\zeta + \varkappa, \; \varkappa = \frac{1}{2}\left[\widehat{\Psi}_-(h_*, H_*) - \widehat{\Psi}_+(h_*, H_*)\right] \tag{3.4.40}$$

and the ϵ-risk of the estimate is (upper-bounded by) Opt.

Problem (3.4.39) is a well-structured convex-concave saddle point problem and as such is beyond the "immediate scope" of the standard Convex Programming software toolbox primarily aimed at solving well-structured convex minimization problems. However, applying conic duality, one can easily eliminate in (3.4.13), (3.4.29) the inner maxima over v, Z and end up with reformulation which can be solved numerically by CVX [77], and this is how (3.4.39) was processed in our experiments.

d, m	Opt, Gaussian case	Opt, sub-Gaussian case	LwBnd
$8, 12$	0.1362(+65%)	0.1382(+67%)	0.0825
$16, 24$	0.1614(+53%)	0.1640(+55%)	0.1058
$32, 48$	0.0687(+46%)	0.0692(+48%)	0.0469

Table 3.4: Upper bound (Opt) on the 0.01-risk of estimate (3.4.40), (3.4.39) vs. lower bound (LwBnd) on 0.01-risk achievable under the circumstances. In the experiments, $\sigma = 0.025$ and $\theta = 10$. Data in parentheses: excess of Opt over LwBnd.

Numerical results. In the experiments to be reported, we used the trivial calibrating function: $\varrho(\cdot) \equiv 0$.

Table 3.4 displays typical numerical results of our experiments. To give an impression of the performance of our approach, we present, along with the upper risk bounds for the estimates yielded by our machinery, simple lower bounds on ϵ-risk achievable under the circumstances. The origin of the lower bounds is as follows. Assume we are speaking about ϵ-risk and have at our disposal a signal $w \in U$, and let $t(w) = \|Aw\|_2$, $\rho = 2\sigma\mathrm{ErfInv}(\epsilon)$, where ErfInv is the inverse error-function:

$$\mathrm{Prob}_{\xi\sim\mathcal{N}(0,1)}\{\xi > \mathrm{ErfInv}(\epsilon)\} = \epsilon.$$

Setting $\theta(w) = \max[1 - \rho/t(w), 0]$, observe that $w' := \theta(w)w \in U$ and $\|Aw - Aw'\|_2 \leq \rho$, which, due to the origin of ρ, implies that there is no way to decide via observation $Au + \xi$, $\xi \sim \mathcal{N}(0, \sigma^2)$, with risk $< \epsilon$ on the two simple hypotheses $u = w$ and $u = w'$. As an immediate consequence, the quantity $\phi(w) := \frac{1}{2}[\|w\|_2^2 - \|w'\|_2^2] = \|w\|_2^2[1 - \theta^2(w)]/2$ is a lower bound on the ϵ-risk, on U, of a whatever estimate of $\|u\|_2^2$. We can now try to maximize the resulting lower risk bound over U, thus arriving at the lower bound

$$\mathrm{LwBnd} = \max_{w\in U}\left\{\frac{1}{2}\|w\|_2^2(1 - \theta^2(w))\right\}.$$

On a closest inspection, the latter problem is not a convex one, which does not prevent building a suboptimal solution to this problem, and this is how the lower risk bounds in Table 3.4 were built (we omit the details). We see that the ϵ-risks of our estimates are within a moderate factor from the optimal ones.

Figure 3.3 shows empirical error distributions of the estimates built in the three experiments reported in Table 3.4. When simulating the observations and estimates, we used $\mathcal{N}(0, \sigma^2 I_d)$ obse4rvation noise and selected signals in U by maximizing over U randomly selected linear forms.

Finally, we note that with our design parameters d, m, θ, σ fixed, we still deal with a family of estimation problems rather than with a single problem, the reason being that our U is ellipsoid with essentially different from each other half-axes, and achievable risks heavily depend on how the right singular vectors of A are oriented with respect to the directions of the half-axes of U, so that the risks of our estimates vary significantly from instance to instance even when the design parameters are fixed. Note also that the "sub-Gaussian experiments" were conducted on exactly the same data as "Gaussian experiments" of the same sizes d, m.

3.4.3 Proof of Proposition 3.4.3

Let us fix $\epsilon \in (0, 1)$. Setting

$$\rho_K = \frac{1}{2}\left[\widehat{\Psi}_+^{K,\epsilon}(\bar{h}, \bar{H}) + \widehat{\Psi}_-^{K,\epsilon}(\bar{h}, \bar{H})\right]$$

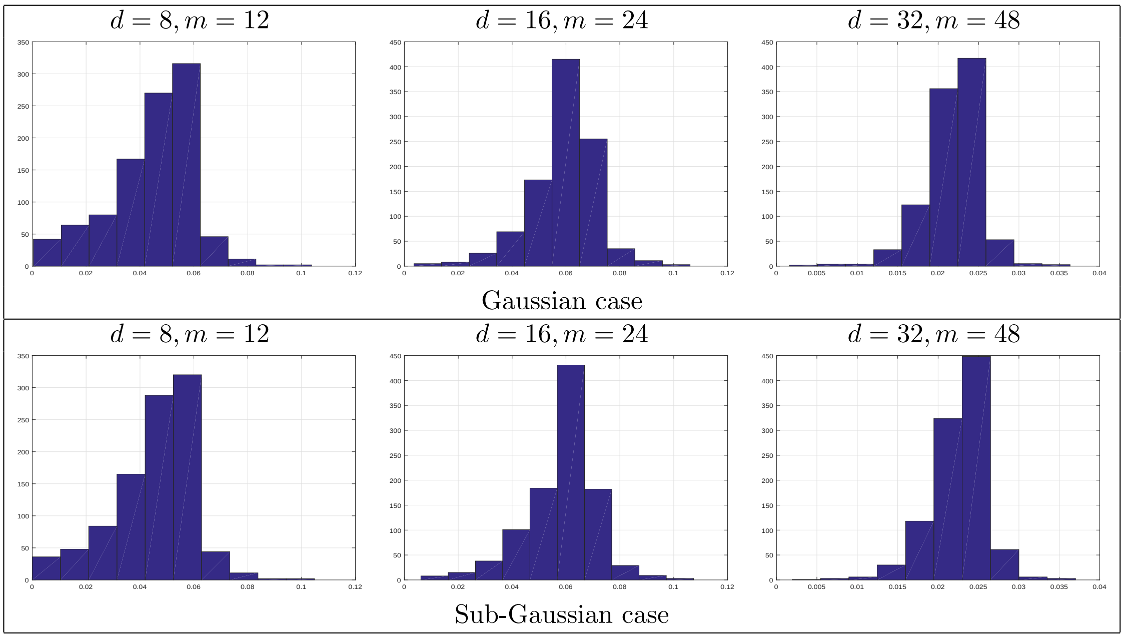


Figure 3.3: Histograms of recovery errors in experiments, data over 1000 simulations per experiment.

and invoking Corollary 3.4.1, all we need to prove is that in the case of A.1-3 one has

$$\lim\sup_{K\to\infty}\left[\widehat{\Psi}_+^{K,\epsilon}(\bar{h},\bar{H})+\widehat{\Psi}_-^{K,\epsilon}(\bar{h},\bar{H})\right]\leq 0. \tag{3.4.41}$$

To this end note that in our current situation, (3.4.8) and (3.4.13) simplify to

$$\begin{array}{l}
\Phi(h,H;Z)=-\frac{1}{2}\ln\mathrm{Det}(I-\Theta_*^{1/2}H\Theta_*^{1/2})+\frac{1}{2}\mathrm{Tr}\Big(Z\underbrace{\left(B^T\left[\left[\begin{array}{c|c}H&h\\\hline h^T&\end{array}\right]+[H,h]^T\left[\Theta_*^{-1}-H\right]^{-1}[H,h]\right]B\right)}_{\mathcal{Q}(h,H)}\Big),\\
\widehat{\Psi}_+^{K,\epsilon}(h,H)=\inf\limits_{\alpha}\left\{\max\limits_{Z\in\mathcal{Z}}\left[\alpha\Phi(h/\alpha,H/\alpha;Z)-\mathrm{Tr}(QZ)+K^{-1}\alpha\ln(2/\epsilon)\right]:\alpha>0,-\gamma\alpha\Theta_*^{-1}\preceq H\preceq\gamma\alpha\Theta_*^{-1}\right\},\\
\widehat{\Psi}_-^{K,\epsilon}(h,H)=\inf\limits_{\alpha}\left\{\max\limits_{Z\in\mathcal{Z}}\left[\alpha\Phi(-h/\alpha,-H/\alpha;Z)+\mathrm{Tr}(QZ)+K^{-1}\alpha\ln(2/\epsilon)\right]:\alpha>0,-\gamma\alpha\Theta_*^{-1}\preceq H\preceq\gamma\alpha\Theta_*^{-1}\right\},
\end{array}$$

whence

$$\begin{array}{l}
\left[\widehat{\Psi}_+^{K}(\bar{h},\bar{H})+\widehat{\Psi}_-^{K}(\bar{h},\bar{H})\right]\leq\inf\limits_{\alpha}\Big\{\max\limits_{Z_1,Z_2\in\mathcal{Z}}\Big[\alpha\Phi(\bar{h}/\alpha,\bar{H}/\alpha;Z_1)-\mathrm{Tr}(QZ_1)+\Phi(-\bar{h}/\alpha,-\bar{H}/\alpha;Z_1)+\mathrm{Tr}(QZ_2)\\
\qquad\qquad+2K^{-1}\alpha\ln(2/\epsilon)\Big]:\alpha>0,-\gamma\alpha\Theta_*^{-1}\preceq\bar{H}\preceq\gamma\alpha\Theta_*^{-1}\Big\}\\
=\inf\limits_{\alpha}\max\limits_{Z_1,Z_2\in\mathcal{Z}}\Big\{-\frac{1}{2}\alpha\ln\mathrm{Det}\left(I-[\Theta_*^{1/2}\bar{H}\Theta_*^{1/2}]^2/\alpha^2\right)+2K^{-1}\alpha\ln(2/\epsilon)+\mathrm{Tr}(Q[Z_2-Z_1])\\
\qquad\qquad+\frac{1}{2}\left[\alpha\mathrm{Tr}\left(Z_1\mathcal{Q}(\bar{h}/\alpha,\bar{H}/\alpha)\right)+\alpha\mathrm{Tr}\left(Z_2\mathcal{Q}(-\bar{h}/\alpha,-\bar{H}/\alpha)\right)\right]:\alpha>0,-\gamma\alpha\Theta_*^{-1}\preceq\bar{H}\preceq\gamma\alpha\Theta_*^{-1}\Big\}\\
=\inf\limits_{\alpha}\max\limits_{Z_1,Z_2\in\mathcal{Z}}\Big\{-\frac{1}{2}\alpha\ln\mathrm{Det}\left(I-[\Theta_*^{1/2}\bar{H}\Theta_*^{1/2}]^2/\alpha^2\right)+2K^{-1}\alpha\ln(2/\epsilon)\\
\qquad\qquad\underbrace{+\mathrm{Tr}(Q[Z_2-Z_1])+\frac{1}{2}\mathrm{Tr}([Z_1-Z_2]B^T\left[\begin{array}{c|c}\bar{H}&\bar{h}\\\hline h^T&\end{array}\right]B)}_{T(Z_1,Z_2)}+\frac{1}{2}\mathrm{Tr}\left(Z_1B^T[\bar{H},\bar{h}]^T[\alpha\Theta_*^{-1}-\bar{H}]^{-1}[\bar{H},\bar{h}]B\right)\\
\qquad\qquad+\frac{1}{2}\mathrm{Tr}\left(Z_2B^T[\bar{H},\bar{h}]^T[\alpha\Theta_*^{-1}+\bar{H}]^{-1}[\bar{H},\bar{h}]B\right):\alpha>0,-\gamma\alpha\Theta_*^{-1}\preceq\bar{H}\preceq\gamma\alpha\Theta_*^{-1}\Big\}
\end{array}\tag{3.4.42}$$

By (3.4.18) we have $\frac{1}{2}B^T\left[\begin{array}{c|c}\bar{H} & \bar{h}\\ \hline \bar{h}^T & \end{array}\right]B = B^T[C^TQC+J]B$, where the only nonzero entry, if any, in $(d+1)\times(d+1)$ matrix J is in the cell $(d+1,d+1)$. Due to the structure of B, see (3.4.8), we conclude that the only nonzero element, if any, in $\bar{J} = B^TJB$ is in the cell $(m+1,m+1)$, and that

$$\frac{1}{2}B^T\left[\begin{array}{c|c}\bar{H} & \bar{h}\\ \hline \bar{h}^T & \end{array}\right]B = (CB)^TQ(CB)+\bar{J} = Q+\bar{J}$$

(recall that $CB = I_{m+1}$). Now, when $Z_1, Z_2 \in \mathcal{Z}$, the entries of Z_1, Z_2 in the cell $(m+1,m+1)$ both are equal to 1, whence

$$\frac{1}{2}\mathrm{Tr}([Z_1-Z_2]B^T\left[\begin{array}{c|c}\bar{H} & \bar{h}\\ \hline \bar{h}^T & \end{array}\right]B) = \mathrm{Tr}([Z_1-Z_2]Q)+\mathrm{Tr}([Z_1-Z_2]\bar{J}) = \mathrm{Tr}([Z_1-Z_2]Q),$$

implying that the quantity $T(Z_1,Z_2)$ in (3.4.42) is zero, provided $Z_1, Z_2 \in \mathcal{Z}$. Consequently, (3.4.42) becomes

$$\begin{array}{rl}\left[\widehat{\Psi}_+^K(\bar{h},\bar{H})+\widehat{\Psi}_-^K(\bar{h},\bar{H})\right] & \leq \inf\limits_{\alpha}\max\limits_{Z_1,Z_2\in\mathcal{Z}}\Big\{-\frac{1}{2}\alpha\ln\mathrm{Det}\left(I-[\Theta_*^{1/2}\bar{H}\Theta_*^{1/2}]^2/\alpha^2\right)+2K^{-1}\alpha\ln(2/\epsilon)\\ & \qquad +\frac{1}{2}\mathrm{Tr}\left(Z_1B^T[\bar{H},h][\alpha\Theta_*^{-1}-\bar{H}]^{-1}[\bar{H},\bar{h}]^TB\right)\\ & \qquad +\frac{1}{2}\mathrm{Tr}\left(Z_2B^T[\bar{H},\bar{h}]^T[\alpha\Theta_*^{-1}+\bar{H}]^{-1}[\bar{H},\bar{h}]B\right):\alpha>0,-\gamma\alpha\Theta_*^{-1}\preceq\bar{H}\preceq\gamma\alpha\Theta_*^{-1}\Big\}\end{array}\tag{3.4.43}$$

Now, for appropriately selected independent of K real c we have

$$\begin{array}{l}-\frac{1}{2}\alpha\ln\mathrm{Det}\left(I-[\Theta_*^{1/2}\bar{H}\Theta_*^{1/2}]^2/\alpha^2\right)\leq c/\alpha,\\ \frac{1}{2}\mathrm{Tr}\left(Z_1B^T[\bar{H},\bar{h}]^T[\alpha\Theta_*^{-1}-\bar{H}]^{-1}[\bar{H},\bar{h}]B\right)\\ \qquad +\frac{1}{2}\mathrm{Tr}\left(Z_2B^T[\bar{H},\bar{h}]^T[\alpha\Theta_*^{-1}+\bar{H}]^{-1}[\bar{H},\bar{h}]B\right)\leq c/\alpha\ \forall Z_1,Z_2\in\mathcal{Z}\end{array}$$

(recall that $\mathcal{Z}$ is bounded). Consequently, given $\omega>0$, we can find $\alpha=\alpha_\omega>0$ large enough to ensure that

$$-\gamma\alpha_\omega\Theta_*^{-1}\preceq\bar{H}\preceq\gamma\alpha_\omega\Theta_*^{-1}\ \&\ 2c/\alpha_\omega\leq\omega,$$

which combines with (3.4.43) to imply that

$$\left[\widehat{\Psi}_+^K(\bar{h},\bar{H})+\widehat{\Psi}_-^K(\bar{h},\bar{H})\right]\leq\omega+2K^{-1}\alpha_\omega\ln(2/\epsilon),$$

and (3.4.41) follows. □

3.5 Exercises for Lecture 3

† marks more difficult exercises.

Exercise 3.1 . The goal of what follows is to refine the change detection procedure (let us refer to it as to "basic") developed in Section 2.8.3.6. The idea is pretty simple. With the notation from Section 2.8.3.6, in basic procedure, when testing the null hypothesis H_0 vs. signal hypothesis H_t^ρ, we looked at the difference $\zeta_t = \omega_t - \omega_1$ and were trying to decide whether the energy of the deterministic component $x_t - x_1$ of ζ_t is 0, as is the case under H_0, or is $\geq \rho^2$, as is the case under H_t^ρ. Note that if $\sigma \in [\underline{\sigma}, \overline{\sigma}]$ is the actual intensity of the observation noise, then the noise component of ζ_t is $\mathcal{N}(0, 2\sigma^2 I_d)$; other things being equal, the large is the noise in ζ_t, the larger should be ρ to allow for a reliable, with a given reliability level, decision of this sort. Now note that under the hypothesis H_t^ρ, we have $x_1 = ... = x_{t-1}$, so that the deterministic component of the difference $\zeta_t = \omega_t - \omega_1$ is exactly the same as for the difference $\widetilde{\zeta}_t = \omega_t - \frac{1}{t-1}\sum_{s=1}^{t-1}\omega_s$, while the noise component in $\widetilde{\zeta}_t$ is $\mathcal{N}(0, \sigma_t^2 I_d)$ with $\sigma_t^2 = \sigma^2 + \frac{1}{t-1}\sigma^2 = \frac{t}{t-1}\sigma^2$; thus, the intensity of noise in

$\widetilde{\zeta}_t$ is at most the one in ζ_t, and this intensity, in contrast to the one for ζ_t, decreases as t grows. Now goes the exercise:

Let reliability tolerances $\epsilon, \varepsilon \in (0,1)$ be given, and let our goal be to design a system of inferences $\mathcal{T}_t$, $t = 2, 3, ..., K$, which, when used in the same fashion as tests $\mathcal{T}_t^\kappa$ were used in Basic procedure, results in false alarm probability at most ϵ and in probability to miss a change of energy $\geq \rho^2$ at most ε; needless to say, we want to achieve this goal with as small ρ as possible. Think how to utilize the above observation to refine Basic procedure by hopefully reducing (and provably not increasing) the required value of ρ. Implement the Basic and the refined change detection procedures and compare their quality (the resulting values of ρ) on, say, the data used in the experiment reported in Section 2.8.3.6.

Exercise 3.2 In the situation of Section 3.3.4, design of a "good" estimate is reduced to solving convex optimization problem (3.3.18). Note that the objective in this problem is, in a sense, "implicit" – the design variable is f, and the objective is obtained from an explicit convex-concave function of f and (x, y) by maximization over (x, y). While there exist solvers capable to process problems of this type efficiently; however, commonly used of-the-shelf solvers, like `cvx`, cannot handle problems like (3.3.18). The goal of the exercise to follow is to reformulate (3.3.18) as a semidefinite program, thus making it amenable for `cvx`.

On an immediate inspection, the situation we are interested in is as follows. We are given

- a nonempty convex compact set $X \subset \mathbf{R}^n$ along with affine function $M(x)$ taking values in $\mathbf{S}^d$ and such that $M(x) \succeq 0$ when $x \in X$, and
- affine function $F(f) : \mathbf{R}^d \to \mathbf{R}^n$.

Given $\gamma > 0$, this data gives rise to the convex function

$$\Psi(f) = \max_{x \in X} \left\{ F^T(f)x + \gamma\sqrt{f^T M(x) f} \right\},$$

and we want to find a "nice" representation of this function, specifically, want to represent the inequality $\tau \geq \Psi(f)$ by a bunch of LMI's in variables τ, f, and perhaps additional variables.

To achieve our goal, we assume in the sequel that the set

$$X^+ = \{(x, M) : x \in X, M = M(x)\}$$

can be described by a system of linear and semidefinite constraints in variables x, M and additional variables, specifically, the system

$$\begin{array}{lll} (a) & s_i - a_i^T x - b_i^T \xi - \mathrm{Tr}(C_i M) \geq 0, i \leq I & \\ (b) & S - \mathcal{A}(x) - \mathcal{B}(\xi) - \mathcal{C}(M) \succeq 0 & [\mathcal{A}(\cdot), \mathcal{B}(\cdot), \mathcal{C}(\cdot) \text{ are affine functions taking values in } \mathbf{S}^N] \\ (c) & M \succeq 0 & \end{array}$$

We assume that this system of constraints is essentially strictly feasible, meaning that there exists a feasible solution at which the semidefinite constraints $(b), (c)$ are satisfied strictly (i.e., the left hand sides of the LMI's are positive definite).

Now goes the exercise:

1. *Check that $\Psi(f)$ is the optimal value in a semidefinite program, specifically,*

$$\Psi(f) = \max_{x, M, \xi, t} \left\{ F^T(f)x + \gamma t : \left\{ \begin{array}{ll} s_i - a_i^T x - b_i^T \xi - \mathrm{Tr}(C_i M) \geq 0, i \leq I & (a) \\ S - \mathcal{A}(x) - \mathcal{B}(\xi) - \mathcal{C}(M) \succeq 0 & (b) \\ M \succeq 0 & (c) \\ \left[\begin{array}{c|c} f^T M f & t \\ \hline t & 1 \end{array}\right] \succeq 0 & (d) \end{array} \right. \right\}. \qquad (P)$$

2. *Passing from (P) to the semidefinite dual of (P), build explicit semidefinite representation of Ψ, that is, an explicit system $\mathcal{S}$ of LMI's in variables f, τ and additional variables u such that*
$$\{\tau \geq \Psi(f)\} \Leftrightarrow \{\exists u : (\tau, f, u) \text{ satisfies } \mathcal{S}\}.$$

Exercise 3.3 † Consider the situation as follows: given an $m \times n$ "sensing matrix" A which is stochastic– with columns from the probabilistic simplex $\Delta_m = \{v \in \mathbf{R}^m : v \geq 0, \sum_i v_i = 1\}$ and a nonempty closed subset U of Δ_n, we observe M-element, $M > 1$, i.i.d. sample $\zeta^M = (\zeta_1, ..., \zeta_M)$ with ζ_k drawn from the discrete distribution Au_*, where u_* is an unknown probabilistic vector ("signal") known to belong to U. We treat the discrete distribution Au, $u \in \Delta_n$, as a distribution on the vertices $e_1, ..., e_m$ of Δ_m, so that possible values of ζ_k are basic orths $e_1, ..., e_m$ in $\mathbf{R}^m$. Our goal is to recover the value at u_* of a given quadratic form

$$F(u) = u^T Q u + 2q^T u.$$

Observe that for $u \in \Delta_n$, we have $u = [uu^T]\mathbf{1}_n$, where $\mathbf{1}_k$ is the all-ones vector in $\mathbf{R}^k$. This observation allows to rewrite $F(u)$ as a homogeneous quadratic form:

$$F(u) = u^T \bar{Q} u, \ \bar{Q} = Q + [q\mathbf{1}_n^T + \mathbf{1}_n q^T]. \tag{3.5.1}$$

The goal of Exercise is to follow the approach developed in Section 3.4.1 for the Gaussian case in order to build an estimate $\widehat{g}(\zeta^M)$ of $F(u)$, specifically, estimate as follows.

Let
$$\mathcal{J}_M = \{(i,j) : 1 \leq i < j \leq M\}, J_M = \mathrm{Card}(\mathcal{J}_M).$$
For $\zeta^M = (\zeta_1, ..., \zeta_M)$ with $\zeta_k \in \{e_1, ..., e_m\}$, $1 \leq k \leq M$, let
$$\omega_{ij}[\zeta^M] = \frac{1}{2}[\zeta_i\zeta_j^T + \zeta_j\zeta_i^T], \ (i,j) \in \mathcal{J}_M.$$

The estimates we are interested in are of the form

$$\widehat{g}(\zeta^M) = \mathrm{Tr}\Bigg(h \underbrace{\left[\frac{1}{J_M}\sum\nolimits_{(i,j)\in\mathcal{J}_M} \omega_{ij}[\zeta^M]\right]}_{\omega[\zeta^M]}\Bigg) + \kappa \tag{3.5.2}$$

where $h \in \mathbf{S}^m$ and $\kappa \in \mathbf{R}$ are the parameters of the estimate.

Now goes the exercise:

1. *Verify that when ζ_k's stem from signal $u \in U$, the expectation of $\omega[\zeta^M]$ is a linear image $Az[u]A^T$ of the matrix $z[u] = uu^T \in \mathbf{S}^n$: denoting by P_u^M the distribution of ζ^M, we have*
$$\mathbf{E}_{\zeta^M \sim P_u^M}\{\omega[\zeta^M]\} = Az[u]A^T. \tag{3.5.3}$$
Check that when setting
$$\mathcal{Z}_k = \{\omega \in \mathbf{S}^k : \omega \succeq 0, \omega \geq 0, \mathbf{1}_k^T \omega \mathbf{1}_k = 1\},$$
where $x \geq 0$ for a matrix x means that x is entrywise nonnegative, the image of $\mathcal{Z}_n$ under the mapping $z \mapsto AzA^T$ is contained in $\mathcal{Z}_m$.

2. Let $\Delta^k = \{z \in \mathbf{S}^k : z \geq 0, \mathbf{1}_n^T z \mathbf{1}_n = 1\}$*, so that* $\mathcal{Z}_k$ *is the set of all positive semidefinite matrices form* Δ^k*. For* $\mu \in \Delta^m$*, let* P_μ *be the distribution of the random matrix* w *taking values in* $\mathbf{S}^m$*, namely, as follows: the possible values of* w *are matrices of the form* $e^{ij} = \frac{1}{2}[e_i e_j^T + e_j e_i^T]$, $1 \leq i \leq j \leq m$*; for every* $i \leq m$*,* w *takes value* e^{ii} *with probability* μ_{ii}*, and for every* i,j *with* $i < j$*,* w *takes value* e^{ij} *with probability* $2\mu_{ij}$*. Further, let us set*

$$\Phi_1(h;\mu) = \ln\left(\sum_{i,j=1}^m \mu_{ij}\exp\{h_{ij}\}\right) : \mathbf{S}^m \times \Delta^m \to \mathbf{R}, \tag{3.5.4}$$

so that Φ_1 *is a continuous convex-concave function on* $\mathbf{S}^m \times \Delta^m$*.*

2.1. *Prove that*

$$\forall (h \in \mathbf{S}^m, \mu \in \mathcal{Z}_m) : \ln\left(\mathbf{E}_{w\sim P_\mu}\{\exp\{\mathrm{Tr}(hw)\}\}\right) = \Phi_1(h;\mu). \tag{3.5.5}$$

2.2. *Derive from 2.1 that setting*

$$K = K(M) = \lfloor M/2 \rfloor,\ \Phi_M(h;\mu) = K\Phi_1(h/K;\mu) : \mathbf{S}^m \times \Delta^m \to \mathbf{R},$$

Φ_M *is a continuous convex-concave function on* $\mathbf{S}^m \times \Delta^m$ *such* $\Phi_K(0;\mu) = 0$ *for all* $\mu \in \mathcal{Z}_m$*, and whenever* $u \in U$*, the following holds true:*

Let P_u^M *be the distribution of* $\zeta^M = (\zeta_1, ..., \zeta_M)$ *with independent blocks* $\zeta_k \sim Au$*, and let* $P_{u,M}$ *is the distribution of* $\omega = \omega[\zeta^M]$*,* $\zeta^M \sim P_u^M$*. Then*

$$\forall (u \in U, h \in \mathbf{S}^m) : \ln\left(\mathbf{E}_{\omega\sim P_{u,M}}\{\exp\{\mathrm{Tr}(h\omega)\}\}\right) \leq \Phi_M(h; Az[u]A^T),\ z[u] = uu^T. \tag{3.5.6}$$

3. Combine the above observations with Corollary 3.3.1 to arrive at the following result:

Proposition 3.5.1 *In the situation in question, let* $\mathcal{Z}$ *be a convex compact subset of* $\mathcal{Z}_n$ *such that* $uu^T \in \mathcal{Z}$ *for all* $u \in U$*. Given* $\epsilon \in (0,1)$*, let*

$$\begin{array}{rcl}
\Psi_+(h,\alpha) &=& \max\limits_{z\in\mathcal{Z}}\left[\alpha\Phi_M(h/\alpha, AzA^T) - \mathrm{Tr}(\bar{Q}z)\right] : \mathbf{S}^m \times \{\alpha > 0\} \to \mathbf{R},\\
\Psi_-(h,\alpha) &=& \max\limits_{z\in\mathcal{Z}}\left[\alpha\Phi_M(-h/\alpha, AzA^T) + \mathrm{Tr}(\bar{Q}z)\right] : \mathbf{S}^m \times \{\alpha > 0\} \to \mathbf{R}\\
\widehat{\Psi}_+(h) &:=& \inf_{\alpha>0}\{\Psi_+(h,\alpha) + \alpha\ln(2/\epsilon)\}\\
&=& \max\limits_{z\in\mathcal{Z}}\inf\limits_{\alpha>0}\left[\alpha\Phi_M(h/\alpha, AzA^T) - \mathrm{Tr}(\bar{Q}z) + \alpha\ln(2/\epsilon)\right]\\
&=& \max\limits_{z\in\mathcal{Z}}\inf\limits_{\beta>0}\left[\beta\Phi_1(h/\beta, AzA^T) - \mathrm{Tr}(\bar{Q}z) + \frac{\beta}{K}\ln(2/\epsilon)\right] \quad [\beta = K\alpha],\\
\widehat{\Psi}_-(h) &:=& \inf\limits_{\alpha>0}\{\Psi_-(h,\alpha) + \alpha\ln(2/\epsilon)\}\\
&=& \max\limits_{z\in\mathcal{Z}}\inf\limits_{\alpha>0}\left[\alpha\Phi_M(-h/\alpha, AzA^T) + \mathrm{Tr}(\bar{Q}z) + \alpha\ln(2/\epsilon)\right]\\
&=& \max\limits_{z\in\mathcal{Z}}\inf\limits_{\beta>0}\left[\beta\Phi_1(-h/\beta, AzA^T) + \mathrm{Tr}(\bar{Q}z) + \frac{\beta}{K}\ln(2/\epsilon)\right] \quad [\beta = K\alpha].
\end{array} \tag{3.5.7}$$

The functions $\widehat{\Psi}_\pm$ *are real valued and convex on* $\mathbf{S}^m$*, and every candidate solution* h *to the convex optimization problem*

$$\mathrm{Opt} = \min_h\left\{\widehat{\Psi}(h) := \frac{1}{2}\left[\widehat{\Psi}_+(h) + \widehat{\Psi}_-(h)\right]\right\}, \tag{3.5.8}$$

induces the estimate

$$\widehat{g}_h(\zeta^M) = \mathrm{Tr}(h\omega[\zeta^M]) + \kappa(h),\ \kappa(h) = \frac{\widehat{\Psi}_-(h) - \widehat{\Psi}_+(h)}{2}$$

of the functional of interest (3.5.1) *via observation* ζ^M *with* ϵ*-risk on* U *not exceeding* $\rho = \widehat{\Psi}(h)$*:*

$$\forall (u \in U) : \mathrm{Prob}_{\zeta^M \sim P_u^M}\{|F(u) - \widehat{g}_h(\zeta^M)| > \rho\} \leq \epsilon.$$

4. Consider an alternative way to estimate $F(u)$, namely, as follows. Let $u \in U$. Given a pair of independent observations ζ_1, ζ_2 drawn from distribution Au, let us convert them into the symmetric matrix $\omega_{1,2}[\zeta^2] = \frac{1}{2}[\zeta_1\zeta_2^T + \zeta_2\zeta_1^T]$. The distribution $P_{u,2}$ of this matrix is exactly the distribution $P_{\mu(z[u])}$, see item B, where $\mu(z) = AzA^T : \Delta^n \to \Delta^m$. Now, given $M = 2K$ observations $\zeta^{2K} = (\zeta_1, ..., \zeta_{2K})$ stemming from signal u, we can split them into K consecutive pairs giving rise to K observations $\omega^K = (\omega_1, ..., \omega_K)$, $\omega_k = \omega[[\zeta_{2k-1}; \zeta_{2k}]]$ drawn independently of each other from probability distribution $P_{\mu(z[u])}$, and the functional of interest (3.5.1) is a linear function $\mathrm{Tr}(\bar{Q}z[u])$ of $z[u]$. Assume that we are given a set $\mathcal{Z}$ as in the premise of Proposition 3.5.1. Observe that we are in the situation as follows:

> Given K independent identically distributed observations $\omega^K = (\omega_1, ..., \omega_K)$ with $\omega_k \sim P_{\mu(z)}$, where z is unknown signal known to belong to $\mathcal{Z}$, we want to recover the value at z of linear function $G(v) = \mathrm{Tr}(\bar{Q}v)$ of $v \in \mathbf{S}^n$. Besides this, we know that P_μ, for every $\mu \in \Delta^m$, satisfies the relation
>
> $$\forall (h \in \mathbf{S}^m) : \ln\left(\mathbf{E}_{\omega \sim P_\mu}\{\exp\{\mathrm{Tr}(h\omega)\}\}\right) \leq \Phi_1(h; \mu).$$

This situation is the one of Section 3.3.3, with the data specified as

$$\mathcal{H} = \mathcal{E}_H = \mathbf{S}^m, \mathcal{M} = \Delta^m \subset \mathcal{E}_M = \mathbf{S}^m, \Phi = \Phi_1, X := \{z[u] : u \in U\} \subset \mathcal{X} := \mathcal{Z} \subset \mathcal{E}_X = \mathbf{S}^n, \mathcal{A}(z) = AzA^T,$$

and we can use the machinery developed in this Section on order to upper-bound ϵ-risk of affine estimate

$$\mathrm{Tr}\left(h\frac{1}{K}\sum_{k=1}^K \omega_k\right) + \kappa$$

of $G(z[u])$ and to build the best, in terms of the upper risk bound, estimate, see Corollary 3.3.2. On a closest inspection (carry it out!), the associated with the above data functions $\widehat{\Psi}_\pm$ arising in (3.3.17) are exactly the functions $\widehat{\Psi}_\pm$ specified in Proposition 3.5.1 for $M = 2K$. Thus, the just outlined approach to estimating $F(u)$ via stemming from $u \in U$ observations ζ^{2K} results in a family of estimates

$$\widetilde{g}_h(\zeta^{2K}) = \mathrm{Tr}\left(h\frac{1}{K}\sum_{k=1}^K \omega[[\zeta_{2k-1}; \zeta_{2k}]]\right) + \kappa(h),\ h \in \mathbf{S}^m$$

and the upper bound on ϵ-risk of estimate $\widetilde{g}_h$ is $\widehat{\Psi}(h)$, where $\widehat{\Psi}(\cdot)$ is associated with $M = 2K$ according to Proposition 3.5.1, that is, is exactly the same as the offered by Proposition upper bound on the ϵ-risk of the estimate $\widehat{g}_h$. Note, however, that the estimates $\widetilde{g}_h$ and $\widehat{g}_h$ are not identical:

$$\begin{array}{rcl}\widetilde{g}_h(\zeta^{2K}) & = & \mathrm{Tr}\left(h\frac{1}{K}\sum_{k=1}^K \omega_{2k-1,2k}[\zeta^{2K}]\right) + \kappa(h), \\ \widehat{g}_h(\zeta^{2K}) & = & \mathrm{Tr}\left(h\frac{1}{K(2K-1)}\sum_{1\leq i<j\leq 2K} \omega_{ij}[\zeta^{2K}]\right) + \kappa(h).\end{array}$$

Now goes the question:

- *Which one of the estimates $\widetilde{g}_h$, $\widehat{g}_h$ would you prefer, that is, which one of these estimates, in your opinion, exhibits better practical performance?*

To check your intuition, test performances of the estimates by simulation. Here is the story underlying the recommended simulation model:

> "Tomorrow, tomorrow not today, all the lazy people say." Does it make sense to be lazy? Imagine you are supposed to do some job, and should decide whether to do it today, or tomorrow. The reward for the job is drawn by nature at random, with unknown to you time-invariant distribution u on n-element set $\{r_1, ..., r_n\}$, with $r_1 \leq r_2 \leq ... \leq r_n$. Given $2K$ historical observations of the rewards, what is better – to do the job today or tomorrow, that is, is the probability of tomorrow reward to be at least the today one greater than 0.5? What is this probability? How to estimate it from historical data?

Pose the above problem as the one of estimating a quadratic functional $u^T\bar{Q}u$ of distribution u from direct observations ($m = n, A = I_n$). Pick $u \in \Delta_n$ at random and run simulations to check which one of the estimates $\widehat{g}_h$, $\widetilde{g}_h$ works better. To avoid the necessity to solve optimization problem (3.5.8)*, you can use $h = \bar{Q}$, resulting in unbiased estimate of $u^T\bar{Q}u$.*

Exercise 3.4 † What follows is a variation of Exercise 3.3. Consider the situation as follows: We observe K realizations η_k, $k \leq K$, of discrete random variable with p possible values, and $L \geq K$ realizations ζ_ℓ, $\ell \leq L$, of discrete random variable with q possible values. All realizations are independent of each other; η_k's are drawn from distribution Pu, and ζ_ℓ – from distribution Qv, where $P \in \mathbf{R}^{p\times r}$, $Q \in \mathbf{R}^{q\times s}$ are given stochastic "sensing matrices," and u, v are unknown "signals" known to belong to given subsets U, resp., V of probabilistic simplexes Δ_r, resp., Δ_s. Our goal is to recover from observations $\{\eta_k, \zeta_\ell\}$ the value at u, v of a given linear function

$$F(u, v) = u^T F v = \mathrm{Tr}(F[uv^T]^T). \tag{3.5.9}$$

> The "covering story" could be as follows. Imagine that there are two possible actions, say, administering to a patient drug A and drug B. Let u is the probability distribution of a (somehow quantified) outcome of the first action, and v be similar distribution for the second action. Observing what happens when the first action is utilized K, and the second – L times, we could ask ourselves what is the probability of an outcome of the first action to be better than an outcome of the second action. This amounts to computing the probability p of the event "$\eta > \zeta$," where η, ζ are independent of each other discrete real-valued random variables with distributions u, resp., v, and p is a linear function of the "joint distribution" uv^T of η, ζ. This story gives rise to the aforementioned estimation problem with the unit sensing matrices P, Q. Assuming that there are "measurement errors" – instead of observing action's outcome "as is," we observe a realization of random variable with distribution depending, in a prescribed fashion, on the outcome.

As always, we encode the p possible values of η_k by the basic orths $e_1, ..., e_p$ in $\mathbf{R}^p$, and the q possible values of ζ – by the basic orths $f_1, ..., f_q$ in $\mathbf{R}^q$.

We intend to focus on estimates of the form

$$\widehat{g}_{h,\kappa}(\eta^K, \zeta^L) = \left[\frac{1}{K}\sum_k \eta_k\right]^T h \left[\frac{1}{L}\sum_\ell \zeta_\ell\right] + \kappa \qquad [h \in \mathbf{R}^{p\times q}, \kappa \in \mathbf{R}]$$

This is what you are supposed to do:

1. (cf. item B in Exercise 3.3) *Denoting by Δ_{mn} the set of nonnegative $m\times n$ matrices with unit sum of all entries (i.e., the set of all probability distributions on $\{1,...,m\}\times\{1,...,n\}$) and assuming $L\geq K$, let us set*

$$\mathcal{A}(z)=PzQ^T:\mathbf{R}^{r\times s}\to\mathbf{R}^{p\times q}$$

and

$$\begin{array}{rcl}\Phi(h;\mu)&=&\ln\left(\sum_{i=1}^p\sum_{j=1}^q\mu_{ij}\exp\{h_{ij}\}\right):\mathbf{R}^{p\times q}\times\Delta_{pq}\to\mathbf{R},\\ \Phi_K(h;\mu)&=&K\Phi(h/K;\mu):\mathbf{R}^{p\times q}\times\Delta_{pq}\to\mathbf{R}.\end{array}$$

Verify that $\mathcal{A}$ maps Δ_{rs} into Δ_{pq}, Φ and Φ_K are continuous convex-concave functions on their domains, and that for every $u\in\Delta_r$, $v\in\Delta_s$, the following holds true:

> *(!) When $\eta^K=(\eta_1,...,\eta_K)$, $\zeta^L=(\zeta_1,...,\zeta_K)$ with mutually independent $\eta_1,...,\zeta_L$ such that $\eta_k\sim Pu$, $\eta_\ell\sim Qv$ for all k, ℓ, we have*
>
> $$\ln\left(\mathbf{E}_{\eta,\zeta}\left\{\exp\left\{\left[\frac{1}{K}\sum_k\eta_k\right]^T h\left[\frac{1}{L}\sum\nolimits_\ell\zeta_\ell\right]\right\}\right\}\right)\leq\Phi_K(h;\mathcal{A}(uv^T)).\qquad(3.5.10)$$

2. *Combine (!) with Corollary 3.3.1 to arrive at the following analogy of Proposition 3.5.1:*

Proposition 3.5.2 *In the situation in question, let $\mathcal{Z}$ be a convex compact subset of Δ_{rs} such that $uv^T\in\mathcal{Z}$ for all $u\in U$, $v\in V$. Given $\epsilon\in(0,1)$, let*

$$\begin{array}{rcl}\Psi_+(h,\alpha)&=&\max\limits_{z\in\mathcal{Z}}\left[\alpha\Phi_K(h/\alpha,PzQ^T)-\mathrm{Tr}(Fz^T)\right]:\mathbf{R}^{p\times q}\times\{\alpha>0\}\to\mathbf{R},\\ \Psi_-(h,\alpha)&=&\max\limits_{z\in\mathcal{Z}}\left[\alpha\Phi_K(-h/\alpha,PzQ^T)+\mathrm{Tr}(Fz^T)\right]:\mathbf{R}^{p\times q}\times\{\alpha>0\}\to\mathbf{R}\\ \widehat{\Psi}_+(h)&:=&\inf_{\alpha>0}\left\{\Psi_+(h,\alpha)+\alpha\ln(2/\epsilon)\right\}\\ &=&\max\limits_{z\in\mathcal{Z}}\inf\limits_{\alpha>0}\left[\alpha\Phi_K(h/\alpha,PzQ^T)-\mathrm{Tr}(Fz^T)+\alpha\ln(2/\epsilon)\right]\\ &=&\max\limits_{z\in\mathcal{Z}}\inf\limits_{\beta>0}\left[\beta\Phi(h/\beta,PzQ^T)-\mathrm{Tr}(Fz^T)+\frac{\beta}{K}\ln(2/\epsilon)\right]\quad[\beta=K\alpha],\\ \widehat{\Psi}_-(h)&:=&\inf\limits_{\alpha>0}\left\{\Psi_-(h,\alpha)+\alpha\ln(2/\epsilon)\right\}\\ &=&\max\limits_{z\in\mathcal{Z}}\inf\limits_{\alpha>0}\left[\alpha\Phi_K(-h/\alpha,PzQ^T)+\mathrm{Tr}(Fz^T)+\alpha\ln(2/\epsilon)\right]\\ &=&\max\limits_{z\in\mathcal{Z}}\inf\limits_{\beta>0}\left[\beta\Phi(-h/\beta,PzQ^T)+\mathrm{Tr}(Fz^T)+\frac{\beta}{K}\ln(2/\epsilon)\right]\quad[\beta=K\alpha].\end{array}\qquad(3.5.11)$$

The functions $\widehat{\Psi}_\pm$ are real valued and convex on $\mathbf{R}^{p\times q}$, and every candidate solution h to the convex optimization problem

$$\mathrm{Opt}=\min_h\left\{\widehat{\Psi}(h):=\frac{1}{2}\left[\widehat{\Psi}_+(h)+\widehat{\Psi}_-(h)\right]\right\},\qquad(3.5.12)$$

induces the estimate

$$\widehat{g}_h(\eta^K,\zeta^L)=\mathrm{Tr}\left(h\left[\left[\frac{1}{K}\sum_k\eta_k\right]\left[\frac{1}{L}\sum_\ell\zeta-\ell\right]^T\right]^T\right)+\kappa(h),\ \kappa(h)=\frac{\widehat{\Psi}_-(h)-\widehat{\Psi}_+(h)}{2}$$

of the functional of interest (3.5.9) *via observation η^K,ζ^L with ϵ-risk on $U\times V$ not exceeding $\rho=\widehat{\Psi}(h)$:*

$$\forall(u\in U,v\in V):\mathrm{Prob}\{|F(u,v)-\widehat{g}_h(\eta^K,\zeta^L)|>\rho\}\leq\epsilon,$$

the probability being taken w.r.t. the distribution of observations η^K,ζ^L stemming from signals u, v.

Exercise 3.5 † [recovering mixture weights] The problem to be addressed in this Exercise is as follows. We are given K probability distributions $P_1, ..., P_K$ on observation space Ω, and let these distributions have densities $p_k(\cdot)$ w.r.t. some reference measure Π on Ω; we assume that $\sum_k p_k(\cdot)$ is positive on Ω. We are given also N independent observations

$$\omega_t \sim P_\mu,\ t = 1, ..., N,$$

drawn from distribution

$$P_\mu = \sum_{k=1}^{K} \mu_k P_k,$$

where μ is unknown "signal known to belong to the probabilistic simplex $\mathbf{\Delta}_K = \{\mu \in \mathbf{R}^K : \mu \geq 0, \sum_k \mu_k = 1\}$. Given $\omega^N = (\omega_1, ..., \omega_N)$, we want to recover the linear image $G\mu$ of μ, where $G \in \mathbf{R}^{\nu \to K}$ is given.

We intend to measure the risk of a candidate estimate $\widehat{G}(\omega^N) : \Omega \times ... \times \Omega \to \mathbf{R}^\nu$ by the quantity

$$\mathrm{Risk}[\widehat{G}(\cdot)] = \sup_{\mu \in \mathbf{\Delta}} \left[\mathbf{E}_{\omega^N \sim P_\mu \times ... \times P_\mu} \left\{ \|\widehat{G}(\omega^N) - G\mu\|_2^2 \right\} \right]^{1/2}$$

3.5.A. Recovering linear form. Let us start with the case when $G = g^T$ is $1 \times K$ matrix.

3.5.A.1. Preliminaries. To motivate the construction to follow, consider the case when Ω is a finite set (obtained, e.g., by "fine discretization" of the "true" observation space). In this situation our problem becomes an estimation problem in Discrete o.s., specifically, as follows: *given stationary N-repeated observation stemming from discrete probability distribution P_μ affinely parameterized by signal $\mu \in \mathbf{\Delta}_K$, we want to recover a linear form of μ.* It is shown in [92] that in this case (same as when recovering linear forms of signals observed via other simple o.s.'s), a nearly optimal in terms of its risk estimate (see [92] for details) is of the form

$$\widehat{g}(\omega^N) = \frac{1}{N} \sum_{t=1}^{N} \Phi(\omega_t), \tag{!}$$

with properly selected Φ; this "proper selection" is obtained by the techniques of Section 3.3 as applied to regular data specifying discrete distributions, see Section 2.8.1.2 The difficulty with this approach is that as far as computations are concerned, optimal design of Φ requires solving convex optimization problem of design dimension of order of the cardinality of G, and this cardinality could be huge already when d is in the range of tens. By this reason, we intend to simplify the outlined approach: the only thing we intend to inherit from the optimality results of [92] is the simple structure (!) of the estimator; taking this structure for granted, we intend to develop an alternative to [92] and the construction from Section 3.3 way to design Φ. With these alternative designs, we have no theoretical guarantees for the resulting estimates to be near-optimal; we sacrifice these guarantees in order to reduce dramatically the computational effort of building the estimates.

3.5.A.2. Generic estimate. Let us select somehow L functions $F_\ell(\cdot)$ on Ω such that

$$\int F_\ell^2(\omega) p_k(\omega) \Pi(d\omega) < \infty,\ 1 \leq \ell \leq L, 1 \leq k \leq K \tag{3.5.13}$$

With $\lambda \in \mathbf{R}^L$, consider estimate of the form

$$\widehat{g}_\lambda(\omega^N) = \frac{1}{N} \sum_{t=1}^{N} \Phi_\lambda(\omega_t),\ \Phi_\lambda(\omega) = \sum_\ell \lambda_\ell F_\ell(\omega). \tag{3.5.14}$$

1. *Prove that*

$$\begin{array}{l}\mathrm{Risk}[\widehat{g}_\lambda] \leq \overline{\mathrm{Risk}}(\lambda)\\ := \max\limits_{k\leq K}\left[\frac{1}{N}\int\left[\sum_\ell \lambda_\ell F_\ell(\omega)\right]^2 p_k(\omega)\Pi(d\omega) + \left[\int\left[\sum_\ell \lambda_\ell F_\ell(\omega)\right]p_k(\omega)\Pi(d\omega) - g^Te_k\right]^2\right]^{1/2}\\ = \max\limits_{k\leq K}\left[\frac{1}{N}\lambda^TW_k\lambda + [e_k^T[M\lambda-g]]^2\right]^{1/2},\end{array} \tag{3.5.15}$$

where

$$\begin{array}{rcl} M &=& \left[M_{k\ell} := \int F_\ell(\omega)p_k(\omega)\Pi(d\omega)\right]_{\substack{k\leq K\\ \ell\leq L}},\\ W_k &=& \left[[W_k]_{\ell\ell'} := \int F_\ell(\omega)F_{\ell'}(\omega)p_k(\omega)\Pi(d\omega)\right]_{\substack{\ell\leq L\\ \ell'\leq L}},\ 1\leq k\leq K.\end{array}$$

and $e_1, ..., e_K$ are the standard basic orths in $\mathbf{R}^K$.

Note that $\overline{\mathrm{Risk}}(\lambda)$ is a convex function of λ; this function is easy to compute, provided the matrices M and W_k, $k \leq K$, are available. Assuming this is the case, we can solve the convex optimization problem

$$\mathrm{Opt} = \min_{\lambda\in\mathbf{R}^K}\overline{\mathrm{Risk}}(\lambda) \tag{3.5.16}$$

and use the estimate (3.5.14) associated with optimal solution to this problem; the risk of this estimate will be upper-bounded by Opt.

3.5.A.3. Implementation. The question we arrive at is the "Measurement Design" question: what is a "presumably good," in terms of the (upper bound Opt on the) risk of the estimate (3.5.14) yielded by an optimal solution to (3.5.16), selection of L and of the functions F_ℓ, $1 \leq \ell \leq L$? We are about to consider three related options – *naive*, *basic*, and *Maximum Likelihood* (ML).

Naive option is to take $F_\ell = p_\ell$, $1 \leq \ell \leq L = K$, assuming that this selection meets (3.5.13). For the sake of definiteness, consider the "Gaussian case," where $\Omega = \mathbf{R}^d$, Π is the Lebesgue measure, and p_k is Gaussian distribution with parameters ν_k, Σ_k:

$$p_k(\omega) = \frac{\exp\{-\frac{1}{2}(\omega-\nu_k)^T\Sigma_k^{-1}(\omega-\nu_k)\}}{\sqrt{(2\pi)^d\mathrm{Det}(\Sigma_k)}}.$$

In this case, the Naive option leads to easily computable matrices M and W_k appearing in (3.5.15).

2. *Check that in the Gaussian case, setting*

$$\begin{array}{rcl}\Sigma_{k\ell} &=& [\Sigma_k^{-1}+\Sigma_\ell^{-1}]^{-1},\\ \Sigma_{k\ell m} &=& [\Sigma_k^{-1}+\Sigma_\ell^{-1}+\Sigma_m^{-1}]^{-1},\\ \chi_k &=& \Sigma_k^{-1}\nu_k,\\ \alpha_{k\ell} &=& \frac{\sqrt{\mathrm{Det}(\Sigma_{k\ell})}}{\sqrt{(2\pi)^d\mathrm{Det}(\Sigma_k)\mathrm{Det}(\Sigma_\ell)}},\\ \beta_{k\ell m} &=& \frac{\sqrt{\mathrm{Det}(\Sigma_{k\ell m})}}{(2\pi)^d\sqrt{\mathrm{Det}(\Sigma_k)\mathrm{Det}(\Sigma_\ell)\mathrm{Det}(\Sigma_m)}},\end{array}$$

we have

$$\begin{array}{rcl} M_{k\ell} &:=& \int p_\ell(\omega)p_k(\omega)\Pi(d\omega)\\ &=& \alpha_{k\ell}\exp\left\{\frac{1}{2}\left[[\chi_k+\chi_\ell]^T\Sigma_{k\ell}[\chi_k+\chi_\ell] - \chi_k^T\Sigma_k\chi_k - \chi_\ell^T\Sigma_\ell\chi_\ell\right]\right\},\\ [W_k]_{\ell m} &:=& \int p_\ell(\omega)p_m(\omega)p_k(\omega)\Pi(d\omega)\\ &=& \beta_{k\ell m}\exp\left\{\frac{1}{2}\left[[\chi_k+\chi_\ell+\chi_m]^T\Sigma_{k\ell m}[\chi_k+\chi_\ell+\chi_m]\right.\right.\\ && \left.\left.-\chi_k^T\Sigma_k\chi_k - \chi_\ell^T\Sigma_\ell\chi_\ell - \chi_m^T\Sigma_m\chi_m\right]\right\}.\end{array}$$

Basic option. On a close inspection, Naive option does not make much sense: when replacing the reference measure Π with another measure Π' which has positive density $\theta(\cdot)$ w.r.t. Π, the densities p_k are updated according to $p_k(\cdot) \mapsto p_k'(\cdot) = \theta(\cdot)p(\cdot)$, so that selecting $F_\ell' = p_\ell'$, the matrices M and W_k become M' and W_k' with

$$\begin{array}{rcl} M'_{k\ell} &=& \int \frac{p_k(\omega)p_\ell(\omega)}{\theta^2(\omega)}\Pi'(d\omega) = \int \frac{p_k(\omega)p_\ell(\omega)}{\theta(\omega)}\Pi(d\omega),\\ {[W_k']}_{\ell m} &=& \int \frac{p_k(\omega)p_\ell(\omega)p_m(\omega)}{\theta^3(\omega)}\Pi'(d\omega) = \int \frac{p_k(\omega)p_\ell(\omega)}{\theta^2(\omega)}\Pi(d\omega). \end{array}$$

We see that in general $M \neq M'$ and $W_k \neq W_k'$, which makes the Naive option unnatural. *Basic* option is to take

$$L = K,\ F_\ell(\omega) = \pi(\omega) := \frac{p_\ell(\omega)}{\sum_k p_k(\omega)}.$$

The motivation is that the functions F_ℓ remain intact when replacing Π with Π', so that here $M = M'$ and $W_k = W_k'$, which is natural. Besides this, there are statistical arguments in favor of Basic option, namely, as follows. Let Π_* be the measure with the density $\sum_k p_k(\cdot)$ w.r.t. Π; taken w.r.t. Π_*, the densities of P_k are exactly the above $\pi_k(\cdot)$, and $\sum_k \pi_k(\omega) \equiv 1$. Now, (3.5.15) says that the risk of estimate $\widehat{g}_\lambda$ can be upper-bounded by the function $\overline{\mathrm{Risk}}(\lambda)$ defined in (3.5.15), and this function, in turn, can be upper-bounded by the function

$$\begin{array}{rcl} \mathrm{Risk}^+(\lambda) &:=& \left[\frac{1}{N}\sum_k \int \left[\sum_\ell \lambda_\ell F_\ell(\omega)\right]^2 p_k(\omega)\Pi(d\omega) + \max_k \left[\int \left[\sum_k \lambda_\ell F_\ell(\omega)\right] p_k(\omega)\Pi(d\omega) - g^T e_k\right]^2\right]^{1/2}\\ &=& \left[\frac{1}{N}\int \left[\sum_\ell \lambda_\ell F_\ell(\omega)\right]^2 \Pi_*(d\omega) + \max_k \left[\int \left[\sum_k \lambda_\ell F_\ell(\omega)\right] \pi_k(\omega)\Pi_*(d\omega) - g^T e_k\right]^2\right]^{1/2}\\ &\leq& \overline{\mathrm{Risk}}(\lambda) \end{array}$$

(we just have said that the maximum of K nonnegatve quantities is at least their sum, and the latter is at most K times the maximum of the quantities). Consequently, the risk of the estimate (3.5.14) stemming from an optimal solution to (3.5.16) can be upper-bounded by the quantity

$$\mathrm{Opt}^+ := \min_\lambda \mathrm{Risk}^+(\lambda) \quad [\geq \mathrm{Opt} := \max_\lambda \overline{\mathrm{Risk}}(\lambda)].$$

Now goes the punchline:

3.1. *Prove that both the quantities* Opt *defined in (3.5.16) and the above* Opt^+ *depend only on the linear span of the functions F_ℓ, $\ell = 1, ..., L$, not on how the functions F_ℓ are selected in this span.*

3.2. *Prove that the selection $F_\ell = \pi_\ell$, $1 \leq \ell \leq L = K$, minimizes* Opt^+ *among all possible selections L, $\{F_\ell\}_{\ell=1}^L$ satisfying (3.5.13).*

Conclude that the selection $F_\ell = \pi_\ell$, $1 \leq \ell \leq L = K$, while not necessary optimal in terms of Opt*, definitely is meaningful: this selection optimizes the natural upper bound* Opt^+ *on* Opt*. Observe that* $\mathrm{Opt}^+ \leq K\mathrm{Opt}$*, so that optimizing instead of* Opt *the upper bound* Opt^+*, although crude, is not completely meaningless.*

A downside of Basic option is that it seems problematic to get closed form expressions for the associated matrices M and W_k, see (3.5.15). For example, in the Gaussian case, Naive choice of F_ℓ's allows to represent M and W_k in an explicit closed form; in contrast to this, when selecting $F_\ell = \pi_\ell$, $\ell \leq L = K$, seemingly the only way to get M and W_k is to use Monte-Carlo simulations. This being said, we indeed can use Monte-Carlo simulations to compute M and W_k, provided we can sample from distributions $P_1, ..., P_K$. In this respect, it should be stressed that with $F_\ell \equiv \pi_\ell$, the entries in M and W_k are expectations, w.r.t. $P_1, ..., P_K$, of *bounded in magnitude by 1*, and thus well-suited for Monte-Carlo simulation, functions of ω.

Maximum Likelihood option. This choice of $\{F_\ell\}_{\ell \leq L}$ follows the idea of discretization Exercise was started with. Specifically, we split Ω into L cells $\Omega_1, ..., \Omega_L$ in such a way that the intersection of any two different cells is of Π-measure zero, and treat as our observations not the actual observations ω_t, but the indexes of the cells ω_t's belong to. With our estimation scheme, this is the same as to select F_ℓ as the characteristic function of Ω_ℓ, $\ell \leq L$ Assuming that for distinct k, k' the densities P_k, $p_{k'}$ differ from each other Π-almost surely, the simplest discretization independent of how the reference measure is selected is the Maximum Likelihood discretization

$$\Omega_\ell = \{\omega : \max_k p_k(\omega) = p_\ell(\omega)\},\ 1 \leq \ell \leq L = K;$$

with the ML option, we take, as F_ℓ's, the characteristic functions of the just defined sets Ω_ℓ, $1 \leq \ell \leq L = K$. Same as with Basic option, the matrices M and W_k associated with ML option can be found by Monte-Carlo simulation.

We have discussed 3 simple options for selecting F_ℓ's. In applications, one can compute the upper risk bounds Opt, see (3.5.16), associated with every one of these three options, and to use the option with the best – the smallest – risk bound. Alternatively, one can take as $\{F_\ell, \ell \leq L\}$ the union of the three collections yielded by the above options (and, perhaps, further extend this union). Note that the larger is the collection of F_ℓ's, the smaller is the associated Opt, so that the only price for combining different selections is in increasing the computational cost of solving (3.5.16).

3.5.A.4. Illustration. Now goes the experimental part of Exercise:

4.1. Run numerical experiments aimed at comparing with each other the estimates yielded by the above three options (Naive, Basic, ML). Recommended setup:

- *$d = 8$, $K = 90$;*
- *Gaussian case with the covariance matrices Σ_k of P_k selected at random:*

$$S_k = \texttt{rand(d,d)},\ \Sigma_k = \frac{S_kS_k^T}{\|S_k\|^2} \qquad [\|\cdot\|\text{: spectral norm}]$$

  and the expectations ν_k of P_k selected at random from $\mathcal{N}(0, \sigma^2 I_d)$, with $\sigma = 0.1$;
- *values of N: $\{10^s, 1 = 0, 1, ..., 5\}$;*
- *linear form to be recovered: $g^T\mu \equiv \mu_1$.*

4.2†. Utilize Cramer-Rao lower risk bound (see Proposition 4.9.8, Exercise 4.24) to upper-bound the level of conservatism $\frac{\mathrm{Opt}}{\mathrm{Risk}_*}$ of the estimates built in item 4.1. Here Risk$_*$ *is the minimax risk in our estimation problem:*

$$\mathrm{Risk}_* = \inf_{\widehat{g}(\cdot)} \sup_{\mu\in\mathbf{\Delta}} \left[\mathbf{E}_{\omega^N\sim P_\mu\times\ldots\times P_\mu}\left\{|\widehat{g}(\omega^N) - g^T\mu|^2\right\}\right]^{1/2},$$

where inf *is taken over all estimates.*

3.5.B. Recovering linear images. Now consider the case when G is a general-type $\nu \times K$ matrix. The analogy of the estimate $\widehat{g}_\lambda(\cdot)$ is now as follows: with somehow chosen $F_1, ..., F_L$ satisfying (3.5.13), we select a $\nu \times L$ matrix $\Lambda = [\lambda_{i\ell}]$, set

$$\Phi_\Lambda(\omega) = [\textstyle\sum_\ell \lambda_{1\ell}F_\ell(\omega); \sum_\ell \lambda_{2\ell}F_\ell(\omega); ...; \sum_\ell \lambda_{\nu\ell}F_\ell(\omega)]$$

and estimate $G\mu$ by

$$\widehat{G}_\Lambda(\omega^N) = \frac{1}{N}\sum_{t=1}^N \Phi_\lambda(\omega_t).$$

5. *Prove the following analogy of the results of item 3.5.A:*

Proposition 3.5.3 *The risk of the proposed estimator can be upper-bounded as follows:*

$$\begin{array}{rcl} \mathrm{Risk}[\widehat{G}_\Lambda] & := & \max_{\mu\in\mathbf{\Delta}_K}\left[\mathbf{E}_{\omega^N\sim P_\mu\times\ldots\times P_\mu}\left\{\|\widehat{G}(\omega^N)-G\mu\|_2^2\right\}\right]^{1/2} \\ & \leq & \overline{\mathrm{Risk}}(\Lambda) := \max_{k\leq K}\overline{\Psi}(\Lambda,e_k), \\ \overline{\Psi}(\Lambda,\mu) & = & \left[\frac{1}{N}\sum_{k=1}^K\mu_k\mathbf{E}_{\omega\sim P_k}\left\{\|\Phi_\Lambda(\omega)\|_2^2\right\}+\|[\psi_\Lambda-G]\mu\|_2^2\right]^{1/2} \\ & = & \left[\|[\psi_\Lambda-G]\mu\|_2^2+\frac{1}{N}\sum_{k=1}^K\mu_k\int[\sum_{i\leq\nu}[\sum_\ell\lambda_{i\ell}F_\ell(\omega)]^2]P_k(d\omega)\right]^{1/2}, \end{array} \tag{3.5.17}$$

where

$$\mathrm{Col}_k[\psi_\Lambda]=\mathbf{E}_{\omega\sim P_k(\cdot)}\Phi_\Lambda(\omega)=\left[\begin{array}{c}\int[\sum_\ell\lambda_{1\ell}F_\ell(\omega)]P_k(d\omega)\\ \cdots\\ \int[\sum_\ell\lambda_{\nu\ell}F_\ell(\omega)]P_k(d\omega)\end{array}\right],\ 1\leq k\leq K$$

and $e_1,...,e_K$ are the standard basic orths in $\mathbf{R}^K$.

Note that exactly the same reasoning as in the case of $G\mu\equiv g^T\mu$ demonstrates that a reasonable way to select L and F_ℓ, $\ell=1,...,L$, is to set $L=K$ and $F_\ell(\cdot)=\pi_\ell(\cdot)$, $1\leq\ell\leq L$.

Lecture 4

Signal Recovery from Gaussian Observations and Beyond

Overview

In this lecture we address one of the most basic problems of High-Dimensional Statistics, specifically, as follows: given positive definite $m \times m$ matrix Γ, $m \times n$ matrix A, $\nu \times n$ matrix B, and indirect noisy observation

$$\begin{array}{c} \omega = Ax + \xi \\ [A : m \times n, \xi \sim \mathcal{N}(0, \Gamma)] \end{array} \tag{4.1.1}$$

of unknown "signal" x known to belong to a given convex compact subset $\mathcal{X}$ of $\mathbf{R}^n$, we want to recover the image $Bx \in \mathbf{R}^\nu$ of x under a given linear mapping. We focus first on the case where the quality of a candidate recovery $\omega \mapsto \widehat{x}(\omega)$ is quantified by its worst-case, over $x \in \mathcal{X}$, expected $\|\cdot\|_2^2$-error, that is, by the risk

$$\mathrm{Risk}[\widehat{x}(\cdot)|\mathcal{X}] = \sup_{x \in \mathcal{X}} \sqrt{\mathbf{E}_{\xi \sim \mathcal{N}(0,\Gamma)} \left\{ \|\widehat{x}(Ax + \xi) - Bx\|_2^2 \right\}}. \tag{4.1.2}$$

The simplest and the most studied type of recovery is affine one: $\widehat{x}(\omega) = H^T\omega + h$; assuming $\mathcal{X}$ symmetric w.r.t. the origin, we lose nothing when passing from affine estimates to linear ones – those of the form $\widehat{x}_H(\omega) = H^T\omega$. An advantage of linear estimates is that under favorable circumstances (e.g., when $\mathcal{X}$ is an ellipsoid), minimizing risk over linear estimates is an efficiently solvable problem, and there exists huge literature on optimal in terms of their risk linear estimates (see, e.g., [101, 102, 140, 141, 46, 59, 134, 5] and references therein). Moreover, in the case of signal recovery from direct observations in white Gaussian noise (the case of $B = A = I_n$, $\Gamma = \sigma^2 I_n$), there is huge body of results on near-optimality of properly selected linear estimates among *all* possible recovery routines, see, e.g., [88, 152] and references therein; a typical result of this type states that when recovering $x \in \mathcal{X}$ from direct observation $\omega = x + \sigma\xi$, $\xi \sim \mathcal{N}(0, I_m)$ and $\mathcal{X}$ being an ellipsoid of the form

$$\{x \in \mathbf{R}^n : \sum_j j^{2\alpha} x_j^2 \leq L^2\},$$

or the box

$$\{x \in \mathbf{R}^n : j^\alpha |x_j| \leq L, j \leq n\},$$

with fixed $L < \infty$ and $\alpha > 0$, the ratio of the risk of a properly selected linear estimate to the *minimax risk*

$$\mathrm{Risk}_{\mathrm{opt}}[\mathcal{X}] := \inf_{\widehat{x}(\cdot)} \mathrm{Risk}[\widehat{x}|\mathcal{X}] \tag{4.1.3}$$

(the infimum is taken over all estimates, not necessarily linear) remains bounded, or even tends to 1, as $\sigma \to +0$, and this happens *uniformly in* n, α and L being fixed. Similar "near-optimality" results are known for "diagonal" case, where $\mathcal{X}$ is the above ellipsoid/box and A, B, Γ are diagonal matrices. To the best of our knowledge, the only "general" (that is, not imposing severe restrictions on how the geometries of $\mathcal{X}$, A, B, Γ are linked to each other) result on optimality of linear estimates is due to D. Donoho who proved [51, 50] that *when recovering a linear form* (i.e., in the case of one-dimensional Bx), the best, over all linear estimates, risk is within the factor 1.2 of the minimax risk.

The goal of this lecture is to establish a rather general result on near-optimality of properly built linear estimates as compared to all possible estimates. A result of this type is bound to impose some restrictions on $\mathcal{X}$, since there are cases (e.g., the one of a high-dimensional $\|\cdot\|_1$-ball $\mathcal{X}$) where linear estimates are *by far* nonoptimal. Our restrictions on $\mathcal{X}$ reduce to the existence of a special type representation of $\mathcal{X}$ and are satisfied, e.g., when $\mathcal{X}$ is the intersection of $K < \infty$ ellipsoids/elliptic cylinders:

$$\begin{array}{c}\mathcal{X} = \{x \in \mathbf{R}^n : x^T S_k x \leq 1, 1 \leq k \leq K\}.\\ {[S_k \succeq 0, \sum_k S_k \succ 0]}\end{array} \tag{4.1.4}$$

in particular, $\mathcal{X}$ can be a symmetric w.r.t. the origin compact polytope given by $2K$ linear inequalities $-1 \leq s_k^T x \leq 1$, $1 \leq k \leq K$, or, equivalently, $\mathcal{X} = \{x : x^T \underbrace{(s_k s_k^T)}_{S_k} x. \leq 1, k \leq K\}$. Another instructive example is a set of the form $\mathcal{X} = \{x : \|Sx\|_p \leq L\}$, where $p \geq 2$ and S is a matrix with trivial kernel. It should be stressed than while imposing some restrictions on $\mathcal{X}$, *we require nothing from* A, B, *and* Γ, *aside of positive definiteness of the latter matrix*. Our main result (Proposition 4.2.2) states, in particular, that with $\mathcal{X}$ given by (4.1.4) and arbitrary A, B, the risk of properly selected linear estimate $\widehat{x}_{H_*}$ with both H_* and the risk efficiently computable, satisfies the bound

$$\mathrm{Risk}[\widehat{x}_{H_*}|\mathcal{X}] \leq O(1)\sqrt{\ln\left(\frac{O(1)\|B\|^2K^2\kappa^{-1}}{\mathrm{Risk}^2_{\mathrm{opt}}[\mathcal{X}]}\right)}\mathrm{Risk}_{\mathrm{opt}}[\mathcal{X}], \tag{$*$}$$

where $\|B\|$ is the spectral norm of B, κ is the minimal eigenvalue of $\sum_k S_k$, $\mathrm{Risk}_{\mathrm{opt}}[\mathcal{X}]$ is the minimax risk, and $O(1)$ is an absolute constant. Note that the outlined result is an "operational" one – the risk of *provably nearly optimal* estimate and the estimate itself are given by efficient computation. This is in sharp contrast with traditional results of non-parametric statistics, where near-optimal estimates and their risks are given in a "closed analytical form," at the price of severe restrictions on the structure of the "data" $\mathcal{X}$, A, B, Γ. This being said, it should be stressed that one of the crucial components in our construction is quite classical – this is the idea, going back to M.S. Pinsker [135], to bound from below the minimax risk via Bayesian risk associated with properly selected Gaussian prior[1].

The main body of the lecture originates from [97] is organized as follows. Section 4.1.1 contains the classical results on the optimal *Bayesian* recovery of a signal, including Gauss-Markov Theorem. Section 4.2 contains problem formulation (Section 4.2.1), construction of the linear estimate we deal with (Section 4.2.2) and the central result on near-optimality of this estimate (Section 4.2.3). We discuss also the "expressive abilities" of the family of sets (we call them *ellitopes*) to which our main result applies. Section 4.3 contains some extensions. Specifically, we present a version of our main result for the case when the usual worst-case expected $\|\cdot\|_2^2$-risk is replaced with properly

[1] [135] addresses the problem of $\|\cdot\|_2$-recovery of a signal x from direct observations ($A = B = I$) in the case when $\mathcal{X}$ is a high-dimensional ellipsoid with "regularly decreasing half-axes," like $\mathcal{X} = \{x \in \mathbf{R}^n : \sum_j j^{2\alpha}x_j^2 \leq L^2\}$ with $\alpha > 0$. In this case Pinsker's construction shows that as $\sigma \to +0$, the risk of properly built linear estimate is, uniformly in n, $(1 + o(1))$ times the minimax risk. This is much stronger than $(*)$, and it seems to be unlikely that a similarly strong result holds true in the general case underlying $(*)$.

defined *relative* risk (Section 4.3.1) and provide a robust, w.r.t. uncertainty in A, B, version of the estimate (Section 4.3.2). In Section 4.3.4, we show that the key argument underlying the proof of our main result can be used beyond the scope of statistics, specifically, when quantifying the approximation ratio of the semidefinite relaxation bound on the maximum of a quadratic form over an ellitope; this result, important by its own right, is then used in Section 4.3.5 to develop computationally efficient ways to build "presumably good" linear estimates in the case when the recovery error is measured in a norm different from the Euclidean one. In Section 4.4 we extend the results of previous sections from ellitopes to their "matrix analogies" – *spectratopes*. Concluding Sections 4.5 and 4.6 deal with variations of our main setting – one where the error of recovery is measured in a norm different from $\|\cdot\|_2$ and the random observation error not necessarily is zero mean Gaussian, and one where the observation error is "uncertain-but-bounded," i.e., is selected "by nature," perhaps in an adversarial fashion, from a given bounded set.

4.1 Preliminaries from Statistics and Optimization

4.1.1 Preliminaries from Statistics: Gauss-Markov Theorem

In our setting of the recovery problem, the signal x underlying observations (4.1.1) is an "uncertain-but-bounded" entity known to belong to a given set $\mathcal{X}$, and the (squared) risk of an estimate (4.1.2) is the worst, over the signals from $\mathcal{X}$, expected $\|\cdot\|_2^2$ recovery error. We could speak also about *Bayesian* risk of recovering, specifically, assume that the signal x is drawn at random from some probability distribution Π (called *Bayesian prior*), and quantify the squared risk of a candidate estimate $\widehat{x}(\cdot)$ as the average over Π, rather than worst-case over $\mathcal{X}$, expected $\|\cdot\|_2^2$ recovery error:

$$\mathrm{RiskB}[\widehat{x}(\cdot)|\Pi] = \sqrt{\mathbf{E}_{(\eta,\zeta)\sim\Pi\times\mathcal{N}(0,\Gamma)}\left\{\|\widehat{x}(A\eta+\zeta)-B\eta\|_2^2\right\}} \tag{4.1.5}$$

we will call this *Bayesian risk*. A good news about Bayesian risk is that it, to some extent, is easy to optimize over the estimate:

Proposition 4.1.1 *Let the distribution Π possess finite second moments, so that the distribution $\Pi\times\mathcal{N}(0,\Gamma)$ possesses finite second moments as well. This distribution induces the probability distribution $\mathcal{D}$ of the random pair $(\eta, \omega := A\eta+\zeta)$, and $\mathcal{D}$, in turn, induces the conditional, ω being given, distribution $P_{|\omega}$ of η and the marginal distribution Q of ω, so that for every (deterministic) estimate $\omega\mapsto\widehat{x}(\omega)$ of Bx one has*

$$\mathrm{RiskB}^2[\widehat{x}(\cdot)|\Pi] = \mathbf{E}_{\omega\sim Q}\left\{\mathbf{E}_{\eta\sim P_{|\omega}}\{\|B\eta-\widehat{x}(\omega)\|_2^2\}\right\}. \tag{4.1.6}$$

The optimal, in terms of the Bayesian risk $\mathrm{RiskB}^2[\widehat{x}(\cdot)|\Pi]$ estimate is just the conditional, ω given, expectation of $B\eta$, that is, the estimate

$$\widehat{x}_\Pi(\omega) = B\mathbf{E}_{P_{|\omega}}\{\eta\} = B\int_{\mathbf{R}^n}\eta P_{|\omega}(d\eta).$$

Proof is immediate: the distribution $\mathcal{D}$ possesses finite second moments along with Π, whence Q possesses finite second moments, and $P_{|\omega}$ possesses finite second moments for Q-almost all ω. Relation (4.1.5) reads

$$\mathrm{RiskB}[\widehat{x}(\cdot)|\Pi] = \int_{\omega\in\mathbf{R}^m}\left[\int_{\mathbf{R}^n}\|\widehat{x}(\omega)-B\eta\|_2^2 P_{|\omega}(d\eta)\right]Q(d\omega),$$

and to minimize the right hand side in $\widehat{x}(\cdot)$ means to take

$$\widehat{x}(\omega)\in\operatorname*{Argmin}_u\left[f_\omega(u) := \int_{\mathbf{R}^n}\|u-B\eta\|_2^2 P_{|\omega}(d\eta)\right].$$

The solution to the resulting, parametric in ω, optimization problem is immediate: for almost all ω (specifically, those for which $f_\omega(u)$ makes sense, that is $P_{|\omega}$ has finite second moments), we have

$$f_\omega(u)=\int_{\mathbf{R}^n}\|u-B\eta\|_2^2P_{|\omega}(d\eta)=u^Tu-2u^T\underbrace{B\int_{\mathbf{R}^n}\eta P_{|\omega}(d\eta)}_{\widehat{x}_\Pi(\omega)}+\int_{\mathbf{R}^n}\eta^TB^TB\eta P_{|\omega}(d\eta),$$

so that $f_\omega(u)$ is a strongly convex quadratic form of u; the gradient in u of this quadratic form vanishes when $u=\widehat{x}_\Pi(\omega)$, so that the latter value of u is the unique global minimizer of $f_\omega(\cdot)$. □

As an immediate consequence of Proposition 4.1.6, we get the following classical

Theorem 4.1.1 [Gauss-Markov Theorem] *Let Θ be a symmetric positive semidefinite $N\times N$ matrix, and U, V be $m\times N$ and $n\times N$ matrices such that the matrix $U^T\Theta U$ is positive definite. Let $\zeta\sim\mathcal{N}(0,\Theta)$, and consider the problem of recovering $V\zeta$ via observation of $U\zeta$. The optimal, in terms of its expected quadratic risk $\mathbf{E}\{\|V\zeta-\widehat{x}(U\zeta)\|_2^2\}$ estimate of $V\zeta$ via $U\zeta$ is the linear estimate*

$$\widehat{x}(U\zeta)=\left[V\Theta U^T(U\Theta U^T)^{-1}\right]U\zeta \tag{4.1.7}$$

Proof. Consider the zero mean random vectors $\Delta:=V\zeta-\left[V\Theta U^T(U\Theta U^T)^{-1}\right]U\zeta$ and $U\zeta$. Their covariance is zero:

$$\begin{array}{rcl}\mathbf{E}\left\{\Delta[U\zeta]^T\right\}&=&\mathbf{E}\left\{V\zeta\zeta^TU^T-\left[V\Theta U^T(U\Theta U^T)^{-1}\right]U\zeta\zeta^TU^T\right\}\\&=&V\Theta U^T-V\Theta U^T(U\Theta U^T)^{-1}U\Theta U^T=0,\end{array}$$

and the joint distribution of these two vectors is Gaussian, implying that the vectors are independent. In other words, $V\zeta=\Delta+\left[V\Theta U^T(U\Theta U^T)^{-1}\right][U\zeta]$ with zero mean Δ independent of $U\zeta$, so that $\left[V\Theta U^T(U\Theta U^T)^{-1}\right][U\zeta]$ is the conditional, $U\zeta$ being given, expectation of $V\zeta$. It remains to apply Proposition 4.1.6. □

In our context, we will need the following corollary of Gauss-Markov Theorem:

Corollary 4.1.1 *Let A be an $m\times n$ matrix, B be a $\nu\times n$ matrix, Q be a positive semidefinite symmetric $n\times n$ matrix, and let $\Gamma\succ 0$. Consider the Gaussian random vector $\zeta=(\xi,\eta)$, with independent of each other $\xi\sim\mathcal{N}(0,\Gamma)$ and $\eta\sim\mathcal{N}(0,Q)$. The best, in terms of its expected quadratic risk $\mathbf{E}\left\{\|B\eta-\widehat{x}(\omega)\|_2^2\right\}$, recovery of $B\eta$ via observation $\omega=A\eta+\xi$ is linear, and the expected quadratic risk of this recovery is*

$$\varphi(Q):=\mathrm{Tr}\left(BQB^T-BQA^T[\Gamma+AQA^T]^{-1}AQB^T\right). \tag{4.1.8}$$

Proof. Setting $\zeta=(\xi,\eta)\sim\mathcal{N}(0,\mathrm{Diag}\{\Gamma,Q\})$, $U\zeta=\xi+A\eta$, $V\zeta=B\eta$, we find ourselves in the situation of Gauss-Markov Theorem, with $U^TQU=\mathbf{E}\{U\zeta\zeta^TU\}=\Gamma+AQA^T$, which is a positive definite matrix, as required by Gauss-Markov Theorem. Applying this theorem, we conclude that the minimum of the expected quadratic risk is achieved on a linear recovery $G(\xi+A\eta)$ of $B\eta$. This expected quadratic risk as a function of G is the quadratic function

$$F(G)=\mathrm{Tr}(G\Gamma G^T)+\mathrm{Tr}\left((B-GA)Q(B^T-A^TG^T)\right),$$

and it remains to optimize this strongly convex quadratic form of G in G. At the minimizer, $\bar{G}$, the directional derivative of F along every direction D should vanish, that is,

$$2\mathrm{Tr}(D\Gamma\bar{G}^T)-2\mathrm{Tr}(DAQ(B^T-A^T\bar{G}^T))=0\,\forall D,$$

so that we should have $[\Gamma+AQA^T]\bar{G}^T=AQB^T$, implying that

$$\bar{G}^T=\underbrace{[\Gamma+AQA^T]^{-1}}_{L}\underbrace{AQB^T}_{E}$$

and therefore

$$\begin{array}{rcl}F(\bar{G}) & = & \mathrm{Tr}\Big(\underbrace{\bar{G}\Gamma\bar{G}^T}_{=E^TL^{-1}\Gamma L^{-1}E} + BQB^T - 2\underbrace{BQA^T\bar{G}^T}_{=E^TL^{-1}E} + \underbrace{\bar{G}AQA^T\bar{G}^T}_{=E^TL^{-1}AQA^TL^{-1}E}\Big)\\ & = & \mathrm{Tr}\Big(E^TL^{-1}\underbrace{[\Gamma + AQA^T]}_{=L}L^{-1}E + BQB^T - 2E^TL^{-1}E\Big)\\ & = & \mathrm{Tr}(BQB^T - E^TL^{-1}E),\end{array}$$

as claimed in (4.1.8). □

4.1.2 Preliminaries from Optimization: Executive Summary on Conic Programming and Conic Duality

4.1.2.1 Cones

A *cone* in Euclidean space E is a nonempty set K which is closed w.r.t. taking *conic* combinations of its elements, that is, linear combinations with nonnegative coefficients. Equivalently: $K \subset E$ is a cone if K is nonempty, and

- $x, y \in K \Rightarrow x + y \in K$;
- $x \in K, \lambda \geq 0 \Rightarrow \lambda x \in K$.

It is immediately seen (check it!) that a cone is a convex set. We call a cone K *regular*, if it is closed, *pointed* (that is, does not contain lines, or, equivalently, $K \bigcap [-K] = \{0\}$) and possesses a nonempty interior.

Given a cone $K \subset E$, we can associate with it its *dual cone* K^* defined as

$$K^* = \{y \in E : \langle y, x\rangle \geq 0\, \forall x \in K\};$$

it is immediately seen that whatever be K, K^* is a closed cone, and $K \subset (K^*)^*$. It is well known that

- if K is a closed cone, it holds $K = (K^*)^*$;
- K is a regular cone if and only if K^* is so.

Examples of "useful in applications" regular cones are as follows:

1. *Nonnegative orthants* $\mathbf{R}^d_+ = \{x \in \mathbf{R}^d : x \geq 0\}$
2. *Lorentz cones* $\mathbf{L}^d_+ = \{x \in \mathbf{R}^d : x_n \geq \sqrt{\sum_{i=1}^{n-1} x_i^2}\}$;
3. *Semidefinite cones* $\mathbf{S}^d_+$ comprised of positive semidefinite symmetric $d \times d$ matrices; Semidefinite cone $\mathbf{S}^d_+$ lives in the space $\mathbf{S}^d$ of symmetric matrices equipped with the Frobenius inner product $\langle A, B\rangle = \mathrm{Tr}(AB^T) = \mathrm{Tr}(AB) = \sum_{i,j=1}^d A_{ij}B_{ij}$, $A, B \in \mathbf{S}^d$.

   All listed so far cones are self-dual.
4. Let $\|\cdot\|$ be a norm on $\mathbf{R}^n$. The set $\{[x; t] \in \mathbf{R}^n \times \mathbf{R} : t \geq \|x\|\}$ is a regular cone, and the dual cone is $\{[y; \tau] : \|y\|_* \leq \tau\}$, where

   $$\|y\|_* = \max_x \{x^Ty : \|x\| \leq 1\}$$

   is the norm on $\mathbf{R}^n$ conjugate to $\|\cdot\|$.

Another useful for the sequel example of a regular cone is the *conic hull* of a convex compact set defined as follows. Let $\mathcal{T}$ be a convex compact set with a nonempty interior in Euclidean space E. We can associate with $\mathcal{T}$ its *conic hull*

$$K(\mathcal{T}) = \mathrm{cl}\underbrace{\left\{[t;\tau] \in E^+ = E \times \mathbf{R} : \tau > 0, t/\tau \in \mathcal{T}\right\}}_{K^o(\mathcal{T})}.$$

It is immediately seen that $K(\mathcal{T})$ is a regular cone (check it!), and that to get this cone, one should add to the convex set $K^o(\mathcal{T})$ the origin in E^+. It is also clear that one can "see $\mathcal{T}$ in $K(\mathcal{T})$:" – $\mathcal{T}$ is nothing but the cross-section of the cone $K(\mathcal{T})$ by the hyperplane $\tau = 1$ in $E^+ = \{[t;\tau]\}$:

$$\mathcal{T} = \{t \in E : [t;1] \in K(\mathcal{T})\}$$

It is easily seen (check it!) that the cone $K^*(\mathcal{T})$ dual to $K(\mathcal{T})$ is given by

$$K^*(\mathcal{T}) = \{[s;\sigma] \in \mathbf{E}^+ : \sigma \geq \phi_{\mathcal{T}}(-s)\},$$

where

$$\phi_{\mathcal{T}}(s) = \max_{t\in\mathcal{T}}\langle s,t\rangle$$

is the support function of $\mathcal{T}$.

4.1.2.2 Conic problems and their duals

Given regular cones $K_i \subset E_i$, $1 \leq i \leq m$, consider optimization problem of the form

$$\mathrm{Opt}(P) = \min\left\{\langle c,x\rangle : \begin{array}{l} A_ix - b_i \in K_i,\ i = 1,...,m \\ Rx = r \end{array}\right\}, \tag{P}$$

where $x \mapsto A_ix - b_i$ are affine mappings acting from some Euclidean space E to the spaces E_i where the cones K_i live. Problem in this form is called a *conic problem on the cones* $K_1,...,K_m$; the constraints $A_ix - b_i \in K_i$ on x are called *conic constraints*. We call a conic problem (P) *strictly feasible*, if it admits a *strictly feasible* solution $\bar{x}$, meaning that $\bar{x}$ satisfies the equality constraints and satisfies *strictly*: $A_i\bar{x} - b_i \in \mathrm{int}\,K_i$ – the conic constraints.

One can associate with conic problem (P) its *dual*, which also is a conic problem. The origin of the dual problem is the desire to obtain in a systematic way – *by linear aggregation of conic constraints* – lover bounds on the optimal value $\mathrm{Opt}(P)$ of the *primal* problem (P). Linear aggregation of constraints works as follows: let us equip every one of conic constraints $A_ix - b_i \in K_i$ with aggregation weight, called *Lagrange multiplier*, y_i restricted to reside in the cone K_i^* dual to K_i. Similarly, we equip the system $Rx = r$ of equality constraints in (P) with Lagrange multiplier z – a vector of the same dimension as r. Now let x be a feasible solution to the conic problem, and let $y_i \in K_i^*$, $i \leq m$, z be Lagrange multipliers. By the definition of the dual cone and due to $A_ix - b_i \in K_i$, $y_i \in K_i^*$ we have

$$\langle y_i, A_ix\rangle \geq \langle y_i, b_i\rangle, 1 \leq i \leq m$$

and of course

$$z^TRx \geq r^Tz.$$

Summing all resulting inequalities up, we arrive at the scalar linear inequality

$$\langle R^*z + \sum_i A_i^*y_i, x\rangle \geq r^Tz + \sum_i\langle b_i, y_i\rangle \tag{!}$$

where A_i^* are the conjugates to A_i: $\langle y, A_i x\rangle_{E_i} \equiv \langle A_i^* y, x\rangle_E$, and R^* is the conjugate of R. By its origin, (!) is a consequence of the system of constraints in (P) and as such is satisfied everywhere on the feasible domain of the problem. If we are lucky to get, as the linear function of x in the left hand side of (!), the objective of (P), that is, if

$$R^*z + \sum_i A_i^* y_i = c,$$

(!) imposes a lower bound on the objective of the primal conic problem (P) everywhere on the feasible domain of the primal problem, and the *conic dual* of (P) is the problem

$$\mathrm{Opt}(D) = \max_{y_i,z}\left\{ r^Tz + \sum_i \langle b_i, y_i\rangle : \begin{array}{c} y_i \in K_i^*,\, 1\leq i\leq m \\ R^*z + \sum_{i=1}^m A_i^* y_i = c \end{array} \right\} \tag{D}$$

of maximizing this lower bound on $\mathrm{Opt}(P)$.

The relations between the primal and the dual conic problems are the subject of the standard *Conic Duality Theorem* as follows:

Theorem 4.1.2 [Conic Duality Theorem] *Consider conic problem (P) (where all K_i are regular cones) along with its dual problem (D). Then*

1. *Duality is symmetric: the dual problem (D) is conic, and the conic dual of (D) is (equivalent to) (P);*

2. *Weak duality: It always holds $\mathrm{Opt}(D) \leq \mathrm{Opt}(P)$*

3. *Strong duality: If one of the problems (P), (D) is strictly feasible and bounded*[2]*, then the other problem in the pair is solvable, and the optimal values of the problems are equal to each other. In particular, if both (P) and (D) are strictly feasible, then both problems are solvable with equal optimal values.*

Remark 4.1.1 *While Conic Duality Theorem in the just presented form meets all our subsequent needs, it makes sense to note that in fact Strong Duality part of the theorem can be strengthened by replacing strict feasibility with "essential strict feasibility" defined as follows: conic problem in the form of (P) (or, which is the same, form of (D)) is called essentially strictly feasible, if it admits a feasible solution $\bar{x}$ which satisfies strictly the non-polyhedral conic constraints, that is, $A_i\bar{x} - b_i \in \mathrm{int}\, K_i$ for all i for which the cone K_i is not polyhedral – is not given by a finite list of homogeneous linear inequality constraints.*

The proof of Conic Duality Theorem can be found in numerous sources, e.g., in [127, Section 7.1.3] or in Section E.3 in Appendix.

4.1.2.3 Schur Complement Lemma

We will use the following extremely useful fact:

Lemma 4.1.1 [Schur Complement Lemma] *Symmetric block matrix* $A = \left[\begin{array}{c|c} P & Q^T \\ \hline Q & R \end{array}\right]$ *with $R \succ 0$ is positive (semi)definite if and only if the matrix $P - Q^TR^{-1}Q$ is so.*

[2] For a minimization problem, boundedness means that the objective is bounded from below on the feasible set, for a maximization problem – that it is bounded from above on the feasible set.

Proof. With u, v of the same sizes as P, respectively, R, we have

$$\min_v [u;v]^T A [u;v] = u^T[P - Q^T R^{-1}Q]u$$

(direct computation utilizing the fact that $R \succ 0$). It follows that the quadratic form associated with A is nonnegative everywhere if and only if the quadratic form with the matrix $[P - Q^T R^{-1}Q]$ is nonnegative everywhere (since the latter quadratic form is obtained from the former one by partial minimization). □

4.2 Near-Optimality of Linear Estimates

4.2.1 Situation and goal

Given $m\times n$ matrix A, $\nu\times n$ matrix B, and $m\times m$ matrix $\Gamma\succ 0$, consider the problem of estimating linear image Bx of unknown signal x known to belong to a given set $\mathcal{X}\subset \mathbf{R}^n$ via noisy observation

$$\omega = Ax+\xi,\ \xi\sim\mathcal{N}(0,\Gamma), \tag{4.2.1}$$

where ξ is the observation noise.

4.2.1.1 Ellitopes

From now on we assume that $\mathcal{X}\subset\mathbf{R}^n$ is a set given by

$$\mathcal{X} = \big\{x\in\mathbf{R}^n : \exists (y\in\mathbf{R}^{\bar n}, t\in\mathcal{T}) :\ x = Py,\ y^T S_k y\le t_k,\ 1\le k\le K\big\}, \tag{4.2.2}$$

where

- P is an $n\times\bar n$ matrix,
- $S_k\succeq 0$ are $\bar n\times\bar n$ matrices with $\sum_k S_k\succ 0$,
- $\mathcal{T}$ is a nonempty computationally tractable[3] convex compact subset of $\mathbf{R}^K_+$ intersecting the interior of $\mathbf{R}^K_+$ and such that $\mathcal{T}$ is monotone, meaning that the relations $0\le\tau\le t$ and $t\in\mathcal{T}$ imply that $\tau\in\mathcal{T}$.[4] Note that under our assumptions $\mathrm{int}\,\mathcal{T}\neq\emptyset$.

In the sequel, we refer to a set of the form (4.2.2) with data $[P,\{S_k, 1\le k\le K\},\mathcal{T}]$ satisfying just formulated assumptions as to an *ellitope*, and to (4.2.2) – as to *ellitopic representation* of $\mathcal{X}$. Here are instructive examples of ellitopes (in all these examples, P is the identity mapping; in the sequel, we call ellitopes of this type *basic* ones):

- when $K=1$, $\mathcal{T}=[0,1]$ and $S_1\succ 0$, $\mathcal{X}$ is the ellipsoid $\{x : x^TS_1x\le 1\}$;
- when $K\ge 1$, $\mathcal{T}=\{t\in\mathbf{R}^K :\ 0\le t_k\le 1,\ k\le K\}$, and $\mathcal{X}$ is the intersection

  $$\bigcap_{1\le k\le K}\{x :\ x^TS_kx\le 1\}$$

  of centered at the origin ellipsoids/elliptic cylinders. In particular, when U is a $K\times n$ matrix of rank n with rows u_k^T, $1\le k\le K$, and $S_k=u_ku_k^T$, $\mathcal{X}$ is symmetric w.r.t. the origin polytope $\{x : \|Ux\|_\infty\le 1\}$;

[3] for all practical purposes, it suffices to assume that $\mathcal{T}$ is given by an explicit *semidefinite representation*

$$\mathcal{T}=\{t : \exists w : A(t,w)\succeq 0\},$$

where $A(t,w)$ is a symmetric and affine in t,w matrix.

[4] The latter relation is "for free" – given a nonempty convex compact set $\mathcal{T}\subset\mathbf{R}^K_+$, the right hand side of (4.2.2) remains intact when passing from $\mathcal{T}$ to its "monotone hull" $\{\tau\in\mathbf{R}^K_+ : \exists t\in\mathcal{T} : \tau\le t\}$ which already is a monotone convex compact set.

- when U, u_k and S_k are as in the latter example and $\mathcal{T} = \{t \in \mathbf{R}^K_+ : \sum_k t_k^{p/2} \leq 1\}$ for some $p \geq 2$, we get $\mathcal{X} = \{x : \|Ux\|_p \leq 1\}$.

It should be added that the family of ellitope-representable sets is quite rich: this family admits a "calculus", so that more ellitopes can be constructed by taking intersections, direct products, linear images (direct and inverse) or arithmetic sums of ellitopes given by the above examples. In fact, the property to be an ellitope is preserved by all basic operations with sets preserving convexity and symmetry w.r.t. the origin, see Section 4.8.

As another instructive, in the context of non-parametric statistics, example of an ellitope, consider the situation where our signals x are discretizations of functions of continuous argument running through a compact d-dimensional domain D, and the functions f we are interested in are those satisfying a Sobolev-type smoothness constraint – an upper bound on the $L_p(D)$-norm of $\mathcal{L}f$, where $\mathcal{L}$ is a linear differential operator with constant coefficients. After discretization, this restriction can be modeled as $\|Lx\|_p \leq 1$, with properly selected matrix L. As we already know from the above example, when $p \geq 2$, the set $\mathcal{X} = \{x : \|Lx\|_p \leq 1\}$ is an ellitope, and as such is captured by our machinery. Note also that by the outlined calculus, imposing on the functions f in question *several Sobolev-type smoothness constraints* with parameters $p \geq 2$, still results in a set of signals which is an ellitope.

4.2.1.2 Estimates and their risks

In the outlined situation, a candidate estimate is a Borel function $\widehat{x}(\cdot) : \mathbf{R}^m \to \mathbf{R}^\nu$; given observation (4.2.1), we recover $w = Bx$ as $\widehat{x}(\omega)$. In the sequel, we quantify the quality of an estimate by its worst-case, over $x \in \mathcal{X}$, expected $\|\cdot\|_2^2$ recovery error:

$$\mathrm{Risk}[\widehat{x}|\mathcal{X}] = \sup_{x \in \mathcal{X}} \left[\mathbf{E}_{\xi \sim \mathcal{N}(0,\Gamma)} \left\{\|\widehat{x}(Ax + \xi) - Bx\|_2^2\right\}\right]^{1/2} \tag{4.2.3}$$

and define the optimal, or the *minimax*, risk as

$$\mathrm{Risk}_{\mathrm{opt}}[\mathcal{X}] = \inf_{\widehat{x}(\cdot)} \mathrm{Risk}[\widehat{x}|\mathcal{X}], \tag{4.2.4}$$

where inf is taken over all Borel candidate estimates.

4.2.1.3 Main goal

Main goal of what follows is to demonstrate that a *linear in* ω estimate

$$\widehat{x}_H(\omega) = H^T\omega \tag{4.2.5}$$

with properly selected efficiently computable matrix H is near-optimal in terms of its risk.

Our initial observation is that when replacing matrices A and B with AP and BP, respectively, we pass from the initial estimation problem of interest – one where the signal set $\mathcal{X}$ is given by (4.2.2), and we want to recover Bx, $x \in \mathcal{X}$, via observation (4.2.1), to the *transformed problem*, where the signal set is

$$\bar{X} = \{y \in \mathbf{R}^{\bar{n}} : \exists t \in \mathcal{T} : \; y^T S_k y \leq t_k, \; 1 \leq k \leq K\},$$

and we want to recover $[BP]y$, $y \in \bar{X}$, via observation

$$\omega = [AP]y + \xi.$$

It is obvious that the considered families of estimates (the family of all linear and the family of all estimates), same as the risks of the estimates, remain intact under this transformation; in particular,

$$\mathrm{Risk}[\widehat{x}|\mathcal{X}] = \sup_{y\in\bar{X}} \left[\mathbf{E}_\xi\{\|\widehat{x}([AP]\,y+\xi)-[BP]\,y\|_2^2\}\right]^{1/2}.$$

Therefore, to save notation, from now on, unless explicitly stated otherwise, we assume that matrix P is identity, so that $\mathcal{X}$ is the basic ellitope

$$\mathcal{X} = \left\{x\in\mathbf{R}^n:\ \exists t\in\mathcal{T},\ x^TS_kx\leq t_k,\ 1\leq k\leq K\right\}. \tag{4.2.6}$$

We assume in the sequel that $B\neq 0$, since otherwise one has $Bx=0$ for all $x\in\mathcal{X}$, and the estimation problem is trivial.

4.2.2 Building linear estimate

We start with building a "presumably good" linear estimate. Restricting ourselves to linear estimates (4.2.5), we may be interested in the estimate with the smallest risk, that is, associated with a $\nu\times m$ matrix H which is an optimal solution to the optimization problem

$$\min_H \left\{R(H):=\mathrm{Risk}^2[\widehat{x}_H|\mathcal{X}]\right\}$$

We have

$$\begin{array}{rcl}R(H) &=& \max\limits_{x\in\mathcal{X}}\mathbf{E}_\xi\{\|H^T\omega-Bx\|_2^2\} = \mathbf{E}_\xi\{\|H^T\xi\|_2^2\}+\max\limits_{x\in\mathcal{X}}\|H^TAx-Bx\|_2^2\\ &=& \mathrm{Tr}(H^T\Gamma H)+\max\limits_{x\in\mathcal{X}}x^T(H^TA-B)^T(H^TA-B)x.\end{array}$$

This function, while convex, can be hard to compute. For this reason, we use a linear estimate yielded by minimizing an *efficiently computable convex upper bound* on $R(H)$ which is built as follows. Let $\phi_\mathcal{T}$ be the support function of $\mathcal{T}$:

$$\phi_\mathcal{T}(\lambda)=\max_{t\in\mathcal{T}}\lambda^Tt:\mathbf{R}^K\to\mathbf{R}.$$

Observe that whenever $\lambda\in\mathbf{R}^K_+$ and H are such that

$$(B-H^TA)^T(B-H^TA)\preceq\sum_k\lambda_kS_k, \tag{4.2.7}$$

for $x\in\mathcal{X}$ it holds

$$\|Bx-H^TAx\|_2^2\leq\phi_\mathcal{T}(\lambda). \tag{4.2.8}$$

Indeed, in the case of (4.2.7) and with $y\in\mathcal{X}$, there exists $t\in\mathcal{T}$ such that $y^TS_ky\leq t_k$ for all t, and consequently the vector $\bar{t}$ with the entries $\bar{t}_k=y^TS_ky$ also belongs to $\mathcal{T}$, whence

$$\|Bx-H^TAx\|_2^2=\|Bx-H^TAx\|_2^2\leq\sum_k\lambda_kx^TS_kx=\lambda^T\bar{t}\leq\phi_\mathcal{T}(\lambda),$$

which combines with (4.2.6) to imply (4.2.8).

From (4.2.8) it follows that if H and $\lambda\geq 0$ are linked by (4.2.7), then

$$\begin{array}{rcl}\mathrm{Risk}^2[\widehat{x}_H|\mathcal{X}] &=& \max\limits_{x\in\mathcal{X}}\mathbf{E}\left\{\|Bx-H^T(Ax+\xi)\|_2^2\right\}=\mathrm{Tr}(H^T\Gamma H)+\max\limits_{x\in\mathcal{X}}\|[B-H^TA]x\|_2^2\\ &\leq& \mathrm{Tr}(H^T\Gamma H)+\phi_\mathcal{T}(\lambda).\end{array}$$

We see that the efficiently computable convex function

$$\widehat{R}(H) = \inf_{\lambda}\left\{\mathrm{Tr}(H^T\Gamma H) + \phi_{\mathcal{T}}(\lambda) : (B - H^TA)^T(B - H^TA) \preceq \sum_k \lambda_k S_k, \lambda \geq 0\right\}$$

(which clearly is well defined due to compactness of $\mathcal{T}$ combined with $\sum_k S_k \succ 0$) is an upper bound on $R(H)$.[5] We have arrived at the following result:

Proposition 4.2.1 *In the situation of this Section, the risk of the "presumably good" linear estimate $\widehat{x}_{H_*}(\omega) = H_*^T\omega$ yielded by an optimal solution (H_*, λ_*) to the (clearly solvable) convex optimization problem*

$$\begin{array}{rcl}\mathrm{Opt} &=& \min_{H,\lambda}\left\{\mathrm{Tr}(H^T\Gamma H) + \phi_{\mathcal{T}}(\lambda) : (B - H^TA)^T(B - H^TA) \preceq \sum_k \lambda_k S_k, \lambda \geq 0\right\}\\ &=& \min_{H,\lambda}\left\{\mathrm{Tr}(H^T\Gamma H) + \phi_{\mathcal{T}}(\lambda) : \left[\begin{array}{c|c}\sum_k \lambda_k S_k & B^T - A^TH\\ \hline B - H^TA & I_\nu\end{array}\right] \succeq 0, \lambda \geq 0\right\}\end{array} \tag{4.2.9}$$

is upper-bounded by $\sqrt{\mathrm{Opt}}$.

4.2.2.1 Illustration: Recovering temperature distribution

Situation: A square steel plate was somehow heated at time 0 and left to cool, the temperature along the perimeter of the plate being all the time kept zero. At time t_1, we measure the temperatures at m points of the plate, and want to recover the distribution of the temperature along the plate at another given time t_0, $0 < t_0 < t_1$.

Physics, after suitable discretization of spatial variables, offers the following model of our situation. We represent the distribution of temperature at time t as $(2N-1) \times (2N-1)$ matrix $U(t) = [u_{ij}(t)]_{i,j=1}^{2N-1}$, where $u_{ij}(t)$ is the temperature, at time t, at the point

$$P_{ij} = (p_i, p_j),\ p_k = k/N - 1, \quad 1 \leq i, j \leq 2N - 1$$

of the plate (in our model, this plate occupies the square $S = \{(p, q) : |p| \leq 1, |q| \leq 1\}$). Here positive integer N is responsible for spatial discretization.

For $1 \leq k \leq 2N - 1$, let us specify functions $\phi_k(s)$ on the segment $-1 \leq s \leq 1$ as follows:

$$\phi_{2\ell-1}(s) = c_{2\ell-1}\cos(\omega_{2\ell-1}s),\ \phi_{2\ell}(s) = c_{2\ell}\sin(\omega_{2\ell}s),\ \omega_{2\ell-1} = (\ell - 1/2)\pi,\ \omega_{2\ell} = \ell\pi,$$

where c_k are readily given by the normalization condition $\sum_{i=1}^{2N-1}\phi_k^2(p_i) = 1$; note that $\phi_k(\pm 1) = 0$. It is immediately seen that the matrices

$$\Phi^{k\ell} = [\phi_k(p_i)\phi_\ell(p_j)]_{i,j=1}^{2N-1},\ 1 \leq k, \ell \leq 2N - 1$$

form an orthonormal basis in the space of $(2N-1) \times (2N-1)$ matrices, so that we can write

$$U(t) = \sum_{k,\ell \leq 2N-1} x_{k\ell}(t)\Phi^{k\ell}.$$

The advantage of representing temperature fields in the basis $\{\Phi^{k\ell}\}_{k,\ell \leq 2N-1}$ stems from the fact that in this basis the heat equation governing evolution of the temperature distribution in time

[5] It is well known that when $K = 1$ (i.e., $\mathcal{X}$ is n ellipsoid), the above bounding scheme is exact: $R(\cdot) \equiv \widehat{R}(\cdot)$. For more complicated $\mathcal{X}$'s, $\widehat{R}(\cdot)$ could be larger than $R(\cdot)$, although the ratio $\widehat{R}(\cdot)/R(\cdot)$ is bounded by $O(\log(K))$, see Section 4.3.4.

becomes extremely simple, just

$$\frac{d}{dt}x_{k\ell}(t) = -(\omega_k^2+\omega_\ell^2)x_{k\ell}(t) \Rightarrow x_{k\ell}(t) = \exp\{-(\omega_k^2+\omega_\ell^2)t\}x_{k\ell} \quad ^6$$

Now we can convert the situation into the one considered in our general estimation scheme, namely, as follows:

- We select somehow the discretization parameter N and treat $x = \{x_{k\ell}(0), 1 \leq k, \ell \leq 2N-1\}$ as the signal underlying our observations.
  In every potential application, we can safely upper-bound the magnitudes of the initial temperatures and thus the magnitude of x, say, by a constraint of the form

  $$\sum_{k,\ell} x_{k\ell}^2(0) \leq R^2$$

  with properly selected R, which allows to specify the domain $\mathcal{X}$ of the signal as the Euclidean ball:

  $$\mathcal{X} = \{x \in \mathbf{R}^{(2N-1)\times(2N-1)} : \|x\|_2^2 \leq R^2\}. \tag{4.2.10}$$

- Let the measurements of the temperature at time t_1 be taken along the points $P_{i(\nu),j(\nu)}$, $1 \leq \nu \leq m$ [7], and let them be affected by $\mathcal{N}(0,\sigma^2 I_m)$-noise, so that our observation is

  $$\omega = A(x) + \xi,\ \xi \sim \mathcal{N}(0,\sigma^2 I_m),$$

  where $x \mapsto A(x)$ is the linear mapping from $\mathbf{R}^{(2N-1)\times(2N-1)}$ into $\mathbf{R}^m$ given by

  $$[A(x)]_\nu = \sum_{k,\ell=1}^{2N-1} \mathrm{e}^{-(\omega_k^2+\omega_\ell^2)t_1}\phi_k(p_{i(\nu)})\phi_\ell(p_{j(\nu)})x_{k\ell}(0). \tag{4.2.11}$$

- What we want to recover, are the temperatures at time t_0 taken along some grid, say, the square $(2K-1)\times(2K-1)$ grid $\{Q_{ij} = (r_i, r_j), 1 \leq i, j \leq 2K-1\}$, where $r_i = i/K - 1$, $1 \leq i \leq 2K-1$. In other words, we want to recover $B(x)$, where the linear mapping $x \mapsto B(x)$ from $\mathbf{R}^{(2N-1)\times(2N-1)}$ into $\mathbf{R}^{(2K-1)\times(2K-1)}$ is given by

  $$[B(x)]_{ij} = \sum_{k,\ell=1}^{2N-1} \mathrm{e}^{-(\omega_k^2+\omega_\ell^2)t_0}\phi_k(r_i)\phi_\ell(r_j)x_{k\ell}(0).$$

[6] The explanation is simple: the functions $\phi_{k\ell}(p,q) = \phi_k(p)\phi_\ell(q)$, $k,\ell = 1,2,...$, form an orthogonal basis in $L_2(S)$ and vanish on the boundary of S, and the heat equation

$$\frac{\partial}{\partial t}u(t;p,q) = \left[\frac{\partial^2}{\partial p^2} + \frac{\partial^2}{\partial q^2}\right]u(t;p,q)$$

governing evolution of the temperature field $u(t;p,q)$, $(p,q) \in S$, with time t, in terms of the coefficients $x_{k\ell}(t)$ of the temperature field in the orthogonal basis $\{\phi_{k\ell}(p,q)\}_{k,\ell}$ becomes

$$\frac{d}{dt}x_{k\ell}(t) = -(\omega_k^2+\omega_\ell^2)x_{k\ell}(t).$$

In our discretization, we truncate the expansion of $u(t;p,q)$, keeping only the terms with $k,\ell \leq 2N-1$, and restrict the spatial variables to reside in the grid $\{P_{ij}, 1 \leq i,j \leq 2N-1\}$.

[7] the construction can be easily extended to allow for measurement points outside of the grid $\{P_{ij}\}$.

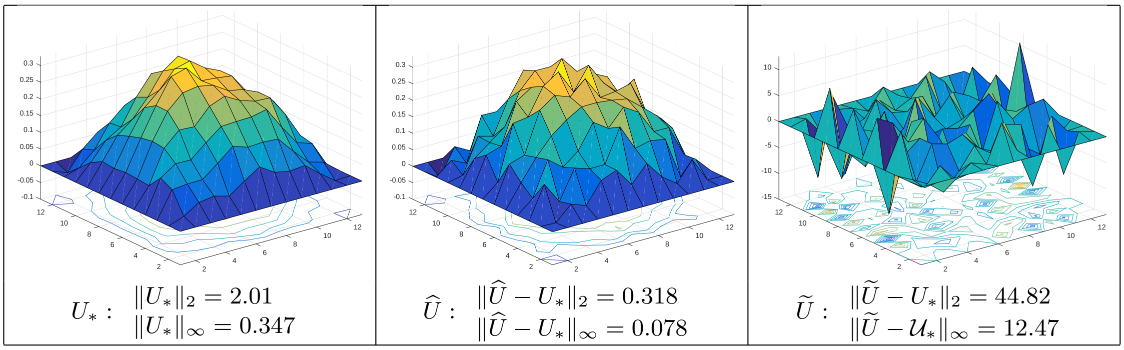


Figure 4.1: . True distribution of temperature $U_* = B(x)$ at time $t_0 = 0.01$ (left) along with its recovery $\widehat{U}$ via optimal linear estimate (center) and the “naive” recovery $\widetilde{U}$ (right).

Ill-posedness. Our problem is a typical example of *ill-posed inverse problem*, where one wants to recover a past state of dynamical system converging exponentially fast to equilibrium and thus “forgetting rapidly” its past. More specifically, in our situation ill-posedness stems from the fact that, as is clearly seen from (4.2.11), contributions of “high frequency” (i.e., with large $\omega_k^2 + \omega_\ell^2$) components $x_{k\ell}(0)$ of the signal to $A(x)$ decrease exponentially fast, with high decay rate, as t_1 grows. As a result, high frequency components $x_{k\ell}(0)$ are impossible to recover from noisy observations of $A(x)$, unless the corresponding time instant t_1 is very small. As a kind of compensation, contributions of high frequency components $x_{k\ell}(0)$ to $B(x)$ are very small, provided t_0 is not too small, implying that there is no necessity to recover well high frequency components, provided they are not huge. Our linear estimate, roughly speaking, seeks for the best tradeoff between these two opposite phenomena, utilizing (4.2.10) as the source of upper bounds on the magnitudes of high frequency components of the signal.

Numerical results. In the experiment to be reported, we used $N = 32$, $m = 100$, $K = 6$, $t_0 = 0.01$, $t_1 = 0.03$ (i.e., temperature is measured at time 0.03 at 100 points selected at random on 63×63 square grid, and we want to recover the temperatures at time 0.01 along 11×11 square grid). We used $R = 15$, that is,

$$\mathcal{X} = \{[x_{k\ell}]_{k,\ell=1}^{63} : \sum_{k,\ell} x_{k\ell}^2 \leq 225\},$$

and $\sigma = 0.001$.

Under the circumstances, the risk of the best linear estimate turns out to be 0.3968. Figure 4.1 shows a sample temperature distribution $B(x) = U_*(t_0)$ at time t_0 stemming from a randomly selected signal $x \in \mathcal{X}$ along with the recovery $\widehat{U}(t_0)$ of U_* by the optimal linear estimate and the naive “least squares” recovery $\widetilde{U}(t_0)$ of U_*. The latter is defined as $B(x_*)$, where x_* is the least squares recovery of signal underlying observation ω:

$$x = x_*(\omega) := \operatorname*{argmin}_x \|A(x) - \omega\|_2.$$

Pay attention to the dramatic difference in performances of the “naive least squares” and the optimal linear estimate

4.2.3 Lower-bounding optimal risk and near-optimality of $\widehat{x}_{H_*}$

Let us consider the convex optimization problem

$$\begin{array}{rl}\mathrm{Opt}_* = \max\limits_{Q,t} \Big\{\varphi(Q) := \mathrm{Tr}\big(B[Q - QA^T(\Gamma + AQA^T)^{-1}AQ]B^T\big), & \\ Q \succeq 0,\, t \in \mathcal{T},\, \mathrm{Tr}(QS_k) \le t_k,\, 1 \le k \le K\Big\} & \quad (4.2.12)\\ = \max\limits_{Q,t} \left\{ \mathrm{Tr}(BQB^T) - \mathrm{Tr}(G) : \begin{array}{l} \left[\begin{array}{cc} G & BQA^T \\ AQB^T & \Gamma + AQA^T \end{array}\right] \succeq 0, \\ Q \succeq 0,\, t \in \mathcal{T},\, \mathrm{Tr}(QS_k) \le t_k,\, 1 \le k \le K \end{array} \right\} & \quad (4.2.13)\end{array}$$

Note that the function $\varphi(Q)$ has a transparent statistical interpretation. Specifically, given an $n \times n$ matrix $Q \succeq 0$, consider two independent Gaussian random vectors, $\xi \sim \mathcal{N}(0, \Gamma)$ and $\eta \sim \mathcal{N}(0, Q)$. We claim that

$$\varphi(Q) = \inf_{G(\cdot):\, \mathbf{R}^m \to \mathbf{R}^\nu} \mathbf{E}_{[\xi,\eta]}\{\|G(\xi + A\eta) - B\eta\|_2^2\}. \quad (4.2.14)$$

Indeed, by the Gauss-Markov Theorem (Theorem 4.1.1), the optimal, in terms of expected $\|\cdot\|_2^2$-error, recovery $G_*(\cdot)$ of $B\eta$ via observation $\xi + A\eta$ – the conditional, given $\xi + A\eta$, expectation of $B\eta$ – is linear, and by Corollary 4.1.1 the corresponding expected $\|\cdot\|_2^2$-error is exactly $\varphi(Q)$.

In the sequel, we set

$$\mathcal{Q} = \{Q \in \mathbf{S}^n : Q \succeq 0, \exists t \in \mathcal{T} : \mathrm{Tr}(QS_k) \le t_k,\, 1 \le k \le K\}. \quad (4.2.15)$$

Note that $\mathcal{Q}$ is a convex compact set due to $\sum_k S_k \succ 0$ combined with compactness of $\mathcal{T}$.

Observe that if (Q, t) is feasible for (4.2.12), then the Gaussian random vector $\eta \sim \mathcal{N}(0, Q)$ belongs to $\mathcal{X}$ "on average" – it satisfies the constraints $\mathbf{E}\{\eta^T S_k \eta\} = \mathrm{Tr}(QS_k) \le t_k$, $k = 1, ..., K$, and $t \in \mathcal{T}$. The lower bounding scheme we intend to implement goes back to Pinsker [135] and heavily relies upon this fact – it bounds from below the minimax, over $x \in \mathcal{X}$, risk of estimating Bx by comparing this risk to the risk of optimal recovery of $B\eta$ in the Gaussian problem, where $\eta \in \mathcal{X}$ with "high probability," as is the case when $Q \in \rho\mathcal{Q}$ with appropriate $\rho < 1$. Specifically, we have the following simple

Lemma 4.2.1 *Given $\rho \in [0, 1]$ and $Q \in \mathcal{Q}_\rho$, let $\eta \sim \mathcal{N}(0, Q)$ and*

$$\delta = \mathrm{Prob}\{\eta \notin \mathcal{X}\}.$$

Then

$$\varphi(Q) \le \mathrm{Risk}^2_{\mathrm{opt}}[\mathcal{X}] + 6M_*^2\sqrt{\delta}, \quad (4.2.16)$$

where

$$M_* = \sqrt{\max_{Q \in \mathcal{Q}} \mathrm{Tr}(BQB^T)}. \quad (4.2.17)$$

For proof, see Section 4.7.1.1.

The next – and main – building block in the proof of near-optimality of linear estimates is

Lemma 4.2.2 *In the situation of this Section (4.2.13) is a conic problem which is strictly feasible and solvable, with the conic dual problem equivalent to (4.2.9). As a consequence, one has*

$$\mathrm{Opt}_* = \mathrm{Opt}. \quad (4.2.18)$$

For proof, fully based on Conic duality, see Section 4.7.1.2.

Let now (Q,t) be an optimal solution to (4.2.12), and let for $0<\rho\leq 1$, $Q_\rho=\rho Q$. Note that $\varphi(Q_\rho)\geq \rho\varphi(Q)=\rho\mathrm{Opt}$, and

$$\mathrm{Tr}(BQ_\rho B^T)=\rho\mathrm{Tr}(BQB^T)\leq \rho M_*^2.$$

In view of Lemma 4.2.1 as applied with Q_ρ in the role of Q, whenever $\rho\in(0,1]$, setting

$$\delta_\rho=\mathrm{Prob}_{\eta\sim\mathcal{N}(0,Q_\rho)}\{\eta\not\in\mathcal{X}\},$$

we have

$$\rho\mathrm{Opt}\leq\varphi(Q_\rho)\leq \mathrm{Risk}^2_{\mathrm{opt}}[\mathcal{X}]+6M_*^2\delta_\rho^{1/2}. \tag{4.2.19}$$

To proceed, we need an upper bound on δ_ρ. It is given by the following simple result (for proof, see Section 4.7.1.3):

Lemma 4.2.3 *Let S and Q be positive semidefinite $n\times n$ matrices with $\rho:=\mathrm{Tr}(SQ)\leq 1$, and let $\eta\sim\mathcal{N}(0,Q)$. Then*

$$\mathrm{Prob}\left\{\eta^TS\eta>1\right\}\leq\inf_\gamma\left\{\exp\left(-\frac{1}{2}\sum_{i=1}^n\ln(1-2\gamma s_i)-\gamma\right):\ 0\leq\gamma<\min_i(2s_i)^{-1}\right\}\leq \mathrm{e}^{-\frac{1-\rho+\rho\ln(\rho)}{2\rho}} \tag{4.2.20}$$

where s_i are the eigenvalues of $Q^{1/2}SQ^{1/2}$.

Now we are done. Indeed, note that the matrix Q_ρ satisfies $\mathrm{Tr}(S_kQ_\rho)\leq\rho t_k$ for some $t\in\mathcal{T}$; applying Lemma 4.2.3 and taking into account (4.2.6), we conclude that

$$\mathrm{Prob}_{\eta\sim\mathcal{N}(0,Q_\rho)}\{\eta\not\in\mathcal{X}\}\leq\sum_{k=1}^K\mathrm{Prob}\{\eta^TS_k\eta>t_k\}\leq K\exp\left\{-\frac{1-\rho+\rho\ln(\rho)}{2\rho}\right\},$$

so that

$$\delta_\rho\leq\min\left[K\exp\left\{-\frac{1-\rho+\rho\ln(\rho)}{2\rho}\right\},1\right]. \tag{4.2.21}$$

It is straightforward to verify that

$$0<\rho\leq 1/6\Rightarrow\exp\left\{-\frac{1-\rho+\rho\ln(\rho)}{2\rho}\right\}\leq\exp\left\{-\frac{1}{4\rho}\right\}. \tag{4.2.22}$$

The latter bound combines with (4.2.19) to yield

$$\rho\,\mathrm{Opt}\leq\mathrm{Risk}^2_{\mathrm{opt}}[\mathcal{X}]+6M_*^2\sqrt{K}\exp\left\{-\frac{1}{8\rho}\right\}\ \forall\rho\in(0,1/6]. \tag{4.2.23}$$

Let us choose $\bar\rho$ according to

$$\frac{1}{\bar\rho}=8\ln\left(\frac{6M_*^2\sqrt{K}}{\mathrm{Risk}^2_{\mathrm{opt}}[\mathcal{X}]}\right)$$

It is immediately seen that $\mathrm{Risk}_{\mathrm{opt}}[\mathcal{X}]\leq M_*$ [8], so that $\bar\rho\in(0,1/6]$ and

$$6M_*^2\sqrt{K}\exp\left\{-\frac{1}{4\bar\rho}\right\}\leq\mathrm{Risk}^2_{\mathrm{opt}}[\mathcal{X}].$$

[8] Indeed, when $x\in\mathcal{X}$, we have $xx^T\in\mathcal{Q}$, see (4.2.15), whence $\|Bx\|\leq M_*$ by (4.2.17), implying that already the risk of the trivial – identically zero – estimate is at most M_*.

Applying (4.2.23) with $\rho = \bar{\rho}$, we get

$$\mathrm{Opt} \leq \frac{2}{\bar{\rho}}\mathrm{Risk}^2_{\mathrm{opt}}[\mathcal{X}] = 16\ln\left(\frac{6M_*^2\sqrt{K}}{\mathrm{Risk}^2_{\mathrm{opt}}[\mathcal{X}]}\right)\mathrm{Risk}^2_{\mathrm{opt}}[\mathcal{X}].$$

Recalling that $\sqrt{\mathrm{Opt}}$ upper-bounds $\mathrm{Risk}[\widehat{x}_{H_*}|\mathcal{X}]$, we have arrived at our main result:

Proposition 4.2.2 *The efficiently computable linear estimate $\widehat{x}_{H_*}(\omega) = H_*^T\omega$ yielded by an optimal solution to the optimization problem* (4.2.9) *is nearly optimal in terms of its risk:*

$$\mathrm{Risk}[\widehat{x}_{H_*}|\mathcal{X}] \leq \sqrt{\mathrm{Opt}} \leq 4\sqrt{\ln\left(\frac{6M_*^2\sqrt{K}}{\mathrm{Risk}^2_{\mathrm{opt}}[\mathcal{X}]}\right)}\mathrm{Risk}_{\mathrm{opt}}[\mathcal{X}] \tag{4.2.24}$$

with M_* given by (4.2.17).

4.2.4 Discussion

The result of Proposition 4.2.2 merits few comments.

4.2.4.1 Simplifying expression for nonoptimality factor

Relation (4.2.24) states that when $\mathcal{X}$ is an ellitope (4.2.2), the risk $\sqrt{\mathrm{Opt}}$ of the efficiently computable linear estimate yielded by (4.2.9) is just by a logarithmic in $\frac{M_*^2K}{\mathrm{Risk}^2_{\mathrm{opt}}[\mathcal{X}]}$ factor worse than the optimal risk $\mathrm{Risk}_{\mathrm{opt}}[\mathcal{X}]$. A minor shortcoming of (4.2.24) is that the "nonoptimality factor" is expressed in terms of unknown to us optimal risk. This can be easily cured. For example, setting

$$\bar{\rho}^{-1} = 40\ln\left(\frac{6M_*^2\sqrt{K}}{\mathrm{Opt}}\right),$$

it is immediately seen that

$$\frac{\bar{\rho}}{2}\mathrm{Opt} \geq 6M_*^2\sqrt{K}\exp\{-\frac{1}{8\rho}\},$$

implying by (4.2.23) that $\frac{1}{2}\bar{\rho}\mathrm{Opt} \leq \mathrm{Risk}^2_{\mathrm{opt}}[\mathcal{X}]$, whence

$$\mathrm{Risk}^2_{\mathrm{opt}}[\mathcal{X}] \geq \left[80\ln\left(\frac{6M_*^2\sqrt{K}}{\mathrm{Opt}}\right)\right]^{-1}\mathrm{Opt}. \tag{4.2.25}$$

Note that all the quantities in the right hand side of (4.2.25) are efficiently computable given the problem data, and that $\sqrt{\mathrm{Opt}}$ is an upper bound on $\mathrm{Risk}[\widehat{x}_{H_*}|\mathcal{X}]$.

Furthermore, if a simple though less precise expression of the factor in terms of this data is required, it can be obtained as follows.

The case of white Gaussian observation noise. To simplify notation, in the rest of Section 4.2.4.1 we assume that the observation noise is white:

$$\Gamma = \sigma^2 I_m \text{ with some } \sigma > 0.$$

Recall that two points $x = x_+$ and $x = -x_+$ of $\mathcal{X}$ can be distinguished through the observation $Ax + \xi$ with maximal probability of error $0 < \alpha < 1$ only if $\|Ax\|_2 \geq c_\alpha\sigma$, $c_\alpha > 0$;[9] by the standard argument one conclude that the risk of estimation of Bx satisfies, for some absolute constant $c > 0$:

$$\mathrm{Risk}^2_{\mathrm{opt}}[\mathcal{X}] \geq \max\left\{\|Bx\|_2 : \|Ax\|_2 \leq c\sigma, \, x \in \mathcal{X}\right\}. \tag{4.2.26}$$

Now let $B = I$, and consider two typical for the traditional non-parametric statistics types of $\mathcal{X}$:

[9] In fact, one can choose $c_\alpha = q_{1-\alpha}$, the $(1-\alpha)$-quantile of the standard normal distribution.

- $\mathcal{X}$ is the ellipsoid $\{x \in \mathbf{R}^n : \sum_i a_i^2 x_i^2 \leq 1\}$ with $0 < a_1 \leq a_2 \leq ... \leq a_n$ (for properly selected a_i this set models the restriction onto a regular n-point grid of functions from a Sobolev ball). Here $K = 1$, $\mathcal{T} = [0,1]$, $S_1 = \mathrm{Diag}\{a_1^2, ..., a_n^2\}$. When choosing $x = te_1$, where e_1 is the first basic orth and $t = \min[1/a_1, c\sigma/\|Ae_1\|_2]$, using (4.2.26) we get $\mathrm{Risk}_{\rm opt}[\mathcal{X}] \geq \min[1/a_1, c\sigma/\|[A]_1\|_2]$ where $[A]_1$ is the first column of A. On the other hand, we have $M_*^2 = a_1^{-2}$, and the simplified risk bound reads
$$\mathrm{Risk}[\widehat{x}_{H_*}|\mathcal{X}] \leq O(1)\sqrt{\ln\left(1 + \frac{\|[A]_1\|_2}{\sigma a_1}\right)}\mathrm{Risk}_{\rm opt}[\mathcal{X}].$$

- $\mathcal{X}$ is the box $\{x \in \mathbf{R}^n : a_i|x_i| \leq 1, 1 \leq i \leq n\}$, where, as above, $0 < a_1 \leq a_2 \leq ... \leq a_n$. Here $K = n$, $\mathcal{T} = [0,1]^n$, $x^T S_k x = a_k^2 x_k^2$, resulting in $M_*^2 = \sum_i a_i^{-2} \leq n a_1^{-2}$. The same bound $\mathrm{Risk}_{\rm opt}[\mathcal{X}] \geq \min[1/a_1, c\sigma/\|[A]_1\|_2]$ holds in this case and, consequently,
$$\mathrm{Risk}[\widehat{x}_{H_*}|\mathcal{X}] \leq O(1)\sqrt{\ln n + \ln\left(1 + \frac{\|[A]_1\|_2}{\sigma a_1}\right)}\mathrm{Risk}_{\rm opt}[\mathcal{X}].$$

Now let B be a general-type matrix, and assume for the sake of simplicity that B has trivial kernel. We associate with the data the following quantities:

- size of $\mathcal{T}$, $T = \max_{t\in\mathcal{T}} \sum_k t_k$, and $\varkappa$ – the minimal eigenvalue of $\sum_k S_k$. Note that for any $x \in \mathcal{X}$, $\sum_k x^T S_k x \leq T$, thus the radius $r(\mathcal{X}) = \max_{x\in\mathcal{X}} \|x\|_2$ of $\mathcal{X}$ satisfies $r(\mathcal{X}) \leq \sqrt{T/\kappa}$;

- *ℓ_1/ℓ_∞-condition number of $\mathcal{T}$*
$$\mathrm{Cond}(\mathcal{T}) = \sqrt{\frac{T}{\max\limits_{t\in\mathcal{T}}\min\limits_{k\leq K} t_k}} = \sqrt{\frac{\max\limits_{t\in\mathcal{T}}\sum_k t_k}{\max\limits_{t\in\mathcal{T}}\min\limits_{k\leq K} t_k}};$$
by our assumptions, $\mathcal{T}$ intersects the interior of $\mathbf{R}^K_+$ and thus $\sqrt{K} \leq \mathrm{Cond}(\mathcal{T}) < \infty$;

- *condition number of B*: $\mathrm{Cond}(B) = \frac{\sigma_{\max}(B)}{\sigma_{\min}(B)}$, where $\sigma_{\max}(B)$ and $\sigma_{\min}(B)$ are, respectively, the largest and the smallest singular values of B.

Corollary 4.2.1 *In the situation of this Section, with $\Gamma = \sigma^2 I_m$ one has*
$$\mathrm{Risk}[\widehat{x}_{H_*}|\mathcal{X}] \leq O(1)\sqrt{\ln\left(K\mathrm{Cond}^2(B)\left[\mathrm{Cond}^2(\mathcal{T}) + \frac{\|A\|^2 T}{\sigma^2\varkappa}\right]\right)}\mathrm{Risk}_{\rm opt}[\mathcal{X}]; \tag{4.2.27}$$
here and in what follows, $O(1)$ stands for a properly selected positive absolute constant.

For proof, see Section 4.7.1.4.

It is worth to note that, surprisingly, the logarithmic factor in (4.2.27) does not depend of the structure of singular spectrum of A, the entity which, as far as the role of A is concerned, is primarily responsible for $\mathrm{Risk}_{\rm opt}[\mathcal{X}]$.

4.2.4.2 Relaxing the symmetry requirement

Sets $\mathcal{X}$ of the form (4.2.2) – we called them ellitopes – are symmetric w.r.t. the origin convex compacts of special structure. This structure is rather flexible, but the symmetry is "built in." We are about to demonstrate that, to some extent, the symmetry requirement can be somehow relaxed. Specifically, assume instead of (4.2.2) that the convex compact set $\mathcal{X}$ known to contain the signals x underlying observations (4.2.1) can be "sandwiched" by two known to us and similar to each other, with coefficient $\alpha \geq 1$, ellitopes:

$$\underbrace{\left\{x \in \mathbf{R}^n : \exists (y \in \mathbf{R}^{\bar{n}}, t \in \mathcal{T}) : x = Py \ \& \ y^T S_k y \leq t_k, \, 1 \leq k \leq K\right\}}_{\underline{\mathcal{X}}} \subset \mathcal{X} \subset \alpha \underline{\mathcal{X}},$$

with S_k and $\mathcal{T}$ possessing the properties postulated in Section 4.2.1. Let Opt and H_* be the optimal value and optimal solution of the optimization problem (4.2.9) associated with the data $S_1, ..., S_K$, $\mathcal{T}$ and matrices $\bar{A} = AP$, $\bar{B} = BP$ in the role of A, B, respectively. It is immediately seen that the risk $\mathrm{Risk}[\widehat{x}_{H_*}|\mathcal{X}]$ of the linear estimate $\widehat{x}_{H_*}(\omega)$ is at most $\alpha\sqrt{\mathrm{Opt}}$. On the other hand, we have $\mathrm{Risk}_{\mathrm{opt}}[\underline{\mathcal{X}}] \leq \mathrm{Risk}_{\mathrm{opt}}[\mathcal{X}]$, and by Proposition 4.2.2 also $\sqrt{\mathrm{Opt}} \leq 4\sqrt{\ln\left(\frac{6M_*^2\sqrt{K}}{\mathrm{Risk}^2_{\mathrm{opt}}[\underline{\mathcal{X}}]}\right)}\mathrm{Risk}_{\mathrm{opt}}[\underline{\mathcal{X}}]$. Taken together, these relations imply that

$$\mathrm{Risk}[\widehat{x}_{H^*}|\mathcal{X}] \leq 4\alpha\sqrt{\ln\left(\frac{6M_*^2\sqrt{K}\alpha}{\mathrm{Risk}^2_{\mathrm{opt}}[\mathcal{X}]}\right)}\mathrm{Risk}_{\mathrm{opt}}[\mathcal{X}]. \tag{4.2.28}$$

In other words, as far as the "level of nonoptimality" of efficiently computable linear estimates is concerned, signal sets $\mathcal{X}$ which can be approximated by ellitopes within a factor α *of order of 1* are nearly as good as the ellitopes. To give an example: it is known that whenever the intersection $\mathcal{X}$ of K elliptic cylinders $\{x : (x - c_k)^T S_k (x - c_k) \leq 1\}$, $S_k \succeq 0$, concentric or not, is bounded and has a nonempty interior, $\mathcal{X}$ can be approximated by an ellipsoid within the factor $\alpha = K + 2\sqrt{K}$ [10]. Assuming w.l.o.g. that the approximating ellipsoid is centered at the origin, the level of nonoptimality of a linear estimate is bounded by (4.2.28) with $O(1)K$ in the role of α. Note that bound (4.2.28) rapidly deteriorates when α grows, and this phenomenon to some extent "reflects the reality." For example, a perfect simplex $\mathcal{X}$ inscribed into the unit sphere in $\mathbf{R}^n$ is in-between two centered at the origin Euclidean balls with the ratio of radii equal to n (i.e. $\alpha = n$). It is immediately seen that with $A = B = I$, $\Gamma = \sigma^2 I$, in the range $\sigma \leq n\sigma^2 \leq 1$ of values of n and σ, we have

$$\mathrm{Risk}_{\mathrm{opt}}[\mathcal{X}] \approx \sqrt{\sigma}, \ \mathrm{Risk}_{\mathrm{opt}}[\widehat{x}_{H_*}|\mathcal{X}] = O(1)\sqrt{n}\sigma,$$

with $\approx$ meaning "up to logarithmic in n/σ factor." In other words, for large $n\sigma$ linear estimates indeed are significantly (albeit not to the full extent of (4.2.28)) outperformed by nonlinear ones.

Another "bad for linear estimates" situation suggested by (4.2.24) is the one where the description (4.2.2) of $\mathcal{X}$, albeit possible, requires a huge value of K. Here again (4.2.24) reflects to some extent the reality: when $\mathcal{X}$ is the unit ℓ_1 ball in $\mathbf{R}^n$, (4.2.2) takes place with $K = 2^{n-1}$; consequently, the factor at $\mathrm{Risk}_{\mathrm{opt}}[\mathcal{X}]$ in the right hand side of (4.2.24) becomes at least $\sqrt{n}$. On the other hand, with $A = B = I$, $\Gamma = \sigma^2 I$, in the range $\sigma \leq n\sigma^2 \leq 1$ of values of n, σ, the risks $\mathrm{Risk}_{\mathrm{opt}}[\mathcal{X}]$, $\mathrm{Risk}_{\mathrm{opt}}[\widehat{x}_{H_*}|\mathcal{X}]$ are basically the same as in the case of $\mathcal{X}$ being the perfect simplex inscribed into the unit sphere in $\mathbf{R}^n$, and linear estimates indeed are "heavily non-optimal" when $n\sigma$ is large.

[10] specifically, setting $F(x) = -\sum_{k=1}^K \ln(1 - (x - c_k)^T S_k (x - c_k)) : \mathrm{int}\,\mathcal{X} \to \mathbf{R}$ and denoting by $\bar{x}$ the *analytic center* $\mathrm{argmin}_{x \in \mathrm{int}\,\mathcal{X}} F(x)$, one has

$$\{x : (x - \bar{x})^T F''(\bar{x})(x - \bar{x}) \leq 1\} \subset \mathcal{X} \subset \{x : (x - \bar{x})^T F''(\bar{x})(x - \bar{x}) \leq [K + 2\sqrt{K}]^2\}.$$

##	X	σ	$\sqrt{\mathrm{Opt}}$	LwB	$\sqrt{\mathrm{Opt}}$/LwB
1	ellipsoid	1.0e-2	0.288	0.153	1.88
2	ellipsoid	1.0e-3	0.103	0.060	1.71
3	ellipsoid	1.0e-4	0.019	0.018	1.06
4	box	1.0e-2	0.698	0.231	3.02
5	box	1.0e-3	0.163	0.082	2.00
6	box	1.0e-4	0.021	0.020	1.06

Table 4.1: Performance of linear estimates (4.2.5), (4.2.9), $m = n = 32$, $B = I$.

4.2.5 Numerical illustration

The "non-optimality factor" θ in the upper bound $\sqrt{\mathrm{Opt}} \leq \theta\mathrm{Risk}_{\mathrm{opt}}[\mathcal{X}]$ from Proposition 4.2.2, while logarithmic, seems to be unpleasantly large. On a closest inspection, one can get less conservative bounds on non-optimality factors. Omitting the details, here are some numerical results. In the six experiments to be reported, we used $n = m = \nu = 32$ and $\Gamma = \sigma^2 I_m$. In the first triple of experiments, $\mathcal{X}$ was the ellipsoid

$$X = \{x \in \mathbf{R}^{32} : \sum_{j=1}^{32} j^2 x_j^2 \leq 1\},$$

that is, P was the identity, $K = 1$, $S_1 = \sum_{j=1}^{32} j^2 e_j e_j^T$ (e_j were basic orths), and $\mathcal{T} = [0,1]$. In the second triple of experiments, $\mathcal{X}$ was the box circumscribed around the above ellipsoid:

$$X = \{x \in \mathbf{R}^{32} : j|x_j| \leq 1,\ 1 \leq j \leq 32\} \qquad [P = I, K = 32, S_k = k^2 e_k e_k^T, k \leq K, \mathcal{T} = [0,1]^K]$$

In all six experiments, B was the identity, and A was a common for all experiments randomly rotated matrix with singular values λ_j, $1 \leq j \leq 32$, forming a geometric progression, with $\lambda_1 = 1$ and $\lambda_{32} = 0.01$. Experiments in a triple differed by the values of σ (0.01,0.001,0.0001).

The results of the experiments are presented in Table 4.1, where, as above, $\sqrt{\mathrm{Opt}}$ is the given by (4.2.9) upper bound on the risk $\mathrm{Risk}[\widehat{x}_{H_*}|X]$ of recovering $Bx = x$, $x \in X$, by the linear estimate yielded by (4.2.5), (4.2.9), LwB is the lower bound on $\mathrm{Risk}_{\mathrm{opt}}[X]$ built as explained above, and the numbers in the last column are (conservative estimates of the) "levels of nonoptimality" of the linear estimates.

4.3 Extensions

4.3.1 Estimation in relative scale

In this Section we consider the setting as follows. Assume that, same as in Section 4.2, we are given a $\nu \times n$ matrix B, and a noisy observation

$$\omega = Ax + \xi,\ \ \xi \sim \mathcal{N}(0, \Gamma),$$

of a signal $x \in \mathcal{X}$ with known $m \times n$ matrix A and $m \times m$ matrix $\Gamma \succ 0$, and we aim to recover $w = Bx$. We are given a positive semidefinite symmetric $n \times n$ matrix S, and we quantify the quality of a candidate estimate $\widehat{x}(\cdot)$ by its *S-risk* – the quantity

$$\mathrm{RiskS}[\widehat{x}|\mathcal{X}] = \inf\left\{\sqrt{\tau} :\ \mathbf{E}\{\|\widehat{x}(Ax+\xi) - Bx\|_2^2\} \leq \tau(1 + x^T Sx)\ \forall x \in \mathcal{X}\right\}. \qquad (4.3.1)$$

The S-risk can be seen as risk with respect to the scale given by the "regularity parameter" $x^T Sx$ of the unknown signal x. In particular, when $S = B^T B$, squared S-risk can be thought of as *relative* risk – the worst, over $x \in \mathbf{R}^n$, expected $\|\cdot\|_2^2$-error of recovering Bx scaled by $\|Bx\|_2^2$; when $S = 0$, we arrive at the usual risk $\mathrm{Risk}[\widehat{x}|\mathcal{X}]$.

Same as in Section 4.2, we assume w.l.o.g. that $\mathcal{X}$ is an ellitope given by (4.2.6) [11]. Besides this, we assume that $B \neq 0$ – otherwise the estimation problem is trivial.

We are about to prove that in the situation in question, efficiently computable linear estimate is near-optimal.

4.3.1.1 Building linear estimate

Given a linear estimate $\widehat{x}_H(\omega) = H^T\omega$ and $\tau \geq 0$, let $\lambda \geq 0$ be such that $[B - H^TA]^T[B - H^TA] \preceq \sum_k \lambda_k S_k + \tau S$, see (4.2.2), implying that for all $x \in \mathcal{X}$, there exists $t = t_x \in \mathcal{T}$ such that

$$\mathbf{E}_\xi\{\|\widehat{x}_H(Ax) - Bx\|_2^2\} \leq x^T[\sum\nolimits_k \lambda_k S_k + \tau S]x + \mathrm{Tr}(H^T\Gamma H) \leq \sum_k t_k\lambda_k + x^TSx + \mathrm{Tr}(H^T\Gamma H),$$

so that for all $x \in \mathcal{X}$

$$\mathbf{E}_\xi\{\|\widehat{x}_H(Ax + \xi) - Bx\|_2^2\} \leq \phi_{\mathcal{T}}(\lambda) + \tau x^TSx + \mathrm{Tr}(H^T\Gamma H),$$

where $\phi_{\mathcal{T}}$ is the support function of $\mathcal{T}$. As a result, whenever H, $\tau \geq 0$ and $\lambda \geq 0$ are such that

$$\mathrm{Tr}(H^T\Gamma H) + \phi_{\mathcal{T}}(\lambda) \leq \tau, \ \ (H^TA - B)^T(H^TA - B) \preceq \sum_k \lambda_k S_k + \tau S,$$

we have

$$\mathrm{RiskS}[\widehat{x}_H|\mathcal{X}] \leq \sqrt{\tau}.$$

We arrive at the convex problem

$$\mathrm{Opt} = \min_{\tau,H,\lambda}\left\{\tau : \left[\begin{array}{cc}\sum_k \lambda_k S_k + \tau S & B^T - A^TH\\ B - H^TA & I_\nu\end{array}\right] \succeq 0,\ \mathrm{Tr}(H^T\Gamma H) + \phi_{\mathcal{T}}(\lambda) \leq \tau,\ \lambda \geq 0\right\}. \quad (4.3.2)$$

The H-component H_* of an optimal solution to this problem yields linear estimate $\widehat{x}_{H_*}(\omega) = H_*^T\omega$ with S-risk $\leq \sqrt{\mathrm{Opt}}$.

4.3.1.2 Lower-bounding the optimal S-risk and near-optimality of $\widehat{x}_{H_*}$

Consider the problem

$$\mathrm{Opt}_* = \max_{W,G,s,v}\left\{\mathrm{Tr}(BWB^T) - \mathrm{Tr}(G) : \begin{array}{l}\left[\begin{array}{cc}G & BWA^T\\ AWB^T & s\Gamma + AWA^T\end{array}\right] \succeq 0,\\ W \succeq 0,\ \mathrm{Tr}(WS_k) \leq v_k,\ 1 \leq k \leq K,\\ \mathrm{Tr}(WS) + s \leq 1,\ [v; s] \in \mathbf{T}\end{array}\right\} \quad (4.3.3)$$

where

$$\mathbf{T} = \mathrm{cl}\{[t;\tau] \in \mathbf{R}^K \times \mathbf{R} : \tau > 0, \tau^{-1}t \in \mathcal{T}\} \subset \mathbf{R}_+^{K+1} \quad (4.3.4)$$

is a closed and pointed convex cone in $\mathbf{R}^{K+1}$ with a nonempty interior. We have the following counterpart of Lemma 4.2.2 for the present setting.

Lemma 4.3.1 *Problem (4.3.3) is strictly feasible and solvable. Furthermore, if $(W, G, [v; s])$ is an optimal solution to (4.3.3), then $s > 0$, and*

$$\mathrm{Opt} = \mathrm{Opt}_* = \mathrm{Tr}\big(B[W - WA^T(s\Gamma + AWA^T)^{-1}AW]B^T\big). \quad (4.3.5)$$

[11] To reduce the general case (4.2.2) to this one with $P = I$ it suffices to "lift" A, B, S to the y-space according to $A \mapsto \bar{A} = AP$, $B \mapsto \bar{B} = BP$, $S \mapsto \bar{S} = P^TSP$ and then replace $\mathcal{X}$ with the set $\mathcal{Y} = \{y \in \mathbf{R}^{\bar{n}} : \exists t \in \mathcal{T} : y^TS_ky \leq t_k, 1 \leq k \leq K\}$.

For proof, see Section 4.7.2.1.

Now let W, v and s stem from an optimal solution to (4.3.3). Then, as we have seen, $s > 0$, and we can set $t = v/s$, so that $t \in \mathcal{T}$. Let also $\rho \in (0, 1]$, and let us put $Q_\rho = \rho W/s$ and $\eta \sim \mathcal{N}(0, Q_\rho)$. We have $s^{-1}W \succeq 0$ and $\mathrm{Tr}(s^{-1}WS_k) \le t_k$, $k \le K$, so that $s^{-1}W \in \mathcal{Q}$ and therefore $Q_\rho \in \rho\mathcal{Q}$. Hence, same as in the case of the usual risk, by Lemma 4.2.3,

$$\mathrm{Prob}\{\eta \notin \mathcal{X}\} \le \delta_\rho := \min\left[K \exp\left\{-\frac{1 - \rho + \rho\ln(\rho)}{2\rho}\right\}, 1\right]. \tag{4.3.6}$$

We also have the following analog of Lemma 4.2.1 (for proof, see Section 4.7.1.1):

Lemma 4.3.2 *Given $\rho \in (0, 1]$, $Q \in \rho\mathcal{Q}$, let $\eta \sim \mathcal{N}(0, Q)$, and let*

$$\delta = \mathrm{Prob}\{\eta \notin \mathcal{X}\}.$$

Then

$$\varphi(Q) \le \mathrm{Risks}^2_{\mathrm{opt}}[\mathcal{X}](1 + \mathrm{Tr}(QS)) + 6M_*^2\sqrt{\delta}, \tag{4.3.7}$$

where M_* is given by (4.2.17), and

$$\mathrm{Risks}_{\mathrm{opt}}[\mathcal{X}] = \inf_{\widehat{x}(\cdot)} \mathrm{RiskS}[\widehat{x}|\mathcal{X}].$$

is the minimax S-risk associated with $\mathcal{X}$.

Now note that

$$\begin{array}{rcl}\varphi(Q_\rho) & = & \mathrm{Tr}\big(B[Q_\rho - Q_\rho A^T(\Gamma + AQ_\rho A^T)^{-1}AQ_\rho]B^T\big)\\ & = & \frac{\rho}{s}\mathrm{Tr}\big(B[W - \rho WA^T(s\Gamma + \rho AWA^T)^{-1}AW]B^T\big) \ge \frac{\rho}{s}\mathrm{Opt}_* = \frac{\rho}{s}\mathrm{Opt}\end{array}$$

(we have used (4.3.5) and the positivity of s). Thus, when applying Lemma 4.3.2 with Q_ρ and δ_ρ in the role of Q and δ, we obtain for all $0 < \rho \le 1$:

$$\begin{array}{rcl}\frac{\rho}{s}\mathrm{Opt} & \le & \mathrm{Risks}^2_{\mathrm{opt}}[\mathcal{X}]\,(1 + \mathrm{Tr}(Q_\rho S)) + 6M_*^2\delta_\rho^{1/2}\\ & = & \mathrm{Risks}^2_{\mathrm{opt}}[\mathcal{X}]\,\big(1 + \frac{\rho}{s}\mathrm{Tr}(WS)\big) + 6M_*^2\delta_\rho^{1/2}.\end{array} \tag{4.3.8}$$

Similarly to Section 4.2.3, taking into account that, same as in the case of usual risk, we have $M_*^2 \ge \mathrm{Risks}^2_{\mathrm{opt}}[\mathcal{X}]$, and setting

$$\bar{\rho}^{-1} = 8\ln\left(\frac{6M_*^2\sqrt{K}}{\mathrm{Risks}^2_{\mathrm{opt}}[\mathcal{X}]}\right)$$

we ensure that

$$6M_*^2\delta_{\bar{\rho}}^{1/2} \le 6M_*^2\sqrt{K}\exp\{-\frac{1}{8\bar{\rho}}\} \le \mathrm{Risks}^2_{\mathrm{opt}}[\mathcal{X}],$$

so that (4.3.8) implies that

$$\frac{\bar{\rho}}{s}\mathrm{Opt} \le 2\mathrm{Risks}^2_{\mathrm{opt}}[\mathcal{X}]\left(1 + \frac{\bar{\rho}}{s}\mathrm{Tr}(WS)\right),$$

that is,

$$\begin{array}{rcl}\bar{\rho}\mathrm{Opt} & \le & 2\mathrm{Risks}^2_{\mathrm{opt}}[\mathcal{X}]\,(s + \bar{\rho}\mathrm{Tr}(WS))\\ & \le & 2\mathrm{Risks}^2_{\mathrm{opt}}[\mathcal{X}]\end{array}$$

(note that $s + \bar{\rho}\mathrm{Tr}(WS) \le s + \mathrm{Tr}(WS) \le 1$ by constraints in (4.3.3)). Recalling that $\sqrt{\mathrm{Opt}}$ upper-bounds $\mathrm{RiskS}[\widehat{x}_{H_*}|\mathcal{X}]$, we arrive at the following

Proposition 4.3.1 *The efficiently computable linear estimate $\widehat{x}_{H_*}(\omega) = H_*^T\omega$ yielded by an optimal solution to the optimization problem in* (4.3.2) *is nearly optimal in terms of S-risk:*

$$\mathrm{RiskS}[\widehat{x}_{H_*}|\mathcal{X}] \leq 4\sqrt{\ln\left(\frac{6M_*^2\sqrt{K}}{\mathrm{Risks}_{\mathrm{opt}}^2[\mathcal{X}]}\right)}\mathrm{Risks}_{\mathrm{opt}}[\mathcal{X}],$$

where M_* is given by (4.2.17).

4.3.1.3 The case of $\mathcal{X} = \mathbf{R}^n$

The problem of minimizing the worst-case, over $x \in \mathcal{X}$, S-risk over linear/all possible estimates makes sense for unbounded $\mathcal{X}$'s as well as for bounded ones. We intend to consider the case where $\mathcal{X} = \mathbf{R}^n$ and to show that in this case an efficiently computable linear estimate is *exactly optimal.*

Similar to (4.3.2), the problem of building the best, in terms of its worst-case over $x \in \mathbf{R}^n$ S-risk, linear estimate reads

$$\mathrm{Opt} = \min_{\tau,H}\left\{\tau : \left[\begin{array}{cc}\tau S & B^T - A^T H\\ B - H^T A & I_\nu\end{array}\right] \succeq 0,\ \mathrm{Tr}(H^T\Gamma H) \leq \tau\right\}; \tag{4.3.9}$$

a feasible solution (τ, H) to this problem produces an estimate $\widehat{x}_H(\omega) = H^T\omega$ with $\mathrm{RiskS}[\widehat{x}_H|\mathbf{R}^n] \leq \sqrt{\tau}$. It turns out (for proof, see Section 4.7.2.2) that

Proposition 4.3.2 *Assuming problem* (4.3.9) *feasible, the problem is solvable, and its optimal solution* (Opt, H_*) *induces linear estimate $\widehat{x}_{H_*}$ which is minimax optimal:*

$$\mathrm{RiskS}[\widehat{x}_{H_*}|\mathbf{R}^n] = \sqrt{\mathrm{Opt}} = \inf_{\widehat{x}(\cdot)} \mathrm{RiskS}[\widehat{x}(\cdot)|\mathbf{R}^n]. \tag{4.3.10}$$

It may be interesting to compare the optimal S-risk $\mathrm{RiskS}[\widehat{x}_{H_*}|\mathbf{R}^n] = \sqrt{\mathrm{Opt}}$ to the maximal risk $\mathrm{Risk}[\widehat{x}_{H^*}|\mathcal{X}_S]$ of the optimal linear estimation of Bx over the ellipsoid $\mathcal{X}_S = \{x \in \mathbf{R}^n : x^T Sx \leq 1\}$, so that H^* is the optimal solution to (4.2.9) with $K = 1$, $S_1 = S$ and $\mathcal{T} = [0, 1]$; note that in this case the optimal value in (4.2.9) is exactly $\mathrm{Risk}[\widehat{x}_{H^*}|\mathcal{X}_S]$, and not just an upper bound on this risk. When comparing (4.2.9) with (4.3.9) one can easily see that both risks are equivalent up to a factor $\sqrt{2}$:

$$\mathrm{RiskS}[\widehat{x}_{H_*}|\mathbf{R}^n] \leq \mathrm{Risk}[\widehat{x}_{H^*}|\mathcal{X}_S] \leq \sqrt{2}\mathrm{RiskS}[\widehat{x}_{H_*}|\mathbf{R}^n].$$

Note also that by the definition of S-risk, we have $\mathrm{Risk}[\widehat{x}_{H_*}|\mathcal{X}_S] \leq \sqrt{2}\mathrm{RiskS}[\widehat{x}_{H_*}|\mathcal{X}_S] \leq \sqrt{2}\mathrm{RiskS}[\widehat{x}_{H_*}|\mathbf{R}^n]$, which combines with the above inequalities to imply that

$$\mathrm{Risk}[\widehat{x}_{H_*}|\mathcal{X}_S] \leq \sqrt{2}\mathrm{Risk}[\widehat{x}_{H^*}|\mathcal{X}_S].$$

4.3.1.4 Numerical illustration

In the above considerations, we treated matrix S as part of the data. In fact, we can make S a variable restricted to reside in a given computationally tractable convex subset $\mathcal{S}$ of the positive semidefinite cone, and look for minimal, over linear estimates *and matrices* $S \in \mathcal{S}$, S-risk. This can be done as follows. We consider a parametric family of problems with τ in (4.3.2) being a parameter rather than a variable, and S being a variable restricted to reside in $\mathcal{S}$; then we apply bisection in τ to find the smallest value of τ for which the problem is feasible. With S and linear estimate yielded by this procedure, the S-risk of the estimate clearly possesses near-optimality properties completely similar to those we have just established for the case of fixed S.

As an illustration of these ideas, consider the following experiment. Let $[r; v]$ be state of pendulum with friction – the 2-dimensional continuous time dynamical system obeying the equations

$$\begin{array}{rcl}\dot{r} & = & v,\\ \dot{v} & = & -\nu^2 r - \kappa v + w,\end{array}$$

where w is the external input. Assuming this input constant on consecutive time intervals of duration Δ, the sequence $z_\tau = [r(\tau\Delta); v(\tau\Delta)]$, $\tau = 0, 1, ...$, obeys finite-difference equation

$$z_\tau = Pz_{\tau-1} + Qw_\tau,\ \ \tau = 1, 2, ...$$

with $P = \exp\left\{\Delta\overbrace{\left[\begin{smallmatrix} 0 & 1 \\ -\nu^2 & -\kappa \end{smallmatrix}\right]}^{\Theta}\right\}$, $Q = \int_0^\Delta \exp\{s\Theta\}[0;1]ds$; here w_τ is the value of $w(\cdot)$ on the (continuous-time) interval $((\tau-1)\Delta, \tau\Delta)$. Assume that we are observing corrupted by noise positions $r_\tau = r(\tau\Delta)$ of the pendulum on the discrete-time horizon $1 \leq \tau \leq T$ and want to recover the inputs w_s, $T - K + 1 \leq s \leq T$. Denoting by $x = [z_0; w_1; w_2; ...; w_T]$ the "signal" underlying our observations, we can easily build a $T \times (T+2)$ matrix A and $1 \times (T+2)$ matrices B_t such that the trajectory $r := [r_1; ...; r_T]$ of pendulum's positions is given by $r = Ax$, and $w_t = B_t x$. What we want to recover from noisy observations

$$\omega = Ax + \xi,\ \xi \sim \mathcal{N}(0, \sigma^2 I_T)$$

of pendulum's (discrete time) trajectory, are the inputs w_t, $1 \leq t \leq T$, and their collections $w^K = [w_{T-K+1}; w_{T-K+2}; ...; w_T] = B^{(K)}x$.[12]

We intend to process our estimation problems by building the best, in terms of its S-risk taken over the entire space $\mathbf{R}^{T+2}$ of signals, estimate; in our design, S is not fixed in advance, but is instead restricted to be positive semidefinite with trace ≤ 1. Thus, the problems we want to solve are of the form (cf. (4.3.9))

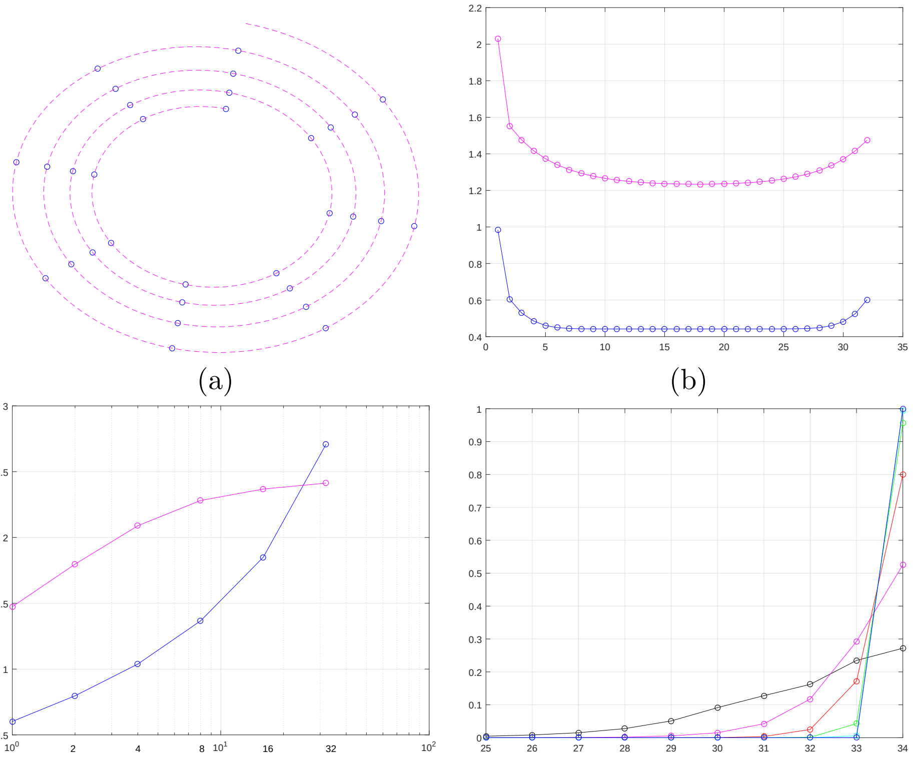
$$\mathrm{Opt}[B] = \min_{\tau,H,S}\left\{\tau : \left[\begin{array}{cc} \tau S & B^T - A^T H \\ B - H^T A & I_T \end{array}\right] \succeq 0,\ \sigma^2\mathrm{Tr}(H^T H) \leq \tau,\ S \succeq 0,\ \mathrm{Tr}(S) \leq 1\right\}, \tag{4.3.11}$$

where B depends on what we want to recover ($B = B_t$ when recovering w_t, and $B = B^{(K)}$ when recovering w^K). By Proposition 4.3.2, the linear estimate $H_{B,*}^T\omega$ yielded by an optimal solution $(\mathrm{Opt}[B], H_{B,*}, S_{B,*})$ to the above (clearly solvable) problem is minimax optimal in terms of its S-risk $\mathrm{RiskS}[\cdot|\mathbf{R}^{T+2}]$ *taken with respect to* $S = S_{B,*}$, and the corresponding minimax optimal risk is exactly $\sqrt{\mathrm{Opt}[B]}$.

> The rationale behind restricting S to have its trace ≤ 1 is as follows. Imagine that we have reasons to believe that the entries in x "are of order of 1;" the simplest way to model this belief is to assume that x is uniformly distributed over the sphere $\mathcal{S}$ of radius $\sqrt{\dim x} = \sqrt{T+2}$. Under this assumption, the claim that an estimate $\widehat{x}(\cdot)$ has S-risk, taken over the entire space w.r.t. a matrix $S \succeq 0$ with $\mathrm{Tr}(S) \leq 1$, at most $\sqrt{\tau}$ means that
>
> $$\mathbf{E}_{\xi\sim\mathcal{N}(0,\sigma^2 I_T)}\{\|\widehat{x}(Ax+\xi) - B_K x\|_2^2\} \leq \tau(1 + x^T S x)\ \forall x.$$
>
> This relation, after taking expectation over the uniformly distributed over $\mathcal{S}$ signal x, implies that the expectation, over both ξ and x, of the squared recovery risk is at most 2τ. Thus, optimising the S-risk over the linear estimates *and* $S \succeq 0$, $\mathrm{Tr}(S) \leq 1$, can be interpreted as a kind of safe minimization of the Bayesian risk taken w.r.t. a specific Bayesian prior (uniform distribution on $\mathcal{S}$). In this context, "safety" means that along with guarantees on the Bayesian risk, we get some meaningful upper bound on the expected $\|\cdot\|_2^2$-error of recovery applicable to *every individual* signal.
>
> In view of the above considerations, in the sequel we take the liberty to refer to the quantity $\sqrt{2\mathrm{Opt}[B]}$, where $\mathrm{Opt}[B]$ is optimal value of (4.3.11), as to the *Bayesian risk* of recovering Bx.

[12] Note that estimating w^K is not the same as "standalone" estimation of each individual entry in w^K.

In the experiment we are about to report, we use $\Delta = 1$, $\kappa = 0.05$ and select ν to make the eigenfrequency of the pendulum equal to 1/8; free motion of the pendulum in the (r, v)-coordinates is shown on Figure 4.2. We used $\sigma = 0.075$, $T = 32$, and solved problem (4.3.11) for several "B-scenarios." The results are presented on Figure 4.2 (b) – (d). Plots (b) and (c) show the bound $\sqrt{2\mathrm{Opt}[B]}$, see above, on the Bayesian risk along with the risk of the best, in terms of its worst-case over signals from the ball $\mathcal{X}$ bounded by $\mathcal{S}$, risk of linear recovery of Bx as given by the optimal values of the associated problems (4.2.9) (blue). Plot (b) shows what happens when recovering individual inputs ($B = B_t$, $t = 1, 2, ..., T$) and displays the risks as functions of t; plot (c) shows the risks of recovering blocks $u^K = B^{(K)}x$ of inputs as functions of $K = 1, 2, 4, ..., 32$. Finally, plot (d) shows the eigenvalues of the S-components of optimal solutions to problems (4.3.11) with $B = B^{(K)}$.[13] Comments are as follows.



(a) (b)

(c) (d)

Figure 4.2: Numerical illustration, Section 4.3.1.4. (a): free motion ($w \equiv 0$) of pendulum in (r, v)-plane in continuous (dashed line) and discrete (circles) time. (b): Bayesian (blue) and worst–case (magenta) risks of recovering w_t vs. $t = 1, 2, ..., 32$. (c): Bayesian (blue) and worst-case (magenta) risks of recovering $w^K := [w_{T-K+1}; w_{T-K+2}; ...; w_T]$ vs. K. (d): eigenvalues $\lambda_i(S_K)$ of S_K ($K = 32$ – black, $K = 16$ – magenta, $K = 8$ – red, $K = 4$ – green, $K = 2$ – cyan, $K = 1$ – blue); we plot 10 largest eigenvalues of the S-matrices; the preceding 24 eigenvalues for all these matrices vanish.
Quiz: How the trajectory on (a) goes – clockwise or counter-clockwise?

- Plot (b) shows that both Bayesian and worst-case risks of recovering $B_t x$ first decrease, and then increase as t grows. Explanation is as follows: our abilities of recovering w_t from observations are affected by two factors. First is that our estimation problem is degenerate – we do not observe the initial state z_0 of the pendulum, so that matrix A has a kernel of dimension 2; the "completely invisible" even in the noiseless case $\sigma = 0$ projection of x on

[13] With $B = B_t$, S-components of optimal solutions to (4.3.11) turn out to be of rank 1 for all t.

Ker A has nontrivial w_t-components, thus making exact recovery impossible even when $\sigma = 0$; however, friction makes our dynamical system stable, so that the influence of unknown initial state z_0 on our abilities to recover w_t deteriorates when t increases. At the same time, "early" inputs contribute to our observations on longer time intervals than the "late" ones, allowing for better suppressing of the observation noise.

- Plot (c) shows that the risks in recovering blocks w^K grow with K, which is how it should be – the larger is the block we want to recover, the larger is the attainable risk of recovery.

- As it could be expected, the optimal matrices $S_{B,*}$ associated with our B's are of low rank (plot (d)); the highest rank (9) corresponds to recovering the entire input; when recovering an individual inputs, the ranks are just 1.

4.3.2 Adding robustness

In this Section we address the situation where the data A, B of problems (4.2.9) and (4.3.2) is not known exactly, and we are looking for estimates which are robust w.r.t. the corresponding data uncertainties. We lose nothing when restricting ourselves with problem (4.3.2), since (4.2.9) is the particular case $S = 0$ of (4.3.2), with ellitope $\mathcal{X}$ given by (4.2.2). We intend to focus on the simplest case of *unstructured norm-bounded uncertainty*

$$[A; B] := \left[\begin{array}{c} A \\ B \end{array}\right] \in \mathcal{U}_r = \left\{[A; B] = [A_*; B_*] + E^T \Delta F : \Delta \in \mathbf{R}^{p\times q}, \|\Delta\| \le r\right\}; \tag{4.3.12}$$

here $A_* \in \mathbf{R}^{m\times n}$, $B_* \in \mathbf{R}^{\nu\times n}$ are given *nominal data*, and $E \in \mathbf{R}^{p\times(m+\nu)}$, $F \in \mathbf{R}^{q\times n}$ are given matrices.[14] Our goal is to solve the *robust counterpart*

$$\begin{array}{r}\mathrm{RobOpt} = \min\limits_{\tau,H,\lambda} \left\{ \tau : \left[\begin{array}{cc} \sum_k \lambda_k S_k + \tau S & B^T - A^T H \\ B - H^T A & I_\nu \end{array}\right] \succeq 0, \ \forall [A; B] \in \mathcal{U} \right. \\ \left. \mathrm{Tr}(H^T \Gamma H) + \phi_{\mathcal{T}}(\lambda) \le \tau, \lambda \ge 0 \right\}\end{array} \tag{4.3.13}$$

of problem (4.3.2). Plugging into (4.3.13) the parametrization of $[A; B]$ via Δ, the uncertainty-affected semidefinite constraint becomes

$$\begin{array}{l} M(\lambda, \tau, H) + \mathcal{E}^T[H]\Delta\mathcal{F} + \mathcal{F}^T\Delta^T\mathcal{E}[H] \succeq 0 \ \ \forall(\Delta : \|\Delta\| \le r), \\ M(\lambda, \tau, H) = \left[\begin{array}{cc} \sum_k \lambda_k S_k + \tau S & B_*^T - A_*^T H \\ B_* - H^T A_* & I_\nu \end{array}\right], \\ \mathcal{E}[H] = [0_{p\times n}, E_B - E_A H], \ \mathcal{F} = [F, 0_{q\times\nu}], \end{array} \tag{4.3.14}$$

where

$$E = [E_A, E_B]$$

is the partitioning of the $p \times (m + \nu)$-matrix E into the blocks comprised by the first m and the last ν columns. A well-known result of [26] (see also [15, Section 8.2.1]) states that when $\mathcal{F} \neq 0$

[14] Recall that in the case of $P \neq I$ we have to replace matrices A, B and S with AP, BP and $P^T SP$, respectively, and modify the definition of $\mathcal{U}_r$ accordingly: namely, when $[A; B]$ runs through the set $\mathcal{U}_r$, $[AP; BP]$ runs through

$$\overline{\mathcal{U}}_r = \left\{[A; B] = [A_*P; B_*P] + E^T \Delta F P : \Delta \in \mathbf{R}^{p\times q}, \|\Delta\| \le r\right\};$$

where A_*, B_* E and F are as in (4.3.12).

(this is the only nontrivial case), the semi-infinite Linear Matrix Inequality in (4.3.14) holds true if and only if there exists μ such that

$$\left[\begin{array}{cc} M(\lambda,\tau,H)-r^2\mu\mathcal{F}^T\mathcal{F} & [\mathcal{E}[H]]^T \\ \mathcal{E}[H] & \mu I_p \end{array}\right]\succeq 0.$$

It follows that the semi-infinite convex problem (4.3.13) is equivalent to the explicit convex program

$$\begin{array}{rcl} \mathrm{RobOpt} & = & \min\limits_{\tau,H,\lambda,\mu}\left\{\tau:\left[\begin{array}{c|c|c} \sum_k\lambda_kS_k+\tau S-\mu r^2F^TF & B_*^T-A_*^TH & \\ \hline B_*-H^TA_* & I_\nu & E_B^T-H^TE_A^T \\ \hline & E_B-E_AH & \mu I_p \end{array}\right]\succeq 0,\right. \\ & & \left.\qquad\qquad\qquad\mathrm{Tr}(H^T\Gamma H)+\phi_{\mathcal{T}}(\lambda)\leq\tau,\lambda\geq 0\right\}. \end{array} \tag{4.3.15}$$

The H-component of optimal solution to (4.3.15) yields robust w.r.t. uncertainty (4.3.12) estimate $H^T\omega$ of Bx via observation $Ax+\xi$, and the expected $\|\cdot\|_2^2$-error of this estimate does not exceed RobOpt, whatever be $x\in\mathcal{X}$ and $[A;B]\in\mathcal{U}$.

4.3.3 Beyond Gaussian observation scheme

Normal observation schemes. Let $\mathcal{M}\subset\mathbf{R}^n$ be a convex set, and let $\mathcal{P}$ be a family of probability distributions on $\mathbf{R}^d$ such that every $p\in\mathcal{P}$ is assigned *parameter value* $\mu(P)\in\mathcal{M}$. For a vector $z\in\mathbf{R}^k$, let $Q[z]=\left[\begin{array}{cc} zz^T & z\\ z^T & 1\end{array}\right]=[z;1][z;1]^T$. Let us call pair $(\mathcal{P},\mathcal{M})$ *normal*, if

$$\begin{array}{rcl} \forall p\in\mathcal{P}: & & \\ \mathbf{E}_{\omega\sim p}\{[\omega;1]\} & = & \ell(\mu(p)),\\ \mathbf{E}_{\omega\sim p}\{Q[\omega]\} & = & \mathcal{L}(Q[\mu(p)]), \end{array}$$

where $\ell(\cdot)$ and $\mathcal{L}(\cdot)$ are affine.

Example 1. Let $\mathcal{M}=\mathbf{R}^m\times\mathbf{R}^{m\times m}$, $\mathcal{P}$ be the family of all Borel probability distributions on $\mathbf{R}^d$ with finite variance, and $\mu(p)=[u;\mathrm{Vec}(U)]$ with u, U such that random variable $\omega\sim p$ can be represented as

$$\omega=u+U\xi$$

with a normalized (zero mean, unit covariance) random vector ξ. For $p\in\mathcal{P}$ with $\mu(p)=[u;\mathrm{Vec}(U)]$ we have

$$\begin{array}{rcl} \mathbf{E}_{\omega\sim p}\{[\omega;1]\} & = & [u;1],\\ \mathbf{E}_{\omega\sim p}\{Q[\omega]\} & = & \left[\begin{array}{cc} uu^T+UU^T & u\\ u^T & 1\end{array}\right], \end{array}$$

and the right hand sides are affine in $\mu(p)$ and $Q[\mu(p)]$, respectively.

Example 2: Poisson observation. In this case $\mathcal{M}=\mathbf{R}^d_+$ and $\mathcal{P}$ is the family of distributions of d-dimensional vectors ω with independent Poisson entries, and $\mu(p)=[\mu_1;...;\mu_d]$ where μ_i is the Poisson parameter of i-th entry. Here for every $p\in\mathcal{P}$ we have

$$\begin{array}{rcl} \mathbf{E}_{\omega\sim p}\{[\omega;1]\} & = & [\mu(p);1],\\ \mathbf{E}_{\omega\sim p}\{Q[\omega]\} & = & \left[\begin{array}{cc} \mu(p)\mu^T(p)+\mathrm{Diag}\{\mu(p)\} & \mu(p)\\ \mu^T(p) & 1\end{array}\right], \end{array} \tag{4.3.16}$$

and the right hand sides are affine in $\mu(p)$ and $Q[\mu(p)]$, respectively.

Example 3: Discrete repeated observation. In this case $\mathcal{M}$ is the probabilistic simplex in $\mathbf{R}^d$, and $\mathcal{P}$ is comprised of distributions p of empirical histograms built from a given number m of observations of a discrete random variable taking values $1,...,d$. In other words, $p\in\mathcal{P}$ is the distribution of random vector $\frac{1}{m}\sum_{i=1}^m e_{\iota_i}$, where m is fixed, e_s is s-th basic orth in $\mathbf{R}^d$, $\iota_1,...,\iota_m$ are i.i.d. random numbers drawn form a distribution π on $\{1,...,d\}$; entries in $\mu(p)\in\mathbf{R}^d$ are probabilities for $\iota\sim\pi$ to take values $s\in\{1,...,m\}$. Here for $p\in\mathcal{P}$ it holds

$$\begin{array}{rcl}\mathbf{E}_{\omega\sim p}\{[\omega;1]\} &=& [\mu;1],\\ \mathbf{E}_{\omega\sim p}\{Q[\omega]\} &=& \left[\begin{array}{cc}\left[1-\frac{1}{m}\right]\mu\mu^T+\frac{1}{m}\mathrm{Diag}\{\mu\} & \mu\\ \mu^T & 1\end{array}\right],\end{array}$$

and the right hand sides are affine in $\mu(p)$ and $Q[\mu(p)]$, respectively.

Estimation problem. Given a nonempty compact set X in $\mathbf{R}^n$, a normal pair $(\mathcal{P},\mathcal{M})$ and an affine mapping

$$x\mapsto\mathcal{A}(x):=A[x;1]$$

which maps X into $\mathcal{M}$, we observe, for some unknown signal x known to belong to X, a realization ω of distribution $p\in\mathcal{P}_x$, where $\mathcal{P}_x=\{p\in\mathcal{P}:\mu(p)=\mathcal{A}(x)\}$, and want to recover from this observation the image Bx of x under a given linear mapping. We intend to use for this purpose *affine estimates* $\widehat{z}_H(\omega)=H[\omega;1]$. Our first goal is to upper-bound the worst case, over $x\in X$, expected $\|\cdot\|_2^2$ error of recovery

$$\mathrm{Risk}[H|X]=\sqrt{\max_{x\in X}\sup_{p\in\mathcal{P}:\mu(p)=A[x;1]}\mathbf{E}_{\omega\sim p}\left\{\|H[\omega;1]-Bx\|_2^2\right\}}.$$

Given H and $G\succeq H^TH$, for $x\in X$ and $p\in\mathcal{P}$ such that $\mu(p)=A[x;1]$ we have

$$\begin{array}{l}\mathbf{E}_{\omega\sim p}\{\|H[\omega;1]-Bx\|_2^2\}=\mathbf{E}_{\omega\sim p}\{\mathrm{Tr}(H^THQ[\omega])-2\mathrm{Tr}([\omega;1]^TH^TBx)+\mathrm{Tr}(x^TB^TBx)\}\\ \leq\mathrm{Tr}(G\mathbf{E}_{\omega\sim p}\{Q[\omega]\})-2\mathrm{Tr}(H^TB\mathbf{E}_{\omega\sim p}\{x[\omega;1]^T\})+\mathrm{Tr}(x^TB^TBx)\}\\ =F_x(H,G):=\mathrm{Tr}(G\mathcal{L}(Q[\mu(p)]))-2\mathrm{Tr}(H^TBx\ell^T(\mu(p)))+\mathrm{Tr}(x^TB^TBx)\\ =\mathrm{Tr}(G\mathcal{L}(Q[\mathcal{A}(x)]))-2\mathrm{Tr}(H^TBx\ell^T(\mathcal{A}(x)))+\mathrm{Tr}(x^TB^TBx),\end{array}$$

We clearly have

$$F_x(H,G)=\mathrm{Tr}(\mathcal{C}(H,G)Q[x]),$$

where $\mathcal{C}(H,G)$ is a symmetric matrix affinely depending on H,G. Assume that X is contained in the set

$$\widehat{X}=\{x:\exists t\in\mathcal{T}:\mathrm{Tr}(S_kQ[x])\leq t_k,\,1\leq k\leq K\}$$

where S_k are symmetric matrices and $\mathcal{T}$ is a convex compact set in $\mathbf{R}^K$ (note that now we do not impose on S_k the restriction $S_k\succeq 0$ and do not require for $\mathcal{T}$ to be contained in $\mathbf{R}^K_+$). Assuming that some combination of matrices S_k with nonnegative coefficients is positive definite, we can use semidefinite relaxation to build a safe tractable approximation of the problem of building minimum risk linear estimate. This approximation is the problem

$$\mathrm{Opt}=\min_{H,G,\lambda}\left\{\phi_{\mathcal{T}}(\lambda):\lambda\geq 0,\mathcal{C}(H,G)\preceq\sum_k\lambda_kS_k,H^TH\preceq G\right\}.$$

4.3.4 Byproduct on semidefinite relaxation

A byproduct of our main observation (Section 4.2.3) we are about to present has nothing to do with statistics; it relates to the quality of the standard semidefinite relaxation. Specifically, given a quadratic from x^TCx and an ellitope $\mathcal{X}$ represented by (4.2.2), consider the problem

$$\mathrm{Opt}_* = \max_{x\in\mathcal{X}} x^TCx = \max_{y\in\bar{X}} y^TP^TCPy. \tag{4.3.17}$$

This problem can be NP-hard (this is already so when $\mathcal{X}$ is the unit box and C is positive semidefinite); however, Opt admits an efficiently computable upper bound given by *semidefinite relaxation* as follows: whenever $\lambda \geq 0$ is such that

$$P^TCP \preceq \sum_{k=1}^K \lambda_k S_k,$$

for $y \in \bar{X}$ we clearly have

$$[Py]^TCPy \leq \sum_k \lambda_k y^TS_ky \leq \phi_{\mathcal{T}}(\lambda)$$

due to the fact that the vector with the entries y^TS_ky, $1 \leq k \leq K$, belongs to $\mathcal{T}$. As a result, the efficiently computable quantity

$$\mathrm{Opt} = \min_\lambda \left\{ \phi_{\mathcal{T}}(\lambda) : \lambda \geq 0, P^TCP \preceq \sum_k \lambda_k S_k \right\} \tag{4.3.18}$$

is an upper bound on Opt_*. We have the following

Proposition 4.3.3 *Let C be a symmetric $n \times n$ matrix and $\mathcal{X}$ be given by ellitopic representation* (4.2.2)*, and let* Opt_* *and* Opt *be given by* (4.3.17) and(4.3.18)*. Then*

$$\frac{\mathrm{Opt}}{4\ln(5K)} \leq \mathrm{Opt}_* \leq \mathrm{Opt}. \tag{4.3.19}$$

For proof, see Section 4.7.2.3.

4.3.5 Linear estimates beyond $\|\cdot\|_2$-risk

Similarly to Section 4.2.1, consider the problem of recovering the image $Bx \in \mathbf{R}^\nu$ of a signal $x \in \mathbf{R}^n$ known to belong to a given ellitope

$$\mathcal{X} = \{x \in \mathbf{R}^n : \exists t \in \mathcal{T} : x^TS_kx \leq t_k,\ 1 \leq k \leq K\} \tag{4.3.20}$$

from noisy observation

$$\omega = Ax + \xi, \tag{4.3.21}$$

where A is a known $m \times n$ matrix, and ξ is a zero mean “standardized” random observation noise with known probability distribution P possessing finite first moment. We are interested in building a linear estimate

$$\widehat{x}_H(\omega) = H^T\omega$$

of Bx. Given a norm $\|\cdot\|$ on $\mathbf{R}^\nu$, we intend to quantify the performance of a candidate estimate $\omega \mapsto \widehat{x}(\omega)$ by the quantity

$$\mathrm{Risk}_{\|\cdot\|}[\widehat{x}(\cdot)|\mathcal{X}] = \sup_{x\in\mathcal{X}} \mathbf{E}_{\xi\sim P}\left\{\|\widehat{x}(Ax+\xi) - Bx\|\right\} \tag{4.3.22}$$

that is, by the worst-case, over $x \in \mathcal{X}$, expected $\|\cdot\|$-error of the recovery.

Our goal is to build a linear estimate with as small risk as possible.

Preliminaries: conjugate norms. Recall that a norm $\|\cdot\|$ on a Euclidean space $\mathcal{E}$, e.g., on $\mathbf{R}^k$, gives rise to its *conjugate* norm

$$\|y\|_* = \max_x\{\langle y,x\rangle : \|x\| \leq 1\},$$

where $\langle\cdot,\cdot\rangle$ is the inner product in $\mathcal{E}$. Equivalently, $\|\cdot\|_*$ is the smallest norm such that

$$\langle x,y\rangle \leq \|x\|\|y\|_*\ \forall x,y. \tag{4.3.23}$$

It is well known that taken twice, norm conjugation recovers the initial norm: $(\|\cdot\|_*)_*$ is exactly $\|\cdot\|$; in other words,

$$\|x\| = \max_y\{\langle x,y\rangle : \|y\|_* \leq 1\}.$$

The standard examples are the conjugates to the standard ℓ_p-norms on $\mathcal{E} = \mathbf{R}^k$, $p \in [1,\infty]$; it turns out that

$$(\|\cdot\|_p)_* = \|\cdot\|_{p*},$$

where $p_* \in [1,\infty]$ is linked to $p \in [1,\infty]$ by the symmetric relation

$$\frac{1}{p} + \frac{1}{p_*} = 1,$$

so that $1_* = \infty$, $\infty_* = 1$, $2_* = 2$; the corresponding version of inequality (4.3.23) is called *Hölder inequality* – an extension of the Cauchy-Schwartz inequality dealing with the case $\|\cdot\| = \|\cdot\|_* = \|\cdot\|_2$.

Processing the problem. Let $H \in \mathbf{R}^{m\times\nu}$. The risk of the linear estimate $\widehat{x}_H(\omega) = H^T\omega$ can be upper-bounded as follows:

$$\begin{array}{rcl}
\mathrm{Risk}_{\|\cdot\|}[\widehat{x}_H|\mathcal{X}] &=& \sup\limits_{x\in\mathcal{X}} \mathbf{E}_{\xi\sim P}\left\{\|H^T(Ax+\xi-Bx\|\right\} \leq \sup\limits_{x\in\mathcal{X}} \mathbf{E}_{\xi\sim P}\left\{\|(B-H^TA)x\| + \|H^T\xi\|\right\}\\
&=& \max\limits_{x\in\mathcal{X}} \|(B-H^TA)x\| + \mathbf{E}_{\xi\sim P}\left\{\|H^T\xi\|\right\}\\
&=& \|B-H^TA\|_{\mathcal{X},\|\cdot\|} + \Psi_P(H),\\
\|V\|_{\mathcal{X},\|\cdot\|} &=& \max_x\left\{\|Vx\| : x\in\mathcal{X}\right\} : \mathbf{R}^{\nu\times n} \to \mathbf{R},\\
\Psi_P(H) &=& \mathbf{E}_{\xi\sim P}\left\{\|H^T\xi\|\right\}.
\end{array} \tag{4.3.24}$$

Note that the upper risk bound (4.3.24) is reasonably tight – at most by factor 2 larger than the actual risk (why?).

By their origin, $\|\cdot\|_{\mathcal{X},\|\cdot\|}$ and $\Psi_P(\cdot)$ are convex real-valued, even and positively homogeneous, of degree 1, functions; a natural course of actions would be to select H as the optimal solution to the convex optimization problem

$$\mathrm{Opt}_\# = \min_H\left\{\|B-H^TA\|_{\mathcal{X},\|\cdot\|} + \Psi_P(H)\right\}. \tag{4.3.25}$$

The difficulty, however, is that the norm $\|\cdot\|_{\mathcal{X},\|\cdot\|}$ and the function Ψ_P could be difficult to compute. For example, as far as computation of the norm $\|\cdot\|_{\mathcal{X},\|\cdot\|}$ is concerned, on a closest inspection, this problem is definitely easy in just three standard cases:

- when $\mathcal{X} = \{x : \|x\|_1 \leq 1\}$ is ℓ_1 ball; this case is of no much interest in our context, since ℓ_1-ball, while being an ellitope, has huge, for large n, "ellitopic size" K: $K = 2^{n-1}$, which makes ellitopic representation of $\mathcal{X}$ useless computationally;
- when $\mathcal{X}$ is an ellipsoid centered at the origin, and $\|\cdot\| = \|\cdot\|_2$; in this case, computing $\|\cdot\|_{\mathcal{X},\|\cdot\|}$ is easy – it reduces to computing spectral norm of a matrix;

- when $\|\cdot\|=\|\cdot\|_\infty$; in this case computing $\|V\|_{\mathcal{X},\|\cdot\|}$ reduces to maximizing k linear forms over $\mathcal{X}$ (why?).

To overcome, to some extent, the above difficulty, from now on we make the following

Assumption A: *The unit ball $\mathcal{B}_*$ of the norm $\|\cdot\|_*$ conjugate to the norm $\|\cdot\|$ participating in the formulation of our estimation problem is an ellitope:*

$$\mathcal{B}_*=\{z=My, y\in\mathcal{Y}\},\ \mathcal{Y}=\{y\in\mathbf{R}^q:\exists r\in\mathcal{R}: y^TQ_\ell y\leq r_\ell,\ 1\leq\ell\leq L\},\qquad(4.3.26)$$

where $Q_\ell\succeq 0$, $\sum_\ell Q_\ell\succ 0$, and $\mathcal{R}$ is a convex compact subset of $\mathbf{R}^L_+$ such that $r\in\mathcal{R}$ implies $r'\in\mathcal{R}$ for all r' satisfying $0\leq r'\leq r$ and, on the top of this, $\mathcal{R}$ contains a strictly positive vector.

Note that Assumption **A** is satisfied, e.g., when $\|\cdot\|=\|\cdot\|_p$ with $p\in[1,2]$; in this case,

$$\mathcal{B}_*=\{y\in\mathbf{R}^\nu:\|y\|_{p_*}\leq 1\}=\{y\in\mathbf{R}^\nu:\exists r\in\mathcal{R}: y^T\underbrace{e_\ell e_\ell^T}_{Q_\ell}y\leq r_\ell, 1\leq\ell\leq L=\nu\},$$

where e_ℓ are the standard basic orths in $\mathbf{R}^\nu$ and

$$\mathcal{R}=\{r\in\mathbf{R}^\nu_+:\sum_{\ell=1}^K r_\ell^{p_*/2}\leq 1\}.$$

The point is that under Assumption **A** we can utilize Proposition 4.3.3 to get a "reasonably tight" efficiently computable convex upper bound on $\|\cdot\|_{\mathcal{X},\|\cdot\|}$; this bound, same as the function $\|\cdot\|_{\mathcal{X},\|\cdot\|}$ itself, is a norm on the space $\mathbf{R}^{\nu\times n}$ of $\nu\times n$ matrices. Specifically, with Assumption **A** in force, consider the ellitope

$$\begin{array}{rcl}\mathcal{Z} &:=& \mathcal{X}\times\mathcal{Y}=\{[x;y]\in\mathbf{R}^n\times\mathbf{R}^q:\exists s=[t;r]\in\mathcal{T}\times\mathcal{R}:\\ && \qquad x^TS_kx\leq t_k,\ 1\leq k\leq K, y^TQ_\ell y\leq r_\ell,\ 1\leq\ell\leq L\}\\ &=& \{w=[x;y]\in\mathbf{R}^n\times\mathbf{R}^q:\exists s=[t;r]\in\mathcal{S}=\mathcal{T}\times\mathcal{R}: w^TU_iw\leq s_i,\ 1\leq i\leq I=K+L\},\end{array}\qquad(4.3.27)$$

with positive semidefinite matrices U_ℓ of size $(n+q)\times(n+q)$ readily given by S_k and Q_ℓ. Now, given a $\nu\times n$ matrix V, let

$$W[V]=\frac{1}{2}\left[\begin{array}{c|c} & V^TM\\ \hline M^TV & \end{array}\right],$$

so that

$$[x;y]^TW[V][x;y]\equiv y^TM^TVx,$$

whence

$$\|V\|_{\mathcal{X},\|\cdot\|}=\max_{x\in\mathcal{X}}\|Vx\|=\max_{x\in\mathcal{X},w\in\mathcal{B}_*}w^TVx=\max_{x\in\mathcal{X},y\in\mathcal{Y}}y^TM^TVx=\max_{z\in\mathcal{Z}}z^TW[V]z.$$

Applying Proposition 4.3.3, we arrive at the following observation:

Corollary 4.3.1 *In the just defined situation, the efficiently computable convex function*

$$\begin{array}{c}\|V\|^+_{\mathcal{X},\|\cdot\|}=\min_{\lambda,\mu}\left\{\phi_{\mathcal{T}}(\lambda)+\phi_{\mathcal{R}}(\mu):\lambda\geq 0,\mu\geq 0,\left[\begin{array}{c|c}\sum_k\lambda_kS_k & \frac{1}{2}V^TM\\ \hline \frac{1}{2}M^TV & \sum_\ell\mu_\ell Q_\ell\end{array}\right]\succeq 0\right\},\\ \left[\phi_{\mathcal{T}}(\lambda)=\max_{t\in\mathcal{T}}\lambda^Tt,\ \phi_{\mathcal{R}}(\mu)=\max_{r\in R}\mu^Tr\right]\end{array}\qquad(4.3.28)$$

is a norm on $\mathbf{R}^{\nu\times n}$, and this norm is a tight upper bound on $\|\cdot\|_{\mathcal{X},\|\cdot\|}$, namely,

$$\forall V\in\mathbf{R}^{\nu\times n}:\|V\|_{\mathcal{X},\|\cdot\|}\leq\|V\|^+_{\mathcal{X},\|\cdot\|}\leq 4\ln(5(K+L))\|V\|_{\mathcal{X},\|\cdot\|}.\qquad(4.3.29)$$

Remark 4.3.1 *Consider the case when $\mathcal{X}$ is the unit $\|\cdot\|_p$-ball and $\|\cdot\| = \|\cdot\|_r$, with $1 \leq r \leq 2 \leq p \leq \infty$, that is,*

$$\begin{array}{rcl}\mathcal{X} & = & \{x \in \mathbf{R}^n : \exists t \in \mathcal{T}_p^n : x^T[e_ie_i^T]x \leq t_i,\, 1 \leq i \leq n\},\\ \mathcal{B}_* & = & \mathcal{Y} := \{y \in \mathbf{R}^\nu : \exists t \in \mathcal{T}_{r_*}^{\nu}\text{`} : y^T[f_if_i^T]y \leq t_i,\, 1 \leq i \leq \nu\}\\ \mathcal{T}_s^\nu & = & \{t \in \mathbf{R}^\nu_+ : \sum_i t_i^{s/2} \leq 1\},\end{array}$$

where e_i, f_i are the standard basic orths in $\mathbf{R}^n$ and in $\mathbf{R}^\nu$, respectively. In this case (4.3.28) reads

$$\|V\|^+_{\mathcal{X},\|\cdot\|} = \min_{\lambda,\mu}\left\{\|\lambda\|_{\frac{p}{p-2}} + \|\mu\|_{\frac{r}{2-r}} : \left[\begin{array}{c|c}\mathrm{Diag}\{\lambda\} & \frac{1}{2}V^T\\ \hline \frac{1}{2}V & \mathrm{Diag}\{\mu\}\end{array}\right] \succeq 0\right\}.$$

Applying technique completely different from the one we have used above, Yu. Nesterov proved [144, Theorems 13.2.4, 13.2.5], among other, that in the case in question the above bound is tight within absolute constant factor $\approx 2.2936...$:

$$\|V\|_{\mathcal{X},\|\cdot\|} \leq \|V\|^+_{\mathcal{X},\|\cdot\|} \leq \frac{1}{\frac{2\sqrt{3}}{\pi} - \frac{2}{3}}\|V\|_{\mathcal{X},\|\cdot\|}.$$

Besides this, $\|V\|_{\mathcal{X},\|\cdot\|} = \|V\|^+_{\mathcal{X},\|\cdot\|}$ when V is entrywise nonnegative (D. Steinberg, see Theorem 3.3. in http://www2.isye.gatech.edu/~nemirovs/Daureen.pdf*).*

Upper-bounding $\Psi_P(\cdot)$. For the sake of simplicity, let us restrict ourselves to the case when

$$\|u\| = \|Ru\|_p,\, p \in [1,2] \;\&\; P = \mathcal{N}(0,Q), \tag{4.3.30}$$

where R is a $\mu \times \nu$ matrix of rank ν [15]. In this case, denoting by $\mathrm{Row}_i[U]$ the transpose of i-th row in matrix U, we have

$$\begin{array}{rcl}\mathbf{E}_{\xi\sim\mathcal{N}(0,Q)}\left\{\|H^T\xi\|^p\right\} & = & \mathbf{E}_{\eta\sim\mathcal{N}(0,I_\ell)}\left\{\|RH^TQ^{1/2}\eta\|_p^p\right\} = \sum_{i=1}^\mu \mathbf{E}_{\eta\sim\mathcal{N}(0,I_\ell)}\left\{\left|\mathrm{Row}_i^T[RH^TQ^{1/2}]\eta\right|^p\right\}\\ & = & c_p^p\sum_{i=1}^\mu \|\mathrm{Row}_i[RH^TQ^{1/2}]\|_2^p,\\ c_p & := & \left(\frac{1}{\sqrt{2\pi}}\int|s|^p\mathrm{e}^{-s^2/2}ds\right)^{1/p} \in [\sqrt{2/\pi},1],\end{array}$$

whence

$$\begin{array}{rcl}\Psi_P(H) & = & \mathbf{E}_{\xi\sim\mathcal{N}(0,Q)}\left\{\|H^T\xi\|\right\} \leq \left[\mathbf{E}_{\xi\sim\mathcal{N}(0,Q)}\left\{\|H^T\xi\|^p\right\}\right]^{1/p}\\ & \leq & c_p\left\|\left[\|\mathrm{Row}_1[RH^TQ^{1/2}]\|_2; \|\mathrm{Row}_2[RH^TQ^{1/2}]\|_2; ...; \|\mathrm{Row}_\mu[RH^TQ^{1/2}]\|_2\right]\right\|_p.\end{array} \tag{4.3.31}$$

Note that (4.3.31) gives a tight, up to factor $c_2/c_1 = \sqrt{\pi/2}$, upper bound on $\Psi_P(\cdot)$; indeed,

$$\begin{array}{rcl}\mathbf{E}_{\xi\sim\mathcal{N}(0,Q)}\left\{\|H^T\xi\|\right\} & = & \mathbf{E}_{\eta\sim\mathcal{N}(0,I_\ell)}\left\{\|RH^TQ^{1/2}\eta\|_p\right\}\\ & = & \mathbf{E}_{\eta\sim\mathcal{N}(0,I_\ell)}\left\{\|[\mathrm{Row}_1^T[RH^TQ^{1/2}]\eta; ...; \mathrm{Row}_\mu^T[RH^TQ^{1/2}]\eta]\|_p\right\}\\ & = & \mathbf{E}_{\eta\sim\mathcal{N}(0,I_m)}\left\{\|[|\mathrm{Row}_1^T[RH^TQ^{1/2}]\eta|; ...; |\mathrm{Row}_\mu^T[RH^TQ^{1/2}]\eta|]\|_p\right\}\\ & \geq & \|[\mathbf{E}_\eta\{|\mathrm{Row}_1^T[RH^TQ^{1/2}]\eta|\}; ...; \mathbf{E}_\eta\{|\mathrm{Row}_\mu^T[RH^TQ^{1/2}]\eta|\}]\|_p\\ & & \text{[Jensen's inequality]}\\ & = & c_1\left\|\left[\|\mathrm{Row}_1[RH^TQ^{1/2}]\|_2; ...; \|\mathrm{Row}_\mu[RH^TQ^{1/2}]\|_2\right]\right\|_p.\end{array}$$

[15] For alternative wider scope scheme of bounding Ψ_P, see Section 4.5.

The bottom line is that in the case of (4.3.30) and under Assumption **A**, the efficiently solvable convex problem

$$\begin{array}{r}\mathrm{Opt}=\min\limits_{H}\Big\{\phi_{\mathcal{T}}(\lambda)+\phi_{\mathcal{R}}(\mu)+c_p\big|\big|[\|\mathrm{Row}_1[RH^TQ^{1/2}]\|_2;...;\|\mathrm{Row}_\mu[RH^TQ^{1/2}]\|_2]\big|\big|_p:\\ \lambda\geq 0,\mu\geq 0,\left[\begin{array}{c|c}\sum_k\lambda_kS_k & \frac{1}{2}V^T\\ \hline \frac{1}{2}V & \sum_\ell Q_\ell\end{array}\right]\succeq 0\Big\}\\ \left[\phi_{\mathcal{T}}(\lambda)=\max_{t\in\mathcal{T}}\lambda^Tt,\ \phi_{\mathcal{R}}(\mu)=\max_{r\in R}\mu^Tr\right]\end{array}$$

is a safe tractable approximation of the problem of building linear estimate of Bx with minimum $\|\cdot\|$-risk, and the risk of the estimate yielded by the optimal solution to the problem is upper-bounded by Opt. Note that now we do *not* claim that the resulting estimate is near-optimal among all possible estimates.

4.4 More extensions – from ellitopes to spectratopes

So far, the domains of signals we dealt with were ellitopes. In this section we demonstrate that basically all our constructions and results can be extended onto a much wider family of signal domains, namely, *spectratopes.*

4.4.1 Spectratopes: definition and examples

We call a set $\mathcal{X}\subset\mathbf{R}^n$ a *basic spectratope*, if it admits *simple spectratopic representation* – representation of the form

$$\mathcal{X}=\left\{x\in\mathbf{R}^n:\exists t\in\mathcal{T}:R_k^2[x]\preceq t_kI_{d_k},1\leq k\leq K\right\}\tag{4.4.1}$$

where

S.1. $R_k[x]=\sum_{i=1}^n x_iR^{ki}$ are symmetric $d_k\times d_k$ matrices linearly depending on $x\in\mathbf{R}^n$ (i,e., "matrix coefficients" R^{ki} belong to $\mathbf{S}^n$)

S.2. $\mathcal{T}\in\mathbf{R}^K_+$ is the set with the same properties as in the definition of an ellitope, that is, $\mathcal{T}$ is a convex compact subset of $\mathbf{R}^K_+$ which contains a positive vector and is monotone:

$$0\leq t'\leq t\in\mathcal{T}\Rightarrow t'\in\mathcal{T}.$$

S.3. Whenever $x\neq 0$, it holds $R_k[x]\neq 0$ for at least one $k\leq K$.

An immediate observation (check it!) is as follows:

Remark 4.4.1 *By Schur Complement Lemma, the set* (4.4.1) *given by data satisfying S.1-2 can be represented as*

$$\mathcal{X}=\left\{x\in\mathbf{R}^n:\exists t\in\mathcal{T}:\left[\begin{array}{c|c}t_kI_{d_k} & R_k[x]\\ \hline R_k[x] & I_{d_k}\end{array}\right]\succeq 0,\ k\leq K\right\}$$

By the latter representation, $\mathcal{X}$ is nonempty, closed, convex, symmetric w.r.t. the origin and contains a neighbourhood of the origin. This set is bounded if and only if the data, in addition to S.1-2, satisfies S.3.

A spectratope $\mathcal{X} \subset \mathbf{R}^\nu$ is a set represented as linear image of a basic spectratope:

$$\mathcal{X} = \{x \in \mathbf{R}^\nu : \exists (y \in \mathbf{R}^n, t \in \mathcal{T}) : x = Py, R_k^2[y] \preceq t_k I_{d_k}, 1 \leq k \leq K\}, \tag{4.4.2}$$

where P is a $\nu \times n$ matrix, and $R_k[\cdot]$, $\mathcal{T}$ are as in S.1-3.

We associate with a basic spectratope (4.4.1), S.1-3 the following entities:

1. The *size*
$$D = \sum_{k=1}^K d_k;$$

2. Linear mappings
$$Q \mapsto \mathcal{R}_k[Q] = \sum_{i,j} Q_{ij} R^{ki} R^{kj} : \mathbf{S}^n \to \mathbf{S}^{d_k} \tag{4.4.3}$$
As is immediately seen, we have
$$\mathcal{R}_k[xx^T] \equiv R_k^2[x], \tag{4.4.4}$$
implying that $\mathcal{R}_k[Q] \succeq 0$ whenever $Q \succeq 0$, whence $\mathcal{R}_k[\cdot]$ is $\succeq$-monotone:
$$Q' \succeq Q \Rightarrow \mathcal{R}_k[Q'] \succeq \mathcal{R}_k[Q]. \tag{4.4.5}$$
Besides this, we have
$$Q \succeq 0 \Rightarrow \mathbf{E}_{\xi \sim \mathcal{N}(0,Q)}\{R_k^2[\xi]\} = \mathbf{E}_{\xi \sim \mathcal{N}(0,Q)}\{\mathcal{R}_k[\xi\xi^T]\} = \mathcal{R}_k[Q], \tag{4.4.6}$$
where the first equality is given by (4.4.4).

3. Linear mappings $\Lambda_k \mapsto \mathcal{R}_k^*[\Lambda_k] : \mathbf{S}^{d_k} \to \mathbf{S}^n$ given by
$$[\mathcal{R}_k^*[\Lambda_k]]_{ij} = \frac{1}{2}\mathrm{Tr}(\Lambda_k[R^{ki}R^{kj} + R^{kj}R^{ki}]),\ 1 \leq i, j \leq n. \tag{4.4.7}$$
It is immediately seen that $\mathcal{R}_k^*[\cdot]$ is the conjugate of $\mathcal{R}_k[\cdot]$:
$$\langle \Lambda_k, \mathcal{R}_k[Q]\rangle_F = \mathrm{Tr}(\Lambda_k \mathcal{R}_k[Q]) = \mathrm{Tr}(\mathcal{R}_k^*[\Lambda_k]Q) = \langle \mathcal{R}_k^*[\Lambda_k], Q\rangle_F, \tag{4.4.8}$$
where $\langle A, B\rangle_F = \mathrm{Tr}(AB)$ is the Frobenius inner product of symmetric matrices. Besides this, we have
$$\Lambda_k \succeq 0 \Rightarrow \mathcal{R}_k^*[\Lambda_k] \succeq 0. \tag{4.4.9}$$
Indeed, $\mathcal{R}_k^*[\Lambda_k]$ is linear in Λ_k, so that it suffices to verify (4.4.9) for dyadic matrices $\Lambda_k = ff^T$; for such a Λ_k, (4.4.7) reads
$$(\mathcal{R}_k^*[ff^T])_{ij} = [R^{ki}f]^T[R^{kj}f],$$
that is, $\mathcal{R}_k^*[ff^T]$ is a Gram matrix and as such is $\succeq 0$. Another way to arrive at (4.4.9) is to note that when $\Lambda_k \succeq 0$ and $Q = xx^T$, the first quantity in (4.4.8) is nonnegative by (4.4.4), and therefore (4.4.8) states that $x^T\mathcal{R}_k^*[\Lambda_k]x \geq 0$ for every x, implying $\mathcal{R}_k^*[\Lambda_k] \succeq 0$.

4. The linear space $\Lambda^K = \mathbf{S}^{d_1} \times ... \times \mathbf{S}^{d_K}$ of all ordered collections $\Lambda = \{\Lambda_k \in \mathbf{S}^{d_k}\}_{k \leq K}$ along with the linear mapping
$$\Lambda \mapsto \lambda[\Lambda] := [\mathrm{Tr}(\Lambda_1); ...; \mathrm{Tr}(\Lambda_K)] : \Lambda^K \to \mathbf{R}^K. \tag{4.4.10}$$

4.4.1.1 Examples of spectratopes

Example 1: Ellitopes. Every ellitope

$$\mathcal{X} = \{x \in \mathbf{R}^\nu : \exists (y \in \mathbf{R}^n, t \in \mathcal{T}) : x = Py, y^T S_k y \leq t_k,\, k \leq K\} \qquad [S_k \succeq 0, \sum_k S_k \succ 0]$$

is a spectratope as well. Indeed, let $S_k = \sum_{j=1}^{r_k} s_{kj}s_{kj}^T$, $r_k = \mathrm{Rank}(S_k)$, be a dyadic representation of the positive semidefinite matrix S_k, so that

$$y^T S_k y = \sum_j (s_{kj}^T y)^2 \;\forall y,$$

and let

$$\widehat{\mathcal{T}} = \{\{t_{kj} \geq 0, 1 \leq j \leq r_k, 1 \leq k \leq K\} : \exists t \in \mathcal{T} : \sum_j t_{kj} \leq t_k\},\;\; R_{kj}[y] = s_{kj}^T y \in \mathbf{S}^1 = \mathbf{R}.$$

We clearly have

$$\mathcal{X} = \{x \in \mathbf{R}^\nu : \exists(\{t_{kj}\} \in \widehat{\mathcal{T}}, y) : x = Py, R_{kj}^2[y] \preceq t_{kj} I_1 \,\forall k, j\}$$

and the right hand side is a legitimate spectratopic representation of $\mathcal{X}$.

Example 2: "Matrix box." Let L be a positive definite $d \times d$ matrix. Then the "matrix box"

$$\begin{array}{rcl}\mathcal{X} & = & \{X \in \mathbf{S}^d : -L \preceq X \preceq L\} = \{X \in \mathbf{S}^d : -I_d \preceq L^{-1/2} X L^{-1/2} \preceq I_d\} \\ & = & \{X \in \mathbf{S}^d : R^2[X] := [L^{-1/2} X L^{-1/2}]^2 \preceq I_d\}\end{array}$$

is a basic spectratope (augment $R_1[\cdot] := R[\cdot]$ with $K = 1$, $\mathcal{T} = [0,1]$). As a result, a *bounded* set $\mathcal{X} \subset \mathbf{R}^\nu$ given by a system of "two-sided" Linear Matrix Inequalities, specifically,

$$\mathcal{X} = \{x \in \mathbf{R}^\nu : \exists t \in \mathcal{T} : -\sqrt{t_k} L_k \preceq S_k[x] \preceq \sqrt{t_k} L_k, k \leq K\}$$

where $S_k[x]$ are symmetric $d_k \times d_k$ matrices linearly depending on x, $L_k \succ 0$ and $\mathcal{T}$ satisfies S.2, is a basic spectratope:

$$\mathcal{X} = \{x \in \mathbf{R}^\nu : \exists t \in \mathcal{T} : R_k^2[x] \leq t_k I_{d_k},\, k \leq K\} \qquad [R_k[x] = L_k^{-1/2} S_k[x] L_k^{-1/2}]$$

Same as ellitopes, spectratopes admit fully algorithmic calculus, see Section 4.8.

4.4.2 Near-optimal linear estimation on spectratopes

4.4.2.1 Situation and goal

Consider the problem of estimating the linear image Bx of unknown signal x, known to belong to a given set $\mathcal{X}$, via observations

$$\omega = Ax + \xi,\; \xi \sim \mathcal{N}(0, \Gamma) \in \mathbf{R}^m$$

of Ax corrupted by Gaussian noise. We have considered this problem, for an ellitope $\mathcal{X}$, in Sections 4.2 and 4.3.1. Now we are about to consider this problem in the case when $\mathcal{X}$ is a spectratope; by the same reasons as in the ellitopic case, we lose nothing when assuming from now on that this spectratope is basic:

$$\mathcal{X} = \{x \in \mathbf{R}^n : \exists t \in \mathcal{T} : R_k^2[x] \preceq t_k I_{d_k}, 1 \leq k \leq K\} \tag{4.4.11}$$

Given positive semidefinite matrix S, we want to build a nearly optimal, in terms of its S-risk on $\mathcal{X}$, among linear estimates, and to show that it is nearly optimal among all estimates, non necessarily linear ones.

As always, we associate with the set $\mathcal{T}$ participating in (4.4.11) its support function

$$\phi_{\mathcal{T}}(\lambda) = \max_{t \in \mathcal{T}} \lambda^T t.$$

From now on, we assume that $B \neq 0$ and $\Gamma \succ 0$.

4.4.2.2 Building linear estimate

Our first step is to build a good linear estimate. In the ellitopic case, it was yielded by an optimal solution to convex problem (4.3.2). The current analogy is given by

Proposition 4.4.1 *Consider convex optimization problem*

$$\mathrm{Opt} = \min_{H,\Lambda=\{\Lambda_k,k\leq K\},\tau}\left\{\tau : \begin{array}{l}(B-H^TA)^T(B-H^TA)\preceq\sum_k\mathcal{R}_k^*(\Lambda_k)+\tau S\\ \Lambda_k\succeq 0, k\leq K, \mathrm{Tr}(H^T\Gamma H)+\phi_{\mathcal{T}}(\lambda[\Lambda])\leq\tau\end{array}\right\}. \tag{4.4.12}$$

The problem is solvable, and a feasible solution (H,λ,τ) to the problem induces linear estimate $\widehat{x}_H = H^T\omega$ of Bx, $x\in\mathcal{X}$, via observation

$$\omega = Ax+\xi,\, \xi\sim\mathcal{N}(0,\Gamma)$$

with S-risk not exceeding $\sqrt{\tau}$.

The crucial observation underlying this result is important by its own right and reads as follows:

Lemma 4.4.1 *Let $\mathcal{X}$ be spectratope (4.4.11), $S\in\mathbf{S}^n_+$ and $Q\in\mathbf{S}^n$. Whenever $\Lambda_k\in\mathbf{S}^{d_k}_+$ and $\tau\geq 0$ satisfy*

$$Q\preceq\sum_k\mathcal{R}_k^*[\Lambda_k]+\tau S,$$

we have

$$x^TQx\leq\tau x^TSx+\phi_{\mathcal{T}}(\lambda[\Lambda]),\ \lambda[\Lambda]=[\mathrm{Tr}(\Lambda_1);...;\mathrm{Tr}(\Lambda_K)].$$

For proofs, see Sections 4.7.3.2, 4.7.3.1.

4.4.2.3 Near-optimality of the estimate

Our current analogy of Proposition 4.3.1 is as follows:

Proposition 4.4.2 *Let $\mathcal{X}$ be given by (4.4.11), and let*

$$\mathcal{Q}=\{Q\in\mathbf{S}^n_+ : \exists t\in\mathcal{T} : \mathcal{R}_k[Q]\preceq t_kI_{d_k}, k\leq K\},\ \mathcal{Q}_\rho=\rho\mathcal{Q}, \rho>0. \tag{4.4.13}$$

The set $\mathcal{Q}$ is a nonempty convex compact set containing a neighbourhood of origin, so that the quantity

$$M_* = \sqrt{\max_{Q\in\mathcal{Q}}\mathrm{Tr}(BQB^T)}, \tag{4.4.14}$$

is well defined and positive. The efficiently computable linear estimate $\widehat{x}_{H_*}(\omega)=H_*^T\omega$ yielded by an optimal solution to the optimization problem in (4.4.12) is nearly optimal in terms of S-risk:

$$\mathrm{RiskS}[\widehat{x}_{H_*}|\mathcal{X}]\leq\sqrt{\mathrm{Opt}}\leq\sqrt{8\ln\left(\frac{6M_*^2\sqrt{2D}}{\mathrm{Risks}^2_{\mathrm{opt}}[\mathcal{X}]}\right)}\mathrm{Risks}_{\mathrm{opt}}[\mathcal{X}],$$

where

$$\mathrm{Risks}_{\mathrm{opt}}[\mathcal{X}]=\inf_{\widehat{x}(\cdot)}\mathrm{RiskS}[\widehat{x}|\mathcal{X}]$$

is the minimax S-risk associated with $\mathcal{X}$, M_* is given by (4.4.14), Opt is given by (4.4.12), and $D=\sum_k d_k$.

For proof, see Section 4.7.3.4. The key argument in the proof is the following analogy of Lemma 4.3.1:

Lemma 4.4.2 *Consider the problem*

$$\mathrm{Opt}_* = \max_{W,G,s,v} \left\{ \mathrm{Tr}(BWB^T) - \mathrm{Tr}(G) : \begin{array}{l} \left[\begin{array}{cc} G & BWA^T \\ AWB^T & s\Gamma + AWA^T \end{array}\right] \succeq 0, \\ W \succeq 0,\, \mathcal{R}_k[W] \preceq v_k I_{d_k},\, 1 \le k \le K, \\ \mathrm{Tr}(WS) + s \le 1,\, [v;s] \in \mathbf{T} \end{array} \right\} \tag{4.4.15}$$

where

$$\mathbf{T} = \mathrm{cl}\{[t;\tau] \in \mathbf{R}^K \times \mathbf{R} : \tau > 0, \tau^{-1}t \in \mathcal{T}\} \subset \mathbf{R}^{K+1}_+ \tag{4.4.16}$$

is a closed and pointed convex cone in $\mathbf{R}^{K+1}$ *with a nonempty interior. Problem* (4.4.15) *is strictly feasible and solvable. Furthermore, if* $(W, G, [v; s])$ *is an optimal solution to* (4.4.15)*, then* $s > 0$*, and*

$$\mathrm{Opt} = \mathrm{Opt}_* = \mathrm{Tr}\big(B[W - WA^T(s\Gamma + AWA^T)^{-1}AW]B^T\big). \tag{4.4.17}$$

For proof, see Section 4.7.3.3.

4.4.3 Maximizing quadratic forms over spectratopes

Now let us extend to our current situation Proposition 4.3.3. The extension reads as follows:

Proposition 4.4.3 *Let* C *be a symmetric* $n \times n$ *matrix and* $\mathcal{X}$ *be given by spectratopic representation*

$$\mathcal{X} = \{x \in \mathbf{R}^n : \exists y \in \mathbf{R}^\mu, t \in \mathcal{T} : x = Py, R_k^2[y] \preceq t_k I_{d_k}, k \le K\},$$

let

$$\mathrm{Opt} = \max_{x \in \mathcal{X}} x^T C x$$

and

$$\mathrm{Opt}_* = \min_{\Lambda = \{\Lambda_k\}_{k \le K}} \left\{ \phi_{\mathcal{T}}(\lambda[\Lambda]) : \Lambda_k \succeq 0, P^T C P \preceq \textstyle\sum_k \mathcal{R}_k^*[\Lambda_k] \right\} \\ [\lambda[\Lambda] = [\mathrm{Tr}(\Lambda_1); ...; \mathrm{Tr}(\Lambda_K)]] \tag{4.4.18}$$

Then (4.4.18) *is solvable, and*

$$\mathrm{Opt} \le \mathrm{Opt}_* \le 2\max[\ln(2D), 1]\mathrm{Opt},\ \ D = \sum_k d_k. \tag{4.4.19}$$

For proof, see Section 4.7.3.5.

4.5 Linear estimates beyond $\|\cdot\|_2$-risk revisited

In Section 4.3.5, we have developed a computationally efficient scheme for building "presumably good" linear estimates of the linear image Bx of unknown signal x known to belong to a given ellitope $\mathcal{X}$ in the case when the risk is defined as the worst, w.r.t. $x \in \mathcal{X}$, expected norm $\|\cdot\|$ of the recovery error. We are about to extend these results to the case when $\mathcal{X}$ is a spectratope, and to demonstrate that the resulting linear estimates are not just "presumably good," but possess near-optimality properties completely similar to those stated in Propositions 4.2.2, 4.3.1, 4.4.2 dealing with the case of $\|\cdot\| = \|\cdot\|_2$. Besides this, in what follows we somehow relax our assumptions on observation noise.

4.5.1 Situation and goal

Same as in Section 4.3.5, we consider the problem of recovering the image $Bx \in \mathbf{R}^\nu$ of a signal $x \in \mathbf{R}^n$ known to belong to a given spectratope

$$\mathcal{X} = \{x \in \mathbf{R}^n : \exists t \in \mathcal{T} : R_k^2[x] \preceq t_k I_{d_k},\, 1 \leq k \leq K\} \tag{4.5.1}$$

from noisy observation

$$\omega = Ax + \xi, \tag{4.5.2}$$

where A is a known $m \times n$ matrix, and ξ is random observation noise.

Observation noise. In typical signal processing applications, the distribution of noise is fixed and is part of the data of the estimation problem. In order to cover some applications (e.g., the one in Section 4.5.3), we allow for "ambiguous" noise distributions; all we know is that this distribution belongs to a family $\mathcal{P}$ of Borel probability distributions on $\mathbf{R}^m$ associated with a given convex compact subset Π of the interior of the cone $\mathbf{S}^m_+$ of positive semidefinite $m \times m$ matrices, "association" meaning that the matrix of second moments of every distribution $P \in \mathcal{P}$ is $\succeq$-dominated by a matrix from Π:

$$P \in \mathcal{P} \Rightarrow \exists Q \in \Pi : \mathrm{Vary}[P] := \mathbf{E}_{\xi\sim P}\{\xi\xi^T\} \preceq Q. \tag{4.5.3}$$

Actual distribution of noise in (4.5.2) is somehow selected from $\mathcal{P}$ by nature (and may, e.g., depend on x).

In the sequel, for a probability distribution P on $\mathbf{R}^m$ we write $P \lll \Pi$ to express the fact that the matrix of second moments of P is $\succeq$-dominated by a matrix from Π:

$$\{P \lll \Pi\} \Leftrightarrow \{\exists \Theta \in \Pi : \mathrm{Vary}[P] \preceq \Theta\}.$$

Quantifying risk. Given Π and a norm $\|\cdot\|$ on $\mathbf{R}^\nu$, we quantify the quality of a candidate estimate $\widehat{x}(\cdot) : \mathbf{R}^m \to \mathbf{R}^\nu$ by its $(\Pi, \|\cdot\|)$-risk on $\mathcal{X}$ defined as

$$\mathrm{Risk}_{\Pi,\|\cdot\|}[\widehat{x}|\mathcal{X}] = \sup_{x\in\mathcal{X}, P\lll\Pi} \mathbf{E}_{\xi\sim P}\left\{\|\widehat{x}(Ax+\xi) - Bx\|\right\}. \tag{4.5.4}$$

Goal. As before, our focus is on *linear estimates* – estimates of the form

$$\widehat{x}_H(\omega) = H^T\omega$$

given by $m \times \nu$ matrices H; our ultimate goal is to demonstrate that under some restrictions on the signal domain $\mathcal{X}$, "presumably good" linear estimate yielded by an optimal solution to an efficiently solvable convex optimization problem is near-optimal in terms of its risk among *all* estimates, linear and nonlinear alike.

4.5.1.1 Assumptions

From now on we make the following assumptions:

> **Assumption A′:** *The unit ball $\mathcal{B}_*$ of the norm $\|\cdot\|_*$ conjugate to the norm $\|\cdot\|$ participating in the formulation of our estimation problem is a spectratope:*
>
> $$\mathcal{B}_* = \{z \in \mathbf{R}^\nu : \exists y \in \mathcal{Y} : z = My\},\ \mathcal{Y} := \{y \in \mathbf{R}^q : \exists r \in \mathcal{R} : S_\ell^2[y] \preceq r_\ell I_{f_\ell},\, 1 \leq \ell \leq L\}, \tag{4.5.5}$$
>
> *where the right hand side data are as required in a spectratopic representation.*

Note that Assumption $\mathbf{A}'$ is weaker than Assumption $\mathbf{A}$ from Section 4.3.5 (recall that ellitopes are spectratopes as well). A potentially useful example of norm $\|\cdot\|$ which obeys Assumption $\mathbf{A}'$ and seemingly does not obey assumption $\mathbf{A}$ is the *nuclear norm* $\|V\|_{\mathrm{Sh},1}$ on the space $\mathbf{R}^\nu = \mathbf{R}^{p\times q}$ of $p\times q$ matrices – the sum of singular values of a matrix V; the conjugate norm is the spectral norm $\|\cdot\| = \|\cdot\|_{\mathrm{Sh},\infty}$ on $\mathbf{R}^\nu = \mathbf{R}^{p\times q}$, and the unit ball of the latter norm is a spectratope:

$$\{X\in\mathbf{R}^{p\times q}: \|X\|\leq 1\} = \{X: \exists t\in\mathcal{T}=[0,1]: R^2[X]\preceq tI_{p+q}\},\ R[X]=\left[\begin{array}{c|c} & X^T\\ \hline X & \end{array}\right].$$

Besides Assumption $\mathbf{A}'$, we make the following regularity assumption:

Assumption R: *All matrices from* Π *are positive definite.*

4.5.2 Building linear estimate

Let $H\in\mathbf{R}^{m\times\nu}$. According to (4.3.24), the risk of the linear estimate $\widehat{x}_H(\omega) = H^T\omega$ can be upper-bounded as follows:

$$\begin{array}{rcl}
\mathrm{Risk}_{\Pi,\|\cdot\|}[\widehat{x}_H(\cdot)|\mathcal{X}] &\leq& \|B-H^TA\|_{\mathcal{X},\|\cdot\|} + \Psi_\Pi(H),\\
\text{where} && \\
\|V\|_{\mathcal{X},\|\cdot\|} &=& \max_x\{\|Vx\|: x\in\mathcal{X}\}: \mathbf{R}^{k\times n}\to\mathbf{R},\\
\Psi_\Pi(H) &=& \sup\limits_{P\lll\Pi}\mathbf{E}_{\xi\sim P}\left\{\|H^T\xi\|\right\}.
\end{array}\qquad(4.5.6)$$

Same as in Section 4.3.5, we need to derive efficiently computable upper bounds on the norm $\|\cdot\|_{\mathcal{X},\|\cdot\|}$ and the function Ψ_Π.

4.5.2.1 Upper-bounding $\|\cdot\|_{\mathcal{X},\|\cdot\|}$

With Assumption $\mathbf{A}'$ in force, consider the spectratope

$$\begin{array}{rcl}
\mathcal{Z} &:=& \mathcal{X}\times\mathcal{Y} = \{[x;y]\in\mathbf{R}^n\times\mathbf{R}^q: \exists s=[t;r]\in\mathcal{T}\times\mathcal{R}:\\
&& \qquad R_k^2[x]\preceq t_kI_{d_k},\ 1\leq k\leq K, S_\ell^2[y]\preceq r_\ell I_{f_\ell},\ 1\leq\ell\leq L\}\\
&=& \{w=[x;y]\in\mathbf{R}^n\times\mathbf{R}^q: \exists s=[t;r]\in\mathcal{S}=\mathcal{T}\times\mathcal{R}: U_i^2[w]\preceq s_iI_{g_i},\\
&& \qquad 1\leq i\leq I=K+L\}
\end{array}\qquad(4.5.7)$$

with $U_\ell[\cdot]$ readily given by $R_k[\cdot]$ and $S_\ell[\cdot]$. Same as in the ellitopic case, given a $\nu\times n$ matrix V, with

$$W[V] = \frac{1}{2}\left[\begin{array}{c|c} & V^TM\\ \hline M^TV & \end{array}\right]$$

it holds

$$\|V\|_{\mathcal{X},\|\cdot\|} = \max_{x\in\mathcal{X}}\|Vx\| = \max_{x\in\mathcal{X},z\in\mathcal{B}_*} z^TVx = \max_{x\in\mathcal{X},y\in\mathcal{Y}} y^TM^TVx = \max_{w\in\mathcal{Z}} w^TW[V]w.$$

Applying Proposition 4.4.3, we arrive at the following version of Corollary 4.3.1:

Corollary 4.5.1 *In the just defined situation, the efficiently computable convex function*

$$\begin{array}{rcl}
\|V\|^+_{\mathcal{X},\|\cdot\|} &=& \min\limits_{\Lambda,\Upsilon}\left\{\phi_{\mathcal{T}}(\lambda[\Lambda]) + \phi_{\mathcal{R}}(\lambda[\Upsilon]): \Lambda=\{\Lambda_k\in\mathbf{S}_+^{d_k}\}_{k\leq K}, \Upsilon=\{\Upsilon_\ell\in\mathbf{S}_+^{f_\ell}\}_{\ell\leq L},\right.\\
&& \left.\qquad\left[\begin{array}{c|c}\sum_k\mathcal{R}_k^*[\Lambda_k] & \frac{1}{2}V^TM\\ \hline \frac{1}{2}M^TV & \sum_\ell\mathcal{S}_\ell^*[\Upsilon_\ell]\end{array}\right]\succeq 0\right\}\\
\multicolumn{3}{l}{\left[\begin{array}{l}
\phi_{\mathcal{T}}(\lambda) = \max\limits_{t\in\mathcal{T}}\lambda^Tt,\ \phi_{\mathcal{R}}(\lambda) = \max\limits_{r\in\mathcal{R}}\lambda^Tr,\ \lambda[\{\Xi_1,...,\Xi_N\}] = [\mathrm{Tr}(\Xi_1);...;\mathrm{Tr}(\Xi_N)],\\
{[\mathcal{R}_k^*[\Lambda_k]]_{ij}} = \frac{1}{2}\mathrm{Tr}(\Lambda_k[R_k^{ki}R_k^{kj}+R_k^{kj}R_k^{ki}]),\ \text{where } R_k[x]=\sum_i x_iR^{ki},\\
{[\mathcal{S}_\ell^*[\Upsilon_\ell]]_{ij}} = \frac{1}{2}\mathrm{Tr}(\Upsilon_\ell[S_\ell^{\ell i}S_\ell^{\ell j}+S_\ell^{\ell j}S_\ell^{\ell i}]),\ \text{where } S_\ell[y]=\sum_i y_iS^{\ell i}.
\end{array}\right]}
\end{array}\qquad(4.5.8)$$

is a norm on $\mathbf{R}^{\nu\times n}$, *and this norm is a tight upper bound on* $\|\cdot\|_{\mathcal{X},\|\cdot\|}$, *namely,*

$$\begin{array}{c}\forall V\in\mathbf{R}^{\nu\times n}: \|V\|_{\mathcal{X},\|\cdot\|}\leq\|V\|^+_{\mathcal{X},\|\cdot\|}\leq 2\max[\ln(2D),1]\|V\|_{\mathcal{X},\|\cdot\|},\\ D=\sum_k d_k+\sum_\ell f_\ell.\end{array}\tag{4.5.9}$$

4.5.2.2 Upper-bounding $\Psi_\Pi(\cdot)$

On the top of upper bounds presented in Section 4.3.5, we derive here another upper bound on Ψ_Π capable to handle any norm obeying Assumption $\mathbf{A}'$. The underlying observation is as follows:

Lemma 4.5.1 *Let V be a $m\times\nu$ matrix, $Q\in\mathbf{S}^m_+$, and P be a probability distribution on $\mathbf{R}^m$ with* $\mathrm{Vary}[P]\preceq Q$. *Let, further, $\|\cdot\|$ be a norm on $\mathbf{R}^\nu$ with the unit ball $\mathcal{B}_*$ of the conjugate norm $\|\cdot\|_*$ given by* (4.5.5). *Finally, let $\Upsilon=\{\Upsilon_\ell\in\mathbf{S}^{f_\ell}_+\}_{\ell\leq L}$ and a matrix $\Theta\in\mathbf{S}^m$ satisfy the constraint*

$$\left[\begin{array}{c|c}\Theta & \frac{1}{2}VM\\ \hline \frac{1}{2}M^TV^T & \sum_\ell\mathcal{S}^*_\ell[\Upsilon_\ell]\end{array}\right]\succeq 0\tag{4.5.10}$$

(for notation, see (4.5.5), (4.5.8)*). Then*

$$\mathbf{E}_{\eta\sim P}\{\|V^T\eta\|\}\leq\mathrm{Tr}(Q\Theta)+\phi_{\mathcal{R}}(\lambda[\Upsilon]).\tag{4.5.11}$$

Proof is immediate. In the case of (4.5.10), we have

$$\begin{array}{rcl}\|V^T\xi\| & = & \max\limits_{z\in\mathcal{B}_*}z^TV^T\xi=\max\limits_{y\in\mathcal{Y}}y^TM^TV^T\xi\\ & \leq & \max\limits_{y\in\mathcal{Y}}\left[\xi^T\Theta\xi+\sum_\ell y^T\mathcal{S}^*_\ell[\Upsilon_\ell]y\right]\ \ [\text{by }\ (4.5.10)]\\ & = & \max\limits_{y\in\mathcal{Y}}\left[\xi^T\Theta\xi+\sum_\ell\mathrm{Tr}(\mathcal{S}^*_\ell[\Upsilon_\ell]yy^T)\right]\\ & = & \max\limits_{y\in\mathcal{Y}}\left[\xi^T\Theta\xi+\sum_\ell\mathrm{Tr}(\Upsilon_\ell S^2_\ell[y])\right]\ \ [\text{by (4.4.4) and (4.4.8)}]\\ & = & \xi^T\Theta\xi+\max\limits_{y,r}\left\{\sum_\ell\mathrm{Tr}(\Upsilon_\ell S^2_\ell[y]):\ S^2_\ell[y]\preceq r_\ell I_{f_\ell},\ell\leq L,r\in\mathcal{R}\right\}\ \ [\text{by (4.5.5)}]\\ & \leq & \xi^T\Theta\xi+\max\limits_{r\in\mathcal{R}}\sum_\ell\mathrm{Tr}(\Upsilon_\ell)r_\ell\ [\text{by }\Upsilon_\ell\succeq 0]\\ & \leq & \xi^T\Theta\xi+\phi_{\mathcal{R}}(\lambda[\Upsilon]).\end{array}$$

Taking expectation of both sides of the resulting inequality w.r.t. distribution P of ξ and taking into account that $\mathrm{Tr}(\mathrm{Vary}[P]\Theta)\leq\mathrm{Tr}(Q\Theta)$ due to $\Theta\succeq 0$ (by (4.5.10)) and $\mathrm{Vary}[P]\preceq Q$, we get (4.5.11). □

Note that when $P=\mathcal{N}(0,Q)$, the smallest possible upper bound on $\mathbf{E}_{\eta\sim P}\{\|V^T\eta\|\}$ which can be extracted from Lemma 4.5.1 (this bound is efficiently computable) is tight, see Lemma 4.5.2 below.

An immediate consequence is

Corollary 4.5.2 *Let*

$$\Gamma(\Theta)=\max_{Q\in\Pi}\mathrm{Tr}(Q\Theta)\tag{4.5.12}$$

and

$$\overline{\Psi}_\Pi(H)=\min_{\{\Upsilon_\ell\}_{\ell\leq L},\Theta\in\mathbf{S}^m}\left\{\Gamma(\Theta)+\phi_{\mathcal{R}}(\lambda[\Upsilon]):\Upsilon_\ell\succeq 0\,\forall\ell,\ \left[\begin{array}{c|c}\Theta & \frac{1}{2}HM\\ \hline \frac{1}{2}M^TH^T & \sum_\ell\mathcal{S}^*_\ell[\Upsilon_\ell]\end{array}\right]\succeq 0\right\}\tag{4.5.13}$$

Then $\overline{\Psi}_\Pi(\cdot):\ \mathbf{R}^{m\times\nu}\to\mathbf{R}$ is efficiently computable convex upper bound on $\Psi_\Pi(\cdot)$.

Indeed, given Lemma 4.5.1, the only non-evident part of the corollary is that $\overline{\Psi}_\Pi(\cdot)$ is a well-defined real-valued function, which is readily given by Lemma 4.7.2, see Section 4.7.3.2.

Remark 4.5.1 When $\Upsilon=\{\Upsilon_\ell\}_{\ell\le L}$, Θ is a feasible solution to the right hand side problem in (4.5.13) and $s>0$, the pair $\Upsilon'=\{s\Upsilon_\ell\}_{\ell\le L}$, $\Theta'=s^{-1}\Theta$ also is a feasible solution; since $\phi_{\mathcal{R}}(\cdot)$ and $\Gamma(\cdot)$ are positive homogeneous of degree 1, we conclude that $\overline{\Psi}_\Pi$ is in fact the infimum of the function

$$2\sqrt{\Gamma(\Theta)\phi_{\mathcal{R}}(\lambda[\Upsilon])}=\inf_{\theta>0}\left[s^{-1}\Gamma(\Theta)+s\phi_{\mathcal{R}}(\lambda[\Upsilon])\right]$$

over Υ,Θ satisfying the constraints of the problem (4.5.13).

In addition, for every feasible solution $\Upsilon=\{\Upsilon_\ell\}_{\ell\le L}$, Θ to the problem (4.5.13) with $\mathcal{M}[\Upsilon]:=\sum_\ell \mathcal{S}_\ell^*[\Upsilon_\ell]\succ 0$, the pair Υ, $\widehat{\Theta}=\frac{1}{4}HM\mathcal{M}^{-1}[\Upsilon]M^TH^T$ is feasible for the problem as well and $0\preceq\widehat{\Theta}\preceq\Theta$ (Schur Complement Lemma), so that $\Gamma(\widehat{\Theta})\le\Gamma(\Theta)$. As a result,

$$\overline{\Psi}_\Pi(H)=\inf_{\Upsilon}\left\{\begin{array}{r}\frac{1}{4}\Gamma(HM\mathcal{M}^{-1}[\Upsilon]M^TH^T)+\phi_{\mathcal{R}}(\lambda[\Upsilon]):\\ \Upsilon=\{\Upsilon_\ell\in\mathbf{S}_+^{f_\ell}\}_{\ell\le L},\mathcal{M}[\Upsilon]\succ 0\end{array}\right\}. \tag{4.5.14}$$

Illustration. Consider the case when $\|u\|=\|u\|_p$ with $p\in[1,2]$, and let us apply the just described scheme for upper-bounding Ψ_Π, assuming $\{Q\}\subset\Pi\subset\{S\in\mathbf{S}_+^m:S\preceq Q\}$ for some given $Q\succ 0$, so that $\Gamma(\Theta)=\mathrm{Tr}(Q\Theta)$, $\Theta\succeq 0$. The unit ball of the norm conjugate to $\|\cdot\|$, that is, the norm $\|\cdot\|_q$, $q=\frac{p}{p-1}\in[2,\infty]$, is the basic spectratope (in fact, ellitope)

$$\mathcal{B}_*=\{y\in\mathbf{R}^\mu:\exists r\in\mathcal{R}:=\{\mathbf{R}_+^\nu:\|r\|_{q/2}\le 1\}:S_\ell^2[y]\le r_\ell,\,1\le\ell\le L=\nu\},\ S_\ell[y]=y_\ell.$$

As a result, Υ's from Remark 4.5.1 are collections of ν positive semidefinite 1×1 matrices, and we can identify them with ν-dimensional nonnegative vectors υ, resulting in $\lambda[\Upsilon]=\upsilon$ and $\mathcal{M}[\Upsilon]=\mathrm{Diag}\{\upsilon\}$. Besides this, for *nonnegative* υ we clearly have $\phi_{\mathcal{R}}(\upsilon)=\|\upsilon\|_{p/(2-p)}$. The optimization problem in (4.5.14) now reads

$$\overline{\Psi}_\Pi(H)=\inf_{\upsilon\in\mathbf{R}^\nu}\left\{\tfrac{1}{4}\mathrm{Tr}(V\mathrm{Diag}^{-1}\{\upsilon\}V^T)+\|\upsilon\|_{p/(2-p)}:\upsilon>0\right\}\qquad[V=Q^{1/2}H]$$

After setting $a_\ell=\|\mathrm{Col}_\ell[V]\|_2$, (4.5.14) becomes

$$\overline{\Psi}_\Pi(H)=\inf_{\upsilon>0}\left\{\frac{1}{4}\sum_\ell\frac{a_\ell^2}{\upsilon_\ell}+\|\upsilon\|_{p/(2-p)}\right\}.$$

This results in $\overline{\Psi}_\Pi(H)=\|[a_1;...;a_\mu]\|_p$. Recalling what a_ℓ and V are, we end up with

$$\forall P,\mathrm{Vary}[P]\preceq Q:\mathbf{E}_{\xi\sim P}\{\|H^T\xi\|\}\le\overline{\Psi}_\Pi(H):=\left\|\left[\|\mathrm{Row}_1[H^TQ^{1/2}]\|_2;\ldots;\|\mathrm{Row}_\nu[H^TQ^{1/2}]\|_2\right]\right\|_p,$$

which is a slightly spoiled version of (4.3.31); in the latter bound the right hand side contains also the factor $c_p\in[\sqrt{2/\pi},1]$. As a compensation, our present bound holds true for every distribution P with $\mathrm{Vary}[P]\preceq Q$, and not for $P=\mathcal{N}(0,Q)$ only.

4.5.2.3 Putting things together: building linear estimate

An immediate summary of Corollaries 4.5.1, 4.5.2 is the following recipe for building "presumably good" linear estimate:

Proposition 4.5.1 *In the situation of Section 4.5.1 and under Assumptions* **A′**, **R** *(see Section 4.5.1.1) consider the convex optimization problem (for notation, see* (4.5.8) *and* (4.5.12)*)*

$$\begin{array}{rcl}\mathrm{Opt}&=&\min\limits_{H,\Lambda,\Upsilon,\Upsilon',\Theta}\left\{\begin{array}{l}\phi_{\mathcal{T}}(\lambda[\Lambda])+\phi_{\mathcal{R}}(\lambda[\Upsilon])+\phi_{\mathcal{R}}(\lambda[\Upsilon'])+\Gamma(\Theta):\\ \Lambda=\{\Lambda_k\succeq 0,k\le K\},\ \Upsilon=\{\Upsilon_\ell\succeq 0,\ell\le L\},\ \Upsilon'=\{\Upsilon'_\ell\succeq 0,\ell\le L\},\\ \left[\begin{array}{c|c}\sum_k\mathcal{R}_k^*[\Lambda_k]&\frac{1}{2}[B^T-A^TH]M\\ \hline \frac{1}{2}M^T[B-H^TA]&\sum_\ell\mathcal{S}_\ell^*[\Upsilon_\ell]\end{array}\right]\succeq 0,\\ \left[\begin{array}{c|c}\Theta&\frac{1}{2}HM\\ \hline \frac{1}{2}M^TH^T&\sum_\ell\mathcal{S}_\ell^*[\Upsilon'_\ell]\end{array}\right]\succeq 0\end{array}\right\}\end{array} \tag{4.5.15}$$

The problem is solvable, and the H-component H_* of its optimal solution yields linear estimate $\widehat{x}_{H_*}(\omega)=H_*^T\omega$ such that

$$\mathrm{Risk}_{\Pi,\|\cdot\|}[\widehat{x}_{H_*}(\cdot)|\mathcal{X}]\leq \mathrm{Opt}. \tag{4.5.16}$$

Note that the only claim in Proposition 4.5.1 which is not an immediate consequence of Corollaries 4.5.1, 4.5.2 is that problem (4.5.15) is solvable; this claim is readily given by the fact that the objective clearly is coercive on the feasible set (recall that $\Gamma(\Theta)$ is coercive on $\mathbf{S}^m_+$ due to $\Pi\subset \mathrm{int}\,\mathbf{S}^m_+$ and that $y\mapsto My$ is an onto mapping, since $\mathcal{B}_*$ is full-dimensional).

4.5.3 Illustration: covariance matrix estimation

Suppose that we observe a sample

$$\eta^T=\{\eta_k=A\xi_k\}_{k\leq T} \tag{4.5.17}$$

where A is a given $m\times n$ matrix, and $\xi_1,...,\xi_T$ are sampled, independently of each other, from zero mean Gaussian distribution with unknown covariance matrix ϑ known to satisfy

$$\gamma\vartheta_*\preceq\vartheta\preceq\vartheta_*, \tag{4.5.18}$$

where $\gamma\geq 0$ and $\vartheta_*\succ 0$ are given. Our goal is to recover ϑ, and the norm on $\mathbf{S}^n$ in which recovery error is measured satisfies Assumption $\mathbf{A}'$.

Processing the problem. We can process the just outlined problem as follows.

1. We represent the set $\{\vartheta\in\mathbf{S}^n_+:\gamma\vartheta_*\preceq\vartheta\preceq\vartheta_*\}$ as the image of the matrix box

$$\mathcal{V}=\{v\in\mathbf{S}^n:\|v\|_{\mathrm{Sh},\infty}\leq 1\}\qquad [\|\cdot\|_{\mathrm{Sh},\infty}:\text{ spectral norm}]$$

under affine mapping, specifically, we set

$$\vartheta_0=\frac{1+\gamma}{2}\vartheta_*,\ \sigma=\frac{1-\gamma}{2}$$

and treat the matrix

$$v=\sigma^{-1}\vartheta_*^{-1/2}(\vartheta-\vartheta_0)\vartheta_*^{-1/2}\quad\left[\Leftrightarrow\vartheta=\vartheta_0+\sigma\vartheta_*^{1/2}v\vartheta_*^{1/2}\right]$$

as the signal underlying our observations. Note that our a priori information on ϑ reduces to $v\in\mathcal{V}$.

2. We pass from observations η_k to "lifted" observations $\eta_k\eta_k^T\in\mathbf{S}^m$, so that

$$\mathbf{E}\{\eta_k\eta_k^T\}=\mathbf{E}\{A\xi_k\xi_k^TA^T\}=A\vartheta A^T=A\underbrace{\left(\vartheta_0+\sigma A\vartheta_*^{1/2}v\vartheta_*^{1/2}\right)}_{\vartheta[v]}A^T,$$

and treat as "actual" observations the matrices

$$\omega_k=\eta_k\eta_k^T-A\vartheta_0A^T.$$

We have[16]

$$\omega_k=\mathcal{A}v+\zeta_k\text{ with }\mathcal{A}v=\sigma A\vartheta_*^{1/2}v\vartheta_*^{1/2}A^T\text{ and }\zeta_k=\eta_k\eta_k^T-A\vartheta[v]A^T. \tag{4.5.19}$$

Observe that random matrices $\zeta_1,...,\zeta_T$ are i.i.d. with zero mean and covariance mapping $\mathcal{Q}[v]$ (that of random matrix-valued variable $\zeta=\eta\eta^T-\mathbf{E}\{\eta\eta^T\}$, $\eta\sim\mathcal{N}(0,A\vartheta[v]A^T)$).

[16] In our current considerations, we need to operate with linear mappings acting from $\mathbf{S}^p$ to $\mathbf{S}^q$. We treat $\mathbf{S}^k$ as Euclidean space equipped with the Frobenius inner product $\langle u,v\rangle=\mathrm{Tr}(uv)$ and denote linear mappings from $\mathbf{S}^p$ into $\mathbf{S}^q$ by capital calligraphic letters, like $\mathcal{A}$, $\mathcal{Q}$, etc. Thus, $\mathcal{A}$ in (4.5.19) denotes the linear mapping which, on a closest inspection, maps matrix $v\in\mathbf{S}^n$ into the matrix $\mathcal{A}v=A[\vartheta[v]-\vartheta[0]]A^T$.

3. Let us $\succeq$-upper-bound the covariance mapping of ζ. Observe that $\mathcal{Q}[v]$ is a symmetric linear mapping of $\mathbf{S}^m$ into itself given by

$$\langle h,\mathcal{Q}[v]h\rangle=\mathbf{E}\{\langle h,\zeta\rangle^2\}=\mathbf{E}\{\langle h,\eta\eta^T\rangle^2\}-\langle h,\mathbf{E}\{\eta\eta^T\}\rangle^2,\ \ h\in\mathbf{S}^m.$$

Given $v\in\mathcal{V}$, let us set $\theta=\vartheta[v]$, so that $0\preceq\theta\preceq\theta_*$, and let $\mathcal{H}(h)=\theta^{1/2}A^ThA\theta^{1/2}$. We have

$$\begin{array}{rcl}\langle h,\mathcal{Q}[v]h\rangle&=&\mathbf{E}_{\xi\sim\mathcal{N}(0,\theta)}\{\mathrm{Tr}^2(hA\xi\xi^TA^T)\}-\mathrm{Tr}^2(h\mathbf{E}_{\xi\sim\mathcal{N}(0,\theta)}\{A\xi\xi^TA^T\})\\&=&\mathbf{E}_{\chi\sim\mathcal{N}(0,I_n)}\{\mathrm{Tr}^2(hA\theta^{1/2}\chi\chi^T\theta^{1/2}A^T))\}-\mathrm{Tr}^2(hA\theta A^T)\\&=&\mathbf{E}_{\chi\sim\mathcal{N}(0,I_n)}\{(\chi^T\mathcal{H}(h)\chi)^2\}-\mathrm{Tr}^2(\mathcal{H}(h)).\end{array}$$

We have $\mathcal{H}(h)=U\mathrm{Diag}\{\lambda\}U^T$ with orthogonal U, so that

$$\begin{array}{l}\mathbf{E}_{\chi\sim\mathcal{N}(0,I_n)}\{(\chi^T\mathcal{H}(h)\chi)^2\}-\mathrm{Tr}^2(\mathcal{H}(h))=\mathbf{E}_{\bar\chi:=U^T\chi\sim\mathcal{N}(0,I_n)}\{(\bar\chi^T\mathrm{Diag}\{\lambda\}\bar\chi)^2\}-(\sum_i\lambda_i)^2\\=\mathbf{E}_{\bar\chi\sim\mathcal{N}(0,I_n)}\{(\sum_i\lambda_i\bar\chi_i^2)^2\}-(\sum_i\lambda_i)^2=\sum_{i\neq j}\lambda_i\lambda_j+3\sum_i\lambda_i^2-(\sum_i\lambda_i)^2=2\sum_i\lambda_i^2=2\mathrm{Tr}([\mathcal{H}(h)]^2).\end{array}$$

Thus,

$$\begin{array}{rcl}\langle h,\mathcal{Q}[v]h\rangle&=&2\mathrm{Tr}([\mathcal{H}(h)]^2)=2\mathrm{Tr}(\theta^{1/2}A^ThA\theta A^ThA\theta^{1/2})\\&\leq&2\mathrm{Tr}(\theta^{1/2}A^ThA\theta_*A^ThA\theta^{1/2})\ \text{[since } 0\preceq\theta\preceq\theta_*]\\&=&2\mathrm{Tr}(\theta_*^{1/2}A^ThA\theta A^ThA\theta_*^{1/2})\leq2\mathrm{Tr}(\theta_*^{1/2}A^ThA\theta_*A^ThA\theta_*^{1/2})\\&=&2\mathrm{Tr}(\theta_*A^ThA\theta_*A^ThA).\end{array}$$

We conclude that

$$\forall v\in\mathcal{V}:\mathcal{Q}[v]\preceq\mathcal{Q},\ \langle e,\mathcal{Q}h\rangle=2\mathrm{Tr}(\vartheta_*A^ThA\vartheta_*A^TeA),\ e,h\in\mathbf{S}^m.\tag{4.5.20}$$

4. To continue, we need to set some additional notation to be used when operating with Euclidean spaces $\mathbf{S}^p$, $p=1,2,...$

- We denote $\bar p=\frac{p(p+1)}{2}=\dim\mathbf{S}^p$, $\mathcal{I}_p=\{(i,j):1\leq i\leq j\leq p\}$, and for $(i,j)\in\mathcal{I}_p$ set

  $$e_p^{ij}=\left\{\begin{array}{ll}e_ie_i^T, & i=j\\ \frac{1}{\sqrt{2}}[e_ie_j^T+e_je_i^T], & i<j\end{array}\right.,$$

  where e_i are the standard basic orths in $\mathbf{R}^p$. Note that $\{e_p^{ij}:(i,j)\in\mathcal{I}_p\}$ is the standard orthonormal basis in $\mathbf{S}^p$. Given $v\in\mathbf{S}^p$, we denote by $\mathrm{X}^p(v)$ the vector of coordinates of v in this basis:

  $$\mathrm{X}^p_{ij}(v)=\mathrm{Tr}(ve_p^{ij})=\left\{\begin{array}{ll}v_{ii}, & i=j\\ \sqrt{2}v_{ij}, & i<j\end{array}\right.,\ (i,j)\in\mathcal{I}_p.$$

  Similarly, for $x\in\mathbf{R}^{\bar p}$, we index the entries in x by pairs ij, $(i,j)\in\mathcal{I}_p$, and set $\mathrm{V}^p(x)=\sum_{(i,j)\in\mathcal{I}_p}x_{ij}e_p^{ij}$, so that $v\mapsto\mathrm{X}^p(v)$ and $x\mapsto\mathrm{V}^p(x)$ are inverse to each other linear norm-preserving maps identifying the Euclidean spaces $\mathbf{S}^p$ and $\mathbf{R}^{\bar p}$ (recall that the inner products on these spaces are, respectively, the Frobenius and the standard one).

- Recall that $\mathcal{V}$ is the matrix box $\{v\in\mathbf{S}^n:v^2\preceq I_n\}=\{v\in\mathbf{S}^n:\exists t\in\mathcal{T}:=[0,1]:v^2\preceq tI_n\}$. We denote by $\mathcal{X}$ the image of $\mathcal{V}$ under the mapping X^n:

  $$\mathcal{X}=\{x\in\mathbf{R}^{\bar n}:\exists t\in\mathcal{T}:R^2[x]\preceq tI_n\},\ R[x]=\sum_{(i,j)\in\mathcal{I}_n}x_{ij}e_n^{ij},\ \ \bar n=\frac{1}{2}n(n+1).$$

  Note that $\mathcal{X}$ is a basic spectratope of size n.

Now we can assume that the signal underlying our observations is $x \in \mathcal{X}$, and the observations themselves are

$$w_k = \mathrm{X}^m(\omega_k) = \underbrace{\mathrm{X}^m(\mathcal{A}\mathrm{V}^n(x))}_{=:\overline{A}x} + z_k,\;\; z_k = \mathrm{X}^m(\zeta_k).$$

Note that $z_k \in \mathbf{R}^{\bar m}$, $1 \le k \le T$, are zero mean i.i.d. random vectors with covariance matrix $Q[x]$ satisfying, in view of (4.5.20), the relation

$$Q[x] \preceq Q, \text{ where } Q_{ij,k\ell} = 2\mathrm{Tr}(\vartheta_* A^T e_m^{ij} A \vartheta_* A^T e_m^{k\ell} A),\ (i,j) \in \mathcal{I}_m, (k,\ell) \in \mathcal{I}_m.$$

Our goal is to estimate $\vartheta[v] - \vartheta[0]$, or, what is the same, to recover

$$\overline{B}x := \mathrm{X}^n(\vartheta[\mathrm{V}^n(x)] - \vartheta[0]).$$

We assume that the norm in which the estimation error is measured is “transferred” from $\mathbf{S}^n$ to $\mathbf{R}^{\bar n}$; we denote the resulting norm on $\mathbf{R}^{\bar n}$ by $\|\cdot\|$ and assume that the unit ball $\mathcal{B}_*$ of the conjugate norm $\|\cdot\|_*$ is given by spectratopic representation:

$$\begin{array}{c}\{u \in \mathbf{R}^{\bar n} : \|u\|_* \le 1\} = \{u \in \mathbf{R}^{\bar n} : \exists y \in \mathcal{Y} : u = My\},\\ \mathcal{Y} := \{y \in \mathbf{R}^q : \exists r \in \mathcal{R} : S_\ell^2[y] \preceq r_\ell I_{f_\ell},\ 1 \le \ell \le L\}.\end{array} \tag{4.5.21}$$

The formulated description of the estimation problem fit the premises of Proposition 4.5.1, specifically:

- the signal x underlying our observation $w^{(T)} = [w_1; ...; w_T]$ is known to belong to basic spectratope $\mathcal{X} \in \mathbf{R}^{\bar n}$, and the observation itself is of the form
$$w^{(T)} = \overline{A}^{(T)}x + z^{(T)},\ \overline{A}^{(T)} = \underbrace{[\overline{A}; ...; \overline{A}]}_{T},\ z^{(T)} = [z_1; ...; z_T];$$
- the noise $z^{(T)}$ is zero mean, and its covariance matrix is $\preceq Q_T := \mathrm{Diag}\{\underbrace{Q, ..., Q}_{T}\}$, which allows to set $\Pi = \{Q_T\}$;
- our goal is to recover $\overline{B}x$, and the norm $\|\cdot\|$ in which the recovery error is measured satisfies (4.5.21).

Proposition 4.5.1 supplies the linear estimate

$$\widehat{x}(w^{(T)}) = \sum_{k=1}^T H_{*k}^T w_k,$$

of $\overline{B}x$ with $H_* = [H_{*1}; ...; H_{*T}]$ stemming from the optimal solution to the convex optimization problem

$$\begin{array}{rcl}\mathrm{Opt} &=& \min\limits_{H=[H_1;...;H_T],\Lambda,\Upsilon} \left\{ \mathrm{Tr}(\Lambda) + \phi_{\mathcal{R}}(\lambda[\Upsilon]) + \overline{\Psi}_{\{Q_T\}}(H_1, ..., H_T) : \right. \\ && \left. \begin{array}{r}\Lambda \in \mathbf{S}^n_+, \Upsilon = \{\Upsilon_\ell \succeq 0, \ell \le L\},\\ \left[\begin{array}{c|c}\mathcal{R}^*[\Lambda] & \frac{1}{2}[\overline{B}^T - \overline{A}^T\sum_k H_k]M \\ \hline \frac{1}{2}M^T[\overline{B} - [\sum_k H_k]^T\overline{A}] & \sum_\ell \mathcal{S}_\ell^*[\Upsilon_\ell]\end{array}\right] \succeq 0\end{array} \right\},\end{array} \tag{4.5.22}$$

where

$$\mathcal{R}^*[\Lambda] \in \mathbf{S}^{\bar n} : \ (\mathcal{R}^*[\Lambda])_{ij,k\ell} = \mathrm{Tr}(\Lambda e_n^{ij} e_n^{k\ell}),\ (i,j) \in \mathcal{I}_n,\ (k,\ell) \in \mathcal{I}_n,$$

and, cf. (4.5.13),

$$\begin{array}{l}\overline{\Psi}_{\{Q_T\}}(H_1, ..., H_T) = \min\limits_{\Upsilon',\Theta} \left\{ \mathrm{Tr}(Q_T\Theta) + \phi_{\mathcal{R}}(\lambda[\Upsilon']) : \ \Theta \in \mathbf{S}^{mT},\ \Upsilon' = \{\Upsilon'_\ell \succeq 0, \ell \le L\}, \right. \\ \qquad\qquad \left. \left[\begin{array}{c|c}\Theta & \frac{1}{2}[H_1M; ...; H_TM] \\ \hline \frac{1}{2}[M^TH_1^T, ..., M^TH_T^T] & \sum_\ell \mathcal{S}_\ell^*[\Upsilon'_\ell]\end{array}\right] \succeq 0 \right\},\end{array}$$

5. Evidently, the function $\overline{\Psi}_{\{Q_T\}}([H_1,...,H_T])$ remains intact when permuting $H_1,...,H_T$; with this in mind, it is clear that permuting $H_1,...,H_T$ and keeping intact Λ and Υ is a symmetry of (4.5.22) – such a transformation maps feasible set onto itself and preserves the value of the objective. Since (4.5.22) is convex and solvable, it follows that there exists an optimal solution to the problem with $H_1 = ... = H_T = H$. On the other hand,

$$\begin{array}{l}
\overline{\Psi}_{\{Q_T\}}(H,...,H)\\
= \min\limits_{\Upsilon',\Theta}\left\{\mathrm{Tr}(Q_T\Theta)+\phi_{\mathcal{R}}(\lambda[\Upsilon'])：\ \Theta\in\mathbf{S}^{mT},\ \Upsilon'=\{\Upsilon'_\ell\succeq 0,\ell\leq L\}\right.\\
\hfill\left.\left[\begin{array}{c|c}\Theta & \frac{1}{2}[HM;...;HM]\\ \hline \frac{1}{2}[M^TH^T,...,M^TH^T] & \sum_\ell\mathcal{S}^*_\ell[\Upsilon'_\ell]\end{array}\right]\succeq 0\right\},\\
= \inf\limits_{\Upsilon',\Theta}\left\{\mathrm{Tr}(Q_T\Theta)+\phi_{\mathcal{R}}(\lambda[\Upsilon'])：\ \Theta\in\mathbf{S}^{mT},\ \Upsilon'=\{\Upsilon'_\ell\succ 0,\ell\leq L\},\right.\\
\hfill\left.\left[\begin{array}{c|c}\Theta & \frac{1}{2}[HM;...;HM]\\ \hline \frac{1}{2}[M^TH^T,...,M^TH^T] & \sum_\ell\mathcal{S}^*_\ell[\Upsilon'_\ell]\end{array}\right]\succeq 0\right\}\\
= \inf\limits_{\Upsilon',\Theta}\left\{\mathrm{Tr}(Q_T\Theta)+\phi_{\mathcal{R}}(\lambda[\Upsilon'])：\ \Theta\in\mathbf{S}^{mT},\ \Upsilon'=\{\Upsilon'_\ell\succ 0,\ell\leq L\},\right.\\
\hfill\left.\Theta\succeq\frac{1}{4}[HM;...;HM]\left[\sum_\ell\mathcal{S}^*_\ell[\Upsilon'_\ell]\right]^{-1}[HM;...;HM]^T\right\}\\
= \inf\limits_{\Upsilon'}\left\{\phi_{\mathcal{R}}(\lambda[\Upsilon'])+\frac{T}{4}\mathrm{Tr}\left(QHM\left[\sum_\ell\mathcal{S}^*_\ell[\Upsilon'_\ell]\right]^{-1}M^TH^T\right)：\Upsilon'=\{\Upsilon'_\ell\succ 0,\ell\leq L\}\right\}\\
\hfill[\text{due to } Q_T=\mathrm{Diag}\{Q,...,Q\}]\\
= \min\limits_{\Upsilon',G}\left\{T\mathrm{Tr}(QG)+\phi_{\mathcal{R}}(\lambda[\Upsilon'])：\ G\in\mathbf{S}^m,\ \Upsilon'=\{\Upsilon'_\ell\succeq 0,\ell\leq L\},\left[\begin{array}{c|c}G & \frac{1}{2}HM\\ \hline \frac{1}{2}M^TH^T & \sum_\ell\mathcal{S}^*_\ell[\Upsilon'_\ell]\end{array}\right]\succeq 0\right\}
\end{array}\tag{4.5.23}$$

(we have used Schur Complement Lemma combined with the fact that $\sum_\ell\mathcal{S}^*_\ell[\Upsilon'_\ell]\succ 0$ whenever $\Upsilon'_\ell\succ 0$ for all ℓ, see Lemma 4.7.2).

In view of the above observations, when replacing variables H and G with $\overline{H}=TH$ and $\overline{G}=T^2G$, respectively, problem (4.5.22), (4.5.23) becomes

$$\begin{array}{rcl}
\mathrm{Opt} & = & \min\limits_{\overline{H},\overline{G},\Lambda,\Upsilon,\Upsilon'}\left\{\mathrm{Tr}(\Lambda)+\phi_{\mathcal{R}}(\lambda[\Upsilon])+\phi_{\mathcal{R}}(\lambda[\Upsilon'])+\frac{1}{T}\mathrm{Tr}(Q\overline{G})：\right.\\
&&\left.\begin{array}{r}
\Lambda\in\mathbf{S}^n_+,\Upsilon=\{\Upsilon_\ell\succeq 0,\ell\leq L\},\Upsilon'=\{\Upsilon'_\ell\succeq 0,\ell\leq L\},\\
\left[\begin{array}{c|c}\mathcal{R}^*[\Lambda] & \frac{1}{2}[\overline{B}^T-\overline{A}^T\overline{H}]M\\ \hline \frac{1}{2}M^T[\overline{B}-\overline{H}^T\overline{A}] & \sum_\ell\mathcal{S}^*_\ell[\Upsilon_\ell]\end{array}\right]\succeq 0,\\
\left[\begin{array}{c|c}\overline{G} & \frac{1}{2}\overline{H}M\\ \hline \frac{1}{2}M^T\overline{H}^T & \sum_\ell\mathcal{S}^*_\ell[\Upsilon'_\ell]\end{array}\right]\succeq 0
\end{array}\right\},
\end{array}\tag{4.5.24}$$

and the estimate

$$\widehat{x}(w^T)=\frac{1}{T}\overline{H}^T\sum_{k=1}^T w_k$$

stemming from an optimal solution to (4.5.24) satisfies

$$\mathrm{Risk}_{\Pi,\|\cdot\|}[\widehat{x}|\mathcal{X}]\leq\mathrm{Opt},$$

where $\Pi=\{Q_T\}$.

4.5.4 Estimation from repeated observations

Consider the special case of the situation from Section 4.5.1, the case where observation ω in (4.5.2) is T-element sample: $\omega=[\bar{\omega}_1;...;\bar{\omega}_T]$ with components

$$\bar{\omega}_t=\bar{A}x+\xi_t,\ t=1,...,T$$

and ξ_t are i.i.d. observation noises with *zero mean* distribution $\bar{P}$ satisfying $\bar{P} \lll \bar{\Pi}$ for some convex compact set $\bar{\Pi} \subset \operatorname{int} \mathbf{S}^{\bar{m}}_+$. In other words, we are in the situation where

$$\begin{array}{l}A = [\underbrace{\bar{A}; ...; \bar{A}}_{T}] \in \mathbf{R}^{m\times n} \text{ for some } \bar{A} \in \mathbf{R}^{\bar{m}\times n} \text{ and } m = T\bar{m},\\ \Pi = \{Q = \mathrm{Diag}\{\underbrace{\bar{Q}, ..., \bar{Q}}_{T}\}, \bar{Q} \in \bar{\Pi}\}\end{array}$$

The same argument as used in item 5 of Section 4.5.3 justifies the following

Proposition 4.5.2 *In the situation in question and under Assumption* $\mathbf{A}'$*, the linear estimate of* Bx *yielded by an optimal solution to problem* (4.5.15) *can be found as follows. We consider the convex optimization problem*

$$\begin{array}{rcl}\overline{\mathrm{Opt}} &=& \min\limits_{\bar{H},\Lambda,\Upsilon,\Upsilon',\bar{\Theta}} \left\{ \phi_{\mathcal{T}}(\lambda[\Lambda]) + \phi_{\mathcal{R}}(\lambda[\Upsilon]) + \phi_{\mathcal{R}}(\lambda[\Upsilon']) + \frac{1}{T}\overline{\Gamma}(\bar{\Theta}) : \right.\\ && \Lambda = \{\Lambda_k \succeq 0, k \leq K\},\ \Upsilon = \{\Upsilon_\ell \succeq 0, \ell \leq L\},\ \Upsilon' = \{\Upsilon'_\ell \succeq 0, \ell \leq L\},\\ && \left.\begin{array}{r}\left[\begin{array}{c|c}\sum_k \mathcal{R}^*_k[\Lambda_k] & \frac{1}{2}[B^T - A^T\bar{H}]M\\ \hline \frac{1}{2}M^T[B - \bar{H}^T A] & \sum_\ell \mathcal{S}^*_\ell[\Upsilon_\ell]\end{array}\right] \succeq 0,\\ \left[\begin{array}{c|c}\bar{\Theta} & \frac{1}{2}\bar{H}M\\ \hline \frac{1}{2}M^T\bar{H}^T & \sum_\ell \mathcal{S}^*_\ell[\Upsilon'_\ell]\end{array}\right] \succeq 0\end{array}\right\}\end{array} \tag{4.5.25}$$

where

$$\overline{\Gamma}(\bar{\Theta}) = \max_{\bar{Q}\in\bar{\Pi}} \mathrm{Tr}(\bar{Q}\bar{\Theta}).$$

The problem is solvable, and the estimate in question is yielded by the $\bar{H}$*-component* $\bar{H}_*$ *of the optimal solution according to*

$$\widehat{x}([\bar{\omega}_1; ...; \bar{\omega}_T]) = \frac{1}{T}\bar{H}_*^T \sum_{t=1}^{T} \bar{\omega}_t.$$

The provided by Proposition 4.5.1 upper bound on the risk $\mathrm{Risk}_{\Pi,\|\cdot\|}[\widehat{x}(\cdot)|\mathcal{X}]$ *of this estimate is* $\overline{\mathrm{Opt}}$*.*

The advantage of this result as compared to what is stated under the circumstances by Proposition 4.5.1 is that the sizes of optimization problem (4.5.25) are independent of T.

4.5.5 Near-optimality in Gaussian case

The risk of the linear estimate $\widehat{x}_{H_*}(\cdot)$ constructed in (4.5.15), (4.5.16) can be compared to the minimax optimal risk of recovering Bx, $x \in \mathcal{X}$, from observations corrupted by zero mean Gaussian noise with covariance matrix from Π; formally, this minimax optimal risk is defined as

$$\mathrm{RiskOpt}_{\Pi,\|\cdot\|}[\mathcal{X}] = \sup_{Q\in\Pi} \inf_{\widehat{x}(\cdot)} \left[\sup_{x\in\mathcal{X}} \mathbf{E}_{\xi\sim\mathcal{N}(0,Q)}\{\|Bx - \widehat{x}(Ax+\xi)\|\}\right] \tag{4.5.26}$$

where the infimum is taken over all estimates.

Proposition 4.5.3 *Under the premise and in the notation of Proposition 4.5.1, let*

$$\begin{array}{rcl}M_*^2 &=& \max\limits_{W} \left\{\mathbf{E}_{\eta\sim\mathcal{N}(0,I_n)}\{\|BW^{1/2}\eta\|^2\} :\right.\\ && \left. W \in \mathcal{Q} := \{W \in \mathbf{S}^n_+ : \exists t \in \mathcal{T} : \mathcal{R}_k[W] \preceq t_k I_{d_k}, 1 \leq k \leq K\}\right\};\end{array} \tag{4.5.27}$$

we have

$$\mathrm{Risk}_{\Pi,\|\cdot\|}[\widehat{x}_{H_*}|\mathcal{X}] \leq \mathrm{Opt} \leq C\sqrt{\ln(2F)\ln\left(\frac{2DM_*^2}{\mathrm{RiskOpt}^2_{\Pi,\|\cdot\|}[\mathcal{X}]}\right)}\mathrm{RiskOpt}_{\Pi,\|\cdot\|}[\mathcal{X}], \tag{4.5.28}$$

where C is a positive absolute constant, and

$$D = \sum_k d_k,\ \ F = \sum_\ell f_\ell. \tag{4.5.29}$$

For the proof, see Section 4.7.4.2. The key component of the proof is the following important by its own right fact (for proof, see Section 4.7.4.1):

Lemma 4.5.2 *Let Y be an $N \times \nu$ matrix, let $\|\cdot\|$ be a norm on $\mathbf{R}^\nu$ such that the unit ball $\mathcal{B}_*$ of the conjugate norm is the spectratope* (4.5.5)*, and let $\zeta \sim \mathcal{N}(0, Q)$ for some positive semidefinite $N \times N$ matrix Q. Then the best upper bound on $\psi_Q(Y) := \mathbf{E}\{\|Y^T\zeta\|\}$ yielded by Lemma 4.5.1, that is, the optimal value* $\mathrm{Opt}[Q]$ *in the convex optimization problem (cf.* (4.5.13)*)*

$$\begin{array}{rcl}\mathrm{Opt}[Q] & = & \min\limits_{\Theta,\Upsilon}\left\{\phi_{\mathcal{R}}(\lambda[\Upsilon]) + \mathrm{Tr}(Q\Theta) : \Upsilon = \{\Upsilon_\ell \succeq 0, 1 \leq \ell \leq L\},\, \Theta \in \mathbf{S}^N,\right.\\ && \left.\left[\begin{array}{c|c}\Theta & \frac{1}{2}YM\\ \hline \frac{1}{2}M^TY^T & \sum_\ell \mathcal{S}_\ell^*[\Upsilon\ell]\end{array}\right] \succeq 0\right\}\end{array} \tag{4.5.30}$$

(for notation, see Lemma 4.5.1 and (4.5.8)*) satisfies the identity*

$$\begin{array}{l}\forall (Q \succeq 0) :\\ \mathrm{Opt}[Q] = \overline{\mathrm{Opt}}[Q] := \min\limits_{G,\Upsilon=\{\Upsilon_\ell,\ell\leq L\}}\left\{\phi_{\mathcal{R}}(\lambda[\Upsilon]) + \mathrm{Tr}(G) : \Upsilon_\ell \succeq 0, \left[\begin{array}{c|c}G & \frac{1}{2}Q^{1/2}YM\\ \hline \frac{1}{2}M^TY^TQ^{1/2} & \sum_\ell \mathcal{S}_\ell^*[\Upsilon_\ell]\end{array}\right] \succeq 0\right\},\end{array} \tag{4.5.31}$$

and is a tight bound on $\psi_Q(Y)$, namely,

$$\psi_Q(Y) \leq \mathrm{Opt}[Q] \leq \frac{8\sqrt{\ln\left(\frac{\sqrt{2}F}{\sqrt{2}-\mathrm{e}^{1/4}}\right)}}{\sqrt{2}-\mathrm{e}^{1/4}}\psi_Q(Y) \leq 62\sqrt{\ln(44F)}\psi_Q(Y), \tag{4.5.32}$$

where $F = \sum_\ell f_\ell$ is the size of the spectratope (4.5.5)*.*

4.6 Linear estimation in the case of uncertain-but-bounded noise

So far, the main subject of our interest was recovering (linear images of) signals via indirect observations of these signals corrupted by random noise. In this section, we focus on alternative observation schemes – those with "uncertain-but-bounded" and with "mixed" noise.

4.6.1 Uncertain-but-bounded noise

Consider recovering problem where one, given observation

$$\omega = Ax + \eta \tag{4.6.1}$$

of unknown signal x known to belong to a given signal set $\mathcal{X}$, wants to recover linear image Bx of x. Here A and B are given $m \times n$ and $\nu \times n$ matrices. The situation looks exactly as before; the difference with our previous considerations is that now we do not assume the observation noise to be random; all we assume about η is that it belongs to a given compact set $\mathcal{H}$ ("uncertain-but-bounded observation noise"). In the situation in question, a natural definition of the risk on $\mathcal{X}$ of a candidate estimate $\omega \mapsto \widehat{x}(\omega)$ is

$$\mathrm{Risk}_\sigma[\widehat{x}|\mathcal{X}] = \sup_{x\in X,\eta\in\mathcal{H}} \|Bx - \widehat{x}(Ax+\eta)\| \tag{4.6.2}$$

("$\mathcal{H}$-risk").

We are about to prove that when $\mathcal{X}$ and $\mathcal{H}$ are spectratopes, and the unit ball of the norm $\|\cdot\|_*$ conjugate to $\|\cdot\|$ is a basic spectratope, an efficiently computable linear estimate is near-optimal in terms of its $\mathcal{H}$-risk.

Our initial observation is that the situation in question reduces straightforwardly to the one where there is no observation noise at all. Indeed, let $\mathcal{Y} = \mathcal{X} \times \mathcal{H}$; then $\mathcal{Y}$ is a spectratope, and we lose nothing when assuming that the signal underlying observation ω is $y = [x; \eta] \in \mathcal{Y}$:

$$\omega = Ax + \eta = \bar{A}y,\ \bar{A} = [A, I_m],$$

while the entity to be recovered is

$$Bx = \bar{B}y,\ \bar{B} = [B, 0_{\nu\times m}].$$

With these conventions, the $\mathcal{H}$-risk of a candidate estimate $\widehat{x}(\cdot) : \mathbf{R}^m \to \mathbf{R}^\nu$ becomes the quantity

$$\mathrm{Risk}_{\|\cdot\|}[\widehat{x}|\mathcal{X}\times\mathcal{H}] = \sup_{y=[x;\eta]\in\mathcal{X}\times\mathcal{H}} \|\bar{B}y - \widehat{x}(\bar{A}y)\|,$$

that is, we indeed arrive at the situation where the observation noise is identically zero.

To avoid messy notation, let us assume that the outlined reduction has been carried out in advance, so that

The problem of interest is to recover the linear image $Bx \in \mathbf{R}^\nu$ of an unknown signal x known to belong to a given spectratope $\mathcal{X}$ from noiseless observation

$$\omega = Ax \in \mathbf{R}^m,$$

and the risk of a candidate estimate is defined as

$$\mathrm{Risk}_{\|\cdot\|}[\widehat{x}|\mathcal{X}] = \sup_{x\in\mathcal{X}} \|Bx - \widehat{x}(Ax)\|,$$

where $\|\cdot\|$ is a given norm with a basic spectratope as the unit ball B_* of the conjugate norm. By our standard argument,

We lose nothing when assuming that the spectratope $\mathcal{X}$ is basic as well, so that

$$\begin{array}{rcl} \mathcal{X} &=& \{x \in \mathbf{R}^n : \exists t \in \mathcal{T} : R_k^2[x] \preceq t_k I_{d_k}, k \le K\}, \\ \mathcal{B}_* &:=& \{u \in \mathbf{R}^\nu : \|u\|_* \le 1\} = \{u \in \mathbf{R}^\nu : \exists r \in \mathcal{R} : S_\ell^2[u] \preceq r_\ell I_{f_\ell}, \ell \le L\} \end{array} \tag{4.6.3}$$

with the standard restrictions on $\mathcal{T}, \mathcal{R}$ and $R_k[\cdot]$, $S_\ell[\cdot]$.

4.6.1.1 Building linear estimate

Let us build a seemingly good linear estimate. For a linear estimate $\widehat{x}_H(\omega) = H^T\omega$, we have

$$\begin{array}{rcl} \mathrm{Risk}_{\|\cdot\|}[\widehat{x}_H|\mathcal{X}] &=& \max\limits_{x\in\mathcal{X}} \|(B - H^TA)x\| \\ &=& \max\limits_{[u;x]\in\mathcal{B}_*\times\mathcal{X}} [u;x]^T \left[\begin{array}{c|c} & \frac{1}{2}(B - H^TA) \\ \hline \frac{1}{2}(B - H^TA)^T & \end{array}\right][u;x]. \end{array}$$

Applying Proposition 4.4.3, we arrive at the following

Proposition 4.6.1 *In the situation of this section, consider the convex optimization problem*

$$\mathrm{Opt}_\# = \min_{H,\Upsilon=\{\Upsilon_\ell\},\Lambda=\{\Lambda_k\}} \left\{ \phi_{\mathcal{R}}(\lambda[\Upsilon]) + \phi_{\mathcal{T}}(\lambda[\Lambda]) : \Upsilon_\ell \succeq 0,\ \Lambda_k \succeq 0,\ \forall(\ell,k) \atop \left[\begin{array}{c|c} \sum_\ell \mathcal{S}_\ell^*[\Upsilon_\ell] & \frac{1}{2}(B-H^TA) \\ \hline \frac{1}{2}(B-H^TA)^T & \sum_k \mathcal{R}_k^*[\Lambda_k] \end{array}\right] \succeq 0 \right\}, \tag{4.6.4}$$

where $\mathcal{R}_k^*[\cdot]$, $\mathcal{S}_\ell^*[\cdot]$ are induced by $R_k[\cdot]$, $S_\ell[\cdot]$, respectively, as explained in Section 4.4.1. The problem is solvable, and the risk of the linear estimate $\widehat{x}_{H_*}(\cdot)$ yielded by the H-component of an optimal solution does not exceed $\mathrm{Opt}_\#$.

For proof, see Section 4.7.5.1.

4.6.1.2 Near-optimality

Proposition 4.6.2 *The linear estimate $\widehat{x}_{H_*}$ yielded by Proposition 4.6.1 is near-optimal in terms of its risk:*

$$\mathrm{Risk}_{\|\cdot\|}[\widehat{x}_{H_*}|\mathcal{X}] \leq \mathrm{Opt}_\# \leq 2\ln(2D)\mathrm{Risk}_{\rm opt}[\mathcal{X}], \quad D = \sum_k d_k + \sum_\ell f_\ell, \tag{4.6.5}$$

where $\mathrm{Risk}_{\rm opt}[\mathcal{X}]$ *is the minimax optimal risk:*

$$\mathrm{Risk}_{\rm opt}[\mathcal{X}] = \inf_{\widehat{x}} \mathrm{Risk}_{\|\cdot\|}[\widehat{x}|\mathcal{X}],$$

where inf *is taken w.r.t. all possible estimates.*

Remark 4.6.1 *When $\mathcal{X}$ and $\mathcal{B}_*$ are basic ellitopes rather than basic spectratopes:*

$$\begin{array}{rcl} \mathcal{X} &=& \{x \in \mathbf{R}^n : \exists t \in \mathcal{T} : x^T R_k x \leq t_k, k \leq K\}, \\ \mathcal{B}_* &:=& \{u \in \mathbf{R}^\nu : \|u\|_* \leq 1\} = \{u \in \mathbf{R}^\nu : \exists r \in \mathcal{R} : u^T S_\ell u \leq r_\ell, \ell \leq L\} \\ && [R_k \succeq 0, \sum_k R_k \succ 0, S_\ell \succeq 0, \sum_\ell S_\ell \succ 0] \end{array} \tag{4.6.6}$$

(4.6.5) *can be strengthened to*

$$\mathrm{Risk}_{\|\cdot\|}[\widehat{x}_{H_*}|\mathcal{X}] \leq \mathrm{Opt}_\# \leq 4\ln(5[K+L])\mathrm{Risk}_{\rm opt}[\mathcal{X}]. \tag{4.6.7}$$

For proof, see Section 4.7.5.2.

4.6.1.3 Nonlinear estimation

Uncertain-but-bounded model of observation error makes it easy to point out an efficiently computable near-optimal *nonlinear* estimate. Specifically, in the situation described in the beginning of Section 4.6.1, assume that the range of observation error η is

$$\mathcal{H} = \{\eta \in \mathbf{R}^m : \|\eta\|_{(m)} \leq \sigma\}, \tag{4.6.8}$$

where $\|\cdot\|_{(m)}$, $\sigma > 0$ are a given norm on $\mathbf{R}^m$ and a given error bound, and let us measure the recovery error by a given norm $\|\cdot\|_{(\nu)}$ on $\mathbf{R}^\nu$. We can immediately point out a (nonlinear) estimate optimal, in terms of its $\mathcal{H}$-risk, within factor 2, specifically, the estimate $\widehat{x}_*$ defined as follows:

Given ω, we solve the feasibility problem

$$\text{find } x \in \mathcal{X} : \|Ax - \omega\|_{(m)} \leq \sigma \tag{$F[\omega]$}$$

find a feasible solution x_ω to the problem, and set $\widehat{x}_*(\omega) = Bx_\omega$.

Note that the estimate is well defined, since ($F[\omega]$) clearly is solvable, with one of the feasible solutions being the true signal underlying observation ω. When $\mathcal{X}$ is a computationally tractable convex compact set, and $\|\cdot\|_{(m)}$ is an efficiently computable norm, a feasible solution to ($F[\omega]$) can be found in a computationally efficient fashion. Let us make the following immediate observation:

Proposition 4.6.3 *The estimate $\widehat{x}_*$ is optimal within factor 2:*

$$\mathrm{Risk}_{\mathcal{H}}[\widehat{x}_*|\mathcal{X}] \leq \mathrm{Opt}_\# := \sup_{x,y}\left\{\|Bx - By\|_{(\nu)} : x, y \in \mathcal{X}, \|A(x-y)\|_{(m)} \leq 2\sigma\right\} \leq 2\mathrm{Risk}_{\mathrm{opt},\mathcal{H}}, \tag{4.6.9}$$

where $\mathrm{Risk}_{\mathrm{opt},\mathcal{H}}$ is the infimum, over all estimates, of the estimate's $\mathcal{H}$-risk.

4.6.2 Recovery under uncertain-but-bounded noise

The proof of Proposition is the subject of Exercise 4.6. Note that Proposition does not impose restrictions on $\mathcal{X}$ and the norms $\|\cdot\|_{(m)}$, $\|\cdot\|_{(\nu)}$.

The only - but essential – shortcoming of the estimate $\widehat{x}_*$ is that we do not know, in general, what is its $\mathcal{H}$-risk. From (4.6.9) it follows that this risk is tightly (namely, within factor 2) upper-bounded by Υ, but this quantity, being the maximum of a convex function over some domain, can be difficult to compute. Aside from handful of special cases where this difficulty does not arise, there is a generic situation when Υ can be tightly upper-bounded by efficient computation. This is the situation where $\mathcal{X}$ is the spectratope defined in (4.6.3), $\|\cdot\|_{(m)}$ is such that the unit ball of this norm is a basic spectratope:

$$B_{(m)} := \{u : \|u\|_{(m)} \leq 1\} = \{u \in \mathbf{R}^m : \exists p \in \mathcal{P} : Q_j^2[u] \preceq p_j I_{e_j}, 1 \leq j \leq J\},$$

and the unit ball of the norm $\|\cdot\|_{(\nu),*}$ *conjugate* to the norm $\|\cdot\|_{(\nu)}$ is a spectratope:

$$B^*_{(\nu)} := \{v \in \mathbf{R}^\nu : \|v\|_{(\nu),*} \leq 1\} = \{v : \exists (w \in \mathbf{R}^N, r \in \mathcal{R}) : v = Mw, S_\ell^2[w] \preceq r_\ell I_{f_\ell}, 1 \leq \ell \leq L\},$$

with our usual restrictions on $\mathcal{P}$, $\mathcal{R}$, $Q_j[\cdot]$, $S_\ell[\cdot]$.

Proposition 4.6.4 *In the situation in question, consider convex optimization problem*

$$\begin{array}{rcl}\mathrm{Opt} & = & \min\limits_{\substack{\Lambda=\{\Lambda_k,k\leq K\},\\ \Upsilon=\{\Upsilon_\ell,\ell\leq L\},\\ \Sigma=\{\Sigma_j,j\leq J\}}} \left\{\phi_{\mathcal{T}}(\lambda[\Lambda]) + \phi_{\mathcal{R}}(\lambda[\Upsilon]) + \sigma^2\phi_{\mathcal{P}}(\lambda[\Sigma]) + \phi_{\mathcal{R}}(\lambda([\Sigma]) : \right.\\ & & \qquad \Lambda_k \succeq 0, \Upsilon_\ell \succeq 0, \Sigma_j \succeq 0 \,\forall(k,\ell,j),\\ & & \qquad \left.\left[\begin{array}{c|c}\sum_\ell \mathcal{S}_\ell^*[\Upsilon_\ell] & M^TB\\ \hline B^TM & \sum_k \mathcal{R}_k^*[\Lambda_k] + A^T[\sum_j \mathcal{Q}_j^*[\Sigma_j]]A\end{array}\right] \succeq 0 \right\}\end{array} \tag{4.6.10}$$

where $\mathcal{R}_k^*[\cdot]$ are associated with the mappings $x \mapsto R_k[x]$ according to (4.4.7), and $\mathcal{S}_\ell^*[\cdot]$ and $\mathcal{Q}_j^*[\cdot]$ are associated in the same fashion with the mappings $w \mapsto S_\ell[w]$ and $u \mapsto Q_j[u]$, respectively, and $\phi_{\mathcal{T}}$, $\phi_{\mathcal{R}}$, $\phi_{\mathcal{P}}$ are the support functions of the corresponding sets $\mathcal{T}$, $\mathcal{R}$, $\mathcal{P}$.

The optimal value in (4.6.10) is an efficiently computable upper bound on the quantity $\mathrm{Opt}_\#$ defined in (4.6.9), and this bound is tight within factor

$$2\max[\ln(2D), 1], \ \ D = \sum_k d_k + \sum_\ell f_\ell + \sum_j e_j. \tag{4.6.11}$$

Proof of Proposition is the subject of Exercise 4.7.

4.6.3 Mixed noise

So far, we have considered separately the cases of random and uncertain-but-bounded observation noises in (4.5.2). Note that both these observation schemes are covered by the following "mixed" scheme:

$$\omega = Ax + \xi + \eta, \tag{4.6.12}$$

where, as above, A is a given $m \times n$ matrix, x us unknown deterministic signal known to belong to a given signal set $\mathcal{X}$, ξ is random noise with distribution known to belong to a family $\mathcal{P}$ of Borel probability distributions on $\mathbf{R}^m$ satisfying (4.5.3) for a given convex compact set $\Pi \subset \operatorname{int} \mathbf{S}^m_+$, and η is "uncertain-but-bounded" observation error known to belong to a given set $\mathcal{H}$. As before, our goal is to recover $Bx \in \mathbf{R}^\nu$ via observation ω. In our present situation, given a norm $\|\cdot\|$ on $\mathbf{R}^\nu$, we can quantify the performance of a candidate estimate $\omega \mapsto \widehat{x}(\omega) : \mathbf{R}^m \to \mathbf{R}^\nu$ by its risk

$$\mathrm{Risk}_{\Pi,\mathcal{H},\|\cdot\|}[\widehat{x}|\mathcal{X}] = \sup_{x\in\mathcal{X}, P\lll\Pi, \eta\in\mathcal{H}} \mathbf{E}_{\xi\sim P}\{\|Bx - \widehat{x}(Ax+\xi+\eta)\|\}.$$

Observe that the estimation problem associated with "mixed" observation scheme straightforwardly reduces to similar problem for random observation scheme, by the same trick we have used in Section 4.6 to eliminate observation noise at all. Indeed, let us treat $x^+ = [x;\eta] \in \mathcal{X}^+ := \mathcal{X} \times \mathcal{H}$ and $\mathcal{X}^+$ as the new signal/signal set underlying our observation, and set $\bar{A}x^+ = Ax + \eta$, $\bar{B}x^+ = Bx$, where $x^+ = [x;\eta]$. With these conventions, the "mixed" observation scheme reduces to

$$\omega = \bar{A}x^+ + \xi,$$

and for every candidate estimate $\widehat{x}(\cdot)$ it clearly holds

$$\mathrm{Risk}_{\Pi,\mathcal{H},\|\cdot\|}[\widehat{x}|\mathcal{X}] = \mathrm{Risk}_{\Pi,\|\cdot\|}[\widehat{x}|\mathcal{X}^+],$$

and we arrive at the situation of Section 4.5.1. Assuming that $\mathcal{X}$ and $\mathcal{H}$ are spectratopes, so is $\mathcal{X}^+$, meaning that all results of Section 4.5 on building presumably good linear estimates and their near-optimality are applicable to our present setup.

An immediate question is: given the reduction we have described, what is the reason for considerations of Section 4.6.1 where we dealt with uncertain-but-bounded noise? The answer is: within its scope, Proposition 4.6.2 is stronger than Proposition 4.5.3, since the "nonoptimality factor" in Proposition 4.6.2 depends (logarithmically) solely on the sizes of the participating spectratopes, while in Proposition 4.5.3 this factor is affected also by the actual minimax risk and deteriorates, albeit just logarithmically, as the minimax risk goes to 0.

4.7 Proofs

4.7.1 Proofs for Section 4.2

4.7.1.1 Proof of Lemmas 4.2.1 and 4.3.2

Since Lemma 4.2.1 is the particular case $S = 0$ of Lemma 4.3.2, we prove here only the latter statement. Let $\widehat{x}(\cdot)$ be an estimate of $w = Bx$, and let R be its S-risk, so that

$$\forall (x \in \mathcal{X}) : \mathbf{E}_{\xi\sim\mathcal{N}(0,\Gamma)}\{\|\widehat{x}(Ax+\xi) - Bx\|_2^2\} \le R^2(1 + x^T S x),$$

see (4.3.1). Our intention is to bound R from below. Observe that $xx^T \in \mathcal{Q}$ when $x \in \mathcal{X}$, whence $\|Bx\|_2 = \sqrt{\mathrm{Tr}(Bxx^T B^T)} \le M_*$ for all $x \in \mathcal{X}$, see (4.2.17). It follows that projecting the estimate onto the $\|\cdot\|_2$-ball of radius M_* centered at the origin, we can only reduce the risk of the estimate.

Besides this, $\|Bx\|_2 \leq M_*$, $x \in \mathcal{X}$, implies that the trivial – identically zero – estimate has risk on $\mathcal{X}$ at most M_*. Consequently, we can assume w.l.o.g. that

$$R \leq M_* \ \& \ \|\widehat{x}(\omega)\|_2 \leq M_* \ \ \forall \omega \in \mathbf{R}^m. \tag{4.7.1}$$

Introducing Gaussian vector $[\eta;\xi]$ with independent $\xi \sim \mathcal{N}(0,\Gamma)$ and $\eta \sim \mathcal{N}(0,Q)$, and taking into account (4.7.1), we have for any $\gamma > 0$

$$\begin{array}{rcl}\varphi(Q) & \leq & \mathbf{E}_{[\xi;\eta]}\left\{\|\widehat{x}(A\eta+\xi)-B\eta\|_2^2\right\} \text{[by (4.2.14)]} \\ & = & \mathbf{E}_{\eta}\left\{\mathbf{E}_{\xi}\left\{\|\widehat{x}(A\eta+\xi)-B\eta\|_2^2\right\}\right\} \\ & = & \mathbf{E}_{\eta}\left\{\mathbf{E}_{\xi}\left\{\|\widehat{x}(A\eta+\xi)-B\eta\|_2^2\right\}1_{\eta\in\mathcal{X}}\right\}+\mathbf{E}_{\eta}\left\{\mathbf{E}_{\xi}\left\{\|\widehat{x}(A\eta+\xi)-B\eta\|_2^2\right\}1_{\eta\notin\mathcal{X}}\right\} \\ & \leq & R^2\mathbf{E}_{\eta}\left\{(1+\eta^TS\eta)1_{\eta\in\mathcal{X}}\right\}+\mathbf{E}_{\eta}\left\{[M_*+\|B\eta\|_2]^21_{\eta\notin\mathcal{X}}\right\} \quad {}^{17} \\ & \leq & R^2\mathbf{E}_{\eta}\left\{(1+\eta^TS\eta)\right\}+2\mathbf{E}_{\eta}\left\{\left[M_*^2+\|B\eta\|_2^4\right]1_{\eta\notin\mathcal{X}}\right\} \\ & \leq & R^2(1+\mathrm{Tr}(QS))+2M_*^2\delta+2\mathbf{E}\left\{\|B\eta\|_2^21_{\eta\notin\mathcal{X}}\right\} \\ & \leq & R^2(1+\mathrm{Tr}(QS))+2M_*^2\delta+2\sqrt{\mathbf{E}\left\{\|B\eta\|_2^41_{\eta\notin\mathcal{X}}\right\}}\sqrt{\delta} \text{ [Cauchy Inequality]} \\ & \leq & R^2(1+\mathrm{Tr}(QS))+2M_*^2\delta+2\sqrt{\mathbf{E}\left\{\|B\eta\|_2^4\right\}}\sqrt{\delta}\end{array} \tag{4.7.2}$$

Now let $BQ^{1/2} = UDV^T$ be the singular value decomposition of $BQ^{1/2}$, so that U and V are orthogonal $\nu \times \nu$, resp., $n \times n$ matrices, and the only nonzero entries in D are $D_{ii} =: \sigma_i \geq 0$, $1 \leq i \leq \min[\nu, n]$. Setting $\theta_i = \sigma_i^2$, $\theta = \sum_i \theta_i = \mathrm{Tr}([BQ^{1/2}][BQ^{1/2}]^T$, $\zeta \sim \mathcal{N}(0, I_n)$ and $\bar{\zeta} = V^T\zeta$, we have

$$\mathbf{E}\left\{\|B\eta\|_2^4\right\} = \mathbf{E}_{\zeta}\left\{\|BQ^{1/2}\zeta\|_2^4\right\} = \mathbf{E}_{\zeta}\left\{(\sum\nolimits_i \theta_i\bar{\zeta}_i^2)^2\right\}.$$

The concluding expression clearly is a convex function of $\theta_1, ..., \theta_n$, whence its maximum over $\theta_i \geq 0$ satisfying $\sum_i \theta_i = \theta$ is achieved when one of θ_i's is equal to θ, and all remaining θ_i's are zeros. Taking into account that $\bar{\zeta}_i \sim \mathcal{N}(0,1)$ (recall that V is orthogonal), we conclude that

$$\mathbf{E}\left\{\|B\eta\|_2^4\right\} \leq \theta^2\mathbf{E}_{s\sim\mathcal{N}(0,1)}\{s^4\} = 3\theta^2.$$

Thus, (4.7.2) implies that

$$\varphi(Q) \leq R^2(1+\mathrm{Tr}(QS)) + 2M_*^2\delta + 2\sqrt{3}\theta = R^2(1+\mathrm{Tr}(QS)) + 2M_*^2\delta + 2\sqrt{3}\mathrm{Tr}(BQB^T)\sqrt{\delta}.$$

Since R can be made arbitrarily close to $\mathrm{Risk}^2_{\mathrm{opt}}[\mathcal{X}]$ and $0 \leq \delta \leq 1$, we conclude that

$$\rho \in [0,1], Q \in \mathcal{Q}_\rho \Rightarrow \varphi(Q) \leq \mathrm{Risks}^2_{\mathrm{opt}}[\mathcal{X}](1+\mathrm{Tr}(QS)) + 2(1+\sqrt{3})M_*^2\sqrt{\delta} \tag{4.7.3}$$

(note that $\mathrm{Tr}(BQB^T) \leq M_*^2$ when $Q \in \mathcal{Q}_\rho \subset \mathcal{Q}$), as required in (4.3.7). □

4.7.1.2 Proof of Lemma 4.2.2

We set (cf. (4.3.4))

$$\mathbf{T} = \mathrm{cl}\{[t;\tau] \in \mathbf{R}^K \times \mathbf{R} : \tau > 0, \tau^{-1}t \in \mathcal{T}\} \subset \mathbf{R}_+^{K+1};$$

recall that $\mathbf{T}$ is a closed and pointed convex cone in $\mathbf{R}^{K+1}$ with a nonempty interior such that

$$\mathcal{T} = \{t : \exists \tau \leq 1 : [t;\tau] \in \mathbf{T}\} = \{t : [t;1] \in \mathbf{T}\}.$$

Note that (4.2.13) is nothing but the conic problem

$$\mathrm{Opt}_* = \max_{Q,G,t}\left\{\mathrm{Tr}(BQB^T) - \mathrm{Tr}(G) : \begin{array}{c}\left[\begin{array}{cc} G & BQA^T \\ AQB^T & \Gamma + AQA^T\end{array}\right] \succeq 0 \\ Q \succeq 0,\ [t;1] \in \mathbf{T},\ \mathrm{Tr}(QS_k) \leq t_k,\ 1 \leq k \leq K\end{array}\right\}. \tag{4.7.4}$$

[17] We use the standard notation: $1_{\eta\in\mathcal{X}}$ is the characteristic function of the event $\eta \in \mathcal{X}$, that is, function of η equal to 1 when $\eta \in \mathcal{X}$ and to 0 otherwise, and similarly for $1_{\eta\notin\mathcal{X}}$.

This problem clearly is strictly feasible (since $\mathrm{int}\,\mathcal{T}$ contains a positive vector) and bounded (the latter is due to $\sum_k S_k \succ 0$), so that its optimal value is equal to the optimal value of its conic dual problem, and all we need in order to prove (4.2.18) is to verify is that the latter problem is equivalent to (4.2.9).

Let us build the dual to (4.7.4) (for "guidelines," see Section 4.1.2.2). Note that the cone dual to $\mathbf{T}$ is

$$\mathbf{T}_* = \{[g;s] : s \geq \phi_{\mathcal{T}}(-g)\}.$$

Denoting the Lagrange multiplier for the first $\succeq$-constraint in (4.7.4) by $\left[\begin{array}{cc} U & V \\ V^T & W \end{array}\right] \succeq 0$, for the second $\succeq$-constraint by $L \succeq 0$, for $\leq$-constraints by $-\lambda$, $\lambda \in \mathbf{R}^K_+$, and for the constraint $[t;1] \in \mathbf{T}$ – by $[g;s] \in \mathbf{T}_*$, multiplying the constraints by the multipliers and summing up the results, we see that the constraints in (4.7.4) imply that on the feasible set of (4.7.4) it holds

$$\begin{array}{l} -\mathrm{Tr}(UG) - \mathrm{Tr}(Q[B^TVA + A^TV^TB]) - \mathrm{Tr}(Q[A^TWA]) - \mathrm{Tr}(LQ) + \sum\limits_k \lambda_k \mathrm{Tr}(QS_k) - \sum\limits_k \lambda_k t_k \\ -\sum\limits_k g_k t_k \leq \mathrm{Tr}(W\Gamma) + s. \end{array} \tag{4.7.5}$$

Now to get the dual to (4.7.4) problem, we need to impose on the Lagrange multipliers the constraint that the left hand side in (4.7.5) is identically in Q, G, t equal to the objective $\mathrm{Tr}(BQB^T) - \mathrm{Tr}(G)$ of (4.7.4), and to minimize over the multipliers under this constraint (in addition to those introduced when specifying the multipliers) the right hand side of (4.7.5). Thus, the problem dual to (4.7.4) is

$$\begin{array}{rcl} [\mathrm{Opt}_* & = &]\min\limits_{U,V,W,L,\lambda,g,s} \left\{ \mathrm{Tr}(W\Gamma) + s : \left[\begin{array}{cc} U & V \\ V^T & W \end{array}\right] \succeq 0,\ L \succeq 0,\ \lambda \geq 0,\ s \geq \phi_{\mathcal{T}}(-g), \right. \\ & & \left. g_k = -\lambda_k,\ 1 \leq k \leq K,\ U = I_\nu,\ -B^TVA - A^TV^TB - A^TWA - L + \sum\limits_k \lambda_k S_k = B^TB \right\} \\ & = & \min\limits_{V,W,\lambda,s} \left\{ \mathrm{Tr}(W\Gamma) + s : \begin{array}{l} W \succeq V^TV,\ \lambda \geq 0,\ s \geq \phi_{\mathcal{T}}(\lambda), \\ \sum_k \lambda_k S_k \succeq B^TB + B^TVA + A^TV^TB + A^TWA \end{array} \right\} \\ & = & \min\limits_{V,W,\lambda} \left\{ \mathrm{Tr}(V^TV\Gamma) + \phi_{\mathcal{T}}(\lambda) : \begin{array}{l} W = V^TV,\ \lambda \geq 0 \\ \sum_k \lambda_k S_k \succeq B^TB + B^TVA + A^TV^TB + A^TWA \end{array} \right\} \\ & = & \min\limits_{V,\lambda} \left\{ \mathrm{Tr}(V\Gamma V^T) + \phi_{\mathcal{T}}(\lambda) : \sum\limits_i \lambda_k S_k \succeq (B+VA)^T(B+VA),\ \lambda \geq 0 \right\}, \end{array}$$

that is, $\mathrm{Opt}_* = \mathrm{Opt}$ (substitute $H = -V^T$ in (4.2.9)). □

4.7.1.3 Proof of Lemma 4.2.3

Representing $\eta = Q^{1/2}\zeta$ with $\zeta \sim \mathcal{N}(0, I_n)$, we reduce the situation to the one where (Q, S) is replaced with $(I_n, \bar{S} = Q^{1/2}SQ^{1/2})$, so that it suffices to prove (4.2.20) in the special case of $Q = I_n$. Moreover, we clearly can assume that S is diagonal with diagonal entries $s_i \geq 0$, $1 \leq i \leq n$, so that $\rho = \sum_i s_i$. Now the relation we should prove reads

$$\mathrm{Prob}_{\eta \sim \mathcal{N}(0,I_n)} \left\{ \sum_{i=1}^n s_i \eta_i^2 > 1 \right\} \leq \mathrm{e}^{-\frac{1-\rho+\rho\ln(\rho)}{2\rho}}.$$

Let $\gamma \geq 0$ be such that $2\gamma \max_i s_i < 1$. Then

$$\ln\left(\mathbf{E}_\eta\{\exp\{\gamma \textstyle\sum_{i=1}^n s_i \eta_i^2\}\}\right) = \sum_{i=1}^n \ln\left(\mathbf{E}_\eta\{\exp\{\gamma s_i \eta_i^2\}\}\right) = -\frac{1}{2}\sum_{i=1}^n \ln(1 - 2\gamma s_i),$$

what implies the first inequality of (4.2.20). Furthermore, for $0 \le \gamma < \frac{1}{2\max_i s_i} \le \frac{1}{2\rho}$,

$$\ln\left(\mathbf{E}_\eta\left\{\exp\left[\gamma\sum_{i=1}^n s_i\eta_i^2\right]\right\}\right) \le -\frac{1}{2}\ln(1-2\gamma\rho)$$

(indeed, the convex function $-\frac{1}{2}\sum_{i=1}^n \ln(1-2\gamma s_i)$ of s varying in the simplex $\{s\ge 0, \sum_i s_i = \rho\}$ attains its maximum at a vertex of the simplex). Specifying $\gamma = \frac{1-\rho}{2\rho}$, we conclude that

$$\begin{array}{rcl}\mathrm{Prob}\left\{\sum_{i=1}^n s_i\eta_i^2 > 1\right\} & \le & \mathbf{E}_\eta\{\exp\{\gamma\sum_{i=1}^n s_i\eta_i^2\}\}\exp\{-\gamma\} \le \exp\{-\frac{1}{2}\ln(1-2\gamma\rho)-\gamma\}\\ & = & \exp\{-\frac{1-\rho+\rho\ln(\rho)}{2\rho}\},\end{array}$$

as claimed. □

4.7.1.4 Proof of Corollary 4.2.1

Observe that $\mathcal{X}$ contains a point $\bar{x}$ with

$$\|\bar{x}\|_2 \ge r := \frac{\sqrt{T}}{\mathrm{Cond}(\mathcal{T})\sqrt{\varkappa}}.$$

Indeed, by definition of $\mathrm{Cond}(\mathcal{T})$, $\mathcal{T}$ contains a vector $\bar{t}$ with all entries $\ge T/\mathrm{Cond}^2(\mathcal{T})$; let now $\bar{x} = re$, where e is the eigenvector of the matrix $S = \sum_{k=1}^K S_k$ corresponding to the minimal eigenvalue $\varkappa$ of this matrix. We have (recall that $S_k \succeq 0$, $k = 1,...,K$)

$$\bar{x}^T S_k \bar{x} \le \varkappa r^2 = T/\mathrm{Cond}^2(\mathcal{T}) \le \bar{t}_k,\ \ 1 \le k \le K,$$

that is, $\bar{x} \in \mathcal{X}$. Selecting the largest $t \in [0,1]$ such that $t\|A\bar{x}\|_2 \le c\sigma$, where c is the positive absolute constant from (4.2.26), we conclude by (4.2.26) that $\mathrm{Risk}_{\rm opt}[\mathcal{X}] \ge t\|B\bar{x}\|_2$, or

$$\mathrm{Risk}_{\rm opt}[\mathcal{X}] \ge \|B\bar{x}\|_2 \min\left[1, \frac{c\sigma}{\|A\bar{x}\|_2}\right].$$

Hence, we get

$$\begin{array}{rcl}\mathrm{Risk}_{\rm opt}[\mathcal{X}] & \ge & \sigma_{\min(B)}\|\bar{x}\|_2 \min\left[1, \frac{c\sigma}{\|A\|\,\|\bar{x}\|_2}\right] \ge \frac{\|B\|}{\mathrm{Cond}(B)}\min\left[\|\bar{x}\|_2, \frac{\sigma}{\|A\|}\right]\\ & \ge & \frac{\|B\|}{\mathrm{Cond}(B)}\min\left[r, \frac{\sigma}{\|A\|}\right] = \frac{\|B\|}{\mathrm{Cond}(B)}\min\left[\frac{\sqrt{T}}{\mathrm{Cond}(\mathcal{T})\sqrt{\varkappa}}, \frac{\sigma}{\|A\|}\right]\end{array}$$

Note that the quantity $M_* = \max_{Q\in\mathcal{Q}} \|BQ^{1/2}\|_2$ admits simple bound:

$$M_* \le \|B\|\sqrt{T/\varkappa}$$

(indeed, since $\sum_k \mathrm{Tr}(QS_k) \le T$ for all $Q \in \mathcal{Q}$, one has $\mathrm{Tr}(Q\sum_k S_k) \le T$, whence $\mathrm{Tr}(Q) \le T/\varkappa$ by the origin of $\varkappa$, and therefore $M_*^2 = \mathrm{Tr}(BQB^T) \le \|B^TB\|\mathrm{Tr}(Q) \le \|B\|^2T/\varkappa$). As a result,

$$\begin{array}{l}\frac{M_*\sqrt{K}}{\mathrm{Risk}_{\rm opt}[\mathcal{X}]} \le c'\mathrm{Cond}(B)\sqrt{\frac{TK}{\varkappa}}\max\left[\mathrm{Cond}(\mathcal{T})\sqrt{\frac{\varkappa}{T}}, \frac{\|A\|}{\sigma}\right]\\ \qquad\qquad \le c'\mathrm{Cond}(B)\sqrt{K}\left[\mathrm{Cond}(\mathcal{T}) + \frac{\|A\|\sqrt{T}}{\sigma\sqrt{\varkappa}}\right]\end{array}$$

with an absolute constant c'; together with (4.2.24) this implies (4.2.27). □

4.7.2 Proofs for Section 4.3

4.7.2.1 Proof of Lemma 4.3.1

1^o. We claim that (4.3.3) is a strictly feasible conic problem with bounded level sets of the objective (the sets where the objective is $\geq a$, for every fixed $a \in \mathbf{R}$); in particular, the problem is solvable.

Indeed, strict feasibility follows from the fact that the interior of the cone $\mathbf{T}$ contains a positive vector, see assumptions on $\mathcal{T}$ in Section 4.2.1. Further, the projections of the feasible set onto the $[v; s]$- and W-spaces are bounded (the first – since at a feasible solution it holds $0 \leq s \leq 1$, and the second – due to the boundedness of the set of v-components of feasible solutions combined with $\sum_k S_k \succ 0$). Boundedness of a level set of the objective follows from the fact that if a sequence of feasible solutions $\{(W_i, G_i, [v^i; s^i]), i = 1, 2, ...\}$ goes to ∞, then, by the above, the sequence $\{W_i, [v^i; s^i]\}$ is bounded, so that $\|G_i\| \to \infty$ as $i \to \infty$; since $G_i \succeq 0$ due to the constraints of the problem, we have $\mathrm{Tr}(G_i) \to \infty$ as $i \to \infty$, which combines with boundedness of $\{W_i\}$ to imply that the objective along our sequence of feasible solutions goes to $-\infty$, which is impossible for a sequence of feasible solutions from a level set of the objective.

2^o. Our next claim is that at an optimal solution $(W, G, [v; s])$ to (4.3.3) one has $s > 0$.

Indeed, otherwise $v = 0$ due to $[v; s] \in \mathbf{T}$ and the origin of $\mathbf{T}$, whence $W = 0$ due to $W \succeq 0$ and $\sum_k S_k \succ 0$; besides this, $G \succeq 0$, so that assuming $s = 0$, we see that $\mathrm{Opt}_* = 0$, which clearly is not the case: $\mathbf{T}$ contains a vector $[\bar{v}; \bar{s}]$ with, say, $\bar{s} = 0.1$ and positive $\bar{v}$, implying that for some $\bar{\tau} > 0$ and all $\tau \in [0, \bar{\tau}]$ tuples

$$W_\tau = \tau I, G_\tau = [\sigma^2 \bar{s}]^{-1}[BW_\tau A^T AW_\tau B^T] = [\sigma^2 \bar{s}]^{-1}\tau^2 BA^T AB^T, [\bar{v}; \bar{s}],$$

where $\sigma^2 > 0$ is the minimal eigenvalue of Γ, are feasible solutions to (4.3.3); since $B \neq 0$, for small positive τ the value of the objective of (4.3.3) at such a solution is positive, which would be impossible when $\mathrm{Opt}_* = 0$.

Furthermore, observe that if (W, G, v, s) is an optimal solution to (4.3.3) (whence, as we already know, $s > 0$), when replacing G with the matrix

$$\bar{G} := BWA^T(s\Gamma + AWA^T)^{-1}AWB^T$$

(so that $G \succeq \bar{G}$ and $(W, \bar{G}, t, s)$ is feasible for (4.3.3)), we keep the solution optimal, thus

$$\mathrm{Opt}_* = \mathrm{Tr}\left(B[W - WA^T(s\Gamma + AWA^T)^{-1}AW]B^T\right).$$

3^o. To complete the proof of the lemma it suffices to show that the conic dual to (4.3.3) is equivalent to (4.3.2); since (4.3.3), as we have already mentioned, is strictly feasible and bounded, this would imply that $\mathrm{Opt} = \mathrm{Opt}_*$.

To build the problem dual to (4.3.3), let the Lagrange multipliers for the constraints be, respectively, $\left[\begin{array}{cc} U & V \\ V^T & Z \end{array}\right] \succeq 0$, $L \succeq 0$, $-\lambda$, $\lambda \in \mathbf{R}^K_+$, $-\tau$, $\tau \geq 0$, and $[g; r] \in \mathbf{T}_*$, where

$$\mathbf{T}_* = \{[g; r] : r \geq \phi_{\mathcal{T}}(-g)\}$$

is the cone dual to $\mathbf{T}$. Taking inner products of the constraints of (4.3.3) with the multipliers and summing up the results, we arrive at the aggregated constraint

$$\begin{array}{r}\mathrm{Tr}(GU) + \mathrm{Tr}(W[A^TV^TB + B^TVA + A^TZA + L - \sum_k \lambda_k S_k - \tau S]) \\ + \sum_k[\lambda_k + g_k]v_k + s[\mathrm{Tr}(Z\Gamma) - \tau + r] + \tau \geq 0\end{array}$$

To get the dual problem, we impose on the multipliers the restriction for the resulting inequality to have the homogeneous in W, G, v, s component identically equal to *minus* the objective of (4.3.3), which amounts to the relations

$$\begin{array}{l} U = I_\nu,\ \tau = r + \mathrm{Tr}(Z\Gamma),\ g_k = -\lambda_k\ \forall k, \\ [A^TV^TB + B^TVA + A^TZA + L - \sum_k \lambda_k S_k - \tau S] = -B^TB. \end{array}$$

Under these relations, the aggregated constraint reads

$$\mathrm{Tr}(BWB^T - G) \le \tau$$

for all feasible solutions to (4.3.3), thus $\mathrm{Opt}_* \le \tau$. Therefore, the problem dual to (4.3.3) is to minimize the resulting upper bound on Opt_*, that is, the dual is

$$\min_{\tau,V,Z,L,\lambda,[g;r]} \left\{ \tau : \begin{array}{l} \left[\begin{array}{cc} I_\nu & V \\ V^T & Z \end{array}\right] \succeq 0,\ L \succeq 0,\ \lambda \ge 0,\ \tau \ge 0,\ r \ge \phi_{\mathcal{T}}(-g) \\ B^TB + A^TV^TB + B^TVA + A^TZA = \sum_k \lambda_k S_k + \tau S - L \\ g = -\lambda, \tau = r + \mathrm{Tr}(Z\Gamma) \end{array} \right\}.$$

Now partial minimization in Z and r results in $Z = V^TV$, $r = \phi_{\mathcal{T}}(-g)$, which, after eliminating L and $[g; r]$, reduces the dual problem to

$$\min_{\tau,V,\lambda} \left\{ \tau : \begin{array}{l} (B + VA)^T(B + VA) \preceq \sum_k \lambda_k S_k + \tau S, \\ \lambda \ge 0,\ \tau \ge \phi_{\mathcal{T}}(\lambda) + \mathrm{Tr}(V^TV\Gamma) \end{array} \right\}.$$

The resulting problem clearly is equivalent to (4.3.2) (substitute $V = -H^T$). Thus, (4.3.5) is proved. □

4.7.2.2 Proof of Proposition 4.3.2

Under the premise of the proposition, the feasible set of (4.3.9) is nonempty, and the objective clearly goes to ∞ along every going to ∞ sequence of feasible solutions (τ_i, H_i), implying that the problem is solvable. The optimal value Opt in the problem clearly is positive due to $\Gamma \succ 0$ and $B \neq 0$. Now assume that (4.3.10) does not hold, so that there exists α and estimate $\widehat{x}_*(\cdot)$ such that

$$\alpha < \mathrm{Opt}\ \&\ \mathbf{E}_{\xi\sim\mathcal{N}(0,\Gamma)}\{\|\widehat{x}_*(Ax + \xi) - Bx\|_2^2\} \le \alpha(1 + x^TSx)\ \forall x \in \mathbf{R}^n, \tag{4.7.6}$$

and let us lead this assumption to contradiction.

Consider the conic problem (cf. (4.3.3))

$$\mathrm{Opt}_* = \max_{W,G,s} \left\{ \mathrm{Tr}(BWB^T) - \mathrm{Tr}(G) : \begin{array}{c} \left[\begin{array}{cc} G & BWA^T \\ AWB^T & s\Gamma + AWA^T \end{array}\right] \succeq 0, \\ W \succeq 0,\ \mathrm{Tr}(WS) + s \le 1,\ s \ge 0 \end{array} \right\}. \tag{4.7.7}$$

This conic problem clearly is strictly feasible; the same argument as in the case of (4.3.3) shows that the conic dual of this problem is equivalent to (4.3.9) and therefore is feasible. By Conic Duality Theorem, it follows that both (4.7.7) and (4.3.9) have equal optimal values, and since $\Gamma \succ 0$, $B \neq 0$, Opt is positive. Thus,

$$\mathrm{Opt}_* = \mathrm{Opt} > 0.$$

This relation, due to $\alpha < \mathrm{Opt}$, implies that there is a feasible solution to (4.7.7) with the value of the objective $> \alpha$. Since the problem is strictly feasible, feasible solutions with $s > 0$ are dense in the feasible set, implying that the above feasible solution, let it be $(\widehat{W}, G, \widehat{s})$, can be selected to have

$\widehat{s} > 0$. Further, keeping $\widehat{W}$ and $\widehat{s}$ intact and replacing G with $\widehat{G} = B\widehat{W}A^T[\widehat{s}\Gamma + A\widehat{W}A^T]^{-1}A\widehat{W}B^T$, we preserve feasibility and can only increase the objective of (4.7.7). The bottom line is that we can point out a feasible solution $(\widehat{W}, \widehat{G}, \widehat{s})$ to (4.7.7) such that

$$\begin{array}{l} \widehat{\alpha} := \mathrm{Tr}(B^T[\widehat{W} - \widehat{W}A^T[\widehat{s}\Gamma + A\widehat{W}A^T]^{-1}A\widehat{W}]B) > \alpha, \\ \widehat{s} > 0,\ \widehat{W} \succeq 0,\ \mathrm{Tr}(\widehat{W}S) + \widehat{s} \leq 1. \end{array} \tag{4.7.8}$$

Observe that

$$\widehat{\alpha} = \widehat{s}\varphi(\widehat{s}^{-1}\widehat{W}) \tag{4.7.9}$$

(see (4.2.12)). Now let $\eta \sim \mathcal{N}(0, \widehat{s}^{-1}\widehat{W})$ be independent of $\xi \sim \mathcal{N}(0, \Gamma)$. We have

$$\begin{array}{rcl} \mathbf{E}_{[\eta;\xi]}\{\|\widehat{x}_*(A\eta + \xi) - B\eta\|_2^2\} & = & \mathbf{E}_\eta\left\{\mathbf{E}_\xi\{\|\widehat{x}_*(A\eta + \xi) - B\eta\|_2^2\}\right\} \\ & \leq & \mathbf{E}_\eta\left\{\alpha(1 + \eta^T S\eta)\right\}\ \text{[by (4.7.6)]} \\ & = & \alpha(1 + \widehat{s}^{-1}\mathrm{Tr}(\widehat{W}S)). \end{array}$$

By (4.2.14), the initial quantity in this chain is $\geq \varphi(\widehat{s}^{-1}\widehat{W}) = \widehat{s}^{-1}\widehat{\alpha}$ (see (4.7.9)), so that the chain yields $\widehat{s}^{-1}\widehat{\alpha} \leq \alpha(1 + \widehat{s}^{-1}\mathrm{Tr}(\widehat{W}S))$, that is,

$$\widehat{\alpha} \leq \alpha(\widehat{s} + \mathrm{Tr}(\widehat{W}S)) \leq \alpha,$$

where the last $\leq$ stems from the last inequality in (4.7.8). The resulting inequality contradicts the first inequality in (4.7.8); we have arrived at the desired contradiction. □

4.7.2.3 Proof of Proposition 4.3.3

We need the following

Lemma 4.7.1 *Let S be a positive semidefinite $\bar{n}\times\bar{n}$ matrix with unit trace and ξ be a Rademacher $\bar{n}$-dimensional random vector (i.e., the entries in ξ are independent and take values ± 1 with probabilities $1/2$). Then*

$$\mathbf{E}\Big\{\exp\left\{\tfrac{1}{4}\xi^T S\xi\right\}\Big\} \leq 3\sqrt{2}, \tag{4.7.10}$$

implying that

$$\mathrm{Prob}\{\xi^T S\xi > s\} \leq 3\sqrt{2}\exp\{-s/4\},\ s \geq 0.$$

Proof. Let $S = \sum_{i=1}^{\bar{n}} \lambda_i g_i g_i^T$ be the eigenvalue decomposition of S, so that $\lambda_i \geq 0$, $\sum_i \lambda_i = 1$ and $\|g_i\|_2 = 1$. Then

$$\mathbf{E}\Big\{\exp\left\{\tfrac{1}{4}\xi^T S\xi\right\}\Big\} = \mathbf{E}\{\exp\{\frac{1}{4}\sum_i \lambda_i (g_i^T\xi)^2\}\}$$

is a convex function of λ and therefore it attains its maximum over nonnegative vectors λ with unit sum of entries at a basic orth. Thus, it suffices to verify (4.7.10) when $S = gg^T$ with unit vector g. By the Hoeffding inequality one has

$$\mathrm{Prob}\{|g^T\xi| > s\} \leq 2\exp\{-s^2/2\}.$$

It follows that $\mathrm{Prob}\{(g^T\xi)^2 > r\} \leq 2\exp\{-r/2\}$, and thus $p(r) := \mathrm{Prob}\{\frac{1}{4}(g^T\xi)^2 \geq r\} \leq 2\exp\{-2r\}$. Consequently,

$$\begin{array}{l} \mathbf{E}\Big\{\exp\{\tfrac{1}{4}(g^T\xi)^2\}\Big\} = \int_0^\infty \exp\{r\}[-dp(r)] = \int_0^\infty \exp\{r\}p(r)dr + 1 \\ \leq \int_0^\infty \exp\{r\}\min[1, 2\exp\{-2r\}]dr + 1 = 1 + \int_0^{\frac{1}{2}\ln(2)} \exp\{r\}dr + 2\int_{\frac{1}{2}\ln(2)}^\infty \exp\{-r\}dr = 3\sqrt{2}. \end{array}$$

□

2^0. The right inequality in (4.3.19) has already been justified. To prove the left inequality in (4.3.19), we, similarly to what was done in Section 4.2.3, introduce the conic problem

$$\mathrm{Opt}_* = \max_{Q,t} \left\{ \mathrm{Tr}(P^TCPQ) : Q \succeq 0, \mathrm{Tr}(QS_k) \leq t_k \, \forall k \leq K, [t;1] \in \mathbf{T} \right\}, \tag{4.7.11}$$

and acting exactly as in the derivation of (4.2.18), we arrive at

$$\mathrm{Opt} = \mathrm{Opt}_*. \tag{4.7.12}$$

Indeed, (4.7.11) is a strictly feasible and bounded conic problem, so that its optimal value is equal to the one in its conic dual, that is,

$$\begin{array}{rcl}
\mathrm{Opt}_* & = & \min\limits_{\lambda,[g;s],L} \left\{ s : \begin{array}{l} \mathrm{Tr}([\sum_k \lambda_k S_k - L]Q) - \sum_k [\lambda_k + g_k] t_k = \mathrm{Tr}(P^TCPQ) \;\; \forall (Q,t), \\ \lambda \geq 0, L \succeq 0, s \geq \phi_{\mathcal{T}}(-g) \end{array} \right\} \\
& = & \min\limits_{\lambda,[g;s],L} \left\{ s : \begin{array}{l} \sum_k \lambda_k S_k - L = P^TCP, \, g = -\lambda, \\ \lambda \geq 0, L \succeq 0, s \geq \phi_{\mathcal{T}}(-g) \end{array} \right\} \\
& = & \min\limits_{\lambda} \left\{ \phi_{\mathcal{T}}(\lambda) : \sum\limits_k \lambda_k S_k \succeq P^TCP, \lambda \geq 0 \right\} = \mathrm{Opt}.
\end{array}$$

3^0. With Lemma 4.7.1 and (4.7.12) at our disposal, we can now complete the proof of Proposition 4.3.3 by adjusting the technique from [124]. Specifically, problem (4.7.11) clearly is solvable; let Q_*, t^* be an optimal solution to the problem. Next, let us set $R_* = Q_*^{1/2}$, $\bar{C} = R_* P^T C P R_*$, let $\bar{C} = UDU^T$ be the eigenvalue decomposition of $\bar{C}$, and let $\bar{S}_k = U^T R_* S_k R_* U$. Observe that

$$\begin{array}{rcl}
\mathrm{Tr}(D) & = & \mathrm{Tr}(R_* P^T C P R_*) = \mathrm{Tr}(Q_* P^T C P) = \mathrm{Opt}_* = \mathrm{Opt}, \\
\mathrm{Tr}(\bar{S}_k) & = & \mathrm{Tr}(R_* S_k R_*) = \mathrm{Tr}(Q_* S_k) \leq t_k^*.
\end{array}$$

Now let ξ be Rademacher random vector. For k with $t_k^* > 0$, applying Lemma 4.7.1 to matrices $\bar{S}_k / t_k^*$, we get for $s > 0$

$$\mathrm{Prob}\{\xi^T \bar{S}_k \xi > s t_k^*\} \leq 3\sqrt{2} \exp\{-s/4\}; \tag{4.7.13}$$

if k is such that $t_k^* = 0$, we have $\mathrm{Tr}(\bar{S}_k) = 0$, that is, $\bar{S}_k = 0$, and (4.7.13) holds true as well. Now let

$$s_* = 4\ln(5K),$$

so that $3\sqrt{2}\exp\{-s_*/4\} < 1/K$. The latter relation combines with (4.7.13) to imply that there exists a realization $\bar{\xi}$ of ξ such that

$$\bar{\xi}^T \bar{S}_k \bar{\xi} \leq s_* t_k^* \, \forall k.$$

Let us set $\bar{y} = \frac{1}{\sqrt{s_*}} R_* U \bar{\xi}$. Then

$$\bar{y}^T S_k \bar{y} = s_*^{-1} \bar{\xi}^T U^T R_* S_k R_* U \bar{\xi} = s_*^{-1} \bar{\xi}^T \bar{S}_k \bar{\xi} \leq t_k^* \;\; \forall k$$

implying that $\bar{y} \in \bar{X}$, and

$$\bar{y}^T P^T C P \bar{y} = s_*^{-1} \bar{\xi}^T U^T R_* C R_* U \bar{\xi} = s_*^{-1} \bar{\xi}^T D \bar{\xi} = s_*^{-1} \mathrm{Tr}(D) = s_*^{-1} \mathrm{Opt}.$$

Thus, $\max_{y \in \bar{X}} y^T P^T C P y \geq s_*^{-1} \mathrm{Opt}$, which is the first inequality in (4.3.19). □

4.7.3 Proofs for Section 4.4

4.7.3.1 Proof of Lemma 4.4.1

Let $x\in\mathcal{X}$, so that for some $t\in\mathcal{T}$ it holds

$$R_k^2[x]\preceq t_kI_{d_k}\;\forall k\leq K$$

We have

$$Q\preceq\sum\mathcal{R}_k^*[\Lambda_k]+\tau S\Rightarrow\begin{array}{l}x^TQx\leq x^T\sum_k\mathcal{R}_k^*[\Lambda_k]x+\tau x^TSx=\sum_k\mathrm{Tr}(\mathcal{R}_k^*[\Lambda_k][xx^T])+\tau x^TSx\\=\sum_k\mathrm{Tr}(\Lambda_k\mathcal{R}_k[xx^T])+\tau x^TSx\text{ [by (4.4.8)]}\\=\sum_k\mathrm{Tr}(\Lambda_kR_k^2[x])+\tau x^TSx\text{ [by (4.4.4)]}\\\leq\sum_k t_k\mathrm{Tr}(\Lambda_kI_{d_k})+\tau x^TSx\leq\phi_{\mathcal{T}}(\lambda[\Lambda])+\tau x^TSx.\end{array}$$ □

4.7.3.2 Proof of Proposition 4.4.1

1^o. Let us start with the following simple observation:

Lemma 4.7.2 *Given spectratope (4.4.11) and a positive definite $n\times n$ matrix Q and setting $\Lambda_k=\mathcal{R}_k[Q]$, we get a collection of positive semidefinite matrices, and $\sum_k\mathcal{R}_k^*(\Lambda_k)$ is positive definite.*

As a corollary, whenever M_k, $k\leq K$, are positive definite matrices, the matrix $\sum_k\mathcal{R}_k^*[M_k]$ is positive definite.

Proof of Lemma. Let us prove the first claim, Assuming the opposite, we would be able to find a nonzero vector y such that $\sum_k y^T\mathcal{R}_k^*(\Lambda_k)y\leq0$, whence

$$0\geq\sum_k y^T\mathcal{R}_k^*(\Lambda_k)y=\sum_k\mathrm{Tr}(\mathcal{R}_k^*[\Lambda_k][yy^T])=\sum_k\mathrm{Tr}(\Lambda_k\mathcal{R}_k[yy^T])$$

(we have used (4.4.8), (4.4.4)). Since $\Lambda_k=\mathcal{R}_k[Q]\succeq0$ due to $Q\succeq0$, see (4.4.5), it follows that $\mathrm{Tr}(\Lambda_k\mathcal{R}_k[yy^T])=0$ for all k. Now, the linear mapping $\mathcal{R}_k[\cdot]$ is $\succeq$-monotone, and Q is positive definite, implying that $Q\succeq r_kyy^T$ for some $r_k>0$, whence $\Lambda_k\succeq r_k\mathcal{R}_k[yy^T]$, and therefore $\mathrm{Tr}(\Lambda_k\mathcal{R}_k[yy^T])=0$ implies that $\mathrm{Tr}(\mathcal{R}_k[yy^T])=0$, that is, $\mathcal{R}_k[yy^T]=R_k^2[y]=0$. Since $R_k[\cdot]$ takes values in $\mathbf{S}^{d_k}$, we get $R_k[y]=0$ for al k, which is impossible due to $y\neq0$ and property S.3, see Section 4.4.1.

The second claim is an immediate consequence of the first one. Indeed, when M_k are positive definite, we can find $\gamma>0$ such that $\Lambda_k\preceq\gamma M_k$ for all $k\leq K$; invoking (4.4.9), we conclude that $\mathcal{R}_k^*[\Lambda_k]\preceq\gamma\mathcal{R}_k^*[M_k]$, whence $\sum_k\mathcal{R}_k^*[M_k]$ is positive definite along with $\sum_k\mathcal{R}_k^*[\Lambda_k]$. □

2^o. Relation $\sum_k\mathcal{R}_k^*(\Lambda_k)\succ0$ combines with $S\succeq0$ to imply that (4.4.12) is feasible. The level sets of (4.4.12) clearly are bounded, which combines with feasibility to imply solvability.

Now let (H,λ,τ) be a feasible solution to (4.4.12), so that

$$Q\preceq\tau S+\sum_k\mathcal{R}_k^*[\Lambda_k]$$

due to the constraints of (4.4.12). By Lemma 4.4.1 as applied to $Q=(B-H^TA)^T(B-H^TA)$, for $x\in\mathcal{X}$ it holds

$$x^T(B-H^TA)^T(B-H^TA)x\leq\tau x^TSx+\phi_{\mathcal{T}}(\lambda[\Lambda])$$

whence

$$\begin{array}{rl}\mathbf{E}_{\omega\sim\mathcal{N}(Ax,\Gamma)}\{\|H^T\omega-Bx\|_2^2\}=&x^T(B-H^TA)^T(B-H^TA)x+\mathrm{Tr}(H^T\Gamma H)\\&\leq\tau x^TSx+\underbrace{[\mathrm{Tr}(H^T\Gamma H)+\phi_{\mathcal{T}}(\lambda[\Lambda])]}_{\leq\tau},\end{array}$$

and therefore $\mathrm{RiskS}[\widehat{x}_H|\mathcal{X}]\leq\sqrt{\tau}$. □

4.7.3.3 Proof of Lemma 4.4.2

The proof follows the one of Lemma 4.3.1.

1^o. We claim that (4.4.15) is a strictly feasible conic problem with bounded level sets of the objective; in particular, the problem is solvable.

Indeed, strict feasibility follows from the fact that the interior of the cone $\mathbf{T}$ contains a positive vector, see assumptions on $\mathcal{T}$ in Section 4.2.1. Further, the projections of the feasible set onto the $[v;s]$- and W-spaces are bounded. Indeed, at a feasible solution it holds $0 \leq s \leq 1$, whence the set of v-components of feasible solutions is bounded due to $[v;s] \in \mathbf{T}$, implying that the set of $[v;s]$-components of feasible solutions is bounded. Second, let us fix $\Lambda_k \in \mathbf{S}^{d_k}$ such that $\Lambda_k \succ 0$, $11 \leq k \leq K$. W-component of a feasible solution $(W, G, [v;s])$ should satisfy

$$\mathrm{Tr}(W \sum_k \mathcal{R}_k^*[\Lambda_k]) = \sum_k \mathrm{Tr}(W \mathcal{R}_k^*[\Lambda_k]) = \sum_k \mathrm{Tr}(\mathcal{R}_k[W]\Lambda_k) \leq \sum_k v_k \mathrm{Tr}(\Lambda_k) \leq C \tag{4.7.14}$$

for properly selected $C \in \mathbf{R}$ (since the $[v;s]$-components of feasible solutions form a bounded set). By Lemma 4.7.2, the matrix $\sum_k \mathcal{R}_k^*[\Lambda_k]$ is positive definite, which combines with (4.7.14) and positive semidefiniteness of W to imply that the set $\mathcal{W}$ of W-components of feasible solutions to (4.4.15) is bounded.

Note that in fact our reasoning justifies the following claim:

(!) *In a feasible solution to* (4.4.15) *with* $v_k = 0, k \leq K$, $W = 0$ *as well.*

Boundedness of a level set of the objective follows from the fact that if a sequence of feasible solutions $\{(W_i, G_i, [v^i; s^i]), i = 1, 2, ...\}$ goes to ∞, then, by the above, the sequence $\{W_i, [v^i; s^i]\}$ is bounded, so that $\|G_i\| \to \infty$ as $i \to \infty$; since $G_i \succeq 0$ due to the constraints of the problem, we have $\mathrm{Tr}(G_i) \to \infty$ as $i \to \infty$, which combines with boundedness of $\{W_i\}$ to imply that the objective along our sequence of feasible solutions goes to $-\infty$, which is impossible for a sequence of feasible solutions from a level set of the objective.

2^o. Our next claim is that at an optimal solution $(W, G, [v;s])$ to (4.4.15) one has $s > 0$.

Indeed, otherwise $v = 0$ due to $[v;s] \in \mathbf{T}$ and the origin of $\mathbf{T}$, whence $W = 0$ due to (!); besides this, $G \succeq 0$, so that assuming $s = 0$, we see that $\mathrm{Opt}_* = 0$, which clearly is not the case: $\mathbf{T}$ contains a vector $[\bar{v}; \bar{s}]$ with, say, $\bar{s} = 0.1$ and positive $\bar{v}$, implying that for some $\bar{\tau} > 0$ and all $\tau \in [0, \bar{\tau}]$ tuples

$$W_\tau = \tau I, G_\tau = [\sigma^2 \bar{s}]^{-1} [B W_\tau A^T A W_\tau B^T] = [\sigma^2 \bar{s}]^{-1} \tau^2 B A^T A B^T, [\bar{v}; \bar{s}],$$

where $\sigma^2 > 0$ is the minimal eigenvalue of Γ, are feasible solutions to (4.4.15); since $B \neq 0$, for small positive τ the value of the objective of (4.4.15) at such a solution is positive, which would be impossible when $\mathrm{Opt}_* = 0$.

Furthermore, observe that if (W, G, v, s) is an optimal solution to (4.4.15) (whence, as we already know, $s > 0$), when replacing G with the matrix

$$\bar{G} := B W A^T (s\Gamma + A W A^T)^{-1} A W B^T$$

(so that $G \succeq \bar{G}$ and $(W, \bar{G}, t, s)$ is feasible for (4.4.15)), we keep the solution optimal, thus

$$\mathrm{Opt}_* = \mathrm{Tr}\left(B[W - W A^T (s\Gamma + A W A^T)^{-1} A W] B^T\right).$$

3^o. To complete the proof of Lemma it suffices to show that the conic dual to (4.4.15) is equivalent to (4.4.12); since (4.4.15), as we have already mentioned, is strictly feasible and bounded, this would imply that $\mathrm{Opt} = \mathrm{Opt}_*$.

To build the problem dual to (4.4.15), let the Lagrange multipliers for the constraints be, respectively, $\left[\begin{array}{cc} U & V \\ V^T & Z \end{array}\right] \succeq 0$, $L \succeq 0$, $\{-\Lambda_k : k \leq K\}$, $\Lambda_k \in \mathbf{S}^{d_k}_+$, $-\tau$, $\tau \geq 0$, and $[g;r] \in \mathbf{T}_*$, where

$$\mathbf{T}_* = \{[g;r] :\ r \geq \phi_{\mathcal{T}}(-g)\}$$

is the cone dual to $\mathbf{T}$. Taking inner products of the constraints of (4.4.15) with the multipliers and summing up the results, we arrive at the aggregated constraint

$$\begin{array}{r} \mathrm{Tr}(GU) + \mathrm{Tr}(W[A^TV^TB + B^TVA + A^TZA + L - \sum_k \mathcal{R}_*[\Lambda_k] - \tau S]) \\ + \sum_k [\mathrm{Tr}(\Lambda_k) + g_k] v_k + s[\mathrm{Tr}(Z\Gamma) - \tau + r] + \tau \geq 0 \end{array}$$

(note that $\mathrm{Tr}(\Lambda_k \mathcal{R}_k[W]) = \mathrm{Tr}(W \mathcal{R}^*_k[\Lambda_k])$ by (4.4.8)). To get the dual problem, we impose on the multipliers the restriction for the resulting inequality to have the homogeneous in W, G, v, s component identically equal to *minus* the objective of (4.4.15), which amounts to the relations

$$\begin{array}{l} U = I_n,\ \tau = r + \mathrm{Tr}(Z\Gamma),\ g_k = -\mathrm{Tr}(\lambda_k)\ \forall k, \\ {[A^TV^TB + B^TVA + A^TZA + L - \sum\limits_k \mathcal{R}^*_k(\Lambda_k) - \tau S] = -B^TB.} \end{array}$$

Under these relations, the aggregated constraint reads

$$\mathrm{Tr}(BWB^T - G) \leq \tau$$

for all feasible solutions to (4.4.15), thus $\mathrm{Opt}_* \leq \tau$. Therefore, the problem dual to (4.4.15) is to minimize the resulting upper bound on Opt_*, that is, the dual is

$$\min_{\tau, V, Z, L, \Lambda, [g;r]} \left\{ \tau : \begin{array}{l} \left[\begin{array}{cc} I_n & V \\ V^T & Z \end{array}\right] \succeq 0,\ L \succeq 0, \Lambda = \{\Lambda_k \succeq 0\}_{k \leq K},\ \tau \geq 0,\ r \geq \phi_{\mathcal{T}}(-g) \\ B^TB + A^TV^TB + B^TVA + A^TZA = \sum_k \mathcal{R}^*_k[\Lambda_k] + \tau S - L \\ g = -\lambda[\Lambda], \tau = r + \mathrm{Tr}(Z\Gamma) \end{array} \right\}.$$

Now partial minimization in Z and r, g results in $Z = V^TV$ and $r = \phi_{\mathcal{T}}(\lambda[\Lambda])$ which, after eliminating L and $[g;r]$, reduces the dual problem to

$$\min_{\tau, V, \lambda} \left\{ \tau : \begin{array}{l} (B + VA)^T(B + VA) \preceq \sum_k \mathcal{R}^*_k[\Lambda_k] + \tau S, \\ \Lambda = [\Lambda_k \succeq 0\}_{k \leq K},\ \tau \geq \phi_{\mathcal{T}}(\lambda) + \mathrm{Tr}(V^TV\Gamma) \end{array} \right\}.$$

The resulting problem clearly is equivalent to (4.4.12) (substitute $V = -H^T$). Thus, (4.4.17) is proved. □

4.7.3.4 Proof of Proposition 4.4.2

A. We are about to use deep result from Functional Analysis ("Noncommutative Khintchine Inequality") due to Lust-Piquard [114], Pisier [136] and Buchholz [29], see [154, Theorem 4.6.1]:

Theorem 4.7.1 *Let $Q_i \in \mathbf{S}^n$, $1 \leq i \leq I$, and let ξ_i, $i = 1, ..., I$, be independent Rademacher (± 1 with probabilities $1/2$) or $\mathcal{N}(0,1)$ random variables. Then for all $t \geq 0$ one has*

$$\mathrm{Prob}\left\{ \left\| \sum_{i=1}^I \xi_i Q_i \right\| \geq t \right\} \leq 2n \exp\left\{ -\frac{t^2}{2v_Q} \right\}$$

where $\|\cdot\|$ is the spectral norm, and $v_Q = \left\| \sum_{i=1}^I Q_i^2 \right\|$.

We need the following immediate consequence of Theorem:

Lemma 4.7.3 *Given spectratope* (4.4.1)*, let $Q \in \mathbf{S}^n_+$ be such that*

$$\mathcal{R}_k[Q] \preceq \rho t_k I_{d_k},\, 1 \le k \le K, \tag{4.7.15}$$

for some $t \in \mathcal{T}$ and some $\rho \in (0,1]$. Then

$$\mathrm{Prob}_{\xi\sim\mathcal{N}(0,Q)}\{\xi \notin \mathcal{X}\} \le \min\left[2D\mathrm{e}^{-\frac{1}{2\rho}}, 1\right],\ D := \sum_{k=1}^K d_k.$$

Proof. When setting $\xi = Q^{1/2}\eta$, $\eta \sim \mathcal{N}(0, I_n)$, we have

$$R_k[\xi] = R_k[Q^{1/2}\eta] =: \sum_{i=1}^n \eta_i \bar{R}^{ki} = \bar{R}_k[\eta]$$

with

$$\sum_i [\bar{R}^{ki}]^2 = \mathbf{E}_{\eta\sim\mathcal{N}(0,I_n)}\left\{\bar{R}_k^2[\eta]\right\} = \mathbf{E}_{\xi\sim\mathcal{N}(0,Q)}\left\{R_k^2[\xi]\right\} = \mathcal{R}_k[Q] \preceq \rho t_k I_{d_k}$$

due to (4.4.6). Hence, by Theorem 4.7.1

$$\mathrm{Prob}_{\xi\sim\mathcal{N}(0,Q)}\{\|R_k[\xi]\|^2 \ge t_k\} = \mathrm{Prob}_{\eta\sim\mathcal{N}(0,I_n)}\{\|\bar{R}_k[\zeta]\|^2 \ge t_k\} \le 2d_k\mathrm{e}^{-\frac{1}{2\rho}}.$$

We conclude that

$$\mathrm{Prob}_{\xi\sim\mathcal{N}(0,Q)}\{\xi \notin \mathcal{X}\} \le \mathrm{Prob}_{\xi\sim\mathcal{N}(0,Q)}\{\exists k : \|R_k[\xi]\|^2 > t_k\} \le 2D\mathrm{e}^{-\frac{1}{2\rho}}. \qquad \square$$

B. Observe that the set $\mathcal{Q}$ given by (4.4.13) is a nonempty convex compact set (compactness is readily given by Lemma 4.7.2, cf. the proof of boundedness of the set $\mathcal{W}$ in item 1^o of the proof of Lemma 4.4.2).

Invoking Lemma 4.7.3, we arrive at the following relation:

$$\rho \in (0,1], Q \in \mathcal{Q}_\rho \Rightarrow \mathrm{Prob}_{\eta\sim\mathcal{N}(0,Q)}\{\eta \notin \mathcal{X}\} \le \delta_\rho := \min\left[2D\exp\left\{-\frac{1}{2\rho}\right\}, 1\right],\ D = \sum_{k\le K} d_k. \tag{4.7.16}$$

We have the following analogy of Lemma 4.3.2:

Lemma 4.7.4 *Consider spectratope* (4.4.11)*. Given $\rho \in (0,1]$, $Q \in \mathcal{Q}_\rho$, and $\delta \le 1$, let $\eta \sim \mathcal{N}(0,Q)$. Assume that*

$$\mathrm{Prob}\{\eta \notin \mathcal{X}\} \le \delta.$$

Then

$$\varphi(Q) \le \mathrm{Risks}^2_{\mathrm{opt}}[\mathcal{X}](1 + \mathrm{Tr}(QS)) + 6M_*^2\sqrt{\delta}, \tag{4.7.17}$$

where M_* is given by (4.4.14) *and*

$$\mathrm{Risks}_{\mathrm{opt}}[\mathcal{X}] = \inf_{\widehat{x}(\cdot)} \mathrm{RiskS}[\widehat{x}|\mathcal{X}].$$

is the minimax S-risk associated with $\mathcal{X}$.

Proof resembles the one of Lemma 4.3.2. Let $\widehat{x}(\cdot)$ be an estimate of $w = Bx$, and let R be its S-risk, so that

$$\forall (x \in \mathcal{X}) : \mathbf{E}_{\xi\sim\mathcal{N}(0,\Gamma)}\{\|\widehat{x}(Ax+\xi) - Bx\|_2^2\} \le R^2(1 + x^T Sx).$$

By exactly the same reasons as in the case of Lemma 4.3.2 we can assume w.l.o.g. that

$$R \le M_* \ \&\ \|\widehat{x}(\omega)\|_2 \le M_* \ \ \forall \omega \in \mathbf{R}^m. \tag{4.7.18}$$

Introducing Gaussian vector $[\eta;\xi]$ with independent $\xi \sim \mathcal{N}(0,\Gamma)$ and $\eta \sim \mathcal{N}(0,Q)$, and taking into account (4.7.18), we have

$$\begin{array}{rcl}
\varphi(Q) & \le & \mathbf{E}_{[\xi;\eta]}\left\{\|\widehat{x}(A\eta+\xi) - B\eta\|_2^2\right\} \text{ [by (4.2.14)]}\\
& = & \mathbf{E}_\eta\left\{\mathbf{E}_\xi\left\{\|\widehat{x}(A\eta+\xi) - B\eta\|_2^2\right\}\right\}\\
& = & \mathbf{E}_\eta\left\{\mathbf{E}_\xi\left\{\|\widehat{x}(A\eta+\xi) - B\eta\|_2^2\right\}1_{\eta\in\mathcal{X}}\right\} + \mathbf{E}_\eta\left\{\mathbf{E}_\xi\left\{\|\widehat{x}(A\eta+\xi) - B\eta\|_2^2\right\}1_{\eta\notin\mathcal{X}}\right\}\\
& \le & R^2\mathbf{E}_\eta\left\{(1+\eta^T S\eta)1_{\eta\in\mathcal{X}}\right\} + \mathbf{E}_\eta\left\{[M_* + \|B\eta\|_2]^2 1_{\eta\notin\mathcal{X}}\right\}\\
& \le & R^2\mathbf{E}_\eta\left\{(1+\eta^T S\eta)\right\} + \mathbf{E}_\eta\left\{\left[2M_*^2 + 2\|B\eta\|_2^2\right]1_{\eta\notin\mathcal{X}}\right\}\\
& \le & R^2(1+\mathrm{Tr}(QS)) + 2M_*^2\delta + 2\mathbf{E}_\eta\left\{\|B\eta\|_2^2 1_{\eta\notin\mathcal{X}}\right\}\\
& \le & R^2(1+\mathrm{Tr}(QS)) + 2M_*^2\delta + 2\sqrt{\mathbf{E}_\eta\left\{\|B\eta\|_2^4 1_{\eta\notin\mathcal{X}}\right\}}\sqrt{\mathbf{E}_\eta\left\{1_{\eta\notin\mathcal{X}}\right\}} \text{ [Cauchy Inequality]}\\
& \le & R^2(1+\mathrm{Tr}(QS)) + 2M_*^2\delta + 2\sqrt{\delta}\sqrt{\mathbf{E}_\eta\left\{\|B\eta\|_2^4\right\}}.
\end{array} \tag{4.7.19}$$

By exactly the same reasons as in the proof of Lemma 4.3.2, see Section 4.7.1.1, we have $\mathbf{E}\{\|B\eta\|_2^4\} \le 3M_*^4$, which combines with (4.7.19) to imply that

$$\varphi(Q) \le R^2(1+\mathrm{Tr}(QS)) + 2M_*^2\delta + \sqrt{3}\rho M_*^2\delta^{1/2} \le R^2(1+\mathrm{Tr}(QS)) + 6M_*^2\sqrt{\delta}$$

(recall that $\delta \in (0,1]$ and $\rho \in (0,1]$). Since R can be made arbitrarily close to $\mathrm{Risk}\mathrm{s}_{\rm opt}[\mathcal{X}]$, (4.7.17) follows. □

C. Let W, v and s stem from an optimal solution to (4.4.15). By Lemma 4.4.2 we have $s > 0$, and we can set $t = v/s$, so that $t \in \mathcal{T}$ due to $[t;s] \in \mathbf{T}$. Let also $\rho \in (0,1]$, and let us put $Q_\rho = \rho W/s$. We have $s^{-1}W \succeq 0$ and $\mathcal{R}_k[s^{-1}W] \le t_k$, $k \le K$, so that $Q_\rho \in \mathcal{Q}_\rho$. Now note that

$$\begin{array}{rcl}
\varphi(Q_\rho) & = & \mathrm{Tr}\big(B[Q_\rho - Q_\rho A^T(\Gamma + AQ_\rho A^T)^{-1}AQ_\rho]B^T\big)\\
& = & \frac{\rho}{s}\mathrm{Tr}\big(B[W - \rho WA^T(s\Gamma + \rho AWA^T)^{-1}AW]B^T\big) \ge \frac{\rho}{s}\mathrm{Opt}_* = \frac{\rho}{s}\mathrm{Opt};\\
0 < \rho \le 1 & \Rightarrow & \mathrm{Prob}_{\xi\sim\mathcal{N}(0,Q_\rho)}\{\xi \notin \mathcal{X}\} \le \delta_\rho := \min\left[2D\exp\{-\frac{1}{2\rho}\}, 1\right]
\end{array} \tag{4.7.20}$$

(we have used (4.4.17), the positivity of s, and (4.7.16)). Thus, when applying Lemma 4.7.4 with Q_ρ and δ_ρ in the role of Q and δ, we obtain for all $0 < \rho \le 1$:

$$\frac{\rho}{s}\mathrm{Opt} \ \le \ \mathrm{Risk}\mathrm{s}^2_{\rm opt}[\mathcal{X}]\,(1 + \mathrm{Tr}(Q_\rho S)) + 6M_*^2\sqrt{2D}\exp\{-\tfrac{1}{4\rho}\}. \tag{4.7.21}$$

Setting

$$\bar{\rho}^{-1} = 4\ln\left(\frac{6M_*^2\sqrt{2D}}{\mathrm{Risk}\mathrm{s}^2_{\rm opt}[\mathcal{X}]}\right)$$

and taking into account that by evident reasons one has $\mathrm{Risk}\mathrm{s}^2_{\rm opt}[\mathcal{X}] \le M_*^2$, we get $\bar{\rho} < 1$ and

$$6M_*^2\sqrt{2D}\exp\left\{-\frac{1}{4\bar{\rho}}\right\} \le \mathrm{Risk}\mathrm{s}^2_{\rm opt}[\mathcal{X}],$$

whence

$$\frac{\bar{\rho}}{s}\mathrm{Opt} \le 2\mathrm{Risk}\mathrm{s}^2_{\rm opt}[\mathcal{X}]\,(1 + \mathrm{Tr}(Q_{\bar{\rho}}S))$$

by (4.7.21). Consequently,

$$\bar{\rho}\mathrm{Opt} \quad \leq \quad 2\mathrm{Risks}^2_{\mathrm{opt}}[\mathcal{X}]\left(s+\bar{\rho}\mathrm{Tr}(WS)\right) \leq 2\mathrm{Risks}^2_{\mathrm{opt}}[\mathcal{X}]$$

(note that $s+\bar{\rho}\mathrm{Tr}(WS) \leq s+\mathrm{Tr}(WS) \leq 1$ by constraints in (4.4.15)). Recalling that $\sqrt{\mathrm{Opt}}$ upper-bounds $\mathrm{RiskS}[\widehat{x}_{H_*}|\mathcal{X}]$, the conclusion of Proposition follows. □

4.7.3.5 Proof of Proposition 4.4.3

Proof follows the lines of the proof of Proposition 4.3.3. First, passing from C to the matrix $\bar{C}=P^TCP$, the situation clearly reduces to the one where $P=I$, which we assume in the sequel. Second, the same arguments as in the proof of Proposition 4.4.1 demonstrate that problem (4.4.18) is solvable, and the left inequality in (4.4.19) is readily given by Lemma 4.4.1. Thus, all we need is to prove the right inequality in (4.4.19).

1^o. Consider the conic problem

$$\mathrm{Opt}_{\#}=\max_{Q,t}\left\{\mathrm{Tr}(\bar{C}Q): Q\succeq 0, \mathcal{R}_k[Q]\preceq t_kI_{d_k}\,\forall k\leq K, [t;1]\in\mathbf{T}\right\}. \tag{4.7.22}$$

This problem clearly is strictly feasible; by the same argument as in item 1^o of the proof of Lemma 4.4.2, the feasible set of the problem is bounded, so that the problem is solvable. We claim that

$$\mathrm{Opt}_{\#}=\mathrm{Opt}_*. \tag{4.7.23}$$

Indeed, (4.7.22) is a strictly feasible and bounded conic problem, so that its optimal value is equal to the one in its conic dual, and this dual is solvable, that is,

$$\begin{array}{rcl}\mathrm{Opt}_{\#} & = & \min\limits_{\Lambda=\{\Lambda_k\}_{k\leq K},[g;s],L}\left\{s: \begin{array}{r}\mathrm{Tr}([\sum_k\mathcal{R}^*_k[\Lambda_k]-L]Q)-\sum_k[\mathrm{Tr}(\Lambda_k)+g_k]t_k\\ =\mathrm{Tr}(\bar{C}Q)\ \ \forall (Q,t),\\ \Lambda_k\succeq 0\,\forall k, L\succeq 0, s\geq\phi_{\mathcal{T}}(-g)\quad\quad\end{array}\right\}\\ & = & \min\limits_{\Lambda,[g;s],L}\left\{s: \begin{array}{l}\sum_k\mathcal{R}^*_k[\Lambda_k]-L=\bar{C}, g=-\lambda[\Lambda],\\ \Lambda_k\succeq 0\,\forall k, L\succeq 0, s\geq\phi_{\mathcal{T}}(-g)\end{array}\right\}\\ & = & \min\limits_{\Lambda}\left\{\phi_{\mathcal{T}}(\lambda[\Lambda]):\sum_k\mathcal{R}^*_k[\Lambda_k]\succeq\bar{C}, \Lambda_k\succeq 0\,\forall k\right\}=\mathrm{Opt}_*.\end{array}$$

and (4.4.18) is solvable along with conic dual to problem (4.7.22).

2^o. Problem (4.7.22), as we already know, is solvable; let Q_*,t^* be an optimal solution to the problem. Next, let us set $R_*=Q_*^{1/2}$, $\widehat{C}=R_*\bar{C}R_*$, and let $\widehat{C}=UDU^T$ be the eigenvalue decomposition of $\widehat{C}$, so that the matrix $D=U^TR_*\bar{C}R_*U$ is diagonal, and the trace of this matrix is $\mathrm{Tr}(R_*\bar{C}R_*)=\mathrm{Tr}(\bar{C}Q_*)=\mathrm{Opt}_{\#}=\mathrm{Opt}_*$, Now let $V=R_*U$, and let $\xi=V\eta$, where η is n-dimensional random Rademacher vector (independent entries taking values ±1 with probabilities 1/2). We have

$$\xi^T\bar{C}\xi=\eta^T[V^T\bar{C}V]\eta=\eta^T[U^TR_*\bar{C}R_*U]\eta=\eta^TD\eta\equiv\mathrm{Tr}(D)=\mathrm{Opt}_*, \tag{4.7.24}$$

(recall that D is diagonal) and

$$\mathbf{E}_\xi\{\xi\xi^T\}=\mathbf{E}_\eta\{V\eta\eta^TV^T\}=VV^T=R_*UU^TR_*=R_*^2=Q_*.$$

From the latter relation,

$$\mathbf{E}_\xi\left\{R_k^2[\xi]\right\}=\mathbf{E}_\xi\left\{\mathcal{R}_k[\xi\xi^T]\right\}=\mathcal{R}_k[\mathbf{E}_\xi\{\xi\xi^T\}]=\mathcal{R}_k[Q_*]\preceq t^*_kI_{d_k}, 1\leq k\leq K. \tag{4.7.25}$$

On the other hand, with properly selected symmetric matrices $\bar{R}^{kj}$ we have

$$R_k[Vy] = \sum_i \bar{R}^{ki} y_i$$

identically in $y \in \mathbf{R}^n$, whence

$$\mathbf{E}_\xi\left\{R_k^2[\xi]\right\} = \mathbf{E}_\eta\left\{R_k^2[V\eta]\right\} = \mathbf{E}_\eta\left\{\left[\sum\nolimits_i \eta_i \bar{R}^{ki}\right]^2\right\} = \sum_{i,j} \mathbf{E}_\eta\{\eta_i\eta_j\}\bar{R}^{ki}\bar{R}^{kj} = \sum_i [\bar{R}^{ki}]^2.$$

This combines with (4.7.25) to imply that

$$\sum_i [\bar{R}^{ki}]^2 \preceq t_k^* I_{d_k},\ 1 \leq k \leq K. \tag{4.7.26}$$

3^o. Let us fix $k \leq K$. Applying Theorem 4.7.1, we derive from (4.7.26) that

$$\mathrm{Prob}\{\eta : \|\bar{R}_k[\eta]\|^2 > t_k^*/\rho\} < 2d_k \mathrm{e}^{-\frac{1}{2\rho}},$$

and recalling the relation between ξ and η, we arrive at

$$\mathrm{Prob}\{\xi : \|R_k[\xi]\|^2 > t_k^*/\rho\} < 2d_k \mathrm{e}^{-\frac{1}{2\rho}}\ \ \forall \rho \in (0,1]. \tag{4.7.27}$$

Note that when $t_k^* = 0$ (4.7.26) implies $\bar{R}^{ki} = 0$ for all i, so that $R_k[\xi] = \bar{R}_k[\eta] = 0$, and (4.7.27) also holds for those k.

Now let us set $\rho = \frac{1}{2\max[\ln(2D),1]}$. For this ρ, the sum over $k \leq K$ of the right hand sides in inequalities (4.7.27) is ≤ 1, implying that there exists a realization $\bar{\xi}$ of ξ such that

$$\|R_k[\bar{\xi}]\|^2 \leq t_k^*/\rho,\ \forall k,$$

or, equivalently,

$$\bar{x} := \rho^{1/2} P\bar{\xi} \in \mathcal{X},$$

implying that

$$\mathrm{Opt} \geq \bar{x}^T C \bar{x} = \rho\xi^T \bar{C}\xi = \rho \mathrm{Opt}_*$$

(the concluding equality is due to (4.7.24)), and we arrive at the right inequality in (4.4.19). □

4.7.4 Proofs for Section 4.5

4.7.4.1 Proof of Lemma 4.5.2

1^o. Let us verify (4.5.31). When $Q \succ 0$, passing from variables (Θ, Υ) in problem (4.5.30) to the variables $(G = Q^{1/2}\Theta Q^{1/2}, \Upsilon)$, the problem becomes exactly the optimization problem in (4.5.31), implying that $\mathrm{Opt}[Q] = \overline{\mathrm{Opt}}[Q]$ when $Q \succ 0$. As it is easily seen, both sides in this equality are continuous in $Q \succeq 0$, and (4.5.31) follows.

2^o. Let us set $\zeta = Q^{1/2}\eta$ with $\eta \sim \mathcal{N}(0, I_N)$ and $Z = Q^{1/2}Y$. All we need to complete the proof of Lemma 4.5.2 is to show that the quantity

$$[\overline{\mathrm{Opt}}[Q] =]\quad \mathrm{Opt} := \min_{\Theta, \Upsilon = \{\Upsilon_\ell, \ell \leq L\}} \left\{\phi_{\mathcal{R}}(\lambda[\Upsilon]) + \mathrm{Tr}(\Theta) : \Upsilon_\ell \succeq 0, \left[\begin{array}{c|c} \Theta & \frac{1}{2}ZM \\ \hline \frac{1}{2}M^T Z^T & \sum_\ell \mathcal{S}_\ell^*[\Upsilon_\ell] \end{array}\right] \succeq 0\right\} \tag{4.7.28}$$

satisfies

$$\mathrm{Opt} \leq \frac{8\sqrt{\ln\left(\frac{4\sqrt{2}F}{\sqrt{2}-\mathrm{e}^{1/4}}\right)}}{\sqrt{2} - \mathrm{e}^{1/4}} \psi_I(Z),\ \ \psi_I(Z) = \mathbf{E}_{\eta \sim \mathcal{N}(0, I_N)}\{\|Z^T\eta\|\}. \tag{4.7.29}$$

3^o. Let us represent Opt as the optimal value of a conic problem. Setting

$$\mathbf{K}=\mathbf{K}[\mathcal{R}]=\mathrm{cl}\{[r;s]:s>0,r/s\in\mathcal{R}\},$$

we ensure that

$$\mathcal{R}=\{r:[r;1]\in\mathbf{K}\},\ \mathbf{K}_*=\{[g;s]:s\geq\phi_{\mathcal{R}}(-g)\},$$

where $\mathbf{K}_*$ is the cone dual to $\mathbf{K}$. Consequently, (4.7.28) reads

$$\mathrm{Opt}=\min_{\Theta,\Upsilon,\theta}\left\{\theta+\mathrm{Tr}(\Theta):\begin{array}{lr}\Upsilon_\ell\succeq 0, 1\leq\ell\leq L & (a)\\ \left[\begin{array}{c|c}\Theta & \frac{1}{2}ZM\\ \hline \frac{1}{2}M^TZ^T & \sum_\ell\mathcal{S}^*_\ell[\Upsilon_\ell]\end{array}\right]\succeq 0 & (b)\\ [-\lambda[\Upsilon];\theta]\in\mathbf{K}_* & (c)\end{array}\right\}. \tag{P}$$

4^o. Now let us prove that there exists matrix $W\in\mathbf{S}^q_+$ and $r\in\mathcal{R}$ such that

$$\mathcal{S}_\ell[W]\preceq r_\ell I_{f_\ell},\ \ell\leq L, \tag{4.7.30}$$

and

$$\mathrm{Opt}\leq\sum_i\sigma_i(ZMW^{1/2}), \tag{4.7.31}$$

where $\sigma_1(\cdot)\geq\sigma_2(\cdot)\geq...$ are singular values.

To get the announced result, let us pass from problem (P) to its conic dual. Applying Lemma 4.7.2 we conclude that (P) is strictly feasible; in addition, (P) clearly is bounded, so that the dual to (P) problem (D) is solvable with optimal value Opt. Let us build (D). Denoting by $\Lambda_\ell\succeq 0,\ell\leq L$, $\left[\begin{array}{c|c}G & -R\\ \hline -R^T & W\end{array}\right]\succeq 0$, $[r;\tau]\in\mathbf{K}$ the Lagrange multipliers for the respective constraints in (P), and aggregating these constraints, the multipliers being the aggregation weights, we arrive at the following aggregated constraint:

$$\mathrm{Tr}(\Theta G)+\mathrm{Tr}(W\textstyle\sum_\ell\mathcal{S}^*_\ell[\Upsilon_\ell])+\sum_\ell\mathrm{Tr}(\Lambda_\ell\Upsilon_\ell)-\sum_\ell r_\ell\mathrm{Tr}(\Upsilon_\ell)+\theta\tau\geq\mathrm{Tr}(ZMR^T).$$

To get the dual problem, we impose on the Lagrange multipliers, in addition to the initial conic constraints like $\Lambda_\ell\succeq 0$, $1\leq\ell\leq L$, the restriction that the left hand side in the aggregated constraint, identically in Θ, Υ_ℓ and θ, is equal to the objective of (P), that is,

$$G=I,\ \mathcal{S}_\ell[W]+\Lambda_\ell-r_\ell I_{f_\ell}=0,\ 1\ \leq\ell\leq L,\ \tau=1,$$

and maximize, under the resulting restrictions, the right-hand side of the aggregated constraint. After immediate simplifications, we arrive at

$$\mathrm{Opt}=\max_{W,R,r}\left\{\mathrm{Tr}(ZMR^T):\ W\succeq R^TR, r\in\mathcal{R},\mathcal{S}_\ell[W]\preceq r_\ell I_{f_\ell}, 1\leq\ell\leq L\right\}$$

(note that $r\in\mathcal{R}$ is equivalent to $[r;1]\in\mathbf{K}$, and $W\succeq R^TR$ is the same as $\left[\begin{array}{c|c}I & -R\\ \hline -R^T & W\end{array}\right]\succeq 0$). Now, to say that $R^TR\preceq W$ is exactly the same as to say that $R=SW^{1/2}$ with the spectral norm $\|S\|_{\mathrm{Sh},\infty}$ of S not exceeding 1, so that

$$\mathrm{Opt}=\max_{W,S,r}\left\{\underbrace{\mathrm{Tr}([ZM[SW^{1/2}]^T)}_{=\mathrm{Tr}([ZMW^{1/2}]S^T)}:W\succeq 0,\|S\|_{\mathrm{Sh},\infty}\leq 1,r\in\mathcal{R},\mathcal{S}_\ell[W]\preceq r_\ell I_{f_\ell},\ell\leq L\right\}$$

and we can immediately eliminate the S-variable, using the well-known fact that for every $p\times q$ matrix J, it holds

$$\max_{S\in\mathbf{R}^{p\times q},\|S\|_{\mathrm{Sh},\infty}\leq 1}\mathrm{Tr}(JS^T)=\|J\|_{\mathrm{Sh},1},$$

where $\|J\|_{\mathrm{Sh},1}$ is the nuclear norm (the sum of singular values) of J. We arrive at

$$\mathrm{Opt}=\max_{W,r}\left\{\|ZMW^{1/2}\|_{\mathrm{Sh},1}:r\in\mathcal{R},W\succeq 0,\mathcal{S}_\ell[W]\preceq r_\ell I_{d_\ell},\,\ell\leq L\right\}.$$

The resulting problem clearly is solvable, and its optimal solution W ensures the target relations (4.7.30), (4.7.31).

5^o. Given W satisfying (4.7.30), (4.7.31), let $UJV=W^{1/2}M^TZ^T$ be the singular value decomposition of $W^{1/2}M^TZ^T$, so that U and V are, respectively, $q\times q$ and $N\times N$ orthogonal matrices, J is $q\times N$ matrix with diagonal $\sigma=[\sigma_1;...;\sigma_p]$, $p=\min[q,N]$, and zero off-diagonal entries; the diagonal entries σ_i, $1\leq i\leq p$ are the singular values of $W^{1/2}M^TZ^T$, or, which is the same, of $ZMW^{1/2}$, so we have

$$\sum_i\sigma_i\geq\mathrm{Opt}.\tag{4.7.32}$$

Now consider the following construction. Let $\eta\sim\mathcal{N}(0,I_N)$; we denote by υ the vector comprised of the first p entries in $V\eta$; note that $\upsilon\sim\mathcal{N}(0,I_p)$, since V is orthogonal. We then augment, if necessary, υ by $q-p$ independent $\mathcal{N}(0,1)$ random variables to obtain a q-dimensional normal vector $\upsilon'\sim\mathcal{N}(0,I_q)$, and set $\chi=U\upsilon'$; because U is orthogonal we also have $\chi\sim\mathcal{N}(0,I_q)$. Observe that

$$\chi^TW^{1/2}M^TZ^T\eta=\chi^TUJV\eta=[\upsilon']^TJ\upsilon=\sum_{i=1}^p\sigma_i\upsilon_i^2.\tag{4.7.33}$$

To continue we need the following simple observations.

1. *One has*

$$\alpha:=\mathrm{Prob}\left\{\sum_{i=1}^p\sigma_i\upsilon_i^2<\frac{1}{2}\sum_{i=1}^p\sigma_i\right\}\leq\frac{\mathrm{e}^{1/4}}{\sqrt{2}}\ [<1].\tag{4.7.34}$$

The claim is evident when $\sigma:=\sum_i\sigma_i=0$. Now let $\sigma>0$, and let us apply the Bernstein bounding scheme. Namely, given $\gamma>0$, consider the random variable

$$\omega=\exp\left\{\frac{1}{2}\gamma\sum_i\sigma_i-\gamma\sum_i\sigma_i\upsilon_i^2\right\}.$$

Note that $\omega>0$ a.s., and is >1 when $\sum_{i=1}^p\sigma_i\upsilon_i^2<\frac{1}{2}\sum_{i=1}^p\sigma_i$, so that $\alpha\leq\mathbf{E}\{\omega\}$, or, equivalently, thanks to $\upsilon\sim\mathcal{N}(0,I_p)$,

$$\ln(\alpha)\leq\ln(\mathbf{E}\{\omega\})=\frac{1}{2}\gamma\sum_i\sigma_i+\sum_i\ln\left(\mathbf{E}\{\exp\{-\gamma\sigma_i\upsilon_i^2\}\}\right)\leq\frac{1}{2}\left[\gamma\sigma-\sum_i\ln(1+2\gamma\sigma_i)\right].$$

Function $-\sum_i\ln(1+2\gamma\sigma_i)$ is convex in $[\sigma_1;...;\sigma_p]\geq 0$, therefore, its maximum over the simplex $\{\sigma_i\geq 0,i\leq p,\sum_i\sigma_i=\sigma\}$ is attained at a vertex, and we get

$$\ln(\alpha)\leq\frac{1}{2}\left[\gamma\sigma-\ln(1+2\gamma\sigma)\right].$$

Minimizing the right hand side in $\gamma>0$, we arrive at (4.7.34).

2. *Whenever* $\varkappa \geq 1$, *one has*

$$\mathrm{Prob}\{\|MW^{1/2}\chi\|_* > \varkappa\} \leq 2F\exp\{-\varkappa^2/2\}, \tag{4.7.35}$$

with F *given by* (4.5.29).

Indeed, setting $\rho = 1/\varkappa^2 \leq 1$ and $\omega = \sqrt{\rho}W^{1/2}\chi$, we get $\omega \sim \mathcal{N}(0, \rho W)$. Let us apply Lemma 4.7.3 to $Q = \rho W$ and to $\mathcal{R}$ in the role of $\mathcal{T}$, L in the role of K, and $\mathcal{S}_\ell[\cdot]$ in the role of $\mathcal{R}_k[\cdot]$. Denoting

$$\mathcal{Y} := \{y : \exists r \in \mathcal{R} : S_\ell^2[y] \preceq r_\ell I_{f_\ell}, \ell \leq L\},$$

we have $\mathcal{S}_\ell[Q] = \rho\mathcal{S}_\ell[W] \preceq \rho r_\ell I_{f_\ell}$, $\ell \leq L$, with $r \in \mathcal{R}$ (see (4.7.30)), so we are under the premise of Lemma 4.7.3. Applying the lemma, we conclude that

$$\mathrm{Prob}\left\{\chi : \varkappa^{-1}W^{1/2}\chi \notin \mathcal{Y}\right\} \leq 2F\exp\{-1/(2\rho)\} = 2F\exp\{-\varkappa^2/2\}.$$

Recalling that $\mathcal{B}_* = M\mathcal{Y}$, we see that $\mathrm{Prob}\{\chi : \varkappa^{-1}MW^{1/2}\chi \notin \mathcal{B}_*\}$ is indeed upper-bounded by the right hand size of (4.7.35), and (4.7.35) follows.

3. *For* $\varkappa \geq 1$, *let*

$$E_\varkappa = \left\{(\chi, \eta) : \ \|MW^{1/2}\chi\|_* \leq \varkappa, \ \sum_i \sigma_i \upsilon_i^2 \geq \frac{1}{2}\sum_i \sigma_i\right\}.$$

Then one has

$$\mathrm{Prob}\{E_\varkappa\} \geq \beta(\varkappa) := 1 - \frac{\mathrm{e}^{1/4}}{\sqrt{2}} - 2F\exp\{-\varkappa^2/2\}. \tag{4.7.36}$$

Indeed, relation (4.7.36) follows from (4.7.34), (4.7.35) due to the union bound.

When $(\chi, \eta) \in E_\varkappa$, we have

$$\varkappa\|Z^T\eta\| \geq \|MW^{1/2}\chi\|_*\|Z^T\eta\| \geq \chi^T W^{1/2}M^T Z^T\eta = \sum_i \sigma_i\upsilon_i^2 \geq \frac{1}{2}\sum_i \sigma_i \geq \frac{1}{2}\mathrm{Opt},$$

(we have used (4.7.33) and (4.7.32)), so that whenever $(\chi, \eta) \in E_\varkappa$ one has $\|Z^T\eta\| \geq \frac{1}{2\varkappa}\mathrm{Opt}$. Hence, finally,

$$2\mathbf{E}_{\eta\sim\mathcal{N}(0,I_N)}\{\|Z^T\eta\|\} \geq \mathrm{Prob}\{(\chi,\eta) \in E_\varkappa\}\varkappa^{-1}\mathrm{Opt} \geq \left[1 - \frac{\mathrm{e}^{1/4}}{\sqrt{2}} - 2F\exp\{-\varkappa^2/2\}\right]\varkappa^{-1}\mathrm{Opt},$$

and we arrive at (4.7.29) when specifying $\varkappa$ as

$$\varkappa = \sqrt{2\ln\left(\frac{4\sqrt{2}F}{\sqrt{2} - \mathrm{e}^{1/4}}\right)}. \qquad \square$$

4.7.4.2 Proof of Proposition 4.5.3

1⁰. Let

$$\Phi(H, \Lambda, \Upsilon, \Upsilon', \Theta; Q) = \phi_{\mathcal{T}}(\lambda[\Lambda]) + \phi_{\mathcal{R}}(\lambda[\Upsilon]) + \phi_{\mathcal{R}}(\lambda[\Upsilon']) + \mathrm{Tr}(Q\Theta) : \mathcal{M} \times \Pi \to \mathbf{R},$$

$$\mathcal{M} = \left\{(H, \Lambda, \Upsilon, \Upsilon', \Theta) : \begin{array}{c} \Lambda = \{\Lambda_k \succeq 0, k \leq K\}, \ \Upsilon = \{\Upsilon_\ell \succeq 0, \ell \leq L\}, \ \Upsilon' = \{\Upsilon'_\ell \succeq 0, \ell \leq L\} \\ \left[\begin{array}{c|c} \sum_k \mathcal{R}_k^*[\Lambda_k] & \frac{1}{2}[B^T - A^T H]M \\ \hline \frac{1}{2}M^T[B - H^T A] & \sum_\ell \mathcal{S}_\ell^*[\Upsilon_\ell] \end{array}\right] \succeq 0 \\ \left[\begin{array}{c|c} \Theta & \frac{1}{2}HM \\ \hline \frac{1}{2}M^T H^T & \sum_\ell \mathcal{S}_\ell^*[\Upsilon'_\ell] \end{array}\right] \succeq 0 \end{array}\right\} \tag{4.7.37}$$

Looking at (4.5.15), we conclude immediately that the optimal value Opt in (4.5.15) is nothing but

$$\mathrm{Opt} = \min_{(H,\Lambda,\Upsilon,\Upsilon',\Theta)\in\mathcal{M}} \left[\overline{\Phi}(H,\Lambda,\Upsilon,\Upsilon',\Theta) := \max_{Q\in\Pi} \Phi(H,\Lambda,\Upsilon,\Upsilon',\Theta;Q)\right]. \tag{4.7.38}$$

Note that the sets $\mathcal{M}$ and Π are closed and convex, Π is compact, and Φ is a continuous convex-concave function on $\mathcal{M}\times\Pi$. In view of these observations, Assumption **R** combines with Sion-Kakutani Theorem to imply that Φ possesses saddle point $(H_*,\Lambda_*,\Upsilon_*,\Upsilon'_*,\Theta_*;Q_*)$ (min in $(H,\Lambda,\Upsilon,\Upsilon',\Theta)$, max in Q) on $\mathcal{M}\times\Pi$, whence Opt is the saddle point value of Φ by (4.7.38). We conclude that for properly selected $Q_*\in\Pi$ it holds

$$\begin{array}{rcl}
\mathrm{Opt} & = & \min\limits_{(H,\Lambda,\Upsilon,\Upsilon',\Theta)\in\mathcal{M}} \Phi(H,\Lambda,\Upsilon,\Upsilon',\Theta;Q_*)\\
& = & \min\limits_{H,\Lambda,\Upsilon,\Upsilon',\Theta} \left\{ \begin{array}{l}\phi_{\mathcal{T}}(\lambda[\Lambda]) + \phi_{\mathcal{R}}(\lambda[\Upsilon]) + \phi_{\mathcal{R}}(\lambda[\Upsilon']) + \mathrm{Tr}(Q_*\Theta) :\\
\qquad \Lambda=\{\Lambda_k\succeq 0, k\leq K\},\ \Upsilon=\{\Upsilon_\ell\succeq 0,\ell\leq L\},\ \Upsilon'=\{\Upsilon'_\ell\succeq 0,\ell\leq L\}\\
\qquad \left[\begin{array}{c|c}\sum_k \mathcal{R}_k^*[\Lambda_k] & \frac{1}{2}[B^T-A^TH]M\\ \hline \frac{1}{2}M^T[B-H^TA] & \sum_\ell \mathcal{S}_\ell^*[\Upsilon_\ell]\end{array}\right]\succeq 0,\\
\qquad \left[\begin{array}{c|c}\Theta & \frac{1}{2}HM\\ \hline \frac{1}{2}M^TH^T & \sum_\ell \mathcal{S}_\ell^*[\Upsilon'_\ell]\end{array}\right]\succeq 0\end{array}\right\}\\
& = & \min\limits_{H,\Lambda,\Upsilon,\Upsilon',G} \left\{ \begin{array}{l}\phi_{\mathcal{T}}(\lambda[\Lambda]) + \phi_{\mathcal{R}}(\lambda[\Upsilon]) + \phi_{\mathcal{R}}(\lambda[\Upsilon']) + \mathrm{Tr}(G) :\\
\qquad \Lambda=\{\Lambda_k\succeq 0, k\leq K\},\ \Upsilon=\{\Upsilon_\ell\succeq 0,\ell\leq L\},\ \Upsilon'=\{\Upsilon'_\ell\succeq 0,\ell\leq L\}\\
\qquad \left[\begin{array}{c|c}\sum_k \mathcal{R}_k^*[\Lambda_k] & \frac{1}{2}[B^T-A^TH]M\\ \hline \frac{1}{2}M^T[B-H^TA] & \sum_\ell \mathcal{S}_\ell^*[\Upsilon_\ell]\end{array}\right]\succeq 0,\\
\qquad \left[\begin{array}{c|c}G & \frac{1}{2}Q_*^{1/2}HM\\ \hline \frac{1}{2}M^TH^TQ_*^{1/2} & \sum_\ell \mathcal{S}_\ell^*[\Upsilon'_\ell]\end{array}\right]\succeq 0\end{array}\right\}\\
& = & \min\limits_{H,\Lambda,\Upsilon} \left\{\phi_{\mathcal{T}}(\lambda[\Lambda]) + \phi_{\mathcal{R}}(\lambda[\Upsilon]) + \overline{\Psi}(H) : \begin{array}{l}\Lambda=\{\Lambda_k\succeq 0, k\leq K\},\ \Upsilon=\{\Upsilon_\ell\succeq 0,\ell\leq L\}\\ \left[\begin{array}{c|c}\sum_k \mathcal{R}_k^*[\Lambda_k] & \frac{1}{2}[B^T-A^TH]M\\ \hline \frac{1}{2}M^T[B-H^TA] & \sum_\ell \mathcal{S}_\ell^*[\Upsilon_\ell]\end{array}\right]\succeq 0\end{array}\right\},\\
\overline{\Psi}(H) & := & \min\limits_{G,\Upsilon'} \left\{\phi_{\mathcal{R}}(\lambda[\Upsilon']) + \mathrm{Tr}(G) : \Upsilon'=\{\Upsilon'_\ell\succeq 0,\ell\leq L\},\ \left[\begin{array}{c|c}G & \frac{1}{2}Q_*^{1/2}HM\\ \hline \frac{1}{2}M^TH^TQ_*^{1/2} & \sum_\ell \mathcal{S}_\ell^*[\Upsilon'_\ell]\end{array}\right]\succeq 0\right\}
\end{array} \tag{4.7.39}$$

where Opt is given by (4.5.15), and the equalities are due to (4.5.30) and (4.5.31).

2^o. From now on we assume that the observation noise ξ in observation (4.5.2) is $\xi\sim\mathcal{N}(0,Q_*)$. Besides this, we assume that $B\neq 0$, since otherwise the conclusion of Proposition 4.5.3 is evident.

3^o. Let W be a positive semidefinite $n\times n$ matrix, let $\eta\sim\mathcal{N}(0,W)$ be random signal, and let $\xi\sim\mathcal{N}(0,Q_*)$ be independent of η; vectors (η,ξ) induce random vector

$$\omega = A\eta+\xi\sim\mathcal{N}(0,AWA^T+Q_*).$$

Now, consider the problem where given ω we are interested to recover $B\eta$, and the Bayesian risk of a candidate estimate $\widehat{x}(\cdot)$ is quantified by $\mathbf{E}_{\eta,\xi}\{\|B\eta-\widehat{x}(A\eta+\xi)\|\}$. Let us set

$$\varrho[W] = \inf_{\widehat{x}(\cdot)} \mathbf{E}_{\eta,\xi}\{\|B\eta-\widehat{x}(A\eta+\xi)\|\}. \tag{4.7.40}$$

Our first observation is that $\varrho[W]$ is "nearly attainable" with a linear estimate. Since $(\omega,B\eta)$ is zero mean Gaussian, the conditional expectation $\mathbf{E}_{|\omega}\{B\eta\}$ of $B\eta$ given ω is linear in ω: $\mathbf{E}_{|\omega}\{B\eta\}=\bar{H}^T\omega$

for some $\bar{H}$ depending on W only. Given an estimate $\widehat{x}(\cdot)$, its Bayesian risk satisfies

$$\varrho = \mathbf{E}_{\eta,\omega}\{\|B\eta - \widehat{x}(\omega)\|\} = \mathbf{E}_\omega\left\{\mathbf{E}_{|\omega}\{\|B\eta - \widehat{x}(\omega)\|\}\right\} \geq \mathbf{E}_\omega\{\|\underbrace{\mathbf{E}_{|\omega}\{B\eta\}}_{=\bar{H}^T\omega} - \widehat{x}(\omega)\|\}$$

by the Jensen inequality. Hence

$$\begin{array}{rcl}\mathbf{E}_{\eta,\xi}\{\|B\eta - \bar{H}^T(A\eta+\xi)\|\} & = & \mathbf{E}_{\eta,\omega}\{\|B\eta - \bar{H}^T\omega\|\} = \mathbf{E}_\omega\{\mathbf{E}_{|\omega}\{\|B\eta - \bar{H}^T\omega\|\}\} \\ & \leq & \mathbf{E}_\omega\left\{\mathbf{E}_{|\omega}\{\|B\eta - \widehat{x}(\omega)\| + \|\widehat{x}(\omega) - \bar{H}^T\omega\|\}\right\} \\ & = & \mathbf{E}_{\eta,\omega}\{\|B\eta - \widehat{x}(\omega)\|\} + \mathbf{E}_\omega\{\|\bar{H}^T\omega - \widehat{x}(\omega)\|\} \leq 2\varrho,\end{array}$$

and thus

$$\mathbf{E}_{\eta,\xi}\{\|B\eta - \bar{H}^T(A\eta+\xi)\|\} = \mathbf{E}_{\eta,\omega}\{\|B\eta - \bar{H}^T\omega\|\} \leq 2\varrho. \tag{4.7.41}$$

(4.7.41) combines with independence of ξ, η and Jensen's inequality to imply that

$$2\varrho \geq \mathbf{E}_\eta\{\mathbf{E}_\xi\{\|B\eta - \bar{H}^T(A\eta+\xi)\|\}\} \geq \mathbf{E}_\eta\{\|\mathbf{E}_\xi\{B\eta - \bar{H}^T(A\eta+\xi)\}\|\} = \mathbf{E}_\eta\{\|(B - \bar{H}^TA)\eta\|\},$$

that is,

$$\mathbf{E}_\eta\{\|(B - \bar{H}^TA)\eta\|\} \leq 2\varrho. \tag{4.7.42}$$

By "symmetric" reasoning,

$$\mathbf{E}_\xi\{\|\bar{H}^T\xi\|\} \leq 2\varrho. \tag{4.7.43}$$

In relations (4.7.42) and (4.7.43), $\bar{H}$ depends solely on W, and ϱ can be made arbitrarily close to $\varrho[W]$, and we arrive at the following

Lemma 4.7.5 *Let W be a positive semidefinite $n\times n$ matrix. Then the risk $\varrho[W]$ defined by* (4.7.40) *satisfies the inequality*

$$\varrho[W] \geq \tfrac{1}{4}\inf_{H\in\mathbf{R}^{m\times\nu}}\left[\mathbf{E}_{\eta\sim\mathcal{N}(0,W)}\{\|[B - H^TA]\eta\|\} + \mathbf{E}_{\xi\sim\mathcal{N}(0,Q_*)}\{\|H^T\xi\|\}\right]. \tag{4.7.44}$$

4^o. Lemma 4.7.5 combines with Lemma 4.5.2 to imply the following result:

Lemma 4.7.6 *Let W be a positive semidefinite $n\times n$ matrix. Then the risk $\varrho[W]$ defined by* (4.7.40) *satisfies the inequality*

$$\begin{array}{rcl}\varrho[W] & \geq & (4\kappa[F])^{-1}\min\limits_{\Upsilon=\{\Upsilon_\ell,\ell\leq L\},G,H}\left\{\mathrm{Tr}\,(WG) + \phi_{\mathcal{R}}(\lambda[\Upsilon]) + \overline{\Psi}(H):\right. \\ & & \left.\Upsilon_\ell\succeq 0\,\forall\ell,\left[\begin{array}{c|c} G & \frac{1}{2}[B^T - A^TH]M \\ \hline \frac{1}{2}M^T[B - H^TA] & \sum_\ell \mathcal{S}_\ell^*[\Upsilon_\ell]\end{array}\right]\succeq 0\right\},\end{array} \tag{4.7.45}$$

where $\overline{\Psi}(H)$ is given by (4.7.39) *and $\kappa[F] = \frac{8}{\sqrt{2}-\mathrm{e}^{1/4}}\sqrt{\ln\left(\frac{4\sqrt{2}F}{\sqrt{2}-\mathrm{e}^{1/4}}\right)}$.*

Proof. Let H be $m\times\nu$ matrix. Applying Lemma 4.5.2 to $N=m$, $Y=H$, $Q=Q_*$, we get

$$\mathbf{E}_{\xi\sim\mathcal{N}(0,Q_*)}\{\|H^T\xi\|\} \geq \kappa^{-1}[F]\,\overline{\Psi}(H). \tag{4.7.46}$$

Applying Lemma 4.5.2 to $N=n$, $Y=(B-H^TA)^T$, $Q=W$, we get

$$\begin{array}{l}\kappa[F]\,\mathbf{E}_{\eta\sim\mathcal{N}(0,W)}\{\|[B - H^TA]\eta\|\} \\ \geq \min\limits_{\Upsilon=\{\Upsilon_\ell\succ 0,\ell\leq L\},G}\left\{\phi_{\mathcal{R}}(\lambda[\Upsilon]) + \mathrm{Tr}(WG):\left[\begin{array}{c|c} G & \frac{1}{2}[B^T - A^TH]M \\ \hline \frac{1}{2}M^T[B - H^TA] & \sum_\ell \mathcal{S}_\ell^*[\Upsilon_\ell]\end{array}\right]\succeq 0\right\}.\end{array}$$

The resulting inequality combines with (4.7.44) and (4.7.46) to imply (4.7.45). □

5^o. For $0<\varkappa\leq 1$, let us set

$$\begin{array}{lrcl}(a) & \mathcal{W}_\varkappa & = & \{W\in\mathbf{S}^n_+ : \exists t\in\mathcal{T}: \mathcal{R}_k[W]\preceq \varkappa t_k I_{d_k}, 1\leq k\leq K\},\\ (b) & \mathcal{Z} & = & \left\{(\Upsilon=\{\Upsilon_\ell,\ell\leq L\},G,H): \begin{array}{c}\Upsilon_\ell\succeq 0\,\forall \ell,\\ \left[\begin{array}{c|c} G & \frac{1}{2}[B^T-A^TH]M\\ \hline \frac{1}{2}M^T[B-H^TA] & \sum_\ell \mathcal{S}^*_\ell[\Upsilon_\ell]\end{array}\right]\succeq 0\end{array}\right\}.\end{array}\qquad (4.7.47)$$

Note that $\mathcal{W}_\varkappa$ is a nonempty convex compact (by Lemma 4.7.2) set such that $\mathcal{W}_\varkappa=\varkappa\mathcal{W}_1$, and $\mathcal{Z}$ is a nonempty closed convex set. Consider the parametric saddle point problem

$$\mathrm{Opt}(\varkappa)=\max_{W\in\mathcal{W}_\varkappa}\min_{(\Upsilon,G,H)\in\mathcal{Z}}\left[E(W;\Upsilon,G,H):=\mathrm{Tr}(WG)+\phi_{\mathcal{R}}(\lambda[\Upsilon])+\overline{\Psi}(H)\right].\qquad (4.7.48)$$

This problem is convex-concave; utilizing the fact that $\mathcal{W}_\varkappa$ is compact and contains positive definite matrices, it is immediately seen that the Sion-Kakutani theorem ensures the existence of a saddle point whenever $\varkappa\in(0,1]$. We claim that

$$0<\varkappa\leq 1\Rightarrow \mathrm{Opt}(\varkappa)\geq\sqrt{\varkappa}\mathrm{Opt}(1).\qquad (4.7.49)$$

Indeed, $\mathcal{Z}$ is invariant w.r.t. scalings

$$(\Upsilon=\{\Upsilon_\ell,\ell\leq L\},G,H)\mapsto(\theta\Upsilon:=\{\theta\Upsilon_\ell,\ell\leq L\},\theta^{-1}G,H),\qquad [\theta>0].$$

When taking into account that $\phi_{\mathcal{R}}(\lambda[\theta\Upsilon])=\theta\phi_{\mathcal{R}}(\lambda[\Upsilon])$, we get

$$\begin{array}{rcl}\underline{E}(W) & := & \min\limits_{(\Upsilon,G,H)\in\mathcal{Z}} E(W;\Upsilon,G,H)=\min\limits_{(\Upsilon,G,H)\in\mathcal{Z}}\inf\limits_{\theta>0}E(W;\theta\Upsilon,\theta^{-1}G,H)\\ & = & \min\limits_{(\Upsilon,G,H)\in\mathcal{Z}}\left[2\sqrt{\mathrm{Tr}(WG)\phi_{\mathcal{R}}(\lambda[\Upsilon])}+\overline{\Psi}(H)\right].\end{array}$$

Because $\overline{\Psi}$ is nonnegative we conclude that whenever $W\succeq 0$ and $\varkappa\in(0,1]$, one has

$$\underline{E}(\varkappa W)\geq\sqrt{\varkappa}\underline{E}(W),$$

which combines with $\mathcal{W}_\varkappa=\varkappa\mathcal{W}_1$ to imply that

$$\mathrm{Opt}(\varkappa)=\max_{W\in\mathcal{W}_\varkappa}\underline{E}(W)=\max_{W\in\mathcal{W}_1}\underline{E}(\varkappa W)\geq\sqrt{\varkappa}\max_{W\in\mathcal{W}_1}\underline{E}(W)=\sqrt{\varkappa}\mathrm{Opt}(1),$$

and (4.7.49) follows.

6^o. We claim that

$$\mathrm{Opt}(1)=\mathrm{Opt},\qquad (4.7.50)$$

where Opt is given by (4.5.15) (and, as we have seen, by (4.7.39) as well). Note that (4.7.50) combines with (4.7.49) to imply that

$$0<\varkappa\leq 1\Rightarrow\mathrm{Opt}(\varkappa)\geq\sqrt{\varkappa}\mathrm{Opt}.\qquad (4.7.51)$$

Verification of (4.7.50) is given by the following computation. We have

$$\begin{array}{l}\mathrm{Opt}(1)=\max\limits_{W\in\mathcal{W}_1}\min\limits_{(\Upsilon,G,H)\in\mathcal{Z}}\left[\mathrm{Tr}(WG)+\phi_{\mathcal{R}}(\lambda[\Upsilon])+\overline{\Psi}(H)\right]\\ \quad=\min\limits_{(\Upsilon,G,H)\in\mathcal{Z}}\max\limits_{W\in\mathcal{W}_1}\left[\mathrm{Tr}(WG)+\phi_{\mathcal{R}}(\lambda[\Upsilon])+\overline{\Psi}(H)\right]\ \text{[by Sion-Kakutani Theorem]}\\ \quad=\min\limits_{(\Upsilon,G,H)\in\mathcal{Z}}\left[\overline{\Psi}(H)+\phi_{\mathcal{R}}(\lambda[\Upsilon])+\max\limits_{W}\{\mathrm{Tr}(GW):W\succeq 0,\exists t\in\mathcal{T}:\mathcal{R}_k[W]\preceq t_kI_{d_k},k\leq K\}\right]\\ \quad=\min\limits_{(\Upsilon,G,H)\in\mathcal{Z}}\left[\overline{\Psi}(H)+\phi_{\mathcal{R}}(\lambda[\Upsilon])+\max\limits_{W,t}\{\mathrm{Tr}(GW):W\succeq 0,[t;1]\in\mathbf{K}[\mathcal{T}],\mathcal{R}_k[W]\preceq t_kI_{d_k},k\leq K\}\right]\end{array}$$

Now, using Conic Duality combined with the fact that $(\mathbf{K}[\mathcal{T}])_* = \{[g;s] : s \geq \phi_{\mathcal{T}}(-g)\}$ we obtain

$$\begin{array}{rl} & \max\limits_{W,t} \{\mathrm{Tr}(GW) : W \succeq 0, [t;1] \in \mathbf{K}[\mathcal{T}], \mathcal{R}_k[W] \preceq t_k I_{d_k}, k \leq K\} \\ = & \min\limits_{Z,[g;s],\Lambda=\{\Lambda_k\}} \left\{ s : \left\{ \begin{array}{l} Z \succeq 0, [g;s] \in (\mathbf{K}[\mathcal{T}])_*, \Lambda_k \succeq 0, k \leq K \\ -\mathrm{Tr}(ZW) - g^T t + \sum_k \mathrm{Tr}(\mathcal{R}_k^*[\Lambda_k] W) \\ -\sum_k t_k \mathrm{Tr}(\Lambda_k) = G \,\forall (W \in \mathbf{S}^n, t \in \mathbf{R}^K) \end{array} \right. \right\} \\ = & \min\limits_{Z,[g;s],\Lambda=\{\Lambda_k\}} \left\{ s : \left\{ \begin{array}{l} Z \succeq 0, s \geq \phi_{\mathcal{T}}(-g), \Lambda_k \succeq 0, k \leq K \\ G = \sum_k \mathcal{R}_k^*[\Lambda_k] - Z, g = -\lambda[\Lambda] \end{array} \right. \right\} \\ = & \min\limits_{\Lambda} \left\{ \phi_{\mathcal{T}}(\lambda[\Lambda]) : \Lambda = \{\Lambda_k \succeq 0, k \leq K\}, G \preceq \sum_k \mathcal{R}_k^*[\Lambda_k] \right\}, \end{array}$$

and we arrive at

$$\begin{array}{l} \mathrm{Opt}(1) = \min\limits_{\Upsilon,G,H,\Lambda} \left[\overline{\Psi}(H) + \phi_{\mathcal{R}}(\lambda[\Upsilon]) + \phi_{\mathcal{T}}(\lambda[\Lambda]) : \begin{array}{l} \Upsilon = \{\Upsilon_\ell \succeq 0, \ell \leq L\}, \Lambda = \{\Lambda_k \succeq 0, k \leq K\}, \\ G \preceq \sum_k \mathcal{R}_k^*[\Lambda_k], \\ \left[\begin{array}{c|c} G & \frac{1}{2}[B^T - A^T H]M \\ \hline \frac{1}{2}M^T[B - H^T A] & \sum_\ell \mathcal{S}_\ell^*[\Upsilon_\ell] \end{array} \right] \succeq 0 \end{array} \right] \\ \qquad = \min\limits_{\Upsilon,H,\Lambda} \left[\overline{\Psi}(H) + \phi_{\mathcal{R}}(\lambda[\Upsilon]) + \phi_{\mathcal{T}}(\lambda[\Lambda]) : \begin{array}{l} \Upsilon = \{\Upsilon_\ell \succeq 0, \ \ell \leq L\}, \Lambda = \{\Lambda_k \succeq 0, k \leq K\} \\ \left[\begin{array}{c|c} \sum_k \mathcal{R}_k^*[\Lambda_k] & \frac{1}{2}[B^T - A^T H]M \\ \hline \frac{1}{2}M^T[B - H^T A] & \sum_\ell \mathcal{S}_\ell^*[\Upsilon_\ell] \end{array} \right] \succeq 0 \end{array} \right] \\ \qquad = \mathrm{Opt} \text{ [see (4.7.39)]}. \end{array}$$

Now we can complete the proof.

7^o. Let us set

$$\varrho_* = \inf_{\widehat{x}(\cdot)} \mathrm{Risk}[\widehat{x}|\mathcal{X}], \ \ \mathrm{Risk}[\widehat{x}|\mathcal{X}] = \sup_{x \in \mathcal{X}} \mathbf{E}_{\xi \sim \mathcal{N}(0,Q_*)} \{\|Bx - \widehat{x}(Ax + \xi)\|\}, \tag{4.7.52}$$

where inf is taken over all estimates. It is immediately seen that $\varrho_* > 0$ due to $Q_* \succ 0$ (recall that $Q_* \in \Pi$ and invoke Assumption **R**) combined with $B \neq 0$ and $0 \in \mathrm{int}\,\mathcal{X}$. Consequently, there is an estimate $\widetilde{x}(\cdot)$ such that $\mathrm{Risk}[\widetilde{x}|\mathcal{X}] \leq \frac{3}{2}\varrho_*$. Further, when $x \in \mathcal{X} \backslash \{0\}$, we have $W := xx^T \in \mathcal{Q}$, see (4.5.27) and (4.4.4), and $W^{1/2} = W/\|x\|_2$, whence for M_* as defined in (4.5.27) we have

$$M_*^2 \geq \mathbf{E}_{\eta \sim \mathcal{N}(0,I_n)} \{\|BW^{1/2}\eta\|^2\} = \|x\|_2^{-2} \|Bx\|^2 \mathbf{E}_{\eta \sim \mathcal{N}(0,I_n)} \{(x^T\eta)^2\} = \|Bx\|^2,$$

and we arrive at

$$x \in \mathcal{X} \Rightarrow \|Bx\| \leq M_*. \tag{4.7.53}$$

Now let us convert the estimate $\widetilde{x}$ into the estimate $\widehat{x}$ defined as follows: $\widehat{x}(\omega)$ is the $\|\cdot\|$-closest to $\widetilde{x}(\omega)$ point of the set $\mathcal{B}_{M_*} = \{u : \|u\| \leq M_*\}$. When $x \in \mathcal{X}$, we have $Bx \in \mathcal{B}_{M_*}$ by (4.7.53), and because, by construction, $\widehat{x}$ is the closest to $\widetilde{x}$ point of $\mathcal{B}_{M_*}$, we have also $\|\widetilde{x}(\omega) - \widehat{x}(\omega)\| \leq \|Bx - \widetilde{x}(\omega)\|$ for all ω. Thus,

$$x \in \mathcal{X} \Rightarrow \|Bx - \widehat{x}(\omega)\| \leq \|Bx - \widetilde{x}(\omega)\| + \|\widetilde{x}(\omega) - \widehat{x}(\omega)\| \leq 2\|Bx - \widetilde{x}(\omega)\|.$$

We conclude that $\|\widehat{x}(\omega)\| \leq M_* \,\forall \omega$, and

$$\mathrm{Risk}[\widehat{x}|\mathcal{X}] \leq 2\mathrm{Risk}[\widetilde{x}|\mathcal{X}] \leq 3\varrho_*. \tag{4.7.54}$$

8^o. For $\varkappa\in(0,1]$, let $W_\varkappa$ be the W-component of a saddle point solution to the saddle point problem (4.7.48). Then, by (4.7.51),

$$\begin{array}{l}\sqrt{\varkappa}\mathrm{Opt}\leq \mathrm{Opt}(\varkappa)=\min\limits_{(\Upsilon,G,H)\in\mathcal{Z}}\left\{\mathrm{Tr}(W_\varkappa G)+\phi_{\mathcal{R}}(\lambda[\Upsilon])+\overline{\Psi}(H)\right\}\\ =\min\limits_{(\Upsilon,G,H)}\left\{\mathrm{Tr}(W_\varkappa G)+\phi_{\mathcal{R}}(\lambda[\Upsilon])+\overline{\Psi}(H):\Upsilon_\ell\succeq 0\,\forall\ell,\left[\begin{array}{c|c}G&\frac{1}{2}[B^T-A^TH]M\\\hline \frac{1}{2}M^T[B-H^TA]&\sum_\ell\mathcal{S}_\ell^*[\Upsilon_\ell]\end{array}\right]\succeq 0\right\}\\ \leq 4\kappa[F]\varrho[W_\varkappa]\end{array}\tag{4.7.55}$$

(we have used (4.7.47.b) and (4.7.45); recall that $\varrho[\cdot]$ is given by (4.7.40)). On the other hand, when applying Lemma 4.7.3 to $Q=W_\varkappa$ we obtain, in view of relations $0<\varkappa\leq 1$, $W_\varkappa\in\mathcal{W}_\varkappa$,

$$\delta(\varkappa):=\mathrm{Prob}_{\eta\sim\mathcal{N}(0,I_n)}\{W_\varkappa^{1/2}\eta\not\in\mathcal{X}\}\leq 2D\exp\{-(2\varkappa)^{-1}\},\tag{4.7.56}$$

with D given by (4.5.29). Setting

$$\mathcal{E}_\varkappa=\{\zeta:\,W_\varkappa^{1/2}\zeta\in\mathcal{X}\},\ \mathcal{E}_\varkappa^c=\mathbf{R}^n\backslash\mathcal{E}_\varkappa,\ \Sigma=\mathrm{Diag}\{I_n,Q_*\},$$

we have by definition of the risk $\varrho[W_\varkappa]$

$$\begin{array}{rcl}\varrho[W_\varkappa]&\leq&\mathbf{E}_{(\eta,\xi)\sim\mathcal{N}(0,\Sigma)}\{\|BW_\varkappa^{1/2}\eta-\widehat{x}(AW_\varkappa^{1/2}\eta+\xi)\|\}\\ &=&\mathbf{E}_{\eta\sim\mathcal{N}(0,I_n)}\left\{\mathbf{E}_{\xi\sim\mathcal{N}(0,Q_*)}\{\|BW_\varkappa^{1/2}\eta-\widehat{x}(AW_\varkappa^{1/2}\eta+\xi)\|\}\right\}\\ &=&\mathbf{E}_{\eta\sim\mathcal{N}(0,I_n)}\left\{\mathbf{E}_{\xi\sim\mathcal{N}(0,Q_*)}\{\|BW_\varkappa^{1/2}\eta-\widehat{x}(AW_\varkappa^{1/2}\eta+\xi)\|\}1\{\eta\in\mathcal{E}_\varkappa\}\right\}\\ &&+\mathbf{E}_{\eta\sim\mathcal{N}(0,I_n)}\left\{\mathbf{E}_{\xi\sim\mathcal{N}(0,Q_*)}\{\|BW_\varkappa^{1/2}\eta-\widehat{x}(AW_\varkappa^{1/2}\eta+\xi)\|\}1\{\eta\in\mathcal{E}_\varkappa^c\}\right\}\\ &\leq&\mathrm{Risk}[\widehat{x}|\mathcal{X}]+\mathbf{E}_{\eta\sim\mathcal{N}(0,I_n)}\left\{(\|BW_\varkappa^{1/2}\eta\|+M_*)1\{\eta\in\mathcal{E}_\varkappa^c\}\right\}\ [\text{since }\|\widehat{x}(\cdot)\|\leq M_*]\\ &\leq&3\varrho_*+M_*\delta(\varkappa)+\mathbf{E}_{\eta\sim\mathcal{N}(0,I_n)}\left\{\|BW_\varkappa^{1/2}\eta\|1\{\eta\in\mathcal{E}_\varkappa^c\}\right\}\ [\text{we have used (4.7.54)}].\end{array}$$

We conclude that

$$\begin{array}{rcl}\varrho[W_\varkappa]&\leq&3\varrho_*+M_*\delta(\varkappa)+\left[\mathbf{E}_{\eta\sim\mathcal{N}(0,I_n)}\left\{\|BW_\varkappa^{1/2}\eta\|^2\right\}\right]^{1/2}\left[\mathrm{Prob}_{\eta\sim\mathcal{N}(0,I_n)}\{\eta\in\mathcal{E}_\varkappa^c\}\right]^{1/2}\\ &\leq&3\varrho_*+M_*[\delta(\varkappa)+\sqrt{\delta(\varkappa)}]\ [\text{by (4.5.27); note that }W_\varkappa\in\mathcal{Q}\text{ due to }\varkappa\leq 1]\\ &\leq&3\varrho_*+2M_*\sqrt{\delta(\varkappa)}\ [\text{since }\delta(\varkappa)\leq 1]\\ &\leq&3\varrho_*+2M_*\sqrt{2D}\exp\{-(4\varkappa)^{-1}\}\ [\text{we have used (4.7.56)}].\end{array}$$

The bottom line here is that

$$0<\varkappa\leq 1\Rightarrow\varrho[W_\varkappa]\leq 3\varrho_*+2M_*\sqrt{2D}\exp\{-\frac{1}{4\varkappa}\}.\tag{4.7.57}$$

Observe that $\varrho_*\leq M_*$, since due to (4.7.53), for the trivial – identically zero – estimate $\bar{x}(\cdot)$ of Bx one has $\mathrm{Risk}[\bar{x}|\mathcal{X}]\leq M_*$. It follows that setting

$$\bar{\varkappa}=\frac{1}{4\ln\left(\frac{2M_*\sqrt{2D}}{\varrho_*}\right)}$$

we ensure that $\bar{\varkappa}\in(0,1]$, whence, by (4.7.57),

$$\varrho[W_{\bar{\varkappa}}]\leq 4\varrho_*.$$

This combines with (4.7.55) to imply that

$$\sqrt{\bar{\varkappa}}\mathrm{Opt}\leq 4\kappa[F]\varrho[W_{\bar{\varkappa}}]\leq 16\kappa[F]\varrho_*,$$

whence finally

$$\mathrm{Opt}\leq\frac{16\kappa[F]}{\sqrt{\bar{\varkappa}}}\varrho_*\leq\frac{128\sqrt{2}}{\sqrt{2}-\mathrm{e}^{1/4}}\sqrt{\ln\left(\frac{4\sqrt{2}F}{\sqrt{2}-\mathrm{e}^{1/4}}\right)\ln\left(\frac{8M_*^2D}{\varrho_*^2}\right)}\varrho_*.$$

Noting that by definition of ϱ_* and $\mathrm{RiskOpt}_{\Pi,\|\cdot\|}[\mathcal{X}]$ we have $\varrho_*\leq\mathrm{RiskOpt}_{\Pi,\|\cdot\|}[\mathcal{X}]\leq M_*$ (the concluding $\leq$ is due to $\|Bx\|\leq M_*$ for $x\in\mathcal{X}$), we arrive at (4.5.28). □

4.7.5 Proofs for Section 4.6

4.7.5.1 Proof of Proposition 4.6.1

The only claim in Proposition which is not an immediate consequence of Proposition 4.4.3 is that problem (4.6.4) is solvable; let us justify this claim. Let $F = \mathrm{Im}A$. Clearly, feasibility of a candidate solution (H, Λ, Υ) to the problem depends solely on the restriction of the linear mapping $z \mapsto H^T z$ onto F, so that adding to the constraints of the problem the requirement that the restriction of this linear mapping on the orthogonal complement of F in $\mathbf{R}^m$ is identically zero, we get an equivalent problem. It is immediately seen that in the resulting problem, the feasible solutions with the value of the objective $\leq a$ for every $a \in \mathbf{R}$ form a compact set, so that the latter problem (and thus – the original one) indeed is solvable. □

4.7.5.2 Proof of Proposition 4.6.2 and justification of Remark 4.6.1

1^o. Observe that setting

$$\mathfrak{R} = \max_x \{\|Bx\| : x \in \mathcal{X}, Ax = 0\}, \tag{4.7.58}$$

we ensure that

$$\mathrm{Risk}_{\mathrm{opt}}[\mathcal{X}] \geq \mathfrak{R}. \tag{4.7.59}$$

Indeed, let $\bar{x}$ be an optimal solution to the (clearly solvable) optimization problem in (4.7.58). Then observation $\omega = 0$ can be obtained from both the signals $x = \bar{x}$ and $x = -\bar{x}$, and therefore the risk of any (deterministic) recovery routine is at least $\|B\bar{x}\| = \mathfrak{R}$, as claimed.

2^o. It may happen that $\mathrm{Ker}\, A = \{0\}$. In this case the situation is trivial: specifying $A^\dagger$ as a partial inverse to A: $A^\dagger A = I_n$ and setting $H^T = BA^\dagger$ (so that $B - H^T A = 0$), $\Upsilon_\ell = 0_{f_\ell \times f_\ell}$, $\ell \leq L$, $\Lambda_k = 0_{dk \times d_k}$, $k \leq K$, we get a feasible solution to the optimization problem in (4.6.4) with zero value of the objective, implying that $\mathrm{Opt}_\# = 0$; consequently, the linear estimate induced by an optimal solution to the problem is with zero risk, and the conclusion of Proposition 4.6.2 is clearly true. with this in mind, we assume from now on that $\mathrm{Ker}\, A \neq \{0\}$. Denoting $\kappa = \dim \mathrm{Ker}\, A$, we can build an $n \times \kappa$ matrix E of rank κ such that $\mathrm{Ker}\, A$ is the image space of E.

3^o. Setting

$$\begin{array}{rcl} \mathcal{Z} &:=& \{z \in \mathbf{R}^\kappa : Ez \in \mathcal{X}\} = \left\{z \in \mathbf{R}^\kappa : \exists (t \in \mathcal{T}) : \bar{R}_k^2[z] \preceq t_k I_{d_k},\, k \leq K\right\},\ \bar{R}_k[z] = R_k[Ez], \\ C &=& \left[\begin{array}{c|c} & \frac{1}{2}BE \\ \hline \frac{1}{2}E^T B^T & \end{array}\right], \end{array} \tag{4.7.60}$$

note that when z runs trough the spectratope $\mathcal{Z}$, Ez runs exactly through the entire set $\{x \in \mathcal{X} : Ax = 0\}$. with this in mind, invoking Proposition 4.4.3, we arrive at

$$\begin{array}{rl} & \mathfrak{R} = \max_{g:\|g\|_* \leq 1} \max_{z \in \mathcal{Z}} g^T BEz = \max_{[u;z] \in \mathcal{B}_* \times \mathcal{Z}} [u; z]^T C [u; z] \\ \leq & \mathrm{Opt} := \min\limits_{\substack{\Upsilon = \{\Upsilon_\ell : \ell \leq L\}, \\ \Lambda = \{\Lambda_k,\, k \leq K\}}} \left\{ \phi_{\mathcal{R}}(\lambda[\Upsilon]) + \phi_{\mathcal{T}}(\lambda[\Lambda]) : \Upsilon_\ell \succeq 0,\ \Lambda_k \succeq 0,\ \forall (\ell, k) \right. \\ & \left. \qquad\qquad \left[\begin{array}{c|c} \sum_\ell S_\ell^*[\Upsilon_\ell] & \frac{1}{2}BE \\ \hline \frac{1}{2}E^T B^T & E^T \left[\sum_k \mathcal{R}_k^*[\Lambda_K]\right] E \end{array}\right] \succeq 0 \right\} \end{array} \tag{4.7.61}$$

(we have used the straightforward identity $\bar{\mathcal{R}}_k^*[\Lambda_k] = E^T \mathcal{R}_k^*[\Lambda_k] E$). By the same Proposition 4.4.3, the optimization problem in (4.7.61) specifying Opt is solvable, and

$$\mathrm{Opt} \leq 2\ln(2D)\mathfrak{R},\ \ D = \sum_k d_k + \sum_\ell f_\ell. \tag{4.7.62}$$

4^o. Let $\bar{\Upsilon}=\{\bar{\Upsilon}_\ell\}$, $\bar{\Lambda}=\{\bar{\Lambda}_k\}$ be an optimal solution to the optimization problem specifying Opt, see (4.7.61), and let

$$\boldsymbol{\Upsilon}=\sum_\ell \mathcal{S}_\ell^*[\bar{\Upsilon}_\ell],\ \boldsymbol{\Lambda}=\sum_k \mathcal{R}_k^*[\bar{\Lambda}_k],$$

so that

$$\mathrm{Opt}=\phi_{\mathcal{R}}(\lambda[\bar{\Upsilon}])+\phi_{\mathcal{T}}(\lambda[\bar{\Lambda}])\ \&\ \left[\begin{array}{c|c}\boldsymbol{\Upsilon} & \frac{1}{2}BE\\ \hline \frac{1}{2}E^TB^T & E^T\boldsymbol{\Lambda}E\end{array}\right]\succeq 0. \tag{4.7.63}$$

We claim that for properly selected $m\times\nu$ matrix H it holds

$$\left[\begin{array}{c|c}\boldsymbol{\Upsilon} & \frac{1}{2}(B-H^TA)\\ \hline \frac{1}{2}(B-H^TA)^T & \boldsymbol{\Lambda}\end{array}\right]\succeq 0. \tag{4.7.64}$$

This claim implies the conclusion of Proposition 4.6.2: by the claim, we have $\mathrm{Opt}_\#\leq\mathrm{Opt}$, which combines with (4.7.62) and (4.7.59) to imply (4.6.5).

In order to justify the claim, assume that it fails to be true, and let us lead this assumption to contradiction.

4^0.a. Consider the semidefinite program

$$\tau_*=\min_{\tau,H}\left\{\tau:\left[\begin{array}{c|c}\boldsymbol{\Upsilon} & \frac{1}{2}(B-H^TA)\\ \hline \frac{1}{2}(B-H^TA)^T & \boldsymbol{\Lambda}\end{array}\right]+\tau I_{\nu+n}\succeq 0\right\}. \tag{4.7.65}$$

The problem clearly is strictly feasible, and the value of the objective at every feasible solution is positive. In addition, the problem is solvable (by exactly the same argument as in the proof of Proposition 4.6.1, see Section 4.7.5.1).

4^0.b. As we have seen, (4.7.65) is a strictly feasible solvable problem with positive optimal value τ_*, so that the problem dual to (4.7.65) is solvable with positive optimal value. Let us build the dual problem. Denoting by $\left[\begin{array}{c|c}U & V\\ \hline V^T & W\end{array}\right]\succeq 0$ the Lagrange multipliers for the semidefinite constraint in (4.7.65) and taking inner product of the left hand side of the constraint with the multiplier, we get the aggregated constraint

$$\mathrm{Tr}(U\boldsymbol{\Upsilon})+\mathrm{Tr}(W\boldsymbol{\Lambda})+\tau[\mathrm{Tr}(U)+\mathrm{Tr}(W)]+\mathrm{Tr}((B-H^TA)V^T)\geq 0;$$

the equality constraints of the dual should make the homogeneous in τ,H part of the left hand side in the aggregated constraint identically equal to τ, which amounts to

$$\mathrm{Tr}(U)+\mathrm{Tr}(W)=1,\ VA^T=0,$$

and the aggregated constraint now reads

$$\tau\geq-\left[\mathrm{Tr}(U\boldsymbol{\Upsilon})+\mathrm{Tr}(W\boldsymbol{\Lambda})+\mathrm{Tr}(BV^T)\right].$$

The dual problem is to maximize the right hand side of the latter constraint over Lagrange multiplier $\left[\begin{array}{c|c}U & V\\ \hline V^T & W\end{array}\right]\succeq 0$ satisfying $AV^T=0$, and its optimal value is $\tau_*>0$, that is, there exists $\left[\begin{array}{c|c}\bar{U} & \bar{V}\\ \hline \bar{V}^T\bar{W} & \end{array}\right]\succeq 0$ such that $A\bar{V}^T=0$ and

$$\mathrm{Tr}(\bar{U}\boldsymbol{\Upsilon})+\mathrm{Tr}(\bar{W}\boldsymbol{\Lambda})+\mathrm{Tr}(B\bar{V}^T)<0; \tag{4.7.66}$$

adding to $\bar{U}$ a small positive multiple of the unit matrix, we can assume, in addition, that $\bar{U} \succ 0$. Now, the relation $A\bar{V}^T = 0$ combines with the definition of E to imply that $\bar{V}^T = EF$ for properly selected matrix F, so that

$$\left[\begin{array}{c|c} \bar{U} & F^TE^T \\ \hline EF & \bar{W} \end{array}\right] \succeq 0,$$

whence, by Schur Complement Lemma,

$$\bar{W} \succeq EF\bar{U}^{-1}F^TE^T,$$

and therefore (4.7.66) combines with $\mathbf{\Lambda} \succeq 0$ to imply that

$$\begin{array}{l} 0 > \mathrm{Tr}(\bar{U}\mathbf{\Upsilon}) + \mathrm{Tr}(\bar{W}\mathbf{\Lambda}) + \mathrm{Tr}(B\bar{V}^T) = \mathrm{Tr}(\bar{U}\mathbf{\Upsilon}) + \mathrm{Tr}(\bar{W}\mathbf{\Lambda}) + \mathrm{Tr}(BEF) \\ \geq \mathrm{Tr}(\bar{U}\mathbf{\Upsilon}) + \mathrm{Tr}(EF\bar{U}^{-1}F^TE^T\mathbf{\Lambda}) + \mathrm{Tr}(BEF) = \mathrm{Tr}\left(\left[\begin{array}{c|c} \mathbf{\Upsilon} & \frac{1}{2}BE \\ \hline \frac{1}{2}E^TB^T & E^T\mathbf{\Lambda}E \end{array}\right]\left[\begin{array}{c|c} \bar{U} & F^T \\ \hline F & F\bar{U}^{-1}F^T \end{array}\right]\right) \end{array}$$

Both matrix factors in the concluding the chain $\mathrm{Tr}(\cdot)$ are positive semidefinite (first – by (4.7.63), and the second - by the Schur Complement Lemma); consequently, the concluding quantity in the chain is nonnegative, which is impossible. We have arrived at a desired contradiction. Proposition 4.6.2 is proved.

Justification of Remark 4.6.1. In the case of (4.6.6), the spectratope $\mathcal{B}_* \times \mathcal{Z}$, see (4.7.60), is an ellitope:

$$\mathcal{B}_* \times \mathcal{Z} = \{[u; z] : \exists (t \in \mathcal{T}, r \in \mathcal{R}) : u^TS_\ell u \leq r_\ell, \ell \leq L, z^T[E^TR_kE]z \leq t_k, k \leq K\}.$$

By the result of Exercise 4.9 as applied to $\mathcal{B}_* \times \mathcal{Z}$ in the role of $\mathcal{X}$, the quantity Opt as defined in (4.7.61) is the same as

$$\mathrm{Opt} = \min_{\lambda,\lambda'} \left\{ \phi_{\mathcal{R}}(\lambda) + \phi_{\mathcal{T}}(\lambda') : \lambda \geq 0, \lambda' \geq 0, C \preceq \sum_\ell \lambda_\ell S_\ell + \sum_k \lambda'_k[E^TR_kE] \right\},$$

whence, by Proposition 4.3.3, along with relation (4.7.62) we have also $\mathrm{Opt} \leq 4\ln(5[K+L])\mathfrak{R}$, implying (4.6.7).

4.8 Appendix: Calculus of Ellitopes/Spectratopes

We present here the rules of the calculus of ellitopes/spectratopes. We formulate these rules for ellitopes; the "spectratopic versions" of the rules are straightforward modifications of the "ellitopic versions."

- Intersection $\mathcal{X} = \bigcap\limits_{i=1}^{I} \mathcal{X}_i$ of ellitopes $\mathcal{X}_i = \{x \in \mathbf{R}^n : \exists (y^i \in \mathbf{R}^{n_i}, t^i \in \mathcal{T}_i) : x = P_iy^i \ \& \ [y^i]^TS_{ik}y^i \leq t^i_k, 1 \leq k \leq K_i\}$, is an ellitope. Indeed, this is evident when $\mathcal{X} = \{0\}$. Assuming $\mathcal{X} \neq \{0\}$, we have

$$\begin{array}{rcl} \mathcal{X} & = & \{x \in \mathbf{R}^n : \exists (y = [y^1; ...; y^I] \in \mathcal{Y}, t = (t^1, ..., t^I) \in \mathcal{T} = \mathcal{T}_1 \times ... \times \mathcal{T}_I) : \\ & & \qquad x = Py := P_1y^1 \ \& \ \underbrace{[y^i]^TS_{ik}y^i}_{y^TS^+_{ik}y} \leq t^i_k, 1 \leq k \leq K_i, 1 \leq i \leq I\}, \\ \mathcal{Y} & = & \{[y^1; ...; y^I] \in \mathbf{R}^{n_1+...+n_I} : P_iy^i = P_1y^1, \ 2 \leq i \leq I\} \end{array}$$

(note that $\mathcal{Y}$ can be identified with $\mathbf{R}^{\bar{n}}$ with a properly selected $\bar{n} > 0$).

- Direct product $\mathcal{X}=\prod\limits_{i=1}^{I}\mathcal{X}_i$ of ellitopes $\mathcal{X}_i=\{x^i\in\mathbf{R}^{n_i}:\exists(y^i\in\mathbf{R}^{\bar n_i},t^i\in\mathcal{T}_i):x^i=P_iy^i,\,1\le i\le I\;\&\;[y^i]^TS_{ik}y^i\le t^i_k,1\le k\le K_i\}$ is an ellitope:

$$\begin{array}{r}\mathcal{X}=\{[x^1;...;x^I]\in\mathbf{R}^{n_1}\times...\times\mathbf{R}^{n_I}:\exists\left(\begin{array}{c}y=[y^1;...;y^I]\in\mathbf{R}^{\bar n_1+...\bar n_I}\\ t=(t^1,...,t^I)\in\mathcal{T}=\mathcal{T}_1\times...\times\mathcal{T}_I\end{array}\right))\\ x=Py:=[P_1y^1;...;P_Iy^I],\,\underbrace{[y^i]^TS_{ik}y^i}_{y^TS^+_{ik}y}\le t^i_k,1\le k\le K_i,1\le i\le I\}\end{array}$$

- The linear image $\mathcal{Z}=\{Rx:x\in\mathcal{X}\}$, $R\in\mathbf{R}^{p\times n}$, of an ellitope $\mathcal{X}=\{x\in\mathbf{R}^n:\exists(y\in\mathbf{R}^{\bar n},t\in\mathcal{T}):x=P_y\;\&\;y^TS_ky\le t_k,\,1\le k\le K\}$ is an ellitope:

$$\mathcal{Z}=\{z\in\mathbf{R}^p:\exists(y\in\mathbf{R}^{\bar n},t\in\mathcal{T}):z=[RP]y\;\&\;y^TS_ky\le t_k,\,1\le k\le K\}.$$

- The inverse linear image $\mathcal{Z}=\{z\in\mathbf{R}^q:Rz\in\mathcal{X}\}$, $R\in\mathbf{R}^{n\times q}$, of an ellitope $\mathcal{X}=\{x\in\mathbf{R}^n:\exists(y\in\mathbf{R}^{\bar n},t\in\mathcal{T}):x=Py\;\&\;y^TS_ky\le t_k,1\le k\le K\}$ under linear mapping $z\mapsto Rz:\mathbf{R}^q\to\mathbf{R}^n$ is an ellitope, *provided that the mapping is an embedding:* $\mathrm{Ker}\,R=\{0\}$. Indeed, setting $E=\{y\in\mathbf{R}^{\bar n}:Py\in\mathrm{Im}R\}$, we get a linear subspace in $\mathbf{R}^{\bar n}$; if $E=\{0\}$, $\mathcal{Z}=\{0\}$ is a spectratope; if $E\ne\{0\}$, we have

$$\begin{array}{rcl}\mathcal{Z}&=&\{z\in\mathbf{R}^q:\exists(y\in E,t\in\mathcal{T}):z=\bar Py\;\&\;y^TS_ky\le t_k,\,1\le k\le K\},\\ \bar P&:&\bar Py=\Pi R,\;\text{ where }\Pi:\mathrm{Im}R\to\mathbf{R}^q\text{ is the inverse of }z\mapsto Rz:\mathbf{R}^q\to\mathrm{Im}R\end{array}$$

(E can be identified with some $\mathbf{R}^k$, and Π is well defined since R is an embedding).

- The arithmetic sum $\mathcal{X}=\{x=\sum_{i=1}^I x^i:x^i\in\mathcal{X}_i,\,1\le i\le I\}$, of ellitopes $\mathcal{X}_i$ is an ellitope, with representation readily given by those of $\mathcal{X}_1,...,\mathcal{X}_I$.

  Indeed, $\mathcal{X}$ is the image of $\mathcal{X}_1\times...\times\mathcal{X}_I$ under the linear mapping $[x^1;...;x^I]\mapsto x^1+....+x^I$, and taking direct products and images under linear mappings preserve ellitopes.

- "$\mathcal{S}$-product." Let $\mathcal{X}_i=\{x^i\in\mathbf{R}^{n_i}:\exists(y^i\in\mathbf{R}^{\bar n_i},t^i\in\mathcal{T}_i):x^i=P_iy^i,\,1\le i\le I\;\&\;[y^i]^TS_{ik}y^i\le t^i_k,1\le k\le K_i\}$ be ellitopes, and let $\mathcal{S}$ be a convex compact set in $\mathbf{R}^I_+$ which intersects the interior of $\mathbf{R}^I_+$ and is monotone: $0\le s'\le s\in\mathcal{S}$ implies $s'\in\mathcal{S}$. We associate with $\mathcal{S}$ the set

$$\mathcal{S}^{1/2}=\left\{s\in\mathbf{R}^I_+:[s_1^2;...;s_I^2]\in\mathcal{S}\right\}$$

  of entrywise square roots of points from $\mathcal{S}$; clearly, $\mathcal{S}^{1/2}$ is a convex compact set.

  $\mathcal{X}_i$ and $\mathcal{S}$ specify the *$\mathcal{S}$-product* of the sets $\mathcal{X}_i$, $i\le I$, defined as the set

$$\mathcal{Z}=\left\{z=[z^1;...;z^I]:\exists(s\in\mathcal{S}^{1/2},x^i\in\mathcal{X}_i,i\le I):z^i=s_ix^i,1\le i\le I\right\},$$

  or, equivalently,

$$\begin{array}{rcl}\mathcal{Z}&=&\left\{z=[z^1;...;z^I]:\exists(r=[r^1;...;r^I]\in\mathcal{R},y^1,...,y^I):\right.\\ &&\qquad\left.z_i=P_iy_i\,\forall i\le I,[y^i]^TS_{ik}y^i\le r^i_k\,\forall(i\le I,k\le K_i)\right\},\\ \mathcal{R}&=&\{[r^1;...;r^I]\ge0:\exists(s\in\mathcal{S}^{1/2},t^i\in\mathcal{T}_i):r^i=s_i^2t^i\,\forall i\le I\}.\end{array}$$

  We claim that $\mathcal{Z}$ is an ellitope. All we need to verify to this end is that the set $\mathcal{R}$ is as it should be in an ellitopic representation, that is, that $\mathcal{R}$ is compact and monotone subset of $\mathbf{R}^{\bar n_1+...+\bar n_I}_+$ containing a strictly positive vector (all this is evident), and that $\mathcal{R}$ is convex. To verify convexity, let $\mathbf{T}_i = \mathrm{cl}\{[t^i;\tau_i] : \tau_i > 0, t^i/\tau_i \in \mathcal{T}_i\}$ be the conic hulls of $\mathcal{T}_i$'s. We clearly have

$$\mathcal{R} = \{[r^1;...;r^I] : \exists s \in \mathcal{S}^{1/2} : [r^i;s_i^2] \in \mathbf{T}_i,\, i \le I\} = \{[r^1;...;r^I] : \exists \sigma \in \mathcal{S} : [r^i;\sigma_i] \in \mathbf{T}_i,\, i \le I\},$$

where the concluding equality is due to the origin of $\mathcal{S}^{1/2}$. The concluding set in the above chain clearly is convex, and we are done.

As an example, consider the situation where the ellitopes $\mathcal{X}_i$ posses nonempty interiors and thus can be thought of as the unit balls of norms $\|\cdot\|_{(i)}$ on the respective spaces $\mathbf{R}^{\bar{n}_i}$, and let $\mathcal{S} = \{s \in \mathbf{R}^I_+ : \|s\|_{p/2} \le 1\}$, where $p \ge 2$. In this situation, $\mathcal{S}^{1/2} = \{s \in \mathbf{R}^I U_+ : \|s\|_p \le 1\}$, whence $\mathcal{Z}$ is the unit ball of the "block p-norm"

$$\|[z^1;...;z^I]\| = \|\,\big[\|z^1\|_{(1)};...;\|z^I\|_{(I)}\big]\,\|_p.$$

Note also that the usual direct product of I ellitopes is their $\mathcal{S}$-product, with $\mathcal{S} = [0,1]^I$.

- "$\mathcal{S}$-weighted sum." Let $\mathcal{X}_i \subset \mathbf{R}^n$ be ellitopes, $1 \le i \le I$, and let $\mathcal{S} \subset \mathbf{R}^I+$, $\mathcal{S}^{1/2}$ be the same as in the previous item. Then the *$\mathcal{S}$-weighted sum* of the sets $\mathcal{X}_i$, defined as

$$\mathcal{X} = \{x : \exists (s \in \mathcal{S}^{1/2}, x^i \in \mathcal{X}_i, i \le I) : x = \sum_i s_i x^i\}$$

is an ellitope. Indeed, the set in question is the image of the $\mathcal{S}$-product of $\mathcal{X}_i$ under the linear mappings $[z^1;...;z^I] \mapsto z^1 + ... + z^I$, and taking $\mathcal{S}$-products and linear images preserves the property to be an ellitope.

It should be stressed that the outlined "calculus rules" are fully algorithmic: representation (4.2.2) of the result of an operation is readily given by the representations (4.2.2) of the operands.

4.9 Exercises for Lecture 4

† marks more difficult exercises.

4.9.1 Linear Estimates vs. Maximum Likelihood

Exercise 4.1 . Consider the problem posed in the beginning of Lecture 4: *Given observation*

$$\omega = Ax + \sigma\xi,\ \xi \sim \mathcal{N}(0, I)$$

of unknown signal x known to belong to a given signal set $\mathcal{X} \subset \mathbf{R}^n$, we want to recover Bx.

Let us restrict ourselves with the case where A is square and invertible matrix, B is the identity, and $\mathcal{X}$ is a computationally tractable convex compact set. As far as computational aspects are concerned, the situation is well suited for utilizing the "magic wand" of Statistics – the *Maximum Likelihood* (ML) estimate where the recovery of x is

$$\widehat{x}_{\mathrm{ML}}(\omega) = \operatorname*{argmin}_{y \in \mathcal{X}} \|\omega - Ay\|_2 \tag{ML}$$

– the signal which maximizes, over $y \in \mathcal{X}$, the likelihood (the probability density) to get the observation we actually got. Indeed, with computationally tractable $\mathcal{X}$, (ML) is an explicit convex, and therefore efficiently solvable, optimization problem. Given the exclusive role played by ML estimate in Statistics, perhaps the first question about our problem of interest is: *how good in the situation in question is the ML estimate?*

The goal of what follows is to demonstrate that *in the situation we are interested in, the ML estimate can be "heavily nonoptimal," and this may happen even when the techniques we develop in Lecture 4 do result in efficiently computable near-optimal linear estimate.*

To justify the claim, investigate the risk (4.1.2) of the ML estimate in the case when

$$\mathcal{X} = \{x \in \mathbf{R}^n : x_1^2 + \epsilon^{-2}\sum_{i=2}^n x_i^2 \leq 1\} \;\&\; A = \mathrm{Diag}\{1, 1/\epsilon, ..., 1/\epsilon\},$$

ϵ and σ are small, and n is large, specifically, $\sigma^2(n-1) \geq 2$. Accompany your theoretical analysis by numerical experiments – compare the empirical risks of the ML estimate with theoretical and empirical risks of the optimal under the circumstances linear estimate.

Recommended setup: *n runs through $\{256, 1024, 2048\}$ and $\epsilon = \sigma$ run through $\{0.01; 0.05; 0.1\}$, and signal x is generated as*

$$x = [\cos(\phi); \sin(\phi)\epsilon\zeta],$$

where $\phi \sim \mathrm{Uniform}[0, 2\pi]$ and random vector ζ is independent of ϕ and is distributed uniformly on the unit sphere in $\mathbf{R}^{n-1}$.

4.9.2 Measurement Design in Signal Recovery

Exercise 4.2 [Measurement Design in Gaussian o.s.] As a preamble to the Exercise, please read the story about possible "physics" of Gaussian o.s. from Section 2.7.3.3. The summary of this story is as follows:

We consider the Measurement Design version of signal recovery in Gaussian o.s., specifically, we are allowed to use observations

$$\omega = A_q x + \sigma\xi \qquad [\xi \sim \mathcal{N}(0, I_m)]$$

where

$$A_q = \mathrm{Diag}\{\sqrt{q_1}, \sqrt{q_2}, ..., \sqrt{q_m}\}A,$$

with a given $A \in \mathbf{R}^{m,n}$ and vector q which we can select in a given convex compact set $\mathcal{Q} \subset \mathbf{R}^m_+$. The signal x underlying the observation is known to belong to a given ellitope $\mathcal{X}$. Your goal is to select $q \in \mathcal{Q}$ and a linear recovery $\omega \mapsto G^T\omega$ of the image Bx of $x \in \mathcal{X}$, with B given, resulting in the minimal worst-case, over $x \in \mathcal{X}$, expected $\|\cdot\|_2^2$ recovery risk. Modify, according to this goal, problem (4.2.9)*. Is it possible to end up with a tractable problem? Work out in full details the case when $\mathcal{Q} = \{q \in \mathbf{R}^m_+ : \sum_i q_i = m\}$.*

Exercise 4.3 [follow-up to Exercise 4.2] *A translucent bar of length $n = 32$ is comprised of 32 consecutive segments of length 1 each, with density ρ_i of i-th segment known to belong to the interval $[\mu - \delta_i, \mu + \delta_i]$.*

Sample translucent bar

The bar is lightened from the left end; when light passes through a segment with density ρ, light's intensity is reduced by factor $\mathrm{e}^{-\alpha\rho}$. The intensity of light at the left endpoint of the bar is 1. You can scan the segments one by one from left to right and measure light intensity ℓ_i at the right endpoint of i-th segment for time q_i; the result z_i of the measurement is $\ell_i \mathrm{e}^{\sigma\xi_i/\sqrt{q_i}}$, where $\xi_i \sim \mathcal{N}(0,1)$ are independent across i. The total time budget is n, and you are interested to recover the $m = n/2$-dimensional vector of densities of the right m segments. Build optimization problem responsible for near-optimal linear recovery with and without Measurement Design (with no Measurement Design, each segment is observed during unit time) and compare the resulting near-optimal risks.

Recommended data:

$$\alpha = 0.01,\ \delta_i = 1.2 + \cos(4\pi(i-1)/n),\ \mu = 1.1\max_i \delta_i,\ \sigma = 0.001.$$

Exercise 4.4 *Let $X \subset \mathbf{R}^n$ be a convex compact set, let $b \in \mathbf{R}^n$, and let A be an $m \times n$ matrix. Consider the problem of affine recovery $\omega \mapsto h^T\omega + c$ of the linear function $Bx = b^Tx$ of $x \in X$ from indirect observation*

$$\omega = Ax + \sigma\xi,\ \xi \sim \mathcal{N}(0, I_m).$$

Given tolerance $\epsilon \in (0,1)$, we are interested to minimize the worst-case, over $x \in X$, width of $(1-\epsilon)$ confidence interval, that is, the smallest ρ such that

$$\mathrm{Prob}\{\xi : b^Tx - f^T(Ax + \sigma\xi) > \rho\} \le \epsilon/2 \ \&\ \mathrm{Prob}\{\xi : b^Tx - f^T(Ax + \sigma\xi) < \rho\} \le \epsilon/2\ \forall x \in X.$$

Pose the problem as a convex optimization problem and consider in details the case where X is the box $\{x \in \mathbf{R}^n : a_j|x_j| \le 1,\ 1 \le j \le n\}$, where $a_j > 0$ for all j.

Exercise 4.5 Let X be an ellitope in $\mathbf{R}^n$:

$$X = \{x \in \mathbf{R}^n : \exists t \in \mathcal{T} : x^TS_kx \le t_k,\ 1 \le k \le K\}$$

with our usual restrictions on S_k and $\mathcal{T}$. Let, further, m be a given positive integer, and $x \mapsto Bx : \mathbf{R}^n \to \mathbf{R}^\nu$ be a given linear mapping. Consider the Measurement Design problem where you are looking for a linear recovery $\omega \mapsto \widehat{x}_H(\omega) := H^T\omega$ of Bx, $x \in X$, from observation

$$\omega = Ax + \sigma\xi \qquad\qquad [\sigma > 0 \text{ is given and } \xi \sim \mathcal{N}(0, I_m)]$$

in which the $m \times n$ sensing matrix A is under your control – it is allowed to be a whatever $m \times n$ matrix of spectral norm not exceeding 1. You are interested to select H and A in order to minimize the worst case, over $x \in X$, expected $\|\cdot\|_2^2$ recovery error. Similarly to (4.2.9), this problem can be posed as

$$\mathrm{Opt} = \min_{H,\lambda,A}\left\{\sigma^2\mathrm{Tr}(H^TH) + \phi_{\mathcal{T}}(\lambda) : \left[\begin{array}{c|c}\sum_k \lambda_k S_k & B^T - A^TH \\ \hline B - H^TA & I_\nu\end{array}\right] \succeq 0,\ \|A\| \le 1,\ \lambda \ge 0\right\}, \quad (4.9.1)$$

where $\|\cdot\|$ stands for the spectral norm. The objective in this problem is the (upper bound on the) squared risk $\mathrm{Risk}^2[\widehat{x}_H|X]$, the sensing matrix being A. The problem is nonconvex, since the matrix participating in the semidefinite constraint is bilinear in H and A.

A natural way to handle an optimization problem with bilinear in the decision variables u, v objective and/or constraints is to use "alternating minimization," where one alternates optimization in v for u fixed and optimization in u for v fixed, where the value of the variable fixed in a round is the result of optimization w.r.t. this variable in the previous round. Alternating minimizations are carried out until the value of the objective (which in the outlined process definitely improves from round to round) stops to improve (or nearly so). Since the algorithm not necessarily converges to the globally optimal solution to the problem of interest, it makes sense to run the algorithm several times from different, say, randomly selected, starting points.

Now goes the Exercise.

1. *Implement Alternating Minimization as applied to* (4.9.1) *and look how it works. You could restrict your experimentation to the case where the sizes m, n, ν are quite moderate, in the range of tens, and X is either the box $\{x : j^{2\gamma}x_j^2 \le 1, 1 \le j \le n\}$, or the ellipsoid $\{x : \sum_{j=1}^n j^{2\gamma}x_j^2 \le 1\}$, where γ is a nonnegative parameter (you could try $\gamma = 0, 1, 2, 3$). As about B, you could generate it at random, or enforce B to have prescribed singular values, say, $\sigma_j = j^{-\theta}$, $1 \le j \le \nu$, and randomly selected system of singular vectors.*

2. *Identify cases where a globally optimal solution to* (4.9.1) *is easy to identify and use this in order to understand how reliable is Alternating minimization in the application in question, reliability meaning the ability to identify near-optimal, in terms of the objective, solutions.*

   If you are not satisfied with Alternating Minimization "as it is," try to improve it.

3. *Modify* (4.9.1) *and your experimentation to cover the cases where the restriction* $\|A\| \leq 1$ *on the sensing matrix is replaced with one of the following restrictions:*

   - $\|\mathrm{Row}_i[A]\|_2 \leq 1,\ 1 \leq i \leq m$
   - $|A_{ij}| \leq 1$ *for all* i, j

   (note that these two types of restrictions mimic what happens if you are interested to recover (linear image of) the vector of parameters in a linear regression model from noisy observations of model's outputs at m *points which you are allowed to select in the unit ball, resp., unit box).*

4. [Embedded Exercise] Recall that a $\nu \times n$ matrix G admits *singular value decomposition* $G = UDV^T$ with orthogonal matrices $U \in \mathbf{R}^{\nu\times\nu}$ and $V \in \mathbf{R}^{n\times n}$ and diagonal $\nu \times n$ matrix D with nonnegative and nonincreasing diagonal entries [18]. These entries are uniquely defined by G and are called *singular values* $\sigma_i(G)$, $1 \leq i \leq \min[\nu, n]$. These singular values admit characterization similar to variational characterization of eigenvalues of a symmetric matrix, see Section A.7.3:

   Theorem 4.9.1 [VCSV - Variational Characterization of Singular Values] *For* $\nu \times n$ *matrix* G *it holds*

   $$\sigma_i(G) = \min_{E \in \mathcal{E}_i} \max_{e \in E, \|e\|_2 = 1} \|GE\|_2,\ 1 \leq i \leq \min[\nu, n], \tag{4.9.2}$$

   where $\mathcal{E}_i$ *is the family of all subspaces in* $\mathbf{R}^n$ *of codimension* $i - 1$.

   Corollary 4.9.1 [SVI - Singular Value Interlacement] *Let* G *and* G' *be* $\nu \times n$ *matrices, and let* $k = \mathrm{Rank}(G'G')$. *Then*

   $$\sigma_i(G) \geq \sigma_{i+k}(G'),\ 1 \leq i \leq \min[\nu, n],$$

   where, by definition, singular values of a $\nu \times n$ *matrix with indexes* $> \min[\nu, n]$ *are zeros.*

   We denote by $\sigma(G)$ the vector of singular values of G arranged in the nonincreasing order. The function $\|G\|_{\mathrm{Sh},p} = \|\sigma(G)\|_p$ is called *Shatten p-norm* of matrix G; this indeed is a norm on the space of $\nu \times n$ matrices, and the conjugate norm is $\|\cdot\|_{\mathrm{Sh},q}$, with $\frac{1}{p} + \frac{1}{q} = 1$. An easy and important consequence of Corollary 4.9.1 is the following fact:

   Corollary 4.9.2 *Given a* $\nu \times n$ *matrix* G*, an integer* k*,* $0 \leq k \leq \min[\nu, n]$*, and* $p \in [1, \infty]$*, (one of) the best approximations of* G *in the Shatten* p*-norm among matrices of rank* $\leq k$ *is obtained from* G *by zeroing our all but* k *largest singular values, that is, the matrix* $G^k = \sum_{i=1}^k \sigma_i(G)\mathrm{Col}_i[U]\mathrm{Col}_i^T[V]$*, where* $G = UDV^T$ *is the singular value decomposition of* G.

   Now goes the Embedded Exercise:

   Prove Theorem 4.9.1 and Corollaries 4.9.1 and 4.9.2.

[18] By definition, diagonality of a rectangular matrix D means that all entries D_{ij} in D with $i \neq j$ are zeros.

5. *Consider the Measurement Design problem* (4.9.1) *in the case when X is an ellipsoid:*

$$X = \{x \in \mathbf{R}^n : \sum_{j=1}^{n} x_i^2/a_j^2 \leq 1\}$$

A is restricted to be $m \times n$ matrix of spectral norm not exceeding 1, and there is no noise in observations: $\sigma = 0$, and find an optimal solution to this problem. Think how this result can be used to get a hopefully good starting point for Alternating Minimization in the case when X is an ellipsoid and σ is small.

Exercise 4.6 *Prove Proposition 4.6.3.*

Exercise 4.7 *Prove Proposition 4.6.4.*

4.9.3 Around semidefinite relaxation

4.9.3.1 Linear estimates of stochastic signals

Exercise 4.8 [19] [recovering stochastic signals] In the recovery problem considered in this Lecture, the signal x underlying observation $\omega = Ax + \xi$ was "deterministic uncertain but bounded" – all a priori information on x was that $x \in \mathcal{X}$ for a given signal set $\mathcal{X}$. There is a well-known alternative model, where the signal x has a random component, specifically,

$$x = [\eta; u]$$

where the "stochastic component" η is random with (partly) known probability distribution P_η, and the "deterministic component" u is known to belong to a given set $\mathcal{X}$. As a typical example, consider linear dynamical system given by

$$\begin{array}{rcl} y_{t+1} & = & P_t y_t + \eta_t + u_t \\ \omega_t & = & C_t y_t + \xi_t \end{array}, 1 \leq t \leq K, \tag{4.9.3}$$

where y_t, η_t, u_t are, respectively, the state, the random "process noise," and the deterministic "uncertain but bounded" disturbance affecting the system at time t, ω_t is the output – it is what we observe at time t, and ξ_t is the observation noise. We assume that the matrices P_t, C_t are known in advance. Note that the trajectory

$$y = [y_1; ...; y_K]$$

of the states depends not only on the trajectories of process noises η_t and disturbances u_t, but also on the initial state y_1, which can be modeled as a realization of either the initial noise η_0, or the initial disturbance u_0. When $u_t \equiv 0$, $y_1 = \eta_0$ and the random vectors $\{\eta_t, 0 \leq t \leq K, \xi_t, 1 \leq t \leq K\}$ are independent of each other zero mean Gaussian, (4.9.3) is the model underlying the famous *Kalman filter*.

Now, given model (4.9.3), we can use the equations of the model to represent the trajectory of the states as linear image of the trajectory of noises $\eta = \{\eta_t\}$ and the trajectory of disturbances $u = \{u_t\}$:

$$y = P\eta + Qu$$

(recall that the initial state is either the component η_0 of η, or the component u_0 of u), and our "full observation" becomes

$$\omega = [\omega_1; ...; \omega_K] = A[\eta; u] + \xi, \;\; \xi = [\xi_1, ..., \xi_K].$$

[19] We are grateful to Dr. Georgios Kotsalis who suggested this Exercise.

A typical statistical problem associated with the outlined situation is to estimate the linear image $B[\eta; u]$ of the "signal" $x = [\eta; u]$ underlying our observation. For example, when speaking about (4.9.3), the goal could be to recover y_{K+1} ("forecast").

We arrive at the following estimation problem:

> Given noisy observation
> $$\omega = Ax + \xi \in \mathbf{R}^m$$
> of signal $x = [\eta; u]$ with random component $\eta \in \mathbf{R}^k$ and deterministic component u known to belong to a given set $\mathcal{X} \subset \mathbf{R}^n$, we want to recover the image $Bx \in \mathbf{R}^\nu$ of the signal. Here A and B are given matrices, η is independent of ξ, and we have a priori (perhaps, incomplete) information on the probability distribution P_η of η, specifically, know that $P_\eta \in \mathcal{P}_\eta$ for a given family $\mathcal{P}_\eta$ of probability distributions. Similarly, we assume that what we know about the noise ξ is that its distribution belongs to a given family $\mathcal{P}_\xi$ of distributions on the observation space.

Given a norm $\|\cdot\|$ on the image space of B, it makes sense to specify the risk of a candidate estimate $\widehat{x}(\omega)$ by taking expectation of the recovery error $\|\widehat{x}(A[\eta; u] + \xi) - B[\eta; u]\|$ over *both* ξ and η and then taking supremum of the result over the allowed distributions of η, ξ and over $u \in \mathcal{X}$:

$$\mathrm{Risk}_{\|\cdot\|}[\widehat{x}] = \sup_{u\in\mathcal{X}} \sup_{P_\xi\sim\mathcal{P}_\xi, P_\eta\sim\mathcal{P}_\eta} \mathbf{E}_{[\xi;\eta]\sim P_\xi\times P_\eta}\left\{\|\widehat{x}(A[\eta; u] + \xi) - B[\eta; u]\|\right\}.$$

When $\|\cdot\| = \|\cdot\|_2$ and all distributions from $\mathcal{P}_\xi$ and $\mathcal{P}_\eta$ are with zero means and finite covariance matrices, it is technically more convenient to operate with the *Euclidean risk*

$$\mathrm{Risk}_{\mathrm{Eucl}}[\widehat{x}] = \left[\sup_{u\in\mathcal{X}} \sup_{P_\xi\sim\mathcal{P}_\xi, P_\eta\sim\mathcal{P}_\eta} \mathbf{E}_{[\xi;\eta]\sim P_\xi\times P_\eta}\left\{\|\widehat{x}(A[\eta; u] + \xi) - B[\eta; u]\|_2^2\right\}\right]^{1/2}.$$

The goal of exercise is to verify that as far as the design of "presumably good" *linear estimates* $\widehat{x}(\omega) = H^T\omega$ is concerned, the techniques developed in this Lecture can be straightforwardly extended from the case of signals with no random component to the one where this component is present.

1. *Let $\mathcal{P}_\xi$ be comprised of all probability distributions P on $\mathbf{R}^m$ with zero mean and covariance matrices* $\mathrm{Vary}[P] = \mathbf{E}_{\xi\sim P}\{\xi\xi^T\}$ *running through a computationally tractable convex compact subset $\mathcal{Q}_\eta \subset \mathrm{int}\,\mathbf{S}^m_+$, and $\mathcal{P}_\eta$ be comprised of all probability distributions P on $\mathbf{R}^k$ with zero mean and covariance matrices running through a computationally tractable convex compact subset $\mathcal{Q}_\eta \subset \mathrm{int}\,\mathbf{S}^k_+$. Let, in addition, $\mathcal{X}$ be a basic spectratope:*
$$\mathcal{X} = \{u \in \mathbf{R}^n : \exists t \in \mathcal{T} : R_k^2[x] \preceq t_k I_{d_k},\ k \le K\}$$
with our standard restrictions on $\mathcal{T}$ and $R_k[\cdot]$. Derive efficiently solvable convex optimization problem "responsible" for presumably good, in terms of its Euclidean risk, linear estimate.

2. *Let $\mathcal{P}_\xi$ be comprised of all probability distributions P on $\mathbf{R}^m$ with matrices of second moments* $\mathrm{Vary}[P] = \mathbf{E}_{\xi\sim P}\{\xi\xi^T\}$ *running through a computationally tractable convex compact subset $\mathcal{Q}_\eta \subset \mathrm{int}\,\mathbf{S}^m_+$, and $\mathcal{P}_\eta$ be comprised of all probability distributions P on $\mathbf{R}^k$ with matrices of second moments running through a computationally tractable convex compact subset $\mathcal{Q}_\eta \subset \mathrm{int}\,\mathbf{S}^k_+$. Let, in addition, $\mathcal{X}$ be a basic spectratope:*
$$\mathcal{X} = \{u \in \mathbf{R}^n : \exists t \in \mathcal{T} : R_k^2[x] \preceq t_k I_{d_k},\ k \le K\},$$
and let $\|\cdot\|$ be such that the unit ball $\mathcal{B}_*$ of the conjugate norm $\|\cdot\|_*$ is a spectratope:
$$\mathcal{B}_* = \{y : \|y\|_* \le 1\} = \{y \in \mathbf{R}^m : \exists (r \in \mathcal{R}, z) : y = Mz, R_\ell^2[z] \preceq r_\ell I_{f_\ell}, \ell \le L\},$$

with our standard restrictions on $\mathcal{T}, \mathcal{R}, R_k[\cdot], S_\ell[\cdot]$. Derive efficiently solvable convex optimization problem "responsible" for presumably good, in terms of its risk $\mathrm{Risk}_{\|\cdot\|}$, *linear estimate.*

What happens when instead of assuming ξ and η independent, we allow for their dependence (and still require from the marginal distributions of ξ and of η to belong to the above families $\mathcal{P}_\xi$, $\mathcal{P}_\eta$) ?

4.9.3.2 More on semidefinite relaxation

Exercise 4.9 Let $\mathcal{X}$ be an ellitope:

$$\mathcal{X} = \{x \in \mathbf{R}^n : \exists (y \in \mathbf{R}^N, t \in \mathcal{T}) : x = Py, y^T S_k y \leq t_k, k \leq K\}$$

with our standard restrictions on $\mathcal{T}$ and S_k. Representing $S_k = \sum_{j=1}^{r_k} s_{kj} s_{kj}^T$, we can pass from initial ellitopic representation of $\mathcal{X}$ to the spectratopic representation of the same set:

$$\begin{array}{c}\mathcal{X} = \{x \in \mathbf{R}^n : \exists (y \in \mathbf{R}^N, t^+ \in \mathcal{T}^+) : x = Py, [s_{kj}^T x]^2 \preceq t^+ kj I_1, 1 \leq k \leq K, 1 \leq j \leq r_k\} \\ \left[\mathcal{T}^+ = \{t^+ = \{t^+_{kj} \geq 0\} : \exists t \in \mathcal{T} : \sum_{j=1}^{r_k} t^+_{kj} \leq t_k,\, 1 \leq k \leq K\}\right]\end{array}$$

If now C is a symmetric $n \times n$ matrix and $\mathrm{Opt} = \max_{x \in \mathcal{X}} x^T C x$, we have

$$\begin{array}{rcl}\mathrm{Opt}_* & \leq & \mathrm{Opt}_e := \min\limits_{\lambda = \{\lambda_k \in \mathbf{R}_+\}} \left\{\phi_{\mathcal{T}}(\lambda) : P^T C P \preceq \sum_k \lambda_k S_k\right\} \\ \mathrm{Opt}_* & \leq & \mathrm{Opt}_s := \min\limits_{\Lambda = \{\Lambda_{kj} \in \mathbf{R}_+\}} \left\{\phi_{\mathcal{T}^+}(\Lambda) : P^T C P \preceq \sum_{k,j} \Lambda_{kj} s_{kj} s_{kj}^T\right\}\end{array}$$

where the first relation is yielded by ellitopic representation of $\mathcal{X}$ and Proposition 4.3.3, and the second, on a closest inspection (carry this inspection out!) – by the spectratopic representation of $\mathcal{X}$ and Proposition 4.4.3. Now goes Exercise:
Prove that $\mathrm{Opt}_e = \mathrm{Opt}_s$.

Exercise 4.10 [estimating Kolmogorov widths of sperctratopes/ellitopes]

4.10.A Preliminaries: Kolmogorov and Gelfand widths. Let $\mathcal{X}$ be a convex compact set in $\mathbf{R}^n$, and let $\|\cdot\|$ be a norm on $\mathbf{R}^n$. Given a linear subspace E in $\mathbf{R}^n$, let

$$\mathrm{dist}_{\|\cdot\|}(x, E) = \min_{z \in E} \|x - z\| : \mathbf{R}^n \to \mathbf{R}_+$$

be the $\|\cdot\|$-distance from x to E. The quantity

$$\mathrm{dist}_{\|\cdot\|}(\mathcal{X}, E) = \max_{x \in \mathcal{X}} \mathrm{dist}_{\|\cdot\|}(x, E)$$

can be viewed as the worst-case $\|\cdot\|$-accuracy to which vectors from $\mathcal{X}$ can be approximated by vectors from E. Given positive integer $m \leq n$ and denoting by $\mathcal{E}_m$ the family of all linear subspaces in $\mathbf{R}^m$ of dimension m, the quantity

$$\delta_m(\mathcal{X}, \|\cdot\|) = \min_{E \in \mathcal{E}_m} \mathrm{dist}_{\|\cdot\|}(\mathcal{X}, E)$$

can be viewed as the best achievable quality of approximation, in $\|\cdot\|$, of vectors from $\mathcal{X}$ by vectors from an m-dimensional linear subspace of $\mathbf{R}^n$; this quantity is called *m-th Kolmogorov width* of $\mathcal{X}$ taken w.r.t. $\|\cdot\|$.

Observe that one has

$$\begin{array}{rl}\mathrm{dist}_{\|\cdot\|}(x, E) = & \max_\xi \{\xi^T x : \|\xi\|_* \leq 1, \xi \in E^\perp\}, \\ \mathrm{dist}_{\|\cdot\|}(\mathcal{X}, E) = & \max\limits_{\substack{x \in \mathcal{X}, \\ \|\xi\|_* \leq 1, \xi \in E^\perp}} \xi^T x\end{array} \qquad (!)$$

where $E^\perp$ is the orthogonal complement to E.

1. *Prove* (!).
   Hint: Represent $\mathrm{dist}_{\|\cdot\|}(x,E)$ *as the optimal value in a conic problem on the cone* $\mathbf{K}=\{[x;t]: t\geq\|x\|\}$ *and use Conic Duality Theorem.*

Now consider the case when $\mathcal{X}$ is the unit ball of some norm $\|\cdot\|_{\mathcal{X}}$. In this case (!) combines with the definition of Kolmogorov width to imply that

$$\begin{array}{rcl}\delta_m(\mathcal{X},\|\cdot\|) &=& \min\limits_{E\in\mathcal{E}_m}\mathrm{dist}_{\|\cdot\|}(x,E)=\min\limits_{E\in\mathcal{E}_m}\max\limits_{x\in\mathcal{X}}\max\limits_{y\in E^\perp,\|y\|_*\leq 1}y^Tx\\ &=& \min\limits_{E\in\mathcal{E}_m}\mathrm{dist}_{\|\cdot\|}(x,E)=\min\limits_{E\in\mathcal{E}_m}\max\limits_{y\in E^\perp,\|y\|_*\leq 1}\max\limits_{x:\|x\|_{\mathcal{X}}\leq 1}y^Tx\\ &=& \min\limits_{F\in\mathcal{E}_{n-m}}\max\limits_{y\in F,\|y\|_*\leq 1}\|y\|_{\mathcal{X},*},\end{array}\tag{4.9.4}$$

where $\|\cdot\|_{\mathcal{X},*}$ is the norm conjugate to $\|\cdot\|_{\mathcal{X}}$. Note that when $\mathcal{Y}$ is a convex compact set in $\mathbf{R}^n$ and $|\cdot|$ is a norm on $\mathbf{R}^n$, the quantity

$$d^m(\mathcal{Y},|\cdot|)=\min_{F\in\mathcal{E}_{n-m}}\max_{y\in\mathcal{Y}\cap F}|y|$$

has a name – it is called the m-th *Gelfand width* of $\mathcal{Y}$ taken w.r.t. $|\cdot|$. "Duality relation" (4.9.4) states that

> *When $\mathcal{X}$, $\mathcal{Y}$ are the unit balls of the respective norms $\|\cdot\|_{\mathcal{X}}$, $\|\cdot\|_{\mathcal{Y}}$, for every $m<n$ m-th Kolmogorov width of $\mathcal{X}$ taken w.r.t. $\|\cdot\|_{\mathcal{Y},*}$ is the same as m-th Gelfand width of $\mathcal{Y}$ taken w.r.t. $\|\cdot\|_{\mathcal{X},*}$.*

The goal of the remaining part of Exercise is to use our results on the quality of semidefinite relaxation on ellitopes/spectratopes to infer efficiently computable upper bounds on Kolmogorov widths of a given set $\mathcal{X}\subset\mathbf{R}^n$. In the sequel we assume that

- $\mathcal{X}$ is a spectratope:

$$\mathcal{X}=\{x\in\mathbf{R}^n:\exists(t\in\mathcal{T},u): x=Pu, R_k^2[u]\preceq t_kI_{d_k}, k\leq K\};$$

- The unit ball $\mathcal{B}_*$ of the norm conjugate to $\|\cdot\|$ is a spectratope:

$$\mathcal{B}_*=\{y:\|y\|_*\leq 1\}=\{y\in\mathbf{R}^n:\exists(r\in\mathcal{R},z): y=Mz, S_\ell^2[z]\preceq r_\ell I_{f_\ell},\ell\leq L\}.$$

with our usual restrictions on $\mathcal{T},\mathcal{R}$ and $R_k[\cdot]$, $S_\ell[\cdot]$.

4.10.B Simple case: $\|\cdot\|=\|\cdot\|_2$. We start with the *simple case* where $\|\cdot\|=\|\cdot\|_2$, so that $\mathcal{B}_*$ is the ellitope $\{y: y^Ty\leq 1\}$.

Let $D=\sum_k d_k$ be the size of the spectratope $\mathcal{X}$, and let

$$\varkappa=2\max[\ln(2D),1].$$

Given integer $m<n$, consider convex optimization problem

$$\mathrm{Opt}(m)=\min_{\Lambda=\{\Lambda_k,k\leq K\},Y}\left\{\phi_{\mathcal{T}}(\lambda[\Lambda]):\Lambda_k\succeq 0\forall k,\sum_k\mathcal{S}_k^*[\Lambda_k]\succeq P^TYP, 0\preceq Y\preceq I_n,\mathrm{Tr}(Y)=n-m\right\}\tag{P_m}$$

2. *Prove the following*

Proposition 4.9.1 *Whenever $1 \leq \mu \leq m < n$, one has*

$$\mathrm{Opt}(m) \leq \varkappa\delta_m^2(\mathcal{X}, \|\cdot\|_2) \ \& \ \delta_m^2(\mathcal{X}, \|\cdot\|_2) \leq \frac{m+1}{m+1-\mu}\mathrm{Opt}(\mu). \tag{4.9.5}$$

Moreover, the above upper bounds on $\delta_m(\mathcal{X}, \|\cdot\|_2)$ are "constructive", meaning that an optimal solution to (P_μ), $\mu \leq m$, can be straightforwardly converted into a linear subspace $E^{m,\mu}$ of dimension m such that

$$\mathrm{dist}_{\|\cdot\|_2}(\mathcal{X}, E^{m,\mu}) \leq \sqrt{\frac{m+1}{m+1-\mu}\mathrm{Opt}(\mu)}.$$

Finally, $\mathrm{Opt}(\mu)$ is nonincreasing in μ.

4.10.C General case. Now consider the case when both $\mathcal{X}$ and the unit ball $\mathcal{B}_*$ of the norm conjugate to $\|\cdot\|$ are spectratopes. As you are about to see, this case is essentially more difficult than the case of $\|\cdot\| = \|\cdot\|_2$, but something still can be done.

3. *Prove the following statement:*

   (!!) Given $m < n$, let Y be an orthoprojector of $\mathbf{R}^n$ of rank $n-m$, and let collections $\Lambda = \{\Lambda_k \succeq 0, k \leq K\}$ and $\Upsilon = \{\Upsilon_\ell \succeq 0, \ell \leq L\}$ satisfy the relation

$$\left[\begin{array}{c|c} \sum_k \mathcal{R}_k^*[\Lambda_k] & \frac{1}{2}P^TYM \\ \hline \frac{1}{2}M^TYP & \sum_\ell \mathcal{S}_\ell^*[\Upsilon_\ell] \end{array}\right] \succeq 0. \tag{4.9.6}$$

   Then

$$\mathrm{dist}_{\|\cdot\|}(\mathcal{X}, \mathrm{Ker}\, Y) \leq \phi_{\mathcal{T}}(\lambda[\Lambda]) + \phi_{\mathcal{R}}(\lambda[\Upsilon]). \tag{4.9.7}$$

   As a result,

$$\begin{array}{rcl} \delta_m(\mathcal{X}, \|\cdot\|) & \leq & \mathrm{dist}_{\|\cdot\|}(\mathcal{X}, \mathrm{Ker}\, Y) \\ & \leq & \mathrm{Opt} := \min\limits_{\Lambda=\{\Lambda_k, k\leq K\}, \Upsilon=\{\Upsilon_\ell, \ell\leq L\}} \Bigg\{ \phi_{\mathcal{T}}(\lambda[\Lambda]) + \phi_{\mathcal{R}}(\lambda[\Upsilon]) : \\ & & \left\{ \begin{array}{l} \Lambda_k \succeq 0\, \forall k, \Upsilon_\ell \succeq 0\, \forall \ell, \\ \left[\begin{array}{c|c} \sum_k \mathcal{R}_k^*[\Lambda_k] & \frac{1}{2}P^TYM \\ \hline \frac{1}{2}M^TYP & \sum_\ell \mathcal{S}_\ell^*[\Upsilon_\ell] \end{array}\right] \succeq 0 \end{array} \right\}. \end{array} \tag{4.9.8}$$

4. *Prove the following statement:*

   (!!!) Let m, n, Y be as in (!!). Then

$$\begin{array}{rcl} \delta_m(\mathcal{X}, \|\cdot\|) & \leq & \mathrm{dist}_{\|\cdot\|}(\mathcal{X}, \mathrm{Ker}\, Y) \\ & \leq & \widehat{\mathrm{Opt}} := \min\limits_{\nu, \Lambda=\{\Lambda_k, k\leq K\}, \Upsilon=\{\Upsilon_\ell, \ell\leq L\}} \Bigg\{ \phi_{\mathcal{T}}(\lambda[\Lambda]) + \phi_{\mathcal{R}}(\lambda[\Upsilon]) : \\ & & \left\{ \begin{array}{l} \nu \geq 0, \Lambda_k \succeq 0\, \forall k, \Upsilon_\ell \succeq 0\, \forall \ell, \\ \left[\begin{array}{c|c} \sum_k \mathcal{R}_k^*[\Lambda_k] & \frac{1}{2}P^TM \\ \hline \frac{1}{2}M^TP & \sum_\ell \mathcal{S}_\ell^*[\Upsilon_\ell] + \nu M^T(I-Y)M \end{array}\right] \succeq 0 \end{array} \right\}. \end{array} \tag{4.9.9}$$

   and $\widehat{\mathrm{Opt}} \leq \mathrm{Opt}$, with Opt given by (4.9.8).

Statements (!!), (!!!) suggest the following policy for upper-bounding the Kolmogorov width $\delta_m(\mathcal{X}, \|\cdot\|)$:

A. First, we select an integer μ, $1 \leq \mu < n$, and solve the convex optimization problem

$$\min_{\Lambda,\Upsilon,Y}\left\{\phi_{\mathcal{T}}(\lambda[\Lambda])+\phi_{\mathcal{R}}(\lambda[\Upsilon]):\left\{\begin{array}{l}\Lambda=\{\Lambda_k\succeq 0,k\leq K\},\Upsilon=\{\Upsilon_\ell\succeq 0,\ell\leq L\},\\ 0\preceq Y\preceq I,\mathrm{Tr}(Y)=n-\mu\\ \left[\begin{array}{c|c}\sum_k\mathcal{R}_k^*[\Lambda_k] & \frac{1}{2}P^TYM\\ \hline \frac{1}{2}M^TYP & \sum_\ell\mathcal{S}_\ell^*[\Upsilon_\ell]\end{array}\right]\succeq 0\end{array}\right.\right\} \qquad (P^\mu)$$

B. Next, we take the Y-component Y^μ of the optimal solution to (I^μ) and "round" it to a orthoprojector Y of rank $n-m$ in the same fashion as in the case of $\|\cdot\|=\|\cdot\|_2$, that is, keep the eigenvectors of Y^μ intact and replace m smallest eigenvalues with zeros, and all remaining eigenvalues with ones.

C. Finally, we solve the convex optimization problem

$$\mathrm{Opt}_{m,\mu}=\min_{\Lambda,\Upsilon,\nu}\left\{\phi_{\mathcal{T}}(\lambda[\Lambda])+\phi_{\mathcal{R}}(\lambda[\Upsilon]):\left\{\begin{array}{l}\nu\geq 0,\Lambda=\{\Lambda_k\succeq 0,k\leq K\},\Upsilon=\{\Upsilon_\ell\succeq 0,\ell\leq L\},\\ \left[\begin{array}{c|c}\sum_k\mathcal{R}_k^*[\Lambda_k] & \frac{1}{2}P^TM\\ \hline \frac{1}{2}M^TP & \sum_\ell\mathcal{S}_\ell^*[\Upsilon_\ell]+\nu M^T(I-Y)M\end{array}\right]\succeq 0\end{array}\right.\right\} \qquad (P^{m,\mu})$$

By (!!!), $\mathrm{Opt}_{m,\mu}$ is an upper bound on the Kolmogorov width $\delta_m(\mathcal{X},\|\cdot\|)$ (and in fact – also on $\mathrm{dist}_{\|\cdot\|}(\mathcal{X},\mathrm{Ker}\,Y)$).

Pay attention to complications incurred by passing from the simple case $\|\cdot\|=\|\cdot\|_2$ to the case of general norm $\|\cdot\|$ with spectratope as the unit ball of the conjugate norm. Indeed, Proposition 4.9.1 gives both a lower bound $\sqrt{\mathrm{Opt}(m)/\varkappa}$ on the m-th Kolmogorov width of $\mathcal{X}$ w.r.t. $\|\cdot\|_2$, and a family of upper bounds $\sqrt{\frac{m+1}{m+1-\mu}\mathrm{Opt}(\mu)}$, $1\leq\mu\leq m$, on this width. As a result, we can approximate $\mathcal{X}$ by m-dimensional subspaces in the Euclidean norm in a "nearly optimal" fashion. Indeed, if for some ϵ and k it holds $\delta_k(\mathcal{X},\|\cdot\|_2)\leq\epsilon$, then $\mathrm{Opt}(k)\leq\varkappa\epsilon^2$ by Proposition 4.9.1 as applied with $m=k$. On the other hand, assuming $k<n/2$, the same Proposition when applied with $m=2k$ and $\mu=k$ says that

$$\mathrm{dist}_{\|\cdot\|_2}(\mathcal{X},E^{m,k})\leq\sqrt{\frac{2k+1}{k+1}\mathrm{Opt}(k)}\leq\sqrt{2}\sqrt{\mathrm{Opt}(k)}\leq\sqrt{2\varkappa}\epsilon.$$

Thus, if "in the nature" $\mathcal{X}$ can be approximated by k-dimensional subspace within $\|\cdot\|_2$-accuracy ϵ, we can efficiently get approximation of "nearly the same quality" ($\sqrt{2\varkappa}\epsilon$ instead of ϵ; recall that $\varkappa$ is just logarithmic in D) and "nearly the same dimension" ($2k$ instead of k).

Neither one of these options is preserved when passing from the Euclidean norm to a general one: in the latter case, we do not have neither lower bounds on Kolmogorov widths, just upper ones, nor understanding of how tight our upper bounds are.

Now – the concluding questions:

5. *Why in step A of the above bounding policy we utilize statement* (!!) *rather than less conservative (since* $\widehat{\mathrm{Opt}}\leq\mathrm{Opt}$*) statement* (!!!) *?*

6. *Implement the above scheme numerically and run experiments. Recommended setup:*

   - *Given positive integers n and κ and a real $\sigma>0$, specify $\mathcal{X}$ as the set of n-dimensional vectors x which can be obtained when restricting a function f of continuous argument $t\in[0,1]$ onto n-point equidistant grid $\{t_i=i/n\}_{i=1}^n$, and impose on f the smoothness restriction that $|f^{(k)}(t)|\leq\sigma^k$, $0\leq t\leq 1$, $k=0,1,2,...,\kappa$; translate this description on f into a bunch of two-sided linear constraints on x, specifically, the constraints*

$$|d_{(k)}^T[x_i;x_{i+1};...;x_{i+k}]|\leq\sigma^k, 1\leq i\leq n-k, 0\leq k\leq\kappa,$$

where $d_{(k)} \in \mathbf{R}^{k+1}$ is the vector of coefficients of finite-difference approximation, with resolution $1/n$, of k-th derivative:

$$d_{(0)} = 1,\ d_{(1)} = n[-1;1],\ d_{(2)} = n^2[1;-2;1],\ d_{(3)} = n^3[-1;3;-3;1],\ d_{(4)} = n^4[1;-4;6;-4;1], ...$$

- *Recommended parameters: $n = 32$, $m = 8$, $\kappa = 5$, $\sigma \in \{0.25, 0.5; 1, 2, 4\}$.*
- *Run experiments with $\|\cdot\| = \|\cdot\|_1$ and $\|\cdot\| = \|\cdot\|_2$.*

Exercise 4.11 † [more on semidefinite relaxation] The goal of this Exercise is to extend SDP relaxation beyond ellitopes/spectratopes.

SDP relaxation is aimed at upper-bounding the quantity

$$\mathrm{Opt}_{\mathcal{X}}(B) = \max_{x \in \mathcal{X}} x^T B x, \qquad [B \in \mathbf{S}^n]$$

where $\mathcal{X} \subset \mathbf{R}^n$ is a given set (which we from now on assume to be nonempty convex compact). To this end we look for a computationally tractable convex compact set $\mathcal{U} \subset \mathbf{S}^n$ such that for every $x \in \mathcal{X}$ it holds $xx^T \in \mathcal{U}$; in this case, we refer to $\mathcal{U}$ as to a set *matching* $\mathcal{X}$ (equivalent wording: "$\mathcal{U}$ matches $\mathcal{X}$"). Given such a set $\mathcal{U}$, the optimal value in the convex optimization problem

$$\overline{\mathrm{Opt}}_{\mathcal{U}}(B) = \max_{U \in \mathcal{U}} \mathrm{Tr}(BU) \tag{4.9.10}$$

is an efficiently computable convex upper bound on $\mathrm{Opt}_{\mathcal{X}}(B)$.

Given $\mathcal{U}$ matching $\mathcal{X}$, we can pass from $\mathcal{U}$ to the conic hull of $\mathcal{U}$ – to the set

$$\mathbf{U}[\mathcal{U}] = \mathrm{cl}\{(U,\mu) \in \mathbf{S}^n \times \mathbf{R}_+ : \mu > 0, U/\mu \in \mathcal{U}\}$$

which, as it is immediately seen, is a closed convex cone contained in $\mathbf{S}^n \times \mathbf{R}_+$; the only point (U,μ) in this cone with $\mu = 0$ has $U = 0$ (since $\mathcal{U}$ is compact), and

$$\mathcal{U} = \{U : (U,1) \in \mathbf{U}\} = \{U : \exists \mu \le 1 : (U,\mu) \in \mathbf{U}\}$$

so that the definition of $\overline{\mathrm{Opt}}_{\mathcal{U}}$ can be rewritten equivalently as

$$\overline{\mathrm{Opt}}_{\mathcal{U}}(B) = \min_{U,\mu} \left\{\mathrm{Tr}(BU) : (U,\mu) \in \mathbf{U}, \mu \le 1\right\}.$$

The question, of course, is where to take a set $\mathcal{U}$ matching $\mathcal{X}$, and the answer depends on what we know about $\mathcal{X}$. For example, when $\mathcal{X}$ is a basic ellitope:

$$\mathcal{X} = \{x \in \mathbf{R}^n : \exists t \in \mathcal{T} : x^T S_k x \le t_k, k \le K\}$$

with our usual restrictions on $\mathcal{T}$ and S_k, it is immediately seen that

$$x \in \mathcal{X} \Rightarrow xx^T \in \mathcal{U} = \{U \in \mathbf{S}^n : U \succeq 0, \exists t \in \mathcal{T} : \mathrm{Tr}(US_k) \le t_k, k \le K\}.$$

Similarly, when $\mathcal{X}$ is a basic spectratope:

$$\mathcal{X} = \{x \in \mathbf{R}^n : \exists t \in \mathcal{T} : S_k^2[x] \preceq t_k I_{d_k}, k \le K\}$$

with our usual restrictions on $\mathcal{T}$ and $S_k[\cdot]$, it is immediately seen that

$$x \in \mathcal{X} \Rightarrow xx^T \in \mathcal{U} = \{U \in \mathbf{S}^n : U \succeq 0, \exists t \in \mathcal{T} : \mathcal{S}_k[U] \preceq t_k I_{d_k}, k \le K\}.$$

One can verify that the semidefinite relaxation bounds on the maximum of a quadratic form on an ellitope/spectratope $\mathcal{X}$ derived in Sections 4.3.4 (for ellitopes) and 4.4.2.2 (for spectratopes, see Lemma 4.4.1) are nothing but the bounds (4.9.10) associated with the just defined $\mathcal{U}$.

4.11.A Matching via absolute norms. There are other ways to specify a set matching $\mathcal{X}$. The seemingly simplest of them is as follows. Let $p(\cdot)$ be an absolute norm on $\mathbf{R}^n$ (recall that it is a norm $p(x)$ which depends solely on abs$[x]$, where abs$[x]$ is the vector comprised of the magnitudes of entries in x). We can convert $p(\cdot)$ into the norm $p^+(\cdot)$ on the space $\mathbf{S}^n$, namely, a

$$p^+(U) = p([p(\mathrm{Col}_1[U]);...;p(\mathrm{Col}_n[U])]) \qquad [U \in \mathbf{S}^n]$$

1.1. *Prove that p^+ indeed is a norm on $\mathbf{S}^n$, and $p^+(xx^T) = p^2(x)$. Denoting by $q(\cdot)$ the norm conjugate to $p(\cdot)$, what is the relation between the norm $(p^+)_*(\cdot)$ conjugate to $p^+(\cdot)$ and the norm $q^+(\cdot)$?*

1.2. *Derive from 1.1 that whenever $p(\cdot)$ is an absolute norm such that $\mathcal{X}$ is contained in the unit ball $\mathcal{B}_{p(\cdot)} = \{x : p(x) \le 1\}$ of the norm p, the set*

$$\mathcal{U}_{p(\cdot)} = \{U \in \mathbf{S}^n : U \succeq 0, p^+(U) \le 1\}$$

is matching $\mathcal{X}$. If, in addition,

$$\mathcal{X} \subset \{x : p(x) \le 1, Px = 0\}, \tag{4.9.11}$$

then the set

$$\mathcal{U}_{p(\cdot),P} = \{U \in \mathbf{S}^n : U \succeq 0, p^+(U) \le 1, PU = 0\}$$

is matching $\mathcal{X}$.

Assume that in addition to $p(\cdot)$, we have at our disposal a computationally tractable closed convex set $\mathcal{D}$ such that whenever $p(x) \le 1$, the vector $[x]^2 := [x_1^2;...;x_n^2]$ belongs to $\mathcal{D}$; in the sequel we call such a set $\mathcal{D}$ *square-dominating* $p(\cdot)$. For example, when $p(\cdot) = \|\cdot\|_r$, we can take

$$\mathcal{D} = \left\{ \begin{array}{ll} \{y \in \mathbf{R}^n_+ : \sum_i y_1 \le 1\}, & r \le 2 \\ \{y \in \mathbf{R}^n_+ : \|y\|_{r/2} \le 1\}, & r > 2 \end{array} \right. .$$

Prove that in this situation the above construction can be refined: whenever $\mathcal{X}$ satisfies (4.9.11)*, the set*

$$\mathcal{U}^{\mathcal{D}}_{p(\cdot),P} = \{U \in \mathbf{S}^n : U \succeq 0, p^+(U) \le 1, PU = 0, \mathrm{diag}(U) \in \mathcal{D}\} \quad [\mathrm{diag}(U) = [U_{11}; U_{22}; ...; U_{nn}]]$$

matches $\mathcal{X}$.

Note: in the sequel, we suppress P in the notation $\mathcal{U}_{p(\cdot),P}$ and $\mathcal{U}^{\mathcal{D}}_{p(\cdot),P}$ when $P = 0$; thus, $\mathcal{U}_{p(\cdot)}$ is the same as $\mathcal{U}_{p(\cdot),0}$.

1.3. *Check that when $p(\cdot) = \|\cdot\|_r$ with $r \in [1, \infty]$, one has*

$$p^+(U) = \|U\|_r := \left\{ \begin{array}{ll} (\sum_{i,j} |U_{ij}|^r)^{1/r}, & 1 \le r < \infty, \\ \max_{i,j} |U_{ij}|, & r = \infty \end{array} \right. ,$$

1.4. *Let $\mathcal{X} = \{x \in \mathbf{R}^n : \|x\|_1 \le 1\}$ and $p(x) = \|x\|_1$, so that $\mathcal{X} \subset \{x : p(x) \le 1\}$, and*

$$\mathrm{Conv}\{[x]^2 : x \in \mathcal{X}\} \subset \mathcal{D} = \{y \in \mathbf{R}^n_+ : \sum_i y_1 = 1\}. \tag{4.9.12}$$

What are the bounds $\overline{\mathrm{Opt}}_{\mathcal{U}_{p(\cdot)}}(B)$ and $\overline{\mathrm{Opt}}_{\mathcal{U}^{\mathcal{D}}_{p(\cdot)}}(B)$? Is it true that the former (the latter) of the bounds is precise? Is it true that the former (the latter) of the bounds is precise when $B \succeq 0$?

1.5. Let $\mathcal{X} = \{x \in \mathbf{R}^n : \|x\|_2 \leq 1\}$ *and* $p(x) = \|x\|_2$, *so that* $\mathcal{X} \subset \{x : p(x) \leq 1\}$ *and* (4.9.12) *holds true. What are the bounds* $\overline{\mathrm{Opt}}_{\mathcal{U}_{p(\cdot)}}(B)$ *and* $\overline{\mathrm{Opt}}_{\mathcal{U}^{\mathcal{D}}_{p(\cdot)}}(B)$ *? Is it true that the former (the latter) of the bounds is precise?*

1.6. Let $\mathcal{X} \subset \mathbf{R}^n_+$ *be closed, convex, bounded, and with a nonempty interior. Verify that the set*

$$\mathcal{X}^+ = \{x \in \mathbf{R}^n : \exists y \in \mathcal{X} : \mathrm{abs}[x] \leq y\}$$

is the unit ball of an absolute norm $p_{\mathcal{X}}$, *and this is the largest absolute norm* $p(\cdot)$ *such that* $\mathcal{X} \subset \{x : p(x) \leq 1\}$. *Derive from this observation that the norm* $p_{\mathcal{X}}(\cdot)$ *is the best (i.e., resulting in the least conservative bounding scheme) among absolute norms which allow to upper-bound* $\mathrm{Opt}_{\mathcal{X}}(B)$ *via the construction from item 1.2.*

4.11.B "Calculus of matchings." Observe that matching we have introduced admits a kind of "calculus." Specifically, consider the situation as follows: for $1 \leq \ell \leq L$, we are given

- nonempty convex compact sets $\mathcal{X}_\ell \subset \mathbf{R}^{n_\ell}$, $0 \in \mathcal{X}_\ell$, along with matching $\mathcal{X}_\ell$ convex compact sets $\mathcal{U}_\ell \subset \mathbf{S}^{n_\ell}$ giving rise to the closed convex cones

$$\mathbf{U}_\ell = \mathrm{clo}\{(U_\ell, \mu_\ell) \in \mathbf{S}^{n_\ell} \times \mathbf{R}_+ : \mu_\ell > 0, \mu_\ell^{-1} U_\ell \in \mathcal{U}_\ell\}$$

  We denote by $\vartheta_\ell(\cdot)$ the Minkovski functions of $\mathcal{X}_\ell$:

$$\vartheta_\ell(y^\ell) = \inf\{t : t > 0, t^{-1} y^\ell \in \mathcal{X}_\ell\} : \mathbf{R}^{n_\ell} \to \mathbf{R} \cup \{+\infty\};$$

  note that $\mathcal{X}_\ell = \{y^\ell : \vartheta_\ell(y^\ell) \leq 1\}$;

- $n_\ell \times n$ matrices A_ℓ such that $\sum_\ell A_\ell^T A_\ell \succ 0$.

On the top of it, we are given a monotone convex set $\mathcal{T} \subset \mathbf{R}^L_+$ intersecting the interior of $\mathbf{R}^L_+$.

These data specify the convex set

$$\mathcal{X} = \{x \in \mathbf{R}^n : \exists t \in \mathcal{T} : \vartheta_\ell^2(A_\ell x) \leq t_\ell, \ell \leq L\} \qquad (*)$$

2.1. Prove the following

Lemma 4.9.1 *In the situation in question, the set*

$$\mathcal{U} = \left\{U \in \mathbf{S}^n : U \succeq 0 \ \&\ \exists t \in \mathcal{T} : (A_\ell U A_\ell^T, t_\ell) \in \mathbf{U}_\ell, \ell \leq L\right\}$$

is a closed and bounded convex set which matches $\mathcal{X}$. *As a result, the efficiently computable quantity*

$$\overline{\mathrm{Opt}}_{\mathcal{U}}(B) = \max_U \{\mathrm{Tr}(BU) : U \in \mathcal{U}\}$$

is an upper bound on

$$\mathrm{Opt}_{\mathcal{X}}(B) = \max_{x \in \mathcal{X}} x^T B x.$$

2.2. Prove that if $\mathcal{X} \subset \mathbf{R}^n$ *is a nonempty convex compact set,* U *is* $m \times n$ *matrix of rank* m, *and* $\mathcal{U}$ *is matching* $\mathcal{X}$, *then the set* $\mathcal{V} = \{V \in \mathcal{S}^m : V \succeq 0, PVP^T \in \mathcal{U}\}$ *matches* $\mathcal{Y} = \{y : \exists x \in \mathcal{X} : y = Px\}$.

2.3. Consider the "direct product" case where $\mathcal{X} = \mathcal{X}_1 \times ... \times \mathcal{X}_L$; specifying A_ℓ as the matrix which "cuts of" a block vector $x = [x^1; ...; x^L] \in \mathbf{R}^{n_1} \times ... \times \mathbf{R}^{n_L}$ ℓ-th block: $A_\ell x = x^\ell$ and setting $\mathcal{T} = [0,1]^L$, we cover this situation by the setup under consideration. In the direct product case, the construction from item 2.1 is as follows: given the sets $\mathcal{U}_\ell$ matching $\mathcal{X}_\ell$, we build the set

$$\mathcal{U} = \{U = [U^{\ell\ell'} \in \mathbf{R}^{n_\ell \times n_{\ell'}}]_{\ell,\ell' \leq L} \in \mathbf{S}^{n_1+...+n_L} : U \succeq 0, U^{\ell\ell} \in \mathcal{U}_\ell, \ell \leq L\}$$

and claim that this set matches $\mathcal{X}$.

Could we be less conservative? While we do not know how to be less conservative in general, we do know how to be less conservative in the special case when $\mathcal{U}_\ell$ are built via the "absolute norm" machinery. Specifically, let $p_\ell(\cdot) : \mathbf{R}^{n_\ell} \to \mathbf{R}_+$, $\ell \leq L$, be absolute norms, let sets $\mathcal{D}_\ell$ be square-dominating $p_\ell(\cdot)$, let

$$\mathcal{X}^\ell \subset \widehat{X}_\ell = \{x^\ell \in \mathbf{R}^{n_\ell} : P_\ell x_\ell = 0, p_\ell(x^\ell) \leq 1\},$$

and let $\mathcal{U}_\ell = \{U \in \mathbf{S}^{n_\ell} : U \succeq 0, P_\ell U = 0, p_\ell^+(U) \leq 1, \mathrm{diag}(U) \in \mathcal{D}_\ell\}$. In this case the above construction results in

$$\mathcal{U} = \left\{U = [U^{\ell\ell'} \in \mathbf{R}^{n_\ell \times n_{\ell'}}]_{\ell,\ell' \leq L} \in \mathbf{S}^{n_1+...+n_L} : U \succeq 0, P_\ell U^{\ell\ell} = 0, p_\ell^+(U^{\ell\ell}) \leq 1, \mathrm{diag}(U^{\ell\ell}) \in \mathcal{D}_\ell, \ell \leq L\right\}.$$

Now let

$$p([x^1; ...; x^L]) = \max[p_1(x^1), ..., p_L(x^L)] : \mathbf{R}^{n_1} \times ... \times \mathbf{R}^{n_L} \to \mathbf{R},$$

so that p is an absolute norm and $\mathcal{X} \subset \{x = [x^1; ...; x^L] : p(x) \leq 1, P_\ell x^\ell = 0, \ell \leq L\}$.

Prove that in fact the set

$$\overline{\mathcal{U}} = \left\{U = [U^{\ell\ell'} \in \mathbf{R}^{n_\ell \times n_{\ell'}}]_{\ell,\ell' \leq L} \in \mathbf{S}^{n_1+...+n_L} : U \succeq 0, P_\ell U^{\ell\ell} = 0, \mathrm{diag}(U^{\ell\ell}) \in \mathcal{D}_\ell, \ell \leq L, p^+(U) \leq 1\right\}$$

matches $\mathcal{X}$, and that we always have $\overline{\mathcal{U}} \subset \mathcal{U}$. Verify that in general this inclusion is strict.

4.11.C Illustration: Nullspace property revisited. Recall sparsity-oriented signal recovery via ℓ_1 minimization from Lecture 1: Given $m \times n$ sensing matrix A and (noiseless) observation $y = Aw$ of unknown signal w known to have at most s nonzero entries, we recover w as

$$\widehat{w} \in \mathop{\mathrm{Argmin}}_z \{\|z\|_1 : Az = y\}.$$

Matrix A is called s-good, if whenever $y = Aw$ with s-sparse w, the only optimal solution to the right hand side optimization problem is w. The (difficult to verify!) necessary and sufficient condition for s-goodness is the Nullspace property:

$$\mathrm{Opt} := \max_z \left\{\|z\|_{(s)} : z \in \mathrm{Ker}\, A, \|z\|_1 \leq 1\right\} < 1/2,$$

where $\|z\|_{(k)}$ is the sum of the k largest entries in the vector abs$[z]$. A verifiable sufficient condition for s-goodness is

$$\widehat{\mathrm{Opt}} := \min_H \max_j \|\mathrm{Col}_j[I - H^T A]\|_{(s)} < 1/2, \qquad (!)$$

the reason being that, as it is immediately seen, $\widehat{\mathrm{Opt}}$ is an upper bound on Opt (see Proposition 1.3.3 with $q = 1$).

An immediate observation is that Opt is nothing but the maximum of quadratic form on appropriate convex compact set. Specifically, let

$$\mathcal{X} = \{[u; v] \in \mathbf{R}^n \times \mathbf{R}^n : Au = 0, \|u\|_1 \leq 1, \sum_i |v_i| \leq s, \|v\|_* \leq 1\}, \; B = \left[\begin{array}{c|c} & \frac{1}{2}I_n \\ \hline \frac{1}{2}I_n & \end{array}\right].$$

Then

$$\begin{array}{rcl}
\mathrm{Opt}_{\mathcal{X}}(B) & = & \max\limits_{[u;v]\in\mathcal{X}} [u;v]^TB[u;v] = \max\limits_{u,v}\left\{u^Tv : Au=0, \|u\|_1\leq 1, \sum_i |v_i|\leq s, \|v\|_\infty\leq 1\right\}\\
& \underbrace{=}_{(a)} & \max_u \left\{\|u\|_{(s)} : Au=0, \|u\|_1\leq 1\right\}\\
& = & \mathrm{Opt},
\end{array}$$

where (a) is due to the well known fact (prove it!) that *whenever s is a positive integer $\leq n$, the extreme points of the set*

$$V=\{v\in\mathbf{R}^n : \sum_i |v_i|\leq s, \|v\|_\infty\leq 1\}$$

are exactly the vectors with at most s nonzero entries, the nonzero entries being ±1; as a result

$$\forall (z\in\mathbf{R}^n) : \max_{v\in V} z^Tv = \|z\|_{(s)}.$$

Now, V is the unit ball of the absolute norm

$$r(v)=\min\left\{t : \|v\|_1\leq st, \|v\|_\infty\leq t\right\},$$

so that $\mathcal{X}$ is contained in the unit ball $\mathcal{B}$ of the absolute norm on $\mathbf{R}^{2n}$ specified as

$$p([u;v])=\max\left\{\|u\|_1, r(v)\right\} \qquad [u,v\in\mathbf{R}^n],$$

specifically,

$$\mathcal{X}=\left\{[u;v] : p([u,v])\leq 1, Au=0\right\}.$$

As a result, whenever $x=[u;v]\in\mathcal{X}$, the matrix

$$U=xx^T=\left[\begin{array}{c|c} U^{11}=uu^T & U^{12}=uv^T\\ \hline U^{21}=vu^T & U^{22}=vv^T\end{array}\right]$$

satisfies the condition $p^+(U)\leq 1$ (see item 1.2 above). In addition, this matrix clearly satisfies the condition

$$A[U^{11},U^{12}]=0.$$

It follows that the set

$$\mathcal{U}=\{U=\left[\begin{array}{c|c} U^{11} & U^{12}\\ \hline U^{21} & U^{22}\end{array}\right]\in\mathbf{S}^{2n} : U\succeq 0, p^+(U)\leq 1, AU^{11}=0, AU^{12}=0\}$$

(which clearly is a nonempty convex compact set) matches $\mathcal{X}$. As a result, the efficiently computable quantity

$$\overline{\mathrm{Opt}}=\max_{U\in\mathcal{U}}\mathrm{Tr}(BU)=\max_U\left\{\mathrm{Tr}(U^{12}) : U=\left[\begin{array}{c|c} U^{11} & U^{12}\\ \hline U^{21} & U^{22}\end{array}\right]\succeq 0, p^+(U)\leq 1, AU^{11}=0, AU^{12}=0\right\} \tag{!!}$$

is an upper bound on Opt, so that the verifiable condition

$$\overline{\mathrm{Opt}}<1/2$$

is sufficient for s-goodness of A.

Now goes the concluding part of Exercise:

3.1. *Prove that $\overline{\mathrm{Opt}}\leq\widehat{\mathrm{Opt}}$, so that* (!!) *is less conservative than* (!).
Hint: *Apply Conic Duality to verify that*

$$\widehat{\mathrm{Opt}}=\max_V\left\{\mathrm{Tr}(V) : V\in\mathbf{R}^{n\times n}, AV=0, \sum_{i=1}^n r(\mathrm{Col}_i[V^T])\leq\right\} \tag{!!!}$$

3.2. *Run simulations with randomly generated Gaussian matrices A and play with different values of s to compare $\widehat{\mathrm{Opt}}$ and $\overline{\mathrm{Opt}}$. To save time, you can use toy sizes m,n, say, $m=18, n=24$.*

4.9.4 Around Propositions 4.2.1 and 4.5.1

4.9.4.1 Optimizing linear estimates on convex hulls of unions of spectratopes

Exercise 4.12 [optimizing linear estimates on convex hull of union of spectratopes] Let

- $\mathcal{X}_1,...,\mathcal{X}_J$ be spectratopes in $\mathbf{R}^n$:

$$\begin{array}{c}\mathcal{X}_j = \{x \in \mathbf{R}^n : \exists (y \in \mathbf{R}^{N_j}, t \in \mathcal{T}_j) : x = P_j y, R_{kj}^2[y] \preceq t_k I_{d_{kj}}, \leq K_j\},\ 1 \leq j \leq J\\ \left[R_{kj}[y] = \sum_{i=1}^{N_j} y_i R^{kji}\right]\end{array}$$

- $A \in \mathbf{R}^{m\times n}$ and $B \in \mathbf{R}^{\nu\times n}$ be given matrices,
- $\|\cdot\|$ be a norm on $\mathbf{R}^\nu$ such that the unit ball $\mathcal{B}_*$ of the conjugate norm $\|\cdot\|_*$ is a spectratope:

$$\begin{array}{c}\mathcal{B}_* := \{u : \|u\|_* \leq 1\} = \{u \in \mathbf{R}^\nu : \exists (z \in \mathbf{R}^N, r \in \mathcal{R}) : u = Mz, S_\ell^2[z] \preceq r_\ell I_{f_\ell}, \ell \leq L\}\\ \left[S_\ell[z] = \sum_{i=1}^{N} z_i S^{\ell i}\right]\end{array}$$

- Π be a convex compact subset of the interior of the positive semidefinite cone $\mathbf{S}^m_+$,

with our standard restrictions on $R_{kj}[\cdot]$, $S_\ell[\cdot]$, $\mathcal{T}_j$, $\mathcal{R}$. Let, further,

$$\mathcal{X} = \mathrm{Conv}\left(\bigcup_j \mathcal{X}_j\right)$$

be the convex hull of the union of spectratopes $\mathcal{X}_j$. Consider the situation where we, given observation

$$\omega = Ax + \xi$$

of unknown signal x known to belong to $\mathcal{X}$, want to recover Bx. We assume that the matrix of second moments of noise is $\succeq$-dominated by a matrix from Π, and quantify the performance of a candidate estimate $\widehat{x}(\cdot)$ by its $\|\cdot\|$-*risk*

$$\mathrm{Risk}_{\Pi,\|\cdot\|}[\widehat{x}|\mathcal{X}] = \sup_{x\in\mathcal{X}} \sup_{P:P\lll\Pi} \mathbf{E}_{\xi\sim P}\left\{\|Bx - \widehat{x}(Ax+\xi)\|\right\}$$

where $P \lll \Pi$ means that the matrix $\mathrm{Vary}[P] = \mathbf{E}_{\xi\sim P}\{\xi\xi^T\}$ of second moments of distribution P is $\succeq$-dominated by a matrix from Π.

Prove the following

Proposition 4.9.2 *In the situation in question, consider convex optimization problem*

$$\begin{array}{rcl}\mathrm{Opt} &=& \min\limits_{H,\Theta,\Lambda^j,\Upsilon^j,\Upsilon'} \left\{ \max\limits_j \left[\phi_{\mathcal{T}_j}(\lambda[\Lambda^j]) + \phi_{\mathcal{R}}(\lambda[\Upsilon^j])\right] + \phi_{\mathcal{R}}(\lambda[\Upsilon']) + \Gamma_\Pi(\Theta) : \right.\\ && \left.\begin{array}{c}\Lambda^j = \{\Lambda^j_k \succeq 0, j \leq K_j\}, j \leq J, \Upsilon^j = \{\Upsilon^j_\ell \succeq 0, \ell \leq L\}, j \leq J, \Upsilon' = \{\Upsilon'_\ell \succeq 0, \ell \leq L\}\\ \left[\begin{array}{c|c}\sum_k \mathcal{R}^*_{kj}[\Lambda^j_k] & \frac{1}{2}P_j^T[B^T - A^T H]M\\ \hline \frac{1}{2}M^T[B - H^T A]P_j & \sum_\ell \mathcal{S}^*_\ell[\Upsilon^j_\ell]\end{array}\right] \succeq 0, j \leq J, \left[\begin{array}{c|c}\Theta & \frac{1}{2}HM\\ \hline \frac{1}{2}M^T H^T & \sum_\ell \mathcal{S}^*_\ell[\Upsilon'_\ell]\end{array}\right] \succeq 0\end{array}\right\},\end{array} \tag{4.9.13}$$

where, as usual,

$$\begin{array}{c}\phi_{\mathcal{T}_j}(\lambda) = \max\limits_{t\in\mathcal{T}_j} t^T\lambda,\ \phi_{\mathcal{R}}(\lambda) = \max\limits_{r\in\mathcal{R}} r^T\lambda,\ \Gamma_\Pi(\Theta) = \max\limits_{Q\in\Pi} \mathrm{Tr}(Q\Theta),\ \lambda[U_1,...,U_s] = [\mathrm{Tr}(U_1);...;\mathrm{Tr}(U_S)],\\ \mathcal{S}^*_\ell[\cdot] : \mathbf{S}^{f_\ell} \to \mathbf{S}^N : \mathcal{S}^*_\ell[U] = \left[\mathrm{Tr}(S^{\ell p} U S^{\ell q})\right]_{p,q\leq N}, \mathcal{R}^*_{kj}[\cdot] : \mathbf{S}^{d_{kj}} \to \mathbf{S}^{N_j} : \mathcal{R}^*_{kj}[U] = \left[\mathrm{Tr}(R^{kjp} U R^{kjq})\right]_{p,q\leq N_j}\end{array}$$

Problem (4.9.13) *is solvable, and H-component H_* of its optimal solution gives rise to linear estimate $\widehat{x}_{H_*}(\omega) = H_*^T\omega$ such that*

$$\mathrm{Risk}_{\Pi,\|\cdot\|}[\widehat{x}_{H_*}|\mathcal{X}] \leq \mathrm{Opt}. \tag{4.9.14}$$

Moreover, the estimate $\widehat{x}_{H_*}$ is near-optimal among linear estimates:

$$\begin{array}{c}\mathrm{Opt} \leq O(1)\ln(D+F)\mathrm{RiskOpt}_{\mathrm{lin}}\\ \left[D = \max_j \sum_{k\leq K_j} d_{kj},\ F = \sum_{\ell\leq L} f_\ell\right]\end{array} \tag{4.9.15}$$

where

$$\mathrm{RiskOpt}_{\mathrm{lin}} = \inf_H \sup_{x\in\mathcal{X},Q\in\Pi} \mathbf{E}_{\xi\sim\mathcal{N}(0,Q)}\left\{\|Bx - H^T(Ax+\xi)\|\right\}$$

is the best risk achievable under the circumstances with linear estimates under zero mean Gaussian noise with covariance matrix restricted to belong to Π.

It should be stressed that convex hull of unions of spectratopes not necessarily is a spectratope, and that Proposition states that the linear estimate stemming from (4.9.13) is near-optimal only among linear, and not among all estimates (the latter can indeed be not the case).

4.9.4.2 Recovering nonlinear vector-valued functions

Exercise 4.13 † [estimating nonlinear vector-valued functions] Consider situation as follows: We are given a noisy observation

$$\omega = Ax + \xi_x \qquad [A \in \mathbf{R}^{\nu\times n}]$$

of the linear image Ax of an unknown signal x known to belong to a given spectratope $\mathcal{X} \subset \mathbf{R}^n$; here ξ_x is the observation noise with distribution P_x which can depend on x. Similarly to Section 4.5, we assume that we are given a computationally tractable convex compact set $\Pi \subset \mathrm{int}\,\mathbf{S}^\nu_+$ such that for every $x \in \mathcal{X}$, $\mathrm{Vary}[P_x] \preceq \Theta$ for some $\Theta \in \Pi$, cf. (4.5.3). What we want is to recover the value $f(x)$ of a given vector-valued function $f : \mathcal{X} \to \mathbf{R}^\nu$, and we measure the recovery error in a given norm $|\cdot|$ on $\mathbf{R}^\nu$.

4.13.A Preliminaries and Main observation. Let $\|\cdot\|$ be a norm on $\mathbf{R}^n$, and $g(\cdot) : \mathcal{X} \to \mathbf{R}^\nu$ be a function. Recall that the function is called *Lipschitz continuous on $\mathcal{X}$ w.r.t. the pair of norms $\|\cdot\|$ on the argument and $|\cdot|$ on the image spaces*, if there exist $L < \infty$ such that

$$|g(x) - g(y)| \leq L\|x - y\|\ \forall (x, y \in \mathcal{X});$$

every L with this property is called Lipschitz constant of g. It is well known that in our finite-dimensional situation, the property of g to be Lipschitz continuous is independent of how the norms $\|\cdot\|$, $|\cdot|$ are selected; this selection affects only the value(s) of Lipschitz constant(s).

Assume from now on that the function of interest f is Lipschitz continuous on $\mathcal{X}$. Let us call a norm $\|\cdot\|$ on $\mathbf{R}^n$ *appropriate* for f, is f is Lipschitz continuous *with constant 1* on $\mathcal{X}$ w.r.t. $\|\cdot\|$, $|\cdot|$. Our immediate observation is as follows:

Observation 4.9.1 *In the situation in question, let $\|\cdot\|$ be appropriate for f. Then recovering $f(x)$ is not more difficult than recovering x in the norm $\|\cdot\|$: every estimate $\widehat{x}(\omega)$ of x via ω which takes all its values in $\mathcal{X}$ induces the "plug-in" estimate*

$$\widehat{f}(\omega) = f(\widehat{x}(\omega))$$

of $f(x)$, and the $\|\cdot\|$-risk

$$\mathrm{Risk}_{\|\cdot\|}[\widehat{x}|\mathcal{X}] = \sup_{x\in\mathcal{X}} \mathbf{E}_{\xi\sim P_x}\left\{\|\widehat{x}(Ax+\xi) - x\|\right\}$$

of estimate $\widehat{x}$ upper-bounds the $|\cdot|$-risk

$$\mathrm{Risk}_{|\cdot|}[\widehat{f}|\mathcal{X}] = \sup_{x\in\mathcal{X}} \mathbf{E}_{\xi\sim P_x}\left\{\|\widehat{f}(Ax+\xi)-f(x)\|\right\}$$

of the induced by $\widehat{x}$ estimate $\widehat{f}$:

$$\mathrm{Risk}_{|\cdot|}[\widehat{f}|\mathcal{X}] \leq \mathrm{Risk}_{\|\cdot\|}[\widehat{x}|\mathcal{X}].$$

When f is defined and Lipschitz continuous with constant 1 w.r.t. $\|\cdot\|, |\cdot|$ on the entire $\mathbf{R}^n$, the conclusion remains true without the assumption that $\widehat{x}$ takes all its values in $\mathcal{X}$.

4.13.B Consequences. Observation 4.9.1 suggests the following simple approach to solving the estimation problem we started with: assuming that we have at our disposal a norm $\|\cdot\|$ on $\mathbf{R}^n$ such that

- $\|\cdot\|$ is appropriate for f, and
- $\|\cdot\|$ is *good*, goodness meaning that the unit ball $\mathcal{B}_*$ of the norm $\|\cdot\|_*$ *conjugate* to $\|\cdot\|$ is a spectratope given by explicit spectratopic representation,

we use the machinery of linear estimation developed in Section 4.5 to build a near-optimal, in terms of its $\|\cdot\|$-risk, linear estimate of x via ω, and convert this estimate in an estimate of $f(x)$; by Observation, the $|\cdot|$- risk of the resulting estimate is upper-bounded by $\|\cdot\|$-risk of the underlying linear estimate. The just outlined construction needs a small correction: in general, the linear estimate $\widetilde{x}(\cdot)$ yielded by Proposition 4.5.1 (same as any nontrivial – not identically zero – *linear* estimate) is *not* guaranteed to take all its values in $\mathcal{X}$, which is, in general, required for Observation to be applicable. This correction is easy: it is enough to convert $\widetilde{x}$ into the estimate $\widehat{x}$ defined by

$$\widehat{x}(\omega) \in \mathop{\mathrm{Argmin}}_{u\in\mathcal{X}} \|u-\widetilde{x}(\omega)\|.$$

This transformation preserves efficient computability of the estimate, and ensures that the corrected estimate takes all its values in $\mathcal{X}$, so that Observation is applicable to $\widehat{x}$; at the same time, "correction" $\widetilde{x}\mapsto\widehat{x}$ nearly preserves the $\|\cdot\|$-risk:

$$\mathrm{Risk}_{\|\cdot\|}[\widehat{x}|\mathcal{X}] \leq 2\mathrm{Risk}_{\|\cdot\|}[\widetilde{x}|\mathcal{X}]. \tag{$*$}$$

Note that when $\|\cdot\|$ is a (general-type) Euclidean norm: $\|x\|^2 = x^TQx$ for some $Q\succ 0$, factor 2 in the right hand side can be discarded.

1. *Justify* $(*)$.

4.13.C How to select $\|\cdot\|$. When implementing the outlined approach, the major question is how to select a norm $\|\cdot\|$ appropriate for f. An ideal for our purposes choice would be to select the smallest among the norms appropriate for f (such a norm does exist under pretty mild assumptions), since the smaller $\|\cdot\|$, the smaller is the $\|\cdot\|$-risk of an estimate of x. This ideal can be achieved in rare cases only: first, it could be difficult to identify the smallest among the norms appropriate for f, and second, our machinery requires from $\|\cdot\|$ to have an explicitly given spectratope as the unit ball of the conjugate norm. Let us look at a couple of "favorable cases," where the just outlined difficulties can be (partially) avoided.

Example 1: a norm-induced f. Let us start with the important by its own right case when f is a scalar functional which itself is a norm, and this norm has a spectratope as the unit ball of the conjugate norm, as is the case when $f(\cdot) = \|\cdot\|_r$, $r \in [1,2]$, or when $f(\cdot)$ is the nuclear norm. In this case the smallest of the norms appropriate for f clearly is f itself, and no one of the outlined difficulties arises. As an extension, when $f(x)$ is obtained from a good norm $\|\cdot\|$ by operations preserving Lipschitz continuity and constant, like $f(x) = \|x-c\|$, or $f(x) = \sum_i a_i\|x-c_i\|$, $\sum_i |a_i| \leq 1$, or

$$f(x) = \sup_{c\in C} / \inf \|x-c\|,$$

or even something like

$$f(x) = \sup_{\alpha\in\mathcal{A}} / \inf \left\{ \sup_{c\in C_\alpha} / \inf \|x-c\| \right\}$$

it seems natural to use this norm in our construction, although now this, perhaps, is not the smallest of the norms appropriate for f.

Now let us address the general case. Note that *in principle* the smallest of the norms appropriate for a given Lipschitz continuous f admits a description. Specifically, assume that $\mathcal{X}$ has a nonempty interior (this is w.l.o.g. – we can always replace $\mathbf{R}^n$ with the linear span of $\mathcal{X}$). A well-known fact of Analysis (Rademacher Theorem) states that in this situation (more generally, when $\mathcal{X}$ is convex with a nonempty interior), a Lipschitz continuous f is differentiable almost everywhere in $\mathcal{X}^o = \operatorname{int}\mathcal{X}$, and f is Lipschitz continuous with constant 1 w.r.t. a norm $\|\cdot\|$ if and only if

$$\|f'(x)\|_{\|\cdot\|\to|\cdot|} \leq 1$$

whenever $x \in \mathcal{X}^o$ is such that the derivative (a.k.a. Jacobian) of f at x exists; here $\|Q\|_{\|\cdot\|\to|\cdot|}$ is the matrix norm of a $\nu\times n$ matrix Q induced by the norms $\|\cdot\|$ on $\mathbf{R}^n$ and $|\cdot|$ on $\mathbf{R}^\nu$:

$$\|Q\|_{\|\cdot\|\to|\cdot|} := \max_{\|x\|\leq 1} |Qx| = \max_{\substack{\|x\|\leq 1\\ |y|_*\leq 1}} y^TQx = \max_{\substack{|y_*|\leq 1\\ [\|x\|_*]_*\leq 1}} x^TQ^Ty = \|Q^T_{|\cdot|_*\to\|\cdot\|_*},$$

where $\|\cdot\|_*$, $|\cdot|_*$ are the conjugates of $\|\cdot\|$, $|\cdot|$.

2. *Prove that a norm $\|\cdot\|$ is appropriate for f if and only if the unit ball of the conjugate to $\|\cdot\|$ norm contains the set*

$$\mathcal{B}_{f,*} = \operatorname{cl}\operatorname{Conv}\{z : \exists (x\in\mathcal{X}_o, y, |y|_*\leq 1) : z = [f'(x)]^Ty\},$$

where $\mathcal{X}_o$ is the set of all $x \in \mathcal{X}^o$ where $f'(x)$ exists. Geometrically: $\mathcal{B}_{f,*}$ is the closed convex hull of the union of all images of the unit ball $\mathcal{B}_*$ of $|\cdot|_*$ under the linear mappings $y \mapsto [f'(x)]^Ty$ stemming from $x \in \mathcal{X}_o$.

Equivalently: $\|\cdot\|$ is appropriate for f if and only if

$$\|u\| \geq \|u\|_f := \max_{z\in\mathcal{B}_{f,*}} z^Tu. \tag{!}$$

Check that $\|u\|_f$ is a norm, provided that $\mathcal{B}_{f,*}$ (this set by construction is a symmetric w.r.t. the origin convex compact set) possesses a nonempty interior; whenever this is the case, $\|u\|_f$ is the smallest of the norms appropriate for f.

Derive from the above that the norms $\|\cdot\|$ we can use in our approach are the norms on $\mathbf{R}^n$ for which the unit ball of the conjugate norm is a spectratope containing $\mathcal{B}_{f,*}$.

Example 2. Consider the case of componentwise quadratic f:

$$f(x) = [\frac{1}{2}x^TQ_1x; \frac{1}{2}x^TQ_2x; ...; \frac{1}{2}x^TQ_\nu x] \qquad [Q_i \in \mathbf{S}^n]$$

and $|u| = \|u\|_q$ with $q \in [1,2]$ [20]. In this case

$$\mathcal{B}_* = \{u \in \mathbf{R}^\nu : \|u\|_p \leq 1\},\ p = \frac{q}{q-1} \in [2,\infty[, \text{ and } f'(x) = \left[x^TQ_1; x^TQ_2; ...; x^TQ_\nu\right].$$

Setting $\mathcal{S} = \{s \in \mathbf{R}^\nu_+ : \|s\|_{p/2} \leq 1\}$ and $\mathcal{S}^{1/2} = \{s \in \mathbf{R}^\nu_+ : [s_1^2; ...; s_\nu^2] \in \mathcal{S}\} = \{s \in \mathbf{R}^\nu_+ : \|s\|_p \leq 1\}$, the set

$$\mathcal{Z} = \{[f'(x)]^Tu : x \in \mathcal{X}, u \in \mathcal{B}_*\}$$

is contained in the set

$$\mathcal{Y} = \{y \in \mathbf{R}^n : \exists (s \in \mathcal{S}^{1/2}, x^i \in \mathcal{X}, i \leq \nu) : y = \sum_i s_iQ_ix_i\},$$

and the set $\mathcal{Y}$ is a spectratope with spectratopic representation readily given by the one of $\mathcal{X}$; indeed, $\mathcal{Y}$ is nothing but the $\mathcal{S}$-sum of the spectratopes $Q_i\mathcal{X}$, $i = 1,...,\nu$, see Section 4.13. As a result, we can use the spectratope $\mathcal{Y}$ (when $\operatorname{int}\mathcal{Y} \neq \emptyset$) or the arithmetic sum of $\mathcal{Y}$ with a small Euclidean ball (when $\operatorname{int}\mathcal{Y} = \emptyset$) to build an estimate of $f(\cdot)$.

3.1. As a simple illustration, work out the problem of recovering the value of a scalar quadratic form

$$f(x) = x^TMx,\ M = \text{Diag}\{i^\alpha, i = 1,...,n\} \qquad [\nu = 1, |\cdot| \textit{ is the usual absolute value}]$$

from noisy observation

$$\omega = Ax + \sigma\eta,\ A = \text{Diag}\{i^\beta, i = 1,...,n\},\ \eta \sim \mathcal{N}(0, I_n)$$

of a signal x known to belong to the ellipsoid

$$\mathcal{X} = \{x \in \mathbf{R}^n : \|Px\|_2 \leq 1\},\ P = \text{Diag}\{i^\gamma, i = 1,...,n\},$$

where α, β, γ are given reals satisfying

$$2\alpha - \gamma - 2\beta < -1.$$

You could start with the simplest unbiased estimate

$$\widetilde{x}(\omega) = [1^{-\beta}\omega_1; 2^{-\beta}\omega_2; ...; n^{-\beta}\omega_n]$$

of x.

3.2. Work out the problem of recovering the norm

$$f(x) = \|Mx\|_p,\ M = \text{Diag}\{i^\alpha, i = 1,...,n\},\ p \in [1,2],$$

from the same observations as in item 3.1 and with

$$\mathcal{X} = \{x : \|Px\|_r \leq 1\},\ P = \text{diag}\{i^\gamma, i = 1,...,n\},\ r \in [2,\infty].$$

[20] to save notation, we assume that the linear parts in the components of f_i are trivial – just zeros. In this respect, note that we always can subtract from f a whatever linear mapping and reduce our estimation problem to those of estimating separately the values at the signal x of the modified f and the linear mapping we have subtracted (we know how to solve the latter problem reasonably well).

4.9.4.3 Suboptimal linear estimation

Exercise 4.14 † [recovery of large-scale signals] When building presumably good linear recovery of the image $Bx \in \mathbf{R}^\nu$ of signal $x \in \mathcal{X}$ from observation

$$\omega = Ax + \sigma\xi \in \mathbf{R}^m$$

in the simplest case where $\mathcal{X} = \{x \in \mathbf{R}^n : x^T Sx \leq 1\}$ is an ellipsoid (so that $S \succ 0$), the recovery error is measured in $\|\cdot\|_2$, and $\xi \sim \mathcal{N}(0, I_m)$, problem (4.2.9) reduces to

$$\mathrm{Opt} = \min_{H,\lambda}\left\{\lambda + \sigma^2\|H\|_F^2 : \left[\begin{array}{c|c}\lambda S & B^T - A^T H \\ \hline B - H^T A & I_\nu\end{array}\right] \succeq 0\right\}, \tag{4.9.16}$$

where $\|\cdot\|_F$ is the Frobenius norm of a matrix. An optimal solution H_* to this problem results in the linear estimate $\widehat{x}_{H_*}(\omega) = H_*^T\omega$ satisfying the risk bound

$$\mathrm{Risk}[\widehat{x}_{H_*}|\mathcal{X}] := \max_{x\in\mathcal{X}}\sqrt{\mathbf{E}\{\|Bx - H_*^T(Ax + \sigma\xi)\|_2^2\}} \leq \sqrt{\mathrm{Opt}}.$$

Now, (4.9.16) is an efficiently solvable convex optimization problem. However, when the sizes m, n of the problem are large, solving the problem by the standard optimization techniques could become prohibitively time-consuming. The goal of what follows is to develop relatively computationally cheap technique for finding a hopefully good *sub*optimal solution to (4.9.16). In the sequel, we assume that $A \neq 0$, otherwise (4.9.16) is trivial.

1. *Prove that problem (4.9.16) can be reduced to similar problem with $S = I_n$ and diagonal positive semidefinite matrix A, the reduction requiring several singular value decompositions and multiplications of matrices of the same sizes as those of A, B, S.*

2. By item 1, we can assume from the very beginning that $S = I$ and $A = \mathrm{Diag}\{\alpha_1, ..., \alpha_n\}$ with $0 \leq \alpha_1 \leq \alpha_2 \leq ... \leq \alpha_n$. Passing in (4.9.16) from variables λ, H to variables $\tau = \sqrt{\lambda}, G = H^T$, the problem becomes

$$\mathrm{Opt} = \min_{G,\tau}\left\{\tau^2 + \sigma^2\|G\|_F^2 : \|B - GA\| \leq \tau\right\}, \tag{4.9.17}$$

where $\|\cdot\|$ is the spectral norm. Now consider the construction as follows:

- We build a partition $\{1, ..., n\} = I_0 \cup I_1 \cup ... \cup I_K$ of the index set $\{1, ..., n\}$ into consecutive segments in such a way that
(a) I_0 is the set of those i, if any, for which $\alpha_i = 0$, and $I_k \neq \emptyset$ when $k \geq 1$,
(b) for $k \geq 1$ the ratios α_j/α_i, $i, j \in I_k$, do not exceed a $\theta > 1$ (θ is the parameter of our construction), while
(c) for $1 \leq k < k' \leq K$, the ratios α_j/α_i, $i \in I_k$, $j \in I_{k'}$, are $> \theta$.
The recipe for building the partition is self-evident, and we clearly have

$$K \leq \ln(\overline{\alpha}/\underline{\alpha})/\ln(\theta) + 1,$$

where $\overline{\alpha}$ is the largest of α_i, and $\underline{\alpha}$ is the smallest of those α_i which are positive.

- For $1 \leq k \leq K$, we denote by i_k the first index in I_k, set $\alpha^k = \alpha_{i_k}$, $n_k = \mathrm{Card}\, I_k$, and define A_k as $n_k \times n_k$ diagonal matrix with diagonal entries α_i, $i \in I_k$.

Now, given $\nu\times n$ matrix C, let us specify C_k, $0 \leq k \leq K$, as $\nu\times n_k$ submatrix of C comprised of columns with indexes from I_k, and consider the following parametric optimization problems:

$$\begin{array}{rcll}\mathrm{Opt}_k^*(\tau) & = & \min_{G_k\in\mathbf{R}^{\nu\times n_k}}\left\{\|G_k\|_F^2 : \|B_k - G_kA_k\| \leq \tau\right\} & (P_k^*[\tau]) \\ \mathrm{Opt}_k(\tau) & = & \min_{G_k\in\mathbf{R}^{\nu\times n_k}}\left\{\|G_k\|_F^2 : \|B_k - \alpha^k G_k\| \leq \tau\right\} & (P_k[\tau])\end{array}$$

where $\tau \geq 0$ is the parameter, and $1 \leq k \leq K$.

Justify the following simple observations:

2.1. G_k is feasible for $(P_k[\tau])$ if and only if the matrix

$$G_k^* = \alpha^k G_k A_k^{-1}$$

is feasible for $(P_k^*[\tau])$, and $\|G_k^*\|_F \leq \|G_k\|_F \leq \theta\|G_k^*\|_F$, implying that

$$\mathrm{Opt}_k^*(\tau) \leq \mathrm{Opt}_k(\tau) \leq \theta^2 \mathrm{Opt}_k^*(\tau);$$

2.2. Problems $(P_k[\tau])$ is easy to solve: if $B_k = U_k D_k V_k^T$ is singular value decomposition of B_k and $\sigma_{k\ell}$, $1 \leq \ell \leq \nu_k := \min[\nu, n_k]$, are diagonal entries of D_k, then an optimal solution to $(P_k[\tau])$ is

$$\widehat{G}_k[\tau] = [\alpha^k]^{-1} U_k D_k[\tau] V_k^T,$$

where $D_k[\tau]$ is obtained from D_k by truncating $\sigma_{k\ell} \mapsto [\sigma_{k\ell} - \tau]_+$ of diagonal entries and keeping zero the off-diagonal entries (from now on, $a_+ = \max[a, 0]$, $a \in \mathbf{R}$). The optimal value in $(P_k[\tau])$ is

$$\mathrm{Opt}_k(\tau) = [\alpha^k]^{-2} \sum_{\ell=1}^{\nu_k} [\sigma_{k\ell} - \tau]_+^2.$$

2.3. If (τ, G) is feasible solution to (4.9.17), then $\tau \geq \underline{\tau} := \|B_0\|$ and the matrices G_k, $1 \leq k \leq K$, are feasible solutions to problems $(P_k^*[\tau])$, implying that

$$\sum_k \mathrm{Opt}_k^*(\tau) \leq \|G\|_F^2,$$

and nearly vice versa: if $\tau \geq \underline{\tau}$, G_k, $1 \leq k \leq K$, are feasible solutions to problems $(P_k^*[\tau])$, and

$$K_+ = \left\{ \begin{array}{ll} K, & I_0 = \emptyset \\ K+1, & I_0 \neq \emptyset \end{array} \right.,$$

then the matrix $G = [0_{\nu \times n_0}, G_1, ..., G_k]$ taken along with $\tau^+ = \sqrt{K_+}\tau$ form a feasible solution to (4.9.17).

Extract from these observations that if τ_* is an optimal solution to the convex optimization problem

$$\min_\tau \left\{ \theta^2 \tau^2 + \sigma^2 \sum_{k=1}^K \mathrm{Opt}_k(\tau) : \tau \geq \underline{\tau} \right\} \tag{4.9.18}$$

and $G_{k,*}$ are optimal solutions to the problems $(P_k[\tau_*])$, then the pair

$$\widehat{\tau} = \sqrt{K_+}\tau_*, \widehat{G} = [0_{\nu \times n_0}, G_{1,*}^*, ..., G_{K,*}^*] \qquad [G_{k,*}^* = \alpha^k G_{k,*} A_k^{-1}]$$

is a feasible solution to (4.9.17), and the value of the objective of the latter problem at this feasible solution is within the factor $\max[K_+, \theta^2]$ of the true optimal value Opt of this problem. As a result, $\widehat{G}$ gives rise to a linear estimate with risk on $\mathcal{X}$ which is within factor $\max[\sqrt{K_+}, \theta]$ of the risk $\sqrt{\mathrm{Opt}}$ of the "presumably good" linear estimate yielded by an optimal solution to (4.9.16).

Pay attention to the facts that

- After carrying out singular value decompositions of matrices B_k, $1 \leq k \leq K$, specifying τ_* and $G_{k,*}$ requires solving univariate convex minimization problem with easy to compute objective, so that the problem can be easily solved, e.g., by bisection;

- The computationally cheap suboptimal solution we end up with is not that bad, since K is "moderate" – just logarithmic in the condition number $\overline{\alpha}/\underline{\alpha}$ of A.

Your next task is a follows:

3. *To get an idea of the performance of the proposed synthesis of "suboptimal" linear estimation, run numerical experiments as follows:*

   - *select somehow n and generate at random the $n \times n$ data matrices S, A, B*
   - *for "moderate" values of n compute both the presumably good linear estimate by solving (4.2.9)*[21] *and the suboptimal estimate as yielded by the above construction and compare their risk bounds and the associated CPU times. For "large" n, where solving (4.2.9) becomes prohibitively time consuming, compute only suboptimal estimate in order to get an impression how the corresponding CPU time grows with n.*

   Recommended setup:

   - *range of n: 50, 100 ("moderate" values), 1000, 2000 ("large" values)*
   - *range of σ: $\{1.0, 0.01, 0.0001\}$*
   - *generation of S, A, B: generate the matrices at random according to*

   $$S = U_S \mathrm{Diag}\{1, 2, ..., n\} U_S^T,\ A = U_A \mathrm{Diag}\{\mu_1, ..., \mu_n\} V_A^T,\ B = U_B \mathrm{Diag}\{\mu_1, ..., \mu_n\} V_B^T,$$

   where U_S, U_A, V_A, U_B, V_B are random orthogonal $n \times n$ matrices, and μ_i form a geometric progression with $\mu_1 = 0.01$ and $\mu_n = 1$.

   You could run the above construction for several values of θ and select the best, in terms of its risk bound, of the resulting suboptimal estimates.

4.14.A Simple case. There is a trivial case where (4.9.17) is really easy; this is the case in the singular value decompositions of A and B the right orthogonal factors are the same, that is, when

$$B = WFV^T,\ A = UDV^T$$

with orthogonal $n \times n$ matrices W, U, V and diagonal F, D. This, at the first glance, very special case is in fact of some importance – it covers the *denoising* situation where $B = A$, so that our goal is to denoise our observation of Ax given a priori information $x \in \mathcal{X}$ on x. In this situation, setting $W^T H^T U = G$, problem (4.9.17) becomes

$$\mathrm{Opt} = \min_G \left\{ \|F - GD\|^2 + \sigma^2 \|G\|_F^2 \right\}. \tag{4.9.19}$$

Now goes the concluding part of Exercise:

4. *Prove that in the situation in question an optimal solution G_* to (4.9.19) can be selected to be diagonal, with diagonal entries γ_i, $1 \le i \le n$, yielded by the optimal solution to the optimization problem*

$$\mathrm{Opt} = \min_\gamma \left\{ f(G) := \max_{i \le n} (\phi_i - \gamma_i \delta_i)^2 + \sigma^2 \sum_{i=1}^n \gamma_i^2 \right\} \qquad [\phi_i = F_{ii}, \delta_i = D_{ii}]$$

[21] When $\mathcal{X}$ is an ellipsoid, semidefinite relaxation bound on the maximum of a quadratic form over $x \in \mathcal{X}$ is exact, so that we are in the case when an optimal solution to (4.2.9) yields the best, in terms of risk on $\mathcal{X}$, linear estimate.

Exercise 4.15 † [image reconstruction – follow-up to Exercise 4.14] A grayscale image can be represented by $m \times n$ matrix $x = [x_{pq}]_{\substack{0\leq p<m,\\ 0\leq q<n}}$ with entries in the range $[-\overline{x}, \overline{x}]$, with $\overline{x} = 255/2$ [22]. Taking picture can be modeled as observing in noise the 2D convolution $x \star \kappa$ of image x with known *blurring kernel* $\kappa = [\kappa_{uv}]_{\substack{0\leq u\leq 2\mu,\\ 0\leq v\leq 2\nu}}$, so that the observation is the random matrix

$$\omega = \Big[\omega_{rs} = \underbrace{\sum_{\substack{0\leq u\leq 2\mu, 0\leq v\leq 2\nu\\ 0\leq p<m, 0\leq q<n:\\ u+p=r, v+q=s}} x_{pq}\kappa_{uv}}_{[x\star\kappa]_{rs}} + \sigma\xi_{rs}\Big]_{\substack{0\leq r<m+2\mu,\\ 0\leq s<n+2\nu}},$$

where independent of each other random variables $\xi_{rs} \sim \mathcal{N}(0,1)$ form observation noise[23]. Our goal is to build a presumably good linear estimate of x via ω. To apply the machinery developed in Section 4.2.2, we need to cover the set of signals x allowed by our a priori assumptions by an ellitope $\mathcal{X}$, to decide in which norm we want to recover x, and then solve the associated optimization problem (4.2.9). The difficulty, however, is that the dimension of this problem formally will be huge – with 256×256 images (a rather poor resolution!), matrix H we are looking for is of the size $\dim \omega \times \dim x = ((256 + 2\mu)(256 + 2\nu)) \times 256^2 \geq 4.295 \times 10^9$; it is impossible just to store such a matrix in the memory of a usual computer, not speaking about optimizing w.r.t. such a matrix. By this reason, in what follows we develop a "practically," and not just theoretically, efficiently computable estimate.

4.15.A The construction. Our key observation is that when passing from representations of x and ω "as they are" to their Discrete Fourier Transforms, the situation simplifies dramatically. Specifically, for matrices y, x of the same sizes, let $y \bullet z$ be the entrywise product of y and z: $[y \bullet z]_{pq} = y_{pq}z_{pq}$. Setting

$$\alpha = 2\mu + m,\ \beta = 2\nu + n,$$

let $F_{\alpha,\beta}$ be the 2D discrete Fourier Transform – a linear mapping from the space $\mathbf{C}^{\alpha\times\beta}$ onto itself given by

$$[F_{\alpha,\beta}y]_{rs} = \frac{1}{\sqrt{\alpha\beta}} \sum_{\substack{0\leq p<\alpha,\\ 0\leq q<\beta}} y_{pq} \exp\left\{-2\pi i r/\alpha - 2\pi i s/\beta\right\},$$

where i is the imaginary unit. It is well known that it is a unitary transformation which is easy-to-compute (it can be computed in $O(\alpha\beta\ln(\alpha\beta))$ arithmetic operations) which "nearly diagonalizes" the convolution: whenever $x \in \mathbf{R}^{m\times n}$, setting

$$x^+ = \left[\begin{array}{c|c} x & 0_{m\times 2\nu} \\ \hline 0_{2\mu\times n} & 0_{2\mu\times 2\nu} \end{array}\right] \in \mathbf{R}^{\alpha\times\beta},$$

we have

$$F_{\alpha,\beta}(x \star \kappa) = \chi \bullet [F_{\alpha,\beta}x^+]$$

with easy-to-compute χ [24]. Now, let δ be another $(2\mu+1)\times(2\nu+1)$ kernel, with the only nonzero entry, equal to 1, in the position (μ,ν) (recall that numeration of indexes starts from 0); then

$$F_{\alpha,\beta}(x \star \delta) = \theta \bullet [F_{\alpha,\beta}x^+]$$

with easy-to-compute θ. Now consider the auxiliary estimation problem as follows:

[22] The actual grayscale image is a matrix with entries, representing pixels' light intensities, in the range $[0, 255]$. It is convenient for us to represent this actual image as the shift, by $\overline{x}$, of a matrix with entries in $[-\overline{x}, \overline{x}]$.

[23] pay attention to the fact that everywhere in this Exercise indexing of elements of 2D arrays starts from 0, and not from 1!

[24] Specifically, $\chi = \sqrt{\alpha\beta}F_{\alpha,\beta}\kappa^+$, where κ^+ is the $\alpha\times\beta$ matrix with κ as $(2\mu+1)\times(2\nu+1)$ North-Western block and zeros outside this block.

Given $R > 0$ and noisy observation

$$\widehat{\omega} = \chi \bullet \widehat{x} + \sigma \underbrace{F_{\alpha,\beta}\xi}_{\eta} \qquad\qquad [\xi = [\xi_{rs}] \text{ with independent } \xi_{rs} \sim \mathcal{N}(0,1)],$$

of signal $\widehat{x} \in \mathbf{C}^{\alpha\times\beta}$ known to satisfy $\|\widehat{x}\|_2 \leq R$, we want to recover, in the Frobenius norm $\|\cdot\|_2$, the matrix $\theta \bullet \widehat{x}$.

Treating signals $\widehat{x}$ and noises η as long vectors rather than matrices and taking into account that $F_{\alpha,\beta}$ is a unitary transformation, we see that our auxiliary problem is nothing but the problem of recovery, in $\|\cdot\|_2$-norm, of the image Θz of signal z known to belong to the centered at the origin Euclidean ball $\mathcal{Z}_R$ of radius R in $\mathbf{C}^{\alpha\beta}$, from noisy observation

$$\zeta = Az + \sigma\eta,$$

where Θ and A are *diagonal* matrices with complex entries, and η is random complex-valued noise with zero mean and unit covariance matrix. Exactly the same argument as in the real case demonstrates that as far as linear estimates $\widehat{z} = H\zeta$ are concerned, we lose nothing when restricting ourselves with diagonal matrices $H = \mathrm{Diag}\{h\}$, and the best, in terms of its worst-case over $z \in \mathcal{Z}_R$ expected $\|\cdot\|_2^2$ error, estimate corresponds to h solving the optimization problem

$$R^2 \max_{\ell\leq\alpha\beta} |\Theta_{\ell\ell} - h_\ell A_{\ell\ell}|^2 + \sigma^2 \sum_{\ell\leq\alpha\beta} |h_\ell|^2.$$

Coming back to the initial setting of our auxiliary estimation problem, we conclude that the best linear recovery of $\theta \bullet \widehat{x}$ via $\widehat{\omega}$ is given by

$$\widehat{z} = h \bullet \widehat{\omega},$$

where h is an optimal solution to the optimization problem

$$\mathrm{Opt} = \min_{h\in\mathbf{C}^{\alpha\times\beta}} \left\{ R^2 \max_{r,s} |\theta_{rs} - h_{rs}\chi_{rs}|^2 + \sigma^2 \sum_{r,s} |h_{rs}|^2 \right\}, \tag{!}$$

and the $\|\cdot\|_2$-risk

$$\mathrm{Risk}_R[\widehat{z}] = \max_{\|\widehat{x}\|_2\leq R} \mathbf{E}\left\{\|\theta \bullet \widehat{x} - h \bullet [\chi \bullet \widehat{x} + \sigma\eta]\|_2\right\}$$

of this estimate does not exceed $\sqrt{\mathrm{Opt}}$.

Now goes your first task:

1.1. *Prove that the above h induces the estimate*

$$\widehat{w}(\omega) = F_{\alpha,\beta}^{-1}\left[h \bullet [F_{\alpha,\beta}\omega]\right]$$

of $x \star \delta$, $x \in \mathcal{X}_R = \{x \in \mathbf{R}^{m\times n} : \|x\|_2 \leq R\}$, via observation $\omega = x \star \kappa + \sigma\xi$, with risk

$$\mathrm{Risk}[\widehat{w}|R] = \max_{x\in\mathbf{R}^{m\times n}:\|x\|_2\leq R} \mathbf{E}\left\{\|x \star \delta - \widehat{w}(x \star \kappa + \sigma\xi)\|_2\right\}$$

not exceeding $\sqrt{\mathrm{Opt}}$. Pay attention to the fact that x itself is nothing but a block in $x \star \delta$; note also that in order for $\mathcal{X}_R$ to cover all images we are interested in, it suffices to take $R = \sqrt{mn}\overline{x}$.

1.2. *Prove that finding optimal solution to (!) is easy – the problem is in fact just one-dimensional one!*

1.3. What are the sources, if any, of the conservatism of the estimate $\widehat{w}$ we have built as compared to the linear estimate given by an optimal solution to (4.2.9) ?

1.4. Think how to incorporate in the above construction a small number L (say, 5-10) of additional a priori constraints on x of the form

$$\|x \star \kappa_\ell\|_2 \leq R_\ell,$$

where $\kappa_\ell \in \mathbf{R}^{(2\mu+1)\times(2\nu+1)}$, and a priori upper bounds u_{rs} on the magnitudes of Fourier coefficients of x^+:

$$|[F_{\alpha\beta}x^+]_{rs}| \leq u_{rs},\ 0 \leq r < \alpha, 0 \leq s < \beta.$$

4.15.B Mimicking Total Variation constraints. For an $m \times n$ image $x \in \mathbf{R}^{m\times n}$, its (anisotropic) total variation is defined as the ℓ_1 norm of the "discrete gradient field" of x:

$$\mathrm{TV}(x) = \underbrace{\sum_{p=0}^{m-1}\sum_{q=0}^{n} |x_{p+1,q} - x_{p,q}|}_{\mathrm{TV}_a(x)} + \underbrace{\sum_{p=0}^{m}\sum_{q=0}^{n-1} |x_{p,q+1} - x_{p,q}|}_{\mathrm{TV}_b(x)}.$$

A well established experimental fact is that for naturally arising images, their total variation is essentially less than what could be expected given the magnitudes of entries in x and the sizes m, n of the image. As a result, it is tempting to incorporate a priori upper bounds on total variation of the image into an image reconstruction procedure. We are about to explain how this can be done in our context. Unfortunately, while an upper bound on total variation is a convex constraint on the image, incorporating this constraint into our construction would completely destroy its "practical computability." What we can do, is to *guess* that bounds on $\mathrm{TV}_{a,b}(x)$ can be somehow mimicked by bounds on the energy of two convolutions: one with kernel $\kappa_a \in \mathbf{R}^{(2\mu+1)\times(2\nu+1)}$ with the only nonzero entries

$$[\kappa_a]_{\mu,\nu} = -1, [\kappa_a]_{\mu+1,\nu} = 1,$$

and the other one with kernel $\kappa_b \in \mathbf{R}^{(2\mu+1)\times(2\nu+1)}$ with the only nonzero entries

$$[\kappa_b]_{\mu,\nu} = -1, [\kappa_b]_{\mu,\nu+1} = 1$$

(recall that the indexes start from 0, and not from 1). Note that $x \star \kappa_a$ and $x \star \kappa_b$ are "discrete partial derivatives" of $x \star \delta$.

For a small library of grayscale $m \times n$ images x we dealt with, experiment shows that, in addition to the energy constraint $\|x\|_2 \leq R = \sqrt{mn}\overline{x}$, the images satisfy the constraints

$$\|x \star \kappa_a\|_2 \leq \gamma R,\ \|x \star \kappa_b\|_2 \leq \gamma_2 R \tag{$*$}$$

with small γ_2, specifically, $\gamma_2 = 0.25$. In addition, it turns out that the ∞-norms of the Fourier transforms of $x \star \kappa_a$ and $x \star \kappa_b$ for these images are much less than one could expect looking at the energy of the transform's argument. Specifically, for all images x from the library it holds

$$\begin{array}{l}\|F_{\alpha\beta}[x \star \kappa_a]\|_\infty \leq \gamma_\infty R,\\ \|F_{\alpha\beta}[x \star \kappa_b]\|_\infty \leq \gamma_\infty R,\end{array},\quad \|\{z_{rs}\}_{r,s}\|_\infty = \max_{r,s} |z_{rs}| \tag{$**$}$$

with $\gamma_\infty = 0.01$ [25]. Now, relations $(**)$ read

$$\max[|\omega^a_{rs}|, |\omega^b_{rs}|]|F_{\alpha\beta}x^+]_{rs}| \leq \gamma_\infty R \forall r, s$$

[25] note that from $(*)$ it follows that $(**)$ holds true with $\gamma_\infty = \gamma_2$, while with our empirical γ's, γ_∞ is 25 times smaller than γ_2.

with easy-to-compute ω^a and ω^b, and in addition $|[F_{\alpha\beta}x^+]_{rs}| \leq R$ due to $\|F_{\alpha\beta}x^+\|_2 = \|x^+\|_2 \leq R$. We arrive at the bounds

$$|[F_{\alpha\beta}x^+]_{rs}| \leq \min\left[1, 1/|\omega^a_{rs}|, 1/|\omega^b_{rs}|\right] R\,\forall r, s.$$

on the magnitudes of entries in $F_{\alpha\beta}x^+$, and can utilize item 1.4 to incorporate these bounds, along with relations $(*)$,

Now goes the exercise:

2. *Write software implementing the outlined deblurring and denoising image reconstruction routine and run numerical experiments.*

   Recommended kernel κ: set $\mu = \rfloor m/32\lfloor$, $\nu = \rfloor n/32\lfloor$, start with

$$\kappa_{uv} = \frac{1}{(2\mu+1)(2\nu+1)} + \left\{\begin{array}{ll}\Delta, & u = \mu, v = \nu \\ 0, & \textit{otherwise}\end{array}\right., 0 \leq u \leq 2\mu, 0 \leq v \leq 2\nu,$$

   and then normalize this kernel to make the sum of entries equal to 1. In this description, $\Delta \geq 0$ is control parameter responsible for well-posedness of the auxiliary estimation problem we end up with: the smaller is Δ, the smaller is $\min_{r,s}|\chi_{rs}|$ (note that when decreasing the magnitudes of χ_{rs}, we increase the optimal value in (!)).

   We recommend to compare what happens when $\Delta = 0$ with what happens when $\Delta = 0.25$, same as compare the estimates accounting and not accounting for the constraints $(*)$, $(**)$. On the top of it, you can compare your results with what is given by "ℓ_1-minimization recovery" described as follows:

> As we remember from item 4.15.A, our problem of interest can be equivalently reformulated as recovering the image Θz of a signal $z \in \mathbf{C}^{\alpha\beta}$ from noisy observation $\widehat{\omega} = Az + \sigma\eta$, where Θ and A are diagonal matrices, and η is the zero mean complex Gaussian noise with unit covariance matrix. In other words, the entries η_ℓ in η are independent of each other real two-dimensional Gaussian vectors with zero mean and the covariance matrix $\frac{1}{2}I_2$. Given a reasonable "reliability tolerance" ϵ, say, $\epsilon = 0.1$, we can easily point out the smallest "confidence radius" ρ such that for $\zeta \sim \mathcal{N}(0, \frac{1}{2}I_2)$ it holds $\mathrm{Prob}\{\|\zeta\|_2 > \rho\} \leq \frac{\epsilon}{\alpha\beta}$, implying that for every ℓ it holds
>
> $$\mathrm{Prob}_\eta\left\{|\widehat{\omega}_\ell - A_\ell z_\ell| > \sigma\rho\right\} \leq \frac{\epsilon}{\alpha\beta},$$
>
> and therefore
>
> $$\mathrm{Prob}_\eta\left\{\|\widehat{\omega} - Az\|_\infty > \sigma\rho\right\} \leq \epsilon.$$
>
> We now can easily find the smallest, in $\|\cdot\|_1$, vector $\widehat{z} = \widehat{z}(\omega)$ which is "compatible with our observation," that is, satisfies the constraint
>
> $$\|\widehat{\omega} - A\widehat{z}\|_\infty \leq \sigma\rho,$$
>
> and take $\Theta\widehat{z}$ as the estimate of the "entity of interest" Θz (cf. Regular ℓ_1 recovery from Section 1.2.3).
>
> Note that this recovery needs no a priori information on z.

Exercise 4.16 [classical periodic nonparametric deconvolution] In classical univariate nonparametric regression, one is interested to recover a function $f(t)$ of continuous argument $t \in [0, 1]$ from noisy observations $\omega_i = f(i/n) + \sigma\eta_i$, $0 \leq i \leq n$, where $\eta_i \sim \mathcal{N}(0, 1)$ are independent across i

observation noises. Usually, a priory restrictions on f are *smoothness assumptions* – existence of $\varkappa$ continuous derivatives satisfying the a priori upper bounds

$$\left(\int_0^1 |f^{(k)}(t)|^{p_k} dt\right)^{1/p_k} \leq L_k,\ 0 \leq\leq \varkappa,$$

on their L_{p_k}-norms. The risk of an estimate is defined as the supremum, over f's of given smoothness, expected L_r-norm of the recovery error; the primary emphasis of classical studies here was how the minimax optimal (i.e., the best, over estimates) risk goes to 0 as the number of observations n goes to infinity, what are near-optimal estimates, etc. Many of these studies were dealing with *periodic case* – one where f can be extended on the entire real axis as $\varkappa$ times continuously differentiable function, or, which is the same, when f is treated as a smooth function on the circumference of length 1 rather than on the unit segment $[0, 1]$. While being slightly simpler for analysis than the general case, the periodic case turned out to be highly instructive: what was established for the latter, usually extended straightforwardly to the former.

What you are about to do in this Exercise, is to apply our machinery of building linear estimates to the outlined recovery of smooth univariate periodic regressing functions.

4.16.A. Setup. What follows is aimed at handling restrictions of smooth functions on the unit (i.e., of unit length) circumference C onto an equidistant n-point grid Γ_n on the circumference. These restrictions form the usual n-dimensional coordinate space $\mathbf{R}^n$; it is convenient to index the entries in $f \in \mathbf{R}^n$ starting from 0 rather than from 1. We equip $\mathbf{R}^n$ with two linear operators:

- *Cyclic shift* (in the sequel – just *shift*) Δ:

$$\Delta \cdot [f_0; f_1; ... : f_{n-2}; f_{n-1}] = [f_{n-1}; f_0; f_1; ...; f_{n-2}],$$

  and

- *Derivative* D:

$$D = n[I - \Delta];$$

Treating $f \in \mathbf{R}^n$ as a restriction of a function F on C onto Γ_n, Df is the finite-difference version of the first order derivative of the function, and the norms

$$|f|_p = n^{-1/p}\|f\|_p,\ p \in [1, \infty]$$

are the discrete versions of the L_p-norms of F.

Next, we can associate with $\chi \in \mathbf{R}^n$ the operator $\sum_{i=0}^{n-1} \chi_i \Delta^i$; the image of $f \in \mathbf{R}^n$ under this operator is denoted $\chi \star f$ and is called (cyclic) *convolution* of χ and f.

The problem we intend to focus on is as follows:

Given are:

- *smoothness data* represented by a nonnegative integer $\varkappa$ and two collections: $\{L_\iota > 0 : 0 \leq \iota \leq \varkappa\}$, $\{p_\iota \in [2, \infty], 0 \leq \iota \leq \varkappa\}$. The smoothness data specify the set

$$\mathcal{F} = \{f \in \mathbf{R}^n : |f|_{p_\iota} \leq L_\iota, 0 \leq \iota \leq \varkappa\}$$

  of signals we are interested in (this is the discrete analogy of *periodic Sobolev ball* – the set of $\varkappa$ times continuously differentiable functions on C with derivatives of orders up to $\varkappa$ bounded, in integral p_ι-norms, by given quantities L_ι;
- two vectors $\alpha \in \mathbf{R}^n$ (*sensing kernel*) and $\beta \in \mathbf{R}^n$ (*decoding kernel*);

- positive integer σ (noise intensity) and a real $q \in [1,2]$.

These data define the estimation problem as follows: given noisy observation

$$\omega = \alpha \star f + \sigma\eta$$

of unknown signal f known to belong to $\mathcal{F}$, where $\eta \in \mathbf{R}^n$ is random observation noise, we want to recover $\beta \star f$ in norm $|\cdot|_q$.

The only assumption on the noise is that

$$\mathrm{Vary}[\eta] := \mathbf{E}\left\{\eta\eta^T\right\} \preceq I_n.$$

The risk of a candidate estimate $\widehat{f}$ is defined as

$$\mathrm{Risk}_r[\widehat{f}|\mathcal{F}] = \sup_{\substack{f\in\mathcal{F},\\ \eta:\mathrm{Cov}[\eta]\preceq I_n}} \mathbf{E}_\eta\left\{|\beta\star f - \widehat{f}(\alpha\star f+\sigma\eta)|_q\right\}.$$

Now goes the exercise:

1. *Check that the situation in question fits the framework of Section 4.5 and figure out to what, under the circumstances, boils down the optimization problem (4.5.15) responsible for the presumably good linear estimate $\widehat{f}_H(\omega) = H^T\omega$.*

2. *Prove that in the case in question the linear estimate yielded by an appropriate optimal solution to (4.5.15) is just the cyclic convolution*

$$\widehat{f}(\omega) = h \star \omega$$

and work out a computationally cheap way to identify h.

3. *Implement your findings in software and run simulations. You could, in particular, consider the denoising problem (that is, the one where $\alpha\star x \equiv \beta\star x \equiv x$) and compare numerically the computed risks of your estimates with the classical result on the limits of performance in recovering smooth univariate regression functions; according to this results, in the situation in question and under the natural assumption that L_ι are nondecreasing in ι, the minimax optimal risk, up to a factor depending solely on $\varkappa$, is $(\sigma^2/n)^{\frac{\varkappa}{2\varkappa+1}}L_\varkappa^{\frac{1}{2\varkappa+1}}$.*

4.9.4.4 Probabilities of large deviations in linear estimation under sub-Gaussian noise

Exercise 4.17 The goal of Exercise is to derive bounds for probabilities of large deviations for estimates yielded by Proposition 4.5.1.

1. *Prove the following fact:*

Lemma 4.9.2 *Let $\Theta, Q \in \mathbf{S}^m_+$, with $Q \succ 0$, and let ξ be sub-Gaussian, with parameters (μ, S), random vector, where μ and S satisfy $\mu\mu^T + S \preceq Q$. Setting $\rho = \mathrm{Tr}(\Theta Q)$, we have*

$$\mathbf{E}_\xi\left\{\exp\{\frac{1}{8\rho}\xi^T\Theta\xi\}\right\} \le \sqrt{2}\exp\{1/4\}. \tag{4.9.20}$$

As a result, for $t > 0$ it holds

$$\mathrm{Prob}\{\sqrt{\xi^T\Theta\xi} \ge t\sqrt{\rho}\} \le \sqrt{2}\exp\{1/4\}\exp\{-t^2/8\},\ t \ge 0. \tag{4.9.21}$$

Hint: *You could use the same trick as in the proof of Lemma 2.8.3.*

2. Recall that (proof of) Proposition 4.5.1 states that in the situation of Section 4.5.1 and under Assumptions $\mathbf{A}'$, $\mathbf{R}$, for every feasible solution $(H, \Lambda, \Upsilon, \Upsilon', \Theta)$ to the optimization problem[26]

$$\mathrm{Opt} = \min_{H,\Lambda,\Upsilon,\Upsilon',\Theta} \left\{ \underbrace{\phi_{\mathcal{T}}(\lambda[\Lambda]) + \phi_{\mathcal{R}}(\lambda[\Upsilon])}_{\mathcal{A}=\mathcal{A}(\Lambda,\Upsilon)} + \underbrace{\phi_{\mathcal{R}}(\lambda[\Upsilon']) + \Gamma_{\Pi}(\Theta)}_{\mathcal{B}=\mathcal{B}(\Theta,\Upsilon')} : \begin{array}{r} \Lambda = \{\Lambda_k \succeq 0, k \le K\},\ \Upsilon = \{\Upsilon_\ell \succeq 0, \ell \le L\},\ \Upsilon' = \{\Upsilon'_\ell \succeq 0, \ell \le L\}, \\ \left[\begin{array}{c|c} \sum_k \mathcal{R}_k^*[\Lambda_k] & \frac{1}{2}[B^T - A^T H]M \\ \hline \frac{1}{2}M^T[B - H^T A] & \sum_\ell \mathcal{S}_\ell^*[\Upsilon_\ell] \end{array}\right] \succeq 0, \\ \left[\begin{array}{c|c} \Theta & \frac{1}{2}HM \\ \hline \frac{1}{2}M^T H^T & \sum_\ell \mathcal{S}_\ell^*[\Upsilon'_\ell] \end{array}\right] \succeq 0 \end{array} \right\}, \tag{4.9.22}$$

one has

$$\max_{x\in\mathcal{X}} \|[B - H^T A]x\| \le \mathcal{A} \ \ \& \max_{P:\mathrm{Var}[P]\lll\Pi} \mathbf{E}_{\xi\sim P}\left\{\|H^T\xi\|\right\} \le \mathcal{B}, \tag{4.9.23}$$

implying that the linear estimate $\widehat{x}_H(\omega) = H^T\omega$ satisfies the risk bound

$$\mathrm{Risk}_{\Pi,\|\cdot\|}[\widehat{x}_H(\cdot)|\mathcal{X}] \le \mathcal{A} + \mathcal{B}. \tag{4.9.24}$$

Prove the following

Proposition 4.9.3 *Let $H, \Lambda, \Upsilon, \Upsilon', \Theta)$ be a feasible solution to (4.9.22), and let $\widehat{x}_H(\omega) = H^T\omega$. Let, further, P be sub-Gaussian, with parameters (μ, S) satisfying*

$$\mu\mu^T + S \lll \Pi$$

probability distribution on $\mathbf{R}^m$. Finally, let $x \in \mathcal{X}$. Then

(i) *One has*

$$\begin{array}{c} \mathbf{E}_{\xi\sim P}\left\{\|Bx - \widehat{x}_H(Ax+\xi)\|\right\} \le \mathcal{A}_* + \mathcal{B}_*, \\ \mathcal{A}_* = \mathcal{A}_*(\Lambda,\Upsilon) := 2\sqrt{\phi_{\mathcal{T}}(\lambda[\Lambda])\phi_{\mathcal{R}}(\lambda[\Upsilon])} \le \mathcal{A}(\Lambda,\Upsilon) := \phi_{\mathcal{T}}(\lambda[\Lambda]) + \phi_{\mathcal{R}}(\lambda[\Upsilon]) \\ \mathcal{B}_* = \mathcal{B}_*(\Theta,\Upsilon') := 2\sqrt{\Gamma_\Pi(\Theta)\phi_{\mathcal{R}}(\lambda[\Upsilon'])} \le \mathcal{B}(\Theta,\Upsilon') := \Gamma_\Pi(\Theta) + \phi_{\mathcal{R}}(\lambda[\Upsilon']) \end{array}$$

(ii) *For every $\epsilon \in (0,1)$ one has*

$$\mathrm{Prob}_{\xi\sim P}\{\xi : \|Bx - \widehat{x}_H(Ax+\xi)\| > \mathcal{A}_* + \theta_\epsilon\mathcal{B}_*\} \le \epsilon,\ \theta_\epsilon = 2\sqrt{2\ln(\sqrt{2}\mathrm{e}^{1/4}/\epsilon)}, \tag{4.9.25}$$

with $\mathcal{A}_*$, $\mathcal{B}_*$ defined in (i).

3. *Assume we are given observation $\omega = Ax + \xi$ of unknown signal x known to belong to a given spectratope $\mathcal{X} \subset \mathbf{R}^n$ and want to recover the signal, quantifying the error of a candidate recovery $\widehat{x}$ as $\max_{k\le K}\|B_k(\widehat{x} - x)\|_{(k)}$, where $B_k \in \mathbf{R}^{\nu_k\times n}$ are given matrices, and $\|\cdot\|_{(k)}$ are given norms on $\mathbf{R}^{\nu_k}$ (for example, x can represent a discretization of a continuous-time signal, and $B_k x$ can be finite-difference approximations of signal's derivatives). As about observation noise ξ, assume, same as in item 2, that it is independent of signal x and is sub-Gaussian with sub-Gaussianity parameters μ, S satisfying $\mu\mu^T + S \preceq Q$, for some given matrix*

[26] for notation, see Section 4.5.1, (4.5.8), and (4.5.12). For reader's convenience, we recall part of this notation: for a probability distribution P on $\mathbf{R}^m$, $\mathrm{Vary}[P] = \mathbf{E}_{\xi\sim P}\{\xi^T\xi\}$, Π is a convex compact subset of $\mathrm{int}\,\mathbf{S}^m_+$, $Q \lll \Pi$ means that $Q \preceq Q'$ for some $Q' \in \Pi$, and $\Gamma_\Pi(\Theta) = \max_{Q\in\Pi}\mathrm{Tr}(\Theta Q)$.

$Q \succ 0$. Finally, assume that the unit balls of the norms conjugate to the norms $\|\cdot\|_{(k)}$ are spectratopes. In this situation, Proposition 4.5.1 provides us with K efficiently computable linear estimates $\widehat{x}_k(\omega) = H_k^T\omega : \mathbf{R}^{\dim\omega} \to \mathbf{R}^{\nu_k}$ along with upper bounds Opt_k on their risks $\max_{x\in\mathcal{X}} \mathbf{E}\left\{\|B_kx - \widehat{x}_k(Ax+\xi)\|_{(k)}\right\}$. Think how, given reliability tolerance $\epsilon \in (0,1)$, assemble these linear estimates into a single estimate $\widehat{x}(\omega) : \mathbf{R}^{\dim\omega} \to \mathbf{R}^n$ such that for every $x \in \mathcal{X}$, the probability of the event

$$\|B_k(\widehat{x}(Ax+\xi) - x)\|_{(k)} \leq \theta\mathrm{Opt}_k,\ 1 \leq k \leq K, \tag{!}$$

is at least $1-\epsilon$, for some moderate (namely, logarithmic in K and $1/\epsilon$) "assembling price" θ.

Exercise 4.18 † *Prove that if ξ is uniformly distributed on the unit sphere $\{x : \|x\|_2 = 1\}$ in $\mathbf{R}^n$, then ξ is sub-Gaussian with parameters $(0, \frac{1}{n}I_n)$.*

4.9.4.5 Linear recovery under signal-dependent noise

Exercise 4.19 [signal recovery in signal-dependent noise] Consider the situation as follows: we observe a realization ω of m-dimensional random vector

$$\omega = Ax + \xi_x,$$

where

- x is unknown signal belonging to a given signal set, specifically, spectratope (which, as always in these cases, we can assume to be basic)

$$\mathcal{X} = \{x \in \mathbf{R}^n : \exists t \in \mathcal{T} : R_k^2[x] \preceq t_kI_{d_k}, k \leq K\}$$

  with our usual restrictions on $\mathcal{T}$ and $R_k[\cdot]$;

- ξ_x is observation noise with distribution which can depend on x; all we know is that

$$\mathrm{Vary}[\xi_x] := \mathbf{E}\{\xi_x\xi_x^T\} \preceq \mathcal{C}[x],$$

  where the entries of symmetric matrix $\mathcal{C}[x]$ are quadratic in x. We assume in the sequel that signals x belong to the subset

$$\mathcal{X}_\mathcal{C} = \{x \in \mathcal{X} : \mathcal{C}[x] \succeq 0\}$$

  of $\mathcal{X}$;

- Our goal is to recover Bx, with given $B \in \mathbf{R}^{\nu\times n}$, in a given norm $\|\cdot\|$ such that the unit ball $\mathcal{B}_*$ of the conjugate norm is a spectratope:

$$\mathcal{B}_* = \{u : \|u\|_* \leq 1\} = M\mathcal{V}, \mathcal{V} = \{v : \exists r \in \mathcal{R} : S_\ell^2[v] \preceq r_\ell I_{f_\ell}, \ell \leq L\}.$$

As always, we quantify the performance of a candidate estimate $\widehat{x}(\omega) : \mathbf{R}^m \to \mathbf{R}^\nu$ by the risk

$$\mathrm{Risk}_{\|\cdot\|}[\widehat{x}|\mathcal{X}_\mathcal{C}] = \sup_{x\in\mathcal{X}_\mathcal{C}} \sup_{\xi_x:\mathrm{Cov}[\xi_x]\preceq\mathcal{C}[x]} \mathbf{E}\left\{\|Bx - \widehat{x}(Ax+\xi_x)\|\right\}.$$

1. *Utilize semidefinite relaxation to build, in a computationally efficient fashion, a "presumably good" linear estimate, specifically, prove the following*

Proposition 4.9.4 *In the situation in question, for $G \in \mathbf{S}^m$ let us define $\alpha_0[G] \in \mathbf{R}$, $\alpha_1[G] \in \mathbf{R}^n$, $\alpha_2[G] \in \mathbf{S}^n$ from the identity*

$$\mathrm{Tr}(\mathcal{C}[x]G) = \alpha_0[G] + \alpha_1^T[G]x + x^T\alpha_2[G]x \;\forall (x \in \mathbf{R}^n, G \in \mathbf{S}^m),$$

so that $\alpha_\chi[G]$ are affine in G. Consider convex optimization problem

$$\begin{array}{l}
\mathrm{Opt} = \min\limits_{H,\mu,D,\Lambda,\Upsilon,\Upsilon',G} \left\{ \mu + \phi_{\mathcal{T}}(\lambda[\Lambda]) + \phi_{\mathcal{R}}(\lambda[\Upsilon]) + \phi_{\mathcal{R}}(\lambda[\Upsilon']) : \right. \\
\qquad \Lambda = \{\Lambda_k \in \mathbf{S}_+^{d_k}, k \le K\}, \Upsilon = \{\Upsilon_\ell \in \mathbf{S}_+^{f_\ell}, \ell \le L\}, \Upsilon' = \{\Upsilon'_\ell \in \mathbf{S}_+^{f_\ell}, \ell \le L\}, D \in \mathbf{S}_+^m \\
\left[\begin{array}{c|c|c} \alpha_0[G] & \frac{1}{2}\alpha_1^T[G] & \\ \hline \frac{1}{2}\alpha_1[G] & \alpha_2[G] & \frac{1}{2}[B^T - A^T H]M \\ \hline & \frac{1}{2}M^T[B - H^T A] & \end{array}\right] \preceq \left[\begin{array}{c|c|c} \mu - \alpha_0[D] & -\frac{1}{2}\alpha_1^T[D] & \\ \hline -\frac{1}{2}\alpha_1[D] & \sum_k \mathcal{R}_k^*[\Lambda_k] - \alpha_2[D] & \\ \hline & & \sum_\ell \mathcal{S}_\ell^*[\Upsilon_\ell] \end{array}\right] \left. \right\} \\
\qquad\qquad \left[\begin{array}{c|c} G & \frac{1}{2}HM \\ \hline \frac{1}{2}M^T H^T & \sum_\ell \mathcal{S}_\ell^*[\Upsilon'_\ell] \end{array}\right] \succeq 0 \\
\left[\begin{array}{c} [\mathcal{R}_k^*[\Lambda_k]]_{ij} = \mathrm{Tr}(\Lambda_k \frac{1}{2}[R^{ki}R^{kj} + R^{kj}R^{ki}]),\ \textit{where}\ R_k[x] = \sum_j x_j R^{kj} \\ {[\mathcal{S}_\ell^*[\Upsilon_\ell]]_{ij}} = \mathrm{Tr}(\Upsilon_\ell \frac{1}{2}[S^{\ell i}S^{\ell j} + S^{\ell j}S^{\ell i}]),\ \textit{where}\ S_\ell[v] = \sum_j v_j S^{\ell j} \\ \lambda[\{Z_i, i \le I\}] = [\mathrm{Tr}(Z_1); ...; \mathrm{Tr}(Z_I)],\ \phi_A(q) = \max_{s \in A} q^T s \end{array}\right]
\end{array}$$

Whenever $H, \mu, D, \Lambda, \Upsilon, \Upsilon', G$ is feasible for the problem, one has

$$\mathrm{Risk}_{\|\cdot\|}[\widehat{H}(\cdot)|\mathcal{X}_{\mathcal{C}}] \le \mu + \phi_{\mathcal{T}}(\lambda[\Upsilon]) + \phi_{\mathcal{R}}(\lambda[\Upsilon]) + \phi_{\mathcal{R}}(\lambda[\Upsilon']).$$

2. *Work out the following special case of the above situation dealing with Poisson Imaging, see Section 2.4.3.2: your observation is m-dimensional random vector with independent Poisson entries, the vector of parameters of the corresponding Poisson distributions being Py; here P is $m \times n$ entrywise nonnegative matrix, and the unknown signal y is known to belong to a given box $Y = \{y \in \mathbf{R}^n : \underline{a} \le y \le \bar{a}\}$, where $0 \le \underline{a} < \bar{a}$. You want to recover y in $\|\cdot\|_p$-norm with given $p \in [1, 2]$.*

4.9.5 Signal recovery in Discrete and Poisson observation schemes

Exercise 4.20 † The goal of what follows is to "transfer" the constructions of linear estimates to the case of multiple indirect observations of discrete random variables. Specifically, we are interested in the situation where

- Our observation is a K-element sample $\omega^K = (\omega_1, .., \omega_K)$ with independent identically distributed components ω_k taking values in m-element set; as always, we encode the points from this m-element set by the standard basic orths $e_1, ..., e_m$ in $\mathbf{R}^m$.

- The (common for all k) probability distribution of ω_k is Ax, where x is unknown "signal" – n-dimensional probabilistic vector known to belong to a closed convex subset $\mathcal{X}$ of n-dimensional probabilistic simplex $\mathbf{\Delta}_n = \{x \in \mathbf{R}^n : x \ge 0, \sum_i x_i = 1\}$, and A is a given $m \times n$ column-stochastic matrix (i.e., entrywise nonnegative matrix with unit column sums).

- Our goal is to recover Bx, where B is a given $\nu \times n$ matrix, and we quantify a candidate estimate $\widehat{x}(\omega^K) : \mathbf{R}^{mK} \to \mathbf{R}^\nu$ by its *risk*

  $$\mathrm{Risk}_{\|\cdot\|}[\widehat{x}|\mathcal{X}] = \sup_{x \in \mathcal{X}} \mathbf{E}_{\omega^K \sim [Ax] \times ... \times [Ax]} \left\{ \|Bx - \widehat{x}(\omega^K)\| \right\},$$

  where $\|\cdot\|$ is a given norm on $\mathbf{R}^\nu$.

What we intend to use are *linear* estimates – estimates of the form

$$\widehat{x}_H(\omega^K) = H^T \underbrace{\left[\frac{1}{K}\sum_{k=1}^{K}\omega_k\right]}_{\widehat{\omega}_K[\omega^K]}, \tag{4.9.26}$$

where $H \in \mathbf{R}^{m\times\nu}$.

1. *In the main body of Lecture 4, $\mathcal{X}$ always was assumed to be symmetric w.r.t. the origin, which easily implies that we gain nothing when passing from linear estimates to affine ones (sums of linear estimates and constants). Now we are in the case when $\mathcal{X}$ can be "heavily asymmetric," which, in general, can make "genuinely affine" estimates more preferable than linear ones. Show that in the case in question, we still lose nothing when restricting ourselves with linear, rather than affine, estimates.*

4.20.A Observation scheme revisited. When observation ω^K stems from a signal $x \in \mathbf{\Delta}_n$, we have

$$\widehat{\omega}_K[\omega^K] = Ax + \xi_x,$$

where

$$\xi_x = \frac{1}{K}\sum_{k=1}^{K}[\omega_k - Ax]$$

is the average of K independent identically distributed zero mean random vectors with common covariance matrix $Q[x]$.

2. *Check that*

$$Q[x] = \mathrm{Diag}\{Ax\} - [Ax][Ax]^T,$$

and derive from this fact that the covariance matrix of ξ_x is

$$Q_K[x] = \frac{1}{K}Q[x].$$

Setting

$$\Pi = \Pi_{\mathcal{X}} = \{Q = \frac{1}{K}\mathrm{Diag}\{Ax\} : x \in \mathcal{X}\},$$

check that $\Pi_{\mathcal{X}}$ is a convex compact subset of the positive semidefinite cone $\mathbf{S}^m_+$, and that whenever $x \in \mathcal{X}$, one has $Q[x] \preceq Q$ for some $Q \in \Pi$.

4.20.B Upper-bounding risk of a linear estimate. We can upper-bound the risk of a linear estimate $\widehat{x}_H$ as follows:

$$\begin{array}{rcl}
\mathrm{Risk}_{\|\cdot\|}[\widehat{x}_H|\mathcal{X}] & = & \sup_{x\in\mathcal{X}} \mathbf{E}_{\omega^K\sim[Ax]\times\ldots\times[Ax]}\left\{\|Bx - H^T\widehat{\omega}_K[\omega^K]\|\right\} \\
& = & \sup_{x\in\mathcal{X}} \mathbf{E}_{\xi_x}\left\{\|[Bx - H^T A]x - H^T\xi_x\|\right\} \\
& \leq & \underbrace{\sup_{x\in\mathcal{X}} \|[B - H^T A]x\|}_{\Phi(H)} + \underbrace{\sup_{\xi:\mathrm{Cov}[\xi]\in\Pi_{\mathcal{X}}} \mathbf{E}_\xi\left\{\|H^T\xi\|\right\}}_{\Psi^{\mathcal{X}}(H)}.
\end{array}$$

As in the main body of Lecture 4, we intend to build a "presumably good" linear estimate by minimizing over H the sum of efficiently computable upper bounds $\overline{\Phi}(H)$ on $\Phi(H)$ and $\overline{\Psi}^{\mathcal{X}}(H)$ on $\Psi^{\mathcal{X}}(H)$.

Assuming from now on that the unit ball $\mathcal{B}_*$ of the norm conjugate to $\|\cdot\|$ is a spectratope:

$$\mathcal{B}_* := \{u : \|u\|_* \leq 1\} = \{u : \exists r \in \mathcal{R}, y : u = My, S_\ell^2[y] \preceq r_\ell I_{f_\ell}, \ell \leq L\}$$

with our usual restrictions of $\mathcal{R}$ and S_ℓ, we can take, as $\overline{\Psi}^{\mathcal{X}}(\cdot)$, the function (4.5.13). What we intend to focus on, is efficient upper-bounding of $\Phi(\cdot)$.

To simplify our task, we from now on focus on the case when $\mathcal{X}$ is cut off $\mathbf{\Delta}_n$ by a bunch of linear inequalities:

$$\mathcal{X} = \{x \in \mathbf{\Delta}_n : Gx \leq g,\, Ex = e\} \qquad [G \in \mathbf{R}^{p\times n}, E \in \mathbf{R}^{q\times n}]$$

Observe that replacing G with $G - \mathbf{1}_p^T g$ and E with $E - \mathbf{1}_q^T e$, we can reduce the situation to the one when all linear constraints in question are homogeneous, that is,

$$\mathcal{X} = \{x \in \mathbf{\Delta}_n : Gx \leq 0,\, Ex = 0\}.$$

which is what we assume from now on. Setting

$$F = [G; E; -E] \in \mathbf{R}^{(p+2q)\times n},$$

we have also

$$\mathcal{X} = \{x \in \mathbf{\Delta}_n : Fx \leq 0\}.$$

We assume also that $\mathcal{X}$ is nonempty. Finally, for the sake of some of the constructions to follow, in addition to what was already assumed about the norm $\|\cdot\|$, let us assume that this norm is *absolute*, that is, $\|u\|$ depends only on the vector of *magnitudes* of entries in u. From this assumption it immediately follows that if $0 \leq u \leq u'$, then $\|u\| \leq \|u'\|$ (why?).

4.20.C Bounding Φ, simple case. Defining the *simple case* as the one where there are no linear constraints (formally, G and E are zero matrices), observe that in this case bounding Φ is trivial:

3. *Prove that in the simple case Φ is convex and efficiently computable "as is:"*

$$\Phi(H) = \max_{i\leq n} \|(B - H^T A)g_i\|,$$

where $g_1, ..., g_n$ are the standard basic orths in $\mathbf{R}^n$.

4.20.D Lagrange upper bound on Φ.

4. *Observing that when $\mu \in \mathbf{R}_+^{p+2q}$, the function*

$$\|(B - H^T A)x\| - \mu^T Fx$$

of x is convex in $x \in \mathbf{\Delta}_n$ and overestimates $\|(B - H^T A)x\|$ everywhere on $\mathcal{X}$, conclude that the efficiently computable convex function

$$\Phi_{\mathrm{L}}(H) = \min_\mu \max_{i\leq n} \{\|(B - H^T A)g_i\| - \mu^T F g_i : \mu \geq 0\}$$

upper-bounds $\Phi(H)$. In the sequel, we call this function the Lagrange upper bound on Φ.

4.20.E Basic upper bound on Φ. For vectors u, v of the same dimension, say, k, let $\mathrm{Max}[u, v]$ stand for the entrywise maximum of u, v:

$$[\mathrm{Max}[u, v]]_i = \max[u_i, v_i],$$

and let

$$[u]_+ = \mathrm{Max}[u, 0_k],$$

where 0_k is the k-dimensional zero vector.

5.1. Let $\Lambda_+ \geq 0$ and $\Lambda_- \geq 0$ be $\nu \times (p + 2q)$ matrices, $\Lambda \geq 0$ meaning that matrix Λ is entrywise nonnegative. Prove that whenever $x \in \mathcal{X}$, one has

$$\begin{array}{l}\|(B - H^T A)x\| \leq \mathcal{B}(x, H, \Lambda_+, \Lambda_-)\\ := \min\limits_t \left\{\|t\| : t \geq \mathrm{Max}\left[[(B - H^T A)x - \Lambda_+ Fx]_+, [-(B - H^T A)x - \Lambda_- Fx]_+\right]\right\}\end{array}$$

and that $\mathcal{B}(x, H, \Lambda_+, \Lambda_-)$ is convex in x.

5.2. Derive from 5.1 that whenever $\Lambda_\pm$ are as in 5.1, one has

$$\Phi(H) \leq \mathcal{B}^+(H, \Lambda_+, \Lambda_-) := \max_{i \leq n} \mathcal{B}(g_i, H, \Lambda_+, \Lambda_-),$$

where, as in item 3, $g_1, ..., g_n$ are the standard basic orths in $\mathbf{R}^n$. Conclude that

$$\Phi(H) \leq \Phi_{\mathrm{B}}(H) = \inf_{\Lambda_\pm} \left\{\mathcal{B}^+(H, \Lambda_+, \Lambda_-) : \Lambda_\pm \in \mathbf{R}_+^{\nu \times (p+2q)}\right\}$$

and that Φ_{B} is convex and real-valued. In the sequel we refer to $\Phi_{\mathrm{B}}(\cdot)$ as to the Basic upper bound on $\Phi(\cdot)$.

4.20.F Sherali-Adams upper bound on Φ. The approach we intend to consider now is the one which we used in Lecture 1, Section 1.3.2, when explaining the origin of the verifiable sufficient condition for s-goodness, see p. 26. Specifically, setting

$$W = \left[\begin{array}{c|c} G & I \\ \hline E & \end{array}\right],$$

let us introduce slack variable $z \in \mathbf{R}^p$ and rewrite the description of $\mathcal{X}$ as

$$\mathcal{X} = \{x \in \mathbf{\Delta}_n : \exists z \geq 0 : W[x; z] = 0\},$$

so that $\mathcal{X}$ is the projection of the polyhedral set

$$\mathcal{X}^+ = \{[x; z] : x \in \mathbf{\Delta}_n, z \geq 0, W[x; z] = 0\}$$

on the x-space. Projection of $\mathcal{X}^+$ on the z-space is a nonempty (since $\mathcal{X}$ is so) and clearly bounded subset of the nonnegative orthant $\mathbf{R}^p_+$, and we can in many ways cover Z by the simplex

$$\Delta[\alpha] = \{z \in \mathbf{R}^p : z \geq 0, \sum_i \alpha_i z_i \leq 1\},$$

where all α_i are positive.

6.1. Let $\alpha > 0$ be such that $Z \subset \Delta[\alpha]$. Prove that

$$\mathcal{X}^+ = \{[x; z] : W[x; z] = 0, [x; z] \in \mathrm{Conv}\{v_{ij} = [g_i; h_j], 1 \leq i \leq n, 0 \leq j \leq p\}\}, \qquad (!)$$

where g_i are the standard basic orhts in $\mathbf{R}^n$,$h_0 = 0 \in \mathbf{R}^p$, and $\alpha_j h_j$, $1 \leq j \leq p$, are the standard basic orths in $\mathbf{R}^p$.

6.2. *Derive from 5.1 that the efficiently computable convex function*

$$\Phi_{\rm SA}(H)=\inf_C \max_{i,j}\left\{\|(B-H^TA)g_i+C^TWv_{ij}\| : C\in\mathbf{R}^{(p+q)\times\nu}\right\}$$

is an upper bound on $\Phi(H)$. In the sequel, we refer to this bound as to the Sherali-Adams one.

4.20.G Combined bound. We can combine the above bounds, specifically, as follows:

7. *Prove that the efficiently computable convex function*

$$\begin{array}{l}
\Phi_{\rm LBS}(H)=\inf\limits_{(\Lambda_\pm,C_\pm,\mu,\mu_+)\in\mathcal{R}} \max_{i,j}\mathcal{G}_{ij}(H,\Lambda_\pm,C_\pm,\mu,\mu_+),\\
\hbox{where}\\
\mathcal{G}_{ij}(H,\Lambda_\pm,C_\pm,\mu,\mu_+):=-\mu^TFg_i+\mu_+^TWv_{ij}+\min\limits_t\Big\{\|t\| : \qquad\qquad (\#)\\
\qquad t\geq \mathrm{Max}\left[[(B-H^TA-\Lambda_+F)g_i+C_+^TWv_{ij}]_+,[(-B+H^TA-\Lambda_-F)g_i+C_-^TWv_{ij}]_+\right]\Big\},\\
\mathcal{R}=\{(\Lambda_\pm,C_\pm,\mu,\mu_+):\Lambda_\pm\in\mathbf{R}_+^{\nu\times(p+2q)},C_\pm\in\mathbf{R}^{(p+q)\times\nu},\mu\in\mathbf{R}_+^{p+2q},\mu_+\in\mathbf{R}^{p+q}\}
\end{array}$$

is an upper bound on $\Phi(H)$, and that this Combined bound is at least as good as the Lagrange, the Basic, and the Sherali-Adams ones.

4.20.H How to select α? A shortcoming of the Sherali-Adams and the combined upper bounds on Φ is the presence of a "degree of freedom" – the positive vector α. Intuitively, we would like to select α to make the simplex $\Delta[\alpha]\supset Z$ to be "as small as possible." It is unclear, however, what "as small as possible" in our context is, not speaking about how to select the required α after we agree how we measure the "size" of $\Delta[\alpha]$. It turns out, however, that we can select efficiently α resulting in the *smallest volume* $\Delta[\alpha]$.

8. *Prove that minimizing the volume of $\Delta[\alpha]\supset Z$ in α reduces to solving the following convex optimization problem:*

$$\inf_{\alpha,u,v}\left\{-\sum_{s=1}^p \ln(\alpha_s) : 0\leq\alpha\leq -v, E^Tu+G^Tv\leq \mathbf{1}_n\right\} \qquad (*)$$

9. *Run numerical experiments to get an impression of the quality of the above bounds. It makes sense to generate problems where we know in advance the actual value of Φ, specifically, to take*

$$\mathcal{X}=\{x\in\mathbf{\Delta}_n : x\geq a\} \qquad (a)$$

with $a\geq0$ such that $\sum_i a_i\leq1$. In this case, we can easily list the extreme point of $\mathcal{X}$ (how?) and thus can easily compute $\Phi(H)$.

In your experiments, you can use the matrices stemming from "presumably good" linear estimates yielded by the optimization problems

$$\mathrm{Opt}=\min_{H,\Upsilon,\Theta}\left\{\overline{\Phi}(H)+\phi_{\mathcal{R}}(\lambda[\Upsilon])+\Gamma_{\mathcal{X}}(\Theta) : \begin{array}{c}\Upsilon=\{\Upsilon_\ell\succeq0,\ell\leq L\}\\ \left[\begin{array}{c|c}\Theta & \frac{1}{2}HM\\ \hline \frac{1}{2}M^TH^T & \sum_\ell \mathcal{S}_\ell^*[\Upsilon_\ell]\end{array}\right]\succeq0\end{array}\right\}, \qquad (P)$$
$$\Gamma_{\mathcal{X}}(\Theta)=\frac{1}{K}\max_{x\in\mathcal{X}}\mathrm{Tr}(\mathrm{Diag}\{Ax\}\Theta),$$

see Corollary 4.5.2, with the actual Φ (which with our $\mathcal{X}$ is available), or the upper bounds on Φ (Lagrange, Basic, Sherali-Adams, and Combined) in the role of $\overline{\Phi}$. Note that it makes *sense to test 7 bounds rather than 4 of them. Specifically, with additional constraints on the optimization variables in (#), we can get, aside of "pure" Lagrange, Basic, and Sherali-Adams bounds and their "three-component combination" (Combined bound), pairwise combinations of the pure bounds as well. For example, to combine Lagrange and Sherali-Adams bound, it suffices to add to (#) the constraints* $\Lambda_\pm = 0$.

Exercise 4.21 [follow-up to Exercise 4.20] In Exercise 4.20, we have built a "presumably good" linear estimate $\widehat{x}_{H_*}(\cdot)$, see (4.9.26), yielded by the H-component H_* of an optimal solution to problem (P), see p. 340; the optimal value Opt in this problem is an upper bound on the risk $\mathrm{Risk}_{\|\cdot\|}[\widehat{x}_{H_*}|\mathcal{X}]$ (here and in what follows we use the same notation and impose the same assumptions as in Exercise 4.20). Now, $\mathrm{Risk}_{\|\cdot\|}$ is the worst, w.r.t. signals $x \in \mathcal{X}$ underlying our observations, expected norm of the recovery error. It makes sense also to provide upper bounds on the probabilities of deviations of error's magnitude from its expected value, and this is the problem we intend to focus on, cf. Exercise 4.17.

Now goes the exercise:

1) *Prove the following*

Lemma 4.9.3 *Let $Q \in \mathbf{S}^m_+$, let K be a positive integer, and let $p \in \mathbf{\Delta}_m$. Let, further, $\omega^K = (\omega_1, ..., \omega_K)$ be i.i.d. random vectors, with ω_k taking the value e_j ($e_1, ..., e_m$ are the standard basic orths in $\mathbf{R}^m$) with probability p_j. Finally, let $\xi_k = \omega_k - \mathbf{E}\{\omega_k\} = \omega_k - p$, and $\widehat{\xi} = \frac{1}{K}\sum_{k=1}^K \xi_k$. Then for every $\epsilon \in (0,1)$ it holds*

$$\mathrm{Prob}\left\{\|\widehat{\xi}\|_2^2 \leq \frac{12\ln(2m/\epsilon)}{K}\right\} \leq \epsilon.$$

Hint: use the classical

Bernstein inequality: *Let $X_1, ..., X_K$ be independent zero mean random variables taking values in $[-M, M]$, and let $\sigma_k^2 = \mathbf{E}\{X_k^2\}$. Then for every $t \geq 0$ one has*

$$\mathrm{Prob}\left\{\sum\nolimits_{k=1}^K X_k \geq t\right\} \leq \exp\{-\frac{t^2}{2[\sum_k \sigma_k^2 + \frac{1}{3}Mt]}\}.$$

2) *Consider the situation described in Exercise 4.20 with $\mathcal{X} = \mathbf{\Delta}_n$, specifically,*

- *Our observation is a sample $\omega^K = (\omega_1, ..., \omega_K)$ with i.i.d. components $\omega_k \sim Ax$, where $X \in \mathbf{\Delta}_n$ is unknown n-dimensional probabilistic vector, A is $m \times n$ stochastic matrix (nonnegative matrix with unit column sums), and $\omega \sim Ax$ means that ω is random vector taking value e_i (e_i are standard basic orths in $\mathbf{R}^m$) with probability $[Ax]_i$, $1 \leq i \leq m$;*
- *Our goal is to recover Bx in a given norm $\|\cdot\|$; here B is a given $\nu \times n$ matrix.*
- *We assume that the unit ball $\mathcal{B}_*$ of the norm $\|\cdot\|_*$ conjugate to $\|\cdot\|$ is a spectratope:*

$$\mathcal{B}_* = \{u = My, y \in \mathcal{Y}\},\ \mathcal{Y} = \{y \in \mathbf{R}^N : \exists r \in \mathcal{R} : S_\ell^2[y] \preceq r_\ell I_{f_\ell},\ \ell \leq L\}.$$

Our goal is to build a presumably good linear estimate

$$\widehat{x}_H(\omega^K) = H^T\widehat{\omega}[\omega^K],\ \widehat{\omega}[\omega^K] = \frac{1}{K}\sum_k \omega_k.$$

Prove the following

Proposition 4.9.5 *Let H,Θ,Υ be a feasible solution to the convex optimization problem*

$$\begin{array}{c}\min_{H,\Theta,\Upsilon}\left\{\Phi(H)+\phi_{\mathcal{R}}(\lambda[\Upsilon])+\Gamma(\Theta)/K:\begin{array}{c}\Upsilon=\{\Upsilon_\ell\succeq 0,\ell\leq L\}\\ \left[\begin{array}{c|c}\Theta & \frac{1}{2}HM\\ \hline \frac{1}{2}M^TH^T & \sum_\ell\mathcal{S}_\ell^*[\Upsilon_\ell]\end{array}\right]\succeq 0\end{array}\right\},\\ \Phi(H)=\max_{j\leq n}\|\mathrm{Col}_j[B-H^TA]\|,\;\Gamma(\Theta)=\max_{x\in\mathbf{\Delta}_n}\mathrm{Tr}(\mathrm{Diag}\{Ax\}\Theta).\end{array}\tag{4.9.27}$$

Then

(i) *For every $x\in\mathbf{\Delta}_n$ it holds*

$$\begin{array}{rcl}\mathbf{E}_{\omega^K\sim Ax\times\ldots\times Ax}\left\{\|Bx-\widehat{x}_H(\omega^K)\|\right\} & \leq & \Phi(H)+2K^{-1/2}\sqrt{\phi_{\mathcal{R}}(\lambda[\Upsilon])\Gamma(\Theta)}\\ & \left[\leq\right. & \left.\Phi(H)+\phi_{\mathcal{R}}(\lambda[\Upsilon])+\Phi(H)+\Gamma(\Theta)/K\right]\end{array}\tag{4.9.28}$$

(ii) *Let $\epsilon\in(0,1)$. For every $x\in\mathbf{\Delta}_n$ one has*

$$\begin{array}{c}\mathrm{Prob}_{\omega^K\sim Ax\times\ldots\times Ax}\left\{\|Bx-\widehat{x}_H(\omega^K)\|>\Phi(H)+2\gamma K^{-1/2}\sqrt{\phi_{\mathcal{R}}(\lambda[\Upsilon])\|\Theta\|_{\mathrm{Sh},\infty}}\right\}\geq 1-\epsilon,\\ \gamma=2\sqrt{3\ln(2m/\epsilon)}.\end{array}\tag{4.9.29}$$

3) *Look what happens when $\nu=m=n$, A and B are the unit matrices, and $H=I$, i.e., we want to understand how good is recovery of a discrete probability distribution by empirical distribute derive from K-element i.i.d. sample drawn from this distribution. Take, as $\|\cdot\|$, the norm $\|\cdot\|_p$ with $p\in[1,2]$, and show that for every $x\in\mathbf{\Delta}_n$ and every $\epsilon\in(0,1)$ one has*

$$\begin{array}{ll}\forall(x\in\mathbf{\Delta}_n): & \\ \mathbf{E}\left\{\|x-\widehat{x}_I(\omega^K)\|_p\right\}\leq n^{\frac{1}{p}-\frac{1}{2}}K^{-\frac{1}{2}} & (a)\\ \mathrm{Prob}\left\{\|x-\widehat{x}_I(\omega^K)\|_p>2\sqrt{3\ln(2n/\epsilon)}n^{\frac{1}{p}-\frac{1}{2}}K^{-\frac{1}{2}}\right\}\geq 1-\epsilon & (b)\end{array}.\tag{4.9.30}$$

Exercise 4.22 [follow-up to Exercise 4.20] Consider the situation as follows. A retailer sells n items by offering customers via internet bundles of $m<n$ items, so that an offer is an m-element subset B of the set $S=\{1,...,n\}$ of the items. A customer has private preferences represented by a subset P of S – customer's *preference set*. We assume that if an offer B intersects with the preference set P of a customer, the latter buys an item drawn at random from the uniform distribution on $B\cap P$, and if $B\cap P=\emptyset$, the customer declines the offer. In the pilot stage we are interested in, the seller learns the market by selecting, one by one, K customers and making offers to them. Specifically, the seller draws k-th customer, $k\leq K$, at random from the uniform distribution on the population of customers, and makes the selected customer an offer drawn at random from the uniform distribution on the set $\mathcal{S}_{m,n}$ of all m-item offers. What is observed in k-th experiment, is the item, if any, bought by customer, and what we want is to make statistical inferences from these observations.

The outlined observation scheme can be formalized as follows. Let $\mathcal{S}$ be the set of all subsets of the n-element set, so that $\mathcal{S}$ is of cardinality $N=2^n$. The population of customers induces a probability distribution p on $\mathcal{S}$: for $P\in\mathcal{S}$, p_P is the fraction of customers with the preference set being P; we refer to p as to the *preference distribution*. An outcome of a single experiment can be represented by a pair (ι,B), where $B\in\mathcal{S}_{m,n}$ is the offer used in the experiment, and ι is either 0 ("nothing is bought", $P\cap B=\emptyset$), or a point from $P\cap B$, the item which was bought, when $P\cap B\neq\emptyset$. Note that A_P is a probability distribution on the $(M=(m+1)\binom{n}{m})$-element set $\Omega=\{(\iota,B)\}$ of possible outcomes. As a result, our observation scheme is fully specified by known to us $M\times N$ column-stochastic matrix A with the columns A_P indexed by $P\in\mathcal{S}$. When a customer is drawn at random from the uniform distribution on the population of customers, the distribution of the outcome clearly is Ap, where p is the (unknown) preference distribution. Our inferences should be based on K-element sample $\omega^K = (\omega_1, ..., \omega_K)$, with $\omega_1, .., \omega_K$ drawn, independently of each other, from the distribution Ap.

Now we can pose various inference problems, e.g., the one of recovering p. We, however, intend to focus on a simpler problem – one of recovering Ap. In terms of our story, this makes sense: when we know Ap, we know, e.g., what is the probability for every offer to be "successful" (something indeed is bought) and/or to result in a specific profit, etc. With this knowledge at hand, the seller can pass from "blind" offering policy (drawing an offer at random from the uniform distribution on the set $\mathcal{S}_{m,n}$) to something more rewarding.

Now goes the exercise:

1. *Use the results of Exercise 4.20 to build "presumably good" linear estimate*

$$\widehat{x}_H(\omega^K) = H^T \left[\frac{1}{K}\sum_{k=1}^K \omega_k\right]$$

*of Ap (*as always, we encode observations ω, which are elements of M-element set Ω, by standard basic orths in $\mathbf{R}^M$*). As the norm $\|\cdot\|$ quantifying the recovery error, use $\|\cdot\|_1$ and/or $\|\cdot\|_2$. In order to avoid computational difficulties, use small m and n (e.g., $m = 3$ and $n = 5$). Compare your results with those for the straightforward estimate $\frac{1}{K}\sum_{k=1}^K \omega_k$ (the empirical distribution of $\omega \sim Ap$).*

2. *Assuming that the "presumably good" linear estimate outperforms the straightforward one, how could this phenomenon be explained? Note that we have no nontrivial a priori information on p!*

Exercise 4.23 [Poisson Imaging] *Poisson Imaging Problem* is to recover an unknown signal observed via Poisson observation scheme. More specifically, assume that our observation is a realization of random vector $\omega \in \mathbf{R}^m_+$ with independent of each other Poisson entries $\omega_i = \text{Poisson}([Ax]_i)$. Here A is a given entrywise nonnegative $m \times n$ matrix, and x is unknown signal known to belong to a given compact convex subset $\mathcal{X}$ of $\mathbf{R}^n_+$. Our goal is to recover in a given norm $\|\cdot\|$ the linear image Bx of x, where B is a given $\nu \times n$ matrix.

We assume in the sequel that $\mathcal{X}$ is a subset cut off the n-dimensional probabilistic simplex $\mathbf{\Delta}_n$ by a bunch of linear equality and inequality constraints. The assumption $\mathcal{X} \subset \mathbf{\Delta}_n$ is not too restrictive. Indeed, assume that we know in advance a linear inequality $\sum_i \alpha_i x_i \leq 1$ with positive coefficients which is valid on $\mathcal{X}$ [27]. Introducing slack variable s given by $\sum_i \alpha_i x_i + s = 1$, we can pass from signal x to the new signal $[\alpha_1 x_1; ...; \alpha_n x_n; s]$, which, after straightforward modification of matrices A and B, brings the situation to the one where $\mathcal{X}$ is a subset of the probabilistic simplex.

Our goal in the sequel is to build a presumably good linear estimate $\widehat{x}_H(\omega) = H^T\omega$ of Bx. Acting in the same fashion as in Exercise 4.20, we start with upper-bounding the risk of a linear estimate. Specifically, representing

$$\omega = Ax + \xi_x,$$

we arrive at zero mean observation noise ξ_x with independent of each other entries $[\xi_x]_i = \omega_i - [Ax]_i$ and covariance matrix $\text{Diag}\{Ax\}$. We now can upper-bound the risk of a linear estimate $\widehat{x}_H(\cdot)$ in the same fashion as in Exercise 4.20. Specifically, denoting by $\Pi_{\mathcal{X}}$ the set of all diagonal matrices

[27] For example, in PET, see Section 2.4.3.2, where x is the density of radioactive tracer injected to the patient taking the PET procedure, we know in advance the total amount $\sum_i v_i x_i$ of the tracer, v_i being the volumes of voxels.

$\mathrm{Diag}\{Ax\}$, $x \in \mathcal{X}$ and by $P_{i,x}$ the Poisson distribution with parameter $[Ax]_i$, we have

$$\begin{array}{rcl}\mathrm{Risk}_{\|\cdot\|}[\widehat{x}_H|\mathcal{X}] &=& \sup_{x\in\mathcal{X}} \mathbf{E}_{\omega\sim P_{1,x}\times\ldots\times P_{m,x}}\left\{\|Bx - H^T\widehat{\omega}_K[\omega^K]\|\right\}\\ &=& \sup_{x\in\mathcal{X}} \mathbf{E}_{\xi_x}\left\{\|[Bx - H^TA]x - H^T\xi_x\|\right\}\\ &\leq& \underbrace{\sup_{x\in\mathcal{X}} \|[B - H^TA]x\|}_{\Phi(H)} + \underbrace{\sup_{\xi:\mathrm{Cov}[\xi]\in\Pi_{\mathcal{X}}} \mathbf{E}_{\xi}\left\{\|H^T\xi\|\right\}}_{\Psi^{\mathcal{X}}(H)}.\end{array}$$

In order to build a presumably good linear estimate, it suffices to build efficiently computable convex in H upper bounds $\overline{\Phi}(H)$ on $\Phi(H)$ and $\overline{\Psi}^{\mathcal{X}}(H)$ on $\Psi^{\mathcal{X}}(H)$. and then take as H an optimal solution to the convex optimization problem

$$\mathrm{Opt} = \min_H \left[\overline{\Phi}(H) + \overline{\Psi}^{\mathcal{X}}(H)\right].$$

Same as in Exercise 4.20, assume from now on that $\|\cdot\|$ is an absolute norm, and the unit ball $\mathcal{B}_*$ of the conjugate norm is a spectratope:

$$\mathcal{B}_* := \{u : \|u\|_* \leq 1\} = \{u : \exists r \in \mathcal{R}, y : u = My, S_\ell^2[y] \preceq r_\ell I_{f_\ell}, \ell \leq L\}$$

Observe that

- In order to build $\overline{\Phi}$, we can use exactly the same techniques as those developed in Exercise 4.20. Indeed, as far as building $\overline{\Phi}$ is concerned, the only difference between our present situation and the one of Exercise4.20 is that in the latter, A was column-stochastic matrix, while now A is just entrywise nonnegative matrix. Note, however, that when upper-bounding Φ in Exercise 4.20, we never used the fact that A is column-stochastic.
- In order to upper-bound $\Psi^{\mathcal{X}}$, we can use the same bound (4.5.13) as in Exercise 4.20.

The bottom line is that in order to build a presumably good linear estimate, we need to solve the convex optimization problem

$$\mathrm{Opt} = \min_{H,\Upsilon,\Theta}\left\{\overline{\Phi}(H) + \phi_{\mathcal{R}}(\lambda[\Upsilon]) + \Gamma_{\mathcal{X}}(\Theta) : \begin{array}{c}\Upsilon = \{\Upsilon_\ell \succeq 0, \ell \leq L\}\\ \left[\begin{array}{c|c}\Theta & \frac{1}{2}HM\\ \hline \frac{1}{2}M^TH^T & \sum_\ell \mathcal{S}_\ell^*[\Upsilon_\ell]\end{array}\right] \succeq 0\end{array}\right\}, \qquad (P)$$
$$\Gamma_{\mathcal{X}}(\Theta) = \max_{x\in\mathcal{X}} \mathrm{Tr}(\mathrm{Diag}\{Ax\}\Theta),$$

(cf. problem (P) on p. 340) with $\overline{\Phi}$ yielded by a whatever construction from Exercise 4.20, e.g., the least conservative Combined upper bound on Φ.

What in our present situation differs significantly from the situation of Exercise 4.20, are the bounds on probabilities of large deviations established in Exercise 4.21, and the goal of what follows is to establish these bounds for Poisson Imaging.

Here is what you are supposed to do:

1. *Let ω be m-dimensional random vector with independent entries $\omega_i \sim \mathrm{Poisson}(\mu_i)$, and let $\mu = [\mu_1; ...; \mu_m]$. Prove that whenever $h \in \mathbf{R}^m$, $\gamma > 0$, and $\delta \geq 0$, one has*
$$\ln\left(\mathrm{Prob}\{h^T\omega > h^T\mu + \delta\}\right) \leq \sum_i [\exp\{\gamma h_i\} - 1]\mu_i - \gamma h^T\mu - \gamma\delta. \qquad (*)$$
2. *Taking for granted that $\mathrm{e}^x \leq 1 + x + \frac{3}{4}x^2$ when $|x| \leq 2/3$, prove that in the situation of item 1 one has*
$$0 \leq \gamma \leq \frac{2}{3\|h\|_\infty} \Rightarrow \ln\left(\mathrm{Prob}\{h^T\omega > h^T\mu + \delta\}\right) \leq \frac{3}{4}\gamma^2 \sum_i h_i^2\mu_i - \gamma\delta. \qquad (\#)$$

Derive from the latter fact that

$$\mathrm{Prob}\left\{h^T\omega > h^T\mu+\delta\right\} \leq \exp\{-\frac{\delta^2}{3[\sum_i h_i^2\mu_i + \|h\|_\infty\delta]}\}. \qquad (\#\#)$$

and conclude that

$$\mathrm{Prob}\left\{|h^T\omega - h^T\mu| > \delta\right\} \leq 2\exp\{-\frac{\delta^2}{3[\sum_i h_i^2\mu_i + \|h\|_\infty\delta]}\}. \qquad (!)$$

3. *Extract from* (!) *the following*

Proposition 4.9.6 *In the situation and under the assumptions of Exercise 4.23, let* Opt *be the optimal value, and* H, Υ, Θ *be a feasible solution to problem* (P)*. Whenever* $x \in \mathcal{X}$ *and* $\epsilon \in (0,1)$*, denoting by* P_x *the distribution of observations stemming from* x *(i.e., the distribution of random vector* ω *with independent entries* $\omega_i \sim \mathrm{Poisson}([Ax]_i)$*), one has*

$$\mathbf{E}\left\{\|Bx - \widehat{x}_H(\omega)\|\right\} \leq \overline{\Phi}(H) + 2\sqrt{\phi_{\mathcal{R}}(\lambda[\Upsilon])\mathrm{Tr}(\mathrm{Diag}(Ax\}\Theta)} \leq \overline{\Phi}(H) + \phi_{\mathcal{R}}(\lambda[\Upsilon]) + \Gamma_{\mathcal{X}}(\Theta) \qquad (4.9.31)$$

and

$$\mathrm{Prob}_{\omega\sim P_x}\Big\{\|Bx - \widehat{x}_H(\omega)\| \leq \overline{\Phi}(H) + 2\sqrt{2}\sqrt{9\ln^2(2m/\epsilon)\mathrm{Tr}(\Theta) + 3\ln(2m/\epsilon)\mathrm{Tr}(\mathrm{Diag}\{Ax\}\Theta)}\sqrt{\phi_{\mathcal{R}}(\lambda[\Upsilon])}\Big\} \geq 1-\epsilon. \qquad (4.9.32)$$

Note that in the case of $[Ax]_i \geq 1$ *for all* $x \in \mathcal{X}$ *and all* i *we have* $\mathrm{Tr}(\Theta) \leq \mathrm{Tr}(\mathrm{Diag}\{Ax\}\Theta)$*, so that in this case the* P_x*-probability of the event*

$$\left\{\omega : \|Bx - \widehat{x}_H(\omega)\| \leq \overline{\Phi}(H) + O(1)\ln(2m/\epsilon)\sqrt{\phi_{\mathcal{R}}(\lambda[\Upsilon])\Gamma_{\mathcal{X}}(\Theta)}\right\}$$

is at least $1-\epsilon$*.*

4.9.6 Numerical lower-bounding minimax risk

Exercise 4.24 † [numerical lower bounding minimax risk]

4.24.A. Motivation. From the theoretical viewpoint, the results on near-optimality of presumably good linear estimates stated in Propositions 4.2.2 and 4.4.2 seem to be pretty strong and general. This being said, for a practically oriented user the "nonoptimality factors" arising in these propositions can be too large to make practical sense. This practical drawback of our theoretical results is not too crucial – what matters in applications, is whether the risk of a proposed estimate is appropriate for the application in question, and not by how much it could be improved were we smart enough to build the "ideal" estimate; results of the latter type from practical viewpoint offer no more than some "moral support." Nevertheless, the "moral support" has its value, and it makes sense to strengthen it by improving the lower risk bounds as compared to those underlying Propositions 4.2.2 and 4.4.2. In this respect, an appealing idea is to pass from lower risk bounds yielded by theoretical considerations to *computation-based* ones. The goal of this exercise is to develop some methodology yielding computation-based lower risk bounds. We start with the main ingredient of this methodology – the classical *Cramer-Rao* bound.

4.24.B. Cramer-Rao bound. Consider the situation as follows: we are given

- an observation space Ω equipped with reference measure Π, basic examples being (A) $\omega = \mathbf{R}^m$ with Lebesgue measure Π, and (B) (finite our countable) discrete set Ω with counting measure Π;
- a convex compact set $\Theta \subset \mathbf{R}^k$ and a family $\Pi = \{p(\omega, \theta) : \theta \in \Theta\}$ of probability densities, taken w.r.t. Π.

Our goal is, given an observation $\omega \sim p(\cdot, \theta)$ stemming from unknown θ known to belong to Θ, to recover θ. We quantify the risk of a candidate estimate $\widehat{\theta}$ as

$$\mathrm{Risk}[\widehat{\theta}|\Theta] = \sup_{\theta\in\Theta} \left(\mathbf{E}_{\omega\sim p(\cdot,\theta)} \left\{\|\widehat{\theta}(\omega) - \theta\|_2^2\right\}\right)^{1/2}, \tag{4.9.33}$$

and define the "ideal" minimax risk as

$$\mathrm{Risk}_{\rm opt} = \inf_{\widehat{\theta}} \mathrm{Risk}[\widehat{\theta}],$$

the infimum being taken w.r.t. all estimates, or, which is the same, all *bounded* estimates (indeed, passing from a candidate estimate $\widehat{\theta}$ to the projected estimate $\widehat{\theta}_\Theta(\omega) = \mathrm{argmin}_{\theta\in\Theta} \|\widehat{\theta}(\omega - \theta\|_2$ we can only reduce the risk of an estimate.

The classical Cramer-Rao inequality, which we intend to use, is certain relation between the covariance matrix of a bounded estimate and its bias; this relation is valid under mild regularity assumptions on the family Π, specifically, as follows:

1) $p(\omega, \theta) > 0$ for all $\omega \in \Omega, \theta \in U$, and $p(\omega, \theta)$ is differentiable in θ, the with $\nabla_\theta p(\omega, \theta)$ continuous in $\theta \in \Theta$;

2) The *Fisher Information matrix*

$$\mathcal{I}(\theta) = \int_\Omega \frac{\nabla_\theta p(\omega, \theta)[\nabla_\theta p(\omega, \theta)]^T}{p(\omega, \theta)} \Pi(d\omega)$$

is well defined for all $\theta \in \Theta$;

3) There exists function $M(\omega) \geq 0$ such that $\int_\Omega M(\omega)\Pi(d\omega) < \infty$ and

$$\|\nabla_\theta p(\omega, \theta)\|_2 \leq M(\omega) \ \forall \omega \in \Omega, \theta \in \Theta.$$

The derivation of the Cramer-Rao bound is as follows. Let $\widehat{\theta}(\omega)$ be a bounded estimate, and let

$$\phi(\theta) = [\phi_1(\theta); ...; \phi_k(\theta)] = \int_\Omega \widehat{\theta}(\omega) p(\omega, \theta)\Pi(d\omega)$$

be the expected value of the estimate. By item 3, $\phi(\theta)$ is differentiable on Θ, with the Jacobian $\phi'(\theta) = \left[\frac{\partial \phi_i(\theta)}{\partial \theta_j}\right]_{i,j\leq k}$ given by

$$\phi'(\theta)h = \int_\Omega \widehat{\theta}(\omega) h^T \nabla_\theta p(\omega, \theta)\Pi(d\omega),\ h \in \mathbf{R}^k.$$

Besides this, recalling that $\int_\Omega p(\omega, \theta)\Pi(d\omega) \equiv 1$ and invoking item 3, we have $\int_\Omega h^T \nabla_\theta p(\omega, \theta)\Pi(d\omega) = 0$, whence, in view of the previous equality,

$$\phi'(\theta)h = \int_\Omega [\widehat{\theta}(\omega) - \phi(\theta)] h^T \nabla_\theta p(\omega, \theta)\Pi(d\omega),\ h \in \mathbf{R}^k.$$

Therefore for all $g,h\in\mathbf{R}^k$ we have

$$\begin{array}{rcl}
[g^T\phi'(\theta)h]^2 &=& \left[\int_\omega [g^T(\widehat{\theta}-\phi(\theta)][h^T\nabla_\theta p(\omega,\theta)/p(\omega,\theta)]p(\omega,\theta)\Pi(d\omega)\right]^2\\
&\leq& \left[\int_\Omega g^T[\widehat{\theta}-\phi(\theta)][\widehat{\theta}-\phi(\theta)]^Tgp(\omega,\theta)\Pi(d\omega)\right]\left[\int_\Omega [h^T\nabla_\theta p(\omega,\theta)/p(\omega,\theta)]^2p(\omega,\theta)\Pi(d\omega)\right]\\
&&\text{[Cauchy's Inequality]}\\
&=& \left[g^T\mathrm{Cov}_{\widehat{\theta}}(\theta)g\right]\left[h^T\mathcal{I}(\theta)h\right],
\end{array}$$

where $\mathrm{Cov}_{\widehat{\theta}}(\theta)$ is the covariance matrix $\mathbf{E}_{\omega\sim p(\cdot,\theta)}\left\{[\widehat{\theta}(\omega)-\phi(\theta)][\widehat{\theta}(\omega)-\phi(\theta)]^T\right\}$ of $\widehat{\theta}(\omega)$ induced by $\omega\sim p(\cdot,\theta)$. We have arrived at the inequality

$$\left[g^T\mathrm{Cov}_{\widehat{\theta}}(\theta)g\right]\left[h^T\mathcal{I}(\theta)h\right]\geq[g^T\phi'(\theta)h]^2\ \forall(g,h\in\mathbf{R}^k,\theta\in\Theta). \tag{$*$}$$

For $\theta\in\Theta$ fixed, let $\mathcal{J}$ be a positive definite matrix such that $\mathcal{J}\succeq\mathcal{I}(\theta)$, whence by $(*)$ it holds

$$\left[g^T\mathrm{Cov}_{\widehat{\theta}}(\theta)g\right]\left[h^T\mathcal{J}h\right]\geq[g^T\phi'(\theta)h]^2\ \forall(g,h\in\mathbf{R}^k). \tag{$**$}$$

For g fixed, the maximum of the right hand side quantity in $(**)$ over h satisfying $h^T\mathcal{J}h\leq 1$ is $g^T\phi'(\theta)\mathcal{J}^{-1}[\phi'(\theta]^Tg$, and we arrive at the *Cramer-Rao inequality*

$$\begin{array}{c}
\forall(\theta\in\Theta,\mathcal{J}\succeq\mathcal{I}(\theta),\mathcal{J}\succ 0):\mathrm{Cov}_{\widehat{\theta}}(\theta)\succeq\phi'(\theta)\mathcal{J}^{-1}[\phi'(\theta]^T\\
\left[\mathrm{Cov}_{\widehat{\theta}}(\theta)=\mathbf{E}_{\omega\sim p(\cdot,\theta)}\left\{[\widehat{\theta}-\phi(\theta)][\widehat{\theta}-\phi(\theta)]^T\right\},\ \phi(\theta)=\mathbf{E}_{\omega\sim p(\cdot),\theta)}\left\{\widehat{\theta}(\omega)\right\}\right]
\end{array} \tag{CR}$$

which holds true for every bounded estimate $\widehat{\theta}(\cdot)$. Note also that for every $\theta\in\Theta$ and every bounded estimate x we have

$$\begin{array}{l}
\mathrm{Risk}^2[\widehat{\theta}]\geq\mathbf{E}_{\omega\sim p(\cdot,\theta)}\left\{\|\widehat{\theta}(\omega)-\theta\|_2^2\right\}=\mathbf{E}_{\omega\sim p(\cdot,\theta)}\left\{\|[\widehat{\theta}(\omega)-\phi(\theta)]+[\phi(\theta)-\theta]\|_2^2\right\}\\
=\mathbf{E}_{\omega\sim p(\cdot,\theta)}\left\{\|\widehat{\theta}(\omega)-\phi(\theta)\|_2^2\right\}-2\underbrace{\mathbf{E}_{\omega\sim p(\cdot,\theta)}\left[[\widehat{\theta}(\omega)-\phi(\theta)]^T[\phi(\theta)-\theta)]\right\}}_{=0}+\|\phi(\theta)-\theta)\|_2^2\\
=\mathrm{Tr}(\mathrm{Cov}_{\widehat{\theta}}(\theta))+\|\phi(\theta)-\theta\|_2^2,
\end{array}$$

whence, in view of (CR), for every bounded estimate $\widehat{\theta}$ it holds

$$\begin{array}{ll}
\forall(\mathcal{J}\succ 0:\mathcal{J}\succeq\mathcal{I}(\theta)\,\forall\theta\in\Theta): & \mathrm{Risk}^2[\widehat{\theta}]\geq\sup_{\theta\in\Theta}\left[\mathrm{Tr}(\phi'(\theta)\mathcal{J}^{-1}[\phi'(\theta)]^T)+\|\phi(\theta)-\theta\|_2^2\right]\\
& \left[\phi(\theta)=\mathbf{E}_{\omega\sim p(\cdot,\theta)}\{\widehat{\theta}(\omega)\}\right]
\end{array} \tag{4.9.34}$$

The fact that we were speaking about estimating "the entire" θ rather than a given vector-valued function $f(\theta):\Theta\to\mathbf{R}^\nu$ plays no special role, and in fact the Cramer-Rao inequality admits the following modification (yielded by a reasoning completely similar to the one we just have carried out):

Proposition 4.9.7 *In the situation described in the beginning of item* **4.24.B** *and under assumptions 1) – 3) of this item, let $f(\cdot):\Theta\to\mathbf{R}^\nu$ be a bounded Borel function, and let $\widehat{f}(\omega)$ be a bounded estimate of $f(\omega)$ via observation $\omega\sim p(\cdot,\theta)$. Then, setting*

$$\phi(\theta)=\mathbf{E}_{\omega\sim p(\cdot,\theta)}\left\{\widehat{f}(\theta)\right\},\ \mathrm{Cov}_{\widehat{f}}(\theta)=\mathbf{E}_{\omega\sim p(\cdot,\theta)}\left\{[\widehat{f}(\omega)-\phi(\theta)][\widehat{f}(\omega)-\phi(\theta)]^T\right\}\qquad[\theta\in\Theta]$$

one has

$$\forall(\theta\in\Theta,\mathcal{J}\succeq\mathcal{I}(\theta),\mathcal{J}\succ 0):\mathrm{Cov}_{\widehat{f}}(\theta)\succeq\phi'(\theta)\mathcal{J}^{-1}[\phi'(\theta)]^T.$$

As a result, setting

$$\mathrm{Risk}[\widehat{f}]=\sup_{\theta\in\Theta}\left[\mathbf{E}_{\omega\sim p(\cdot,\theta)}\left\{\|\widehat{f}(\omega)-f(\theta)\|_2^2\right\}\right]^{1/2},$$

it holds

$$\forall(\mathcal{J}\succ 0:\mathcal{J}\succeq\mathcal{I}(\theta)\,\forall\theta\in\Theta):\quad \mathrm{Risk}^2[\widehat{f}]\geq\sup_{\theta\in\Theta}\left[\mathrm{Tr}(\phi'(\theta)\mathcal{J}^{-1}[\phi'(\theta)]^T)+\|\phi(\theta)-f(\theta)\|_2^2\right]$$

Now goes the first part of the exercise:

1. *Derive from (4.9.34) the following*

Proposition 4.9.8 *In the situation of item 4.24.B, let*

- *$\Theta \subset \mathbf{R}^k$ be $\|\cdot\|_2$-ball of radius $r>0$,*
- *the family $\mathcal{P}$ be such that $\mathcal{I}(\theta)\preceq\mathcal{J}$ for some $\mathcal{J}\succ 0$ and all $\theta\in\Theta$.*

Then the minimax optimal risk satisfies the bound

$$\mathrm{Risk}_{\rm opt}\geq \frac{rk}{r\sqrt{\mathrm{Tr}(\mathcal{J})}+k}. \tag{4.9.35}$$

In particular, when $\mathcal{J}=\alpha^{-1}I_k$, we have

$$\mathrm{Risk}_{\rm opt}\geq \frac{r\sqrt{\alpha k}}{r+\sqrt{\alpha k}}. \tag{4.9.36}$$

Hint. Assuming w.l.o.g. that Θ is centered at the origin, and given a bounded estimate $\widehat{\theta}$ with risk $\mathfrak{R}$, let $\phi(\theta)$ be associated with the estimate via (4.9.34). Select $\gamma\in(0,1)$ and consider two cases: (a): there exists $\theta\in\partial\Theta$ such that $\|\phi(\theta)-\theta\|_2>\gamma r$, and (b): $\|\phi(\theta)-\theta\|_2\leq\gamma r$ for all $\theta\in\partial\Theta$. In the case of (a), lower-bound $\mathfrak{R}$ by $\max_{\theta\in\Theta}\|\phi(\theta)-\theta\|_2$, see (4.9.34). In the case of(b), lower-bound $\mathfrak{R}^2$ by $\max_{\theta\in\Theta}\mathrm{Tr}(\phi'(\theta)\mathcal{J}^{-1}[\phi'(\theta)]^T)$, see (4.9.34), and use Divergence theorem to lower-bound the latter quantity in terms of the flux of the vector field $\phi(\cdot)$ over $\partial\Theta$.

When implementing the above strategy, you could find useful the following fact (prove it!)

Lemma 4.9.4 *Let Φ be an $n\times n$ matrix, and $\mathcal{J}$ be a positive semidefinite $n\times n$ matrix. Then*

$$\mathrm{Tr}(\Phi\mathcal{J}^{-1}\Phi^T)\geq \mathrm{Tr}^2(\Phi)/\mathrm{Tr}(\mathcal{J}).$$

4.24.C. Application to signal recovery. Proposition 4.9.8 allows to build computation-based lower risk bounds in the signal recovery problem considered in Section 4.2, specifically, the problem where one wants to recover the linear image Bx of unknown signal x known to belong to a given ellitope

$$\mathcal{X}=\{x\in\mathbf{R}^n:\exists t\in\mathcal{T}:x^TS_\ell x\leq t_\ell,\ell\leq L\}$$

(with our usual restriction on S_ℓ and $\mathcal{T}$) via observation

$$\omega=Ax+\sigma\xi,\;\xi\sim\mathcal{N}(0,I_m),$$

and the risk of a candidate estimate, same as in Section 4.2, is defined according to (4.9.33)[28]. It is convenient to assume that the matrix B (which in our general setup can be an arbitrary $\nu\times n$ matrix) is a *nonsingular* $n\times n$ matrix[29] Under this assumption, setting

$$\mathcal{Y}=B^{-1}\mathcal{X}=\{y\in\mathbf{R}^n:\exists t\in\mathcal{T}:y^T[B^{-1}]^TS_\ell B^{-1}y\leq t_\ell,\ell\leq L\}$$

[28] In fact, the approach to be developed can be applied to signal recovery problems involving Discrete/Poisson observation schemes, different from $\|\cdot\|_2$ norms used to measure the recovery error, signal-dependent noises, etc.

[29] This assumption is nonrestrictive. Indeed, when $B\in\mathbf{R}^{\nu\times n}$ with $\nu<n$, we can add to B $n-\nu$ zero rows, which keeps our estimation problem intact. When $\nu\geq n$, we can add to B a small perturbation to ensure $\mathrm{Ker}\,B=\{0\}$, which, for small enough perturbation, again keeps our estimation problem basically intact. It remains to note that when $\mathrm{Ker}\,B=\{0\}$. we can replace $\mathbf{R}^\nu$ with the image space of B, which again does not affect the estimation problem we are interested in.

and $\bar{A} = AB^{-1}$, we lose nothing when replacing the sensing matrix A with $\bar{A}$ and treating as our signal $y \in \mathcal{Y}$ rather than $\mathcal{X}$; thus, we have reduced the situation to the one where A is replaced with $\bar{A}$, $\mathcal{X}$ with $\mathcal{Y}$, and B with the unit matrix I_n. For the sake of simplicity, we assume from now on that A (and therefore $\bar{A}$) is with trivial kernel. Finally, let $\tilde{S}_\ell \succeq S_\ell$ be close to S_k positive definite matrices, e.g., $\tilde{S}_\ell = S_\ell + 10^{-100} I_n$; setting $\bar{S}_\ell = [B^{-1}]^T \tilde{S}_\ell B^{-1}$ and

$$\bar{\mathcal{Y}} = \{y \in \mathbf{R}^n : \exists t \in \mathcal{T} : y^T \bar{S}_\ell y \leq t_\ell, \ell \leq L\},$$

observe that $\bar{S}_\ell \succ 0$ and $\bar{\mathcal{Y}} \subset \mathcal{Y}$; this, any lower bound on the $\|\cdot\|_2$-risk of recovery $y \in \bar{\mathcal{Y}}$ via observation $\omega = AB^{-1}y + \sigma\xi$, $\xi \sim \mathcal{N}(0, I_m)$, automatically is a lower bound on the minimax risk $\mathrm{Risk}_{\mathrm{opt}}$ corresponding to our original problem of interest.

Now assume that we can point out a k-dimensional linear subspace E in $\mathbf{R}^n$ and positive reals r, γ such that

(i) the centered at the origin $\|\cdot\|_2$-ball $\Theta = \{\theta \in E : \|\theta\|_2 \leq r\}$ is contained in $\bar{Y}$;

(ii) The restriction $\bar{A}_E$ of $\bar{A}$ onto E satisfies the relation

$$\mathrm{Tr}(\bar{A}_E^* \bar{A}_E) \leq \gamma$$

($\bar{A}_E^* : \mathbf{R}^m \to E$ is the conjugate of the linear map $\bar{A}_E : E \to \mathbf{R}^m$).

Consider the auxiliary estimation problem obtained from the (reformulated) problem of interest by replacing the signal set $\bar{\mathcal{Y}}$ with Θ. Since $\Theta \subset \bar{\mathcal{Y}}$, the minimax risk in the auxiliary problem is a lower bound on the minimax risk $\mathrm{Risk}_{\mathrm{opt}}$ we are interested in. On the other hand, the auxiliary problem is nothing but the problem of recovering parameter $\theta \in \Theta$ from observation $\omega \sim \mathcal{N}(\bar{A}\theta, \sigma^2 I)$, which is nothing but a special case of the problem considered in item 4.24.B; as is immediately seen, the Fisher Information matrix in this problem is independent of θ and is $\sigma^{-2}\bar{A}_E^*\bar{A}_E$:

$$e^T \mathcal{I}(\theta) e = \sigma^{-2} e^T \bar{A}_E^* \bar{A}_E e, \; e \in E.$$

Invoking Proposition 4.9.8, we arrive at the lower bound on the minimax risk in the auxiliary problem (and thus – in the problem of interest as well):

$$\mathrm{Risk}_{\mathrm{opt}} \geq \frac{r\sigma k}{r\sqrt{\gamma} + \sigma k}. \tag{4.9.37}$$

The resulting risk bound depends on r, k, γ and is the larger the smaller is γ and the larger are k and r.

Lower-bounding $\mathrm{Risk}_{\mathrm{opt}}$**.** In order to extract from the just outlined bounding scheme its best, we need a mechanism which allows to generate k-dimensional "disks" $\Theta \subset \bar{\mathcal{Y}}$ along with associated quantities r, γ. In order to design such a mechanism, it is convenient to represent k-dimensional linear subspaces of $\mathbf{R}^n$ as the image spaces of orthogonal $n \times n$ projectors P of rank k. Such a projector P gives rise to the contained in $\bar{\mathcal{Y}}$ disk Θ_P of the radius $r = r_P$, where r_P is the largest ρ such that the set $\{y \in \mathbf{R}^n : y^T P y \leq \rho^2\}$ is contained in $\bar{\mathcal{Y}}$ ("condition $\mathcal{C}(r)$"), and we can equip the disk with γ satisfying (ii) if and only if

$$\mathrm{Tr}(P\bar{A}^T \bar{A} P) \leq \gamma,$$

or, which is the same (recall that P is orthogonal projector)

$$\mathrm{Tr}(\bar{A} P \bar{A}^T) \leq \gamma \tag{4.9.38}$$

(“condition $\mathcal{D}(\gamma)$”). Now, when P is a nonzero orthogonal projector, the simplest sufficient condition for the validity of $\mathcal{C}(r)$ is the existence of $t \in \mathcal{T}$ such that

$$\forall (y \in \mathbf{R}^n, \ell \leq L) : y^T P \bar{S}_\ell P y \leq t_\ell r^{-2} y^T P y,$$

or, which is the same,

$$\exists s : r^2 s \in \mathcal{T} \ \&\ P\bar{S}_\ell P \preceq s_\ell P,\ \ell \leq L. \tag{4.9.39}$$

We are about to rewrite (4.9.38), (4.9.39) as a system of *linear* matrix inequalities. This is what you are supposed to do:

2.1. Prove the following simple fact:

Observation 4.9.2 *Let Q be a positive definite and R be a nonzero positive semidefinite matrix, and s be a real. Then*

$$RQR \preceq sR$$

if and only if

$$sQ^{-1} \succeq R.$$

2.2. Extract from Observation the conclusion as follows. Let $\mathbf{T}$ be the conic hull of $\mathcal{T}$:

$$\mathbf{T} = \mathrm{cl}\{[s;\tau] : \tau > 0, s/\tau \in \mathcal{T}\} = \{[s;\tau] : \tau > 0, s/\tau \in \mathcal{T}\} \cup \{0\}.$$

Consider the system of constraints

$$\begin{array}{c} s_\ell \bar{S}_\ell^{-1} \succeq P, \ell \leq L \ \&\ \mathrm{Tr}(\bar{A}P\bar{A}^T) \leq \gamma \\ P \text{ is orthogonal projector of rank } k \geq 1 \end{array} \tag{\#}$$

in variables $[s;\tau] \in \mathbf{T}$, k, γ and P. Every feasible solution to this system gives rise to k-dimensional Euclidean subspace $E \subset \mathbf{R}^n$ (the image space of P) such that the centered at the origin Euclidean ball Θ in E of radius

$$r = 1/\sqrt{\tau}$$

taken along with γ satisfy the conditions (i) - (ii). Consequently, this feasible solution yields the lower bound

$$\mathrm{Risk}_{\rm opt} \geq \psi_{\sigma,k}(\gamma,\tau) := \frac{\sigma k}{\sqrt{\gamma} + \sigma\sqrt{\tau}k}$$

on the minimax risk in the problem of interest.

An “ideal” way to utilize item 2.2 to lower-bound $\mathrm{Risk}_{\rm opt}$ would be to look through $k = 1, ..., n$ and for every k to maximize the lower risk bound $\psi_{\sigma,k}(\gamma,\tau)$ under constraints (#), thus arriving at the problem

$$\min_{[s;\tau],\gamma,P} \left\{ \frac{\sigma}{\psi_{\sigma,k}(\gamma,\tau)} = \sqrt{\gamma}/k + \sigma\sqrt{\tau} : \begin{array}{l} s_\ell \bar{S}_\ell^{-1} \succeq P, \ell \leq L \ \&\ \mathrm{Tr}(\bar{A}P\bar{A}^T) \leq \gamma \\ P \text{ is orthogonal projector of rank } k \end{array} \right\} \tag{P_k}$$

This problem seems to be computationally intractable, since the constraints of (P_k) include the nonconvex restriction on P to be an orthogonal projector of rank k. A natural convex relaxation of this restriction is

$$0 \preceq P \preceq I_n,\ \mathrm{Tr}(P) = k.$$

The (minor) remaining difficulty is that the objective in (P) is nonconvex. Note, however, that to minimize $\sqrt{\gamma}/k + \sigma\sqrt{\tau}$ is basically the same as to minimize the convex function $\gamma/k^2 + \sigma^2\tau$ which is a tight "proxy" of the squared objective of (P_k). We arrive at convex "proxy" of (P_k) – the problem

$$\min_{[s;\tau],\gamma,P}\left\{\gamma/k^2+\sigma^2\tau:\begin{array}{l}[s;\tau]\in\mathbf{T}, 0\preceq P\preceq I_n, \mathrm{Tr}(P)=k\\ s_\ell\bar{S}_\ell^{-1}\succeq P, \ell\leq L, \mathrm{Tr}(\bar{A}P\bar{A}^T)\leq\gamma\end{array}\right\}\tag{$P[k]$}$$

$k=1,...,n$. Problem $(P[k])$ clearly is solvable, and the P-component $P^{(k)}$ of its optimal solution gives rise to a bunch of orthogonal projectors $P_\kappa^{(k)}$, $\kappa=1,...,n$ obtained from $P^{(k)}$ by "rounding" – to get $P_\kappa^{(k)}$, we replace the κ leading eigenvalues of $P^{(k)}$ with ones, and the remaining eigenvalues – with zeros, while keeping the eigenvectors intact. We can now for every $\kappa=1,...,n$ fix the P-variable in (P_k) as $P_\kappa^{(k)}$ and solve the resulting problem in the remaining variables $[s;\tau]$ and γ, which is easy – with P fixed, the problem clearly reduces to the one of minimizing τ under the convex constraints

$$s_\ell\bar{S}_\ell^{-1}\succeq P, \ell\leq L, [s;\tau]\in\mathbf{T}$$

on $[s;\tau]$. As a result, for every $k\in\{1,...,n\}$, we get n lower bounds on $\mathrm{Risk}_{\rm opt}$, that is, total of n^2 lower risk bounds, of which we select the best – the largest.

Now goes the next part of the exercise:

3. *Implement the outlined methodology numerically and compare the lower bound on the minimax risk with the upper risk bounds of presumably good linear estimates yielded by Proposition 4.2.1.*

   Recommended setup:

   - *Sizes:* $m=n=\nu=16$
   - A, B: $B=I_n$, $A=\mathrm{Diag}\{a_1,...,a_n\}$ *with* $a_i=i^{-\alpha}$ *and* α *running through* $\{0,1,2\}$;
   - $\mathcal{X}=\{x\in\mathbf{R}^n: x^TS_\ell x\leq 1, \ell\leq L\}$ *(i.e.,* $\mathcal{T}=[0,1]^L$*) with randomly generated* S_ℓ.
     Range of L: $\{1,4,16\}$. *For* L *in this range, you can generate* S_ℓ, $\ell\leq L$, *as* $S_\ell=R_\ell R_\ell^T$ *with* $R_\ell=\mathtt{randn}(n,p)$, *where* $p=\rfloor n/L\lfloor$.
   - *Range of* σ: $\{1.0, 0.1, 0.01, 0.001, 0.0001\}$

4.24.D. More on Cramer-Rao risk bound. Let us fix $\mu\in(1,\infty)$ and a norm $\|\cdot\|$ on $\mathbf{R}^k$, and let $\|\cdot\|_*$ be the norm conjugate to $\|\cdot\|$, and $\mu_*=\frac{\mu}{\mu-1}$. Assume that we are in the situation of item 4.24.B and under assumptions 1) and 3) from this item; as about assumption 2) we now replace it with the assumption that the quantity

$$\mathcal{I}_{\|\cdot\|_*,\mu_*}(\theta):=\left[\mathbf{E}_{\omega\sim p(\cdot,\theta)}\left\{\|\nabla_\theta p(\omega,\theta)\|_*^{\mu_*}\right\}\right]^{1/\mu_*}$$

is well defined and bounded on Θ; in the sequel, we set

$$\mathcal{I}_{\|\cdot\|_*,\mu_*}=\sup_{\theta\in\Theta}\mathcal{I}_{\|\cdot\|_*,\mu_*}(\theta).$$

4. *Prove the following variant of Cramer-Rao risk hound:*

Proposition 4.9.9 *In the situation described in the beginning of item 4.24.D, let* $\Theta\subset\mathbf{R}^k$ *be a* $\|\cdot\|$*-ball of radius* r. *Then the minimax* $\|\cdot\|$*-risk of recovering* $\theta\in\Theta$ *via observation* $\omega\sim p(\cdot,\theta)$ *can be lower-bounded as*

$$\begin{array}{l}\mathrm{Risk}_{\mathrm{opt},\|\cdot\|}[\Theta]:=\inf\limits_{\widehat{\theta}(\cdot)}\sup\limits_{\theta\in\Theta}\left[\mathbf{E}_{\omega\sim p(\cdot,\theta)}\left\{\|\widehat{\theta}(\omega)-\theta\|^\mu\right\}\right]^{1/\mu}\geq\frac{rk}{r\mathcal{I}_{\|\cdot\|_*,\mu_*}+k},\\ \mathcal{I}_{\|\cdot\|_*,\mu_*}=\max\limits_{\theta\in\Theta}\left[\mathcal{I}_{\|\cdot\|_*,\mu_*}(\theta):=\left[\mathbf{E}_{\omega\sim p(\cdot,\theta)}\left\{\|\nabla_\theta\ln(p(\omega,\theta))\|_*^{\mu_*}\right\}\right]^{1/\mu_*}\right]\end{array}\tag{4.9.40}$$

Example I: Gaussian case, estimating shift. Let $\mu = 2$, and let $p(\omega,\theta) = \mathcal{N}(A\theta,\sigma^2 I_m)$ with $A \in \mathbf{R}^{m\times k}$. Then

$$\begin{array}{l}
\nabla_\theta \ln(p(\omega,\theta)) = \sigma^{-2}A^T(\omega - A\theta) \Rightarrow \\
\int \|\nabla_\theta \ln(p(\omega,\theta))\|_*^2 p(\omega,\theta)d\omega = \sigma^{-4}\int \|A^T(\omega - A\theta)\|_*^2 p(\omega,\theta)d\omega \\
= \sigma^{-4}\frac{1}{[\sqrt{2\pi}\sigma]^m}\int \|A^T\omega\|_*^2 \exp\{-\frac{\omega^T\omega}{2\sigma^2}\}d\omega \\
= \sigma^{-4}\frac{1}{[2\pi]^{m/2}}\int \|A^T\sigma\xi\|_*^2 \exp\{-\xi^T\xi/2\}d\xi \\
= \sigma^{-2}\frac{1}{[2\pi]^{m/2}}\int \|A^T\xi\|_*^2 \exp\{-\xi^T\xi/2\}d\xi
\end{array}$$

whence

$$\mathcal{I}_{\|\cdot\|_*,2} = \sigma^{-1}\underbrace{\left[\mathbf{E}_{\xi\sim\mathcal{N}(0,I_m)}\left\{\|A^T\xi\|_*^2\right\}\right]^{1/2}}_{\gamma_{\|\cdot\|}(A)}.$$

Consequently, assuming Θ to be $\|\cdot\|$-ball of radius r in $\mathbf{R}^k$, lower bound (4.9.40) becomes

$$\mathrm{Risk}_{\mathrm{opt},\|\cdot\|}[\Theta] \geq \frac{rk}{r\mathcal{I}_{\|\cdot\|_*} + k} = \frac{rk}{r\sigma^{-1}\gamma_{\|\cdot\|}(A) + k} = \frac{r\sigma k}{r\gamma_{\|\cdot\|}(A) + \sigma k}. \tag{4.9.41}$$

The case of direct observations. Just to see how it works, consider the case $m = k$, $A = I_k$ of direct observations, and let $\Theta = \{\theta \in \mathbf{R}^k : \|\theta\| \leq r\}$. Then

- We have $\gamma_{\|\cdot\|_1}(I_k) \leq O(1)\sqrt{\ln(n)}$, whence the $\|\cdot\|_1$-risk bound is

$$\mathrm{Risk}_{\mathrm{opt},\|\cdot\|_1}[\Theta] \geq O(1)\frac{r\sigma k}{r\sqrt{\ln(n)} + \sigma k}; \qquad [\Theta = \{\theta \in \mathbf{R}^k : \|\theta - a\|_1 \leq r\}]$$

- We have $\gamma_{\|\cdot\|_2}(I_k) = \sqrt{k}$, whence the $\|\cdot\|_2$-risk bound is

$$\mathrm{Risk}_{\mathrm{opt},\|\cdot\|_2}[\Theta] \geq \frac{r\sigma\sqrt{k}}{r + \sigma\sqrt{k}}; \qquad [\Theta = \{\theta \in \mathbf{R}^k : \|\theta - a\|_2 \leq r\}]$$

- We have $\gamma_{\|\cdot\|_\infty}(I_k) \leq O(1)k$, whence the $\|\cdot\|$-risk bound is

$$\mathrm{Risk}_{\mathrm{opt},\|\cdot\|_2}[\Theta] \geq O(1)\frac{r\sigma}{r + \sigma}. \qquad [\Theta = \{\theta \in \mathbf{R}^k : \|\theta - a\|_\infty \leq r\}]$$

In fact, the above examples are basically covered by the following

Observation 4.9.3 *Let $\|\cdot\|$ be a norm on $\mathbf{R}^k$, and let*

$$\Theta = \{\theta \in \mathbf{R}^k : \|\theta\| \leq r\}.$$

Consider the problem of recovering signal $\theta \in \Theta$ via observation $\omega \sim \mathcal{N}(\theta,\sigma^2 I_k)$, let

$$\mathrm{Risk}_{\|\cdot\|}[\widehat{\theta}|\Theta] = \sup_{\theta\in\Theta}\left(\mathbf{E}_{\omega\sim\mathcal{N}(\theta,\sigma^2 I)}\left\{\|\widehat{\theta}(\omega) - \theta\|^2\right\}\right)^{1/2}$$

be the $\|\cdot\|$-risk of an estimate $\widehat{\theta}(\cdot)$, and let

$$\mathrm{Risk}_{\mathrm{opt},\|\cdot\|}[\Theta] = \inf_{\widehat{\theta}(\cdot)} \mathrm{Risk}_{\|\cdot\|}[\widehat{\theta}|\Theta]$$

be the associated minimax risk.

Assume that the norm $\|\cdot\|$ is absolute and symmetric w.r.t permutation of coordinates. Then

$$\mathrm{Risk}_{\mathrm{opt},\|\cdot\|}[\Theta] \geq \frac{r\sigma k}{2\sqrt{\ln(ek)}r\alpha_* + \sigma k}, \quad \alpha_* = \|[1;...;1]\|_*. \tag{4.9.42}$$

Here is the concluding part of the exercise:

5. Prove Observation and compare the lower risk bound from Observation with the $\|\cdot\|$-risk of the "plug-in" estimate $\widehat{\chi}(\omega) \equiv \omega$.

Example II: Gaussian case, estimating covariance. Let $\mu = 2$, let K be a positive integer, and let our observation ω be a collection of K i.i.d. samples $\omega_t \sim \mathcal{N}(0, \theta)$, $1 \le t \le K$, with unknown θ known to belong to a given convex compact subset Θ of the interior of the positive semidefinite cone $\mathbf{S}^n_+$. Given ω_1,...,ω_K, we want to recover θ in the Shatten norm $\|\cdot\|_{\mathrm{Sh},s}$ with $s \in [1,\infty]$. Our estimation problem is covered by the setup of Exercise 4.24 with $\mathcal{P}$ comprised of the product probability densities $p(\omega,\theta) = \prod_{t=1}^K g(\omega_t,\theta)$, $\theta \in \Theta$, where $g(\cdot,\theta)$ is the density of $\mathcal{N}(0,\theta)$. We have

$$\begin{array}{rcl}\nabla_\theta \ln(p(\omega,\theta)) &=& \frac{1}{2}\sum_t \nabla_\theta \ln(g(\omega_t,\theta)) = \frac{1}{2}\sum_t \left[\theta^{-1}\omega_t\omega_t^T\theta^{-1} - \theta^{-1}\right] \\ &=& \frac{1}{2}\theta^{-1/2}\left[\sum_t \left[[\theta^{-1/2}\omega_t][\theta^{-1/2}\omega_t]^T - I_n\right]\right]\theta^{-1/2}\end{array} \tag{4.9.43}$$

With some effort it can be proved that when

$$K \ge n,$$

which we assume from now on, for independent across t random vectors $\xi_1, ..., \xi_K$ sampled from the standard Gaussian distribution $\mathcal{N}(0, I_n)$ for every $u \in [1,\infty]$ one has

$$\left[\mathbf{E}\left\{\|\sum_{t=1}^K [\xi_t\xi_t^T - I_n]\|^2_{\mathrm{Sh},u}\right\}\right]^{1/2} \le C n^{\frac{1}{2}+\frac{1}{u}}\sqrt{K} \tag{4.9.44}$$

with appropriate *absolute constant* C. Consequently, for $\theta \in \Theta$ and all $u \in [1,\infty]$ we have

$$\begin{array}{l}\mathbf{E}_{\omega\sim p(\cdot,\theta)}\left\{\|\nabla_\theta \ln(p(\omega,\theta))\|^2_{\mathrm{Sh},u}\right\} = \frac{1}{4}\mathbf{E}_{\omega\sim p(\cdot,\theta)}\left\{\|\theta^{-1/2}\left[\sum_t \left[[\theta^{-1/2}\omega_t][\theta^{-1/2}\omega_t]^T - I_n\right]\right]\theta^{-1/2}\|^2_{\mathrm{Sh},u}\right\} \\ \quad \text{[by (4.9.43)]} \\ = \frac{1}{4}\mathbf{E}_{\xi\sim p(\cdot,I_n)}\left\{\|\theta^{-1/2}\left[\sum_t \left[\xi_t\xi_t^T - I_n\right]\right]\theta^{-1/2}\|^2_{\mathrm{Sh},u}\right\} \quad \text{[setting } \theta^{-1/2}\omega_t = \xi_t\text{]} \\ \le \frac{1}{4}\|\theta^{-1/2}\|^4_{\mathrm{Sh},\infty}\mathbf{E}_{\xi\sim p(\cdot,I_n)}\left\{\|\sum_t \left[\xi_t\xi_t^T - I_n\right]\|^2_{\mathrm{Sh},u}\right\} \quad \text{[since } \|AB\|_{\mathrm{Sh},u}\| \le \|A\|_{\mathrm{Sh},\infty}\|B\|_{\mathrm{Sh},u}\text{]} \\ \le \frac{1}{4}\|\theta^{-1/2}\|^4_{\mathrm{Sh},\infty}\left[Cn^{\frac{1}{2}+\frac{1}{u}}\sqrt{K}\right]^2 \quad \text{[by (4.9.44)]}\end{array}$$

and we arrive at

$$\left[\mathbf{E}_{\omega\sim p(\cdot,\theta)}\left\{\|\nabla_\theta \ln(p(\omega,\theta))\|^2_{\mathrm{Sh},u}\right\}\right]^{1/2} \le \frac{C}{2}\|\theta^{-1}\|_{\mathrm{Sh},\infty} n^{\frac{1}{2}+\frac{1}{u}}\sqrt{K}. \tag{4.9.45}$$

Now assume that Θ is $\|\cdot\|_{\mathrm{Sh},s}$-ball of radius $r < 1$ centered at I_n:

$$\Theta = \{\theta \in \mathbf{S}^n : \|\theta - I_n\|_{\mathrm{Sh},s} \le r\}. \tag{4.9.46}$$

In this case the estimation problem from Example II is the scope of Proposition 4.9.9, and the quantity $I_{\|\cdot\|_*,2}$ as defined in (4.9.40) can be upper-bounded as follows:

$$\begin{array}{rcl}I_{\|\cdot\|_*,2} &=& \max\limits_{\theta\in\Theta}\left[\mathbf{E}_{\omega\sim p(\cdot,\theta)}\left\{\|\nabla_\theta \ln(p(\omega,\theta))\|^2_{\mathrm{Sh},s_*}\right\}\right]^{1/2} \\ &\le& O(1)n^{\frac{1}{2}+\frac{1}{s_*}}\sqrt{K}\max_{\theta\in\Theta}\|\theta^{-1}\|_{\mathrm{Sh},\infty} \text{ [see (4.9.45)]} \\ &\le& O(1)\frac{n^{\frac{1}{2}+\frac{1}{s_*}}\sqrt{K}}{1-r}.\end{array}$$

We can now use Proposition 4.9.9 to lower-bound the minimax $\|\cdot\|_{\mathrm{Sh},s}$-risk, thus arriving at

$$\mathrm{Risk}_{\mathrm{opt},\|\cdot\|_{\mathrm{Sh},s}}[\Theta] \ge O(1)\frac{n(1-r)r}{\sqrt{K}n^{\frac{1}{2}-\frac{1}{s}}r + n(1-r)} \tag{4.9.47}$$

(note that we are in the case of $k = \dim\theta = \frac{n(n+1)}{2}$).

Let us compare this lower risk bound with the $\|\cdot\|_{\mathrm{Sh},s}$-risk of the "plug-in" estimate

$$\widehat{\theta}(\omega)=\frac{1}{K}\sum_{t=1}^{K}\omega_t\omega_t^T.$$

Assuming $\theta\in\Theta$, we have

$$\begin{array}{l}
\mathbf{E}_{\omega\sim p(\cdot,\theta)}\left\{\|K[\widehat{\theta}(\omega)-\theta]\|^2_{\mathrm{Sh},s}\right\}=\mathbf{E}_{\omega\sim p(\cdot,\theta)}\left\{\|\sum_t[\omega_t\omega_t^T-\theta]\|^2_{\mathrm{Sh},s}\right\}\\
=\mathbf{E}_{\omega\sim p(\cdot,\theta)}\left\{\|\theta^{1/2}\left[\sum_t[[\theta^{-1/2}\omega_t][\theta^{-1/2}\omega_t]^T-I_n]\right]\theta^{1/2}\|^2_{\mathrm{Sh},s}\right\}\\
=\mathbf{E}_{\xi\sim p(\cdot,I_n)}\left\{\|\theta^{1/2}\left[\sum_t[\xi_t\xi_t^T-I_n]\right]\theta^{1/2}\|^2_{\mathrm{Sh},s}\right\}\\
\leq\|\theta^{1/2}\|^4_{\mathrm{Sh},\infty}\mathbf{E}_{\xi\sim p(\cdot,I_n)}\left\{\|\sum_t[\xi_t\xi_t^T-I_n]\|^2_{\mathrm{Sh},s}\right\}\\
\leq\|\theta^{1/2}\|^4_{\mathrm{Sh},\infty}\left[Cn^{\frac{1}{2}+\frac{1}{s}}\sqrt{K}\right]^2,\ \text{[see (4.9.44)]}
\end{array}$$

and we arrive at

$$\mathrm{Risk}_{\|\cdot\|_{\mathrm{Sh},s}}[\widehat{\theta}|\Theta]\leq O(1)\max_{\theta\in\Theta}\|\theta\|_{\mathrm{Sh},\infty}\frac{n^{\frac{1}{2}+\frac{1}{s}}}{\sqrt{K}}.\tag{4.9.48}$$

In the case of (4.9.46), the latter bound becomes

$$\mathrm{Risk}_{\|\cdot\|_{\mathrm{Sh},s}}[\widehat{\theta}|\Theta]\leq O(1)\max_{\theta\in\Theta}\|\theta\|_{\mathrm{Sh},\infty}\frac{n^{\frac{1}{2}+\frac{1}{s}}}{\sqrt{K}}.\tag{4.9.49}$$

For the sake of simplicity, assume that r in (4.9.46) is 1/2 (what actually matters below is that $r\in(0,1)$ is bounded away from 0 and from 1). In this case the lower bound (4.9.47) on the minimax $\|\cdot\|_{\mathrm{Sh},s}$-risk reads

$$\mathrm{Risk}_{\mathrm{opt},\|\cdot\|_{\mathrm{Sh},s}}[\Theta]\geq O(1)\min\left[\frac{n^{\frac{1}{2}+\frac{1}{s}}}{\sqrt{K}},1\right].$$

When K is "large:" $K\geq n^{1+\frac{2}{s}}$, this lower bound matches, within an absolute constant factor, the upper bound (4.9.49) on the risk of the plug-in estimate, so that the latter estimate is near-optimal. When $K<n^{1+\frac{2}{s}}$, the lower risk bound becomes $O(1)$, so that here a nearly optimal estimate is the trivial estimate $\widehat{\theta}(\omega)\equiv I_n$.

Exercise 4.25 † [follow-up to Exercise 4.24]

1. *Prove the following version of Proposition 4.9.8:*

Proposition 4.9.10 *In the situation of item 4.24.B and under assumptions 1) – 3) from this item, let*

- *$\|\cdot\|$ be a norm on $\mathbf{R}^k$ such that*

$$\|\theta\|_2\leq\kappa\|\theta\|\ \forall\theta\in\mathbf{R}^k$$

- *$\Theta\subset\mathbf{R}^k$ be $\|\cdot\|$-ball of radius $r>0$,*
- *the family $\mathcal{P}$ be such that $\mathcal{I}(\theta)\preceq\mathcal{J}$ for some $\mathcal{J}\succ0$ and all $\theta\in\Theta$.*

Then the minimax optimal risk

$$\mathrm{Risk}_{\mathrm{opt},\|\cdot\|}=\inf_{\widehat{\theta}(\cdot)}\left(\sup_{\theta\in\Theta}\mathbf{E}_{\omega\sim p(\cdot,\theta)}\left\{\|\theta-\widehat{\theta}(\omega)\|^2\right\}\right)^{1/2}$$

of recovering parameter $\theta \in \Theta$ from observation $\omega \sim p(\cdot, \theta)$ in the norm $\|\cdot\|$ satisfies the bound

$$\mathrm{Risk}_{\mathrm{opt},\|\cdot\|} \geq \frac{rk}{r\kappa\sqrt{\mathrm{Tr}(\mathcal{J})} + k}. \tag{4.9.50}$$

In particular, when $\mathcal{J} = \alpha^{-1} I_k$, we get

$$\mathrm{Risk}_{\mathrm{opt},\|\cdot\|} \geq \frac{r\sqrt{\alpha k}}{r\kappa + \sqrt{\alpha k}}. \tag{4.9.51}$$

2. *Apply Proposition 4.9.10 to get lower bounds on the minimax $\|\cdot\|$-risk in the following estimation problems:*

   2.1. *Given indirect observation $\omega = A\theta + \sigma\xi$, $\xi \sim \mathcal{N}(0, I_m)$ of unknown vector θ known to belong to $\Theta = \{\theta \in \mathbf{R}^k : \|\theta\|_p \leq r\}$ with given A, $\mathrm{Ker}\, A = \{0\}$, $p \in [2, \infty]$, $r > 0$, we want to recover θ in $\|\cdot\|_p$.*

   2.2. *Given indirect observation $\omega = L\theta R + \sigma\xi$, where θ is unknown $\mu \times \nu$ matrix known to belong to the Shatten norm ball $\Theta \in \mathbf{R}^{\mu\times\nu} : \|\theta\|_{\mathrm{Sh},p} \leq r$, we want to recover θ in $\|\cdot\|_{\mathrm{Sh},p}$. Here $L \in \mathbf{R}^{m\times\mu}, \mathrm{Ker}\, L = \{0\}$ and $R = \mathbf{R}^{\nu\times n}, \mathrm{Ker}\, R^T = \{0\}$ are given matrices, $p \in [2, \infty]$, and ξ is random Gaussian $m \times n$ matrix (i.e., the entries in ξ are independent of each other $\mathcal{N}(0, 1)$ random variables).*

   2.3. *Given K-repeated observation $\omega^K = (\omega_1, ..., \omega_K)$ with i.i.d. components $\omega_t \sim \mathcal{N}(0, \theta)$, $1 \leq t \leq K$, with unknown $\theta \in \mathbf{S}^n$ known to belong to the matrix box $\Theta = \{\theta : \beta_- I_n \preceq \theta \preceq \beta_+ I_n\}$ with given $0 < \beta_- < \beta_+ < \infty$, we want to recover θ in the spectral norm.*

4.9.7 Around $\mathcal{S}$-Lemma

Exercise 4.26 *Proposition 4.3.3 provides us with upper bound on the quality of semidefinite relaxation as applied to the problem of upper-bounding the maximum of a homogeneous quadratic form over an ellitope. Extend the construction to the case when an inhomogeneous quadratic form is maximized over a shifted ellitope, so that quantity to upper-bound is*

$$\mathrm{Opt} = \max_{x\in X} \left[f(x) := x^T Ax + 2b^T x + c\right],\ X = \{x : \exists (y, t \in \mathcal{T}) : x = Py+p, y^T S_k y \leq t_k, 1 \leq k \leq K\}$$

with our standard assumptions on S_k and $\mathcal{T}$.

Note: X is centered at p, and a natural upper bound on Opt is

$$\mathrm{Opt} \leq f(p) + \widehat{\mathrm{Opt}},$$

where $\widehat{\mathrm{Opt}}$ is an upper bound on the quantity

$$\overline{\mathrm{Opt}} = \max_{x\in X} [f(x) - f(p)];$$

what you are interested to upper-bound, is the ratio $\widehat{\mathrm{Opt}}/\overline{\mathrm{Opt}}$.

$\mathcal{S}$-Lemma is a classical result of extreme importance in Semidefinite Optimization. Basically, Lemma states that when the ellitope $\mathcal{X}$ in Proposition 4.3.3 is just an ellipsoid, (4.3.19) can be strengthen to $\mathrm{Opt} = \mathrm{Opt}_*$. In fact, $\mathcal{S}$-Lemma is even stronger:

Lemma 4.9.5 [$\mathcal{S}$-Lemma] *Consider two quadratic forme $f(x) = x^TAx+2a^Tx+\alpha$, $g(x) = x^TBx+2b^Tx+\beta$ such that $g(\bar{x}) < 0$ for some $\bar{x}$. Then the implication*

$$g(x) \leq 0 \Rightarrow f(x) \leq 0$$

takes place if and only if for some $\lambda \geq 0$ it holds $f(x) \leq \lambda g(x)$ for all x, or, which is the same, if and only if Linear Matrix Inequality

$$\left[\begin{array}{c|c} \lambda B - A & \lambda b - a \\ \hline \lambda b^T - a^T & \lambda\beta - \alpha \end{array}\right] \succeq 0$$

in scalar variable λ has a nonnegative solution.

Proof of $\mathcal{S}$-Lemma can be found, e.g., in [14, Section 3.5.2]

The goal of subsequent exercises is to get "tight" tractable outer approximations of sets obtained from ellitopes by quadratic lifting. We fix an ellitope

$$X = \{x \in \mathbf{R}^n : \exists t \in \mathcal{T} : x^TS_kx \leq t_k, 1 \leq k \leq K\} \tag{4.9.52}$$

where, as always, S_k are positive semidefinite matrices with positive definite sum, and $\mathcal{T}$ is a computationally tractable convex compact subset in $\mathbf{R}^k_+$ such that $t \in \mathcal{T}$ implies $t' \in \mathcal{T}$ whenever $0 \leq t' \leq t$ and $\mathcal{T}$ contains a positive vector.

Exercise 4.27 *Let us associate with ellitope X given by (4.9.52) the sets*

$$\mathcal{X} = \mathrm{Conv}\{xx^T : x \in X\},\ \widehat{\mathcal{X}} = \{Y \in \mathbf{S}^n : Y \succeq 0, \exists t \in \mathcal{T} : \mathrm{Tr}(S_kY) \leq t_k, 1 \leq k \leq K\},$$

so that $\mathcal{X}$, $\widehat{\mathcal{X}}$ are convex compact sets containing the origin, and $\widehat{\mathcal{X}}$ is computationally tractable along with $\mathcal{T}$. Prove that

1. *When $K = 1$, we have $\mathcal{X} = \widehat{\mathcal{X}}$;*

2. *We always have $\mathcal{X} \subset \widehat{\mathcal{X}} \subset 4\ln(5K)\mathcal{X}$.*

Exercise 4.28 *For $x \in \mathbf{R}^n$ let $Z(x) = [x; 1][x; 1]^T$, $Z^o[x] = \left[\begin{array}{c|c} xx^T & x \\ \hline x^T & \end{array}\right]$. Let*

$$C = \left[\begin{array}{c|c} & \\ \hline & 1 \end{array}\right],$$

and let us associate with ellitope X given by (4.9.52) the sets

$$\begin{array}{rcl} \mathcal{X}^+ & = & \mathrm{Conv}\{Z^o[x] : x \in X\}, \\ \widehat{\mathcal{X}}^+ & = & \{Y = \left[\begin{array}{c|c} U & u \\ \hline u^T & \end{array}\right] \in \mathbf{S}^{n+1} : Y + C \succeq 0, \exists t \in \mathcal{T} : \mathrm{Tr}(S_kU) \leq t_k, 1 \leq k \leq K\}, \end{array}$$

so that $\mathcal{X}^+$, $\widehat{\mathcal{X}}^+$ are convex compact sets containing the origin, and $\widehat{\mathcal{X}}^+$ is computationally tractable along with $\mathcal{T}$. Prove that

1. *When $K = 1$, we have $\mathcal{X}^+ = \widehat{\mathcal{X}}^+$;*

2. *We always have $\mathcal{X}^+ \subset \widehat{\mathcal{X}}^+ \subset 4\ln(5(K+1))\mathcal{X}^+$.*

4.9.8 Beyond the scope of linear estimates

Exercise 4.29 † [beyond the scope of linear estimates] In Lecture 4, we were considering the problem of recovering the image Bx of unknown signal x known to belong to a given signal set $\mathcal{X}$ from a noisy observation

$$\omega = Ax + \xi.$$

We have seen that if $\mathcal{X}$ is an ellitope/spectratope, then, under reasonable assumptions on how we measure the recovery error, an appropriate efficiently computable *linear in* ω estimate is near-optimal. Note that the ellitopic/spectratopic structure of $\mathcal{X}$ is crucial here. Indeed, consider the simply-looking problem of recovering $Bx = x$ in $\|\cdot\|_2$-norm from *direct* observations ($Ax = x$) corrupted by the standard Gaussian noise $\xi \sim \mathcal{N}(0, \sigma^2 I)$, and let $\mathcal{X}$ be the unit $\|\cdot\|_1$=ball:

$$\mathcal{X} = \{x \in \mathbf{R}^n : \sum_i |x_i| \leq 1\}.$$

In this situation, building the optimal, in terms of the worst-case, over $x \in \mathcal{X}$, expected squared risk linear estimate $\widehat{x}_H(\omega) = H^T\omega$ is extremely simple:

$$\begin{array}{rcl}\mathrm{Risk}^2[\widehat{x}_H|\mathcal{X}] & := & \max_{x\in\mathcal{X}} \mathbf{E}\left\{\|\widehat{x}_H(\omega - Bx\|_2^2\right\} = \max_{x\in\mathcal{X}} \left\{\|[I - H^T]x\|_2^2 + \sigma^2 \mathrm{Tr}(HH^T)\right\} \\ & = & \max_{i\leq n} \|\mathrm{Col}_i[I - H^T]\|_2^2 + \sigma^2\mathrm{Tr}(HH^T).\end{array}$$

Clearly, the optimal H is just a scalar matrix hI, the optimal h is the minimizer of the univariate quadratic function $(1-h)^2 + \sigma^2 nh^2$, and the best achievable with linear estimates squared risk is

$$R^2 = \min_h \left[(1-h)^2 + \sigma^2 nh^2\right] = \frac{n\sigma^2}{1+n\sigma^2}.$$

On the other hand, consider *nonlinear* estimate as follows. Given σ, "safety factor" $\rho \geq 1$ and observation ω, specify the estimate $\widehat{x}(\omega)$ as an optimal solution to the optimization problem

$$\mathrm{Opt}(\omega) = \min_{y\in\mathcal{X}} \|y - \omega\|_\infty.$$

Note that the probability for the true signal to satisfy $\|x - \omega\|_\infty \leq \rho\sigma$ is at least $1 - p$, $p = 2n\exp\{-\rho^2/2\}$, and if this event $\mathcal{E}$ happens, then both x and $\widehat{x}$ belong to the box $\{y : \|y - \omega\|_\infty \leq \rho\sigma\}$, implying that $\|x - \widehat{x}\|_\infty \leq 2\rho\sigma$; in addition, we always have $\|x - \widehat{x}\|_2 \leq \|x - \widehat{x}\|_1 \leq 2$, since $x \in \mathcal{X}$ and $\widehat{x} \in \mathcal{X}$. We therefore have

$$\|x - \widehat{x}\|_2 \leq \sqrt{\|x - \widehat{x}\|_\infty \|x - \widehat{x}\|_1} \leq \left\{\begin{array}{ll} 2\sqrt{\rho\sigma}, & \omega \in \mathcal{E} \\ 2, & \omega \notin \mathcal{E}\end{array}\right.,$$

whence

$$\mathbf{E}\left\{\|\widehat{x} - x\|_2^2\right\} \leq 4\rho\sigma + 4p \leq 4\rho\sigma + 8n\exp\{-\rho^2/2\}. \tag{$*$}$$

Assuming $\sigma \leq 2n\exp\{-1/2\}$ and specifying ρ as $\sqrt{2\ln(2n/\sigma)}$, we get $\rho \geq 1$ and $2n\exp\{-\rho^2/2\} \leq \sigma$, implying that the right hand side in ($*$) is at most $8\rho\sigma$. In other words, for our nonlinear estimate $\widehat{x}(\omega)$ it holds

$$\mathrm{Risk}^2[\widehat{x}|\mathcal{X}] \leq 8\sqrt{\ln(2n/\sigma)}\sigma.$$

When $n\sigma^2$ is of order of 1, the latter bound on the squared risk is of order of $\sigma\sqrt{\ln(1/\sigma)}$, while the best squared risk achievable with linear estimates under the circumstances is of order of 1. We conclude that when σ is small and n is large (specifically, is of order of $1/\sigma^2$), the best linear estimate is *by far* inferior as compared to our nonlinear estimate – the ratio of the corresponding squared risks is as large as $\frac{O(1)}{\sigma\sqrt{\ln(1/\sigma)}}$ – the factor which is "by far" worse than the nonoptimality factor in the case of ellitope/spectratope $\mathcal{X}$.

4.29.A Generic polyhedral estimate. The nonlinear estimate $\widehat{x}$ which we have built (in fact, this estimate is nearly optimal under the circumstances in a meaningful range of values of n and σ) is given by a construction of the following generic form (for the sake of simplicity, we restrict our subsequent considerations to the case $\xi \sim \mathcal{N}(0, \sigma^2 I)$):

> Given the data $A \in \mathbf{R}^{m\times n}, B \in \mathbf{R}^{\nu\times n}, \mathcal{X} \subset \mathbf{R}^n, \sigma$ of our recovery problem (where $\mathcal{X}$ is a computationally tractable convex compact set), we specify somehow positive integer N, confidence tolerance $\epsilon \in (0,1)$, and N linear forms $g_i^T x$ on the space $\mathbf{R}^n$ of signals. We then use the machinery of Section 3.3 to build affine in ω estimates
>
> $$\widehat{g}_i(\omega) = \overline{g}_i^T \omega + \theta_i,\, i \leq N,$$
>
> along with "confidence widths" $\rho_i > 0$ such that
>
> $$\mathrm{Prob}_{\xi\sim\mathcal{N}(0,\sigma^2 I)}\left\{|\widehat{g}_i(Ax+\xi) - g_i^T x| > \rho_i\right\} \leq \epsilon/N,\, i = 1, ..., N,\ \forall x \in \mathcal{X}. \tag{4.9.53}$$
>
> In the sequel, we set
>
> $$S = \mathrm{Diag}\{\rho_1^{-1}, \rho_2^{-1}, ..., \rho_N^{-1}\},\ G = [g_1^T; g_2^T; ...; g_N^T],\ \widehat{g}(\omega) = [\widehat{g}_1(\omega); \widehat{g}_2(\omega); ...\widehat{g}_N(\omega)].$$
>
> Note that due to the origin of ρ_i's (see (4.9.53)), for every $x \in \mathcal{X}$ we have
>
> $$\mathrm{Prob}\{\mathcal{E}_x := \{\xi : \|S(Gx - \widehat{g}(Ax+\xi))\|_\infty \leq 1\}\} \geq 1 - \epsilon. \tag{4.9.54}$$
>
> Our estimate is as follows: given ω, we specify $\overline{x}(\omega)$ as (any) optimal solution to the optimization problem
>
> $$\min_{y\in\mathcal{X}} \|S(Gy - \widehat{g}(\omega))\|_\infty,$$
>
> and estimate Bx by $\widehat{x}(\omega) = B\overline{x}(\omega)$.

In the sequel, we refer to the just specified estimate as to the *polyhedral* one. The rationale behind this approach is that in the case of Gaussian observation scheme (same as in the cases of other simple o.s.'s, e.g. Poisson and Discrete ones), the machinery from Section 3.3 allows for *near-optimal* estimation of linear forms on a *whatever* computationally tractable convex compact signal set $\mathcal{X}$ (see Proposition 3.3.4 and [92]), which makes it natural to try to reduce more complex estimating problems to those of estimating linear forms. To the best of our knowledge, this approach was first used in [125, Section 2] in connection with recovering (restrictions on regular grids of) multivariate functions from Sobolev balls from direct observations.

4.29.B Upper-bounding ϵ-risk of polyhedral estimate. Our local goal is to upper-bound ϵ-risk

$$\mathrm{Risk}_{\epsilon,\|\cdot\|}[\widehat{x}|\mathcal{X}] = \inf_\rho \left\{\rho : \mathrm{Prob}_{\xi\sim\mathcal{N}(0,\sigma^2 I)}\left\{\|Bx - \widehat{x}(Ax+\xi)\| > \rho\right\} \leq \epsilon\, \forall x \in \mathcal{X}\right\}$$

of the just defined polyhedral estimate $\widehat{x}(\cdot)$; here $\|\cdot\|$ is some given norm on the space $\mathbf{R}^\nu$ where Bx takes values.

1.1. Let

$$H = \frac{1}{2}SG = [\frac{1}{2\rho_1}g_1^T; ...; \frac{1}{2\rho_N}g_N^T]$$

and

$$\mathrm{Opt} = \max_{x,y\in\mathcal{X}} \{\|B(x-y)\| : x \in \mathcal{X}, y \in \mathcal{X}, \|H(x-y)\|_\infty \leq 1\}$$

Prove that

$$\mathrm{Risk}_{\epsilon,\|\cdot\|}[\widehat{x}|\mathcal{X}] \leq \mathrm{Opt}. \tag{4.9.55}$$

1.2. Assume that we have at our disposal a set $\mathcal{U}_{\mathcal{X}}$ matching $\mathcal{X}$ and a set $\mathcal{U}_{\|\cdot\|_*}$ matching the unit ball of the norm $\|\cdot\|_*$ conjugate to $\|\cdot\|$ (for definition of matching and related facts, see Exercise 4.11). Prove that

$$\begin{array}{l}
\mathrm{Risk}_{\epsilon,\|\cdot\|}[\widehat{x}|\mathcal{X}] \leq \max\limits_{\substack{U_{uu},U_{ux},U_{uy},\\ U_{xx},U_{xy},U_{yy}}} \left\{ \mathrm{Tr}\left(\overline{B}\left[\begin{array}{c|c|c} U_{uu} & U_{ux} & U_{uy} \\ \hline U_{ux}^T & U_{xx} & U_{xy} \\ \hline U_{uy}^T & U_{xy}^T & U_{yy}\end{array}\right]\right) : \right. \\
\left. \left[\begin{array}{c|c|c} U_{uu} & U_{ux} & U_{uy} \\ \hline U_{ux}^T & U_{xx} & U_{xy} \\ \hline U_{uy}^T & U_{xy}^T & U_{yy}\end{array}\right] \succeq 0, U_{uu} \in \mathcal{U}_{\|\cdot\|_*}, U_{xx} \in \mathcal{U}_{\mathcal{X}}, U_{yy} \in \mathcal{U}_{\mathcal{X}}, \|[H,-H]\left[\begin{array}{c|c} U_{xx} & U_{xy} \\ \hline U_{xy}^T & U_{yy}\end{array}\right][H^T;-H^T]\|_\infty \leq 1 \right\}, \\
\qquad\qquad\qquad \overline{B} := \left[\begin{array}{c|c|c} & \frac{1}{2}B & -\frac{1}{2}B \\ \hline \frac{1}{2}B^T & & \\ \hline -\frac{1}{2}B^T & & \end{array}\right].
\end{array}$$

Note that the right hand side quantity is efficiently computable.

Refine the above upper risk bound in the case when

$$\mathcal{X} \subset \{x \in \mathbf{R}^n : p(x) \leq 1, Px = 0\},$$

$p(\cdot)$ being an absolute norm, $\|\cdot\|$ is an absolute norm, so that $q(\cdot) \equiv \|\cdot\|_*$ is an absolute norm as well, and

$$\begin{array}{rcl}
\mathcal{U}_{\mathcal{X}} & = & \{U \in \mathbf{S}^n : U \succeq 0, PU = 0, \mathrm{diag}(U) \in \mathcal{D}_{p(\cdot)}, p^+(U) \leq 1\}, \\
\mathcal{U}_{\|\cdot\|_*} & = & \{V \in \mathbf{S}^\nu : V \succeq 0, \mathrm{diag}(V) \in \mathcal{D}_{q(\cdot)}, q^+(V) \leq 1\},
\end{array}$$

where the sets $\mathcal{D}_{p(\cdot)}$, $\mathcal{D}_{q(\cdot)}$ are square-dominating $p(\cdot)$, resp., $q(\cdot)$, see Exercise 4.11.

4.29.C Simple case. The major conceptual issue with the outlined generic estimation scheme is how to specify the linear forms $g_i^T x$. The goal of what follows is to investigate this scheme in the extremely simple case where the above conceptual issue can be easily resolved, and the resulting estimate can be proved to be near-optimal. This simple case is as follows:

- $\mathcal{X}$ is "scaled r-ball:"
$$\mathcal{X} = \{x \in \mathbf{R}^n : \sum_i a_i^r |x_i|^r \leq 1\}$$
with $a_i > 0$ and some $r \in [1,2)$ (when $r \geq 2$, the just defined set $\mathcal{X}$ is an ellitope, so that we know better than to use the above scheme);

- both A and B are unit matrices ("denoising signal from direct observations"), and the recovery error is measured in $\|\cdot\|_p$-norm with $p \geq r$;

- $\xi \sim \mathcal{N}(0,\sigma^2 I)$.

Clearly, we can assume w.l.o.g. that $a_1 \leq a_2 \leq ... \leq a_n$. Besides this, a straightforward scaling of x and σ allows to assume that $a_1 = 1$.

In the situation in question it is natural to implement the above generic scheme as follows:

- we set $N = n$ and $g_i^T x = x_i$;

- whatever easy under the circumstances is to specify $\widehat{g}_i(\omega)$ via the techniques from Section 3.3, we intend to use the simplest estimates $\widehat{g}(\omega)$ of $g_i^T x = x_i$ – just the plug-in estimates $\widehat{g}_i(\omega) = \omega_i$;

- Same as in the above motivating example, instead of introducing confidence tolerance ϵ, we operate with safety parameter $\rho \geq 1$ and set

$$\rho_i = \sigma\rho,$$

resulting in

$$G = I_n, S = \frac{1}{\sigma\rho} I_n, H = \frac{1}{2\sigma\rho} I_n, \widehat{g}(\omega) \equiv \omega$$

and ensuring that

$$\forall x \in \mathbf{R}^n : \mathrm{Prob}_{\xi\sim\mathcal{N}(0,\sigma^2 I)} \{\|H(x-\omega)\|_\infty > 1/2\} \leq p_n(\rho) := 2n \exp\{-\rho^2/2\}.$$

- Given ω, we estimate x as an optimal solution to the optimization problem

$$\min_{x\in\mathcal{X}} \|SG(x-\omega)\|_\infty,$$

or, which is the same, of the problem

$$\min_{x\in\mathcal{X}} \|x-\omega\|_\infty.$$

4.29.C.1 Risk analysis. Let us fix signal $x \in \mathcal{X}$ underlying observation $\omega = x+\xi$, and let $\mathcal{E}$ be the event $\|\xi\|_\infty \leq \sigma\rho$, so that $\mathrm{Prob}\{\mathcal{E}\} \geq 1-p_n(\rho)$.

2.1. Prove that setting $\Delta = \Delta[\xi] := x - \widehat{x}(x+\xi)$ (Δ is the recovery error), one has

$$\begin{array}{lll} \forall \xi : & \sum_i a_i^r |\Delta_i|^r \leq 2^r & (a), \\ \forall \xi \in \mathcal{E} : & \|\Delta_i\|_\infty \leq 2\rho\sigma & (b) \end{array} \tag{4.9.56}$$

2.2. For $s \geq 0$, let $M(s)$ be the largest $m \leq n$ such that $s[\sum_{i=1}^m a_i^r]^{1/r} \leq 1$. Prove that

$$\xi \in \mathcal{E} \Rightarrow \|\Delta\|_p \leq 2\rho\sigma\,[1+M(\rho\sigma)]^{1/p}. \tag{!}$$

2.3. Derive from 2.2 that specifying $\rho = \sqrt{2\ln(2n/\sigma)}$ we arrive at estimate $\widehat{x}_\sigma(\cdot)$ such that in the range $\sigma \leq n$ it holds

$$\mathrm{Risk}_{\|\cdot\|_p}[\widehat{x}_\sigma|\mathcal{X}] := \max_{x\in\mathcal{X}} \mathbf{E}_{\xi\sim\mathcal{N}(0,\sigma^2 I)} \{\|\widehat{x}_\sigma(x+\xi)-x\|_p\} \leq 4\sigma\sqrt{2\ln(2n/\sigma)} \left[1+M(\sigma\sqrt{2\ln(2n/\sigma)})\right]^{1/p}.$$

4.29.C.2 Near-optimality. The goal of what follows is to demonstrate that in the simple situation under consideration, the estimate $\widehat{x}_\sigma$ is nearly minimax optimal. To this end we need the following simple result of Nonparametric Statistics:

> (#) *Consider the situation as follows: we are given $N \geq 2$ vectors $f^i \in \mathbf{R}^m$, $1 \leq i \leq N$, a positive σ such that*
>
> $$\forall (i,j) : \frac{\|f^i - f^j\|_2^2}{2\sigma^2} < \frac{1}{2}\ln(N-1) - \ln(2) \tag{4.9.57}$$
>
> *and a norm $\|\cdot\|$ on $\mathbf{R}^n$. Assume now that a signal f known to belong to the set $\mathcal{F} = \{f^1,...,f^N\}$ is observed according to*
>
> $$\omega = f + \xi,\ \xi \sim \mathcal{N}(0,\sigma^2 I).$$
>
> *Then for every estimate $\widehat{f} : \mathbf{R}^m \to \mathbf{R}^m$ its $\|\cdot\|$-risk on $\mathcal{F}$*
>
> $$\mathrm{Risk}_{\|\cdot\|}[\widehat{f}|\mathcal{F}] = \sup_{f\in\mathcal{F}} \mathbf{E}_{\xi\sim\mathcal{N}(0,\sigma^2 I)} \left\{\|\widehat{f}(f+\xi) - f\|\right\}$$
>
> *satisfies the lower bound*
>
> $$\mathrm{Risk}_{\|\cdot\|}[\widehat{f}|\mathcal{F}] \geq \frac{1}{4}\min_{i\neq j} \|f^i - f^j\|. \tag{4.9.58}$$

This is a straightforward adaptation to the situation we are interested in of Proposition 1.2.3 in [125]; this Proposition, in turn, is a direct consequence of the basic *Fano Inequality* from Information Theory.

Our course of actions is as follows.

A. Given $\sigma \in (0,n)$, let us specify ρ according to item 1.3, and let $m = M(\rho\sigma)$. We assume from now on that $m \geq 1$, meaning that σ is not too large, specifically,

$$\rho\sigma = \sigma\sqrt{2\ln(2n/\sigma)} \leq 1.$$

(recall that $a_1 = 1$). As a result, $m+1 \leq 2m$, and the risk bound from item 1.3 implies that

$$\mathrm{Risk}_{\|\cdot\|_p}[\widehat{x}_\sigma|\mathcal{X}] \leq O(1)\sigma\sqrt{2\ln(2n/\sigma)}M^{1/p}(\sigma\sqrt{2\ln(2n/\sigma)}). \tag{4.9.59}$$

B. Let $B = \{f \in \mathbf{R}^m : \|f\|_\infty \leq \sigma\}$. By definition of $M(\cdot)$, $\mathcal{X}$ contains the set

$$\{x = [f; 0_{n-m}] : f \in B\},$$

so that the problem of recovering $f \in B$ from observation $f + \zeta$, $\zeta \sim (0, \sigma^2 I_m)$ clearly reduces to the problem of recovering $x \in \mathcal{X}$ from observation $x + \xi$, $\xi \sim \mathcal{N}(0, \sigma^2 I_n)$. As a result, the minimax $\|\cdot\|_p$-risk

$$\mathrm{Risk}_* = \inf_{\widehat{f}(\cdot)} \mathrm{Risk}_{\|\cdot\|_p}[\widehat{f}|B]$$

in the former problem is a *lower bound* on the minimax $\|\cdot\|_p$-risk

$$\mathrm{Risk}^* = \inf_{\widehat{x}}\left[\mathrm{Risk}_{\|\cdot\|_p}[\widehat{x}|\mathcal{X}] := \sup_{x\in\mathcal{X}} \mathbf{E}_{\xi\sim\mathcal{N}(0,\sigma^2 I)}\left\{\|\widehat{x}(x+\xi) - x\|_p\right\}\right]$$

in the problem of interest. What we are about to do, is to demonstrate that we can select $N = \exp\{O(1)m\}$ vectors $f^i \in B$ such that they are not too far away from each other, in $\|\cdot\|_2$, so that (4.9.57) holds true, and at the same time the pairwise $\|\cdot\|_p$-distances between distinct f_i's are not too small, so that (#) gives a "reasonable" lower bound on Risk_* (and thus – on $\mathrm{Risk}^* \geq \mathrm{Risk}_*$). As we shall see, in this way we can build a matching (4.9.59) lower bound on the minimax risk Risk^* in the estimation problem of interest.

Let us implement the outlined strategy.

3.1. Let ζ be an m-dimensional random Rademacher vector. Prove that

$$\mathrm{Prob}\left\{\mathrm{Card}(\{i : \zeta_i = 1\}) > \frac{3m}{4}\right\} \leq \exp\{-m/8\}.$$

3.2. Let ζ^1, ζ^2 be two independent m-dimensional random Rademacher vectors. Prove that

$$\mathrm{Prob}\left\{\|\zeta^1 - \zeta^2\|_p < 2(m/4)^{1/p}\right\} \leq \exp\{-m/8\}.$$

3.3. Let N be a positive integer such that $\frac{N(N-1)}{2} < \exp\{m/8\}$. Prove that there exists a collection of N vectors $g^i \in \mathbf{R}^m$, $i \leq N$, such that $\|g^i\|_\infty \leq 1$ for all i and $\|g^i - g^j\|_p \geq 2(m/4)^{1/p}$ whenever $i, j \leq N$ and $i \neq j$.

In order to allow for varying values of n, assume that $a_1, ..., a_n$ are the initial n terms of a given sequence $a^\infty = \{1 = a_1 \leq a_2 \leq a_3 \leq ...\}$, and assume that $m = M(\rho\sigma) = M(\sigma\sqrt{2\ln(2n/\sigma)})$ is such that the quantity

$$N(m) := \lfloor \exp\{m/16\} \rfloor$$

satisfies

$$\frac{1}{2}\ln(N(m) - 1) - 2\ln 2 > m/36; \tag{4.9.60}$$

this assumption clearly is satisfied when m is large enough, that is, it is satisfied when σ is not too large: $\sigma \leq \overline{\sigma}(a^\infty)$, and n is not too small: $n \geq \underline{n}(a^\infty)$, for properly selected $\overline{\sigma}(a^\infty) > 0$ and $\underline{n}(a^\infty)$.

3. *Derive from the above that in the range $\sigma \leq \overline{\sigma}(a^\infty)$, $n \geq \underline{n}(a^\infty)$, the estimate $\widehat{x}_\sigma$ is minimax optimal on $\mathcal{X}$ within a logarithmic factor:*

$$\mathrm{Risk}_{\|\cdot\|_p}[\widehat{x}_\sigma|\mathcal{X}] \leq O(1)\sqrt{2\ln(2n/\sigma)}\mathrm{Risk}^*$$

with absolute constant factor $O(1)$.

4.29.D. General case. What is on our agenda now, is a "presumably good" design of a polyhedral estimate in the general situation described in item 4.29.A. Let us make the same assumption as in item 4.29.B:

> We have at our disposal a set $\mathcal{U}_\mathcal{X}$ matching $\mathcal{X}$ and a set $\mathcal{U}_{\|\cdot\|_*}$ matching the unit ball of the norm $\|\cdot\|_*$ conjugate to the norm $\|\cdot\|$ in which we measure the recovery error.

4.29.D.1 The construction. Observe that every vector $h \in \mathbf{R}^m$ of unit $\|\cdot\|_2$-norm produces an "easy to estimate" linear form on $\mathbf{R}^n$, specifically, the form

$$[\underbrace{A^Th}_{g_h}]^Tx;$$

the simplest "plugin" estimate of this form $\widehat{g}_h(\omega) = h^T\omega$ is unbiased:

$$\mathbf{E}_{\xi\sim\mathcal{N}(0,\sigma^2 I)}\widehat{g}_h(Ax+\xi) = g_h^Tx \ \forall x$$

and satisfies the error bound

$$\mathrm{Prob}_{\xi\sim\mathcal{N}(0,\sigma^2 I)}\{|\widehat{g}_h(x) - g_h^Tx| > \alpha\sigma\} \leq 2\mathrm{Erf}(\alpha) \ \forall (x \in \mathbf{R}^n, \alpha \geq 0).$$

As a result, selecting $N \geq m$ unit vectors $h_i \in \mathbf{R}^m$ and setting

$$\rho = \rho_{N,\epsilon} = \mathrm{ErfInv}\left(\frac{\epsilon}{2N}\right)\sigma,$$

we ensure that

$$\mathrm{Prob}_{\xi\sim\mathcal{N}(0,\sigma^2 I)}\left\{\xi : |h_i^T(Ax+\xi) - g_{h_i}^Tx| \leq \rho, 1 \leq i \leq N, \forall x \in \mathbf{R}^n\right\} \geq 1-\epsilon. \tag{4.9.61}$$

Same as in item 4.29.A, the vectors $h_1, ..., h_N$ give rise to the polyhedral estimate $\widehat{x}(\omega)$ of Bx of a signal $x \in \mathcal{X}$ via observation $\omega = Ax + \xi$, specifically, the estimate $\widehat{x}(\omega) = B\bar{x}(\omega)$, where $\bar{x}(\omega)$ is an optimal solution to the convex optimization problem

$$\min_{y\in\mathcal{X}} \|[g_{h_1}^T; ...; g_{h_N}^T]y - [h_1^T; ...; h_N^T]\omega\|_\infty \left[\equiv \|H^T(Ay-\omega)\|_\infty, \ H = [h_1, ..., h_N]\right]. \tag{4.9.62}$$

4.29.D.2 Bounding risk. We are about to bound from above the risk $\mathrm{Risk}_{\epsilon,\|\cdot\|}$ of the just defined estimate. The bound is similar, but not identical, to the bound we have built in item 4.29.B.

Observe that $\mathcal{U}_\mathcal{X}$ gives rise to the closed convex cone

$$\mathbf{U}_*[\mathcal{U}_\mathcal{X}] = \{(U,\mu) \in \mathbf{S}^n_+ \times \mathbf{R} : \mathrm{Tr}(VU) \leq \mu \forall V \in \mathcal{U}_\mathcal{X}\}$$

which is computationally tractable along with $\mathcal{U}_\mathcal{X}$. Note that since $xx^T \in \mathcal{U}_\mathcal{X}$ whenever $x \in \mathcal{X}$, we have

$$x^TUx = \mathrm{Tr}(U[xx^T]) \leq \mu \ \forall (U,\mu) \in \mathbf{U}_*[\mathcal{U}_\mathcal{X}]$$

(in words: the U-component of a pair $(U,\mu) \in \mathcal{B}_*[\mathcal{U}_\mathcal{X}]$ is a positive semidefinite matrix such that the quadratic form x^TUx is upper-bounded by μ everywhere on $\mathcal{X}$).

Now goes the exercise:

4.1. Let $p(\cdot)$ be an absolute norm on $\mathbf{R}^n$ with the unit ball containing $\mathcal{X}$, and let $r(\cdot)$ be a norm on $\mathbf{R}^n$ such that whenever $p(u) \leq 1$, we have also $r([u]^2) \leq 1$, where $[u]^2 = [u_1^2; ...; u_n^2]$. In this situation the set

$$\mathcal{U}_{p,r} = \left\{U \in \mathbf{S}^n_+ : \|U\|_{p^+} \leq 1, r(\mathrm{diag}(U)) \leq 1\right\} \qquad [\mathrm{diag}(U) = [U_{11}; ...; U_{nn}]]$$

matches $\mathcal{X}$ (for notation and justification, see Exercise 4.11). Prove that

$$\mathbf{U}_*[\mathcal{U}_{p,r}] = \left\{(U, \mu) \in \mathbf{S}^n_+ \times \mathbf{R} : \exists V \in \mathbf{S}^n, v \in \mathbf{R}^n : U \preceq V + \mathrm{Diag}\{v\}, \|V\|_{p^+,*} + r_*(v) \leq \mu\right\},$$

where $\|\cdot\|_{p^+,*}$ is the norm on $\mathbf{S}^n$ conjugate to $\|\cdot\|_{p^+}$, and $r_*(\cdot)$ is the norm conjugate to $r(\cdot)$.

4.2. Verify that in the situation of item 4.29.D.1 for every $x, y \in \mathcal{X}$ and every $(U, \mu) \in \mathbf{U}_*[\mathcal{U}_\mathcal{X}]$ one has

$$(x - y)^T U (x - y) \leq 4\mu.$$

4.3. Verify the following

Lemma 4.9.6 *In the situation of item 4.29.D, consider the convex optimization problem*

$$\mathrm{Opt}[h_1, ..., h_N] = \min_{\lambda, U, \mu, V, \nu} \left\{ 4\rho^2 \sum_{i=1}^N \lambda_i + 4\mu + \nu : \begin{array}{c} \lambda \geq 0, (U, \mu) \in \mathbf{U}_*[\mathcal{U}_\mathcal{X}], (V, \nu) \in \mathbf{U}_*[\mathcal{U}_{\|\cdot\|_*}] \\ \left[\begin{array}{c|c} V & \frac{1}{2}B \\ \hline \frac{1}{2}B^T & A^T[\sum_{i=1}^N \lambda_i h_i h_i^T]A + U \end{array}\right] \succeq 0 \end{array} \right\}. \tag{4.9.63}$$

The optimal value of the problem is an upper bound on $\mathrm{Risk}_{\epsilon,\|\cdot\|}[\widehat{x}|\mathcal{X}]$*, where $\widehat{x}(\cdot)$ is the polyhedral estimate associated with $h_1, ..., h_N$ via the construction from item 4.29.D.1.*

4.29.D.3 Bottom line. Observe that the matrix $\sum_{i=1}^n \lambda_i h_i h_i^T$ participating in (4.9.63) is positive semidefinite with trace $\sum_i \lambda_i$ (recall that h_i are unit vectors). It follows that the optimal value in the convex optimization problem

$$\mathrm{Opt} = \min_{G, U, \mu, V, \nu} \left\{ 4\rho^2 \mathrm{Tr}(G) + 4\mu + \nu : \begin{array}{c} G \succeq 0, (U, \mu) \in \mathbf{U}_*[\mathcal{U}_\mathcal{X}], (V, \nu) \in \mathbf{U}_*[\mathcal{U}_{\|\cdot\|_*}] \\ \left[\begin{array}{c|c} V & \frac{1}{2}B \\ \hline \frac{1}{2}B^T & A^T G A + U \end{array}\right] \succeq 0 \end{array} \right\} \tag{4.9.64}$$

is $\leq \mathrm{Opt}[h_1, ..., h_N]$ *for all collections of unit vectors $h_1, ..., h_N$*. On the other hand, the G-component of a feasible solution to (4.9.64) always can be represented as $\sum_i \lambda_i h_i h_i^T$ with unit h_i and nonnegative λ_i such that $\sum_i \lambda_i = \mathrm{Tr}(G)$ (look at the eigenvalue decomposition of G and bear in mind that $N \geq m$). It follows that *we can efficiently optimize the polyhedral estimate in question in* $h_1, ..., h_N$, or, more exactly, we can optimize in $h_1, ..., h_N$ the natural under the circumstances upper bound $\mathrm{Opt}[h_1, ..., h_N]$ on the risk $\mathrm{Risk}_{\epsilon,\|\cdot\|}[\widehat{x}|\mathcal{X}]$ of this estimate – to this end, it suffices to find (near)-optimal solution to (4.9.64) and take, as $h_1, ..., h_N$, the normalized eigenvectors of the G-component of this solution (if $N > m$, we could augment these eigenvectors with $N - m$ arbitrary unit vectors). As a byproduct of our resoning, we conclude that with the proposed synthesis scheme, there is no need to consider the values of N which are greater than m.

4.29.D.4 "Globalization." Assume that $\mathcal{X}$ is the ball $\{x \in \mathbf{R}^n : q(x) \leq R\}$ associated with a norm $q(\cdot)$. On a straightforward inspection, our upper risk bounds remain intact when we pass from the polyhedral estimate built in item 4.29.D.1 to the estimate $\widehat{x}_{H,q}(\omega) = B\widetilde{x}_{H,q}(\omega)$, where $\widetilde{x}_{H,q}(\omega)$ is an optimal solution to the convex optimization problem

$$\min_y \left\{q(y) : \|H(Ay - \omega)\|_\infty \leq \rho\right\}, \quad H = [h_1, ..., h_N]$$

($\widehat{x}_{H,q}(\omega)$ is undefined when the latter problem is infeasible). When $q(\cdot) = \|\cdot\|$, *we arrive at6 regular ℓ_1 recovery*, see Section 1.2.3.

A common shortcoming of all estimation schemes developed in Lecture 4, is its sensitivity to the a priori given signal set $\mathcal{X}$; note, however, that *for H fixed*, the estimate we are speaking now depends solely on the norm $q(\cdot)$, but not on $\mathcal{X}$. This observation suggests the following course of actions:

> We specify $H = [h_1, ..., h_m]$ by selecting $h_1, ..., h_m$ as the orthonormal collection of eigenvectors of the G-component of an optimal solution to the convex optimization problem (4.9.64) associated with $N = m$ and $\mathcal{X} = \bar{\mathcal{X}} := \{x : q(x) \leq \bar{R}\}$ with somehow selected $\bar{R} > 0$, and use the estimate $\widehat{x}_{H,q}(\omega)$, with no care of whether the signal underlying our observation belongs or does not belong to $\mathcal{X}$.

Now goes the exercise:

5. *Prove that whatever be the signal x underlying our observation, one has*

$$\mathrm{Prob}_{\xi\sim\mathcal{N}(0,\sigma^2 I)}\{\xi : \|Bx - \widehat{x}_{H,q}(Ax+\xi)\| > \mathrm{Risk}(x)\} \leq \epsilon,\ \mathrm{Risk}(x) = \left\{\begin{array}{ll} \mathrm{Opt}, & q(x) \leq \bar{R} \\ \frac{q(x)}{\bar{R}}\mathrm{Opt}, & q(x) \geq \bar{R} \end{array}\right.,$$

where Opt *is the optimal value in the problem* (4.9.64) *giving rise to H (let us call this problem "nominal;" it is immediately seen that the nominal problem is solvable, so that H is well defined).*

4.29.D.5 Spectratopic case. Now consider the case where $\mathcal{X}$ and the unit ball of the norm $\|\cdot\|_*$ are spectratopes (as always, we can assume w.l.o.g. that $\mathcal{X}$ is a basic spectratope):

$$\begin{array}{rcl} \mathcal{X} & = & \{x \in \mathbf{R}^n : \exists t \in \mathcal{T} : \exists t \in \mathcal{T} : R_k^2[x] \preceq t_k I_{d_k}, k \leq K\}, \\ \mathcal{B}_* & := & \{v \in \mathbf{R}^\nu : \|v\|_* \leq 1\} = \{v \in \mathbf{R}^\nu : \exists (u \in \mathbf{R}^N, r \in \mathcal{R}) : v = Mu,\ S_\ell^2[u] \preceq r_\ell I_{f_\ell}, \ell \leq L\}, \end{array}$$

with our usual restrictions on $\mathcal{T}, \mathcal{R}, R_k[\cdot], S_\ell[\cdot]$. Observe that it is easy to equip a spectratope

$$\mathcal{Y} = \{y \in \mathbf{R}^N : \exists (z \in \mathbf{R}^L, v \in \mathcal{V}) : y = Qz, F_j^2[z] \preceq v_j I_{p_j}, j \leq J\}$$

with a computationally tractable set $\mathcal{U}_{\mathcal{Y}}$ matching $\mathcal{Y}$ – it suffices to set

$$\begin{array}{c} \mathcal{U}_{\mathcal{Y}} = \{Y \in \mathbf{S}^N_+ : \exists (Z \in \mathbf{S}^L_+, v \in \mathcal{V}) : Y = QZQ^T,\ \mathcal{F}_j[Z] \preceq v_j I_{p_j}, j \leq J\} \\ \left[F_j[z] = \sum_s z_s F^{js} \Rightarrow \mathcal{F}_j[Z] = \sum_{s,s'} Z_{ss'} F^{js} F^{js'}\right] \end{array}$$

It is immediately seen (check it!) that

$$\begin{array}{c} \mathbf{U}_*[\mathcal{U}_{\mathcal{Y}}] \supset \widehat{\mathbf{U}}_*[\mathcal{U}_{\mathcal{Y}}], \\ \widehat{\mathbf{U}}_*[\mathcal{U}_{\mathcal{Y}}] := \left\{(U, \mu) \in \mathbf{S}^N_+ \times \mathbf{R} : \exists \Xi = \{\Xi_j \in \mathbf{S}^{p_j}_+, j \leq J\} : Q^T U Q \preceq \sum_j \mathcal{F}_j^*[\Xi_j], \phi_{\mathcal{V}}(\lambda[\Xi]) \leq \mu\right\}\} \\ \left[\Lambda[\Xi] = [Tr(\Xi_1); ...; \mathrm{Tr}(\Xi_J)],\ \phi_{\mathcal{V}}(\lambda) = \max_{v\in\mathcal{V}} \lambda^T v\right] \end{array} \tag{4.9.65}$$

As a result, in the situation of this item we can equip the spectratopes $\mathcal{X}$ and $\mathcal{B}_*$ with the just described sets $\mathcal{U}_{\mathcal{X}}$, $\mathcal{U}_{\|\cdot\|_*} := \mathcal{U}_{\mathcal{B}_*}$, giving rise to the "optimized polyhedral estimate" yielded by the construction of item 4.29.D.3; the ϵ-risk $\mathrm{Risk}_{\epsilon,\|\cdot\|}$ of this estimate is upper-bounded by the optimal value Opt in the optimization problem (4.9.64) associated with the just defined $\mathcal{U}_{\mathcal{X}}$ and $\mathcal{U}_{\|\cdot\|_*}$. Let us set

$$\mathrm{Opt}_\# = \min_{\Theta,U,\mu,V,\nu}\left\{\sigma^2\mathrm{Tr}(\Theta) + \mu + \nu : \begin{array}{l} \Theta \succeq 0, (U,\mu) \in \mathbf{U}_*[\mathcal{U}_{\mathcal{X}}], (V,\nu) \in \mathbf{U}_*[\mathcal{U}_{\|\cdot\|_*}] \\ \left[\begin{array}{c|c} V & \frac{1}{2}B \\ \hline \frac{1}{2}B^T & A^T\Theta A + U \end{array}\right] \succeq 0 \end{array}\right\} \tag{4.9.66}$$

and consider the problem (4.5.15) responsible under the circumstances for the near-optimal linear estimate (see Proposition 4.5.1):

$$\begin{array}{rcl}\mathrm{Opt}_{\mathrm{lin}} & = & \min\limits_{H,\Lambda,\Upsilon,\Upsilon',\Theta}\left\{\phi_{\mathcal{T}}(\lambda[\Lambda])+\phi_{\mathcal{R}}(\lambda[\Upsilon])+\phi_{\mathcal{R}}(\lambda[\Upsilon'])+\sigma^2\mathrm{Tr}(\Theta):\right.\\ && \Lambda=\{\Lambda_k\succeq 0,k\le K\},\ \Upsilon=\{\Upsilon_\ell\succeq 0,\ell\le L\},\ \Upsilon'=\{\Upsilon'_\ell\succeq 0,\ell\le L\},\\ && \left[\begin{array}{c|c}\sum_k\mathcal{R}_k^*[\Lambda_k] & \frac{1}{2}[B^T-A^TH]M\\ \hline \frac{1}{2}M^T[B-H^TA] & \sum_\ell\mathcal{S}_\ell^*[\Upsilon_\ell]\end{array}\right]\succeq 0,\\ && \left.\left[\begin{array}{c|c}\Theta & \frac{1}{2}HM\\ \hline \frac{1}{2}M^TH^T & \sum_\ell\mathcal{S}_\ell^*[\Upsilon'_\ell]\end{array}\right]\succeq 0\right\}\end{array}\qquad (4.9.67)$$

Now comes a big surprize:

6.1. Let $\varkappa = 2\max[\mathrm{ErfInv}(\frac{\epsilon}{2m}),1]$. *Prove that the optimal value* Opt *in the problem (4.9.64) associated with our* $\mathcal{U}_{\mathcal{Y}}$ *and* $\mathcal{U}_{\|\cdot\|_*}$ *is within factor* $\varkappa$ *of* $\mathrm{Opt}_\#$*:*

$$\mathrm{Opt}\le\varkappa\mathrm{Opt}_\#.$$

6.2. Prove that $\mathrm{Opt}_\#\le\mathrm{Opt}_{\mathrm{lin}}$.

In other words, our optimized polyhedral estimate is nearly at least as good as the near-optimal linear estimate yielded by Proposition 4.5.1: ϵ-risk of the former estimate is within logarithmic in m,ϵ factor $\varkappa$ of the worst case, w.r.t. $x\in\mathcal{X}$, expected $\|\cdot\|$-error of the latter estimate.

7. Prove that if ξ is a sub-Gaussian random vector with sub-Gaussianity parameters $(0,\sigma^2I_m)$ and $h\in\mathbf{R}^m$ is a $\|\cdot\|_2$-unit vector, then

$$\forall r\ge 0:\mathrm{Prob}\{|h^T\xi|>r\}\le\exp\{-\frac{r^2}{2\sigma^2}\}.$$

Derive from this observation that all preceding results in this Exercise remain valid in the case when observation noise is $(0,\sigma^2I_m)$-sub-Gaussian rather than $\mathcal{N}(0,\sigma^2I_m)$, provided that ErfInv(...) *in the constructions and claims is replaced with* $\sqrt{2\ln(\frac{1}{...})}$.

Exercise 4.30 † [polyhedral estimation of distributions] The goal of this Exercise is to adjust the polyhedral estimate from Exercise 4.29 to the situation as follows. Given are

- Convex compact set $\mathcal{X}$ in $\mathbf{R}^n$ along with a computationally tractable closed convex cone $\mathbf{U}_*^{\mathcal{X}}\subset\mathbf{S}_+^n\times\mathbf{R}_+$ such that whenever $(U,\mu)\in\mathbf{U}_*^{\mathcal{X}}$, we have

  $$x^TUx\le\mu\ \forall x\in\mathcal{X}$$

  and, in addition, $\mathbf{U}_*$ contains a pair (U,μ) with $U\succ 0$;

- $m\times n$ sensing matrix A such that $A\mathcal{X}\in\mathbf{\Delta}_m$ (i.e., for every $x\in\mathcal{X}$, Ax is a probabilistic m-dimensional vector), and $\nu\times n$ matrix B

- a norm $\|\cdot\|$ on $\mathbf{R}^\nu$ along with spectratopic representation of the unit ball of the conjugate norm:

  $$\mathcal{B}_*:=\{v\in\mathbf{R}^\nu:\|v\|_*\le 1\}=\{v\in\mathbf{R}^\nu:\exists(u\in\mathbf{R}^N,r\in\mathcal{R}):v=Mu,\ S_\ell^2[u]\preceq r_\ell I_{f_\ell},\ell\le L\}.$$

  This representation gives rise to the convex cone

  $$\mathbf{V}_*=\{(V,\theta)\in\mathbf{S}_+^\nu\times\mathbf{R}^+:\exists\Upsilon=\{\Upsilon_\ell\in\mathbf{S}_+^{f_\ell},\ell\le L\}:V\preceq\sum_\ell\mathcal{S}_\ell^*[\Upsilon_\ell],\phi_{\mathcal{R}}(\lambda[\Upsilon])\le\theta\}$$

  such that

  $$(V,\theta)\in\mathbf{V}_*\Rightarrow u^TVu\le\theta\ \forall(u:\exists r\in\mathcal{R}:S_\ell^2[u]\preceq r_\ell I_{f_\ell},\ell\le L).$$

What we observe, are K independent of each other realizations of a discrete random variable with probability distribution Ax, where x is unknown signal known to belong to $\mathcal{X}$, and our goal is to recover Bx from these observations.

As always in similar cases, we encode observations stemming from a signal $x \in \mathcal{X}$ as independent across $k \leq K$ realizations $\xi^1, ..., \xi^K$ of random variable ξ_x taking values e_i, $1 \leq i \leq m$ (e_i are the standard basic orths in $\mathbf{R}^m$) with probabilities $[Ax]_i$. We set

$$\omega^K = \frac{1}{K}\sum_{k=1}^{K} \xi^k$$

and denote by P_x the distribution of ω^K.

4.30.A Preliminaries. Prove the following lemmas:

1.1. **Lemma 4.9.7** *Let $h \in \mathbf{R}^n$, and let ξ be random variable taking values $e_1, ..., e_n$ with probabilities $p_1, ..., p_n$ (e_i are the standard basic orths). Let, further, $\eta = \xi - \mathbf{E}\{\xi\} = \xi - p$ and $\eta^t = \xi^t - p$, where ξ^t are independent across t copies of ξ. Then*

$$0 \leq r \leq 2 \Rightarrow \mathrm{Prob}\{|\frac{1}{N}h^T\sum_{t=1}^{N}\eta^t| > r\|h\|_\infty\} \leq 2\exp\{-\frac{r^2 N}{4}\}.$$

1.2. **Lemma 4.9.8** *Let Q be an $m \times m$ matrix such that $Q \succeq 0$ and $Q_{ii} \leq 1$, $1 \leq i \leq M$. Then we can find reliably a decomposition*

$$Q = GG^T$$

with $m \times \bar{m}$ matrix G, where $\bar{m} = 2^k$ with the smallest integer k such that $2^k \geq m$, such that

$$\|G\|_\infty := \max_{i,j}|G_{ij}| \leq \kappa/\sqrt{\bar{m}},\ \kappa = 2\sqrt{\ln(2\bar{m})}.$$

Hint: *Reduce the situation to the one where $\bar{m} = m$, set $R = Q^{1/2}$ and look at the magnitudes of entries in the random matrix*

$$G_\xi = R\mathrm{Diag}\{\xi\}H,$$

where H is the orthonormal scaling of $\bar{m}\times\bar{m}$ Hadamard matrix, and ξ is Rademacher random vector.

4.30.B The polyhedral estimate we intend to use is as follows: given reliability level $1-\epsilon$, $\epsilon \in (0,1)$, we fix J vectors $h^1, ..., h^J$ from $\mathbf{R}^m$ and specify a positive real ρ such that

$$\forall x \in \mathcal{X} : \mathrm{Prob}_{\omega^K\sim P_x}\left\{|[\omega^K - Ax]^T h^j| \leq \rho, 1 \leq j \leq J\right\} \geq 1-\epsilon.$$

Given observations, we compute the associated ω^K and solve the optimization problem

$$\min_{y\in\mathcal{X}} \|H^T[Ay - \omega^K]\|_\infty \qquad [H = [h^1, ..., h^J]] (P[\omega^K])$$

An optimal solution $\overline{x}(\omega^K)$ to the problem gives rise to the estimate $\widehat{x}(\omega^K) = B\overline{x}(\omega^K)$ of Bx.

Prove the following results:

2.1. **Proposition 4.9.11** *In the situation in question, let*

$$\mathfrak{R} = \max_z\left\{\|Bz\| : \|H^T Az\|_\infty \leq 2\rho, z \in \mathcal{X} - \mathcal{X}\right\},$$

The quantity $\mathfrak{R}$ is an upper bound on the ϵ-risk of the polyhedral estimate we have just built:

$$\mathfrak{R} \geq \mathrm{Risk}_{\epsilon,\|\cdot\|}[\widehat{x}|\mathcal{X}] := \inf\left\{r : \mathrm{Prob}_{\omega^k\sim P_x}\{\|Bx - \widehat{x}(\omega^K)\| > r\} \leq \epsilon\,\forall x \in \mathcal{X}\right\}.$$

2.2. **Proposition 4.9.12** *In the situation described in the beginning of Exercise and given $\epsilon \in (0,1)$, let us set*

$$\bar{m} = 2^{\mathrm{Ceil}(\log_2(m))},\ \varkappa = \frac{2\sqrt{\ln(2\bar{m})}}{\sqrt{\bar{m}}},\ \varrho_{K,m}(\epsilon) = \frac{2\sqrt{\ln(2\bar{m}/\epsilon)}}{\sqrt{K}},$$
$$\rho_{K,m}(\epsilon) = \varkappa\varrho_{K,m}(\epsilon) = \frac{4\sqrt{\ln(2\bar{m}/\epsilon)\ln(2\bar{m})}}{\sqrt{K\bar{m}}},\ \vartheta_{K,m}(\epsilon) = \sqrt{\bar{m}}\rho_{K,m}(\epsilon) = \frac{4\sqrt{\ln(2\bar{m}/\epsilon)\ln(2\bar{m})}}{\sqrt{K}},$$

and let ϵ, K, m be such that $\varrho_{K,m}(\epsilon) \leq 2$. Consider the convex optimization problem

$$\mathrm{Opt} = \min_{\Theta,(U,\mu),(V,\theta)} \left\{ 4\vartheta^2_{K,M}(\epsilon)\chi(\Theta) + 4\mu + \theta : \begin{array}{c} \Theta \succeq 0, (U,\mu) \in \mathbf{U}_*, (V,\theta) \in \mathbf{V}_* \\ \left[\begin{array}{c|c} V & \frac{1}{2}M^TB \\ \hline \frac{1}{2}B^TM & A^T\Theta A + U \end{array}\right] \succeq 0 \end{array} \right\},$$
$$\chi(\Theta) = \max_i \Theta_{ii} \tag{4.9.68}$$

A feasible solution $(\Theta, (U,\mu), (V,\theta))$ to this problem gives rise to $m \times \bar{m}$ matrix H such that

$$\Theta = \chi(\Theta)HH^T \ \&\ \|H\|_\infty \leq \varkappa \tag{4.9.69}$$

(Lemma 4.9.8). Then for every $x \in \mathcal{X}$ it holds

$$\mathrm{Prob}_{\omega^K \sim P_x}\left\{\|H^T[\omega^K - Ax]\|_\infty \leq \rho_{K,m}(\epsilon)\right\} \geq 1 - \epsilon, \tag{4.9.70}$$

and the pair $(H, \rho := \rho_{K,m}(\epsilon))$ gives rise to a polyhedral estimate $\widehat{x}(\omega^K)$ satisfying

$$\mathrm{Risk}_{\epsilon,\|\cdot\|}[\widehat{x}|\mathcal{X}] \leq \mathfrak{R} := \max_z\{\|Bz\| : \|H^TAz\|_\infty \leq 2\rho, z \in \mathcal{X} - \mathcal{X}\} \leq 4\vartheta^2_{K,M}(\epsilon)\chi(\Theta) + 4\mu + \theta.$$

‘

Bibliography

[1] E.D. Andersen, K.D. Andersen. *The MOSEK optimization toolbox for MATLAB manual. Version 7.0*, 2013. `http://docs.mosek.com/7.0/toolbox/`.

[2] K.E. Andersen, M.B. Hansen. Multiplicative censoring: density estimation by a series expansion approach. *Journal of Statistical Planning and Inference*, 98(1-2):137–155, 2001.

[3] T.W. Anderson. The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities. *Proceedings of the American Mathematical Society*, 6(2):170–176, 1955.

[4] A. Antoniadis, I. Gijbels. Detecting abrupt changes by wavelet methods. *Journal of Nonparametric Statistics*, 14(1-2):7–29, 2002.

[5] B.F. Arnold, P. Stahlecker. Another view of the Kuks-Olman estimator. *Journal of Statistical Planning and Inference* **89** (2000), 169-174

[6] T. Augustin, R. Hable. On the impact of robust statistics on imprecise probability models: a review. *Structural Safety*, 32(6):358–365, 2010.

[7] R. Bakeman, J.M. Gottman. *Observing Interaction: An Introduction to Sequential Analysis.* Cambridge University Press, 1997.

[8] G.A. Barnard. Sequential tests in industrial statistics. *Supplement to the Journal of the Royal Statistical Society*, pages 1–26, 1946.

[9] M. Basseville. Detecting changes in signals and systems – a survey. *Automatica*, 24(3):309–326, 1988.

[10] M. Basseville, I.V. Nikiforov. *Detection of Abrupt Changes: Theory and Application.* Prentice-Hall, Englewood Cliffs, N.J., 1993.

[11] T. Bednarski. Binary experiments, minimax tests and 2-alternating capacities. *The Annals of Statistics*, 10(1):226–232, 1982.

[12] D. Belomestny, A. Goldenschluger. Nonparametric density estimation from observations with multiplicative measurement errors. *arXiv preprint arXiv:1709.00629*, 2017.

[13] A. Ben-Tal, A. Nemirovski. *Lectures on modern convex optimization: analysis, algorithms, and engineering applications*, SIAM Series in Optimization volume 2, SIAM, 2001.

[14] A. Ben-Tal, A. Nemirovski. *Lectures on modern convex optimization.* Lecture Notes, `http://www2.isye.gatech.edu/~nemirovs/Lect_ModConvOpt.pdf`

[15] A. Ben-Tal, L. El Ghaoui, A. Nemirovski. *Robust Optimization*, Princeton University Press, 2009.

[16] M. Bertero, P. Boccacci. Image restoration methods for the Large Binocular Telescope (LBT). *Astronomy and Astrophysics Supplementary Series*, 147(2):323333, 2000.

[17] M. Bertero, P. Boccacci. Application of the OS-EM method to the restoration of LBT images. *Astronomy and Astrophysics Supplementary Series*, 144(1):181-186, 2000.

[18] E. Betzig, G.H. Patterson, R. Sougrat, O.W. Lindwasser, S. Olenych, J.S. Bonifacino, M.W. Davidson, J. Lippincott-Schwartz, and H.F. Hess. Imaging intracellular fluorescent proteins at nanometer resolution. *Science*, 313(5793):1642-1645, 2006.

[19] L. Birgé. *Approximation dans les spaces métriques et théorie de l'estimation: inégalités de Cràmer-Chernoff et théorie asymptotique des tests.* PhD thesis, Université Paris VII, 1980.

[20] L. Birgé. Vitesses maximales de décroissance des erreurs et tests optimaux associés. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 55(3):261–273, 1981.

[21] L. Birgé. Sur un théorème de minimax et son application aux tests. *Probab. Math. Stat.*, 3:259–282, 1982.

[22] L. Birgé. Approximation dans les espaces métriques et théorie de l'estimation. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 65(2):181–237, 1983.

[23] L. Birgé. Robust testing for independent non identically distributed variables and Markov chains. In *Specifying Statistical Models*, pages 134–162. Springer, 1983.

[24] L. Birgé. Model selection via testing: an alternative to (penalized) maximum likelihood estimators. In *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, volume 42, pages 273–325. Elsevier, 2006.

[25] L. Birgé. Robust tests for model selection. In M. Banerjee, F. Bunea, J. Huang, V. Koltchinskii, , and M. Maathuis, editors, *From Probability to Statistics and Back: High-Dimensional Models and Processes – A Festschrift in Honor of Jon A. Wellner*, pages 47–64. Institute of Mathematical Statistics, 2013.

[26] S. Boyd, L. El Ghaoui, E. Feron, V. Balakrishnan. *Linear Matrix Inequalities in System and Control Theory.* SIAM, Philadelphia, 1994.

[27] E. Brodsky, B.S. Darkhovsky. *Nonparametric methods in change point problems.* Springer, 1993.

[28] E. Brunel, F. Comte, V. Genon-Catalot. Nonparametric density and survival function estimation in the multiplicative censoring model. *Test*, 25(3):570–590, 2016.

[29] A. Buchholz. Operator Khintchine inequality in the non-commutative probability. *Mathematische Annalen* 391:1–16, 2001.

[30] A. Buja. On the Huber-Strassen theorem. *Probability Theory and Related Fields*, 73(1):149–152, 1986.

[31] M. Burnashev. On the minimax detection of an imperfectly known signal in a white noise background. *Theory Probab. Appl.*, 24:107–119, 1979.

[32] M. Burnashev. Discrimination of hypotheses for gaussian measures and a geometric characterization of the gaussian distribution. *Math. Notes*, 32:757–761, 1982.

[33] T.T. Cai, M.G. Low. A note on nonparametric estimation of linear functionals. *Annals of statistics*, pages 1140–1153, 2003.

[34] T.T. Cai, M.G. Low. Minimax estimation of linear functionals over nonconvex parameter spaces. *The Annals of statistics*, 32(2):552–576, 2004.

[35] T.T. Cai, M.G. Low. On adaptive estimation of linear functionals. *The Annals of Statistics*, 33(5):2311–2343, 2005.

[36] E. Candes, J. Romberg, T. Tao. Signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.* **59:8** (2005), 1207-1223.

[37] E. Candes. Compressive sampling. In: Javier, S., Juan, L. V., Joan, V. (eds). International Congress of Mathematicians, Madrid 2006, vol. III, pp. 14371452. European Mathematical Society Publishing House (2006)

[38] E. Candes, T. Tao. Decoding by linear programming. *IEEE Trans. Inf. Theory* **51** (2006), 4203-4215.

[39] E. Candes, T. Tao. Near-optimal signal recovery from random projections and universal encoding strategies. *IEEE Trans. Inf. Theory* **52** (2006), 5406-5425.

[40] E. Candes, T. Tao. The Dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics* **35** (2007), 2313-2351.

[41] E. Candes. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus de lAcad. Des Sci. Ser. I* **346** (2008), 589-592.

[42] Y. Cao, V. Guigues, A. Juditsky, A. Nemirovski, Y. Xie. Change Detection via Affine and Quadratic Detectors. *Electronic Journal of Statistics* **12:1** (2018), 1-57, arXiv:1608.00524 `https://arxiv.org/pdf/1608.00524.pdf`

[43] J. Chen, A.K. Gupta. *Parametric Statistical Change Point Analysis: With Applications to Genetics, Medicine, and Finance.* Birkhauser, 2012.

[44] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, pages 493–507, 1952.

[45] H. Chernoff. *Sequential Analysis and Optimal Design.* SIAM, 1972.

[46] N. Christopeit, K. Helmes. Linear Minimax Estimation with Ellipsoidal Constraints. *Acta Applicandae Mathematicae* **43** (1996), 3-15.

[47] I. Dattner, A. Goldenshluger, A. Juditsky. On deconvolution of distribution functions. *The Annals of Statistics*, 39(5):2477–2501, 2011.

[48] D. Donoho, R. Liu. Geometrizing rate of convergence, I. Technical report, Tech. Report 137a, Dept. of Statist., University of California, Berkeley, 1987.

[49] D. Donoho, R. Liu, B. MacGibbon. Minimax risk over hyperrectangles, and implications. *The Annals of Mathematical Statistics*, **18**:1416–1437, 1990.

[50] D. Donoho, R. Liu. Geometrizing rates of convergence, II. *The Annals of Statistics* 19 (1991), 633-667.

[51] Donoho, D. “Statistical estimation and optimal recovery” – *The Annals of Statistics* 22:1 (1994), 238–270.

[52] D. Donoho, R. Liu. Geometrizing rates of convergence, III. *The Annals of Statistics*, pages 668–701, 1991.

[53] D. Donoho, M. Low. Renormalization exponents and optimal pointwise rates of convergence. *The Annals of Mathematical Statistics* **20**:944–970, 1992.

[54] D. Donoho. De-noising by soft-thresholding. *IEEE Transactions on Information Theory*, 41(3):613–627, 1995.

[55] D. Donoho, I.M. Johnstone. Minimax estimation via wavelet shrinkage. *The Annals of Statistics*, 26(3):879–921, 1998.

[56] D. Donoho, X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Trans. Inf. Theory* **47:7** (2001), 2845-2862.

[57] D. Donoho. Neighborly polytopes and sparse solutions of underdetermined linear equations. Technical report, Department of Statistics, Stanford University (2004).

[58] D. Donoho, M. Elad, V. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Inf. Theory* **52** (2006), 6-18.

[59] H. Drugas. Spectral Methods in Linear Minimax Estimation. *Acta Applicandae Mathematicae* **43** (1996), 17-42.

[60] S. Efromovich. *Nonparametric curve estimation: methods, theory, and applications.* Springer Science & Business Media, 2008.

[61] S. Efromovich, M. Pinsker. Sharp-optimal and adaptive estimation for heteroscedastic nonparametric regression. *Statistica Sinica*, pages 925–942, 1996.

[62] F. Enikeeva, Z. Harchaoui High-dimensional change-point detection with sparse alternatives. *arXiv:1312.1900*, 2014.

[63] J. Fan. On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics*, 19:3 (1991), 1257-1272.

[64] G. Fellouris, G. Sokolov. Second-order asymptotic optimality in multisensor sequential change detection. arXiv:1410.3815, 2014.

[65] J.J. Fuchs. On sparse representations in arbitrary redundant bases. *IEEE Trans. Inf. Theory* **50** (2004), 1341-1344.

[66] J.J. Fuchs. Recovery of exact sparse representations in the presence of bounded noise. *IEEE Trans. Inf. Theory* **51** (2005), 3601-3608.

[67] W.R. Gaffey. A consistent estimator of a component of a convolution. *The Annals of Mathematical Statistics*, 30(1):198–205, 1959.

[68] N.H. Gholson, RL. Moose. Maneuvering target tracking using adaptive state estimation. *IEEE Transactions on Aerospace and Electronic Systems*, 13(3):310–317, 1977.

[69] B.K. Ghosh. A brief history of sequential analysis. *Handbook of Sequential Analysis*, pages 1–19, 1991.

[70] A. Goldenshluger, A. Juditsky, A. Tsybakov, A. Zeevi. Change–point estimation from indirect observations. 1. minimax complexity. *Ann. Inst. Henri Poincare Probab. Stat.*, 44:787–818, 2008.

[71] A. Goldenshluger, A. Juditsky, A. Tsybakov, A. Zeevi. Change-point estimation from indirect observations. 2. adaptation. *Ann. Inst. H. Poincare Probab. Statist*, 44(5):819–836, 2008.

[72] A. Goldenshluger. A universal procedure for aggregating estimators. *The Annals of Statistics*, pages 542–568, 2009.

[73] A. Goldenshluger, A. Juditsky, A. Nemirovski. Hypothesis testing by convex optimization. *Electron. J. Statist.* 9(2):1645-1712, 2015. arXiv:1311.6765 `https://arxiv.org/pdf/1311.6765.pdf`

[74] A. Goldenshluger, A. Juditsky, A. Nemirovski. Rejoinder of "hypothesis testing by convex optimization". *Electronic Journal of Statistics*, 9(2):1744–1748, 2015.

[75] Y.K. Golubev, B.Y. Levit, A.B. Tsybakov. Asymptotically efficient estimation of analytic functions in gaussian noise. *Bernoulli*, pages 167–181, 1996.

[76] L. Gordon, M. Pollak. An efficient sequential nonparametric scheme for detecting a change of distribution. *The Annals of Statistics*, pages 763–804, 1994.

[77] M. Grant, S. Boyd. *The* `CVX` *Users Guide. Release 2.1*, 2014. `http://web.cvxr.com/cvx/doc/CVX.pdf`.

[78] V. Guigues, A. Juditsky, A. Nemirovski. Hypothesis Testing via Euclidean Separation. `https://arxiv.org/pdf/1705.07196.pdf`

[79] F. Gustafsson. *Adaptive filtering and change detection*, volume 1. Wiley New York, 2000.

[80] S.W. Hell, J. Wichmann. Breaking the diffraction resolution limit by stimulated emission: stimulated-emission-depletion fluorescence microscopy. *Optic Letters* 19(11):780-782, 1994.

[81] S.W. Hell. Toward fluorescence nanoscopy. *Nature Biotechnology*, 21(11):1347-1355, 2003.

[82] S.W. Hell. Microscopy and its focal switch. *Nature Methods*, 6(1):24-32, 2008.

[83] S.T. Hess, T. Girirajan, M.D. Mason. Ultra-high resolution imaging by fluorescence photoactivation localization microscopy. *Biophysical Journal*, 91(11):4258–4272, 2006.

[84] J.-B. Hiriart-Urruty, C. Lemarechal. *Convex Analysis and Minimization Algorithms I: Fundamentals; II: Advanced Theory and Bundle Methods (Grundlehren Der Mathematischen Wissenschaften).* Springer, 1993.

[85] P.J. Huber. A robust version of the probability ratio test. *The Annals of Mathematical Statistics*, 36(6):1753–1758, 1965.

[86] P.J. Huber, V. Strassen. Minimax tests and the Neyman-Pearson lemma for capacities. *The Annals of Statistics*, 1(2):251–263, 1973.

[87] P.J. Huber, V. Strassen. Note: Correction to minimax tests and the Neyman-Pearson lemma for capacities. *The Annals of Statistics*, 2(1):223–224, 1974.

[88] I.A. Ibragimov, R.Z. Khasminskii. *Theory of Statistical Estimation: Asymptotic Theory.* Springer-Verlag, Berlin, New York, 1981.

[89] I.A. Ibragimov, R.Z. Khasminskii. On nonparametric estimation of the value of a linear functional in gaussian white noise. *Theory of Probability & Its Applications*, 29(1):18–32, 1985.

[90] I.A. Ibragimov, R.Z. Khasminskii. Estimation of linear functionals in gaussian noise. *Theory of Probability & Its Applications*, 32(1):30–39, 1988.

[91] Y. Ingster, I.A. Suslina. Nonparametric goodness-of-fit testing under Gaussian models. volume 169 of *Lecture Notes in Statistics*. Springer, 2002.

[92] A. Juditsky, A. Nemirovski. Nonparametric estimation by convex programming. *The Annals of Statistics*, 37(5a):2278–2300, 2009. `https://arxiv.org/pdf/0908.3108.pdf`

[93] A. Juditsky, F. Kilinç Karzan, A. Nemirovski, B. Polyak. Accuracy Guarantees for ℓ_1 recovery of block-sparse signals. *The Annals of Statistics* **40:6** (2012), 3077-3107. `https://arxiv.org/pdf/1111.2546.pdf`

[94] A. Juditsky, A. Nemirovski. On sequential hypotheses testing via convex optimization. *Avtomatika i Telemekhanika* 2015 No. 5, 100-120 (in Russian; English translation *Automation and Remote Control* **76:5** (2015), 809-825. `https://arxiv.org/pdf/1412.1605.pdf`

[95] A. Juditsky, A. Nemirovski. Hypothesis Testing via affine detectors. *Electron. J. Statist.* 10(2):2204-2242, 2016. `https://arxiv.org/pdf/1604.02576.pdf`

[96] A. Juditsky, A. Nemirovski (2016). Estimating linear and quadratic forms via indirect observations. `https://arxiv.org/pdf/1612.01508.pdf`

[97] A. Juditsky, A. Nemirovski (2016). Near-Optimality of Linear Recovery in Gaussian Observation Scheme under $\|\cdot\|_2^2$-Loss. To appear in *The Annals of Statistics*. `https://arxiv.org/pdf/1602.01355.pdf`

[98] A. Juditsky, A. Nemirovski (2017). Near-Optimality of Linear Recovery from Indirect Observations. To appear in *Mathematical Statistics and Learning*. `https://arxiv.org/pdf/1704.00835.pdf`

[99] A. Juditsky, A. Nemirovski (2017). Estimating Linear and Quadratic forms via Indirect Observations. `https://arxiv.org/pdf/1612.01508.pdf`

[100] C. Kraft. Some conditions for consistency and uniform consistency of statistical procedures. *Univ. of California Publ. Statist.*, 2:493–507, 1955.

[101] J. Kuks, W. Olman. Minimax Linear Estimation of Regression Coefficients, I. *Iswestija Akademija Nauk Estonskoj SSR* **20** (1971), 480-482 (in Russian)

[102] J. Kuks, W. Olman. Minimax Linear Estimation of Regression Coefficients, II. *Iswestija Akademija Nauk Estonskoj SSR* **21** (1972), 66-72 (in Russian)

[103] V. Kuznetsov. Stable detection when signal and spectrum of normal noise are inaccurately known. *Telecommunications and radio engineering*, 30(3):58–64, 1976.

[104] T.L. Lai. Sequential changepoint detection in quality control and dynamical systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 613–658, 1995.

[105] T.L. Lai. Information bounds and quick detection of parameter changes in stochastic systems. *IEEE Transactions on Information Theory*, 44(7):2917–2929, 1998.

[106] T.L. Lai. Sequential analysis: some classical problems and new challenges. *Statistica Sinica*, 11(2):303–350, 2001.

[107] A. Lakhina, M. Crovella, C. Diot. Diagnosing network-wide traffic anomalies. In *Proc. of SIGCOMM*, 2004.

[108] L. Le Cam. On the assumptions used to prove asymptotic normality of maximum likelihood estimates. *The Annals of Mathematical Statistics*, pages 802–828, 1970.

[109] L. Le Cam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, pages 38–53, 1973.

[110] L. Le Cam. On local and global properties in the theory of asymptotic normality of experiments. *Stochastic processes and related topics*, 1:13–54, 1975.

[111] L. Le Cam. *Asymptotic Methods in Statistical Decision Theory.* Springer Series in Statistics. Springer, 1986.

[112] K. Liu, R. Zhang, Y. Mei. Scalable sum-shrinkage schemes for distributed monitoring large-scale data streams. `https://arxiv.org/pdf/1603.08652.pdf`

[113] G. Lorden. Procedures for reacting to a change in distribution. *Annals of Mathematical Statistics*, 42(6):1897–1908, Dec. 1971.

[114] F. Lust-Piquard. Inégalités de Khintchine dans C_p $(1 < p < \infty)$. *Comptes Rendus de l'Académie des Sciences de Paris, Série I* 393(7):289–292, 1986.

[115] L. Mackey, M. Jordan, R. Chen, B. Farrell, J. Tropp. Matrix concentration inequalities via the method of exchangeable pairs. *The Annals of Probability*, 42(3):906–945, 2014.

[116] E. Mazor, A. Averbuch, Y. Bar-Shalom, J. Dayan. Interacting multiple model methods in target tracking: a survey. *IEEE Trans. on Aerospace and Electronic Systems*, 34(1):103–123, 1998.

[117] Y. Mei. Asymptotic optimality theory for decentralized sequential hypothesis testing in sensor networks. *IEEE Trans. Inf. Theory*, 54(5):2072–2089, 2008.

[118] A. Meister. *Deconvolution problems in nonparametric statistics*, volume 193. Springer, 2009.

[119] C. Micchelli, T. Rivlin. A survey of optimal recovery. In *Optimal estimation in approximation theory*, pages 1–54. Springer, 1977.

[120] C. Micchelli, T. Rivlin. *Lectures on optimal recovery.* In *Numerical analysis, Lancaster 1984, Lecture Notes in Mathematics*, volume 1129, pages 21–93. Springer, 1985.

[121] G.V. Moustakides. Optimal stopping times for detecting changes in distributions. *Ann. Statist.*, 14:1379–1387, 1986.

[122] H.-G. Müller, U. Stadtmüller. Discontinuous versus smooth regression. *The Annals of Statistics*, 27(1):299–337, 1999.

[123] A. Nemirovskii. On nonparametric estimation of smooth regression functions. (in Russian) *Izvestia AN SSSR, Ser. Tekhnicheskaya Kibernetika*, 1985, No. 3 (the journal is translated into English as *Engineering Cybernetics. Soviet J. Computer & Systems Sci.*)

[124] A. Nemirovski, C. Roos, T. Terlaky. On maximization of quadratic form over intersection of ellipsoids with common center. *Mathematical Programming* **86** (1999), 463-473.

[125] A. Nemirovski, *Topics in Non-parametric Statistics*, in: M. Emery, A. Nemirovski, D. Voiculescu, Lectures on Probability Theory and Statistics, Ecole d'Eteé de Probabilités de Saint-Flour XXVII – 1998, Ed. P. Bernard. - Lecture Notes in Mathematics v. 1738, Springer (2000), 87–285. `http://www2.isye.gatech.edu/~nemirovs/snf00.pdf`

[126] A. Nemirovski, A. Juditsky, G. Lan, A. Shapiro. Stochastic Approximation Approach to Stochastic Programming. *SIAM J. on Optimization*, 19(4):1574–1609, 2009.

[127] A. Nemirovski. *Introduction to Linear Optimization*, Lecture Notes. https://www2.isye.gatech.edu/~nemirovs/OPTI_LectureNotes2015.pdf

[128] A. Nemirovski. *Interior Point polynomial time methods in Convex Programming*, Lecture Notes. https://www.isye.gatech.edu/~nemirovs/Lect_IPM.pdf

[129] A. Nemirovski. Sums of random symmetric matrices and quadratic optimization under orthogonality constraints. *Mathematical programming*, 109(2):283–317, 2007.

[130] Yu. Nesterov, A. Nemirovski. *Interior Point Polynomial Algorithms in Convex Programming.* SIAM, 1994.

[131] M.H. Neumann. Optimal change-point estimation in inverse problems. *Scandinavian Journal of Statistics*, 24(4):503–521, 1997.

[132] F. Österreicher. On the construction of least favourable pairs of distributions. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 43(1):49–55, 1978.

[133] E.S. Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, June 1954.

[134] J. Pilz. Minimax linear regression estimation with symmetric parameter restrictions. *J. of Statistical Planning and Inference* **13** (1986), 297-318.

[135] M.S. Pinsker. Optimal filtration of square-integrable signals in Gaussian noise. *Problems of Information Transmission* **16:2** (1980), 52-68 (in Russian).

[136] G. Pisier. Non-commutative vector valued L_p spaces and completely p-summing maps. - Astérisque 247, 1998.

[137] M. Pollak. Optimal detection of a change in distribution. *Ann. Statist.*, 13:206–227, 1985.

[138] M. Pollak. Average run lengths of an optimal method of detecting a change in distribution. *Ann. Statist.*, 1987.

[139] V.H. Poor, O. Hadjiliadis. *Quickest Detection.* Cambridge University Press, 2009.

[140] C.R. Rao. *Linear Statistical Inference and its Applications*, Wiley, New York, 1973.

[141] C.R. Rao. Estimation of parameters in a linear model. *The Annals of Statistics* **4:6** (1976), 1023-1037.

[142] H. Rieder. Least favorable pairs for special capacities. *The Annals of Statistics*, pages 909–921, 1977.

[143] M.J. Rust, M. Bates, X. Zhuang. Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). *Nature Methods*, 3(10):793–796, 2006.

[144] R. Saigal, H. Wolkowitcz, L. Vandenberghe, Eds. *Handbook on Semidefinite Programming*, Kluwer Academic Publishers, 2000.

[145] A.W. Shewhart. Economic control of quality of manufactured product. *Preprinted by ASQC quality press*, 1931.

[146] W.A. Shiryaev. On optimal methods in quickest detection problems. *Theory Prob. Appl.*, 8:22–46, Jan. 1963.

[147] D. Siegmund. *Sequential Analysis: Tests and Confidence Intervals.* Springer Series in Statistics. Springer, Aug. 1985.

[148] D. Siegmund, B. Yakir. *The Statistics of Gene Mapping.* Springer, 2007.

[149] A.G. Tartakovsky, V.V. Veeravalli. Change-point detection in multichannel and distributed systems. *Applied Sequential Methodologies: Real-World Examples with Data Analysis*, 173:339–370, 2004.

[150] A.G. Tartakovsky, V.V. Veeravalli. Asymptotically optimal quickest change detection in distributed sensor systems. *Sequential Analysis*, 27(4):441–475, 2008.

[151] A. Tartakovsky, I. Nikiforov, M. Basseville. *Sequential Analysis: Hypothesis Testing and Changepoint Detection.* Chapman and Hall/CRC, 2014.

[152] A.B. Tsybakov. *Introduction to Nonparametric Estimation* Springer Series in Statistics, Springer 2008.

[153] J.A. Tropp. The random paving property for uniformly bounded matrices. *Studia Mathematica* 185:67–82, 2008.

[154] J.A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

[155] Y. Vardi. Multiplicative censoring, renewal processes, deconvolution and decreasing density: nonparametric estimation. *Biometrika*, 76(4):751–761, 1989.

[156] Y. Vardi, L. Shepp, L. Kaufman. A statistical model for positron emission tomography. *Journal of the American statistical Association*, 80(389):8–20, 1985.

[157] V.V. Veeravalli, T. Banerjee. Quickest change detection. In *E-Reference Signal Processing.* Elsevier, 2013.

[158] A. Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186, 1945.

[159] A. Wald. *Sequential Analysis.* John Wiley and Sons, NY, 1947.

[160] A. Wald, J. Wolfowitz. Optimum character of the sequential probability ratio test. *The Annals of Mathematical Statistics*, pages 326–339, 1948.

[161] Y. Wang. Jump and sharp cusp detection by wavelets. *Biometrika*, 82(2):385–397, 1995.

[162] L. Wasserman. *All of nonparametric statistics.* Springer Science & Business Media, 2006.

[163] A.S. Willsky. *Detection of abrupt changes in dynamic systems.* Springer, 1985.

[164] Y. Xie, D. Siegmund. Sequential multi-sensor change-point detection. *Annals of Statistics*, 41(2):670–692, 2013.

[165] Y. Yin. Detection of the number, locations and magnitudes of jumps. *Communications in Statistics. Stochastic Models*, 4(3):445–455, 1988.

[166] C.-H. Zhang. Fourier methods for estimating mixing densities and distributions. *The Annals of Statistics*, pages 806–831, 1990.

Appendix A

Prerequisites from Linear Algebra and Analysis

Regarded as mathematical entities, the objective and the constraints in a Mathematical Programming problem are functions of several real variables; therefore before entering the Optimization Theory and Methods, we need to recall several basic notions and facts about the spaces $\mathbf{R}^n$ where these functions live, same as about the functions themselves. The reader is supposed to know most of the facts to follow, so he/she should not be surprised by a "cooking book" style which we intend to use in this Lecture.

A.1 Space $\mathbf{R}^n$: algebraic structure

Basically all events and constructions to be considered will take place in the *space* $\mathbf{R}^n$ *of n-dimensional real vectors*. This space can be described as follows.

A.1.1 A point in $\mathbf{R}^n$

A point in $\mathbf{R}^n$ (called also an *n-dimensional vector*) is an ordered collection $x = (x_1, ..., x_n)$ of n reals, called the *coordinates*, or *components*, or *entries* of vector x; the space $\mathbf{R}^n$ itself is the set of all collections of this type.

A.1.2 Linear operations

$\mathbf{R}^n$ is equipped with two *basic operations*:

- *Addition of vectors.* This operation takes on input two vectors $x = (x_1, ..., x_n)$ and $y = (y_1, ..., y_n)$ and produces from them a new vector

  $$x + y = (x_1 + y_1, ..., x_n + y_n)$$

  with entries which are sums of the corresponding entries in x and in y.

- *Multiplication of vectors by reals.* This operation takes on input a real λ and an n-dimensional vector $x = (x_1, ..., x_n)$ and produces from them a new vector

  $$\lambda x = (\lambda x_1, ..., \lambda x_n)$$

  with entries which are λ times the entries of x.

The as far as addition and multiplication by reals are concerned, the arithmetic of $\mathbf{R}^n$ inherits most of the common rules of Real Arithmetic, like $x+y = y+x$, $(x+y)+z = x+(y+z)$, $(\lambda+\mu)(x+y) = \lambda x+\mu x+\lambda y+\mu y$, $\lambda(\mu x) = (\lambda\mu)x$, etc.

A.1.3 Linear subspaces

Linear subspaces in $\mathbf{R}^n$ are, by definition, nonempty subsets of $\mathbf{R}^n$ which are closed with respect to addition of vectors and multiplication of vectors by reals:

$$L \subset \mathbf{R}^n \text{ is a linear subspace } \Leftrightarrow \left\{ \begin{array}{c} L \neq \emptyset; \\ x, y \in L \Rightarrow x + y \in L; \\ x \in L, \lambda \in \mathbf{R} \Rightarrow \lambda x \in L. \end{array} \right.$$

A.1.3.1 Examples of linear subspaces

Examples of linear subspaces are:

1. The entire $\mathbf{R}^n$;
2. The *trivial* subspace containing the single zero vector $0 = (0, ..., 0)$ [1]; (this vector/point is called also the origin)
3. The set $\{x \in \mathbf{R}^n : x_1 = 0\}$ of all vectors x with the first coordinate equal to zero.

   The latter example admits a natural extension:
4. The set of all solutions to a *homogeneous* (i.e., with zero right hand side) system of linear equations

$$\left\{ x \in \mathbf{R}^n : \begin{array}{rcl} a_{11}x_1 + ... + a_{1n}x_n & = & 0 \\ a_{21}x_1 + ... + a_{2n}x_n & = & 0 \\ & ... & \\ a_{m1}x_1 + ... + a_{mn}x_n & = & 0 \end{array} \right\} \tag{A.1.1}$$

   always is a linear subspace in $\mathbf{R}^n$. This example is "generic", that is, *every* linear subspace in $\mathbf{R}^n$ is the solution set of a (finite) system of homogeneous linear equations, see Proposition A.3.6 below.
5. *Linear span of a set of vectors.* Given a nonempty set X of vectors, one can form a linear subspace $\mathrm{Lin}(X)$, called the *linear span* of X; this subspace consists of all vectors x which can be represented as *linear combinations* $\sum\limits_{i=1}^{N} \lambda_i x_i$ of vectors from X (in $\sum\limits_{i=1}^{N} \lambda_i x_i$, N is an arbitrary positive integer, λ_i are reals and x_i belong to X). Note that

   > $\mathrm{Lin}(X)$ *is the smallest linear subspace which contains* X*: if* L *is a linear subspace such that* $L \supset X$*, then* $L \supset L(X)$ (why?).

   The "linear span" example also is generic:

   > *Every linear subspace in* $\mathbf{R}^n$ *is the linear span of an appropriately chosen finite set of vectors from* $\mathbf{R}^n$.

   (see Theorem A.1.2.(i) below).

A.1.3.2 Sums and intersections of linear subspaces

Let $\{L_\alpha\}_{\alpha \in I}$ be a family (finite or infinite) of linear subspaces of $\mathbf{R}^n$. From this family, one can build two sets:

1. *The sum* $\sum\limits_{\alpha} L_\alpha$ of the subspaces L_α which consists of all vectors which can be represented as finite sums of vectors taken each from its own subspace of the family;
2. *The intersection* $\bigcap\limits_{\alpha} L_\alpha$ of the subspaces from the family.

[1] Pay attention to the notation: we use the same symbol 0 to denote the real zero and the n-dimensional vector with all coordinates equal to zero; these two zeros are not the same, and one should understand from the context (it always is very easy) which zero is meant.

Theorem A.1.1 *Let $\{L_\alpha\}_{\alpha\in I}$ be a family of linear subspaces of $\mathbf{R}^n$. Then*

(i) *The sum $\sum\limits_\alpha L_\alpha$ of the subspaces from the family is itself a linear subspace of $\mathbf{R}^n$; it is the smallest of those subspaces of $\mathbf{R}^n$ which contain every subspace from the family;*

(ii) *The intersection $\bigcap\limits_\alpha L_\alpha$ of the subspaces from the family is itself a linear subspace of $\mathbf{R}^n$; it is the largest of those subspaces of $\mathbf{R}^n$ which are contained in every subspace from the family.*

A.1.4 Linear independence, bases, dimensions

A collection $X = \{x^1, ..., x^N\}$ of vectors from $\mathbf{R}^n$ is called *linearly independent*, if no nontrivial (i.e., with at least one nonzero coefficient) linear combination of vectors from X is zero.

> *Example of linearly independent set:* the collection of n *standard basic orths* $e_1 = (1, 0, ..., 0)$, $e_2 = (0, 1, 0, ..., 0)$, ..., $e_n = (0, ..., 0, 1)$.
>
> *Examples of linearly dependent sets:* (1) $X = \{0\}$; (2) $X = \{e_1, e_1\}$; (3) $X = \{e_1, e_2, e_1 + e_2\}$.

A collection of vectors $f^1, ..., f^m$ is called a *basis* in $\mathbf{R}^n$, if

1. The collection is linearly independent;
2. Every vector from $\mathbf{R}^n$ is a linear combination of vectors from the collection (i.e., $\mathrm{Lin}\{f^1, ..., f^m\} = \mathbf{R}^n$).

> *Example of a basis:* The collection of standard basic orths $e_1, ..., e_n$ is a basis in $\mathbf{R}^n$.
>
> *Examples of non-bases:* (1) The collection $\{e_2, ..., e_n\}$. This collection is linearly independent, but not every vector is a linear combination of the vectors from the collection; (2) The collection $\{e_1, e_1, e_2, ..., e_n\}$. Every vector is a linear combination of vectors form the collection, but the collection is not linearly independent.

Besides the bases of the entire $\mathbf{R}^n$, one can speak about the bases of linear subspaces:

A collection $\{f^1, ..., f^m\}$ of vectors is called a *basis of a linear subspace* L, if

1. The collection is linearly independent,
2. $L = \mathrm{Lin}\{f^1, ..., f^m\}$, i.e., all vectors f^i belong to L, and every vector from L is a linear combination of the vectors $f^1, ..., f^m$.

In order to avoid trivial remarks, it makes sense to agree once for ever that

> *An empty set of vectors is linearly independent, and an empty linear combination of vectors $\sum\limits_{i\in\emptyset} \lambda_i x_i$ equals to zero.*

With this convention, the trivial linear subspace $L = \{0\}$ also has a basis, specifically, an empty set of vectors.

Theorem A.1.2 (i) *Let L be a linear subspace of $\mathbf{R}^n$. Then L admits a (finite) basis, and all bases of L are comprised of the same number of vectors; this number is called the dimension of L and is denoted by* $\dim(L)$.

> We have seen that $\mathbf{R}^n$ admits a basis comprised of n elements (the standard basic orths). From (i) it follows that *every* basis of $\mathbf{R}^n$ contains exactly n vectors, and the dimension of $\mathbf{R}^n$ is n.

(ii) *The larger is a linear subspace of $\mathbf{R}^n$, the larger is its dimension: if $L \subset L'$ are linear subspaces of $\mathbf{R}^n$, then $\dim(L) \leq \dim(L')$, and the equality takes place if and only if $L = L'$.*

> We have seen that the dimension of $\mathbf{R}^n$ is n; according to the above convention, the trivial linear subspace $\{0\}$ of $\mathbf{R}^n$ admits an empty basis, so that its dimension is 0. Since $\{0\} \subset L \subset \mathbf{R}^n$ for every linear subspace L of $\mathbf{R}^n$, it follows from (ii) that the dimension of a linear subspace in $\mathbf{R}^n$ is an integer between 0 and n.

(iii) *Let L be a linear subspace in $\mathbf{R}^n$. Then*

(iii.1) *Every linearly independent subset of vectors from L can be extended to a basis of L;*

(iii.2) *From every spanning subset X for L – i.e., a set X such that $\mathrm{Lin}(X) = L$ – one can extract a basis of L.*

It follows from (iii) that

– every linearly independent subset of L contains at most $\dim(L)$ vectors, and if it contains exactly $\dim(L)$ vectors, it is a basis of L;

– every spanning set for L contains at least $\dim(L)$ vectors, and if it contains exactly $\dim(L)$ vectors, it is a basis of L.

(iv) *Let L be a linear subspace in $\mathbf{R}^n$, and $f^1, ..., f^m$ be a basis in L. Then every vector $x \in L$ admits exactly one representation*

$$x = \sum_{i=1}^{m} \lambda_i(x) f^i$$

as a linear combination of vectors from the basis, and the mapping

$$x \mapsto (\lambda_1(x), ..., \lambda_m(x)) : L \to \mathbf{R}^m$$

is a one-to-one mapping of L onto $\mathbf{R}^m$ which is linear, i.e. for every $i = 1, ..., m$ one has

$$\begin{array}{rcl} \lambda_i(x+y) & = & \lambda_i(x) + \lambda_i(y) \quad \forall (x, y \in L); \\ \lambda_i(\nu x) & = & \nu \lambda_i(x) \quad \forall (x \in L, \nu \in \mathbf{R}). \end{array} \tag{A.1.2}$$

The reals $\lambda_i(x)$, $i = 1, ..., m$, are called the coordinates of $x \in L$ in the basis $f^1, ..., f^m$.

E.g., the coordinates of a vector $x \in \mathbf{R}^n$ in the *standard basis* $e_1, ..., e_n$ of $\mathbf{R}^n$ – the one comprised of the standard basic orths – are exactly the entries of x.

(v) [Dimension formula] *Let L_1, L_2 be linear subspaces of $\mathbf{R}^n$. Then*

$$\dim(L_1 \cap L_2) + \dim(L_1 + L_2) = \dim(L_1) + \dim(L_2).$$

A.1.5 Linear mappings and matrices

A function $\mathcal{A}(x)$ (another name – *mapping*) defined on $\mathbf{R}^n$ and taking values in $\mathbf{R}^m$ is called *linear*, if it preserves linear operations:

$$\mathcal{A}(x+y) = \mathcal{A}(x) + \mathcal{A}(y) \quad \forall (x, y \in \mathbf{R}^n); \quad \mathcal{A}(\lambda x) = \lambda \mathcal{A}(x) \quad \forall (x \in \mathbf{R}^n, \lambda \in \mathbf{R}).$$

It is immediately seen that a linear mapping from $\mathbf{R}^n$ to $\mathbf{R}^m$ can be represented as multiplication by an $m \times n$ matrix:

$$\mathcal{A}(x) = Ax,$$

and this matrix is uniquely defined by the mapping: the columns A_j of A are just the images of the standard basic orths e_j under the mapping $\mathcal{A}$:

$$A_j = \mathcal{A}(e_j).$$

Linear mappings from $\mathbf{R}^n$ into $\mathbf{R}^m$ can be added to each other:

$$(\mathcal{A} + \mathcal{B})(x) = \mathcal{A}(x) + \mathcal{B}(x)$$

and multiplied by reals:

$$(\lambda \mathcal{A})(x) = \lambda \mathcal{A}(x),$$

and the results of these operations again are linear mappings from $\mathbf{R}^n$ to $\mathbf{R}^m$. The addition of linear mappings and multiplication of these mappings by reals correspond to the same operations with the matrices representing the mappings: adding/multiplying by reals mappings, we add, respectively, multiply by reals the corresponding matrices.

Given two linear mappings $\mathcal{A}(x) : \mathbf{R}^n \to \mathbf{R}^m$ and $\mathcal{B}(y) : \mathbf{R}^m \to \mathbf{R}^k$, we can build their superposition

$$\mathcal{C}(x) \equiv \mathcal{B}(\mathcal{A}(x)) : \mathbf{R}^n \to \mathbf{R}^k,$$

which is again a linear mapping, now from $\mathbf{R}^n$ to $\mathbf{R}^k$. In the language of matrices representing the mappings, the superposition corresponds to matrix multiplication: the $k \times n$ matrix C representing the mapping $\mathcal{C}$ is the product of the matrices representing $\mathcal{A}$ and $\mathcal{B}$:

$$\mathcal{A}(x) = Ax, \; \mathcal{B}(y) = By \Rightarrow \mathcal{C}(x) \equiv \mathcal{B}(\mathcal{A}(x)) = B \cdot (Ax) = (BA)x.$$

Important convention. When speaking about adding n-dimensional vectors and multiplying them by reals, it is absolutely unimportant whether we treat the vectors as the column ones, or the row ones, or write down the entries in rectangular tables, or something else. However, when matrix operations (matrix-vector multiplication, transposition, etc.) become involved, it is important whether we treat our vectors as columns, as rows, or as something else. For the sake of definiteness, *from now on we treat all vectors as column ones*, independently of how we refer to them in the text. For example, when saying for the first time what a vector is, we wrote $x = (x_1, ..., x_n)$, which might suggest that we were speaking about row vectors. We stress that it is not the case, and the only reason for using the notation $x = (x_1, ..., x_n)$ instead of the "correct" one $x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$ is to save space and to avoid ugly formulas like $f(\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix})$ when speaking about functions with vector arguments. After we have agreed that *there is no such thing as a row vector in this Lecture course*, we can use (and do use) without any harm whatever notation we want.

Exercise A.1 1. *Mark in the list below those subsets of* $\mathbf{R}^n$ *which are linear subspaces, find out their dimensions and point out their bases:*

(a) $\mathbf{R}^n$

(b) $\{0\}$

(c) $\emptyset$

(d) $\{x \in \mathbf{R}^n : \sum\limits_{i=1}^{n} i x_i = 0\}$

(e) $\{x \in \mathbf{R}^n : \sum\limits_{i=1}^{n} i x_i^2 = 0\}$

(f) $\{x \in \mathbf{R}^n : \sum\limits_{i=1}^{n} i x_i = 1\}$

(g) $\{x \in \mathbf{R}^n : \sum\limits_{i=1}^{n} i x_i^2 = 1\}$

2. *It is known that L is a subspace of* $\mathbf{R}^n$ *with exactly one basis. What is L?*

3. *Consider the space* $\mathbf{R}^{m\times n}$ *of $m \times n$ matrices with real entries. As far as linear operations – addition of matrices and multiplication of matrices by reals – are concerned, this space can be treated as certain* $\mathbf{R}^N$.

   (a) *Find the dimension of* $\mathbf{R}^{m\times n}$ *and point out a basis in this space*

   (b) *In the space* $\mathbf{R}^{n\times n}$ *of square $n \times n$ matrices, there are two interesting subsets: the set* $\mathbf{S}^n$ *of symmetric matrices* $\{A = [A_{ij}] : A_{ij} = A_{ij}\}$ *and the set* $\mathbf{J}^n$ *of skew-symmetric matrices* $\{A = [A_{ij}] : A_{ij} = -A_{ji}\}$.

      i. *Verify that both* $\mathbf{S}^n$ *and* $\mathbf{J}^n$ *are linear subspaces of* $\mathbf{R}^{n\times n}$

      ii. *Find the dimension and point out a basis in* $\mathbf{S}^n$

      iii. *Find the dimension and point out a basis in* $\mathbf{J}^n$

      iv. *What is the sum of* $\mathbf{S}^n$ *and* $\mathbf{J}^n$*? What is the intersection of* $\mathbf{S}^n$ *and* $\mathbf{J}^n$*?*

A.2 Space $\mathbf{R}^n$: Euclidean structure

So far, we were interested solely in the algebraic structure of $\mathbf{R}^n$, or, which is the same, in the properties of the *linear* operations (addition of vectors and multiplication of vectors by scalars) the space is endowed with. Now let us consider another structure on $\mathbf{R}^n$ – the *standard Euclidean structure* – which allows to speak about distances, angles, convergence, etc., and thus makes the space $\mathbf{R}^n$ a much richer mathematical entity.

A.2.1 Euclidean structure

The standard Euclidean structure on $\mathbf{R}^n$ is given by the *standard inner product* – an operation which takes on input two vectors x, y and produces from them a real, specifically, the real

$$\langle x, y\rangle \equiv x^T y = \sum_{i=1}^n x_i y_i$$

The basic properties of the inner product are as follows:

1. [bi-linearity]: The real-valued function $\langle x, y\rangle$ of two vector arguments $x, y \in \mathbf{R}^n$ is linear with respect to every one of the arguments, the other argument being fixed:

$$\begin{array}{rcll} \langle \lambda u + \mu v, y\rangle & = & \lambda\langle u, y\rangle + \mu\langle v, y\rangle & \forall(u, v, y \in \mathbf{R}^n, \lambda, \mu \in \mathbf{R}) \\ \langle x, \lambda u + \mu v\rangle & = & \lambda\langle x, u\rangle + \mu\langle x, v\rangle & \forall(x, u, v \in \mathbf{R}^n, \lambda, \mu \in \mathbf{R}) \end{array}$$

2. [symmetry]: The function $\langle x, y\rangle$ is symmetric:

$$\langle x, y\rangle = \langle y, x\rangle \quad \forall(x, y \in \mathbf{R}^n).$$

3. [positive definiteness]: The quantity $\langle x, x\rangle$ always is nonnegative, and it is zero iff ("if and only if") x is zero.

Remark A.2.1 The outlined 3 properties – bi-linearity, symmetry and positive definiteness – form a definition of an Euclidean inner product, and there are infinitely many different from each other ways to satisfy these properties; in other words, there are infinitely many different Euclidean inner products on $\mathbf{R}^n$. The standard inner product $\langle x, y\rangle = x^T y$ is just a particular case of this general notion. Although in the sequel we normally work with the standard inner product, the reader should remember that the facts we are about to recall are valid for all Euclidean inner products, and not only for the standard one.

The notion of an inner product underlies a number of purely algebraic constructions, in particular, those of *inner product representation of linear forms* and of *orthogonal complement*.

A.2.2 Inner product representation of linear forms on $\mathbf{R}^n$

A *linear form* on $\mathbf{R}^n$ is a real-valued function $f(x)$ on $\mathbf{R}^n$ which is additive ($f(x + y) = f(x) + f(y)$) and homogeneous ($f(\lambda x) = \lambda f(x)$)

Example of linear form: $f(x) = \sum_{i=1}^n i x_i$

Examples of non-linear functions: (1) $f(x) = x_1 + 1$; (2) $f(x) = x_1^2 - x_2^2$; (3) $f(x) = \sin(x_1)$.

When adding/multiplying by reals linear forms, we again get linear forms (scientifically speaking: "linear forms on $\mathbf{R}^n$ form a linear space"). *Euclidean structure allows to identify linear forms on $\mathbf{R}^n$ with vectors from $\mathbf{R}^n$*:

Theorem A.2.1 *Let $\langle\cdot,\cdot\rangle$ be a Euclidean inner product on $\mathbf{R}^n$.*

(i) *Let $f(x)$ be a linear form on $\mathbf{R}^n$. Then there exists a uniquely defined vector $f \in \mathbf{R}^n$ such that the form is just the inner product with f:*

$$f(x) = \langle f, x\rangle \quad \forall x$$

(ii) *Vice versa, every vector $f \in \mathbf{R}^n$ defines, via the formula*

$$f(x) \equiv \langle f, x\rangle,$$

a linear form on $\mathbf{R}^n$;

(iii) *The above one-to-one correspondence between the linear forms and vectors on $\mathbf{R}^n$ is linear: adding linear forms (or multiplying a linear form by a real), we add (respectively, multiply by the real) the vector(s) representing the form(s).*

A.2.3 Orthogonal complement

An Euclidean structure allows to associate with a linear subspace $L \subset \mathbf{R}^n$ another linear subspace $L^\perp$ – the *orthogonal complement* (or the *annulator*) of L; by definition, $L^\perp$ consists of all vectors which are orthogonal to every vector from L:

$$L^\perp = \{f : \langle f, x\rangle = 0 \quad \forall x \in L\}.$$

Theorem A.2.2 (i) *Twice taken, orthogonal complement recovers the original subspace: whenever L is a linear subspace of $\mathbf{R}^n$, one has*

$$(L^\perp)^\perp = L;$$

(ii) *The larger is a linear subspace L, the smaller is its orthogonal complement: if $L_1 \subset L_2$ are linear subspaces of $\mathbf{R}^n$, then $L_1^\perp \supset L_2^\perp$*

(iii) *The intersection of a subspace and its orthogonal complement is trivial, and the sum of these subspaces is the entire $\mathbf{R}^n$:*

$$L \cap L^\perp = \{0\}, \quad L + L^\perp = \mathbf{R}^n.$$

Remark A.2.2 From Theorem A.2.2.(iii) and the Dimension formula (Theorem A.1.2.(v)) it follows, first, that for every subspace L in $\mathbf{R}^n$ one has

$$\dim(L) + \dim(L^\perp) = n.$$

Second, every vector $x \in \mathbf{R}^n$ admits a unique decomposition

$$x = x_L + x_{L^\perp}$$

into a sum of two vectors: the first of them, x_L, belongs to L, and the second, $x_{L^\perp}$, belongs to $L^\perp$. This decomposition is called the *orthogonal decomposition* of x *taken with respect to* $L, L^\perp$; x_L is called the *orthogonal projection* of x onto L, and $x_{L^\perp}$ – the orthogonal projection of x onto the orthogonal complement of L. Both projections depend on x linearly, for example,

$$(x+y)_L = x_L + y_L, \quad (\lambda x)_L = \lambda x_L.$$

The mapping $x \mapsto x_L$ is called the *orthogonal projector* onto L.

A.2.4 Orthonormal bases

A collection of vectors $f^1, ..., f^m$ is called *orthonormal* w.r.t. Euclidean inner product $\langle\cdot,\cdot\rangle$, if distinct vector from the collection are orthogonal to each other:

$$i \neq j \Rightarrow \langle f^i, f^j\rangle = 0$$

and inner product of every vector f^i with itself is unit:

$$\langle f^i, f^i\rangle = 1, \ i = 1, ..., m.$$

Theorem A.2.3 (i) *An orthonormal collection $f^1, ..., f^m$ always is linearly independent and is therefore a basis of its linear span $L = \mathrm{Lin}(f^1, ..., f^m)$ (such a basis in a linear subspace is called orthonormal). The coordinates of a vector $x \in L$ w.r.t. an orthonormal basis $f^1, ..., f^m$ of L are given by explicit formulas:*

$$x = \sum_{i=1}^m \lambda_i(x) f^i \Leftrightarrow \lambda_i(x) = \langle x, f^i\rangle.$$

Example of an orthonormal basis in $\mathbf{R}^n$*:* The standard basis $\{e_1, ..., e_n\}$ is orthonormal *with respect to the standard inner product* $\langle x, y\rangle = x^T y$ on $\mathbf{R}^n$ (but is not orthonormal w.r.t. other Euclidean inner products on $\mathbf{R}^n$).

Proof of (i): Taking inner product of both sides in the equality

$$x = \sum_j \lambda_j(x) f^j$$

with f^i, we get

$$\begin{array}{rcll} \langle x, f_i\rangle & = & \langle\sum\limits_j \lambda_j(x) f^j, f^i\rangle & \\ & = & \sum\limits_j \lambda_j(x)\langle f^j, f^i\rangle & \text{[bilinearity of inner product]} \\ & = & \lambda_i(x) & \text{[orthonormality of } \{f^i\}] \end{array}$$

Similar computation demonstrates that if 0 is represented as a linear combination of f^i with certain coefficients λ_i, then $\lambda_i = \langle 0, f^i\rangle = 0$, i.e., all the coefficients are zero; this means that an orthonormal system is linearly independent.

(ii) *If* $f^1, ..., f^m$ *is an orthonormal basis in a linear subspace* L*, then the inner product of two vectors* $x, y \in L$ *in the coordinates* $\lambda_i(\cdot)$ *w.r.t. this basis is given by the standard formula*

$$\langle x, y\rangle = \sum_{i=1}^m \lambda_i(x)\lambda_i(y).$$

Proof:

$$\begin{array}{rcll} x & = & \sum\limits_i \lambda_i(x) f^i,\ y = \sum\limits_i \lambda_i(y) f^i & \\ \Rightarrow \langle x, y\rangle & = & \langle\sum\limits_i \lambda_i(x) f^i, \sum\limits_i \lambda_i(y) f^i\rangle & \\ & = & \sum\limits_{i,j} \lambda_i(x)\lambda_j(y)\langle f^i, f^j\rangle & \text{[bilinearity of inner product]} \\ & = & \sum\limits_i \lambda_i(x)\lambda_i(y) & \text{[orthonormality of } \{f^i\}] \end{array}$$

(iii) *Every linear subspace* L *of* $\mathbf{R}^n$ *admits an orthonormal basis; moreover, every orthonormal system* $f^1, ..., f^m$ *of vectors from* L *can be extended to an orthonormal basis in* L.

Important corollary: *All Euclidean spaces of the same dimension are "the same".* Specifically, *if* L *is an* m*-dimensional space in a space* $\mathbf{R}^n$ *equipped with an Euclidean inner product* $\langle\cdot,\cdot\rangle$*, then there exists a one-to-one mapping* $x \mapsto A(x)$ *of* L *onto* $\mathbf{R}^m$ *such that*

- *The mapping preserves linear operations:*

$$A(x + y) = A(x) + A(y) \quad \forall (x, y \in L); A(\lambda x) = \lambda A(x) \quad \forall (x \in L, \lambda \in \mathbf{R});$$

- *The mapping converts the* $\langle\cdot,\cdot\rangle$ *inner product on* L *into the standard inner product on* $\mathbf{R}^m$*:*

$$\langle x, y\rangle = (A(x))^T A(y) \quad \forall x, y \in L.$$

Indeed, by (iii) L admits an orthonormal basis $f^1, ..., f^m$; using (ii), one can immediately check that the mapping

$$x \mapsto A(x) = (\lambda_1(x), ..., \lambda_m(x))$$

which maps $x \in L$ into the m-dimensional vector comprised of the coordinates of x in the basis $f^1, ..., f^m$, meets all the requirements.

Proof of (iii) is given by important by its own right *Gram-Schmidt orthogonalization process* as follows. We start with an arbitrary basis $h^1, ..., h^m$ in L and step by step convert it into an orthonormal basis $f^1, ..., f^m$. At the beginning of a step t of the construction, we already have an orthonormal collection $f^1, ..., f^{t-1}$ such that $\mathrm{Lin}\{f^1, ..., f^{t-1}\} = \mathrm{Lin}\{h^1, ..., h^{t-1}\}$. At a step t we

1. Build the vector
$$g^t = h^t - \sum_{j=1}^{t-1} \langle h^t, f^j\rangle f^j.$$
It is easily seen (check it!) that
   (a) One has
$$\mathrm{Lin}\{f^1, ..., f^{t-1}, g^t\} = \mathrm{Lin}\{h^1, ..., h^t\}; \tag{A.2.1}$$
   (b) $g^t \neq 0$ (derive this fact from (A.2.1) and the linear independence of the collection $h^1, ..., h^m$);
   (c) g^t is orthogonal to $f^1, ..., f^{t-1}$
2. Since $g^t \neq 0$, the quantity $\langle g^t, g^t\rangle$ is positive (positive definiteness of the inner product), so that the vector
$$f^t = \frac{1}{\sqrt{\langle g^t, g^t\rangle}} g^t$$
is well defined. It is immediately seen (check it!) that the collection $f^1, ..., f^t$ is orthonormal and
$$\mathrm{Lin}\{f^1, ..., f^t\} = \mathrm{Lin}\{f^1, ..., f^{t-1}, g^t\} = \mathrm{Lin}\{h^1, ..., h^t\}.$$
Step t of the orthogonalization process is completed.

After m steps of the optimization process, we end up with an orthonormal system $f^1, ..., f^m$ of vectors from L such that
$$\mathrm{Lin}\{f^1, ..., f^m\} = \mathrm{Lin}\{h^1, ..., h^m\} = L,$$
so that $f^1, ..., f^m$ is an orthonormal basis in L.

The construction can be easily modified (do it!) to extend a given orthonormal system of vectors from L to an orthonormal basis of L.

Exercise A.2 *1. What is the orthogonal complement (w.r.t. the standard inner product) of the subspace $\{x \in \mathbf{R}^n : \sum\limits_{i=1}^n x_i = 0\}$ in $\mathbf{R}^n$?*

2. Find an orthonormal basis (w.r.t. the standard inner product) in the linear subspace $\{x \in \mathbf{R}^n : x_1 = 0\}$ of $\mathbf{R}^n$

3. Let L be a linear subspace of $\mathbf{R}^n$, and $f^1, ..., f^m$ be an orthonormal basis in L. Prove that for every $x \in \mathbf{R}^n$, the orthogonal projection x_L of x onto L is given by the formula
$$x_L = \sum_{i=1}^m (x^T f^i) f^i.$$

4. Let L_1, L_2 be linear subspaces in $\mathbf{R}^n$. Verify the formulas
$$(L_1 + L_2)^\perp = L_1^\perp \cap L_2^\perp; \quad (L_1 \cap L_2)^\perp = L_1^\perp + L_2^\perp.$$

5. Consider the space of $m \times n$ matrices $\mathbf{R}^{m\times n}$, and let us equip it with the "standard inner product" (called the Frobenius inner product*)*
$$\langle A, B\rangle = \sum_{i,j} A_{ij} B_{ij}$$
(as if we were treating $m \times n$ matrices as mn-dimensional vectors, writing the entries of the matrices column by column, and then taking the standard inner product of the resulting long vectors).

(a) *Verify that in terms of matrix multiplication the Frobenius inner product can be written as*

$$\langle A, B\rangle = \mathrm{Tr}(AB^T),$$

where $\mathrm{Tr}(C)$ *is the trace (the sum of diagonal elements) of a square matrix* C*.*

(b) *Build an orthonormal basis in the linear subspace* $\mathbf{S}^n$ *of symmetric* $n\times n$ *matrices*

(c) *What is the orthogonal complement of the subspace* $\mathbf{S}^n$ *of symmetric* $n\times n$ *matrices in the space* $\mathbf{R}^{n\times n}$ *of square* $n\times n$ *matrices?*

(d) *Find the orthogonal decomposition, w.r.t.* $\mathbf{S}^2$*, of the matrix* $\left[\begin{array}{cc}1 & 2\\ 3 & 4\end{array}\right]$

A.3 Affine subspaces in $\mathbf{R}^n$

Many of events to come will take place not in the entire $\mathbf{R}^n$, but in its *affine subspaces* which, geometrically, are planes of different dimensions in $\mathbf{R}^n$. Let us become acquainted with these subspaces.

A.3.1 Affine subspaces and affine hulls

Definition of an affine subspace. Geometrically, a linear subspace L of $\mathbf{R}^n$ is a special plane – the one passing through the origin of the space (i.e., containing the zero vector). To get an arbitrary plane M, it suffices to subject an appropriate special plane L to a translation – to add to all points from L a fixed *shifting vector* a. This geometric intuition leads to the following

Definition A.3.1 [Affine subspace] *An affine subspace (a plane) in* $\mathbf{R}^n$ *is a set of the form*

$$M = a + L = \{y = a + x \mid x \in L\}, \tag{A.3.1}$$

where L *is a linear subspace in* $\mathbf{R}^n$ *and* a *is a vector from* $\mathbf{R}^n$ [2]*.*

E.g., shifting the linear subspace L comprised of vectors with zero first entry by a vector $a = (a_1, ..., a_n)$, we get the set $M = a + L$ of all vectors x with $x_1 = a_1$; according to our terminology, this is an affine subspace.

Immediate question about the notion of an affine subspace is: what are the "degrees of freedom" in decomposition (A.3.1) – how "strict" M determines a and L? The answer is as follows:

Proposition A.3.1 *The linear subspace* L *in decomposition (A.3.1) is uniquely defined by* M *and is the set of all differences of the vectors from* M*:*

$$L = M - M = \{x - y \mid x, y \in M\}. \tag{A.3.2}$$

The shifting vector a *is not uniquely defined by* M *and can be chosen as an arbitrary vector from* M*.*

A.3.2 Intersections of affine subspaces, affine combinations and affine hulls

An immediate conclusion of Proposition A.3.1 is as follows:

Corollary A.3.1 *Let* $\{M_\alpha\}$ *be an arbitrary family of affine subspaces in* $\mathbf{R}^n$*, and assume that the set* $M = \cap_\alpha M_\alpha$ *is nonempty. Then* M_α *is an affine subspace.*

From Corollary A.3.1 it immediately follows that for every nonempty subset Y of $\mathbf{R}^n$ there exists the smallest affine subspace containing Y – the intersection of all affine subspaces containing Y. This smallest affine subspace containing Y is called the *affine hull* of Y (notation: $\mathrm{Aff}(Y)$).

All this resembles a lot the story about linear spans. Can we further extend this analogy and to get a description of the affine hull $\mathrm{Aff}(Y)$ in terms of elements of Y similar to the one of the linear span ("linear span of X is the set of all linear combinations of vectors from X")? Sure we can!

[2] according to our convention on arithmetic of sets, we were supposed to write in (A.3.1) $\{a\} + L$ instead of $a + L$ – we did not define arithmetic sum of a vector and a set. Usually people ignore this difference and omit the brackets when writing down singleton sets in similar expressions: we shall write $a + L$ instead of $\{a\} + L$, $\mathbf{R}d$ instead of $\mathbf{R}\{d\}$, etc.

Let us choose somehow a point $y_0 \in Y$, and consider the set

$$X = Y - y_0.$$

All affine subspaces containing Y should contain also y_0 and therefore, by Proposition A.3.1, can be represented as $M = y_0 + L$, L being a linear subspace. It is absolutely evident that an affine subspace $M = y_0 + L$ contains Y iff the subspace L contains X, and that the larger is L, the larger is M:

$$L \subset L' \Rightarrow M = y_0 + L \subset M' = y_0 + L'.$$

Thus, to find the smallest among *affine subspaces containing* Y, it suffices to find the smallest among the *linear subspaces containing* X and to translate the latter space by y_0:

$$\mathrm{Aff}(Y) = y_0 + \mathrm{Lin}(X) = y_0 + \mathrm{Lin}(Y - y_0). \tag{A.3.3}$$

Now, we know what is $\mathrm{Lin}(Y - y_0)$ – this is a set of all linear combinations of vectors from $Y - y_0$, so that a generic element of $\mathrm{Lin}(Y - y_0)$ is

$$x = \sum_{i=1}^{k} \mu_i (y_i - y_0) \quad [k \text{ may depend of } x]$$

with $y_i \in Y$ and real coefficients μ_i. It follows that the generic element of $\mathrm{Aff}(Y)$ is

$$y = y_0 + \sum_{i=1}^{k} \mu_i (y_i - y_0) = \sum_{i=0}^{k} \lambda_i y_i,$$

where

$$\lambda_0 = 1 - \sum_i \mu_i, \ \lambda_i = \mu_i, \ i \geq 1.$$

We see that a generic element of $\mathrm{Aff}(Y)$ is a linear combination of vectors from Y. Note, however, that the coefficients λ_i in this combination are not completely arbitrary: their sum is equal to 1. Linear combinations of this type – with the unit sum of coefficients – have a special name – they are called affine combinations.

We have seen that every vector from $\mathrm{Aff}(Y)$ is an affine combination of vectors of Y. Whether the inverse is true, i.e., whether $\mathrm{Aff}(Y)$ contains all affine combinations of vectors from Y? The answer is positive. Indeed, if

$$y = \sum_{i=1}^{k} \lambda_i y_i$$

is an affine combination of vectors from Y, then, using the equality $\sum_i \lambda_i = 1$, we can write it also as

$$y = y_0 + \sum_{i=1}^{k} \lambda_i (y_i - y_0),$$

y_0 being the "marked" vector we used in our previous reasoning, and the vector of this form, as we already know, belongs to $\mathrm{Aff}(Y)$. Thus, we come to the following

Proposition A.3.2 [Structure of affine hull]

$$\mathrm{Aff}(Y) = \{\textit{the set of all affine combinations of vectors from } Y\}.$$

When Y itself is an affine subspace, it, of course, coincides with its affine hull, and the above Proposition leads to the following

Corollary A.3.2 *An affine subspace M is closed with respect to taking affine combinations of its members – every combination of this type is a vector from M. Vice versa, a nonempty set which is closed with respect to taking affine combinations of its members is an affine subspace.*

A.3.3 Affinely spanning sets, affinely independent sets, affine dimension

Affine subspaces are closely related to linear subspaces, and the basic notions associated with linear subspaces have natural and useful affine analogies. Here we introduce these notions and discuss their basic properties.

Affinely spanning sets. Let $M = a + L$ be an affine subspace. We say that a subset Y of M is *affinely spanning* for M (we say also that Y spans M affinely, or that M is affinely spanned by Y), if $M = \mathrm{Aff}(Y)$, or, which is the same due to Proposition A.3.2, if every point of M is an affine combination of points from Y. An immediate consequence of the reasoning of the previous Section is as follows:

Proposition A.3.3 *Let $M = a + L$ be an affine subspace and Y be a subset of M, and let $y_0 \in Y$. The set Y affinely spans M – $M = \mathrm{Aff}(Y)$ – if and only if the set*

$$X = Y - y_0$$

spans the linear subspace L: $L = \mathrm{Lin}(X)$.

Affinely independent sets. A linearly independent set $x_1, ..., x_k$ is a set such that no nontrivial linear combination of $x_1, ..., x_k$ equals to zero. An equivalent definition is given by Theorem A.1.2.(iv): $x_1, ..., x_k$ are linearly independent, if the coefficients in a linear combination

$$x = \sum_{i=1}^{k} \lambda_i x_i$$

are uniquely defined by the value x of the combination. This equivalent form reflects the essence of the matter – what we indeed need, is the uniqueness of the coefficients in expansions. Accordingly, this equivalent form is the prototype for the notion of an affinely independent set: we want to introduce this notion in such a way that the coefficients λ_i in an *affine* combination

$$y = \sum_{i=0}^{k} \lambda_i y_i$$

of "affinely independent" set of vectors $y_0, ..., y_k$ would be uniquely defined by y. *Non*-uniqueness would mean that

$$\sum_{i=0}^{k} \lambda_i y_i = \sum_{i=0}^{k} \lambda'_i y_i$$

for two different collections of coefficients λ_i and λ'_i with unit sums of coefficients; if it is the case, then

$$\sum_{i=0}^{m} (\lambda_i - \lambda'_i) y_i = 0,$$

so that y_i's are linearly dependent and, moreover, there exists a nontrivial zero combination of then with *zero sum of coefficients* (since $\sum_i (\lambda_i - \lambda'_i) = \sum_i \lambda_i - \sum_i \lambda'_i = 1 - 1 = 0$). Our reasoning can be inverted – if there exists a nontrivial linear combination of y_i's with zero sum of coefficients which is zero, then the coefficients in the representation of a vector as an affine combination of y_i's are not uniquely defined. Thus, in order to get uniqueness we should for sure forbid relations

$$\sum_{i=0}^{k} \mu_i y_i = 0$$

with nontrivial zero sum coefficients μ_i. Thus, we have motivated the following

Definition A.3.2 [Affinely independent set] *A collection $y_0, ..., y_k$ of n-dimensional vectors is called affinely independent, if no nontrivial linear combination of the vectors with zero sum of coefficients is zero:*

$$\sum_{i=1}^{k} \lambda_i y_i = 0, \ \sum_{i=0}^{k} \lambda_i = 0 \Rightarrow \lambda_0 = \lambda_1 = ... = \lambda_k = 0.$$

With this definition, we get the result completely similar to the one of Theorem A.1.2.(iv):

Corollary A.3.3 *Let $y_0, ..., y_k$ be affinely independent. Then the coefficients λ_i in an affine combination*

$$y = \sum_{i=0}^{k} \lambda_i y_i \quad [\sum_i \lambda_i = 1]$$

of the vectors $y_0, ..., y_k$ are uniquely defined by the value y of the combination.

Verification of affine independence of a collection can be immediately reduced to verification of linear independence of closely related collection:

Proposition A.3.4 *$k+1$ vectors $y_0, ..., y_k$ are affinely independent if and only if the k vectors $(y_1 - y_0), (y_2 - y_0), ..., (y_k - y_0)$ are linearly independent.*

From the latter Proposition it follows, e.g., that the collection $0, e_1, ..., e_n$ comprised of the origin and the standard basic orths is affinely independent. Note that this collection is linearly dependent (as every collection containing zero). You should definitely know the difference between the two notions of independence we deal with: linear independence means that no nontrivial linear combination of the vectors can be zero, while affine independence means that no nontrivial linear combination *from certain restricted class of them* (with zero sum of coefficients) can be zero. Therefore, there are more affinely independent sets than the linearly independent ones: a linearly independent set is for sure affinely independent, but not vice versa.

Affine bases and affine dimension. Propositions A.3.2 and A.3.3 reduce the notions of affine spanning/affine independent sets to the notions of spanning/linearly independent ones. Combined with Theorem A.1.2, they result in the following analogies of the latter two statements:

Proposition A.3.5 [Affine dimension] *Let $M = a + L$ be an affine subspace in $\mathbf{R}^n$. Then the following two quantities are finite integers which are equal to each other:*

(i) *minimal # of elements in the subsets of M which affinely span M;*

(ii) *maximal # of elements in affine independent subsets of M.*

The common value of these two integers is by 1 more than the dimension dim L *of L.*

By definition, the *affine dimension* of an affine subspace $M = a + L$ is the dimension dim L of L. Thus, if M is of affine dimension k, then the minimal cardinality of sets affinely spanning M, same as the maximal cardinality of affine independent subsets of M, is $k + 1$.

Theorem A.3.1 [Affine bases] *Let $M = a + L$ be an affine subspace in $\mathbf{R}^n$.*

A. *Let $Y \subset M$. The following three properties of X are equivalent:*

(i) *Y is an affine independent set which affinely spans M;*

(ii) *Y is affine independent and contains $1 + \dim L$ elements;*

(iii) *Y affinely spans M and contains $1 + \dim L$ elements.*

A subset Y of M possessing the indicated equivalent to each other properties is called an affine basis of M. Affine bases in M are exactly the collections $y_0, ..., y_{\dim L}$ such that $y_0 \in M$ and $(y_1 - y_0), ..., (y_{\dim L} - y_0)$ is a basis in L.

B. *Every affinely independent collection of vectors of M either itself is an affine basis of M, or can be extended to such a basis by adding new vectors. In particular, there exists affine basis of M.*

C. *Given a set Y which affinely spans M, you can always extract from this set an affine basis of M.*

We already know that the standard basic orths $e_1, ..., e_n$ form a basis of the entire space $\mathbf{R}^n$. And what about affine bases in $\mathbf{R}^n$? According to Theorem A.3.1.**A**, you can choose as such a basis a collection $e_0, e_0 + e_1, ..., e_0 + e_n$, e_0 being an arbitrary vector.

Barycentric coordinates. Let M be an affine subspace, and let $y_0, ..., y_k$ be an affine basis of M. Since the basis, by definition, affinely spans M, every vector y from M is an affine combination of the vectors of the basis:

$$y = \sum_{i=0}^{k} \lambda_i y_i \quad [\sum_{i=0}^{k} \lambda_i = 1],$$

and since the vectors of the affine basis are affinely independent, the coefficients of this combination are uniquely defined by y (Corollary A.3.3). These coefficients are called *barycentric coordinates* of y with respect to the affine basis in question. In contrast to the usual coordinates with respect to a (linear) basis, the barycentric coordinates could not be quite arbitrary: their sum should be equal to 1.

A.3.4 Dual description of linear subspaces and affine subspaces

To the moment we have introduced the notions of linear subspace and affine subspace and have presented a scheme of generating these entities: to get, e.g., a linear subspace, you start from an arbitrary nonempty set $X \subset \mathbf{R}^n$ and add to it all linear combinations of the vectors from X. When replacing linear combinations with the affine ones, you get a way to generate affine subspaces.

The just indicated way of generating linear subspaces/affine subspaces resembles the approach of a worker building a house: he starts with the base and then adds to it new elements until the house is ready. There exists, anyhow, an approach of an artist creating a sculpture: he takes something large and then deletes extra parts of it. Is there something like "artist's way" to represent linear subspaces and affine subspaces? The answer is positive and very instructive.

A.3.4.1 Affine subspaces and systems of linear equations

Let L be a linear subspace. According to Theorem A.2.2.(i), it is an orthogonal complement – namely, the orthogonal complement to the linear subspace $L^\perp$. Now let $a_1, ..., a_m$ be a finite spanning set in $L^\perp$. A vector x which is orthogonal to $a_1, ..., a_m$ is orthogonal to the entire $L^\perp$ (since every vector from $L^\perp$ is a linear combination of $a_1, ..., a_m$ and the inner product is bilinear); and of course vice versa, a vector orthogonal to the entire $L^\perp$ is orthogonal to $a_1, ..., a_m$. We see that

$$L = (L^\perp)^\perp = \{x \mid a_i^T x = 0, \, i = 1, ..., k\}. \tag{A.3.4}$$

Thus, we get a very important, although simple,

Proposition A.3.6 ["Outer" description of a linear subspace] *Every linear subspace L in $\mathbf{R}^n$ is a set of solutions to a homogeneous linear system of equations*

$$a_i^T x = 0, \, i = 1, ..., m, \tag{A.3.5}$$

given by properly chosen m and vectors $a_1, ..., a_m$.

Proposition A.3.6 is an "iff " statement: as we remember from Example 4, solution set to a homogeneous system of linear equations with n variables always is a linear subspace in $\mathbf{R}^n$.

From Proposition A.3.6 and the facts we know about the dimension we can easily derive several important consequences:

- Systems (A.3.5) which define a given linear subspace L are exactly the systems given by the vectors $a_1, ..., a_m$ which span $L^\perp$ [3]
- The smallest possible number m of equations in (A.3.5) is the dimension of $L^\perp$, i.e., by Remark A.2.2, is codim $L \equiv n - \dim L$ [4]

[3] the reasoning which led us to Proposition A.3.6 says that $[a_1, ..., a_m$ span $L^\perp] \Rightarrow$ [(A.3.5) defines L]; now we claim that the inverse also is true

[4] to make this statement true also in the extreme case when $L = \mathbf{R}^n$ (i.e., when codim $L = 0$), we from now on make a convention that an *empty* set of equations or inequalities defines, as the solution set, the entire space

Now, an affine subspace M is, by definition, a translation of a linear subspace: $M = a + L$. As we know, vectors x from L are exactly the solutions of certain *homogeneous* system of linear equations

$$a_i^T x = 0,\ i = 1, ..., m.$$

It is absolutely clear that adding to these vectors a fixed vector a, we get exactly the set of solutions to the *inhomogeneous* solvable linear system

$$a_i^T x = b_i \equiv a_i^T a,\ i = 1, ..., m.$$

Vice versa, the set of solutions to a *solvable* system of linear equations

$$a_i^T x = b_i,\ i = 1, ..., m,$$

with n variables is the sum of a particular solution to the system and the solution set to the corresponding homogeneous system (the latter set, as we already know, is a linear subspace in $\mathbf{R}^n$), i.e., is an affine subspace. Thus, we get the following

Proposition A.3.7 [“Outer” description of an affine subspace]
Every affine subspace $M = a + L$ in $\mathbf{R}^n$ is a set of solutions to a solvable linear system of equations

$$a_i^T x = b_i,\ i = 1, ..., m, \tag{A.3.6}$$

given by properly chosen m and vectors $a_1, ..., a_m$.

Vice versa, the set of all solutions to a solvable system of linear equations with n variables is an affine subspace in $\mathbf{R}^n$.

The linear subspace L associated with M is exactly the set of solutions of the homogeneous (with the right hand side set to 0) version of system (A.3.6).

We see, in particular, that an affine subspace always is closed.

Comment. The “outer” description of a linear subspace/affine subspace – the “artist’s” one – is in many cases much more useful than the “inner” description via linear/affine combinations (the “worker’s” one). E.g., with the outer description it is very easy to check whether a given vector belongs or does not belong to a given linear subspace/affine subspace, which is not that easy with the inner one[5)]. In fact both descriptions are “complementary” to each other and perfectly well work in parallel: what is difficult to see with one of them, is clear with another. The idea of using “inner” and “outer” descriptions of the entities we meet with – linear subspaces, affine subspaces, convex sets, optimization problems – the general idea of *duality* – is, we would say, the main driving force of Convex Analysis and Optimization, and in the sequel we would all the time meet with different implementations of this fundamental idea.

A.3.5 Structure of the simplest affine subspaces

This small subsection deals mainly with terminology. According to their dimension, affine subspaces in $\mathbf{R}^n$ are named as follows:

- Subspaces of dimension 0 are translations of the only 0-dimensional linear subspace $\{0\}$, i.e., are singleton sets – vectors from $\mathbf{R}^n$. These subspaces are called *points*; a point is a solution to a square system of linear equations with nonsingular matrix.
- Subspaces of dimension 1 (lines). These subspaces are translations of one-dimensional linear subspaces of $\mathbf{R}^n$. A one-dimensional linear subspace has a single-element basis given by a nonzero vector d and is comprised of all multiples of this vector. Consequently, line is a set of the form

$$\{y = a + td \mid t \in \mathbf{R}\}$$

[5)] in principle it is not difficult to certify that a given point belongs to, say, a linear subspace given as the linear span of some set – it suffices to point out a representation of the point as a linear combination of vectors from the set. But how could you certify that the point does *not* belong to the subspace?

given by a pair of vectors a (the origin of the line) and d (the direction of the line), $d \neq 0$. The origin of the line and its direction are not uniquely defined by the line; you can choose as origin any point on the line and multiply a particular direction by nonzero reals.

In the barycentric coordinates a line is described as follows:

$$l = \{\lambda_0 y_0 + \lambda_1 y_1 \mid \lambda_0 + \lambda_1 = 1\} = \{\lambda y_0 + (1-\lambda) y_1 \mid \lambda \in \mathbf{R}\},$$

where y_0, y_1 is an affine basis of l; you can choose as such a basis any pair of distinct points on the line.

The "outer" description of a line is as follows: it is the set of solutions to a linear system with n variables and $n-1$ linearly independent equations.

- Subspaces of dimension > 2 and $< n-1$ have no special names; sometimes they are called affine planes of such and such dimension.

- Affine subspaces of dimension $n-1$, due to important role they play in Convex Analysis, have a special name – they are called *hyperplanes*. The outer description of a hyperplane is that a hyperplane is the solution set of a *single* linear equation

$$a^T x = b$$

with nontrivial left hand side ($a \neq 0$). In other words, a hyperplane is the level set $a(x) = \text{const}$ of a nonconstant linear form $a(x) = a^T x$.

- The "largest possible" affine subspace – the one of dimension n – is unique and is the entire $\mathbf{R}^n$. This subspace is given by an empty system of linear equations.

A.4 Space $\mathbf{R}^n$: metric structure and topology

Euclidean structure on the space $\mathbf{R}^n$ gives rise to a number of extremely important *metric* notions – distances, convergence, etc. For the sake of definiteness, we associate these notions with the standard inner product $x^T y$.

A.4.1 Euclidean norm and distances

By positive definiteness, the quantity $x^T x$ always is nonnegative, so that the quantity

$$|x| \equiv \|x\|_2 = \sqrt{x^T x} = \sqrt{x_1^2 + x_2^2 + ... + x_n^2}$$

is well-defined; this quantity is called the (standard) *Euclidean norm* of vector x (or simply the norm of x) and is treated as the distance from the origin to x. The distance between two arbitrary points $x, y \in \mathbf{R}^n$ is, by definition, the norm $|x - y|$ of the difference $x - y$. The notions we have defined satisfy all basic requirements on the general notions of a norm and distance, specifically:

1. *Positivity of norm: The norm of a vector always is nonnegative; it is zero if and only is the vector is zero*:
$$|x| \geq 0 \quad \forall x; \quad |x| = 0 \Leftrightarrow x = 0.$$

2. *Homogeneity of norm: When a vector is multiplied by a real, its norm is multiplied by the absolute value of the real*:
$$|\lambda x| = |\lambda| \cdot |x| \quad \forall (x \in \mathbf{R}^n, \lambda \in \mathbf{R}).$$

3. *Triangle inequality: Norm of the sum of two vectors is $\leq$ the sum of their norms*:
$$|x + y| \leq |x| + |y| \quad \forall (x, y \in \mathbf{R}^n).$$

In contrast to the properties of positivity and homogeneity, which are absolutely evident, the Triangle inequality is not trivial and definitely requires a proof. The proof goes through a fact which is extremely important by its own right – the *Cauchy Inequality*, which perhaps is the most frequently used inequality in Mathematics:

Theorem A.4.1 [Cauchy's Inequality] *The absolute value of the inner product of two vectors does not exceed the product of their norms:*

$$|x^Ty| \leq |x||y| \quad \forall (x, y \in \mathbf{R}^n)$$

and is equal to the product of the norms if and only if one of the vectors is proportional to the other one:

$$|x^Ty| = |x||y| \Leftrightarrow \{\exists \alpha : x = \alpha y \text{ or } \exists \beta : y = \beta x\}$$

Proof is immediate: we may assume that both x and y are nonzero (otherwise the Cauchy inequality clearly is equality, and one of the vectors is constant times (specifically, zero times) the other one, as announced in Theorem). Assuming $x, y \neq 0$, consider the function

$$f(\lambda) = (x - \lambda y)^T(x - \lambda y) = x^Tx - 2\lambda x^Ty + \lambda^2 y^Ty.$$

By positive definiteness of the inner product, this function – which is a second order polynomial – is nonnegative on the entire axis, whence the discriminant of the polynomial

$$(x^Ty)^2 - (x^Tx)(y^Ty)$$

is nonpositive:

$$(x^Ty)^2 \leq (x^Tx)(y^Ty).$$

Taking square roots of both sides, we arrive at the Cauchy Inequality. We also see that the inequality is equality if and only if the discriminant of the second order polynomial $f(\lambda)$ is zero, i.e., if and only if the polynomial has a (multiple) real root; but due to positive definiteness of inner product, $f(\cdot)$ has a root λ if and only if $x = \lambda y$, which proves the second part of Theorem. □

From Cauchy's Inequality to the Triangle Inequality: Let $x, y \in \mathbf{R}^n$. Then

$$\begin{array}{rcll} |x+y|^2 & = & (x+y)^T(x+y) & \text{[definition of norm]} \\ & = & x^Tx + y^Ty + 2x^Ty & \text{[opening parentheses]} \\ & \leq & \underbrace{x^Tx}_{|x|^2} + \underbrace{y^Ty}_{|y|^2} + 2|x||y| & \text{[Cauchy's Inequality]} \\ & = & (|x|+|y|)^2 & \\ \Rightarrow |x+y| & \leq & |x|+|y| & \square \end{array}$$

The properties of norm (i.e., of the distance to the origin) we have established induce properties of the distances between pairs of arbitrary points in $\mathbf{R}^n$, specifically:

1. *Positivity of distances: The distance $|x - y|$ between two points is positive, except for the case when the points coincide ($x = y$), when the distance between x and y is zero;*

2. *Symmetry of distances: The distance from x to y is the same as the distance from y to x:*

$$|x - y| = |y - x|;$$

3. *Triangle inequality for distances:* For every three points x, y, z, the distance from x to z does not exceed the sum of distances between x and y and between y and z:

$$|z - x| \leq |y - x| + |z - y| \quad \forall (x, y, z \in \mathbf{R}^n)$$

A.4.2 Convergence

Equipped with distances, we can define the fundamental notion of *convergence of a sequence of vectors*. Specifically, we say that *a sequence $x^1, x^2, ...$ of vectors from $\mathbf{R}^n$ converges to a vector $\bar{x}$*, or, equivalently, that *$\bar{x}$ is the limit of the sequence $\{x^i\}$* (notation: $\bar{x} = \lim\limits_{i\to\infty} x^i$), if the distances from $\bar{x}$ to x^i go to 0 as $i \to \infty$:

$$\bar{x} = \lim_{i\to\infty} x^i \Leftrightarrow |\bar{x} - x^i| \to 0, i \to \infty,$$

or, which is the same, for every $\epsilon > 0$ there exists $i = i(\epsilon)$ such that the distance between every point x^i, $i \geq i(\epsilon)$, and $\bar{x}$ does not exceed ϵ:

$$\left\{|\bar{x} - x^i| \to 0, i \to \infty\right\} \Leftrightarrow \left\{\forall \epsilon > 0 \exists i(\epsilon) : i \geq i(\epsilon) \Rightarrow |\bar{x} - x^i| \leq \epsilon\right\}.$$

Exercise A.3 *Verify the following facts:*

1. $\bar{x} = \lim\limits_{i\to\infty} x^i$ *if and only if for every* $j = 1, ..., n$ *the coordinates # j of the vectors* x^i *converge, as* $i \to \infty$, *to the coordinate # j of the vector* $\bar{x}$;
2. *If a sequence converges, its limit is uniquely defined;*
3. *Convergence is compatible with linear operations:*
   — *if* $x^i \to x$ *and* $y^i \to y$ *as* $i \to \infty$, *then* $x^i + y^i \to x + y$ *as* $i \to \infty$;
   — *if* $x^i \to x$ *and* $\lambda_i \to \lambda$ *as* $i \to \infty$, *then* $\lambda_i x^i \to \lambda x$ *as* $i \to \infty$.

A.4.3 Closed and open sets

After we have at our disposal distance and convergence, we can speak about *closed* and *open* sets:

- A set $X \subset \mathbf{R}^n$ is called *closed*, if it contains limits of all converging sequences of elements of X:

$$\left\{x^i \in X, x = \lim_{i\to\infty} x^i\right\} \Rightarrow x \in X$$

- A set $X \subset \mathbf{R}^n$ is called *open*, if whenever x belongs to X, all points close enough to x also belong to X:

$$\forall(x \in X)\exists(\delta > 0) : |x' - x| < \delta \Rightarrow x' \in X.$$

  An open set containing a point x is called a *neighbourhood* of x.

  Examples of closed sets: (1) $\mathbf{R}^n$; (2) $\emptyset$; (3) the sequence $x^i = (i, 0, ..., 0)$, $i = 1, 2, 3, ...$; (4) $\{x \in \mathbf{R}^n : \sum\limits_{i=1}^{n} a_{ij}x_j = 0,\ i = 1, ..., m\}$ (in other words: a linear subspace in $\mathbf{R}^n$ always is closed, see Proposition A.3.6);(5) $\{x \in \mathbf{R}^n : \sum\limits_{i=1}^{n} a_{ij}x_j = b_i,\ i = 1, ..., m\}$ (in other words: an affine subset of $\mathbf{R}^n$ always is closed, see Proposition A.3.7);; (6) Any finite subset of $\mathbf{R}^n$

  Examples of non-closed sets: (1) $\mathbf{R}^n\backslash\{0\}$; (2) the sequence $x^i = (1/i, 0, ..., 0)$, $i = 1, 2, 3, ...$; (3) $\{x \in \mathbf{R}^n : x_j > 0, j = 1, ..., n\}$; (4) $\{x \in \mathbf{R}^n : \sum\limits_{i=1}^{n} x_j > 5\}$.

  Examples of open sets: (1) $\mathbf{R}^n$; (2) $\emptyset$; (3) $\{x \in \mathbf{R}^n : \sum\limits_{j=1}^{n} a_{ij}x_j > b_j, i = 1, ..., m\}$; (4) complement of a finite set.

  Examples of non-open sets: (1) A nonempty finite set; (2) the sequence $x^i = (1/i, 0, ..., 0)$, $i = 1, 2, 3, ...$, and the sequence $x^i = (i, 0, 0, ..., 0)$, $i = 1, 2, 3, ...$; (3) $\{x \in \mathbf{R}^n : x_j \geq 0, j = 1, ..., n\}$; (4) $\{x \in \mathbf{R}^n : \sum\limits_{i=1}^{n} x_j \geq 5\}$.

Exercise A.4 *Mark in the list to follows those sets which are closed and those which are open:*

1. *All vectors with integer coordinates*
2. *All vectors with rational coordinates*
3. *All vectors with positive coordinates*
4. *All vectors with nonnegative coordinates*
5. $\{x : |x| < 1\}$;
6. $\{x : |x| = 1\}$;

7. $\{x : |x| \leq 1\}$;
8. $\{x : |x| \geq 1\}$:
9. $\{x : |x| > 1\}$;
10. $\{x : 1 < |x| \leq 2\}$.

Verify the following facts

1. *A set $X \subset \mathbf{R}^n$ is closed if and only if its complement $\bar{X} = \mathbf{R}^n \backslash X$ is open;*
2. *Intersection of every family (finite or infinite) of closed sets is closed. Union of every family (finite of infinite) of open sets is open.*
3. *Union of finitely many closed sets is closed. Intersection of finitely many open sets is open.*

A.4.4 Local compactness of $\mathbf{R}^n$

A fundamental fact about convergence in $\mathbf{R}^n$, which in certain sense is characteristic for this series of spaces, is the following

Theorem A.4.2 *From every bounded sequence $\{x^i\}_{i=1}^{\infty}$ of points from $\mathbf{R}^n$ one can extract a converging subsequence $\{x^{i_j}\}_{j=1}^{\infty}$. Equivalently: A closed and bounded subset X of $\mathbf{R}^n$ is compact, i.e., a set possessing the following two equivalent to each other properties:*

(i) *From every sequence of elements of X one can extract a subsequence which converges to certain point of X;*

(ii) *From every open covering of X (i.e., a family $\{U_\alpha\}_{\alpha \in A}$ of open sets such that $X \subset \bigcup\limits_{\alpha \in A} U_\alpha$) one can extract a finite sub-covering, i.e., a finite subset of indices $\alpha_1, ..., \alpha_N$ such that $X \subset \bigcup\limits_{i=1}^{N} U_{\alpha_i}$.*

A.5 Continuous functions on $\mathbf{R}^n$

A.5.1 Continuity of a function

Let $X \subset \mathbf{R}^n$ and $f(x) : X \to \mathbf{R}^m$ be a function (another name – mapping) defined on X and taking values in $\mathbf{R}^m$.

1. f is called *continuous at a point* $\bar{x} \in X$, if for every sequence x^i of points of X converging to $\bar{x}$ the sequence $f(x^i)$ converges to $f(\bar{x})$. Equivalent definition:

   $f : X \to \mathbf{R}^m$ is continuous at $\bar{x} \in X$, if for every $\epsilon > 0$ there exists $\delta > 0$ such that
   $$x \in X, |x - \bar{x}| < \delta \Rightarrow |f(x) - f(\bar{x})| < \epsilon.$$

2. f is called *continuous on* X, if f is continuous at every point from X. Equivalent definition: f preserves convergence: whenever a sequence of points $x^i \in X$ converges to a point $x \in X$, the sequence $f(x^i)$ converges to $f(x)$.

   Examples of continuous mappings:

   1. An *affine* mapping
   $$f(x) = \begin{bmatrix} \sum\limits_{j=1}^{m} A_{1j} x_j + b_1 \\ \vdots \\ \sum\limits_{j=1}^{m} A_{mj} x_j + b_m \end{bmatrix} \equiv Ax + b : \mathbf{R}^n \to \mathbf{R}^m$$
   is continuous on the entire $\mathbf{R}^n$ (and thus – on every subset of $\mathbf{R}^n$) (check it!).

2. The norm $|x|$ is a continuous on $\mathbf{R}^n$ (and thus – on every subset of $\mathbf{R}^n$) real-valued function (check it!).

Exercise A.5 • *Consider the function*

$$f(x_1,x_2) = \begin{cases} \frac{x_1^2-x_2^2}{x_1^2+x_2^2}, & (x_1,x_2) \neq 0 \\ 0, & x_1 = x_2 = 0 \end{cases} : \mathbf{R}^2 \to \mathbf{R}.$$

Check whether this function is continuous on the following sets:

1. $\mathbf{R}^2$;
2. $\mathbf{R}^2\backslash\{0\}$;
3. $\{x \in \mathbf{R}^2 : x_1 = 0\}$;
4. $\{x \in \mathbf{R}^2 : x_2 = 0\}$;
5. $\{x \in \mathbf{R}^2 : x_1 + x_2 = 0\}$;
6. $\{x \in \mathbf{R}^2 : x_1 - x_2 = 0\}$;
7. $\{x \in \mathbf{R}^2 : |x_1 - x_2| \leq x_1^4 + x_2^4\}$;

- *Let $f : \mathbf{R}^n \to \mathbf{R}^m$ be a continuous mapping. Mark those of the following statements which always are true:*

1. *If U is an open set in $\mathbf{R}^m$, then so is the set $f^{-1}(U) = \{x : f(x) \in U\}$;*
2. *If U is an open set in $\mathbf{R}^n$, then so is the set $f(U) = \{f(x) : x \in U\}$;*
3. *If F is a closed set in $\mathbf{R}^m$, then so is the set $f^{-1}(F) = \{x : f(x) \in F\}$;*
4. *If F is an closed set in $\mathbf{R}^n$, then so is the set $f(F) = \{f(x) : x \in F\}$.*

A.5.2 Elementary continuity-preserving operations

All "elementary" operations with mappings preserve continuity. Specifically,

Theorem A.5.1 *Let X be a subset in $\mathbf{R}^n$.*

(i) [stability of continuity w.r.t. linear operations] *If $f_1(x), f_2(x)$ are continuous functions on X taking values in $\mathbf{R}^m$ and $\lambda_1(x)$, $\lambda_2(x)$ are continuous real-valued functions on X, then the function*

$$f(x) = \lambda_1(x)f_1(x) + \lambda_2(x)f_2(x) : X \to \mathbf{R}^m$$

is continuous on X;

(ii) [stability of continuity w.r.t. superposition] *Let*

- *$X \subset \mathbf{R}^n$, $Y \subset \mathbf{R}^m$;*
- *$f : X \to \mathbf{R}^m$ be a continuous mapping such that $f(x) \in Y$ for every $x \in X$;*
- *$g : Y \to \mathbf{R}^k$ be a continuous mapping.*

Then the composite mapping

$$h(x) = g(f(x)) : X \to \mathbf{R}^k$$

is continuous on X.

A.5.3 Basic properties of continuous functions on $\mathbf{R}^n$

The basic properties of continuous functions on $\mathbf{R}^n$ can be summarized as follows:

Theorem A.5.2 *Let X be a nonempty closed and bounded subset of $\mathbf{R}^n$.*

(i) *If a mapping $f : X \to \mathbf{R}^m$ is continuous on X, it is bounded on X: there exists $C < \infty$ such that $|f(x)| \leq C$ for all $x \in X$.*

Proof. Assume, on the contrary to what should be proved, that f is unbounded, so that for every i there exists a point $x^i \in X$ such that $|f(x^i)| > i$. By Theorem A.4.2, we can extract from the sequence $\{x^i\}$ a subsequence $\{x^{i_j}\}_{j=1}^{\infty}$ which converges to a point $\bar{x} \in X$. The real-valued function $g(x) = |f(x)|$ is continuous (as the superposition of two continuous mappings, see Theorem A.5.1.(ii)) and therefore its values at the points x^{i_j} should converge, as $j \to \infty$, to its value at $\bar{x}$; on the other hand, $g(x^{i_j}) \geq i_j \to \infty$ as $j \to \infty$, and we get the desired contradiction.

(ii) *If a mapping $f : X \to \mathbf{R}^m$ is continuous on X, it is uniformly continuous: for every $\epsilon > 0$ there exists $\delta > 0$ such that*

$$x, y \in X, |x - y| < \delta \Rightarrow |f(x) - f(y)| < \epsilon.$$

Proof. Assume, on the contrary to what should be proved, that there exists $\epsilon > 0$ such that for every $\delta > 0$ one can find a pair of points x, y in X such that $|x - y| < \delta$ and $|f(x) - f(y)| \geq \epsilon$. In particular, for every $i = 1, 2, ...$ we can find two points x^i, y^i in X such that $|x^i - y^i| \leq 1/i$ and $|f(x^i) - f(y^i)| \geq \epsilon$. By Theorem A.4.2, we can extract from the sequence $\{x^i\}$ a subsequence $\{x^{i_j}\}_{j=1}^{\infty}$ which converges to certain point $\bar{x} \in X$. Since $|y^{i_j} - x^{i_j}| \leq 1/i_j \to 0$ as $j \to \infty$, the sequence $\{y^{i_j}\}_{j=1}^{\infty}$ converges to the same point $\bar{x}$ as the sequence $\{x^{i_j}\}_{j=1}^{\infty}$ (why?) Since f is continuous, we have

$$\lim_{j\to\infty} f(y^{i_j}) = f(\bar{x}) = \lim_{j\to\infty} f(x^{i_j}),$$

whence $\lim\limits_{j\to\infty} (f(x^{i_j}) - f(y^{i_j})) = 0$, which contradicts the fact that $|f(x^{i_j}) - f(y^{i_j})| \geq \epsilon > 0$ for all j.

(iii) *Let f be a real-valued continuous function on X. The f attains its minimum on X:*

$$\operatorname*{Argmin}_X f \equiv \{x \in X : f(x) = \inf_{y\in X} f(y)\} \neq \emptyset,$$

same as f attains its maximum at certain points of X:

$$\operatorname*{Argal}_X f \equiv \{x \in X : f(x) = \sup_{y\in X} f(y)\} \neq \emptyset.$$

Proof: Let us prove that f attains its maximum on X (the proof for minimum is completely similar). Since f is bounded on X by (i), the quantity

$$f^* = \sup_{x\in X} f(x)$$

is finite; of course, we can find a sequence $\{x^i\}$ of points from X such that $f^* = \lim\limits_{i\to\infty} f(x^i)$. By Theorem A.4.2, we can extract from the sequence $\{x^i\}$ a subsequence $\{x^{i_j}\}_{j=1}^{\infty}$ which converges to certain point $\bar{x} \in X$. Since f is continuous on X, we have

$$f(\bar{x}) = \lim_{j\to\infty} f(x^{i_j}) = \lim_{i\to\infty} f(x^i) = f^*,$$

so that the maximum of f on X indeed is achieved (e.g., at the point $\bar{x}$).

Exercise A.6 *Prove that in general no one of the three statements in Theorem A.5.2 remains valid when X is closed, but not bounded, same as when X is bounded, but not closed.*

A.6 Differentiable functions on $\mathbf{R}^n$

A.6.1 The derivative

The reader definitely is familiar with the notion of derivative of a real-valued function $f(x)$ of real variable x:

$$f'(x) = \lim_{\Delta x \to 0} \frac{f(x+\Delta x) - f(x)}{\Delta x}$$

This definition does not work when we pass from functions of single real variable to functions of several real variables, or, which is the same, to functions with vector arguments. Indeed, in this case the shift in the argument Δx should be a vector, and we do not know what does it mean to *divide* by a vector...

A proper way to extend the notion of the derivative to real- and vector-valued functions of vector argument is to realize what in fact is the meaning of the derivative in the univariate case. What $f'(x)$ says to us is *how to approximate f in a neighbourhood of x by a linear function*. Specifically, *if $f'(x)$ exists, then the linear function $f'(x)\Delta x$ of Δx approximates the change $f(x+\Delta x) - f(x)$ in f up to a remainder which is of highest order as compared with Δx as $\Delta x \to 0$:*

$$|f(x+\Delta x) - f(x) - f'(x)\Delta x| \leq \bar{o}(|\Delta x|) \text{ as } \Delta x \to 0.$$

In the above formula, we meet with the notation $\bar{o}(|\Delta x|)$, and here is the explanation of this notation:

> *$\bar{o}(|\Delta x|)$ is a common name of all functions $\phi(\Delta x)$ of Δx which are well-defined in a neighbourhood of the point $\Delta x = 0$ on the axis, vanish at the point $\Delta x = 0$ and are such that*
>
> $$\frac{\phi(\Delta x)}{|\Delta x|} \to 0 \text{ as } \Delta x \to 0.$$
>
> For example,
>
> 1. $(\Delta x)^2 = \bar{o}(|\Delta x|)$, $\Delta x \to 0$,
> 2. $|\Delta x|^{1.01} = \bar{o}(|\Delta x|)$, $\Delta x \to 0$,
> 3. $\sin^2(\Delta x) = \bar{o}(|\Delta x|)$, $\Delta x \to 0$,
> 4. $\Delta x \neq \bar{o}(|\Delta x|)$, $\Delta x \to 0$.
>
> Later on we shall meet with the notation "$\bar{o}(|\Delta x|^k)$ as $\Delta x \to 0$", where k is a positive integer. The definition is completely similar to the one for the case of $k = 1$:
>
> *$\bar{o}(|\Delta x|^k)$ is a common name of all functions $\phi(\Delta x)$ of Δx which are well-defined in a neighbourhood of the point $\Delta x = 0$ on the axis, vanish at the point $\Delta x = 0$ and are such that*
>
> $$\frac{\phi(\Delta x)}{|\Delta x|^k} \to 0 \text{ as } \Delta x \to 0.$$

Note that if $f(\cdot)$ is a function defined in a neighbourhood of a point x on the axis, then there perhaps are many linear functions $a\Delta x$ of Δx which well approximate $f(x+\Delta x) - f(x)$, in the sense that the remainder in the approximation

$$f(x+\Delta x) - f(x) - a\Delta x$$

tends to 0 as $\Delta x \to 0$; among these approximations, however, there exists *at most one* which approximates $f(x+\Delta x) - f(x)$ "very well" – so that the remainder is $\bar{o}(|\Delta x|)$, and not merely tends to 0 as $\Delta x \to 0$. Indeed, if

$$f(x+\Delta x) - f(x) - a\Delta x = \bar{o}(|\Delta x|),$$

then, dividing both sides by Δx, we get

$$\frac{f(x+\Delta x) - f(x)}{\Delta x} - a = \frac{\bar{o}(|\Delta x|)}{\Delta x};$$

by definition of $\bar{o}(\cdot)$, the right hand side in this equality tends to 0 as $\Delta x \to 0$, whence

$$a = \lim_{\Delta x \to 0} \frac{f(x+\Delta x) - f(x)}{\Delta x} = f'(x).$$

Thus, *if* a linear function $a\Delta x$ of Δx approximates the change $f(x+\Delta x)-f(x)$ in f up to the remainder which is $\bar{o}(|\Delta x|)$ as $\Delta x \to 0$, *then* a is the derivative of f at x. You can easily verify that the inverse statement also is true: *if* the derivative of f at x exists, *then* the linear function $f'(x)\Delta x$ of Δx approximates the change $f(x+\Delta x)-f(x)$ in f up to the remainder which is $\bar{o}(|\Delta x|)$ as $\Delta x \to 0$.

The advantage of the "$\bar{o}(|\Delta x|)$"-definition of derivative is that it can be naturally extended onto vector-valued functions of vector arguments (you should just replace "axis" with $\mathbf{R}^n$ in the definition of $\bar{o}$) and enlightens the *essence* of the notion of derivative: when it exists, this is exactly *the linear function of* Δx *which approximates the change* $f(x+\Delta x)-f(x)$ *in* f *up to a remainder which is* $\bar{o}(|\Delta x|)$. The precise definition is as follows:

Definition A.6.1 [Frechet differentiability] *Let f be a function which is well-defined in a neighbourhood of a point $x \in \mathbf{R}^n$ and takes values in $\mathbf{R}^m$. We say that f is differentiable at x, if there exists a linear function $Df(x)[\Delta x]$ of $\Delta x \in \mathbf{R}^n$ taking values in $\mathbf{R}^m$ which approximates the change $f(x+\Delta x)-f(x)$ in f up to a remainder which is $\bar{o}(|\Delta x|)$:*

$$|f(x+\Delta x)-f(x)-Df(x)[\Delta x]| \le \bar{o}(|\Delta x|). \tag{A.6.1}$$

Equivalently: a function f which is well-defined in a neighbourhood of a point $x \in \mathbf{R}^n$ and takes values in $\mathbf{R}^m$ is called differentiable at x, if there exists a linear function $Df(x)[\Delta x]$ of $\Delta x \in \mathbf{R}^n$ taking values in $\mathbf{R}^m$ such that for every $\epsilon > 0$ there exists $\delta > 0$ satisfying the relation

$$|\Delta x| \le \delta \Rightarrow |f(x+\Delta x)-f(x)-Df(x)[\Delta x]| \le \epsilon|\Delta x|.$$

A.6.2 Derivative and directional derivatives

We have defined what does it mean that a function $f: \mathbf{R}^n \to \mathbf{R}^m$ is differentiable at a point x, but did not say yet what is the derivative. The reader could guess that the derivative is exactly the "linear function $Df(x)[\Delta x]$ of $\Delta x \in \mathbf{R}^n$ taking values in $\mathbf{R}^m$ which approximates the change $f(x+\Delta x)-f(x)$ in f up to a remainder which is $\le \bar{o}(|\Delta x|)$" participating in the definition of differentiability. The guess is correct, but we cannot merely call the entity participating in the definition the derivative – why do we know that this entity is unique? Perhaps there are many different linear functions of Δx approximating the change in f up to a remainder which is $\bar{o}(|\Delta x|)$. In fact there is no more than a single linear function with this property due to the following observation:

> *Let f be differentiable at x, and $Df(x)[\Delta x]$ be a linear function participating in the definition of differentiability. Then*
>
> $$\forall \Delta x \in \mathbf{R}^n: \quad Df(x)[\Delta x] = \lim_{t\to+0} \frac{f(x+t\Delta x)-f(x)}{t}. \tag{A.6.2}$$
>
> *In particular,* <u>*the derivative*</u> *$Df(x)[\cdot]$ is uniquely defined by f and x.*
>
> **Proof.** We have
>
> $$\begin{array}{cl} |f(x+t\Delta x)-f(x)-Df(x)[t\Delta x]| \le \bar{o}(|t\Delta x|) & \\ \Downarrow & \\ |\frac{f(x+t\Delta x)-f(x)}{t} - \frac{Df(x)[t\Delta x]}{t}| \le \frac{\bar{o}(|t\Delta x|)}{t} & \\ \Updownarrow & [\text{since } Df(x)[\cdot] \text{ is linear}] \\ |\frac{f(x+t\Delta x)-f(x)}{t} - Df(x)[\Delta x]| \le \frac{\bar{o}(|t\Delta x|)}{t} & \\ \Downarrow & \\ Df(x)[\Delta x] = \lim\limits_{t\to+0} \frac{f(x+t\Delta x)-f(x)}{t} & \left[\begin{array}{l}\text{passing to limit as } t\to+0; \\ \text{note that } \frac{\bar{o}(|t\Delta x|)}{t} \to 0, t\to+0\end{array}\right] \end{array}$$

Pay attention to important remarks as follows:

1. The right hand side limit in (A.6.2) is an important entity called the *directional derivative of f taken at x along (a direction) Δx*; note that this quantity is defined in the "purely univariate" fashion – by dividing the change in f by the magnitude of a shift in a direction Δx and passing to limit as the magnitude of the shift approaches 0. Relation (A.6.2) says that the derivative, if exists, is, at every Δx, nothing that the directional derivative of f taken at x along Δx. Note, however, that

differentiability is much more than the existence of directional derivatives along all directions Δx; differentiability requires also *the directional derivatives to be "well-organized" – to depend linearly on the direction* Δx. It is easily seen that just existence of directional derivatives does not imply their "good organization": for example, the Euclidean norm

$$f(x) = |x|$$

at $x = 0$ possesses directional derivatives along all directions:

$$\lim_{t\to+0} \frac{f(0+t\Delta x) - f(0)}{t} = |\Delta x|;$$

these derivatives, however, depend *non-linearly* on Δx, so that the Euclidean norm is *not* differentiable at the origin (although is differentiable everywhere outside the origin, but this is another story).

2. It should be stressed that the derivative, if exists, is what it is: *a linear function of* $\Delta x \in \mathbf{R}^n$ *taking values in* $\mathbf{R}^m$. As we shall see in a while, we can *represent* this function by something "tractable", like a vector or a matrix, and can understand how to compute such a representation; however, an intelligent reader should bear in mind that *a* representation is not exactly the same as *the* represented entity. Sometimes the difference between derivatives and the entities which represent them is reflected in the terminology: what we call the *derivative*, is also called the *differential*, while the word "derivative" is reserved for the vector/matrix representing the differential.

A.6.3 Representations of the derivative

indexderivatives!representation ofBy definition, the derivative of a mapping $f : \mathbf{R}^n \to \mathbf{R}^m$ at a point x is a linear function $Df(x)[\Delta x]$ taking values in $\mathbf{R}^m$. How could we represent such a function?

Case of $m = 1$ – the gradient. Let us start with real-valued functions (i.e., with the case of $m = 1$); in this case the derivative is a *linear* real-valued function on $\mathbf{R}^n$. As we remember, the standard Euclidean structure on $\mathbf{R}^n$ allows to represent every linear function on $\mathbf{R}^n$ as the inner product of the argument with certain fixed vector. In particular, the derivative $Df(x)[\Delta x]$ of a scalar function can be represented as

$$Df(x)[\Delta x] = [\text{vector}]^T \Delta x;$$

what is denoted "vector" in this relation, is called the *gradient* of f at x and is denoted by $\nabla f(x)$:

$$Df(x)[\Delta x] = (\nabla f(x))^T \Delta x. \tag{A.6.3}$$

How to compute the gradient? The answer is given by (A.6.2). Indeed, let us look what (A.6.3) and (A.6.2) say when Δx is the i-th standard basic orth. According to (A.6.3), $Df(x)[e_i]$ is the i-th coordinate of the vector $\nabla f(x)$; according to (A.6.2),

$$\left.\begin{array}{r} Df(x)[e_i] = \lim\limits_{t\to+0} \frac{f(x+te_i)-f(x)}{t}, \\ Df(x)[e_i] = -Df(x)[-e_i] = -\lim\limits_{t\to+0} \frac{f(x-te_i)-f(x)}{t} = \lim\limits_{t\to-0} \frac{f(x+te_i)-f(x)}{t} \end{array}\right\} \Rightarrow Df(x)[e_i] = \frac{\partial f(x)}{\partial x_i}.$$

Thus,

If a real-valued function f is differentiable at x, then the first order partial derivatives of f at x exist, and the gradient of f at x is just the vector with the coordinates which are the first order partial derivatives of f taken at x:

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix}.$$

The derivative of f, taken at x, is the linear function of Δx given by

$$Df(x)[\Delta x] = (\nabla f(x))^T \Delta x = \sum_{i=1}^{n} \frac{\partial f(x)}{\partial x_i} (\Delta x)_i.$$

General case – the Jacobian. Now let $f : \mathbf{R}^n \to \mathbf{R}^m$ with $m \geq 1$. In this case, $Df(x)[\Delta x]$, regarded as a function of Δx, is a linear mapping from $\mathbf{R}^n$ to $\mathbf{R}^m$; as we remember, the standard way to represent a linear mapping from $\mathbf{R}^n$ to $\mathbf{R}^m$ is to represent it as the multiplication by $m \times n$ matrix:

$$Df(x)[\Delta x] = [m \times n \text{ matrix}] \cdot \Delta x. \tag{A.6.4}$$

What is denoted by "matrix" in (A.6.4), is called *the Jacobian* of f at x and is denoted by $f'(x)$. How to compute the entries of the Jacobian? Here again the answer is readily given by (A.6.2). Indeed, on one hand, we have

$$Df(x)[\Delta x] = f'(x)\Delta x, \tag{A.6.5}$$

whence

$$[Df(x)[e_j]]_i = (f'(x))_{ij},\ i = 1, ..., m, j = 1, ..., n.$$

On the other hand, denoting

$$f(x) = \begin{bmatrix} f_1(x) \\ \vdots \\ f_m(x) \end{bmatrix},$$

the same computation as in the case of gradient demonstrates that

$$[Df(x)[e_j]]_i = \frac{\partial f_i(x)}{\partial x_j}$$

and we arrive at the following conclusion:

If a vector-valued function $f(x) = (f_1(x), ..., f_m(x))$ is differentiable at x, then the first order partial derivatives of all f_i at x exist, and the Jacobian of f at x is just the $m \times n$ matrix with the entries $[\frac{\partial f_i(x)}{\partial x_j}]_{i,j}$ (so that the rows in the Jacobian are $[\nabla f_1(x)]^T$,..., $[\nabla f_m(x)]^T$. The derivative of f, taken at x, is the linear vector-valued function of Δx given by

$$Df(x)[\Delta x] = f'(x)\Delta x = \begin{bmatrix} [\nabla f_1(x)]^T \Delta x \\ \vdots \\ [\nabla f_m(x)]^T \Delta x \end{bmatrix}.$$

Remark A.6.1 Note that for a real-valued function f we have defined both the gradient $\nabla f(x)$ and the Jacobian $f'(x)$. These two entities are "nearly the same", but not exactly the same: the Jacobian is a vector-row, and the gradient is a vector-column linked by the relation

$$f'(x) = (\nabla f(x))^T.$$

Of course, both these representations of the derivative of f yield the same linear approximation of the change in f:

$$Df(x)[\Delta x] = (\nabla f(x))^T \Delta x = f'(x)\Delta x.$$

A.6.4 Existence of the derivative

We have seen that the existence of the derivative of f at a point implies the existence of the first order partial derivatives of the (components $f_1, ..., f_m$ of) f. The inverse statement is not exactly true – the existence of all first order partial derivatives $\frac{\partial f_i(x)}{\partial x_j}$ not necessarily implies the existence of the derivative; we need a bit more:

Theorem A.6.1 [Sufficient condition for differentiability] *Assume that*

1. *The mapping $f = (f_1, ..., f_m) : \mathbf{R}^n \to \mathbf{R}^m$ is well-defined in a neighbourhood U of a point $x_0 \in \mathbf{R}^n$,*
2. *The first order partial derivatives of the components f_i of f exist everywhere in U,*

   and
3. *The first order partial derivatives of the components f_i of f are continuous at the point x_0.*

Then f is differentiable at the point x_0.

A.6.5 Calculus of derivatives

The calculus of derivatives is given by the following result:

Theorem A.6.2 (i) [Differentiability and linear operations] *Let $f_1(x)$, $f_2(x)$ be mappings defined in a neighbourhood of $x_0 \in \mathbf{R}^n$ and taking values in $\mathbf{R}^m$, and $\lambda_1(x), \lambda_2(x)$ be real-valued functions defined in a neighbourhood of x_0. Assume that $f_1, f_2, \lambda_1, \lambda_2$ are differentiable at x_0. Then so is the function $f(x) = \lambda_1(x)f_1(x) + \lambda_2(x)f_2(x)$, with the derivative at x_0 given by*

$$\begin{array}{rcl} Df(x_0)[\Delta x] & = & [D\lambda_1(x_0)[\Delta x]]f_1(x_0) + \lambda_1(x_0)Df_1(x_0)[\Delta x] \\ & & +[D\lambda_2(x_0)[\Delta x]]f_2(x_0) + \lambda_2(x_0)Df_2(x_0)[\Delta x] \\ & & \Downarrow \\ f'(x_0) & = & f_1(x_0)[\nabla\lambda_1(x_0)]^T + \lambda_1(x_0)f_1'(x_0) \\ & & +f_2(x_0)[\nabla\lambda_2(x_0)]^T + \lambda_2(x_0)f_2'(x_0). \end{array}$$

(ii) [chain rule] *Let a mapping $f : \mathbf{R}^n \to \mathbf{R}^m$ be differentiable at x_0, and a mapping $g : \mathbf{R}^m \to \mathbf{R}^n$ be differentiable at $y_0 = f(x_0)$. Then the superposition $h(x) = g(f(x))$ is differentiable at x_0, with the derivative at x_0 given by*

$$\begin{array}{c} Dh(x_0)[\Delta x] = Dg(y_0)[Df(x_0)[\Delta x]] \\ \Downarrow \\ h'(x_0) = g'(y_0)f'(x_0) \end{array}$$

If the outer function g is real-valued, then the latter formula implies that

$$\nabla h(x_0) = [f'(x_0)]^T \nabla g(y_0)$$

(recall that for a real-valued function ϕ, $\phi' = (\nabla\phi)^T$).

A.6.6 Computing the derivative

Representations of the derivative via first order partial derivatives normally allow to compute it by the standard Calculus rules, in a completely mechanical fashion, not thinking at all of *what* we are computing. The examples to follow (especially the third of them) demonstrate that it often makes sense to bear in mind *what* is the derivative; this sometimes yield the result much faster than blind implementing Calculus rules.

Example 1: The gradient of an affine function. An *affine* function

$$f(x) = a + \sum_{i=1}^n g_i x_i \equiv a + g^T x : \mathbf{R}^n \to \mathbf{R}$$

is differentiable at every point (Theorem A.6.1) and its gradient, of course, equals g:

$$\begin{array}{rcll} (\nabla f(x))^T \Delta x & = & \lim\limits_{t\to+0} t^{-1}[f(x+t\Delta x) - f(x)] & [(\text{A.6.2})] \\ & = & \lim\limits_{t\to+0} t^{-1}[tg^T\Delta x] & [\text{arithmetics}] \end{array}$$

and we arrive at

$$\boxed{\nabla(a + g^T x) = g}$$

Example 2: The gradient of a quadratic form. For the time being, let us define a homogeneous quadratic form on $\mathbf{R}^n$ as a function

$$f(x) = \sum_{i,j} A_{ij}x_i x_j = x^T A x,$$

where A is an $n \times n$ matrix. Note that the matrices A and A^T define the same quadratic form, and therefore the *symmetric* matrix $B = \frac{1}{2}(A + A^T)$ also produces the same quadratic form as A and A^T. It follows that we always may assume (and do assume from now on) that the matrix A producing the quadratic form in question is symmetric.

A quadratic form is a simple polynomial and as such is differentiable at every point (Theorem A.6.1). What is the gradient of f at a point x? Here is the computation:

$$\begin{array}{rcl}
(\nabla f(x))^T\Delta x & = & Df(x)[\Delta x]\\
& = & \lim\limits_{t\to+0}\left[(x+t\Delta x)^TA(x+t\Delta x)-x^TAx\right]\\
&& \hfill [(A.6.2)]\\
& = & \lim\limits_{t\to+0}\left[x^TAx+t(\Delta x)^TAx+tx^TA\Delta x+t^2(\Delta x)^TA\Delta x-x^TAx\right]\\
&& \hfill [\text{opening parentheses}]\\
& = & \lim\limits_{t\to+0} t^{-1}\left[2t(Ax)^T\Delta x+t^2(\Delta x)^TA\Delta x\right]\\
&& \hfill [\text{arithmetics + symmetry of } A]\\
& = & 2(Ax)^T\Delta x
\end{array}$$

We conclude that

$$\boxed{\nabla(x^TAx)=2Ax}$$

(recall that $A=A^T$).

Example 3: The derivative of the log-det barrier. Let us compute the derivative of the *log-det barrier* (playing an extremely important role in modern optimization)

$$F(X)=\ln\mathrm{Det}(X);$$

here X is an $n\times n$ matrix (or, if you prefer, n^2-dimensional vector). Note that $F(X)$ is well-defined and differentiable in a neighbourhood of every point $\bar X$ with positive determinant (indeed, $\mathrm{Det}(X)$ is a polynomial of the entries of X and thus – is everywhere continuous and differentiable with continuous partial derivatives, while the function $\ln(t)$ is continuous and differentiable on the positive ray; by Theorems A.5.1.(ii), A.6.2.(ii), F is differentiable at every X such that $\mathrm{Det}(X)>0$). The reader is kindly asked to try to find the derivative of F by the standard techniques; if the result will not be obtained in, say, 30 minutes, please look at the 8-line computation to follow (in this computation, $\mathrm{Det}(\bar X)>0$, and $G(X)=\mathrm{Det}(X)$):

$$\begin{array}{rll}
& DF(\bar X)[\Delta X] & \\
= & D\ln(G(\bar X))[DG(\bar X)[\Delta X]] & [\text{chain rule}]\\
= & G^{-1}(\bar X)DG(\bar X)[\Delta X] & [\ln'(t)=t^{-1}]\\
= & \mathrm{Det}^{-1}(\bar X)\lim\limits_{t\to+0}t^{-1}\left[\mathrm{Det}(\bar X+t\Delta X)-\mathrm{Det}(\bar X)\right] & [\text{definition of } G \text{ and (A.6.2)}]\\
= & \mathrm{Det}^{-1}(\bar X)\lim\limits_{t\to+0}t^{-1}\left[\mathrm{Det}(\bar X(I+t\bar X^{-1}\Delta X))-\mathrm{Det}(\bar X)\right] & \\
= & \mathrm{Det}^{-1}(\bar X)\lim\limits_{t\to+0}t^{-1}\left[\mathrm{Det}(\bar X)(\mathrm{Det}(I+t\bar X^{-1}\Delta X)-1)\right] & [\mathrm{Det}(AB)=\mathrm{Det}(A)\mathrm{Det}(B)]\\
= & \lim\limits_{t\to+0}t^{-1}\left[\mathrm{Det}(I+t\bar X^{-1}\Delta X)-1\right] & \\
= & \mathrm{Tr}(\bar X^{-1}\Delta X)=\sum\limits_{i,j}[\bar X^{-1}]_{ji}(\Delta X)_{ij} &
\end{array}$$

where the concluding equality

$$\lim_{t\to+0}t^{-1}[\mathrm{Det}(I+tA)-1]=\mathrm{Tr}(A)\equiv\sum_i A_{ii} \tag{A.6.6}$$

is immediately given by recalling what is the determinant of $I+tA$: this is a polynomial of t which is the sum of products, taken along all diagonals of a $n\times n$ matrix and assigned certain signs, of the entries of $I+tA$. At every one of these diagonals, except for the main one, there are at least two cells with the entries proportional to t, so that the corresponding products do not contribute to the constant and the linear in t terms in $\mathrm{Det}(I+tA)$ and thus do not affect the limit in (A.6.6). The only product which does contribute to the linear and the constant terms in $\mathrm{Det}(I+tA)$ is the product $(1+tA_{11})(1+tA_{22})...(1+tA_{nn})$ coming from the main diagonal; it is clear that in this product the constant term is 1, and the linear in t term is $t(A_{11}+...+A_{nn})$, and (A.6.6) follows.

A.6.7 Higher order derivatives

Let $f:\mathbf{R}^n\to\mathbf{R}^m$ be a mapping which is well-defined and differentiable at every point x from an open set U. The Jacobian of this mapping $J(x)$ is a mapping from $\mathbf{R}^n$ to the space $\mathbf{R}^{m\times n}$ matrices, i.e., is a mapping taking values in certain $\mathbf{R}^M$ ($M = mn$). The derivative of this mapping, if it exists, is called the *second derivative* of f; it again is a mapping from $\mathbf{R}^n$ to certain $\mathbf{R}^M$ and as such can be differentiable, and so on, so that we can speak about the second, the third, ... derivatives of a vector-valued function of vector argument. A *sufficient* condition for the existence of k derivatives of f in U is that f is C^k in U, i.e., that all partial derivatives of f of orders $\leq k$ exist and are continuous everywhere in U (cf. Theorem A.6.1).

We have explained what does it mean that f has k derivatives in U; note, however, that according to the definition, highest order derivatives at a point x are just long vectors; say, the second order derivative of a scalar function f of 2 variables is the Jacobian of the mapping $x \mapsto f'(x) : \mathbf{R}^2 \to \mathbf{R}^2$, i.e., a mapping from $\mathbf{R}^2$ to $\mathbf{R}^{2\times 2} = \mathbf{R}^4$; the third order derivative of f is therefore the Jacobian of a mapping from $\mathbf{R}^2$ to $\mathbf{R}^4$, i.e., a mapping from $\mathbf{R}^2$ to $\mathbf{R}^{4\times 2} = \mathbf{R}^8$, and so on. The question which should be addressed now is: *What is a natural and transparent way to represent the highest order derivatives?*

The answer is as follows:

> ($*$) *Let $f : \mathbf{R}^n \to \mathbf{R}^m$ be C^k on an open set $U \subset \mathbf{R}^n$. The derivative of order $\ell \leq k$ of f, taken at a point $x \in U$, can be naturally identified with a function*
>
> $$D^\ell f(x)[\Delta x^1, \Delta x^2, ..., \Delta x^\ell]$$
>
> *of ℓ vector arguments $\Delta x^i \in \mathbf{R}^n$, $i = 1, ..., \ell$, and taking values in $\mathbf{R}^m$. This function is linear in every one of the arguments Δx^i, the other arguments being fixed, and is symmetric with respect to permutation of arguments $\Delta x^1, ..., \Delta x^\ell$.*
>
> *In terms of f, the quantity $D^\ell f(x)[\Delta x^1, \Delta x^2, ..., \Delta x^\ell]$ (full name: "the ℓ-th derivative (or differential) of f taken at a point x along the directions $\Delta x^1, ..., \Delta x^\ell$") is given by*
>
> $$D^\ell f(x)[\Delta x^1, \Delta x^2, ..., \Delta x^\ell] = \frac{\partial^\ell}{\partial t_\ell \partial t_{\ell-1} ... \partial t_1}\Big|_{t_1 = ... = t_\ell = 0} f(x + t_1 \Delta x^1 + t_2 \Delta x^2 + ... + t_\ell \Delta x^\ell). \tag{A.6.7}$$

The explanation to our claims is as follows. Let $f : \mathbf{R}^n \to \mathbf{R}^m$ be C^k on an open set $U \subset \mathbf{R}^n$.

1. When $\ell = 1$, ($*$) says to us that the first order derivative of f, taken at x, is a linear function $Df(x)[\Delta x^1]$ of $\Delta x^1 \in \mathbf{R}^n$, taking values in $\mathbf{R}^m$, and that the value of this function at every Δx^1 is given by the relation
$$Df(x)[\Delta x^1] = \frac{\partial}{\partial t_1}\Big|_{t_1=0} f(x + t_1 \Delta x^1) \tag{A.6.8}$$
(cf. (A.6.2)), which is in complete accordance with what we already know about the derivative.

2. To understand what is the second derivative, let us take the first derivative $Df(x)[\Delta x^1]$, *let us temporarily fix somehow the argument* Δx^1 and treat the derivative as a function of x. As a function of x, Δx^1 being fixed, the quantity $Df(x)[\Delta x^1]$ is again a mapping which maps U into $\mathbf{R}^m$ and is differentiable by Theorem A.6.1 (provided, of course, that $k \geq 2$). The derivative of this mapping is certain linear function of $\Delta x \equiv \Delta x^2 \in \mathbf{R}^n$, depending on x as on a parameter; and of course it depends on Δx^1 as on a parameter as well. Thus, the derivative of $Df(x)[\Delta x^1]$ in x is certain function
$$D^2 f(x)[\Delta x^1, \Delta x^2]$$
of $x \in U$ and $\Delta x^1, \Delta x^2 \in \mathbf{R}^n$ and taking values in $\mathbf{R}^m$. What we know about this function is that it is linear in Δx^2. In fact, it is also linear in Δx^1, since it is the derivative in x of certain function (namely, of $Df(x)[\Delta x^1]$) *linearly depending on the parameter* Δx^1, so that the derivative of the function *in* x is linear in the parameter Δx^1 as well (differentiation is a linear operation with respect to a function we are differentiating: summing up functions and multiplying them by real constants, we sum up, respectively, multiply by the same constants, the derivatives). Thus, $D^2 f(x)[\Delta x^1, \Delta x^2]$ is linear in Δx^1 when x and Δx^2 are fixed, and is linear in Δx^2 when x and Δx^1 are fixed. Moreover, we have
$$\begin{array}{rcll} D^2 f(x)[\Delta x^1, \Delta x^2] & = & \frac{\partial}{\partial t_2}\big|_{t_2=0} Df(x + t_2 \Delta x^2)[\Delta x^1] & \text{[cf. (A.6.8)]} \\ & = & \frac{\partial}{\partial t_2}\big|_{t_2=0} \frac{\partial}{\partial t_1}\big|_{t_1=0} f(x + t_2 \Delta x^2 + t_1 \Delta x^1) & \text{[by (A.6.8)]} \\ & = & \frac{\partial^2}{\partial t_2 \partial t_1}\Big|_{t_1=t_2=0} f(x + t_1 \Delta x^1 + t_2 \Delta x^2) & \end{array} \tag{A.6.9}$$

as claimed in (A.6.7) for $\ell = 2$. The only piece of information about the second derivative which is contained in $(*)$ and is not justified yet is that $D^2 f(x)[\Delta x^1, \Delta x^2]$ is symmetric in $\Delta x^1, \Delta x^2$; but this fact is readily given by the representation (A.6.7), since, as they prove in Calculus, if a function ϕ possesses *continuous* partial derivatives of orders $\leq \ell$ in a neighbourhood of a point, then these derivatives in this neighbourhood are independent of the order in which they are taken; it follows that

$$\begin{array}{rcll} D^2 f(x)[\Delta x^1, \Delta x^2] & = & \frac{\partial^2}{\partial t_2 \partial t_1}\Big|_{t_1=t_2=0} \underbrace{f(x + t_1\Delta x^1 + t_2 \Delta x^2)}_{\phi(t_1,t_2)} & [\text{(A.6.9)}] \\ & = & \frac{\partial^2}{\partial t_1 \partial t_2}\Big|_{t_1=t_2=0} \phi(t_1, t_2) & \\ & = & \frac{\partial^2}{\partial t_1 \partial t_2}\Big|_{t_1=t_2=0} f(x + t_2\Delta x^2 + t_1 \Delta x^1) & \\ & = & D^2 f(x)[\Delta x^2, \Delta x^1] & [\text{the same (A.6.9)}] \end{array}$$

3. Now it is clear how to proceed: to define $D^3 f(x)[\Delta x^1, \Delta x^2, \Delta x^3]$, we fix in the second order derivative $D^2 f(x)[\Delta x^1, \Delta x^2]$ the arguments $\Delta x^1, \Delta x^2$ and treat it as a function of x only, thus arriving at a mapping which maps U into $\mathbf{R}^m$ and depends on $\Delta x^1, \Delta x^2$ as on parameters (linearly in every one of them). Differentiating the resulting mapping in x, we arrive at a function $D^3 f(x)[\Delta x^1, \Delta x^2, \Delta x^3]$ which by construction is linear in every one of the arguments Δx^1, Δx^2, Δx^3 and satisfies (A.6.7); the latter relation, due to the Calculus result on the symmetry of partial derivatives, implies that $D^3 f(x)[\Delta x^1, \Delta x^2, \Delta x^3]$ is symmetric in $\Delta x^1, \Delta x^2, \Delta x^3$. After we have at our disposal the third derivative $D^3 f$, we can build from it in the already explained fashion the fourth derivative, and so on, until k-th derivative is defined.

Remark A.6.2 Since $D^\ell f(x)[\Delta x^1, ..., \Delta x^\ell]$ is linear in every one of Δx^i, we can expand the derivative in a multiple sum:

$$\begin{array}{rcl} & & \Delta x^i = \sum\limits_{j=1}^n \Delta x^i_j e_j \\ & & \Downarrow \\ D^\ell f(x)[\Delta x^1, ..., \Delta x^\ell] & = & D^\ell f(x)[\sum\limits_{j_1=1}^n \Delta x^1_{j_1} e_{j_1}, ..., \sum\limits_{j_\ell=1}^n \Delta x^\ell_{j_\ell} e_{j_\ell}] \\ & = & \sum\limits_{1 \leq j_1,...,j_\ell \leq n} D^\ell f(x)[e_{j_1}, ..., e_{j_\ell}] \Delta x^1_{j_1} ... \Delta x^\ell_{j_\ell} \end{array} \tag{A.6.10}$$

What is the origin of the coefficients $D^\ell f(x)[e_{j_1}, ..., e_{j_\ell}]$? According to (A.6.7), one has

$$\begin{array}{rcl} D^\ell f(x)[e_{j_1}, ..., e_{j_\ell}] & = & \frac{\partial^\ell}{\partial t_\ell \partial t_{\ell-1} ... \partial t_1}\Big|_{t_1=...=t_\ell=0} f(x + t_1 e_{j_1} + t_2 e_{j_2} + ... + t_\ell e_{j_\ell}) \\ & = & \frac{\partial^\ell}{\partial x_{j_\ell} \partial x_{j_{\ell-1}} ... \partial x_{j_1}} f(x). \end{array}$$

so that the coefficients in (A.6.10) are nothing but the partial derivatives, of order ℓ, of f.

Remark A.6.3 An important particular case of relation (A.6.7) is the one when $\Delta x^1 = \Delta x^2 = ... = \Delta x^\ell$; let us call the common value of these ℓ vectors d. According to (A.6.7), we have

$$D^\ell f(x)[d, d, ..., d] = \frac{\partial^\ell}{\partial t_\ell \partial t_{\ell-1} ... \partial t_1}\Big|_{t_1=...=t_\ell=0} f(x + t_1 d + t_2 d + ... + t_\ell d).$$

This relation can be interpreted as follows: consider the function

$$\phi(t) = f(x + td)$$

of a real variable t. Then (check it!)

$$\phi^{(\ell)}(0) = \frac{\partial^\ell}{\partial t_\ell \partial t_{\ell-1} ... \partial t_1}\Big|_{t_1=...=t_\ell=0} f(x + t_1 d + t_2 d + ... + t_\ell d) = D^\ell f(x)[d, ..., d].$$

In other words, $D^\ell f(x)[d, ..., d]$ is what is called *ℓ-th directional derivative of f taken at x along the direction d*; to define this quantity, we pass from function f of several variables to the univariate function $\phi(t) = f(x+td)$ – restrict f onto the line passing through x and directed by d – and then take the "usual" derivative of order ℓ of the resulting function of single real variable t at the point $t = 0$ (which corresponds to the point x of our line).

Representation of higher order derivatives. k-th order derivative $D^k f(x)[\cdot,...,\cdot]$ of a C^k function $f:\mathbf{R}^n\to\mathbf{R}m$ is what it is – it is a symmetric k-linear mapping on $\mathbf{R}^n$ taking values in $\mathbf{R}^m$ and depending on x as on a parameter. Choosing somehow coordinates in $\mathbf{R}^n$, we can represent such a mapping in the form

$$D^k f(x)[\Delta x_1,...,\Delta x_k]=\sum_{1\leq i_1,...,i_k\leq n}\frac{\partial^k f(x)}{\partial x_{i_k}\partial x_{i_{k-1}}...\partial x_{i_1}}(\Delta x_1)_{i_1}...(\Delta x_k)_{i_k}.$$

We may say that the derivative can be represented by k-index collection of m-dimensional vectors $\frac{\partial^k f(x)}{\partial x_{i_k}\partial x_{i_{k-1}}...\partial x_{i_1}}$. This collection, however, is a difficult-to-handle entity, so that such a representation does not help. There is, however, a case when the collection becomes an entity we know to handle; this is the case of the second-order derivative of a scalar function ($k=2,m=1$). In this case, the collection in question is just a symmetric matrix $H(x)=\left[\frac{\partial^2 f(x)}{\partial x_i\partial x_j}\right]_{1\leq i,j\leq n}$. This matrix is called the *Hessian* of f at x. Note that

$$D^2 f(x)[\Delta x_1,\Delta x_2]=\Delta x_1^T H(x)\Delta x_2.$$

A.6.8 Calculus of C^k mappings

The calculus of C^k mappings can be summarized as follows:

Theorem A.6.3 (i) *Let U be an open set in $\mathbf{R}^n$, $f_1(\cdot),f_2(\cdot):\mathbf{R}^n\to\mathbf{R}^m$ be C^k in U, and let real-valued functions $\lambda_1(\cdot),\lambda_2(\cdot)$ be C^k in U. Then the function*

$$f(x)=\lambda_1(x)f_1(x)+\lambda_2(x)f_2(x)$$

is C^k in U.

(ii) *Let U be an open set in $\mathbf{R}^n$, V be an open set in $\mathbf{R}^m$, let a mapping $f:\mathbf{R}^n\to\mathbf{R}^m$ be C^k in U and such that $f(x)\in V$ for $x\in U$, and, finally, let a mapping $g:\mathbf{R}^m\to\mathbf{R}^p$ be C^k in V. Then the superposition*

$$h(x)=g(f(x))$$

is C^k in U.

Remark A.6.4 For higher order derivatives, in contrast to the first order ones, there is no simple "chain rule" for computing the derivative of superposition. For example, the second-order derivative of the superposition $h(x)=g(f(x))$ of two C^2-mappings is given by the formula

$$Dh(x)[\Delta x^1,\Delta x^2]=Dg(f(x))[D^2 f(x)[\Delta x^1,\Delta x^2]]+D^2 g(x)[Df(x)[\Delta x^1],Df(x)[\Delta x^2]]$$

(check it!). We see that both the first- and the second-order derivatives of f and g contribute to the second-order derivative of the superposition h.

The only case when there does exist a simple formula for high order derivatives of a superposition is the case when the inner function is affine: if $f(x)=Ax+b$ and $h(x)=g(f(x))=g(Ax+b)$ with a C^ℓ mapping g, then

$$D^\ell h(x)[\Delta x^1,...,\Delta x^\ell]=D^\ell g(Ax+b)[A\Delta x^1,...,A\Delta x^\ell]. \tag{A.6.11}$$

A.6.9 Examples of higher-order derivatives

Example 1: Second-order derivative of an affine function $f(x)=a+b^Tx$ is, of course, identically zero. Indeed, as we have seen,

$$Df(x)[\Delta x^1]=b^T\Delta x^1$$

is independent of x, and therefore the derivative of $Df(x)[\Delta x^1]$ in x, which should give us the second derivative $D^2 f(x)[\Delta x^1,\Delta x^2]$, is zero. Clearly, the third, the fourth, etc., derivatives of an affine function are zero as well.

Example 2: Second-order derivative of a homogeneous quadratic form $f(x) = x^TAx$ (A is a symmetric $n \times n$ matrix). As we have seen,

$$Df(x)[\Delta x^1] = 2x^TA\Delta x^1.$$

Differentiating in x, we get

$$D^2f(x)[\Delta x^1, \Delta x^2] = \lim_{t\to+0} t^{-1}\left[2(x+t\Delta x^2)^TA\Delta x^1 - 2x^TA\Delta x^1\right] = 2(\Delta x^2)^TA\Delta x^1,$$

so that

$$\boxed{D^2f(x)[\Delta x^1, \Delta x^2] = 2(\Delta x^2)^TA\Delta x^1}$$

Note that the second derivative of a quadratic form is independent of x; consequently, the third, the fourth, etc., derivatives of a quadratic form are identically zero.

Example 3: Second-order derivative of the log-det barrier $F(X) = \ln \mathrm{Det}(X)$. As we have seen, this function of an $n\times n$ matrix is well-defined and differentiable on the set U of matrices with positive determinant (which is an open set in the space $\mathbf{R}^{n\times n}$ of $n \times n$ matrices). In fact, this function is C^∞ in U. Let us compute its second-order derivative. As we remember,

$$DF(X)[\Delta X^1] = \mathrm{Tr}(X^{-1}\Delta X^1). \tag{A.6.12}$$

To differentiate the right hand side in X, let us first find the derivative of the mapping $G(X) = X^{-1}$ which is defined on the open set of non-degenerate $n \times n$ matrices. We have

$$\begin{array}{rcl}
DG(X)[\Delta X] & = & \lim\limits_{t\to+0} t^{-1}\left[(X+t\Delta X)^{-1} - X^{-1}\right] \\
& = & \lim\limits_{t\to+0} t^{-1}\left[(X(I+tX^{-1}\Delta X))^{-1} - X^{-1}\right] \\
& = & \lim\limits_{t\to+0} t^{-1}\left[(I+t\underbrace{X^{-1}\Delta X}_{Y})^{-1}X^{-1} - X^{-1}\right] \\
& = & \left[\lim\limits_{t\to+0} t^{-1}\left[(I+tY)^{-1} - I\right]\right]X^{-1} \\
& = & \left[\lim\limits_{t\to+0} t^{-1}\left[I - (I+tY)\right](I+tY)^{-1}\right]X^{-1} \\
& = & \left[\lim\limits_{t\to+0} [-Y(I+tY)^{-1}]\right]X^{-1} \\
& = & -YX^{-1} \\
& = & -X^{-1}\Delta XX^{-1}
\end{array}$$

and we arrive at the important by its own right relation

$$\boxed{D(X^{-1})[\Delta X] = -X^{-1}\Delta XX^{-1}, \quad [X \in \mathbf{R}^{n\times n}, \mathrm{Det}(X) \neq 0]}$$

which is the “matrix extension” of the standard relation $(x^{-1})' = -x^{-2}$, $x \in \mathbf{R}$.

Now we are ready to compute the second derivative of the log-det barrier:

$$\begin{array}{rcl}
& F(X) = \ln\mathrm{Det}(X) & \\
& \Downarrow & \\
& DF(X)[\Delta X^1] = \mathrm{Tr}(X^{-1}\Delta X^1) & \\
& \Downarrow & \\
D^2F(X)[\Delta X^1, \Delta X^2] & = & \lim\limits_{t\to+0} t^{-1}\left[\mathrm{Tr}((X+t\Delta X^2)^{-1}\Delta X^1) - \mathrm{Tr}(X^{-1}\Delta X^1)\right] \\
& = & \lim\limits_{t\to+0} \mathrm{Tr}\left(t^{-1}(X+t\Delta X^2)^{-1}\Delta X^1 - X^{-1}\Delta X^1\right) \\
& = & \lim\limits_{t\to+0} \mathrm{Tr}\left(\left[t^{-1}(X+t\Delta X^2)^{-1} - X^{-1}\right]\Delta X^1\right) \\
& = & \mathrm{Tr}\left([-X^{-1}\Delta X^2X^{-1}]\Delta X^1\right),
\end{array}$$

and we arrive at the formula

$$\boxed{D^2F(X)[\Delta X^1, \Delta X^2] = -\mathrm{Tr}(X^{-1}\Delta X^2X^{-1}\Delta X^1) \quad [X \in \mathbf{R}^{n\times n}, \mathrm{Det}(X) > 0]}$$

Since $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$ (check it!) for all matrices A, B such that the product AB makes sense and is square, the right hand side in the above formula is symmetric in ΔX^1, ΔX^2, as it should be for the second derivative of a C^2 function.

A.6.10 Taylor expansion

Assume that $f : \mathbf{R}^n \to \mathbf{R}^m$ is C^k in a neighbourhood U of a point $\bar{x}$. The *Taylor expansion of order* k of f, built at the point $\bar{x}$, is the function

$$\begin{array}{rcl} F_k(x) & = & f(\bar{x}) + \frac{1}{1!}Df(\bar{x})[x-\bar{x}] + \frac{1}{2!}D^2f(\bar{x})[x-\bar{x},x-\bar{x}] \\ & & +\frac{1}{3!}D^2f(\bar{x})[x-\bar{x},x-\bar{x},x-\bar{x}] + ... + \frac{1}{k!}D^kf(\bar{x})[\underbrace{x-\bar{x},...,x-\bar{x}}_{k \text{ times}}] \end{array} \tag{A.6.13}$$

We are already acquainted with the Taylor expansion of order 1

$$F_1(x) = f(\bar{x}) + Df(\bar{x})[x-\bar{x}]$$

– this is the affine function of x which approximates "very well" $f(x)$ in a neighbourhood of $\bar{x}$, namely, within approximation error $\bar{o}(|x-\bar{x}|)$. Similar fact is true for Taylor expansions of higher order:

Theorem A.6.4 *Let $f : \mathbf{R}^n \to \mathbf{R}^m$ be C^k in a neighbourhood of $\bar{x}$, and let $F_k(x)$ be the Taylor expansion of f at $\bar{x}$ of degree k. Then*

(i) *$F_k(x)$ is a vector-valued polynomial of full degree $\leq k$ (i.e., every one of the coordinates of the vector $F_k(x)$ is a polynomial of $x_1, ..., x_n$, and the sum of powers of x_i's in every term of this polynomial does not exceed k);*

(ii) *$F_k(x)$ approximates $f(x)$ in a neighbourhood of $\bar{x}$ up to a remainder which is $\bar{o}(|x-\bar{x}|^k)$ as $x \to \bar{x}$:*

For every $\epsilon > 0$, there exists $\delta > 0$ such that

$$|x-\bar{x}| \leq \delta \Rightarrow |F_k(x) - f(x)| \leq \epsilon|x-\bar{x}|^k.$$

$F_k(\cdot)$ is the unique polynomial with components of full degree $\leq k$ which approximates f up to a remainder which is $\bar{o}(|x-\bar{x}|^k)$.

(iii) *The value and the derivatives of F_k of orders $1, 2, ..., k$, taken at $\bar{x}$, are the same as the value and the corresponding derivatives of f taken at the same point.*

As stated in Theorem, $F_k(x)$ approximates $f(x)$ for x close to $\bar{x}$ up to a remainder which is $\bar{o}(|x-\bar{x}|^k)$. In many cases, it is not enough to know that the reminder is "$\bar{o}(|x-\bar{x}|^k)$" — we need an explicit bound on this remainder. The standard bound of this type is as follows:

Theorem A.6.5 *Let k be a positive integer, and let $f : \mathbf{R}^n \to \mathbf{R}^m$ be C^{k+1} in a ball $B_r = B_r(\bar{x}) = \{x \in \mathbf{R}^n : |x-\bar{x}| < r\}$ of a radius $r > 0$ centered at a point $\bar{x}$. Assume that the directional derivatives of order $k+1$, taken at every point of B_r along every unit direction, do not exceed certain $L < \infty$:*

$$|D^{k+1}f(x)[d, ..., d]| \leq L \quad \forall(x \in B_r)\forall(d, |d| = 1).$$

Then for the Taylor expansion F_k of order k of f taken at $\bar{x}$ one has

$$|f(x) - F_k(x)| \leq \frac{L|x-\bar{x}|^{k+1}}{(k+1)!} \quad \forall(x \in B_r).$$

Thus, in a neighbourhood of $\bar{x}$ the remainder of the $\underline{k\text{-th order}}$ Taylor expansion, taken at $\bar{x}$, is of order of $L|x-\bar{x}|^{k+1}$, where L is the maximal (over all unit directions and all points from the neighbourhood) magnitude of the directional derivatives $\underline{\text{of order } k+1}$ of f.

A.7 Symmetric matrices

A.7.1 Spaces of matrices

Let $\mathbf{S}^m$ be the space of symmetric $m \times m$ matrices, and $\mathbf{M}^{m,n}$ be the space of rectangular $m \times n$ matrices with real entries. From the viewpoint of their linear structure (i.e., the operations of addition and multiplication by reals) $\mathbf{S}^m$ is just the arithmetic linear space $\mathbf{R}^{m(m+1)/2}$ of dimension $\frac{m(m+1)}{2}$: by arranging the elements of a symmetric $m \times m$ matrix X in a single column, say, in the row-by-row order, you get a usual m^2-dimensional column vector; multiplication of a matrix by a real and addition of matrices correspond to the same operations with the "representing vector(s)". When X runs through $\mathbf{S}^m$, the vector representing X runs through $m(m+1)/2$-dimensional subspace of $\mathbf{R}^{m^2}$ consisting of vectors satisfying the "symmetry condition" – the coordinates coming from symmetric to each other pairs of entries in X are equal to each other. Similarly, $\mathbf{M}^{m,n}$ as a linear space is just $\mathbf{R}^{mn}$, and it is natural to equip $\mathbf{M}^{m,n}$ with the inner product defined as the usual inner product of the vectors representing the matrices:

$$\langle X, Y\rangle = \sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}Y_{ij} = \mathrm{Tr}(X^TY).$$

Here Tr stands for the *trace* – the sum of diagonal elements of a (square) matrix. With this inner product (called the *Frobenius inner product*), $\mathbf{M}^{m,n}$ becomes a legitimate Euclidean space, and we may use in connection with this space all notions based upon the Euclidean structure, e.g., the (Frobenius) norm of a matrix

$$\|X\|_2 = \sqrt{\langle X, X\rangle} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}^2} = \sqrt{\mathrm{Tr}(X^TX)}$$

and likewise the notions of orthogonality, orthogonal complement of a linear subspace, etc. The same applies to the space $\mathbf{S}^m$ equipped with the Frobenius inner product; of course, the Frobenius inner product of symmetric matrices can be written without the transposition sign:

$$\langle X, Y\rangle = \mathrm{Tr}(XY),\;\; X, Y \in \mathbf{S}^m.$$

A.7.2 Main facts on symmetric matrices

Let us focus on the space $\mathbf{S}^m$ of symmetric matrices. The most important property of these matrices is as follows:

Theorem A.7.1 [Eigenvalue decomposition] *$n \times n$ matrix A is symmetric iff it admits an orthonormal system of eigenvectors: there exist orthonormal basis $\{e_1, ..., e_n\}$ such that*

$$Ae_i = \lambda_i e_i,\; i = 1, ..., n, \tag{A.7.1}$$

for reals λ_i.

In connection with Theorem A.7.1, it is worthy to recall the following notions and facts:

A.7.2.1 Eigenvectors and eigenvalues

An *eigenvector* of an $n \times n$ matrix A is a *nonzero* vector e (real or complex) such that $Ae = \lambda e$ for (real or complex) scalar λ; this scalar is called the *eigenvalue* of A *corresponding to the eigenvector* e.

Eigenvalues of A are exactly the roots of the *characteristic polynomial*

$$\pi(z) = \det(zI - A) = z^n + b_1 z^{n-1} + b_2 z^{n-2} + ... + b_n$$

of A.

Theorem A.7.1 states, in particular, that for a symmetric matrix A, all eigenvalues are real, and the corresponding eigenvectors can be chosen to be real and to form an orthonormal basis in $\mathbf{R}^n$.

A.7.2.2 Eigenvalue decomposition of a symmetric matrix

Theorem A.7.1 admits equivalent reformulation as follows (check the equivalence!):

Theorem A.7.2 *An $n \times n$ matrix A is symmetric iff it can be represented in the form*

$$A = U\Lambda U^T, \tag{A.7.2}$$

where

- *U is an orthogonal matrix: $U^{-1} = U^T$ (or, which is the same, $U^TU = I$, or, which is the same, $UU^T = I$, or, which is the same, the columns of U form an orthonormal basis in $\mathbf{R}^n$, or, which is the same, the columns of U form an orthonormal basis in $\mathbf{R}^n$).*

- *Λ is the diagonal matrix with the diagonal entries $\lambda_1, ..., \lambda_n$.*

Representation (A.7.2) with orthogonal U and diagonal Λ is called the *eigenvalue decomposition* of A. In such a representation,

- The columns of U form an orthonormal system of eigenvectors of A;
- The diagonal entries in Λ are the eigenvalues of A corresponding to these eigenvectors.

A.7.2.3 Vector of eigenvalues

When speaking about eigenvalues $\lambda_i(A)$ of a symmetric $n \times n$ matrix A, we always arrange them in the non-ascending order:

$$\lambda_1(A) \geq \lambda_2(A) \geq ... \geq \lambda_n(A);$$

$\lambda(A) \in \mathbf{R}^n$ denotes the vector of eigenvalues of A taken in the above order.

A.7.2.D. Freedom in eigenvalue decomposition. Part of the data Λ, U in the eigenvalue decomposition (A.7.2) is uniquely defined by A, while the other data admit certain "freedom". Specifically, the sequence $\lambda_1, ..., \lambda_n$ of eigenvalues of A (i.e., diagonal entries of Λ) is exactly the sequence of roots of the characteristic polynomial of A (every root is repeated according to its multiplicity) and thus is uniquely defined by A (provided that we arrange the entries of the sequence in the non-ascending order). The columns of U are not uniquely defined by A. What is uniquely defined, are the *linear spans* $E(\lambda)$ of the columns of U corresponding to all eigenvalues equal to certain λ; such a linear span is nothing but the *spectral subspace* $\{x : Ax = \lambda x\}$ of A corresponding to the eigenvalue λ. There are as many spectral subspaces as many different eigenvalues; spectral subspaces corresponding to different eigenvalues of symmetric matrix are orthogonal to each other, and their sum is the entire space. When building an orthogonal matrix U in the spectral decomposition, one chooses an orthonormal eigenbasis in the spectral subspace corresponding to the largest eigenvalue and makes the vectors of this basis the first columns in U, then chooses an orthonormal basis in the spectral subspace corresponding to the second largest eigenvalue and makes the vector from this basis the next columns of U, and so on.

A.7.2.E. "Simultaneous" decomposition of commuting symmetric matrices. Let $A_1, ..., A_k$ be $n \times n$ symmetric matrices. It turns out that *the matrices commute with each other ($A_iA_j = A_jA_i$ for all i, j) iff they can be "simultaneously diagonalized", i.e., there exist a single orthogonal matrix U and diagonal matrices Λ_1,...,Λ_k such that*

$$A_i = U\Lambda_i U^T,\ i = 1, ..., k.$$

You are welcome to prove this statement by yourself; to simplify your task, here are two simple and important by their own right statements which help to reach your target:

> **A.7.2.E.1:** *Let λ be a real and A, B be two commuting $n \times n$ matrices. Then the spectral subspace $E = \{x : Ax = \lambda x\}$ of A corresponding to λ is invariant for B (i.e., $Be \in E$ for every $e \in E$).*
>
> **A.7.2.E.2:** *If A is an $n \times n$ matrix and L is an invariant subspace of A (i.e., L is a linear subspace such that $Ae \in L$ whenever $e \in L$), then the orthogonal complement $L^\perp$ of L is invariant for the matrix A^T. In particular, if A is symmetric and L is invariant subspace of A, then $L^\perp$ is invariant subspace of A as well.*

A.7.3 Variational characterization of eigenvalues

Theorem A.7.3 [VCE – Variational Characterization of Eigenvalues] *Let A be a symmetric matrix. Then*

$$\lambda_\ell(A) = \min_{E \in \mathcal{E}_\ell} \max_{x \in E, x^T x = 1} x^T Ax,\ \ell = 1, ..., n, \tag{A.7.3}$$

where $\mathcal{E}_\ell$ is the family of all linear subspaces in $\mathbf{R}^n$ of the dimension $n - \ell + 1$.

VCE says that to get the largest eigenvalue $\lambda_1(A)$, you should maximize the quadratic form $x^T Ax$ over the unit sphere $S = \{x \in \mathbf{R}^n : x^T x = 1\}$; the maximum is exactly $\lambda_1(A)$. To get the second largest eigenvalue $\lambda_2(A)$, you should act as follows: you choose a linear subspace E of dimension $n-1$ and maximize the quadratic form $x^T Ax$ over the cross-section of S by this subspace; the maximum value of the form depends on E, and you minimize this maximum over linear subspaces E of the dimension $n-1$; the result is exactly $\lambda_2(A)$. To get $\lambda_3(A)$, you replace in the latter construction subspaces of the dimension $n-1$ by those of the dimension $n-2$, and so on. In particular, the smallest eigenvalue $\lambda_n(A)$ is just the minimum, over all linear subspaces E of the dimension $n-n+1=1$, i.e., over all lines passing through the origin, of the quantities $x^T Ax$, where $x \in E$ is unit ($x^T x = 1$); in other words, $\lambda_n(A)$ is just the minimum of the quadratic form $x^T Ax$ over the unit sphere S.

Proof of the VCE is pretty easy. Let $e_1, ..., e_n$ be an orthonormal eigenbasis of A: $Ae_\ell = \lambda_\ell(A)e_\ell$. For $1 \leq \ell \leq n$, let $F_\ell = \mathrm{Lin}\{e_1, ..., e_\ell\}$, $G_\ell = \mathrm{Lin}\{e_\ell, e_{\ell+1}, ..., e_n\}$. Finally, for $x \in \mathbf{R}^n$ let $\xi(x)$ be the vector of coordinates of x in the orthonormal basis $e_1, ..., e_n$. Note that

$$x^T x = \xi^T(x)\xi(x),$$

since $\{e_1, ..., e_n\}$ is an orthonormal basis, and that

$$\begin{array}{rcl} x^T Ax & = & x^T A \sum\limits_i \xi_i(x)e_i) = x^T \sum\limits_i \lambda_i(A)\xi_i(x)e_i = \\ \sum\limits_i \lambda_i(A)\xi_i(x) \underbrace{(x^T e_i)}_{\xi_i(x)} & & \\ & = & \sum\limits_i \lambda_i(A)\xi_i^2(x). \end{array} \tag{A.7.4}$$

Now, given ℓ, $1 \leq \ell \leq n$, let us set $E = G_\ell$; note that E is a linear subspace of the dimension $n-\ell+1$. In view of (A.7.4), the maximum of the quadratic form $x^T Ax$ over the intersection of our E with the unit sphere is

$$\max \left\{ \sum_{i=\ell}^n \lambda_i(A)\xi_i^2 : \sum_{i=\ell}^n \xi_i^2 = 1 \right\},$$

and the latter quantity clearly equals to $\max\limits_{\ell \leq i \leq n} \lambda_i(A) = \lambda_\ell(A)$. Thus, for appropriately chosen $E \in \mathcal{E}_\ell$, the inner maximum in the right hand side of (A.7.3) equals to $\lambda_\ell(A)$, whence the right hand side of (A.7.3) is $\leq \lambda_\ell(A)$. It remains to prove the opposite inequality. To this end, consider a linear subspace E of the dimension $n-\ell+1$ and observe that it has nontrivial intersection with the linear subspace F_ℓ of the dimension ℓ (indeed, $\dim E + \dim F_\ell = (n-\ell+1)+\ell > n$, so that $\dim(E \cap F) > 0$ by the Dimension formula). It follows that there exists a unit vector y belonging to both E and F_ℓ. Since y is a unit vector from F_ℓ, we have $y = \sum\limits_{i=1}^{\ell} \eta_i e_i$ with $\sum\limits_{i=1}^{\ell} \eta_i^2 = 1$, whence, by (A.7.4),

$$y^T Ay = \sum_{i=1}^{\ell} \lambda_i(A)\eta_i^2 \geq \min_{1 \leq i \leq \ell} \lambda_i(A) = \lambda_\ell(A).$$

Since y is in E, we conclude that

$$\max_{x \in E: x^T x = 1} x^T Ax \geq y^T Ay \geq \lambda_\ell(A).$$

Since E is an arbitrary subspace form $\mathcal{E}_\ell$, we conclude that the right hand side in (A.7.3) is $\geq \lambda_\ell(A)$. □

A simple and useful byproduct of our reasoning is the relation (A.7.4):

Corollary A.7.1 *For a symmetric matrix A, the quadratic form $x^T Ax$ is weighted sum of squares of the coordinates $\xi_i(x)$ of x taken with respect to an orthonormal eigenbasis of A; the weights in this sum are exactly the eigenvalues of A:*

$$x^T Ax = \sum_i \lambda_i(A)\xi_i^2(x).$$

A.7.3.1 Corollaries of the VCE

VCE admits a number of extremely important corollaries as follows: matrix A is called positive definite (notation: $A \succ 0$), if it is symmetric and the quadratic form x^TAx is positive outside the origin; A is called positive semidefinite (notation: $A \succeq 0$), if A is symmetric and the quadratic form x^TAx is nonnegative everywhere. VCE provides us with the following eigenvalue characterization of positive (semi)definite matrices:

Proposition A.7.1 : *A symmetric matrix A is positive semidefinite if and only if its eigenvalues are nonnegative; A is positive definite if and only if all eigenvalues of A are positive*

Indeed, A is positive definite, iff the minimum value of x^TAx over the unit sphere is positive, and is positive semidefinite, iff this minimum value is nonnegative; it remains to note that by VCE, the minimum value of x^TAx over the unit sphere is exactly the minimum eigenvalue of A.

A.7.3.2 $\succeq$-Monotonicity of the vector of eigenvalues

Let us write $A \succeq B$ ($A \succ B$) to express that A, B are symmetric matrices of the same size such that $A - B$ is positive semidefinite (respectively, positive definite).

Proposition A.7.2 *If $A \succeq B$, then $\lambda(A) \geq \lambda(B)$, and if $A \succ B$, then $\lambda(A) > \lambda(B)$.*

Indeed, when $A \succeq B$, then, of course,

$$\max_{x\in E: x^Tx=1} x^TAx \geq \max_{x\in E: x^Tx=1} x^TBx$$

for every linear subspace E, whence

$$\lambda_\ell(A) = \min_{E\in\mathcal{E}_\ell} \max_{x\in E: x^Tx=1} x^TAx \geq \min_{E\in\mathcal{E}_\ell} \max_{x\in E: x^Tx=1} x^TBx = \lambda_\ell(B),\ \ \ell = 1, ..., n,$$

i.e., $\lambda(A) \geq \lambda(B)$. The case of $A \succ B$ can be considered similarly.

A.7.3.3 Eigenvalue Interlacement Theorem

We shall formulate this extremely important theorem as follows:

Theorem A.7.4 [Eigenvalue Interlacement Theorem] *Let A be a symmetric $n \times n$ matrix and $\bar{A}$ be the angular $(n-k) \times (n-k)$ submatrix of A. Then, for every $\ell \leq n-k$, the ℓ-th eigenvalue of $\bar{A}$ separates the ℓ-th and the $(\ell+k)$-th eigenvalues of A:*

$$\lambda_\ell(A) \succeq \lambda_\ell(\bar{A}) \succeq \lambda_{\ell+k}(A). \tag{A.7.5}$$

Indeed, by VCE, $\lambda_\ell(\bar{A}) = \min_{E\in\bar{\mathcal{E}}_\ell} \max_{x\in E: x^Tx=1} x^TAx$, where $\bar{\mathcal{E}}_\ell$ is the family of all linear subspaces of the dimension $n-k-\ell+1$ contained in the linear subspace $\{x \in \mathbf{R}^n : x_{n-k+1} = x_{n-k+2} = ... = x_n = 0\}$. Since $\bar{\mathcal{E}}_\ell \subset \mathcal{E}_{\ell+k}$, we have

$$\lambda_\ell(\bar{A}) = \min_{E\in\bar{\mathcal{E}}_\ell} \max_{x\in E: x^Tx=1} x^TAx \geq \min_{E\in\mathcal{E}_{\ell+k}} \max_{x\in E: x^Tx=1} x^TAx = \lambda_{\ell+k}(A).$$

We have proved the left inequality in (A.7.5). Applying this inequality to the matrix $-A$, we get

$$-\lambda_\ell(\bar{A}) = \lambda_{n-k-\ell}(-\bar{A}) \geq \lambda_{n-\ell}(-A) = -\lambda_\ell(A),$$

or, which is the same, $\lambda_\ell(\bar{A}) \leq \lambda_\ell(A)$, which is the first inequality in (A.7.5).

A.7.4 Positive semidefinite matrices and the semidefinite cone

A.7.4.A. Positive semidefinite matrices. Recall that an $n\times n$ matrix A is called *positive semidefinite* (notation: $A\succeq 0$), if A is symmetric and produces nonnegative quadratic form:

$$A\succeq 0\Leftrightarrow\{A=A^T \quad\text{and}\quad x^TAx\geq 0\quad\forall x\}.$$

A is called positive *definite* (notation: $A\succ 0$), if it is positive semidefinite and the corresponding quadratic form is positive outside the origin:

$$A\succ 0\Leftrightarrow\{A=A^T \quad\text{and}\quad x^TAx>00\quad\forall x\neq 0\}.$$

It makes sense to list a number of equivalent definitions of a positive semidefinite matrix:

Theorem A.7.5 *Let A be a symmetric $n\times n$ matrix. Then the following properties of A are equivalent to each other:*

(i) $A\succeq 0$
(ii) $\lambda(A)\geq 0$
(iii) $A=D^TD$ *for certain rectangular matrix* D
(iv) $A=\Delta^T\Delta$ *for certain upper triangular* $n\times n$ *matrix* Δ
(v) $A=B^2$ *for certain symmetric matrix* B;
(vi) $A=B^2$ *for certain* $B\succeq 0$.

The following properties of a symmetric matrix A also are equivalent to each other:

(i′) $A\succ 0$
(ii′) $\lambda(A)>0$
(iii′) $A=D^TD$ *for certain rectangular matrix* D *of rank* n
(iv′) $A=\Delta^T\Delta$ *for certain nondegenerate upper triangular* $n\times n$ *matrix* Δ
(v′) $A=B^2$ *for certain nondegenerate symmetric matrix* B;
(vi′) $A=B^2$ *for certain* $B\succ 0$.

Proof. (i)⇔(ii): this equivalence is stated by Proposition A.7.1.

(ii)⇔(vi): Let $A=U\Lambda U^T$ be the eigenvalue decomposition of A, so that U is orthogonal and Λ is diagonal with nonnegative diagonal entries $\lambda_i(A)$ (we are in the situation of (ii) !). Let $\Lambda^{1/2}$ be the diagonal matrix with the diagonal entries $\lambda_i^{1/2}(A)$; note that $(\Lambda^{1/2})^2=\Lambda$. The matrix $B=U\Lambda^{1/2}U^T$ is symmetric with nonnegative eigenvalues $\lambda_i^{1/2}(A)$, so that $B\succeq 0$ by Proposition A.7.1, and

$$B^2=U\Lambda^{1/2}\underbrace{U^TU}_{I}\Lambda^{1/2}U^T=U(\Lambda^{1/2})^2U^T=U\Lambda U^T=A,$$

as required in (vi).

(vi)⇒(v): evident.

(v)⇒(iv): Let $A=B^2$ with certain symmetric B, and let b_i be i-th column of B. Applying the Gram-Schmidt orthogonalization process (see proof of Theorem A.2.3.(iii)), we can find an orthonormal system of vectors $u_1,...,u_n$ and lower triangular matrix L such that $b_i=\sum\limits_{j=1}^{i}L_{ij}u_j$, or, which is the same, $B^T=LU$, where U is the orthogonal matrix with the rows $u_1^T,...,u_n^T$. We now have $A=B^2=B^T(B^T)^T=LUU^TL^T=LL^T$. We see that $A=\Delta^T\Delta$, where the matrix $\Delta=L^T$ is upper triangular.

(iv)⇒(iii): evident.

(iii)⇒(i): If $A=D^TD$, then $x^TAx=(Dx)^T(Dx)\geq 0$ for all x.

We have proved the equivalence of the properties (i) – (vi). Slightly modifying the reasoning (do it yourself!), one can prove the equivalence of the properties (i′) – (vi′). □

Remark A.7.1 (i) [Checking positive semidefiniteness] Given an $n\times n$ symmetric matrix A, one can check whether it is positive semidefinite by a purely algebraic finite algorithm (the so called *Lagrange diagonalization of a quadratic form*) which requires at most $O(n^3)$ arithmetic operations. Positive definiteness of a matrix can be checked also by the *Choleski factorization algorithm* which finds the decomposition in (iv′), if it exists, in approximately $\frac{1}{6}n^3$ arithmetic operations.

There exists another useful algebraic criterion (Sylvester's criterion) for positive semidefiniteness of a matrix; according to this criterion, a symmetric matrix A is positive definite if and only if its angular minors are positive, and A is positive semidefinite if and only if all its principal minors are nonnegative. For example, a symmetric 2×2 matrix $A = \left[\begin{array}{cc} a & b \\ b & c \end{array}\right]$ is positive semidefinite iff $a \geq 0$, $c \geq 0$ and $\det(A) \equiv ac - b^2 \geq 0$.

(ii) [Square root of a positive semidefinite matrix] By the first chain of equivalences in Theorem A.7.5, a symmetric matrix A is $\succeq 0$ if and only if A is the square of a positive semidefinite matrix B. The latter matrix is uniquely defined by $A \succeq 0$ and is called the *square root* of A (notation: $A^{1/2}$).

A.7.4.1 The semidefinite cone

When adding symmetric matrices and multiplying them by reals, we add, respectively multiply by reals, the corresponding quadratic forms. It follows that

Observation A.7.1 *The sum of positive semidefinite matrices and a product of a positive semidefinite matrix and a nonnegative real is positive semidefinite*

or, which is the same (see Section B.1.4),

Observation A.7.2 *$n \times n$ positive semidefinite matrices form a cone $\mathbf{S}^n_+$ in the Euclidean space $\mathbf{S}^n$ of symmetric $n \times n$ matrices, the Euclidean structure being given by the Frobenius inner product $\langle A, B\rangle = \mathrm{Tr}(AB) = \sum\limits_{i,j} A_{ij}B_{ij}$.*

The cone $\mathbf{S}^n_+$ is called the *semidefinite* cone of size n. It is immediately seen that the semidefinite cone $\mathbf{S}^n_+$ is "good" (see Lecture 5), specifically,

- $\mathbf{S}^n_+$ is closed: the limit of a converging sequence of positive semidefinite matrices is positive semidefinite;
- $\mathbf{S}^n_+$ is pointed: the only $n \times n$ matrix A such that both A and $-A$ are positive semidefinite is the zero $n \times n$ matrix;
- $\mathbf{S}^n_+$ possesses a nonempty interior which is comprised of positive definite matrices.

Note that the relation $A \succeq B$ means exactly that $A - B \in \mathbf{S}^n_+$, while $A \succ B$ is equivalent to $A - B \in \mathrm{int}\,\mathbf{S}^n_+$. The "matrix inequalities" $A \succeq B$ ($A \succ B$) match the standard properties of the usual scalar inequalities, e.g.:

$$\begin{array}{lr}
A \succeq A & \text{[reflexivity]} \\
A \succeq B,\, B \succeq A \Rightarrow A = B & \text{[antisymmetry]} \\
A \succeq B,\, B \succeq C \Rightarrow A \succeq C & \text{[transitivity]} \\
A \succeq B,\, C \succeq D \Rightarrow A + C \succeq B + D & \text{[compatibility with linear operations, I]} \\
A \succeq B,\, \lambda \geq 0 \Rightarrow \lambda A \succeq \lambda B & \text{[compatibility with linear operations, II]} \\
A_i \succeq B_i, A_i \to A, B_i \to B \text{ as } i \to \infty \Rightarrow A \succeq B & \text{[closedness]}
\end{array}$$

with evident modifications when $\succeq$ is replaced with $\succ$, or

$$A \succeq B, C \succ D \Rightarrow A + C \succ B + D,$$

etc. Along with these standard properties of inequalities, the inequality $\succeq$ possesses a nice additional property:

Observation A.7.3 *In a valid $\succeq$-inequality*

$$A \succeq B$$

one can multiply both sides from the left and by the right by a (rectangular) matrix and its transpose:

$$\begin{array}{c}
A, B \in \mathbf{S}^n, \quad A \succeq B, \quad V \in \mathbf{M}^{n,m} \\
\Downarrow \\
V^T A V \succeq V^T B V
\end{array}$$

Indeed, we should prove that if $A - B \succeq 0$, then also $V^T(A - B)V \succeq 0$, which is immediate – the quadratic form $y^T[V^T(A-B)V]y = (Vy)^T(A-B)(Vy)$ of y is nonnegative along with the quadratic form $x^T(A-B)x$ of x.

An important additional property of the semidefinite cone is its *self-duality*:

Theorem A.7.6 *A symmetric matrix Y has nonnegative Frobenius inner products with all positive semidefinite matrices iff Y itself is positive semidefinite.*

Proof. “if” part: Assume that $Y \succeq 0$, and let us prove that then $\mathrm{Tr}(YX) \geq 0$ for every $X \succeq 0$. Indeed, the eigenvalue decomposition of Y can be written as

$$Y = \sum_{i=1}^{n} \lambda_i(Y) e_i e_i^T,$$

where e_i are the orthonormal eigenvectors of Y. We now have

$$\begin{array}{rcl}\mathrm{Tr}(YX) & = & \mathrm{Tr}((\sum\limits_{i=1}^{n} \lambda_i(Y) e_i e_i^T)X) = \sum\limits_{i=1}^{n} \lambda_i(Y)\mathrm{Tr}(e_i e_i^T X) \\ & = & \sum\limits_{i=1}^{n} \lambda_i(Y)\mathrm{Tr}(e_i^T X e_i),\end{array} \tag{A.7.6}$$

where the concluding equality is given by the following well-known property of the trace:

Observation A.7.4 *Whenever matrices A, B are such that the product AB makes sense and is a square matrix, one has*

$$\mathrm{Tr}(AB) = \mathrm{Tr}(BA).$$

Indeed, we should verify that if $A \in \mathbf{M}^{p,q}$ and $B \in \mathbf{M}^{q,p}$, then $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$. The left hand side quantity in our hypothetic equality is $\sum\limits_{i=1}^{p}\sum\limits_{j=1}^{q} A_{ij}B_{ji}$, and the right hand side quantity is $\sum\limits_{j=1}^{q}\sum\limits_{i=1}^{p} B_{ji}A_{ij}$; they indeed are equal.

Looking at the concluding quantity in (A.7.6), we see that it indeed is nonnegative whenever $X \succeq 0$ (since $Y \succeq 0$ and thus $\lambda_i(Y) \geq 0$ by P.7.5).

”only if” part: We are given Y such that $\mathrm{Tr}(YX) \geq 0$ for all matrices $X \succeq 0$, and we should prove that $Y \succeq 0$. This is immediate: for every vector x, the matrix $X = xx^T$ is positive semidefinite (Theorem A.7.5.(iii)), so that $0 \leq \mathrm{Tr}(Yxx^T) = \mathrm{Tr}(x^T Y x) = x^T Y x$. Since the resulting inequality $x^T Y x \geq 0$ is valid for every x, we have $Y \succeq 0$. □

A.8 Singular values and Singular Value Decomposition

It can be proved that every $m \times n$ matrix A admits *singular value decomposition*

$$A = UDV^T,$$

where U is an orthogonal $m \times m$ matrix, V is orthogonal $n \times n$ matrix, and D is diagonal $m \times n$ matrix, diagonality meaning that the only nonzero entries in D are the diagonal entries D_{ii}, $1 \leq i \leq \min[m, n]$; all these diagonal entries are nonnegative. Moreover, while A can admit various singular value decompositions, the diagonal entries in D are defined by A uniquely, up to permutation; they are called *singular values* of A. The collection of *nonzero* (i.e., positive) singular values of A is exactly the same as similar collection for A^T, and both these collections can be obtained by taking square roots of the positive eigenvalues of A^TA (or AA^T). In addition, the positive singular values of A are closely related to the eigenvalues of the symmetric matrix $A^+ = \left[\begin{array}{c|c} & A \\ \hline A^T & \end{array}\right]$; specifically, the collection of nonzero eigenvalues of A^+ is the union of two collections: one comprised by positive singular values of A, and the other one comprised by minus positive singular values of A.

Singular values of a symmetric matrix are just magnitudes of eignvalues of the matrix.

Appendix B

Convex sets in $\mathbf{R}^n$

B.1 Definition and basic properties

B.1.1 A convex set

In the school geometry a figure is called convex if it contains, along with every pair of its points x, y, also the entire segment $[x, y]$ linking the points. This is exactly the definition of a convex set in the multidimensional case; all we need is to say what does it mean "the segment $[x, y]$ linking the points $x, y \in \mathbf{R}^n$". This is said by the following

Definition B.1.1 [Convex set]
1) *Let x, y be two points in $\mathbf{R}^n$. The set*

$$[x, y] = \{z = \lambda x + (1 - \lambda)y : 0 \leq \lambda \leq 1\}$$

is called a segment with the endpoints x, y.

2) *A subset M of $\mathbf{R}^n$ is called convex, if it contains, along with every pair of its points x, y, also the entire segment $[x, y]$:*

$$x, y \in M,\ 0 \leq \lambda \leq 1 \Rightarrow \lambda x + (1 - \lambda)y \in M.$$

Note that by this definition an empty set is convex (by convention, or better to say, by the exact sense of the definition: for the empty set, you cannot present a counterexample to show that it is not convex).

B.1.2 Examples of convex sets

B.1.2.1 Affine subspaces and polyhedral sets

Example B.1 *A linear/affine subspace of $\mathbf{R}^n$ is convex.*

Convexity of affine subspaces immediately follows from the possibility to represent these sets as solution sets of systems of linear equations (Proposition A.3.7), due to the following simple and important fact:

Proposition B.1.1 *The solution set of an arbitrary (possibly, infinite) system*

$$a_\alpha^T x \leq b_\alpha,\ \alpha \in \mathcal{A} \tag{!}$$

of nonstrict linear inequalities with n unknowns x – the set

$$S = \{x \in \mathbf{R}^n : a_\alpha^T x \leq b_\alpha,\ \alpha \in \mathcal{A}\}$$

is convex.

In particular, the solution set of a finite system

$$Ax \leq b$$

of m nonstrict inequalities with n variables (A is $m \times n$ matrix) is convex; a set of this latter type is called polyhedral.

Exercise B.1 *Prove Proposition B.1.1.*

Remark B.1.1 *Note that every set given by Proposition B.1.1 is not only convex, but also closed (why?). In fact, from Separation Theorem (Theorem B.2.9 below) it follows that*

> *Every closed convex set in $\mathbf{R}^n$ is the solution set of a (perhaps, infinite) system of nonstrict linear inequalities.*

Remark B.1.2 *Note that replacing some of the nonstrict linear inequalities $a_\alpha^T x \le b_\alpha$ in (!) with their strict versions $a_\alpha^T x < b_\alpha$, we get a system with the solution set which still is convex (why?), but now not necessary is closed.*

B.1.2.2 Unit balls of norms

Let $\|\cdot\|$ be a norm on $\mathbf{R}^n$ i.e., a real-valued function on $\mathbf{R}^n$ satisfying the three characteristic properties of a norm (Section A.4.1), specifically:

A. [positivity] $\|x\| \ge 0$ for all $x \in \mathbf{R}^n$; $\|x\| = 0$ is and only if $x = 0$;

B. [homogeneity] For $x \in \mathbf{R}^n$ and $\lambda \in \mathbf{R}$, one has
$$\|\lambda x\| = |\lambda|\|x\|;$$

C. [triangle inequality] For all $x, y \in \mathbf{R}^n$ one has
$$\|x + y\| \le \|x\| + \|y\|.$$

Example B.2 *The unit ball of the norm $\|\cdot\|$ – the set*
$$\{x \in E : \|x\| \le 1\},$$
same as every other $\|\cdot\|$-ball
$$\{x : \|x - a\| \le r\}$$
($a \in \mathbf{R}^n$ and $r \ge 0$ are fixed) is convex.

In particular, Euclidean balls ($\|\cdot\|$-balls associated with the standard Euclidean norm $\|x\|_2 = \sqrt{x^T x}$) are convex.

The standard examples of norms on $\mathbf{R}^n$ are the ℓ_p-norms
$$\|x\|_p = \begin{cases} \left(\sum\limits_{i=1}^n |x_i|^p\right)^{1/p}, & 1 \le p < \infty \\ \max\limits_{1 \le i \le n} |x_i|, & p = \infty \end{cases}$$

These indeed are norms (which is not clear in advance; for proof, see p. 486). When $p = 2$, we get the usual Euclidean norm; of course, you know how the Euclidean ball looks. When $p = 1$, we get
$$\|x\|_1 = \sum_{i=1}^n |x_i|,$$
and the unit ball is the *hyperoctahedron*
$$V = \{x \in \mathbf{R}^n : \sum_{i=1}^n |x_i| \le 1\}$$

When $p = \infty$, we get
$$\|x\|_\infty = \max_{1 \le i \le n} |x_i|,$$
and the unit ball is the *hypercube*
$$V = \{x \in \mathbf{R}^n : -1 \le x_i \le 1, 1 \le i \le n\}.$$

Exercise B.2 *Prove that unit balls of norms on* $\mathbf{R}^n$ *are exactly the same as convex sets* V *in* $\mathbf{R}^n$ *satisfying the following three properties:*

1. V *is symmetric w.r.t. the origin:* $x \in V \Rightarrow -x \in V$*;*
2. V *is bounded and closed;*
3. V *contains a neighbourhood of the origin.*

A set V *satisfying the outlined properties is the unit ball of the norm*

$$\|x\| = \inf\left\{t \geq 0 : t^{-1}x \in V\right\}.$$

Hint: You could find useful to verify and to exploit the following facts:

1. A norm $\|\cdot\|$ on $\mathbf{R}^n$ is Lipschitz continuous with respect to the standard Euclidean distance: there exists $C_{\|\cdot\|} < \infty$ such that $|\|x\| - \|y\|| \leq C_{\|\cdot\|}\|x-y\|_2$ for all x, y
2. Vice versa, the Euclidean norm is Lipschitz continuous with respect to a given norm $\|\cdot\|$: there exists $c_{\|\cdot\|} < \infty$ such that $|\|x\|_2 - \|y\|_2| \leq c_{\|\cdot\|}\|x-y\|$ for all x, y

B.1.2.3 Ellipsoids

Example B.3 [Ellipsoid] *Let* Q *be a* $n \times n$ *matrix which is symmetric* ($Q = Q^T$) *and positive definite* ($x^TQx \geq 0$*, with* $\geq$ *being* $=$ *iff* $x = 0$)*. Then, for every nonnegative* r*, the* Q*-ellipsoid of radius* r *centered at* a *– the set*

$$\{x : (x-a)^TQ(x-a) \leq r^2\}$$

is convex.

> To see that an ellipsoid $\{x : (x-a)^TQ(x-a) \leq r^2\}$ is convex, note that since Q is positive definite, the matrix $Q^{1/2}$ is well-defined and positive definite. Now, if $\|\cdot\|$ is a norm on $\mathbf{R}^n$ and P is a nonsingular $n \times n$ matrix, the function $\|Px\|$ is a norm along with $\|\cdot\|$ (why?). Thus, the function $\|x\|_Q \equiv \sqrt{x^TQx} = \|Q^{1/2}x\|_2$ is a norm along with $\|\cdot\|_2$, and the ellipsoid in question clearly is just $\|\cdot\|_Q$-ball of radius r centered at a.

B.1.2.4 Neighbourhood of a convex set

Example B.4 *Let* M *be a convex set in* $\mathbf{R}^n$*, and let* $\epsilon > 0$*. Then, for every norm* $\|\cdot\|$ *on* $\mathbf{R}^n$*, the* ϵ*-neighbourhood of* M*, i.e., the set*

$$M_\epsilon = \{y \in \mathbf{R}^n : \mathrm{dist}_{\|\cdot\|}(y, M) \equiv \inf_{x \in M} \|y - x\| \leq \epsilon\}$$

is convex.

Exercise B.3 *Justify the statement of Example B.4.*

B.1.3 Inner description of convex sets: Convex combinations and convex hull

B.1.3.1 Convex combinations

Recall the notion of *linear combination* y of vectors $y_1, ..., y_m$ – this is a vector represented as

$$y = \sum_{i=1}^m \lambda_i y_i,$$

where λ_i are real coefficients. Specifying this definition, we have come to the notion of an *affine combination* - this is a linear combination with the sum of coefficients equal to one. The last notion in this genre is the one of *convex combination*.

Definition B.1.2 *A convex combination of vectors $y_1, ..., y_m$ is their affine combination with nonnegative coefficients, or, which is the same, a linear combination*

$$y = \sum_{i=1}^{m} \lambda_i y_i$$

with nonnegative coefficients with unit sum:

$$\lambda_i \geq 0, \ \sum_{i=1}^{m} \lambda_i = 1.$$

The following statement resembles those in Corollary A.3.2:

Proposition B.1.2 *A set M in $\mathbf{R}^n$ is convex iff it is closed with respect to taking all convex combinations of its elements, i.e., iff every convex combination of vectors from M again is a vector from M.*

Exercise B.4 *Prove Proposition B.1.2.*
Hint: Assuming $\lambda_1, ..., \lambda_m > 0$, one has

$$\sum_{i=1}^{m} \lambda_i y_i = \lambda_1 y_1 + (\lambda_2 + \lambda_3 + ... + \lambda_m) \sum_{i=2}^{m} \mu_i y_i, \quad \mu_i = \frac{\lambda_i}{\lambda_2 + \lambda_3 + ... + \lambda_m}.$$

B.1.3.2 Convex hull

Same as the property to be linear/affine subspace, the property to be convex is preserved by taking intersections (why?):

Proposition B.1.3 *Let $\{M_\alpha\}_\alpha$ be an arbitrary family of convex subsets of $\mathbf{R}^n$. Then the intersection*

$$M = \cap_\alpha M_\alpha$$

is convex.

As an immediate consequence, we come to the notion of *convex hull* $\mathrm{Conv}(M)$ of a nonempty subset in $\mathbf{R}^n$ (cf. the notions of linear/affine hull):

Corollary B.1.1 [Convex hull]
Let M be a nonempty subset in $\mathbf{R}^n$. Then among all convex sets containing M (these sets do exist, e.g., $\mathbf{R}^n$ itself) there exists the smallest one, namely, the intersection of all convex sets containing M. This set is called the convex hull of M [notation: $\mathrm{Conv}(M)$].

The linear span of M is the set of all linear combinations of vectors from M, the affine hull is the set of all affine combinations of vectors from M. As you guess,

Proposition B.1.4 [Convex hull via convex combinations] *For a nonempty $M \subset \mathbf{R}^n$:*

$$\mathrm{Conv}(M) = \{\textit{the set of all convex combinations of vectors from } M\}.$$

Exercise B.5 *Prove Proposition B.1.4.*

B.1.3.3 Simplex

The convex hull of $m+1$ affinely independent points $y_0, ..., y_m$ (Section A.3.3) is called *m-dimensional simplex with the vertices* $y_0, .., y_m$. By results of Section A.3.3, every point x of an m-dimensional simplex with vertices $y_0, ..., y_m$ admits exactly one representation as a convex combination of the vertices; the corresponding coefficients form the unique solution to the system of linear equations

$$\sum_{i=0}^{m} \lambda_i x_i = x, \sum_{i=0}^{m} \lambda_i = 1.$$

This system is solvable iff $x \in M = \mathrm{Aff}(\{y_0, .., y_m\})$, and the components of the solution (the barycentric coordinates of x) are affine functions of $x \in \mathrm{Aff}(M)$; the simplex itself is comprised of points from M with nonnegative barycentric coordinates.

B.1.4 Cones

A nonempty subset M of $\mathbf{R}^n$ is called *conic*, if it contains, along with every point $x \in M$, the entire ray $\mathbf{R}x = \{tx : t \geq 0\}$ spanned by the point:

$$x \in M \Rightarrow tx \in M \ \forall t \geq 0.$$

A convex conic set is called *a cone.*

Proposition B.1.5 *A nonempty subset M of $\mathbf{R}^n$ is a cone iff it possesses the following pair of properties:*

- *is conic: $x \in M, t \geq 0 \ \Rightarrow tx \in M$;*
- *contains sums of its elements: $x, y \in M \ \Rightarrow x + y \in M$.*

Exercise B.6 *Prove Proposition B.1.5.*

As an immediate consequence, we get that a cone is closed with respect to taking linear combinations *with nonnegative coefficients* of the elements, and vice versa – a nonempty set closed with respect to taking these combinations is a cone.

Example B.5 *The solution set of an arbitrary (possibly, infinite) system*

$$a_\alpha^T x \leq 0, \ \ \alpha \in \mathcal{A}$$

of homogeneous linear inequalities with n unknowns x – the set

$$K = \{x : a_\alpha^T x \leq 0 \ \forall \alpha \in \mathcal{A}\}$$

– is a cone.

In particular, the solution set to a homogeneous finite system of m homogeneous linear inequalities

$$Ax \leq 0$$

(A is $m \times n$ matrix) is a cone; a cone of this latter type is called polyhedral.

Note that the cones given by systems of linear homogeneous nonstrict inequalities necessarily are closed. From Separation Theorem B.2.9 it follows that, vice versa, every closed convex cone is the solution set to such a system, so that Example B.5 is the generic example of a closed convex cone.

> Cones form a very important family of convex sets, and one can develop theory of cones absolutely similar (and in a sense, equivalent) to that one of all convex sets. E.g., introducing the notion of *conic combination* of vectors $x_1, ..., x_k$ as a linear combination of the vectors with nonnegative coefficients, you can easily prove the following statements completely similar to those for general convex sets, with conic combination playing the role of convex one:
>
> - A set is a cone iff it is nonempty and is closed with respect to taking all conic combinations of its elements;
> - Intersection of a family of cones is again a cone; in particular, for every nonempty set $M \subset \mathbf{R}^n$ there exists the smallest cone containing M – its conic!hull Cone (M), and this conic hull is comprised of all conic combinations of vectors from M.
>
> In particular, the conic hull of a nonempty finite set $M = \{u_1, ..., u_N\}$ of vectors in $\mathbf{R}^n$ is the cone
>
> $$\text{Cone}\,(M) = \{\sum_{i=1}^{N} \lambda_i u_i : \lambda_i \geq 0, \ i = 1, ..., N\}.$$

B.1.5 Calculus of convex sets

Proposition B.1.6 *The following operations preserve convexity of sets:*

1. *Taking intersection: if M_α, $\alpha \in \mathcal{A}$, are convex sets, so is the set $\bigcap\limits_\alpha M_\alpha$.*
2. *Taking direct product: if $M_1 \subset \mathbf{R}^{n_1}$ and $M_2 \subset \mathbf{R}^{n_2}$ are convex sets, so is the set*
$$M_1 \times M_2 = \{y = (y_1, y_2) \in \mathbf{R}^{n_1} \times \mathbf{R}^{n_2} = \mathbf{R}^{n_1+n_2} : y_1 \in M_1, y_2 \in M_2\}.$$
3. *Arithmetic summation and multiplication by reals: if $M_1, ..., M_k$ are convex sets in $\mathbf{R}^n$ and $\lambda_1, ..., \lambda_k$ are arbitrary reals, then the set*
$$\lambda_1 M_1 + ... + \lambda_k M_k = \{\sum_{i=1}^k \lambda_i x_i : x_i \in M_i,\ i = 1, ..., k\}$$
is convex.
4. *Taking the image under an affine mapping: if $M \subset \mathbf{R}^n$ is convex and $x \mapsto \mathcal{A}(x) \equiv Ax + b$ is an affine mapping from $\mathbf{R}^n$ into $\mathbf{R}^m$ (A is $m \times n$ matrix, b is m-dimensional vector), then the set*
$$\mathcal{A}(M) = \{y = \mathcal{A}(x) \equiv Ax + a : x \in M\}$$
is a convex set in $\mathbf{R}^m$;
5. *Taking the inverse image under affine mapping: if $M \subset \mathbf{R}^n$ is convex and $y \mapsto Ay + b$ is an affine mapping from $\mathbf{R}^m$ to $\mathbf{R}^n$ (A is $n \times m$ matrix, b is n-dimensional vector), then the set*
$$\mathcal{A}^{-1}(M) = \{y \in \mathbf{R}^m : \mathcal{A}(y) \in M\}$$
is a convex set in $\mathbf{R}^m$.

Exercise B.7 *Prove Proposition B.1.6.*

B.1.6 Topological properties of convex sets

Convex sets and closely related objects - convex functions - play the central role in Optimization. To play this role properly, the convexity alone is insufficient; we need convexity plus closedness.

B.1.6.1 The closure

It is clear from definition of a closed set (Section A.4.3) that the intersection of a family of closed sets in $\mathbf{R}^n$ is also closed. From this fact it, as always, follows that for every subset M of $\mathbf{R}^n$ there exists the smallest closed set containing M; this set is called the *closure* of M and is denoted $\operatorname{cl} M$. In Analysis they prove the following inner description of the closure of a set in a metric space (and, in particular, in $\mathbf{R}^n$):

The closure of a set $M \subset \mathbf{R}^n$ is exactly the set comprised of the limits of all converging sequences of elements of M.

With this fact in mind, it is easy to prove that, e.g., the closure of the open Euclidean ball
$$\{x : |x - a| < r\} \quad [r > 0]$$
is the closed ball $\{x : \|x - a\|_2 \leq r\}$. Another useful application example is the closure of a set
$$M = \{x : a_\alpha^T x < b_\alpha,\ \alpha \in \mathcal{A}\}$$
given by strict linear inequalities: *if such a set is nonempty*, then its closure is given by the nonstrict versions of the same inequalities:
$$\operatorname{cl} M = \{x : a_\alpha^T x \leq b_\alpha,\ \alpha \in \mathcal{A}\}.$$

Nonemptiness of M in the latter example is essential: the set M given by two strict inequalities
$$x < 0, \quad -x < 0$$
in $\mathbf{R}$ clearly is empty, so that its closure also is empty; in contrast to this, applying formally the above rule, we would get wrong answer
$$\operatorname{cl} M = \{x : x \leq 0, x \geq 0\} = \{0\}.$$

B.1.6.2 The interior

Let $M \subset \mathbf{R}^n$. We say that a point $x \in M$ is an *interior* point of M, if some neighbourhood of the point is contained in M, i.e., there exists centered at x ball of positive radius which belongs to M:

$$\exists r > 0 \quad B_r(x) \equiv \{y : \|y - x\|_2 \leq r\} \subset M.$$

The set of all interior points of M is called the *interior* of M [notation: $\operatorname{int} M$].

E.g.,

- The interior of an open set is the set itself;
- The interior of the closed ball $\{x : \|x - a\|_2 \leq r\}$ is the open ball $\{x : \|x - a\|_2 < r\}$ (why?)
- The interior of a polyhedral set $\{x : Ax \leq b\}$ with matrix A not containing zero rows is the set $\{x : Ax < b\}$ (why?)

> The latter statement is not, generally speaking, valid for sets of solutions of infinite systems of linear inequalities. E.g., the system of inequalities
>
> $$x \leq \frac{1}{n},\ n = 1, 2, ...$$
>
> in $\mathbf{R}$ has, as a solution set, the nonpositive ray $\mathbf{R}_- = \{x \leq 0\}$; the interior of this ray is the negative ray $\{x < 0\}$. At the same time, strict versions of our inequalities
>
> $$x < \frac{1}{n},\ n = 1, 2, ...$$
>
> define the same nonpositive ray, not the negative one.

It is also easily seen (this fact is valid for arbitrary metric spaces, not for $\mathbf{R}^n$ only), that

- the interior of an arbitrary set is open

The interior of a set is, of course, contained in the set, which, in turn, is contained in its closure:

$$\operatorname{int} M \subset M \subset \operatorname{cl} M. \tag{B.1.1}$$

The complement of the interior in the closure – the set

$$\partial M = \operatorname{cl} M \backslash \operatorname{int} M$$

– is called the *boundary* of M, and the points of the boundary are called *boundary points* of M (Warning: these points not necessarily belong to M, since M can be less than $\operatorname{cl} M$; in fact, all boundary points belong to M iff $M = \operatorname{cl} M$, i.e., iff M is closed).

The boundary of a set clearly is closed (as the intersection of two closed sets $\operatorname{cl} M$ and $\mathbf{R}^n \backslash \operatorname{int} M$; the latter set is closed as a complement to an open set). From the definition of the boundary,

$$M \subset \operatorname{int} M \cup \partial M \quad [= \operatorname{cl} M],$$

so that a point from M is either an interior, or a boundary point of M.

B.1.6.3 The relative interior

Many of the constructions in Optimization possess nice properties in the interior of the set the construction is related to and may lose these nice properties at the boundary points of the set; this is why in many cases we are especially interested in interior points of sets and want the set of these points to be "enough massive". What to do if it is not the case – e.g., there are no interior points at all (look at a segment in the plane)? It turns out that in these cases we can use a good surrogate of the "normal" interior – the *relative interior* defined as follows.

Definition B.1.3 [Relative interior] *Let $M \subset \mathbf{R}^n$. We say that a point $x \in M$ is relative interior for M, if M contains the intersection of a small enough ball centered at x with* $\operatorname{Aff}(M)$*:*

$$\exists r > 0 \quad B_r(x) \cap \operatorname{Aff}(M) \equiv \{y : y \in \operatorname{Aff}(M),\ \|y - x\|_2 \leq r\} \subset M.$$

The set of all relative interior points of M is called its relative interior [*notation:* $\operatorname{ri} M$].

E.g. the relative interior of a singleton is the singleton itself (since a point in the 0-dimensional space is the same as a ball of a positive radius); more generally, the relative interior of an affine subspace is the set itself. The interior of a segment $[x, y]$ ($x \neq y$) in $\mathbf{R}^n$ is empty whenever $n > 1$; in contrast to this, the relative interior is nonempty independently of n and is the interval (x, y) – the segment with deleted endpoints. Geometrically speaking, the relative interior is the interior we get when regard M as a subset of its affine hull (the latter, geometrically, is nothing but $\mathbf{R}^k$, k being the affine dimension of $\mathrm{Aff}(M)$).

Exercise B.8 *Prove that the relative interior of a simplex with vertices $y_0, ..., y_m$ is exactly the set $\{x = \sum\limits_{i=0}^{m} \lambda_i y_i : \lambda_i > 0, \sum\limits_{i=0}^{m} \lambda_i = 1\}$.*

We can play with the notion of the relative interior in basically the same way as with the one of interior, namely:

- since $\mathrm{Aff}(M)$, as every affine subspace, is closed and contains M, it contains also the smallest closed sets containing M, i.e., $\mathrm{cl}\, M$. Therefore we have the following analogies of inclusions (B.1.1):
$$\mathrm{ri}\, M \subset M \subset \mathrm{cl}\, M \quad [\subset \mathrm{Aff}(M)]; \tag{B.1.2}$$
- we can define the *relative boundary* $\partial_{\mathrm{ri}} M = \mathrm{cl}\, M \backslash \mathrm{ri}\, M$ which is a closed set contained in $\mathrm{Aff}(M)$, and, as for the "actual" interior and boundary, we have
$$\mathrm{ri}\, M \subset M \subset \mathrm{cl}\, M = \mathrm{ri}\, M + \partial_{\mathrm{ri}} M.$$

Of course, if $\mathrm{Aff}(M) = \mathbf{R}^n$, then the relative interior becomes the usual interior, and similarly for boundary; this for sure is the case when $\mathrm{int}\, M \neq \emptyset$ (since then M contains a ball B, and therefore the affine hull of M is the entire $\mathbf{R}^n$, which is the affine hull of B).

B.1.6.4 Nice topological properties of convex sets

An arbitrary set M in $\mathbf{R}^n$ may possess very pathological topology: both inclusions in the chain
$$\mathrm{ri}\, M \subset M \subset \mathrm{cl}\, M$$
can be very "non-tight". E.g., let M be the set of rational numbers in the segment $[0, 1] \subset \mathbf{R}$. Then $\mathrm{ri}\, M = \mathrm{int}\, M = \emptyset$ – since every neighbourhood of every rational real contains irrational reals – while $\mathrm{cl}\, M = [0, 1]$. Thus, $\mathrm{ri}\, M$ is "incomparably smaller" than M, $\mathrm{cl}\, M$ is "incomparably larger", and M is contained in its relative boundary (by the way, what is this relative boundary?).

The following proposition demonstrates that the topology of a *convex* set M is much better than it might be for an arbitrary set.

Theorem B.1.1 *Let M be a convex set in $\mathbf{R}^n$. Then*

(i) The interior $\mathrm{int}\, M$, the closure $\mathrm{cl}\, M$ and the relative interior $\mathrm{ri}\, M$ are convex;

(ii) If M is nonempty, then the relative interior $\mathrm{ri}\, M$ of M is nonempty

(iii) The closure of M is the same as the closure of its relative interior:
$$\mathrm{cl}\, M = \mathrm{cl}\, \mathrm{ri}\, M$$
(in particular, every point of $\mathrm{cl}\, M$ is the limit of a sequence of points from $\mathrm{ri}\, M$)

(iv) The relative interior remains unchanged when we replace M with its closure:
$$\mathrm{ri}\, M = \mathrm{ri}\, \mathrm{cl}\, M.$$

Proof. (i): prove yourself!

(ii): Let M be a nonempty convex set, and let us prove that $\mathrm{ri}\, M \neq \emptyset$. By translation, we may assume that $0 \in M$. Further, we may assume that the linear span of M is the entire $\mathbf{R}^n$. Indeed, as far as linear operations and the Euclidean structure are concerned, the linear span L of M, as every other linear subspace in $\mathbf{R}^n$, is equivalent to certain $\mathbf{R}^k$; since the notion of relative interior deals only with linear and Euclidean structures, we lose nothing thinking of $\mathrm{Lin}(M)$ as of $\mathbf{R}^k$ and taking it as our universe instead of the original universe $\mathbf{R}^n$. Thus, in the rest of the proof of (ii) we assume that $0 \in M$ and $\mathrm{Lin}(M) = \mathbf{R}^n$; what we should prove is that the interior of M (which in the case in question is the same as relative interior) is nonempty. Note that since $0 \in M$, we have $\mathrm{Aff}(M) = \mathrm{Lin}(M) = \mathbf{R}^n$.

Since $\mathrm{Lin}(M) = \mathbf{R}^n$, we can find in M n linearly independent vectors $a_1, .., a_n$. Let also $a_0 = 0$. The $n+1$ vectors $a_0, ..., a_n$ belong to M, and *since M is convex*, the convex hull of these vectors also belongs to M. This convex hull is the set

$$\Delta = \{x = \sum_{i=0}^{n} \lambda_i a_i : \lambda \geq 0, \sum_i \lambda_i = 1\} = \{x = \sum_{i=1}^{n} \mu_i a_i : \mu \geq 0, \sum_{i=1}^{n} \mu_i \leq 1\}.$$

We see that Δ is the image of the *standard full-dimensional simplex*

$$\{\mu \in \mathbf{R}^n : \mu \geq 0, \sum_{i=1}^{n} \mu_i \leq 1\}$$

under linear transformation $\mu \mapsto A\mu$, where A is the matrix with the columns $a_1, ..., a_n$. The standard simplex clearly has a nonempty interior (comprised of all vectors $\mu > 0$ with $\sum\limits_i \mu_i < 1$); since A is nonsingular (due to linear independence of $a_1, ..., a_n$), multiplication by A maps open sets onto open ones, so that Δ has a nonempty interior. Since $\Delta \subset M$, the interior of M is nonempty. □

(iii): We should prove that the closure of $\mathrm{ri}\, M$ is exactly the same that the closure of M. In fact we shall prove even more:

Lemma B.1.1 *Let $x \in \mathrm{ri}\, M$ and $y \in \mathrm{cl}\, M$. Then all points from the half-segment $[x, y)$,*

$$[x, y) = \{z = (1 - \lambda)x + \lambda y : 0 \leq \lambda < 1\}$$

belong to the relative interior of M.

Proof of the Lemma. Let $\mathrm{Aff}(M) = a + L$, L being linear subspace; then

$$M \subset \mathrm{Aff}(M) = x + L.$$

Let B be the unit ball in L:

$$B = \{h \in L : \|h\|_2 \leq 1\}.$$

Since $x \in \mathrm{ri}\, M$, there exists positive radius r such that

$$x + rB \subset M. \tag{B.1.3}$$

Now let $\lambda \in [0, 1)$, and let $z = (1 - \lambda)x + \lambda y$. Since $y \in \mathrm{cl}\, M$, we have $y = \lim\limits_{i\to\infty} y_i$ for certain sequence of points from M. Setting $z_i = (1 - \lambda)x + \lambda y_i$, we get $z_i \to z$ as $i \to \infty$. Now, from (B.1.3) and the convexity of M is follows that the sets $Z_i = \{u = (1 - \lambda)x' + \lambda y_i : x' \in x + rB\}$ are contained in M; clearly, Z_i is exactly the set $z_i + r'B$, where $r' = (1 - \lambda)r > 0$. Thus, z is the limit of sequence z_i, and r'-neighbourhood (in $\mathrm{Aff}(M)$) of every one of the points z_i belongs to M. For every $r'' < r'$ and for all i such that z_i is close enough to z, the r'-neighbourhood of z_i contains the r''-neighbourhood of z; thus, a neighbourhood (in $\mathrm{Aff}(M)$) of z belongs to M, whence $z \in \mathrm{ri}\, M$. □

A useful byproduct of Lemma B.1.1 is as follows:

Corollary B.1.2 *Let M be a convex set. Then every convex combination*

$$\sum_i \lambda_i x_i$$

of points $x_i \in \mathrm{cl}\, M$ where at least one term with positive coefficient corresponds to $x_i \in \mathrm{ri}\, M$ is in fact a point from $\mathrm{ri}\, M$.

(iv): The statement is evidently true when M is empty, so assume that M is nonempty. The inclusion $\mathrm{ri}\, M \subset \mathrm{ri}\, \mathrm{cl}\, M$ is evident, and all we need is to prove the inverse inclusion. Thus, let $z \in \mathrm{ri}\, \mathrm{cl}\, M$, and let us prove that $z \in \mathrm{ri}\, M$. Let $x \in \mathrm{ri}\, M$ (we already know that the latter set is nonempty). Consider the segment $[x, z]$; since z is in the relative interior of $\mathrm{cl}\, M$, we can extend a little bit this segment through the point z, not leaving $\mathrm{cl}\, M$, i.e., there exists $y \in \mathrm{cl}\, M$ such that $z \in [x, y)$. We are done, since by Lemma B.1.1 from $z \in [x, y)$, with $x \in \mathrm{ri}\, M$, $y \in \mathrm{cl}\, M$, it follows that $z \in \mathrm{ri}\, M$. □

We see from the proof of Theorem B.1.1 that to get a closure of a (nonempty) convex set, it suffices to subject it to the "radial" closure, i.e., to take a point $x \in \mathrm{ri}\, M$, take all rays in $\mathrm{Aff}(M)$ starting at x and look at the intersection of such a ray l with M; such an intersection will be a convex set on the line which contains a one-sided neighbourhood of x, i.e., is either a segment $[x, y_l]$, or the entire ray l, or a half-interval $[x, y_l)$. In the first two cases we should not do anything; in the third we should add y to M. After all rays are looked through and all "missed" endpoints y_l are added to M, we get the closure of M. To understand what is the role of convexity here, look at the *nonconvex* set of rational numbers from $[0,1]$; the interior ($\equiv$ relative interior) of this "highly percolated" set is empty, the closure is $[0,1]$, and there is no way to restore the closure in terms of the interior.

B.2 Main theorems on convex sets

B.2.1 Caratheodory Theorem

Let us call the *affine dimension* (or simple *dimension* of a nonempty set $M \subset \mathbf{R}^n$ (notation: $\dim M$) the affine dimension of $\mathrm{Aff}(M)$.

Theorem B.2.1 [Caratheodory] *Let $M \subset \mathbf{R}^n$, and let $\dim \mathrm{Conv}M = m$. Then every point $x \in \mathrm{Conv}M$ is a convex combination of at most $m+1$ points from M.*

Proof. Let $x \in \mathrm{Conv}M$. By Proposition B.1.4 on the structure of convex hull, x is convex combination of certain points $x_1, ..., x_N$ from M:

$$x = \sum_{i=1}^{N} \lambda_i x_i, \quad [\lambda_i \geq 0, \sum_{i=1}^{N} \lambda_i = 1].$$

Let us choose among all these representations of x as a convex combination of points from M the one with the smallest possible N, and let it be the above combination. We claim that $N \leq m+1$ (this claim leads to the desired statement). Indeed, if $N > m+1$, then the system of $m+1$ homogeneous equations

$$\begin{array}{rcl} \sum\limits_{i=1}^{N} \mu_i x_i & = & 0 \\ \sum\limits_{i=1}^{N} \mu_i & = & 0 \end{array}$$

with N unknowns $\mu_1, ..., \mu_N$ has a nontrivial solution $\delta_1, ..., \delta_N$:

$$\sum_{i=1}^{N} \delta_i x_i = 0, \sum_{i=1}^{N} \delta_i = 0, \ (\delta_1, ..., \delta_N) \neq 0.$$

It follows that, for every real t,

$$(*) \quad \sum_{i=1}^{N} [\lambda_i + t\delta_i] x_i = x.$$

What is to the left, is an affine combination of x_i's. When $t = 0$, this is a convex combination - all coefficients are nonnegative. When t is large, this is *not* a convex combination, since some of δ_i's are negative (indeed, not all of them are zero, and the sum of δ_i's is 0). There exists, of course, the largest t for which the combination (*) has nonnegative coefficients, namely

$$t^* = \min_{i:\delta_i < 0} \frac{\lambda_i}{|\delta_i|}.$$

For this value of t, the combination (*) is with nonnegative coefficients, and at least one of the coefficients is zero; thus, we have represented x as a convex combination of less than N points from M, which contradicts the definition of N. □

B.2.2 Radon Theorem

Theorem B.2.2 [Radon] *Let S be a set of at least $n+2$ points $x_1, ..., x_N$ in $\mathbf{R}^n$. Then one can split the set into two nonempty subsets S_1 and S_2 with intersecting convex hulls: there exists partitioning $I \cup J = \{1, ..., N\}$, $I \cap J = \emptyset$, of the index set $\{1, ..., N\}$ into two nonempty sets I and J and convex combinations of the points $\{x_i, i \in I\}$, $\{x_j, j \in J\}$ which coincide with each other, i.e., there exist α_i, $i \in I$, and β_j, $j \in J$, such that*

$$\sum_{i\in I} \alpha_i x_i = \sum_{j \in J} \beta_j x_j; \quad \sum_i \alpha_i = \sum_j \beta_j = 1; \quad \alpha_i, \beta_j \geq 0.$$

Proof. Since $N > n+1$, the homogeneous system of $n+1$ scalar equations with N unknowns $\mu_1, ..., \mu_N$

$$\begin{array}{rcl} \sum\limits_{i=1}^N \mu_i x_i & = & 0 \\ \sum\limits_{i=1}^N \mu_i & = & 0 \end{array}$$

has a nontrivial solution $\lambda_1, ..., \lambda_N$:

$$\sum_{i=1}^N \mu_i x_i = 0, \sum_{i=1}^N \lambda_i = 0, \ [(\lambda_1, ..., \lambda_N) \neq 0].$$

Let $I = \{i : \lambda_i \geq 0\}$, $J = \{i : \lambda_i < 0\}$; then I and J are nonempty and form a partitioning of $\{1, ..., N\}$. We have

$$a \equiv \sum_{i \in I} \lambda_i = \sum_{j \in J} (-\lambda_j) > 0$$

(since the sum of all λ's is zero and not all λ's are zero). Setting

$$\alpha_i = \frac{\lambda_i}{a}, i \in I, \ \beta_j = \frac{-\lambda_j}{a}, j \in J,$$

we get

$$\alpha_i \geq 0, \ \beta_j \geq 0, \sum_{i \in I} \alpha_i = 1, \sum_{j \in J} \beta_j = 1,$$

and

$$[\sum_{i \in I} \alpha_i x_i] - [\sum_{j \in J} \beta_j x_j] = a^{-1} \left([\sum_{i \in I} \lambda_i x_i] - [\sum_{j \in J} (-\lambda_j) x_j] \right) = a^{-1} \sum_{i=1}^N \lambda_i x_i = 0. \quad \square$$

B.2.3 Helley Theorem

Theorem B.2.3 [Helley, I] *Let $\mathcal{F}$ be a finite family of convex sets in $\mathbf{R}^n$. Assume that every $n+1$ sets from the family have a point in common. Then all the sets have a point in common.*

Proof. Let us prove the statement by induction on the number N of sets in the family. The case of $N \leq n+1$ is evident. Now assume that the statement holds true for all families with certain number $N \geq n+1$ of sets, and let $S_1, ..., S_N, S_{N+1}$ be a family of $N+1$ convex sets which satisfies the premise of the Helley Theorem; we should prove that the intersection of the sets $S_1, ..., S_N, S_{N+1}$ is nonempty.

Deleting from our $N+1$-set family the set S_i, we get N-set family which satisfies the premise of the Helley Theorem and thus, by the inductive hypothesis, the intersection of its members is nonempty:

$$(\forall i \leq N+1) : \ T^i = S_1 \cap S_2 \cap ... \cap S_{i-1} \cap S_{i+1} \cap ... \cap S_{N+1} \neq \emptyset.$$

Let us choose a point x_i in the (nonempty) set T^i. We get $N+1 \geq n+2$ points from $\mathbf{R}^n$. By Radon's Theorem, we can partition the index set $\{1, ..., N+1\}$ into two nonempty subsets I and J in such a way that certain convex combination x of the points x_i, $i \in I$, is a convex combination of the points x_j, $j \in J$, as well. Let us verify that x belongs to all the sets $S_1, ..., S_{N+1}$, which will complete the proof. Indeed, let i^* be an index from our index set; let us prove that $x \in S_{i^*}$. We have either $i^* \in I$, or $i^* \in J$. In the first case all the sets T^j, $j \in J$, are contained in S_{i^*} (since S_{i^*} participates in all intersections which give T^i with $i \neq i^*$).

Consequently, all the points x_j, $j \in J$, belong to S_{i^*}, and therefore x, which is a convex combination of these points, also belongs to S_{i^*} (all our sets are convex!), as required. In the second case similar reasoning says that all the points x_i, $i \in I$, belong to S_{i^*}, and therefore x, which is a convex combination of these points, belongs to S_{i^*}. □

Exercise B.9 *Let $S_1, ..., S_N$ be a family of N convex sets in $\mathbf{R}^n$, and let m be the affine dimension of* $\mathrm{Aff}(S_1 \cup ... \cup S_N)$. *Assume that every $m+1$ sets from the family have a point in common. Prove that all sets from the family have a point in common.*

In the aforementioned version of the Helley Theorem we dealt with finite families of convex sets. To extend the statement to the case of infinite families, we need to strengthen slightly the assumption. The resulting statement is as follows:

Theorem B.2.4 [Helley, II] *Let $\mathcal{F}$ be an arbitrary family of convex sets in $\mathbf{R}^n$. Assume that*

(a) every $n+1$ sets from the family have a point in common,

and

(b) every set in the family is closed, and the intersection of the sets from certain finite subfamily of the family is bounded (e.g., one of the sets in the family is bounded).

Then all the sets from the family have a point in common.

Proof. By the previous theorem, all finite subfamilies of $\mathcal{F}$ have nonempty intersections, and these intersections are convex (since intersection of a family of convex sets is convex, Theorem B.1.3); in view of (a) these intersections are also closed. Adding to $\mathcal{F}$ all intersections of finite subfamilies of $\mathcal{F}$, we get a larger family $\mathcal{F}'$ comprised of closed convex sets, and a finite subfamily of this larger family again has a nonempty intersection. Besides this, from (b) it follows that this new family contains a bounded set Q. Since all the sets are closed, the family of sets

$$\{Q \cap Q' : Q' \in \mathcal{F}\}$$

is a *nested family of compact sets* (i.e., a family of compact sets with nonempty intersection of sets from every finite subfamily); by the well-known Analysis theorem such a family has a nonempty intersection[1]. □

B.2.4 Polyhedral representations and Fourier-Motzkin Elimination

B.2.4.1 Polyhedral representations

Recall that by definition a polyhedral set X in $\mathbf{R}^n$ is the set of solutions of a finite system of nonstrict linear inequalities in variables $x \in \mathbf{R}^n$:

$$X = \{x \in \mathbf{R}^n : Ax \le b\} = \{x \in \mathbf{R}^n : a_i^T x \le b_i,\ 1 \le i \le m\}.$$

We shall call such a representation of X its *polyhedral description*. A polyhedral set always is convex and closed (Proposition B.1.1). Now let us introduce the notion of *polyhedral representation* of a set $X \subset \mathbf{R}^n$.

Definition B.2.1 We say that a set $X \subset \mathbf{R}^n$ is *polyhedrally representable*, if it admits a representation as follows:

$$X = \{x \in \mathbf{R}^n : \exists u \in \mathbf{R}^k : Ax + Bu \le c\} \tag{B.2.1}$$

where A, B are $m \times n$ and $m \times k$ matrices and $c \in \mathbf{R}^m$. A representation of X of the form of (B.2.1) is called a *polyhedral representation of* X, and variables u in such a representation are called *slack variables*.

Geometrically, a polyhedral representation of a set $X \subset \mathbf{R}^n$ is its representation as the *projection* $\{x : \exists u : (x, u) \in Y\}$ of a *polyhedral* set $Y = \{(x, u) : Ax + Bu \le c\}$ in the space of $n + k$ variables

[1] here is the proof of this Analysis theorem: assume, on contrary, that the compact sets Q_α, $\alpha \in \mathcal{A}$, have empty intersection. Choose a set Q_{α^*} from the family; for every $x \in Q_{\alpha^*}$ there is a set Q^x in the family which does not contain x - otherwise x would be a common point of all our sets. Since Q^x is closed, there is an open ball V_x centered at x which does not intersect Q^x. The balls V_x, $x \in Q_{\alpha^*}$, form an open covering of the compact set Q_{α^*}, and therefore there exists a finite subcovering $V_{x_1}, ..., V_{x_N}$ of Q_{α^*} by the balls from the covering. Since Q^{x_i} does not intersect V_{x_i}, we conclude that the intersection of the finite subfamily $Q_{\alpha^*}, Q^{x_1}, ..., Q^{x_N}$ is empty, which is a contradiction

($x \in \mathbf{R}^n, u \in \mathbf{R}^k$) under the linear mapping (the projection) $(x,u) \mapsto x : \mathbf{R}^{n+k}_{x,u} \to \mathbf{R}^n_x$ of the $n+k$-dimensional space of (x,u)-variables (the space where Y lives) to the n-dimensional space of x-variables where X lives.

Note that every polyhedrally representable set is the image under linear mapping (even a projection) of a polyhedral, and thus convex, set. It follows that *a polyhedrally representable set definitely is convex* (Proposition B.1.6).

Examples: 1) Every polyhedral set $X = \{x \in \mathbf{R}^n : Ax \le b\}$ is polyhedrally representable – a polyhedral description of X is nothing but a polyhedral representation with no slack variables ($k = 0$). Vice versa, a polyhedral representation of a set X with no slack variables ($k = 0$) clearly is a polyhedral description of the set (which therefore is polyhedral).

2) Looking at the set $X = \{x \in \mathbf{R}^n : \sum_{i=1}^n |x_i| \le 1\}$, we cannot say immediately whether it is or is not polyhedral; at least the initial description of X is *not* of the form $\{x : Ax \le b\}$. However, X admits a polyhedral representation, e.g., the representation

$$X = \{x \in \mathbf{R}^n : \exists u \in \mathbf{R}^n : \underbrace{-u_i \le x_i \le u_i}_{\Leftrightarrow |x_i| \le u_i}, 1 \le i \le n, \sum_{i=1}^n u_i \le 1\}. \tag{B.2.2}$$

Note that the set X in question can be described by a system of linear inequalities *in x-variables only*, namely, as

$$X = \{x \in \mathbf{R}^n : \sum_{i=1}^n \epsilon_i x_i \le 1\,, \forall(\epsilon_i = \pm 1,\, 1 \le i \le n)\},$$

that is, X is polyhedral. However, the above polyhedral description of X (which in fact is minimal in terms of the number of inequalities involved) requires 2^n inequalities — an astronomically large number when n is just few tens. In contrast to this, the polyhedral representation (B.2.2) of the same set requires just n slack variables u and $2n+1$ linear inequalities on x, u – the "complexity" of this representation is just linear in n.

3) Let $a_1, ..., a_m$ be given vectors in $\mathbf{R}^n$. Consider the conic hull of the finite set $\{a_1, ..., a_m\}$ – the set $\mathrm{Cone}\,\{a_1, ..., a_m\} = \{x = \sum\limits_{i=1}^m \lambda_i a_i : \lambda \ge 0\}$ (see Section B.1.4). It is absolutely unclear whether this set is polyhedral. In contrast to this, its polyhedral representation is immediate:

$$\begin{array}{rl}\mathrm{Cone}\,\{a_1, ..., a_m\} = & \{x \in \mathbf{R}^n : \exists \lambda \ge 0 : x = \sum_{i=1}^m \lambda_i a_i\} \\ & = \{x \in \mathbf{R}^n : \exists \lambda \in \mathbf{R}^m : \left\{\begin{array}{l} -\lambda \le 0 \\ x - \sum_{i=1}^m \lambda_i a_i \le 0 \\ -x + \sum_{i=1}^m \lambda_i a_i \le 0 \end{array}\right. \}\end{array}$$

In other words, the original description of X is nothing but its polyhedral representation (in slight disguise), with λ_i's in the role of slack variables.

B.2.4.2 Every polyhedrally representable set is polyhedral! (Fourier-Motzkin elimination)

A surprising and deep fact is that the situation in the Example 2) above is quite general:

Theorem B.2.5 *Every polyhedrally representable set is polyhedral.*

Proof: Fourier-Motzkin Elimination. Recalling the definition of a polyhedrally representable set, our claim can be rephrased equivalently as follows: *the projection of a polyhedral set Y in a space $\mathbf{R}^{n+k}_{u,v}$ of (x,u)-variables on the subspace $\mathbf{R}^n_x$ of x-variables is a polyhedral set in $\mathbf{R}^n$.* All we need is to prove this claim in the case of exactly one slack variable (since the projection which reduces the dimension by k — "kills k slack variables" — is the result of k subsequent projections, every one reducing the dimension by 1 ("killing one slack variable each")).

Thus, let

$$Y = \{(x,u) \in \mathbf{R}^{n+1} : a_i^T x + b_i u \le c_i, 1 \le i \le m\}$$

be a polyhedral set with n variables x and single variable u; we want to prove that the projection

$$X = \{x : \exists u : Ax + bu \le c\}$$

of Y on the space of x-variables is polyhedral. To see it, let us split the inequalities defining Y into three groups (some of them can be empty):
— "black" inequalities — those with $b_i = 0$; these inequalities do not involve u at all;
— "red" inequalities – those with $b_i > 0$. Such an inequality can be rewritten equivalently as $u \leq b_i^{-1}[c_i - a_i^T x]$, and it imposes a (depending on x) *upper bound* on u;
— "green" inequalities – those with $b_i < 0$. Such an inequality can be rewritten equivalently as $u \geq b_i^{-1}[c_i - a_i^T x]$, and it imposes a (depending on x) *lower bound* on u.
Now it is clear when $x \in X$, that is, when x can be extended, by some u, to a point (x, u) from Y: this is the case iff , first, x satisfies all black inequalities, and, second, the red upper bounds on u specified by x are compatible with the green lower bounds on u specified by x, meaning that every lower bound is $\leq$ every upper bound (the latter is necessary and sufficient to be able to find a value of u which is $\geq$ all lower bounds and $\leq$ all upper bounds). Thus,

$$X = \left\{ x : \left\{ \begin{array}{l} a_i^T x \leq c_i \text{ for all "black" indexes } i \text{ – those with } b_i = 0 \\ b_j^{-1}[c_j - a_j^T x] \leq b_k^{-1}[c_k - a_k^T x] \text{ for all "green" (i.e., with } b_j < 0) \text{ indexes } j \\ \text{ and all "red" (i.e., with } b_k > 0) \text{ indexes } k \end{array} \right. \right\}.$$

We see that X is given by finitely many nonstrict linear inequalities in x-variables only, as claimed. □

The outlined procedure for building polyhedral descriptions (i.e., polyhedral representations not involving slack variables) for projections of polyhedral sets is called *Fourier-Motzkin elimination.*

B.2.4.3 Some applications

As an immediate application of Fourier-Motzkin elimination, let us take a linear program $\min\limits_x \{c^T x : Ax \leq b\}$ and look at the set T of values of the objective at all feasible solutions, if any:

$$T = \{t \in \mathbf{R} : \exists x : c^T x = t, Ax \leq b\}.$$

Rewriting the linear equality $c^T x = t$ as a pair of opposite inequalities, we see that T is polyhedrally representable, and the above definition of T is nothing but a polyhedral representation of this set, with x in the role of the vector of slack variables. By Fourier-Motzkin elimination, T is polyhedral – this set is given by a finite system of nonstrict linear inequalities in variable t only. As such, as it is immediately seen, T is
— either empty (meaning that the LP in question is infeasible),
— or is a below unbounded nonempty set of the form $\{t \in \mathbf{R} : -\infty \leq t \leq b\}$ with $b \in \mathbf{R} \cup \{+\infty\}$ (meaning that the LP is feasible and unbounded),
— or is a below bounded nonempty set of the form $\{t \in \mathbf{R} : -a \leq t \leq b\}$ with $a \in \mathbf{R}$ and $+\infty \geq b \geq a$. In this case, the LP is feasible and bounded, and a is its optimal value.
Note that given the list of linear inequalities defining T (this list can be built algorithmically by Fourier-Motzkin elimination as applied to the original polyhedral representation of T), we can easily detect which one of the above cases indeed takes place, i.e., to identify the feasibility and boundedness status of the LP and to find its optimal value. When it is finite (case 3 above), we can use the Fourier-Motzkin elimination backward, starting with $t = a \in T$ and extending this value to a pair (t, x) with $t = a = c^T x$ and $Ax \leq b$, that is, we can augment the optimal value by an optimal solution. Thus, we can say that Fourier-Motzkin elimination is a finite Real Arithmetics algorithm which allows to check whether an LP is feasible and bounded, and when it is the case, allows to find the optimal value and an optimal solution. An unpleasant fact of life is that this algorithm is completely impractical, since the elimination process can blow up exponentially the number of inequalities. Indeed, from the description of the process it is clear that if a polyhedral set is given by m linear inequalities, then eliminating one variable, we can end up with as much as $m^2/4$ inequalities (this is what happens if there are $m/2$ red, $m/2$ green and no black inequalities). Eliminating the next variable, we again can "nearly square" the number of inequalities, and so on. Thus, the number of inequalities in the description of T can become astronomically large when even when the dimension of x is something like 10. The actual importance of Fourier-Motzkin elimination is of theoretical nature. For example, the LP-related reasoning we have just carried out shows that *every feasible and bounded LP program is solvable – has an optimal solution* (we shall revisit this result in more details in Section B.2.11.B). This is a fundamental fact for LP, and the above reasoning (even with the justification of the elimination "charged" to it) is the shortest and most transparent way to prove this fundamental fact. Another application of the fact that polyhedrally representable sets are polyhedral is the Homogeneous Farkas Lemma to be stated and proved in Section B.2.5.A; this lemma will be instrumental in numerous subsequent theoretical developments.

B.2.4.4 Calculus of polyhedral representations

The fact that polyhedral sets are exactly the same as polyhedrally representable ones does not nullify the notion of a polyhedral representation. The point is that a set can admit "quite compact" polyhedral representation involving slack variables and require astronomically large, completely meaningless for any practical purpose number of inequalities in its polyhedral description (think about the set (B.2.2) when $n = 100$). Moreover, polyhedral representations admit a kind of "fully algorithmic calculus." Specifically, it turns out that all basic convexity-preserving operations (cf. Proposition B.1.6) as applied to polyhedral operands preserve polyhedrality; moreover, polyhedral representations of the results are readily given by polyhedral representations of the operands. Here is "algorithmic polyhedral analogy" of Proposition B.1.6:

1. *Taking finite intersection*: Let M_i, $1 \leq i \leq m$, be polyhedral sets in $\mathbf{R}^n$ given by their polyhedral representations
$$M_i = \{x \in \mathbf{R}^n : \exists u^i \in \mathbf{R}^{k_i} : A_i x + B_i u^i \leq c^i\},\ 1 \leq i \leq m.$$
Then the intersection of the sets M_i is polyhedral with an explicit polyhedral representation, specifically,
$$\bigcap_{i=1}^m M_i = \{x \in \mathbf{R}^n : \exists u = (u^1, ..., u^m) \in \mathbf{R}^{k_1+...+k_m} : \underbrace{A_i x + B_i u^i \leq c^i,\ 1 \leq i \leq m}_{\substack{\text{system of nonstrict linear}\\ \text{inequalities in } x,u}}\}$$

2. *Taking direct product*: Let $M_i \subset \mathbf{R}^{n_i}$, $1 \leq i \leq m$, be polyhedral sets given by polyhedral representations
$$M_i = \{x^i \in \mathbf{R}^{n_i} : \exists u^i \in \mathbf{R}^{k_i} : A_i x^i + B_i u^i \leq c^i\},\ 1 \leq i \leq m.$$
Then the direct product $M_1 \times ... \times M_m := \{x = (x^1, ..., x^m) : x^i \in M_i,\ 1 \leq i \leq m\}$ of the sets is a polyhedral set with explicit polyhedral representation, specifically,
$$\begin{array}{rl} M_1 \times ... \times M_m = & \{x = (x^1, ..., x^m) \in \mathbf{R}^{n_1+...+n_m} : \\ & \exists u = (u^1, ..., u^m) \in \mathbf{R}^{k_1+...+k_m} : A_i x^i + B_i u^i \leq c^i,\ 1 \leq i \leq m\}. \end{array}$$

3. *Arithmetic summation and multiplication by reals*: Let $M_i \subset \mathbf{R}^n$, $1 \leq i \leq m$, be polyhedral sets given by polyhedral representations
$$M_i = \{x \in \mathbf{R}^n : \exists u^i \in \mathbf{R}^{k_i} : A_i x + B_i u^i \leq c^i\},\ 1 \leq i \leq m,$$
and let $\lambda_1, ..., \lambda_k$ be reals. Then the set $\lambda_1 M_1 + ... + \lambda_m M_m := \{x = \lambda_1 x_1 + ... + \lambda_m x_m : x_i \in M_i,\ 1 \leq i \leq m\}$ is polyhedral with explicit polyhedral representation, specifically,
$$\begin{array}{rl} \lambda_1 M_1 + ... + \lambda_m M_m = & \{x \in \mathbf{R}^n : \exists (x^i \in \mathbf{R}^n, u^i \in \mathbf{R}^{k_i}, 1 \leq i \leq m) : \\ & x \leq \sum_i \lambda_i x^i, x \geq \sum_i \lambda_i x^i,\ A_i x^i + B_i u^i \leq c^i,\ 1 \leq i \leq m\}. \end{array}$$

4. Taking the image under an affine mapping: Let $M \subset \mathbf{R}^n$ be a polyhedral set given by polyhedral representation
$$M = \{x \in \mathbf{R}^n : \exists u \in \mathbf{R}^k : Ax + Bu \leq c\}$$
and let $\mathcal{P}(x) = Px + p : \mathbf{R}^n \to \mathbf{R}^m$ be an affine mapping. Then the image $\mathcal{P}(M) := \{y = Px + p : x \in M\}$ of M under the mapping is polyhedral set with explicit polyhedral representation, specifically,
$$\mathcal{P}(M) = \{y \in \mathbf{R}^m : \exists (x \in \mathbf{R}^n, u \in \mathbf{R}^k) : y \leq Px + p, y \geq Px + p, Ax + Bu \leq c\}.$$

5. Taking the inverse image under affine mapping: Let $M \subset \mathbf{R}^n$ be polyhedral set given by polyhedral representation
$$M = \{x \in \mathbf{R}^n : \exists u \in \mathbf{R}^k : Ax + Bu \leq c\}$$
and let $\mathcal{P}(y) = Py + p : \mathbf{R}^m \to \mathbf{R}^n$ be an affine mapping. Then the inverse image $\mathcal{P}^{-1}(M) := \{y : Py + p \in M\}$ of M under the mapping is polyhedral set with explicit polyhedral representation, specifically,
$$\mathcal{P}^{-1}(M) = \{y \in \mathbf{R}^m : \exists u : A(Py + p) + Bu \leq c\}.$$

Note that rules for intersection, taking direct products and taking inverse images, as applied to polyhedral *descriptions* of operands, lead to polyhedral descriptions of the results. In contrast to this, the rules for taking sums with coefficients and images under affine mappings heavily exploit the notion of polyhedral *representation*: even when the operands in these rules are given by polyhedral descriptions, there are no simple ways to point out polyhedral *descriptions* of the results.

Exercise B.10 *Justify the above calculus rules.*

Finally, we note that the problem of minimizing a linear form $c^T x$ over a set M given by polyhedral representation:

$$M = \{x \in \mathbf{R}^n : \exists u \in \mathbf{R}^k : Ax + Bu \leq c\}$$

can be immediately reduced to an explicit LP program, namely,

$$\min_{x,u} \left\{c^T x : Ax + Bu \leq c\right\}.$$

A reader with some experience in Linear Programming definitely used a lot the above "calculus of polyhedral representations" when building LPs (perhaps without clear understanding of what in fact is going on, same as Molière's Monsieur Jourdain all his life has been speaking prose without knowing it).

B.2.5 General Theorem on Alternative and Linear Programming Duality

B.2.5.1 Homogeneous Farkas Lemma

Let $a_1, ..., a_N$ be vectors from $\mathbf{R}^n$, and let a be another vector. Here we address the question: when a belongs to the cone spanned by the vectors $a_1, ..., a_N$, i.e., when a can be represented as a linear combination of a_i with *nonnegative* coefficients? A *necessary* condition is evident: if

$$a = \sum_{i=1}^{n} \lambda_i a_i \quad [\lambda_i \geq 0,\, i = 1, ..., N]$$

then every vector h which has nonnegative inner products with all a_i should also have nonnegative inner product with a:

$$a = \sum_i \lambda_i a_i \;\&\; \lambda_i \geq 0\, \forall i \;\&\; h^T a_i \geq 0\, \forall i \Rightarrow h^T a \geq 0.$$

The Homogeneous Farkas Lemma says that this evident necessary condition is also sufficient:

Lemma B.2.1 [Homogeneous Farkas Lemma] *Let $a, a_1, ..., a_N$ be vectors from $\mathbf{R}^n$. The vector a is a conic combination of the vectors a_i (linear combination with nonnegative coefficients) iff every vector h satisfying $h^T a_i \geq 0$, $i = 1, ..., N$, satisfies also $h^T a \geq 0$. In other words, a homogeneous linear inequality*

$$a^T h \geq 0$$

in variable h is consequence of the system

$$a_i^T h \geq 0,\, 1 \leq i \leq N$$

of homogeneous linear inequalities iff it can be obtained from the inequalities of the system by "admissible linear aggregation" – taking their weighted sum with nonnegative weights.

Proof. The necessity – the "only if" part of the statement – was proved before the Farkas Lemma was formulated. Let us prove the "if" part of the Lemma. Thus, assume that every vector h satisfying $h^T a_i \geq 0$ $\forall i$ satisfies also $h^T a \geq 0$, and let us prove that a is a conic combination of the vectors a_i.

An "intelligent" proof goes as follows. The set Cone $\{a_1, ..., a_N\}$ of all conic combinations of $a_1, ..., a_N$ is polyhedrally representable (Example 3 in Section B.2.5.A.1) and as such is polyhedral (Theorem B.2.5):

$$\text{Cone}\,\{a_1, ..., a_N\} = \{x \in \mathbf{R}^n : p_j^T x \geq b_j, 1 \leq j \leq J\}. \tag{!}$$

Observing that $0 \in \text{Cone}\,\{a_1, ..., a_N\}$, we conclude that $b_j \leq 0$ for all j; and since $\lambda a_i \in \text{Cone}\,\{a_1, ..., a_N\}$ for every i and every $\lambda \geq 0$, we should have $\lambda p_j^T a_i \geq 0$ for all i, j and all $\lambda \geq 0$, whence $p_j^T a_i \geq 0$ for all i and

j. For every j, relation $p_j^T a_i \geq 0$ for all i implies, by the premise of the statement we want to prove, that $p_j^T a \geq 0$, and since $b_j \leq 0$, we see that $p_j^T a \geq b_j$ for all j, meaning that a indeed belongs to Cone $\{a_1, ..., a_N\}$ due to (!). □

An interested reader can get a better understanding of the power of Fourier-Motzkin elimination, which ultimately is the basis for the above intelligent proof, by comparing this proof with the one based on Helley's Theorem.

Proof based on Helley's Theorem. As above, we assume that every vector h satisfying $h^T a_i \geq 0$ $\forall i$ satisfies also $h^T a \geq 0$, and we want to prove that a is a conic combination of the vectors a_i.

There is nothing to prove when $a = 0$ – the zero vector of course is a conic combination of the vectors a_i. Thus, from now on we assume that $a \neq 0$.

1^0. Let

$$\Pi = \{h : a^T h = -1\},$$

and let

$$A_i = \{h \in \Pi : a_i^T h \geq 0\}.$$

Π is a hyperplane in $\mathbf{R}^n$, and every A_i is a polyhedral set contained in this hyperplane and is therefore convex.

2^0. What we know is that the intersection of all the sets A_i, $i = 1, ..., N$, is empty (since a vector h from the intersection would have nonnegative inner products with all a_i and the inner product -1 with a, and we are given that no such h exists). Let us choose the smallest, in the number of elements, of those sub-families of the family of sets $A_1, ..., A_N$ which still have empty intersection of their members; without loss of generality we may assume that this is the family $A_1, ..., A_k$. Thus, the intersection of all k sets $A_1, ..., A_k$ is empty, but the intersection of every $k-1$ sets from the family $A_1, ..., A_k$ is nonempty.

3^0. We claim that

(A) $a \in \mathrm{Lin}(\{a_1, ..., a_k\})$;

(B) The vectors $a_1, ..., a_k$ are linearly independent.

> (A) is easy: assuming that $a \notin E = \mathrm{Lin}(\{a_1, ..., a_k\})$, we conclude that the orthogonal projection f of the vector a onto the orthogonal complement $E^\perp$ of E is nonzero. The inner product of f and a is the same as $f^T f$, is.e., is positive, while $f^T a_i = 0$, $i = 1, ..., k$. Taking $h = -(f^T f)^{-1} f$, we see that $h^T a = -1$ and $h^T a_i = 0$, $i = 1, ..., k$. In other words, h belongs to every set A_i, $i = 1, ..., k$, by definition of these sets, and therefore the intersection of the sets $A_1, ..., A_k$ is nonempty, which is a contradiction.
>
> (B) is given by the Helley Theorem I. (B) is evident when $k = 1$, since in this case linear dependence of $a_!, ..., a_k$ would mean that $a_1 = 0$; by (A), this implies that $a = 0$, which is not the case. Now let us prove (B) in the case of $k > 1$. Assume, on the contrary to what should be proven, that $a_1, ..., a_k$ are linearly dependent, so that the dimension of $E = \mathrm{Lin}(\{a_1, ..., a_k\})$ is certain $m < k$. We already know from A. that $a \in E$. Now let $A_i' = A_i \cap E$. We claim that every $k-1$ of the sets A_i' have a nonempty intersection, while all k these sets have empty intersection. The second claim is evident – since the sets $A_1, ..., A_k$ have empty intersection, the same is the case with their parts A_i'. The first claim also is easily supported: let us take $k-1$ of the dashed sets, say, $A_1', ..., A_{k-1}'$. By construction, the intersection of $A_1, ..., A_{k-1}$ is nonempty; let h be a vector from this intersection, i.e., a vector with nonnegative inner products with $a_1, ..., a_{k-1}$ and the product -1 with a. When replacing h with its orthogonal projection h' on E, we do not vary all these inner products, since these are products with vectors from E; thus, h' also is a common point of $A_1, ..., A_{k-1}$, and since this is a point from E, it is a common point of the dashed sets $A_1', ..., A_{k-1}'$ as well.
>
> Now we can complete the proof of (B): the sets $A_1', ..., A_k'$ are convex sets belonging to the *hyperplane* $\Pi' = \Pi \cap E = \{h \in E : a^T h = -1\}$ (Π' indeed is a hyperplane in E, since $0 \neq a \in E$) *in the m-dimensional linear subspace* E. Π' is an affine subspace of the affine dimension $\ell = \dim E - 1 = m - 1 < k - 1$ (recall that we are in the situation when $m = \dim E < k$),

and every $\ell+1 \leq k-1$ subsets from the family $A'_1,...,A'_k$ have a nonempty intersection. From the Helley Theorem I (see Exercise B.9) it follows that all the sets $A'_1,...,A'_k$ have a point in common, which, as we know, is not the case. The contradiction we have got proves that $a_1,...,a_k$ are linearly independent.

4^0. With (A) and (B) in our disposal, we can easily complete the proof of the "if" part of the Farkas Lemma. Specifically, by (A), we have

$$a = \sum_{i=1}^{k} \lambda_i a_i$$

with some real coefficients λ_i, and all we need is to prove that these coefficients are nonnegative. Assume, on the contrary, that, say, $\lambda_1 < 0$. Let us extend the (linearly independent in view of (B)) system of vectors $a_1,...,a_k$ by vectors $f_1,...,f_{n-k}$ to a basis in $\mathbf{R}^n$, and let $\xi_i(x)$ be the coordinates of a vector x in this basis. The function $\xi_1(x)$ is a linear form of x and therefore is the inner product with certain vector:

$$\xi_1(x) = f^T x \quad \forall x.$$

Now we have

$$f^T a = \xi_1(a) = \lambda_1 < 0$$

and

$$f^T a_i = \begin{cases} 1, & i = 1 \\ 0, & i = 2,...,k \end{cases}$$

so that $f^T a_i \geq 0$, $i = 1,...,k$. We conclude that a proper normalization of f – namely, the vector $|\lambda_1|^{-1} f$ – belongs to $A_1,...,A_k$, which is the desired contradiction – by construction, this intersection is empty. □

B.2.6 General Theorem on Alternative

B.2.6.1 Certificates for solvability and insolvability

Consider a (finite) system of scalar inequalities with n unknowns. To be as general as possible, we do not assume for the time being the inequalities to be linear, and we allow for both non-strict and strict inequalities in the system, as well as for equalities. Since an equality can be represented by a pair of non-strict inequalities, our system can always be written as

$$f_i(x)\ \Omega_i\ 0,\ i = 1,...,m, \tag{$\mathcal{S}$}$$

where every Ω_i is either the relation " $>$ " or the relation " $\geq$ ".

The basic question about ($\mathcal{S}$) is

(?) Whether ($\mathcal{S}$) has a solution or not.

Knowing how to answer the question (?), we are able to answer many other questions. E.g., to verify whether a given real a is a lower bound on the optimal value c^* of (LP) is the same as to verify whether the system

$$\begin{cases} -c^T x + a & > 0 \\ \quad Ax - b & \geq \quad 0 \end{cases}$$

has no solutions.

The general question above is too difficult, and it makes sense to pass from it to a seemingly simpler one:

(??) How to certify that ($\mathcal{S}$) has, or does not have, a solution.

Imagine that you are very smart and know the correct answer to (?); how could you convince me that your answer is correct? What could be an "evident for everybody" validity certificate for your answer?

If your claim is that ($\mathcal{S}$) is solvable, a certificate could be just to point out a solution x^* to ($\mathcal{S}$). Given this certificate, one can substitute x^* into the system and check whether x^* indeed is a solution.

Assume now that your claim is that ($\mathcal{S}$) has no solutions. What could be a "simple certificate" of this claim? How one could certify a *negative* statement? This is a highly nontrivial problem not just for mathematics; for example, in criminal law: how should someone accused in a murder prove his innocence? The "real life" answer to the question "how to certify a negative statement" is discouraging: such a statement normally *cannot* be certified (this is where the rule “a person is presumed innocent until proven guilty” comes from). In mathematics, however, the situation is different: in some cases there exist “simple certificates” of negative statements. E.g., in order to certify that $(\mathcal{S})$ has no solutions, it suffices to demonstrate that a consequence of $(\mathcal{S})$ is a contradictory inequality such as

$$-1 \geq 0.$$

For example, assume that λ_i, $i = 1, ..., m$, are nonnegative weights. Combining inequalities from $(\mathcal{S})$ with these weights, we come to the inequality

$$\sum_{i=1}^{m} \lambda_i f_i(x) \;\Omega\; 0 \tag{Comb(λ)}$$

where Ω is either " $>$ " (this is the case when the weight of at least one strict inequality from $(\mathcal{S})$ is positive), or " $\geq$ " (otherwise). Since the resulting inequality, due to its origin, is a consequence of the system $(\mathcal{S})$, i.e., it is satisfied by every solution to $(\mathcal{S})$, it follows that if (Comb(λ)) has no solutions at all, we can be sure that $(\mathcal{S})$ has no solution. Whenever this is the case, we may treat the corresponding vector λ as a “simple certificate” of the fact that $(\mathcal{S})$ is infeasible.

Let us look what does the outlined approach mean when $(\mathcal{S})$ is comprised of *linear* inequalities:

$$(\mathcal{S}): \quad \{a_i^T x \;\Omega_i\; b_i,\ i = 1, ..., m\} \qquad \left[\Omega_i = \left\{ \begin{array}{c} " > " \\ " \geq " \end{array} \right. \qquad \right]$$

Here the “combined inequality” is linear as well:

$$(\text{Comb}(\lambda)): \qquad (\sum_{i=1}^{m} \lambda a_i)^T x \;\Omega\; \sum_{i=1}^{m} \lambda b_i$$

(Ω is " $>$ " whenever $\lambda_i > 0$ for at least one i with $\Omega_i =$ " $>$ ", and Ω is " $\geq$ " otherwise). Now, when can a *linear* inequality

$$d^T x \;\Omega\; e$$

be contradictory? Of course, it can happen only when $d = 0$. Whether in this case the inequality is contradictory, it depends on what is the relation Ω: if $\Omega =$ " $>$ ", then the inequality is contradictory iff $e \geq 0$, and if $\Omega =$ " $\geq$ ", it is contradictory iff $e > 0$. We have established the following simple result:

Proposition B.2.1 *Consider a system of linear inequalities*

$$(\mathcal{S}): \qquad \left\{ \begin{array}{lcl} a_i^T x & > & b_i,\ i = 1, ..., m_s, \\ a_i^T x & \geq & b_i,\ i = m_s + 1, ..., m. \end{array} \right.$$

with n-dimensional vector of unknowns x. Let us associate with $(\mathcal{S})$ two systems of linear inequalities and equations with m-dimensional vector of unknowns λ:

$$\mathcal{T}_{\rm I}: \quad \left\{ \begin{array}{lrcl} (a) & \lambda & \geq & 0; \\ (b) & \sum\limits_{i=1}^{m} \lambda_i a_i & = & 0; \\ (c_{\rm I}) & \sum\limits_{i=1}^{m} \lambda_i b_i & \geq & 0; \\ \hline (d_{\rm I}) & \sum\limits_{i=1}^{m_s} \lambda_i & > & 0. \end{array} \right.$$

$$\mathcal{T}_{\rm II}: \quad \left\{ \begin{array}{lrcl} (a) & \lambda & \geq & 0; \\ (b) & \sum\limits_{i=1}^{m} \lambda_i a_i & = & 0; \\ \hline (c_{\rm II}) & \sum\limits_{i=1}^{m} \lambda_i b_i & > & 0. \end{array} \right.$$

Assume that at least one of the systems $\mathcal{T}_{\rm I}$, $\mathcal{T}_{\rm II}$ is solvable. Then the system $(\mathcal{S})$ is infeasible.

B.2.6.2 General Theorem on Alternative

Proposition B.2.1 says that in some cases it is easy to certify infeasibility of a linear system of inequalities: a "simple certificate" is a solution to another system of linear inequalities. Note, however, that the existence of a certificate of this latter type is to the moment only a *sufficient*, but not a *necessary*, condition for the infeasibility of $(\mathcal{S})$. A fundamental result in the theory of linear inequalities is that the sufficient condition in question is in fact also necessary:

Theorem B.2.6 [General Theorem on Alternative] *In the notation from Proposition B.2.1, system $(\mathcal{S})$ has no solutions iff either $\mathcal{T}_{\rm I}$, or $\mathcal{T}_{\rm II}$, or both these systems, are solvable.*

Proof. GTA is a more or less straightforward corollary of the Homogeneous Farkas Lemma. Indeed, in view of Proposition B.2.1, all we need to prove is that if $(\mathcal{S})$ has no solution, then at least one of the systems $\mathcal{T}_{\rm I}$, or $\mathcal{T}_{\rm II}$ is solvable. Thus, assume that $(\mathcal{S})$ has no solutions, and let us look at the consequences. Let us associate with $(\mathcal{S})$ the following system of *homogeneous* linear inequalities in variables x, τ, ϵ:

$$\begin{array}{llllll}(a) & & \tau & -\epsilon & \geq & 0\\ (b) & a_i^Tx & -b_i\tau & -\epsilon & \geq & 0,\ i=1,...,m_{\rm s}\\ (c) & a_i^Tx & -b_i\tau & & \geq & 0,\ i=m_{\rm s}+1,...,m\end{array}\tag{B.2.3}$$

We claim that

> (!) *For every solution to* (B.2.3), *one has* $\epsilon\leq 0$.
>
> Indeed, assuming that (B.2.3) has a solution x,τ,ϵ with $\epsilon>0$, we conclude from (B.2.3.a) that $\tau>0$; from (B.2.3.$b-c$) it now follows that $\tau^{-1}x$ is a solution to $(\mathcal{S})$, while the latter system is unsolvable.

Now, (!) says that the homogeneous linear inequality

$$-\epsilon\geq 0\tag{B.2.4}$$

is a consequence of the system of homogeneous linear inequalities (B.2.3). By Homogeneous Farkas Lemma, it follows that there exist nonnegative weights ν, λ_i, $i=1,...,m$, such that the vector of coefficients of the variables x,τ,ϵ in the left hand side of (B.2.4) is linear combination, with the coefficients $\nu,\lambda_1,...,\lambda_m$, of the vectors of coefficients of the variables in the inequalities from (B.2.3):

$$\begin{array}{lrcl}(a) & \sum\limits_{i=1}^m\lambda_ia_i & = & 0\\ (b) & -\sum\limits_{i=1}^m\lambda_ib_i+\nu & = & 0\\ (c) & -\sum\limits_{i=1}^{m_{\rm s}}\lambda_i-\nu & = & -1\end{array}\tag{B.2.5}$$

Recall that by their origin, ν and all λ_i are nonnegative. Now, it may happen that $\lambda_1,...,\lambda_{m_{\rm s}}$ are zero. In this case $\nu>0$ by (B.2.5.c), and relations (B.2.5$a-b$) say that $\lambda_1,...,\lambda_m$ solve $\mathcal{T}_{\rm II}$. In the remaining case (that is, when not all $\lambda_1,...,\lambda_{m_{\rm s}}$ are zero, or, which is the same, when $\sum\limits_{i=1}^{m_{\rm s}}\lambda_i>0$), the same relations say that $\lambda_1,...,\lambda_m$ solve $\mathcal{T}_{\rm I}$. □

B.2.6.3 Corollaries of the Theorem on Alternative

We formulate here explicitly two very useful principles following from the Theorem on Alternative:

> **A.** *A system of linear inequalities*
>
> $$a_i^Tx\ \Omega_i\ b_i,\ i=1,...,m$$
>
> *has no solutions iff one can combine the inequalities of the system in a linear fashion (i.e., multiplying the inequalities by nonnegative weights, adding the results and passing, if necessary, from an inequality $a^Tx>b$ to the inequality $a^Tx\geq b$) to get a contradictory inequality, namely, either the inequality $0^Tx\geq 1$, or the inequality $0^Tx>0$.*

B. *A linear inequality*

$$a_0^T x \; \Omega_0 \; b_0$$

is a consequence of a solvable *system of linear inequalities*

$$a_i^T x \; \Omega_i \; b_i, \;\; i = 1, ..., m$$

iff it can be obtained by combining, in a linear *fashion, the inequalities of the system and the trivial inequality* $0 > -1$.

It should be stressed that the above principles are highly nontrivial and very deep. Consider, e.g., the following system of 4 linear inequalities with two variables u, v:

$$\begin{array}{c} -1 \le u \le 1 \\ -1 \le v \le 1. \end{array}$$

From these inequalities it follows that

$$u^2 + v^2 \le 2, \qquad (!)$$

which in turn implies, by the Cauchy inequality, the linear inequality $u + v \le 2$:

$$u + v = 1 \times u + 1 \times v \le \sqrt{1^2 + 1^2}\sqrt{u^2 + v^2} \le (\sqrt{2})^2 = 2. \qquad (!!)$$

The concluding inequality is linear and is a consequence of the original system, but in the demonstration of this fact both steps (!) and (!!) are "highly nonlinear". It is absolutely unclear a priori why the same consequence can, as it is stated by Principle **A**, be derived from the system in a linear manner as well [of course it can – it suffices just to add two inequalities $u \le 1$ and $v \le 1$].

Note that the Theorem on Alternative and its corollaries **A** and **B** heavily exploit the fact that we are speaking about *linear* inequalities. E.g., consider the following 2 quadratic and 2 linear inequalities with two variables:

$$\begin{array}{lrcl} (a) & u^2 & \ge & 1; \\ (b) & v^2 & \ge & 1; \\ (c) & u & \ge & 0; \\ (d) & v & \ge & 0; \end{array}$$

along with the quadratic inequality

$$\begin{array}{lrcl} (e) & uv & \ge & 1. \end{array}$$

The inequality (e) is clearly a consequence of (a) – (d). However, if we extend the system of inequalities (a) – (b) by all "trivial" (i.e., identically true) linear and quadratic inequalities with 2 variables, like $0 > -1$, $u^2 + v^2 \ge 0$, $u^2 + 2uv + v^2 \ge 0$, $u^2 - uv + v^2 \ge 0$, etc., and ask whether (e) can be derived in a *linear* fashion from the inequalities of the extended system, the answer will be negative. Thus, Principle **A** fails to be true already for quadratic inequalities (which is a great sorrow – otherwise there were no difficult problems at all!)

B.2.7 Application: Linear Programming Duality

We are about to use the Theorem on Alternative to obtain the basic results of the LP duality theory.

B.2.7.1 Dual to an LP program: the origin

The motivation for constructing the problem dual to an LP program

$$c^* = \min_x \left\{ c^T x : Ax - b \ge 0 \right\} \quad \left[A = \begin{bmatrix} a_1^T \\ a_2^T \\ \dots \\ a_m^T \end{bmatrix} \in \mathbf{R}^{m \times n} \right] \qquad \text{(LP)}$$

is the desire to generate, in a systematic way, lower bounds on the optimal value c^* of (LP). An evident way to bound from below a given function $f(x)$ in the domain given by system of inequalities

$$g_i(x) \geq b_i,\ i = 1, ..., m, \tag{B.2.6}$$

is offered by what is called the *Lagrange duality* and is as follows:
Lagrange Duality:

- *Let us look at all inequalities which can be obtained from* (B.2.6) *by linear aggregation, i.e., at the inequalities of the form*

$$\sum_i y_i g_i(x) \geq \sum_i y_i b_i \tag{B.2.7}$$

with the "aggregation weights" $y_i \geq 0$. Note that the inequality (B.2.7)*, due to its origin, is valid on the entire set X of solutions of* (B.2.6)*.*

- *Depending on the choice of aggregation weights, it may happen that the left hand side in* (B.2.7) *is $\leq f(x)$ $\underline{\text{for all}}$ $x \in \mathbf{R}^n$. Whenever it is the case, the right hand side $\sum\limits_i y_i b_i$ of* (B.2.7) *is a lower bound on f in X.*

> Indeed, $\underline{\text{on } X}$ the quantity $\sum\limits_i y_i b_i$ is a lower bound on $y_i g_i(x)$, and for y in question the latter function of x is $\underline{\text{everywhere}}$ $\leq f(x)$.

It follows that

- *The optimal value in the problem*

$$\max_y \left\{ \sum_i y_i b_i : \begin{array}{rr} y \geq 0, & (a) \\ \sum\limits_i y_i g_i(x) \leq f(x)\ \forall x \in \mathbf{R}^n & (b) \end{array} \right\} \tag{B.2.8}$$

is a lower bound on the values of f on the set of solutions to the system (B.2.6).

Let us look what happens with the Lagrange duality when f and g_i are homogeneous linear functions: $f = c^T x$, $g_i(x) = a_i^T x$. In this case, the requirement (B.2.8.b) merely says that $c = \sum\limits_i y_i a_i$ (or, which is the same, $A^T y = c$ due to the origin of A). Thus, problem (B.2.8) becomes the Linear Programming problem

$$\max_y \left\{ b^T y : A^T y = c,\ y \geq 0 \right\}, \tag{LP*}$$

which is nothing but the LP dual of (LP).

By the construction of the dual problem,

> [Weak Duality] *The optimal value in* (LP*) *is less than or equal to the optimal value in* (LP).

In fact, the "less than or equal to" in the latter statement is "equal", provided that the optimal value c^* in (LP) is a number (i.e., (LP) is feasible and below bounded). To see that this indeed is the case, note that a real a is a lower bound on c^* iff $c^T x \geq a$ whenever $Ax \geq b$, or, which is the same, iff the system of linear inequalities

$$(\mathcal{S}_a): \qquad -c^T x > -a, Ax \geq b$$

has no solution. We know by the Theorem on Alternative that the latter fact means that some other system of linear equalities (more exactly, at least one of a certain pair of systems) does have a solution. More precisely,

> (*) $(\mathcal{S}_a)$ *has no solutions iff at least one of the following two systems with $m+1$ unknowns:*

$$\mathcal{T}_{\rm I}: \left\{ \begin{array}{lrcl} (a) & \lambda = (\lambda_0, \lambda_1, ..., \lambda_m) & \geq & 0; \\ (b) & -\lambda_0 c + \sum\limits_{i=1}^m \lambda_i a_i & = & 0; \\ \hline (c_{\rm I}) & -\lambda_0 a + \sum\limits_{i=1}^m \lambda_i b_i & \geq & 0; \\ (d_{\rm I}) & \lambda_0 & > & 0, \end{array} \right.$$

or

$$\mathcal{T}_{\rm II}: \quad \left\{ \begin{array}{lrcl} (a) & \lambda = (\lambda_0, \lambda_1, ..., \lambda_m) & \geq & 0; \\ (b) & -\lambda_0 c - \sum\limits_{i=1}^m \lambda_i a_i & = & 0; \\ \hline (c_{\rm II}) & -\lambda_0 a - \sum\limits_{i=1}^m \lambda_i b_i & > & 0 \end{array} \right.$$

– has a solution.

Now assume that (LP) *is feasible.* We claim that *under this assumption* $(\mathcal{S}_a)$ *has no solutions if and only if* $\mathcal{T}_{\rm I}$ *has a solution.*

> The implication "$\mathcal{T}_{\rm I}$ has a solution $\Rightarrow$ $(\mathcal{S}_a)$ has no solution" is readily given by the above remarks. To verify the inverse implication, assume that $(\mathcal{S}_a)$ has no solutions and the system $Ax \leq b$ has a solution, and let us prove that then $\mathcal{T}_{\rm I}$ has a solution. If $\mathcal{T}_{\rm I}$ has no solution, then by (*) $\mathcal{T}_{\rm II}$ has a solution and, moreover, $\lambda_0 = 0$ for (every) solution to $\mathcal{T}_{\rm II}$ (since a solution to the latter system with $\lambda_0 > 0$ solves $\mathcal{T}_{\rm I}$ as well). But the fact that $\mathcal{T}_{\rm II}$ has a solution λ with $\lambda_0 = 0$ is independent of the values of a and c; if this fact would take place, it would mean, by the same Theorem on Alternative, that, e.g., the following instance of $(\mathcal{S}_a)$:
>
> $$0^T x \geq -1, Ax \geq b$$
>
> has no solutions. The latter means that the system $Ax \geq b$ has no solutions – a contradiction with the assumption that (LP) is feasible. □

Now, if $\mathcal{T}_{\rm I}$ has a solution, this system has a solution with $\lambda_0 = 1$ as well (to see this, pass from a solution λ to the one λ/λ_0; this construction is well-defined, since $\lambda_0 > 0$ for every solution to $\mathcal{T}_{\rm I}$). Now, an $(m+1)$-dimensional vector $\lambda = (1, y)$ is a solution to $\mathcal{T}_{\rm I}$ if and only if the m-dimensional vector y solves the system of linear inequalities and equations

$$\begin{array}{rcl} y & \geq & 0; \\ A^T y \equiv \sum\limits_{i=1}^m y_i a_i & = & c; \\ b^T y & \geq & a \end{array} \tag{D}$$

Summarizing our observations, we come to the following result.

Proposition B.2.2 *Assume that system* (D) *associated with the LP program* (LP) *has a solution* (y, a)*. Then* a *is a lower bound on the optimal value in* (LP)*. Vice versa, if* (LP) *is feasible and* a *is a lower bound on the optimal value of* (LP)*, then* a *can be extended by a properly chosen* m*-dimensional vector* y *to a solution to* (D)*.*

We see that the entity responsible for lower bounds on the optimal value of (LP) is the system (D): every solution to the latter system induces a bound of this type, and *in the case when* (LP) *is feasible*, all lower bounds can be obtained from solutions to (D). Now note that if (y, a) is a solution to (D), then the pair $(y, b^T y)$ also is a solution to the same system, and the lower bound $b^T y$ on c^* is not worse than the lower bound a. Thus, as far as lower bounds on c^* are concerned, we lose nothing by restricting ourselves to the solutions (y, a) of (D) with $a = b^T y$; the best lower bound on c^* given by (D) is therefore the optimal value of the problem $\max_y \left\{ b^T y : A^T y = c, y \geq 0 \right\}$, which is nothing but the dual to (LP) problem (LP*). Note that (LP*) is also a Linear Programming program.

All we know about the dual problem to the moment is the following:

Proposition B.2.3 *Whenever* y *is a feasible solution to* (LP*)*, the corresponding value of the dual objective* $b^T y$ *is a lower bound on the optimal value* c^* *in* (LP)*. If* (LP) *is feasible, then for every* $a \leq c^*$ *there exists a feasible solution* y *of* (LP*) *with* $b^T y \geq a$*.*

B.2.7.2 Linear Programming Duality Theorem

Proposition B.2.3 is in fact equivalent to the following

Theorem B.2.7 [Duality Theorem in Linear Programming] *Consider a linear programming program*

$$\min_x \left\{c^T x : Ax \geq b\right\} \tag{LP}$$

along with its dual

$$\max_y \left\{b^T y : A^T y = c, y \geq 0\right\} \tag{LP*}$$

Then

1) *The duality is symmetric: the problem dual to dual is equivalent to the primal;*

2) *The value of the dual objective at every dual feasible solution is* $\leq$ *the value of the primal objective at every primal feasible solution*

3) *The following 5 properties are equivalent to each other:*

(i) *The primal is feasible and bounded below.*

(ii) *The dual is feasible and bounded above.*

(iii) *The primal is solvable.*

(iv) *The dual is solvable.*

(v) *Both primal and dual are feasible.*

Whenever (i) $\equiv$ (ii) $\equiv$ (iii) $\equiv$ (iv) $\equiv$ (v) *is the case, the optimal values of the primal and the dual problems are equal to each other.*

Proof. 1) is quite straightforward: writing the dual problem (LP*) in our standard form, we get

$$\min_y \left\{ -b^T y : \begin{bmatrix} I_m \\ A^T \\ -A^T \end{bmatrix} y - \begin{bmatrix} 0 \\ -c \\ c \end{bmatrix} \geq 0 \right\},$$

where I_m is the m-dimensional unit matrix. Applying the duality transformation to the latter problem, we come to the problem

$$\max_{\xi,\eta,\zeta} \left\{ 0^T\xi + c^T\eta + (-c)^T\zeta : \begin{array}{rcl} \xi & \geq & 0 \\ \eta & \geq & 0 \\ \zeta & \geq & 0 \\ \xi - A\eta + A\zeta & = & -b \end{array} \right\},$$

which is clearly equivalent to (LP) (set $x = \eta - \zeta$).

2) is readily given by Proposition B.2.3.

3):

(i)$\Rightarrow$(iv): If the primal is feasible and bounded below, its optimal value c^* (which of course is a lower bound on itself) can, by Proposition B.2.3, be (non-strictly) majorized by a quantity $b^T y^*$, where y^* is a feasible solution to (LP*). In the situation in question, of course, $b^T y^* = c^*$ (by already proved item 2)); on the other hand, in view of the same Proposition B.2.3, the optimal value in the dual is $\leq c^*$. We conclude that the optimal value in the dual is attained and is equal to the optimal value in the primal.

(iv)$\Rightarrow$(ii): evident;

(ii)$\Rightarrow$(iii): This implication, in view of the primal-dual symmetry, follows from the implication (i)$\Rightarrow$(iv).

(iii)$\Rightarrow$(i): evident.

We have seen that (i)$\equiv$(ii)$\equiv$(iii)$\equiv$(iv) and that the first (and consequently each) of these 4 equivalent properties implies that the optimal value in the primal problem is equal to the optimal value in the dual one. All which remains is to prove the equivalence between (i)–(iv), on one hand, and (v), on the other hand. This is immediate: (i)–(iv), of course, imply (v); vice versa, in the case of (v) the primal is not only feasible, but also bounded below (this is an immediate consequence of the feasibility of the dual problem, see 2)), and (i) follows. □

An immediate corollary of the LP Duality Theorem is the following *necessary and sufficient* optimality condition in LP:

Theorem B.2.8 [Necessary and sufficient optimality conditions in linear programming] *Consider an LP program* (LP) *along with its dual* (LP*). *A pair* (x, y) *of primal and dual feasible solutions is comprised of optimal solutions to the respective problems iff*

$$y_i[Ax - b]_i = 0,\ \ i = 1, ..., m, \qquad \text{[complementary slackness]}$$

likewise as iff

$$c^T x - b^T y = 0 \qquad \text{[zero duality gap]}$$

Indeed, the "zero duality gap" optimality condition is an immediate consequence of the fact that the value of primal objective at every primal feasible solution is $\geq$ the value of the dual objective at every dual feasible solution, while the optimal values in the primal and the dual are equal to each other, see Theorem B.2.7. The equivalence between the "zero duality gap" and the "complementary slackness" optimality conditions is given by the following computation: whenever x is primal feasible and y is dual feasible, the products $y_i[Ax - b]_i$, $i = 1, ..., m$, are nonnegative, while the sum of these products is precisely the duality gap:

$$y^T[Ax - b] = (A^T y)^T x - b^T y = c^T x - b^T y.$$

Thus, the duality gap can vanish at a primal-dual feasible pair (x, y) iff all products $y_i[Ax - b]_i$ for this pair are zeros.

B.2.8 Separation Theorem

B.2.8.1 Separation: definition

Recall that a *hyperplane* M in $\mathbf{R}^n$ is, by definition, an affine subspace of the dimension $n-1$. By Proposition A.3.7, hyperplanes are exactly the same as level sets of nontrivial linear forms:

$$\begin{array}{c} M \subset \mathbf{R}^n \text{ is a hyperplane} \\ \Updownarrow \\ \exists a \in \mathbf{R}^n, b \in \mathbf{R}, a \neq 0 : \quad M = \{x \in \mathbf{R}^n : a^T x = b\} \end{array}$$

We can, consequently, associate with the hyperplane (or, better to say, with the associated linear form a; this form is defined uniquely, up to multiplication by a nonzero real) the following sets:

- "upper" and "lower" open half-spaces $M^{++} = \{x \in \mathbf{R}^n : a^T x > b\}$, $M^{--} = \{x \in \mathbf{R}^n : a^T x < b\}$;

  these sets clearly are convex, and since a linear form is continuous, and the sets are given by strict inequalities on the value of a continuous function, they indeed are open.

  Note that since a is uniquely defined by M, up to multiplication by a nonzero real, these open half-spaces are uniquely defined by the hyperplane, up to swapping the "upper" and the "lower" ones (which half-space is "upper", it depends on the particular choice of a);

- "upper" and "lower" closed half-spaces $M^{+} = \{x \in \mathbf{R}^n : a^T x \geq b\}$, $M^{-} = \{x \in \mathbf{R}^n : a^T x \leq b\}$;

  these are also convex sets, now closed (since they are given by non-strict inequalities on the value of a continuous function). It is easily seen that the closed upper/lower half-space is the closure of the corresponding open half-space, and M itself is the boundary (i.e., the complement of the interior to the closure) of all four half-spaces.

It is clear that our half-spaces and M itself partition $\mathbf{R}^n$:

$$\mathbf{R}^n = M^{--} \cup M \cup M^{++}$$

(partitioning by disjoint sets),

$$\mathbf{R}^n = M^{-} \cup M^{+}$$

(M is the intersection of the right hand side sets).

Now we define the basic notion of *separation* of two convex sets T and S by a hyperplane.

Definition B.2.2 [separation] *Let* S, T *be two nonempty convex sets in* $\mathbf{R}^n$.

- *A hyperplane*
$$M = \{x \in \mathbf{R}^n : a^T x = b\} \quad [a \neq 0]$$
is said to separate S and T, if, first,
$$S \subset \{x : a^T x \leq b\}, \quad T \subset \{x : a^T x \geq b\}$$
(i.e., S and T belong to the opposite closed half-spaces into which M splits $\mathbf{R}^n$), and, second, at least one of the sets S, T is not contained in M itself:
$$S \cup T \not\subset M.$$
The separation is called strong, if there exist b', b'', $b' < b < b''$, such that
$$S \subset \{x : a^T x \leq b'\}, \quad T \subset \{x : a^T x \geq b''\}.$$

- *A linear form $a \neq 0$ is said to separate (strongly separate) S and T, if for properly chosen b the hyperplane $\{x : a^T x = b\}$ separates (strongly separates) S and T.*
- *We say that S and T can be (strongly) separated, if there exists a hyperplane which (strongly) separates S and T.*

E.g.,

- the hyperplane $\{x : a^T x \equiv x_2 - x_1 = 1\}$ in $\mathbf{R}^2$ strongly separates convex polyhedral sets $T = \{x \in \mathbf{R}^2 : 0 \leq x_1 \leq 1, 3 \leq x_2 \leq 5\}$ and $S = \{x \in \mathbf{R}^2 : x_2 = 0; x_1 \geq -1\}$;
- the hyperplane $\{x : a^T x \equiv x = 1\}$ in $\mathbf{R}^1$ separates (but not strongly separates) the convex sets $S = \{x \leq 1\}$ and $T = \{x \geq 1\}$;
- the hyperplane $\{x : a^T x \equiv x_1 = 0\}$ in $\mathbf{R}^2$ separates (but not strongly separates) the sets $S = \{x \in \mathbf{R}^2 :, x_1 < 0, x_2 \geq -1/x_1\}$ and $T = \{x \in \mathbf{R}^2 : x_1 > 0, x_2 > 1/x_1\}$;
- the hyperplane $\{x : a^T x \equiv x_2 - x_1 = 1\}$ in $\mathbf{R}^2$ does *not* separate the convex sets $S = \{x \in \mathbf{R}^2 : x_2 \geq 1\}$ and $T = \{x \in \mathbf{R}^2 : x_2 = 0\}$;
- the hyperplane $\{x : a^T x \equiv x_2 = 0\}$ in $\mathbf{R}^2$ does not separate the sets $S = \{x \in \mathbf{R}^2 : x_2 = 0, x_1 \leq -1\}$ and $T = \{x \in \mathbf{R}^2 : x_2 = 0, x_1 \geq 1\}$.

The following Exercise presents an equivalent description of separation:

Exercise B.11 *Let S, T be nonempty convex sets in $\mathbf{R}^n$. Prove that a linear form a separates S and T iff*
$$\sup_{x \in S} a^T x \leq \inf_{y \in T} a^T y$$
and
$$\inf_{x \in S} a^T x < \sup_{y \in T} a^T y.$$
This separation is strong iff
$$\sup_{x \in S} a^T x < \inf_{y \in T} a^T y.$$

Exercise B.12 *Whether the sets $S = \{x \in \mathbf{R}^2 : x_1 > 0, x_2 \geq 1/x_1\}$ and $T = \{x \in \mathbf{R}^2 : x_1 < 0, x_2 \geq -1/x_1\}$ can be separated? Whether they can be strongly separated?*

B.2.8.2 Separation Theorem

Theorem B.2.9 [Separation Theorem] *Let S and T be nonempty convex sets in $\mathbf{R}^n$.*
(i) *S and T can be separated iff their relative interiors do not intersect: $\operatorname{ri} S \cap \operatorname{ri} T = \emptyset$.*
(ii) *S and T can be strongly separated iff the sets are at a positive distance from each other:*
$$\operatorname{dist}(S, T) \equiv \inf\{\|x - y\|_2 : x \in S, y \in T\} > 0.$$
In particular, if S, T are closed nonempty non-intersecting convex sets and one of these sets is compact, S and T can be strongly separated.

Proof takes several steps.

(i), Necessity. Assume that S, T can be separated, so that for certain $a \neq 0$ we have

$$\inf_{x\in S} a^T x \leq \inf_{y\in T} a^T y; \quad \inf_{x\in S} a^T x < \sup_{y\in T} a^T y. \tag{B.2.9}$$

We should lead to a contradiction the assumption that $\mathrm{ri}\, S$ and $\mathrm{ri}\, T$ have in common certain point $\bar{x}$. Assume that it is the case; then from the first inequality in (B.2.9) it is clear that $\bar{x}$ maximizes the linear function $f(x) = a^T x$ on S and simultaneously minimizes this function on T. Now, we have the following simple and important

Lemma B.2.2 *A linear function $f(x) = a^T x$ can attain its maximum/minimum over a convex set Q at a point $x \in \mathrm{ri}\, Q$ if and only if the function is constant on Q.*

Proof. "if" part is evident. To prove the "only if" part, let $\bar{x} \in \mathrm{ri}\, Q$ be, say, a minimizer of f over Q and y be an arbitrary point of Q; we should prove that $f(\bar{x}) = f(y)$. There is nothing to prove if $y = \bar{x}$, so let us assume that $y \neq \bar{x}$. Since $\bar{x} \in \mathrm{ri}\, Q$, the segment $[y, \bar{x}]$, which is contained in M, can be extended a little bit through the point $\bar{x}$, not leaving M (since $\bar{x} \in \mathrm{ri}\, Q$), so that there exists $z \in Q$ such that $\bar{x} \in [y, z)$, i.e., $\bar{x} = (1-\lambda)y + \lambda z$ with certain $\lambda \in (0, 1]$; since $y \neq \bar{x}$, we have in fact $\lambda \in (0, 1)$. Since f is linear, we have

$$f(\bar{x}) = (1-\lambda)f(y) + \lambda f(z);$$

since $f(\bar{x}) \leq \min\{f(y), f(z)\}$ and $0 < \lambda < 1$, this relation can be satisfied only when $f(\bar{x}) = f(y) = f(z)$. □

By Lemma B.2.2, $f(x) = f(\bar{x})$ on S and on T, so that $f(\cdot)$ is constant on $S \cup T$, which yields the desired contradiction with the second inequality in (B.2.9). □

(i), Sufficiency. The proof of sufficiency part of the Separation Theorem is much more instructive. There are several ways to prove it, and We choose the one which goes via the Homogeneous Farkas Lemma B.2.1, which is extremely important in its own right.

(i), Sufficiency, Step 1: Separation of a convex polytope and a point outside the polytope. Let us start with seemingly very particular case of the Separation Theorem – the one where S is the convex full points $x_1, ..., x_N$, and T is a singleton $T = \{x\}$ which does not belong to S. We intend to prove that in this case there exists a linear form which separates x and S; in fact we shall prove even the existence of strong separation.

Let us associate with n-dimensional vectors $x_1, ..., x_N, x$ the $(n+1)$-dimensional vectors $a = \begin{bmatrix} x \\ 1 \end{bmatrix}$ and $a_i = \begin{bmatrix} x_i \\ 1 \end{bmatrix}$, $i = 1, ..., N$. We claim that a does not belong to the conic hull of $a_1, ..., a_N$. Indeed, if a would be representable as a linear combination of $a_1, ..., a_N$ with nonnegative coefficients, then, looking at the last, $(n+1)$-st, coordinates in such a representation, we would conclude that the sum of coefficients should be 1, so that the representation, actually, represents x as a convex combination of $x_1, ..., x_N$, which was assumed to be impossible.

Since a does not belong to the conic hull of $a_1, ..., a_N$, by the Homogeneous Farkas Lemma (Lemma B.2.1) there exists a vector $h = \begin{bmatrix} f \\ \alpha \end{bmatrix} \in \mathbf{R}^{n+1}$ which "separates" a and $a_1, ..., a_N$ in the sense that

$$h^T a > 0, \ h^T a_i \leq 0, \ i = 1, ..., N,$$

whence, of course,

$$h^T a > \max_i h^T a_i.$$

Since the components in all the inner products $h^T a, h^T a_i$ coming from the $(n+1)$-st coordinates are equal to each other , we conclude that the n-dimensional component f of h separates x and $x_1, ..., x_N$:

$$f^T x > \max_i f^T x_i.$$

Since for every convex combination $y = \sum\limits_i \lambda_i x_i$ of the points x_i one clearly has $f^T y \leq \max\limits_i f^T x_i$, we conclude, finally, that

$$f^T x > \max_{y \in \mathrm{Conv}(\{x_1,...,x_N\})} f^T y,$$

so that f strongly separates $T = \{x\}$ and $S = \mathrm{Conv}(\{x_1, ..., x_N\})$. □

(i), Sufficiency, Step 2: Separation of a convex set and a point outside of the set. Now consider the case when S is an arbitrary nonempty convex set and $T = \{x\}$ is a singleton outside S (the difference with Step 1 is that now S is not assumed to be a polytope).

First of all, without loss of generality we may assume that S contains 0 (if it is not the case, we may subject S and T to translation $S \mapsto p + S$, $T \mapsto p + T$ with $p \in -S$). Let L be the linear span of S. If $x \notin L$, the separation is easy: taking as f the orthogonal to L component of x, we shall get

$$f^T x = f^T f > 0 = \max_{y \in S} f^T y,$$

so that f strongly separates S and $T = \{x\}$.

It remains to consider the case when $x \in L$. Since $S \subset L$, $x \in L$ and $x \notin S$, L is a nonzero linear subspace; w.l.o.g., we can assume that $L = \mathbf{R}^n$.

Let $\Sigma = \{h : \|h\|_2 = 1\}$ be the unit sphere in $L = \mathbf{R}^n$. This is a closed and bounded set in $\mathbf{R}^n$ (boundedness is evident, and closedness follows from the fact that $\|\cdot\|_2$ is continuous). Consequently, Σ is a compact set. Let us prove that there exists $f \in \Sigma$ which separates x and S in the sense that

$$f^T x \geq \sup_{y \in S} f^T y. \tag{B.2.10}$$

Assume, on the contrary, that no such f exists, and let us lead this assumption to a contradiction. Under our assumption for every $h \in \Sigma$ there exists $y_h \in S$ such that

$$h^T y_h > h^T x.$$

Since the inequality is strict, it immediately follows that there exists a neighbourhood U_h of the vector h such that

$$(h')^T y_h > (h')^T x \quad \forall h' \in U_h. \tag{B.2.11}$$

The family of open sets $\{U_h\}_{h \in \Sigma}$ covers Σ; since Σ is compact, we can find a finite subfamily $U_{h_1}, ..., U_{h_N}$ of the family which still covers Σ. Let us take the corresponding points $y_1 = y_{h_1}, y_2 = y_{h_2}, ..., y_N = y_{h_N}$ and the polytope $S' = \mathrm{Conv}(\{y_1, ..., y_N\})$ spanned by the points. Due to the origin of y_i, all of them are points from S; since S is convex, the polytope S' is contained in S and, consequently, does not contain x. By Step 1, x can be strongly separated from S': there exists a such that

$$a^T x > \sup_{y \in S'} a^T y. \tag{B.2.12}$$

By normalization, we may also assume that $\|a\|_2 = 1$, so that $a \in \Sigma$. Now we get a contradiction: since $a \in \Sigma$ and $U_{h_1}, ..., U_{h_N}$ form a covering of Σ, a belongs to certain U_{h_i}. By construction of U_{h_i} (see (B.2.11)), we have

$$a^T y_i \equiv a^T y_{h_i} > a^T x,$$

which contradicts (B.2.12) – recall that $y_i \in S'$.

The contradiction we get proves that there exists $f \in \Sigma$ satisfying (B.2.10). We claim that f separates S and $\{x\}$; in view of (B.2.10), all we need to verify our claim is to show that the linear form $f(y) = f^T y$ is non-constant on $S \cup T$, which is evident: we are in the situation when $0 \in S$ and $L \equiv \mathrm{Lin}(S) = \mathbf{R}^n$ and $f \neq 0$, so that $f(y)$ is non-constant already on S. □

Mathematically oriented reader should take into account that the simple-looking reasoning underlying Step 2 in fact brings us into a completely new world. Indeed, the considerations at Step 1 and in the proof of Homogeneous Farkas Lemma are "pure arithmetic" – we never used things like convergence, compactness, etc., and used rational arithmetic only – no square roots, etc. It means that the Homogeneous Farkas Lemma and the result stated a Step 1 remain valid if we, e.g., replace our universe $\mathbf{R}^n$ with the space $\mathbf{Q}^n$ of n-dimensional rational vectors (those with rational coordinates; of course, the multiplication by reals in this space should be restricted to multiplication by rationals). The "rational" Farkas Lemma or the possibility to separate a rational vector from a "rational" polytope by a rational linear form, which is the "rational" version of the result of Step 1, definitely are of interest (e.g., for Integer Programming). In contrast to these "purely arithmetic" considerations, at Step 2 we used compactness – something heavily exploiting the fact that our universe is $\mathbf{R}^n$ and not, say, $\mathbf{Q}^n$ (in the latter space bounded and closed sets not necessary are compact). Note also that we could not avoid things like compactness arguments at Step 2, since the very fact we are proving is true in $\mathbf{R}^n$ but not in $\mathbf{Q}^n$. Indeed, consider the "rational plane" – the universe comprised of all 2-dimensional vectors with rational entries, and let S be the half-plane in this rational plane given by the linear inequality

$$x_1 + \alpha x_2 \leq 0,$$

where α is irrational. S clearly is a "convex set" in $\mathbf{Q}^2$; it is immediately seen that a point outside this set cannot be separated from S by a rational linear form.

(i), Sufficiency, Step 3: Separation of two nonempty and non-intersecting convex sets. Now we are ready to prove that two nonempty and non-intersecting convex sets S and T can be separated. To this end consider the arithmetic difference

$$\Delta = S - T = \{x - y : x \in S, y \in T\}.$$

By Proposition B.1.6.3, Δ is convex (and, of course, nonempty) set; since S and T do not intersect, Δ does not contain 0. By Step 2, we can separate Δ and $\{0\}$: there exists $f \neq 0$ such that

$$f^T 0 = 0 \geq \sup_{z\in\Delta} f^T z \;\&\; f^T 0 > \inf_{z\in\Delta} f^T z.$$

In other words,

$$0 \geq \sup_{x\in S, y\in T} [f^T x - f^T y] \;\&\; 0 > \inf_{x\in S, y\in T} [f^T x - f^T y],$$

which clearly means that f separates S and T. □

(i), Sufficiency, Step 4: Separation of nonempty convex sets with non-intersecting relative interiors. Now we are able to complete the proof of the "if" part of the Separation Theorem. Let S and T be two nonempty convex sets with non-intersecting relative interiors; we should prove that S and T can be properly separated. This is immediate: as we know from Theorem B.1.1, the sets $S' = \operatorname{ri} S$ and $T' = \operatorname{ri} T$ are nonempty and convex; since we are given that they do not intersect, they can be separated by Step 3: there exists f such that

$$\inf_{x\in T'} f^T x \geq \sup_{y\in S'} f^T x \;\&\; \sup_{x\in T'} f^T x > \inf_{y\in S'} f^T x. \tag{B.2.13}$$

It is immediately seen that in fact f separates S and T. Indeed, the quantities in the left and the right hand sides of the first inequality in (B.2.13) clearly remain unchanged when we replace S' with $\operatorname{cl} S'$ and T' with $\operatorname{cl} T'$; by Theorem B.1.1, $\operatorname{cl} S' = \operatorname{cl} S \supset S$ and $\operatorname{cl} T' = \operatorname{cl} T \supset T$, and we get $\inf\limits_{x\in T} f^T x = \inf\limits_{x\in T'} f^T x$, and similarly $\sup\limits_{y\in S} f^T y = \sup\limits_{y\in S'} f^T y$. Thus, we get from (B.2.13)

$$\inf_{x\in T} f^T x \geq \sup_{y\in S} f^T y.$$

It remains to note that $T' \subset T$, $S' \subset S$, so that the second inequality in (B.2.13) implies that

$$\sup_{x\in T} f^T x > \inf_{y\in S} f^T x. \quad \square$$

(ii), Necessity: prove yourself.

(ii), Sufficiency: Assuming that $\rho \equiv \inf\{\|x-y\|_2 : x \in S, y \in T\} > 0$, consider the sets $S' = \{x : \inf\limits_{y\in S} \|x-y\|_2 \leq \rho\}$. Note that S' is convex along with S (Example B.4) and that $S' \cap T = \emptyset$ (why?) By (i), S' and T can be separated, and if f is a linear form which separates S' and T, then the same form strongly separates S and T (why?). The "in particular" part of (ii) readily follows from the just proved statement due to the fact that if two closed nonempty sets in $\mathbf{R}^n$ do not intersect and one of them is compact, then the sets are at positive distance from each other (why?). □

Exercise B.13 *Derive the statement in Remark B.1.1 from the Separation Theorem.*

Exercise B.14 *Implement the following alternative approach to the proof of Separation Theorem:*

1. *Prove that if x is a point in $\mathbf{R}^n$ and S is a nonempty closed convex set in $\mathbf{R}^n$, then the problem*

$$\min_y \{\|x-y\|_2 : y \in S\}$$

has a unique optimal solution $\bar{x}$.

2. *In the situation of 1), prove that if $x \notin S$, then the linear form $e = x - \bar{x}$ strongly separates $\{x\}$ and S:*

$$\max_{y\in S} e^T y = e^T \bar{x} = e^T x - e^T e < e^T x,$$

thus getting a direct proof of the possibility to separate strongly a nonempty closed convex set and a point outside this set.

3. *Derive from 2) the Separation Theorem.*

B.2.8.3 Supporting hyperplanes

By the Separation Theorem, a closed and nonempty convex set M is the intersection of all closed half-spaces containing M. Among these half-spaces, the most interesting are the "extreme" ones – those with the boundary hyperplane touching M. The notion makes sense for an arbitrary (not necessary closed) convex set, but we shall use it for closed sets only, and include the requirement of closedness in the definition:

Definition B.2.3 [Supporting plane] *Let M be a convex closed set in $\mathbf{R}^n$, and let x be a point from the relative boundary of M. A hyperplane*

$$\Pi = \{y : a^T y = a^T x\} \quad [a \neq 0]$$

is called supporting to M at x, if it separates M and $\{x\}$, i.e., if

$$a^T x \geq \sup_{y\in M} a^T y \quad \& \quad a^T x > \inf_{y\in M} a^T y. \tag{B.2.14}$$

Note that since x is a point from the relative boundary of M and therefore belongs to $\mathrm{cl}\, M = M$, the first inequality in (B.2.14) in fact is equality. Thus, an equivalent definition of a supporting plane is as follows:

> *Let M be a closed convex set and x be a relative boundary point of M. The hyperplane $\{y : a^T y = a^T x\}$ is called supporting to M at x, if the linear form $a(y) = a^T y$ attains its maximum on M at the point x and is nonconstant on M.*

E.g., the hyperplane $\{x_1 = 1\}$ in $\mathbf{R}^n$ clearly is supporting to the unit Euclidean ball $\{x : |x| \leq 1\}$ at the point $x = e_1 = (1, 0, ..., 0)$.

The most important property of a supporting plane is its existence:

Proposition B.2.4 [Existence of supporting hyperplane] *Let M be a convex closed set in $\mathbf{R}^n$ and x be a point from the relative boundary of M. Then*

(i) *There exists at least one hyperplane which is supporting to M at x;*

(ii) *If Π is supporting to M at x, then the intersection $M \cap \Pi$ is of affine dimension less than the one of M (recall that the affine dimension of a set is, by definition, the affine dimension of the affine hull of the set).*

Proof. (i) is easy: if x is a point from the relative boundary of M, then it is outside the relative interior of M and therefore $\{x\}$ and $\mathrm{ri}\, M$ can be separated by the Separation Theorem; the separating hyperplane is exactly the desired supporting to M at x hyperplane.

To prove (ii), note that if $\Pi = \{y : a^T y = a^T x\}$ is supporting to M at $x \in \partial_{\mathrm{ri}} M$, then the set $M' = M \cap \Pi$ is a nonempty (it contains x) convex set, and the linear form $a^T y$ is constant on M' and therefore (why?) on $\mathrm{Aff}(M')$. At the same time, the form is nonconstant on M by definition of a supporting plane. Thus, $\mathrm{Aff}(M')$ is a proper (less than the entire $\mathrm{Aff}(M)$) subset of $\mathrm{Aff}(M)$, and therefore the affine dimension of $\mathrm{Aff}(M')$ (i.e., the affine dimension of M') is less than the affine dimension of $\mathrm{Aff}(M)$ (i.e., than the affine dimension of M). [2]. □

B.2.9 Polar of a convex set and Milutin-Dubovitski Lemma

B.2.9.1 Polar of a convex set

Let M be a nonempty convex set in $\mathbf{R}^n$. The *polar* $\mathrm{Polar}\,(M)$ of M is the set of all linear forms which do not exceed 1 on M, i.e., the set of all vectors a such that $a^T x \le 1$ for all $x \in M$:

$$\mathrm{Polar}\,(M) = \{a : a^T x \le 1 \forall x \in M\}.$$

For example, $\mathrm{Polar}\,(\mathbf{R}^n) = \{0\}$, $\mathrm{Polar}\,(\{0\}) = \mathbf{R}^n$; if L is a liner subspace in $\mathbf{R}^n$, then $\mathrm{Polar}\,(L) = L^\perp$ (why?).

The following properties of the polar are evident:

1. $0 \in \mathrm{Polar}\,(M)$;
2. $\mathrm{Polar}\,(M)$ is convex;
3. $\mathrm{Polar}\,(M)$ is closed.

It turns out that these properties characterize polars:

Proposition B.2.5 *Every closed convex set M containing the origin is polar, specifically, it is polar of its polar:*

$$\begin{array}{c} M \text{ is closed and convex, } 0 \in M \\ \Updownarrow \\ M = \mathrm{Polar}\,(\mathrm{Polar}\,(M)) \end{array}$$

Proof. All we need is to prove that if M is closed and convex and $0 \in M$, then $M = \mathrm{Polar}\,(\mathrm{Polar}\,(M))$. By definition,

$$y \in \mathrm{Polar}\,(M), x \in M \Rightarrow y^T x \le 1,$$

so that $M \subset \mathrm{Polar}\,(\mathrm{Polar}\,(M))$. To prove that this inclusion is in fact equality, assume, on the contrary, that there exists $\bar{x} \in \mathrm{Polar}\,(\mathrm{Polar}\,(M)) \backslash M$. Since M is nonempty, convex and closed and $\bar{x} \notin M$, the point $\bar{x}$ can be strongly separated from M (Separation Theorem, (ii)). Thus, for appropriate b one has

$$b^T \bar{x} > \sup_{x \in M} b^T x.$$

Since $0 \in M$, the left hand side quantity in this inequality is positive; passing from b to a proportional vector $a = \lambda b$ with appropriately chosen positive λ, we may ensure that

$$a^T \bar{x} > 1 \ge \sup_{x \in M} a^T x.$$

This is the desired contradiction, since the relation $1 \ge \sup\limits_{x \in M} a^T x$ implies that $a \in \mathrm{Polar}\,(M)$, so that the relation $a^T \bar{x} > 1$ contradicts the assumption that $\bar{x} \in \mathrm{Polar}\,(\mathrm{Polar}\,(M))$. □

Exercise B.15 *Let M be a convex set containing the origin, and let M' be the polar of M. Prove the following facts:*

[2] In the latter reasoning we used the following fact: if $P \subset Q$ are two affine subspaces, then the affine dimension of P is $\le$ the one of Q, with $\le$ being $=$ iff $P = Q$. Please prove this fact

1. $\mathrm{Polar}\,(M) = \mathrm{Polar}\,(\mathrm{cl}\,M)$*;*
2. *M is bounded iff $0 \in \mathrm{int}\,M'$;*
3. *$\mathrm{int}\,M \neq \emptyset$ iff M' does not contain straight lines;*
4. *M is a closed cone of and only if M' is a closed cone. If M is a cone (not necessarily closed), then*
$$M' = \{a : a^T x \leq 0 \forall x \in M\}. \tag{B.2.15}$$

B.2.9.2 Dual cone

Let $M \subset \mathbf{R}^n$ be a cone. By Exercise B.15.4, the polar M' of M is a closed cone given by (B.2.15). The set $M_* = -M'$ (which also is a closed cone), that is, the set

$$M_* = \{a : a^T x \geq 0 \forall x \in M\}$$

of all vectors which have nonnegative inner products with all vectors from M, is called the cone *dual* to M. By Proposition B.2.5 and Exercise B.15.4, the family of closed cones in $\mathbf{R}^n$ is closed with respect to passing to a dual cone, and the duality is symmetric: for a closed cone M, M_* also is a closed cone, and $(M_*)_* = M$.

Exercise B.16 *Let M be a closed cone in $\mathbf{R}^n$, and M_* be its dual cone. Prove that*

1. *M is pointed (i.e., does not contain lines) if and only M_* has a nonempty interior. Derive from this fact that M is a closed pointed cone with a nonempty interior if and only if the dual cone has the same properties.*
2. *Prove that $a \in \mathrm{int}\,M_*$ iff $a^T x > 0$ for all nonzero vectors $x \in M$.*

B.2.9.3 Dubovitski-Milutin Lemma

Let $M_1, ..., M_k$ be cones (not necessarily closed), and M be their intersection; of course, M also is a cone. How to compute the cone dual to M?

Proposition B.2.6 *Let $M_1, ..., M_k$ be cones. The cone M' dual to the intersection M of the cones M_1,...,M_k contains the arithmetic sum $\widetilde{M}$ of the cones M'_1,...,M'_k dual to M_1,...,M_k. If all the cones $M_1, ..., M_k$ are closed, then M' is equal to $\mathrm{cl}\,\widetilde{M}$. In particular, for closed cones M_1,...,M_k, M' coincides with $\widetilde{M}$ if and only if the latter set is closed.*

Proof. Whenever $a_i \in M'_i$ and $x \in M$, we have $a_i^T x \geq 0$, $i = 1, ..., k$, whence $(a_1 + ... + a_k)^T x \geq 0$. Since the latter relation is valid for all $x \in M$, we conclude that $a_1 + ... + a_k \in M'$. Thus, $\widetilde{M} \subset M'$.

Now assume that the cones $M_1, ..., M_k$ are closed, and let us prove that $M = \mathrm{cl}\,\widetilde{M}$. Since M' is closed and we have seen that $\widetilde{M} \subset M'$, all we should prove is that if $a \in M'$, then $a \in \widehat{M} = \mathrm{cl}\,\widetilde{M}$ as well. Assume, on the contrary, that $a \in M' \backslash \widehat{M}$. Since the set $\widetilde{M}$ clearly is a cone, its closure $\widehat{M}$ is a closed cone; by assumption, a does not belong to this closed cone and therefore, by Separation Theorem (ii), a can be strongly separated from $\widehat{M}$ and therefore – from $\widetilde{M} \subset \widehat{M}$. Thus, for some x one has

$$a^T x < \inf_{b \in \widetilde{M}} b^T x = \inf_{a_i \in M'_i, i=1,...,k} (a_1 + ... + a_k)^T x = \sum_{i=1}^k \inf_{a_i \in M'_i} a_i^T x. \tag{B.2.16}$$

From the resulting inequality it follows that $\inf\limits_{a_i \in M'_i} a_i^T x > -\infty$; since M'_i is a cone, the latter is possible iff $\inf\limits_{a_i \in M'_i} a_i^T x = 0$, i.e., iff for every i one has $x \in \mathrm{Polar}\,(M'_i) = M_i$ (recall that the cones M_i are closed). Thus, $x \in M_i$ for all i, and the concluding quantity in (B.2.16) is 0. We see that $x \in M = \cap_i M_i$, and that (B.2.16) reduces to $a^T x < 0$. This contradicts the inclusion $a \in M'$. □

Note that in general $\widetilde{M}$ can be non-closed even when all the cones $M_1, ..., M_k$ are closed. Indeed, take $k = 2$, and let M'_1 be the ice-cream cone $\{(x, y, z) \in \mathbf{R}^3 : z \geq \sqrt{x^2 + y^2}\}$, and M'_2 be the ray $\{z = x \leq 0, y = 0\}$ in bR^3. Observe that the points from $\widetilde{M} \equiv M'_1 + M'_2$ are exactly the points of the form $(x - t, y, z - t)$ with $t \geq 0$ and $z \geq \sqrt{x^2 + y^2}$. In particular, for x positive the points $(0, 1, \sqrt{x^2 + 1} - x) = (x - x, 1, \sqrt{x^2 + 1} - x)$

belong to $\widetilde{M}$; as $x \to \infty$, these points converge to $p = (0, 1, 0)$, and thus $p \in \mathrm{cl}\,\widetilde{M}$. On the other hand, there clearly do not exist x, y, z, t with $t \geq 0$ and $z \geq \sqrt{x^2 + y^2}$ such that $(x - t, y, z - t) = (0, 1, 0)$, that is, $p \notin \widetilde{M}$.

Dubovitski-Milutin Lemma presents a simple sufficient condition for $\widetilde{M}$ to be closed and thus to coincide with M':

Proposition B.2.7 [Dubovitski-Milutin Lemma] *Let $M_1, ..., M_k$ be cones such that M_k is closed and the set $M_k \cap \mathrm{int}\, M_1 \cap \mathrm{int}\, M_2 \cap ... \cap \mathrm{int}\, M_{k-1}$ is nonempty, and let $M = M_1 \cap ... \cap M_k$. Let also M_i' be the cones dual to M_i. Then*

(i) $\mathrm{cl}\, M = \bigcap\limits_{i=1}^{k} \mathrm{cl}\, M_i$;

(ii) *the cone $\widetilde{M} = M_1' + ... + M_k'$ is closed, and thus coincides with the cone M' dual to $\mathrm{cl}\, M$ (or, which is the same by Exercise B.15.1, with the cone dual to M). In other words, every linear form which is nonnegative on M can be represented as a sum of k linear forms which are nonnegative on the respective cones M_1,...,M_k.*

Proof. (i): We should prove that under the premise of the Dubovitski-Milutin Lemma, $\mathrm{cl}\, M = \bigcap\limits_i \mathrm{cl}\, M_i$. The right hand side here contains M and is closed, so that all we should prove is that every point x in $\bigcap\limits_{i=1}^{k} \mathrm{cl}\, M_i$ is the limit of an appropriate sequence $x_t \in M$. By premise of the Lemma, there exists a point $\bar{x} \in M_k \cap \mathrm{int}\, M_1 \cap \mathrm{int}\, M_2 \cap ... \cap \mathrm{int}\, M_{k-1}$; setting $x_t = t^{-1}\bar{x} + (1 - t^{-1})x$, $t = 1, 2, ...$, we get a sequence converging to x as $t \to \infty$; at the same time, $x_t \in M_k$ (since x, $\bar{x}$ are in $\mathrm{cl}\, M_k = M_k$) and $x_t \in M_i$ for every $i < k$ (by Lemma B.1.1; note that for $i < k$ one has $\bar{x} \in \mathrm{int}\, M_i$ and $x \in \mathrm{cl}\, M_i$), and thus $x_t \in M$. Thus, every point $x \in \bigcap\limits_{i=1}^{k} \mathrm{cl}\, M_i$ is the limit of a sequence from M. □

(ii): Under the premise of the Lemma, when replacing the cones $M_1, ..., M_k$ with their closures, we do not vary the polars M_i' of the cones (and thus do not vary $\widetilde{M}$) and replace the intersection of the sets $M_1, ..., M_k$ with its closure (by (i)), thus not varying the polar of the intersection. And of course when replacing the cones $M_1, ..., M_k$ with their closures, we preserve the premise of Lemma. Thus, we lose nothing when assuming, in addition to the premise of Lemma, that the cones $M_1, ..., M_k$ are closed. To prove the lemma for closed cones M_1,...,M_k, we use induction in $k \geq 2$.

Base $k = 2$: Let a sequence $\{f_t + g_t\}_{t=1}^{\infty}$ with $f_t \in M_1'$ and $g_t \in M_2'$ converge to certain h; we should prove that $h = f + g$ for appropriate $f \in M_1'$ and $g \in M_2'$. To achieve our goal, it suffices to verify that for an appropriate subsequence t_j of indices there exists $f \equiv \lim\limits_{j\to\infty} f_{t_j}$. Indeed, if this is the case, then $g = \lim\limits_{j\to\infty} g_{t_j}$ also exists (since $f_t + g_t \to h$ as $t \to \infty$ and $f + g = h$; besides this, $f \in M_1'$ and $g \in M_2'$, since both the cones in question are closed. In order to verify the existence of the desired subsequence, it suffices to lead to a contradiction the assumption that $\|f_t\|_2 \to \infty$ as $t \to \infty$. Let the latter assumption be true. Passing to a subsequence, we may assume that the unit vectors $\phi_t = f_t/\|f_t\|_2$ have a limit ϕ as $t \to \infty$; since M_1' is a closed cone, ϕ is a unit vector from M_1'. Now, since $f_t + g_t \to h$ as $t \to \infty$, we have $\phi = \lim\limits_{t\to\infty} f_t/\|f_t\|_2 = -\lim\limits_{t\to\infty} g_t/\|f_t\|_2$ (recall that $\|f_t\|_2 \to \infty$ as $t \to \infty$, whence $h/\|f_t\|_2 \to 0$ as $t \to \infty$). We see that the vector $-\phi$ belongs to M_2'. Now, by assumption M_2 intersects the interior of the cone M_1; let $\bar{x}$ be a point in this intersection. We have $\phi^T\bar{x} \geq 0$ (since $\bar{x} \in M_1$ and $\phi \in M_1'$) and $\phi^T\bar{x} \leq 0$ (since $-\phi \in M_2'$ and $\bar{x} \in M_2$). We conclude that $\phi^T\bar{x} = 0$, which contradicts the facts that $0 \neq \phi \in M_1'$ and $\bar{x} \in \mathrm{int}\, M_1$ (see Exercise B.16.2). □

Inductive step: Assume that the statement we are proving is valid in the case of $k - 1 \geq 2$ cones, and let M_1,...,M_k be k cones satisfying the premise of the Dubovitski-Milutin Lemma. By this premise, the cone $M^1 = M_1 \cap ... \cap M_{k-1}$ has a nonempty interior, and M_k intersects this interior. Applying to the pair of cones M^1, M_k the already proved 2-cone version of the Lemma, we see that the set $(M^1)' + M_k'$ is closed; here $(M^1)'$ is the cone dual to M^1. Further, the cones $M_1, ..., M_{k-1}$ satisfy the premise of the $(k-1)$-cone version of the Lemma; by inductive hypothesis, the set $M_1' + ... + M_{k-1}'$ is closed and therefore, by Proposition B.2.6, equals to $(M^1)'$. Thus, $M_1' + ... + M_k' = (M^1)' + M_k'$, and we have seen that the latter set is closed. □

B.2.10 Extreme points and Krein-Milman Theorem

Supporting planes are useful tool to prove existence of *extreme points* of convex sets. Geometrically, an extreme point of a convex set M is a point in M which cannot be obtained as a convex combination of other points of the set; and the importance of the notion comes from the fact (which we shall prove in the mean time) that the set of all extreme points of a "good enough" convex set M is the "shortest worker's instruction for building the set" – this is the smallest set of points for which M is the convex hull.

B.2.10.1 Extreme points: definition

The exact definition of an extreme point is as follows:

Definition B.2.4 [extreme points] *Let M be a nonempty convex set in $\mathbf{R}^n$. A point $x \in M$ is called an extreme point of M, if there is no nontrivial (of positive length) segment $[u, v] \in M$ for which x is an interior point, i.e., if the relation*

$$x = \lambda u + (1 - \lambda)v$$

with $\lambda \in (0, 1)$ and $u, v \in M$ is valid iff

$$u = v = x.$$

E.g., the extreme points of a segment are exactly its endpoints; the extreme points of a triangle are its vertices; the extreme points of a (closed) circle on the 2-dimensional plane are the points of the circumference.

An equivalent definitions of an extreme point is as follows:

Exercise B.17 *Let M be a convex set and let $x \in M$. Prove that*

1. *x is extreme iff the only vector h such that $x \pm h \in M$ is the zero vector;*
2. *x is extreme iff the set $M \backslash \{x\}$ is convex.*

B.2.10.2 Krein-Milman Theorem

It is clear that a convex set M not necessarily possesses extreme points; as an example you may take the open unit ball in $\mathbf{R}^n$. This example is not interesting – the set in question is not closed; when replacing it with its closure, we get a set (the closed unit ball) with plenty of extreme points – these are all points of the boundary. There are, however, *closed* convex sets which do not possess extreme points – e.g., a line or an affine subspace of larger dimension. A nice fact is that the absence of extreme points in a closed convex set M always has the standard reason – the set contains a line. Thus, a closed and nonempty convex set M which does not contain lines for sure possesses extreme points. And if M is nonempty convex compact set, it possesses a quite representative set of extreme points – their convex hull is the entire M.

Theorem B.2.10 *Let M be a closed and nonempty convex set in $\mathbf{R}^n$. Then*
(i) *The set $\mathrm{Ext}(M)$ of extreme points of M is nonempty iff M does not contain lines;*
(ii) *If M is bounded, then M is the convex hull of its extreme points:*

$$M = \mathrm{Conv}(\mathrm{Ext}(M)),$$

so that every point of M is a convex combination of the points of $\mathrm{Ext}(M)$.

Part (ii) of this theorem is the finite-dimensional version of the famous *Krein-Milman Theorem*.

Proof. Let us start with (i). The "only if" part is easy, due to the following simple

> **Lemma B.2.3** *Let M be a closed convex set in $\mathbf{R}^n$. Assume that for some $\bar{x} \in M$ and $h \in \mathbf{R}^n$ M contains the ray*
>
> $$\{\bar{x} + th : t \geq 0\}$$
>
> *starting at $\bar{x}$ with the direction h. Then M contains also all parallel rays starting at the points of M:*
>
> $$(\forall x \in M) : \{x + th : t \geq 0\} \subset M.$$
>
> *In particular, if M contains certain line, then it contains also all parallel lines passing through the points of M.*

Comment. For a closed convex set M, the set of all directions h such that $x+th \in M$ for some x and all $t \geq 0$ (i.e., by Lemma – such that $x+th \in M$ for all $x \in M$ and all $t \geq 0$) is called *the recessive cone of M* [notation: $\mathrm{Rec}(M)$]. With Lemma B.2.3 it is immediately seen (prove it!) that $\mathrm{Rec}(M)$ indeed is a closed cone, and that

$$M + \mathrm{Rec}(M) = M.$$

Directions from $\mathrm{Rec}(M)$ are called recessive for M.

Proof of the lemma is immediate: if $x \in M$ and $\bar{x}+th \in M$ for all $t \geq 0$, then, due to convexity, for any fixed $\tau \geq 0$ we have

$$\epsilon(\bar{x} + \frac{\tau}{\epsilon}h) + (1-\epsilon)x \in M$$

for all $\epsilon \in (0,1)$. As $\epsilon \to +0$, the left hand side tends to $x+\tau h$, and since M is closed, $x+\tau h \in M$ for every $\tau \geq 0$. □

Exercise B.18 *Let M be a closed nonempty convex set. Prove that* $\mathrm{Rec}(M) \neq \{0\}$ *iff M is unbounded.*

Lemma B.2.3, of course, resolves all our problems with the "only if" part. Indeed, here we should prove that if M possesses extreme points, then M does not contain lines, or, which is the same, that if M contains lines, then it has no extreme points. But the latter statement is immediate: if M contains a line, then, by Lemma, there is a line in M passing through every given point of M, so that no point can be extreme. □

Now let us prove the "if" part of (i). Thus, from now on we assume that M does not contain lines; our goal is to prove that then M possesses extreme points. Let us start with the following

Lemma B.2.4 *Let Q be a nonempty closed convex set, $\bar{x}$ be a relative boundary point of Q and Π be a hyperplane supporting to Q at $\bar{x}$. Then all extreme points of the nonempty closed convex set $\Pi \cap Q$ are extreme points of Q.*

Proof of the Lemma. First, the set $\Pi \cap Q$ is closed and convex (as an intersection of two sets with these properties); it is nonempty, since it contains $\bar{x}$ (Π contains $\bar{x}$ due to the definition of a supporting plane, and Q contains $\bar{x}$ due to the closedness of Q). Second, let a be the linear form associated with Π:

$$\Pi = \{y : a^T y = a^T \bar{x}\},$$

so that

$$\inf_{x\in Q} a^T x < \sup_{x\in Q} a^T x = a^T \bar{x} \tag{B.2.17}$$

(see Proposition B.2.4). Assume that y is an extreme point of $\Pi \cap Q$; what we should do is to prove that y is an extreme point of Q, or, which is the same, to prove that

$$y = \lambda u + (1-\lambda)v$$

for some $u, v \in Q$ and $\lambda \in (0,1)$ is possible only if $y = u = v$. To this end it suffices to demonstrate that under the above assumptions $u, v \in \Pi \cap Q$ (or, which is the same, to prove that $u, v \in \Pi$, since the points are known to belong to Q); indeed, we know that y is an extreme point of $\Pi \cap Q$, so that the relation $y = \lambda u + (1-\lambda)v$ with $\lambda \in (0,1)$ *and* $u, v \in \Pi \cap Q$ does imply $y = u = v$.

To prove that $u, v \in \Pi$, note that since $y \in \Pi$ we have

$$a^T y = a^T \bar{x} \geq \max\{a^T u, a^T v\}$$

(the concluding inequality follows from (B.2.17)). On the other hand,

$$a^T y = \lambda a^T u + (1-\lambda) a^T v;$$

combining these observations and taking into account that $\lambda \in (0,1)$, we conclude that

$$a^T y = a^T u = a^T v.$$

But these equalities imply that $u, v \in \Pi$. □

Equipped with the Lemma, we can easily prove (i) by induction on the dimension of the convex set M (recall that this is nothing but the affine dimension of the affine span of M, i.e., the linear dimension of the linear subspace L such that $\mathrm{Aff}(M) = a + L$).

There is nothing to do if the dimension of M is zero, i.e., if M is a point – then, of course, $M = \mathrm{Ext}(M)$. Now assume that we already have proved the nonemptiness of $\mathrm{Ext}(T)$ for all nonempty closed and not containing lines convex sets T of certain dimension k, and let us prove that the same statement is valid for the sets of dimension $k+1$. Let M be a closed convex nonempty and not containing lines set of dimension $k+1$. Since M does not contain lines and is of positive dimension, it differs from $\mathrm{Aff}(M)$ and therefore it possesses a relative boundary point $\bar{x}$ [3]. According to Proposition B.2.4, there exists a hyperplane $\Pi = \{x : a^T x = a^T \bar{x}\}$ which supports M at $\bar{x}$:

$$\inf_{x\in M} a^T x < \max_{x\in M} a^T x = a^T \bar{x}.$$

By the same Proposition, the set $T = \Pi \cap M$ (which is closed, convex and nonempty) is of affine dimension less than the one of M, i.e., of the dimension $\leq k$. T clearly does not contain lines (since even the larger set M does not contain lines). By the inductive hypothesis, T possesses extreme points, and by Lemma B.2.4 all these points are extreme also for M. The inductive step is completed, and (i) is proved. □

Now let us prove (ii). Thus, let M be nonempty, convex, closed and bounded; we should prove that

$$M = \mathrm{Conv}(\mathrm{Ext}(M)).$$

What is immediately seen is that the right hand side set is contained in the left hand side one. Thus, all we need is to prove that every $x \in M$ is a convex combination of points from $\mathrm{Ext}(M)$. Here we again use induction on the affine dimension of M. The case of 0-dimensional set M (i.e., a point) is trivial. Assume that the statement in question is valid for all k-dimensional convex closed and bounded sets, and let M be a convex closed and bounded set of dimension $k+1$. Let $x \in M$; to represent x as a convex combination of points from $\mathrm{Ext}(M)$, let us pass through x an arbitrary line $\ell = \{x + \lambda h : \lambda \in \mathbf{R}\}$ $(h \neq 0)$ in the affine span $\mathrm{Aff}(M)$ of M. Moving along this line from x in each of the two possible directions, we eventually leave M (since M is bounded); as it was explained in the proof of (i), it means that there exist nonnegative λ_+ and λ_- such that the points

$$\bar{x}_\pm = x + \lambda_\pm h$$

both belong to the relative boundary of M. Let us verify that $\bar{x}_\pm$ are convex combinations of the extreme points of M (this will complete the proof, since x clearly is a convex combination of the two points $\bar{x}_\pm$). Indeed, M admits supporting at $\bar{x}_+$ hyperplane Π; as it was explained in the proof of (i), the set $\Pi \cap M$ (which clearly is convex, closed and bounded) is of affine dimension less than that one of M; by the inductive hypothesis, the point $\bar{x}_+$ of this set is a convex combination of extreme points of the set, and by Lemma B.2.4 all these extreme points are extreme points of M as well. Thus, $\bar{x}_+$ is a convex combination of extreme points of M. Similar reasoning is valid for $\bar{x}_-$. □

B.2.10.3 Example: Extreme points of a polyhedral set.

Consider a polyhedral set

$$K = \{x \in \mathbf{R}^n : Ax \leq b\},$$

A being a $m \times n$ matrix and b being a vector from $\mathbf{R}^m$. What are the extreme points of K? The answer is given by the following

Theorem B.2.11 [Extreme points of polyhedral set]
Let $x \in K$. The vector x is an extreme point of K iff some n linearly independent (i.e., with linearly independent vectors of coefficients) inequalities of the system $Ax \leq b$ are equalities at x.

[3] Indeed, there exists $z \in \mathrm{Aff}(M) \backslash M$, so that the points

$$x_\lambda = x + \lambda(z - x)$$

(x is an arbitrary fixed point of M) do not belong to M for some $\lambda \geq 1$, while $x_0 = x$ belongs to M. The set of those $\lambda \geq 0$ for which $x_\lambda \in M$ is therefore nonempty and bounded from above; this set clearly is closed (since M is closed). Thus, there exists the largest $\lambda = \lambda^*$ for which $x_\lambda \in M$. We claim that x_{λ^*} is a relative boundary point of M. Indeed, by construction this is a point from M. If it would be a point from the relative interior of M, then all the points x_λ with close to λ^* and greater than λ^* values of λ would also belong to M, which contradicts the origin of λ^*

Proof. Let a_i^T, $i = 1, ..., m$, be the rows of A.

The "only if" part: let x be an extreme point of K, and let I be the set of those indices i for which $a_i^T x = b_i$; we should prove that the set F of vectors $\{a_i : i \in I\}$ contains n linearly independent vectors, or, which is the same, that $\mathrm{Lin}(F) = \mathbf{R}^n$. Assume that it is not the case; then the orthogonal complement to F contains a nonzero vector h (since the dimension of $F^\perp$ is equal to $n - \dim \mathrm{Lin}(F)$ and is therefore positive). Consider the segment $\Delta_\epsilon = [x - \epsilon h, x + \epsilon h]$, $\epsilon > 0$ being the parameter of our construction. Since h is orthogonal to the "active" vectors a_i – those with $i \in I$, all points y of this segment satisfy the relations $a_i^T y = a_i^T x = b_i$. Now, if i is a "nonactive" index – one with $a_i^T x < b_i$ – then $a_i^T y \leq b_i$ for all $y \in \Delta_\epsilon$, provided that ϵ is small enough. Since there are finitely many nonactive indices, we can choose $\epsilon > 0$ in such a way that all $y \in \Delta_\epsilon$ will satisfy all "nonactive" inequalities $a_i^T x \leq b_i$, $i \notin I$. Since $y \in \Delta_\epsilon$ satisfies, as we have seen, also all "active" inequalities, we conclude that with the above choice of ϵ we get $\Delta_\epsilon \subset K$, which is a contradiction: $\epsilon > 0$ and $h \neq 0$, so that Δ_ϵ is a nontrivial segment with the midpoint x, and no such segment can be contained in K, since x is an extreme point of K. □

To prove the "if" part, assume that $x \in K$ is such that among the inequalities $a_i^T x \leq b_i$ which are equalities at x there are n linearly independent, say, those with indices $1, ..., n$, and let us prove that x is an extreme point of K. This is immediate: assuming that x is not an extreme point, we would get the existence of a nonzero vector h such that $x \pm h \in K$. In other words, for $i = 1, ..., n$ we would have $b_i \pm a_i^T h \equiv a_i^T (x \pm h) \leq b_i$, which is possible only if $a_i^T h = 0$, $i = 1, ..., n$. But the only vector which is orthogonal to n linearly independent vectors in $\mathbf{R}^n$ is the zero vector (why?), and we get $h = 0$, which was assumed not to be the case. □

.

Corollary B.2.1 *The set of extreme points of a polyhedral set is finite.*

Indeed, according to the above Theorem, every extreme point of a polyhedral set $K = \{x \in \mathbf{R}^n : Ax \leq b\}$ satisfies the equality version of certain n-inequality subsystem of the original system, the matrix of the subsystem being nonsingular. Due to the latter fact, an extreme point is uniquely defined by the corresponding subsystem, so that the number of extreme points does not exceed the number C_m^n of $n \times n$ submatrices of the matrix A and is therefore finite. □

Note that C_m^n is nothing but an upper (ant typically very conservative) bound on the number of extreme points of a polyhedral set given by m inequalities in $\mathbf{R}^n$: some $n \times n$ submatrices of A can be singular and, what is more important, the majority of the nonsingular ones normally produce "candidates" which do not satisfy the remaining inequalities.

> **Remark B.2.1** The result of Theorem B.2.11 is very important, in particular, for the theory of the Simplex method – the traditional computational tool of Linear Programming. When applied to the LP program in the standard form
>
> $$\min_x \left\{c^T x : Px = p,\, x \geq 0\right\} \quad [x \in \mathbf{R}^n],$$
>
> with $k \times n$ matrix P, the result of Theorem B.2.11 is that extreme points of the feasible set are exactly *the basic feasible solutions* of the system $Px = p$, i.e., nonnegative vectors x such that $Px = p$ and the set of columns of P associated with positive entries of x is linearly independent. Since the feasible set of an LP program in the standard form clearly does not contain lines, among the optimal solutions (if they exist) to an LP program in the standard form at least one is an extreme point of the feasible set (Theorem B.2.14.(ii)). Thus, in principle we could look through the finite set of all extreme points of the feasible set ($\equiv$ through all basic feasible solutions) and to choose the one with the best value of the objective. This recipe allows to find a feasible solution in finitely many arithmetic operations, provided that the program is solvable, and is, basically, what the Simplex method does; this latter method, of course, looks through the basic feasible solutions in a smart way which normally allows to deal with a negligible part of them only.
>
> Another useful consequence of Theorem B.2.11 is that if all the data in an LP program are rational, then every extreme point of the feasible domain of the program is a vector with rational entries. In particular, a solvable standard form LP program with rational data has at least one rational optimal solution.

B.2.10.4 Illustration: Birkhoff Theorem

An $n\times n$ matrix X is called *double stochastic*, if its entries are nonnegative and all column and row sums are equal to 1. These matrices (treated as elements of $\mathbf{R}^{n^2}=\mathbf{R}^{n\times n}$) form a bounded polyhedral set, specifically, the set

$$\Pi_n=\{X=[x_{ij}]_{i,j=1}^n : x_{ij}\geq 0\,\forall i,j, \sum_i x_{ij}=1\,\forall j, \sum_j x_{ij}=1\,\forall i\}$$

By Krein-Milman Theorem, Π_n is the convex hull of its extreme points. What are these extreme points? The answer is given by important

Theorem B.2.12 [Birkhoff Theorem] *Extreme points of Π_n are exactly the permutation matrices of order n, i.e., $n\times n$ Boolean (i.e., with 0/1 entries) matrices with exactly one nonzero element (equal to 1) in every row and every column.*

Exercise B.19 [Easy part] *Prove the easy part of the Theorem, specifically, that every $n\times n$ permutation matrix is an extreme point of Π_n.*

Proof of difficult part. Now let us prove that every extreme point of Π_n is a permutation matrix. To this end let us note that the $2n$ linear equations in the definition of Π_n — those saying that all row and column sums are equal to 1 - are linearly dependent, and dropping one of them, say, $\sum_i x_{in}=1$, we do not alter the set. Indeed, the remaining equalities say that all row sums are equal to 1, so that the total sum of all entries in X is n, and that the first $n-1$ column sums are equal to 1, meaning that the last column sum is $n-(n-1)=1$. Thus, we lose nothing when assuming that there are just $2n-1$ equality constraints in the description of Π_n. Now let us prove the claim by induction in n. The base $n=1$ is trivial. Let us justify the inductive step $n-1\Rightarrow n$. Thus, let X be an extreme point of Π_n. By Theorem B.2.11, among the constraints defining Π_n (i.e., $2n-1$ equalities and n^2 inequalities $x_{ij}\geq 0$) there should be n^2 linearly independent which are satisfied at X as equations. Thus, at least $n^2-(2n-1)=(n-1)^2$ entries in X should be zeros. It follows that at least one of the columns of X contains ≤ 1 nonzero entries (since otherwise the number of zero entries in X would be at most $n(n-2)<(n-1)^2$). Thus, there exists at least one column with at most 1 nonzero entry; since the sum of entries in this column is 1, this nonzero entry, let it be $x_{\bar{i}\bar{j}}$, is equal to 1. Since the entries in row $\bar{i}$ are nonnegative, sum up to 1 and $x_{\bar{i}\bar{j}}=1$, $x_{\bar{i}\bar{j}}=1$ is the only nonzero entry in its row and its column. Eliminating from X the row $\bar{i}$ and the column $\bar{j}$, we get an $(n-1)\times(n-1)$ double stochastic matrix. By inductive hypothesis, this matrix is a convex combination of $(n-1)\times(n-1)$ permutation matrices. Augmenting every one of these matrices by the column and the row we have eliminated, we get a representation of X as a convex combination of $n\times n$ permutation matrices: $X=\sum_\ell \lambda_\ell P_\ell$ with nonnegative λ_ℓ summing up to 1. Since $P_\ell\in\Pi_n$ and X is an extreme point of Π_n, in this representation all terms with nonzero coefficients λ_ℓ must be equal to $\lambda_\ell X$, so that X is one of the permutation matrices P_ℓ and as such is a permutation matrix. □

B.2.11 Structure of polyhedral sets

B.2.11.1 Main result

By definition, a polyhedral set M is the set of all solutions to a finite system of nonstrict linear inequalities:

$$M=\{x\in\mathbf{R}^n : Ax\leq b\},\tag{B.2.18}$$

where A is a matrix of the column size n and certain row size m and b is m-dimensional vector. This is an "outer" description of a polyhedral set. We are about to establish an important result on the equivalent "inner" representation of a polyhedral set.

Consider the following construction. Let us take two finite nonempty set of vectors V ("vertices") and R ("rays") and build the set

$$M(V,R)=\mathrm{Conv}(V)+\mathrm{Cone}\,(R)=\{\sum_{v\in V}\lambda_v v+\sum_{r\in R}\mu_r r : \lambda_v\geq 0, \mu_r\geq 0, \sum_v \lambda_v=1\}.$$

Thus, we take all vectors which can be represented as sums of convex combinations of the points from V and conic combinations of the points from R. The set $M(V,R)$ clearly is convex (as the arithmetic sum of two convex sets $\mathrm{Conv}(V)$ and $\mathrm{Cone}\,(R)$). The promised inner description polyhedral sets is as follows:

Theorem B.2.13 [Inner description of a polyhedral set] *The sets of the form $M(V,R)$ are exactly the nonempty polyhedral sets: $M(V,R)$ is polyhedral, and every nonempty polyhedral set M is $M(V,R)$ for properly chosen V and R.*

The polytopes $M(V,\{0\}) = \mathrm{Conv}(V)$ are exactly the nonempty and bounded polyhedral sets. The sets of the type $M(\{0\},R)$ are exactly the polyhedral cones (sets given by finitely many nonstrict homogeneous linear inequalities).

Remark B.2.2 In addition to the results of the Theorem, it can be proved that in the representation of a nonempty polyhedral set M as $M = \mathrm{Conv}(V) + \mathrm{Cone}\,(R)$

– the "conic" part $\mathrm{Conv}(R)$ (not the set R itself!) is uniquely defined by M and is the recessive cone of M (see Comment to Lemma B.2.3);

– if M does not contain lines, then V can be chosen as the set of all extreme points of M.

Postponing temporarily the proof of Theorem B.2.13, let us explain why this theorem is that important – why it is so nice to know both inner and outer descriptions of a polyhedral set.

Consider a number of natural questions:

- A. Is it true that the inverse image of a polyhedral set $M \subset \mathbf{R}^n$ under an affine mapping $y \mapsto \mathcal{P}(y) = Py + p : \mathbf{R}^m \to \mathbf{R}^n$, i.e., the set
$$\mathcal{P}^{-1}(M) = \{y \in \mathbf{R}^m : Py + p \in M\}$$
is polyhedral?
- B. Is it true that the image of a polyhedral set $M \subset \mathbf{R}^n$ under an affine mapping $x \mapsto y = \mathcal{P}(x) = Px + p : \mathbf{R}^n \to \mathbf{R}^m$ – the set
$$\mathcal{P}(M) = \{Px + p : x \in M\}$$
is polyhedral?
- C. Is it true that the intersection of two polyhedral sets is again a polyhedral set?
- D. Is it true that the arithmetic sum of two polyhedral sets is again a polyhedral set?

The answers to all these questions are positive; one way to see it is to use calculus of polyhedral representations along with the fact that polyhedrally representable sets are exactly the same as polyhedral sets (Section B.2.4). Another very instructive way is to use the just outlined results on the structure of polyhedral sets, which we intend to do now.

It is very easy to answer affirmatively to A, starting from the original – outer – definition of a polyhedral set: if $M = \{x : Ax \le b\}$, then, of course,
$$\mathcal{P}^{-1}(M) = \{y : A(Py + p) \le b\} = \{y : (AP)y \le b - Ap\}$$
and therefore $\mathcal{P}^{-1}(M)$ is a polyhedral set.

An attempt to answer affirmatively to B via the same definition fails – there is no easy way to convert the linear inequalities defining a polyhedral set into those defining its image, and it is absolutely unclear why the image indeed is given by finitely many linear inequalities. Note, however, that there is no difficulty to answer affirmatively to B with the inner description of a nonempty polyhedral set: if $M = M(V,R)$, then, evidently,
$$\mathcal{P}(M) = M(\mathcal{P}(V), PR),$$
where $PR = \{Pr : r \in R\}$ is the image of R under the action of the homogeneous part of $\mathcal{P}$.

Similarly, positive answer to C becomes evident, when we use the outer description of a polyhedral set: taking intersection of the solution sets to two systems of nonstrict linear inequalities, we, of course, again get the solution set to a system of this type – you simply should put together all inequalities from the original two systems. And it is very unclear how to answer positively to D with the outer definition of a polyhedral set – what happens with inequalities when we add the solution sets? In contrast to this, the inner description gives the answer immediately:
$$\begin{array}{rcl} M(V,R) + M(V',R') & = & \mathrm{Conv}(V) + \mathrm{Cone}\,(R) + \mathrm{Conv}(V') + \mathrm{Cone}\,(R') \\ & = & [\mathrm{Conv}(V) + \mathrm{Conv}(V')] + [\mathrm{Cone}\,(R) + \mathrm{Cone}\,(R')] \\ & = & \mathrm{Conv}(V + V') + \mathrm{Cone}\,(R \cup R') \\ & = & M(V + V', R \cup R'). \end{array}$$

Note that in this computation we used two rules which should be justified: $\mathrm{Conv}(V)+\mathrm{Conv}(V') = \mathrm{Conv}(V+V')$ and $\mathrm{Cone}\,(R) + \mathrm{Cone}\,(R') = \mathrm{Cone}\,(R \cup R')$. The second is evident from the definition of the conic hull, and only the first needs simple reasoning. To prove it, note that $\mathrm{Conv}(V) + \mathrm{Conv}(V')$ is a convex set which contains $V + V'$ and therefore contains $\mathrm{Conv}(V + V')$. The inverse inclusion is proved as follows: if

$$x = \sum_i \lambda_i v_i,\ y = \sum_j \lambda'_j v'_j$$

are convex combinations of points from V, resp., V', then, as it is immediately seen (please check!),

$$x + y = \sum_{i,j} \lambda_i \lambda'_j (v_i + v'_j)$$

and the right hand side is a convex combination of points from $V + V'$.

We see that it is extremely useful to keep in mind both descriptions of polyhedral sets – what is difficult to see with one of them, is absolutely clear with another.

As a seemingly “more important” application of the developed theory, let us look at Linear Programming.

B.2.11.2 Theory of Linear Programming

A general Linear Programming program is the problem of maximizing a linear objective function over a polyhedral set:

$$\text{(P)} \quad \max_x \left\{ c^T x : x \in M = \{x \in \mathbf{R}^n : Ax \leq b\} \right\};$$

here c is a given n-dimensional vector – the objective, A is a given $m \times n$ *constraint matrix* and $b \in \mathbf{R}^m$ is the right hand side vector. Note that (P) is called “Linear Programming program in the canonical form”; there are other equivalent forms of the problem.

B.2.11.3 Solvability of a Linear Programming program

According to the Linear Programming terminology which you for sure know, (P) is called

- feasible, if it admits a feasible solution, i.e., the system $Ax \leq b$ is solvable, and infeasible otherwise;
- bounded, if it is feasible and the objective is above bounded on the feasible set, and unbounded, if it is feasible, but the objective is not bounded from above on the feasible set;
- solvable, if it is feasible and the optimal solution exists – the objective attains its maximum on the feasible set.

If the program is bounded, then the upper bound of the values of the objective on the feasible set is a real; this real is called the optimal value of the program and is denoted by c^*. It is convenient to assign optimal value to unbounded and infeasible programs as well – for an unbounded program it, by definition, is $+\infty$, and for an infeasible one it is $-\infty$.

Note that our terminology is aimed to deal with maximization programs; if the program is to minimize the objective, the terminology is updated in the natural way: when defining bounded/unbounded programs, we should speak about below boundedness rather than about the above boundedness of the objective, etc. E.g., the optimal value of an unbounded minimization program is $-\infty$, and of an infeasible one it is $+\infty$. This terminology is consistent with the usual way of converting a minimization problem into an equivalent maximization one by replacing the original objective c with $-c$: the properties of feasibility – boundedness – solvability remain unchanged, and the optimal value in all cases changes its sign.

We have said that you for sure know the above terminology; this is not exactly true, since you definitely have heard and used the words “infeasible LP program”, “unbounded LP program”, but hardly used the words “bounded LP program” – only the “solvable” one. This indeed is true, although absolutely unclear in advance – a bounded LP program always is solvable. We have already established this fact, even twice — via Fourier-Motzkin elimination (Section B.2.4 and via the LP Duality Theorem). Let us reestablish this fundamental for Linear Programming fact with the tools we have at our disposal now.

Theorem B.2.14 (i) *A Linear Programming program is solvable iff it is bounded.*

(ii) *If the program is solvable and the feasible set of the program does not contain lines, then at least one of the optimal solutions is an extreme point of the feasible set.*

Proof. (i): The "only if" part of the statement is tautological: the definition of solvability includes boundedness. What we should prove is the "if" part – that a bounded program is solvable. This is immediately given by the inner description of the feasible set M of the program: this is a polyhedral set, so that being nonempty (as it is for a bounded program), it can be represented as

$$M(V,R) = \mathrm{Conv}(V) + \mathrm{Cone}\,(R)$$

for some nonempty finite sets V and R. We claim first of all that since (P) is bounded, the inner product of c with every vector from R is nonpositive. Indeed, otherwise there would be $r \in R$ with $c^T r > 0$; since $M(V,R)$ clearly contains with every its point x the entire ray $\{x + tr : t \geq 0\}$, and the objective evidently is unbounded on this ray, it would be above unbounded on M, which is not the case.

Now let us choose in the *finite and nonempty* set V the point, let it be called v^*, which maximizes the objective on V. We claim that v^* is an optimal solution to (P), so that (P) is solvable. The justification of the claim is immediate: v^* clearly belongs to M; now, a generic point of $M = M(V,R)$ is

$$x = \sum_{v\in V} \lambda_v v + \sum_{r\in R} \mu_r r$$

with nonnegative λ_v and μ_r and with $\sum\limits_v \lambda_v = 1$, so that

$$\begin{array}{rcll} c^T x & = & \sum\limits_v \lambda_v c^T v + \sum\limits_r \mu_r c^T r & \\ & \leq & \sum\limits_v \lambda_v c^T v & [\text{since } \mu_r \geq 0 \text{ and } c^T r \leq 0,\ r \in R] \\ & \leq & \sum\limits_v \lambda_v c^T v^* & [\text{since } \lambda_v \geq 0 \text{ and } c^T v \leq c^T v^*] \\ & = & c^T v^* & [\text{since } \sum\limits_v \lambda_v = 1]\ \square \end{array}$$

(ii): if the feasible set of (P), let it be called M, does not contain lines, it, being convex and closed (as a polyhedral set) possesses extreme points. It follows that (ii) is valid in the trivial case when the objective of (ii) is constant on the entire feasible set, since then every extreme point of M can be taken as the desired optimal solution. The case when the objective is nonconstant on M can be immediately reduced to the aforementioned trivial case: if x^* is an optimal solution to (P) and the linear form $c^T x$ is nonconstant on M, then the hyperplane $\Pi = \{x : c^T x = c^*\}$ is supporting to M at x^*; the set $\Pi \cap M$ is closed, convex, nonempty and does not contain lines, therefore it possesses an extreme point x^{**} which, on one hand, clearly is an optimal solution to (P), and on another hand is an extreme point of M by Lemma B.2.4. □

B.2.11.4 Structure of a polyhedral set: proofs

B.2.11.5 Structure of a bounded polyhedral set

Let us start with proving a significant part of Theorem B.2.13 – the one describing *bounded* polyhedral sets.

Theorem B.2.15 [Structure of a bounded polyhedral set] *A bounded and nonempty polyhedral set M in $\mathbf{R}^n$ is a polytope, i.e., is the convex hull of a finite nonempty set:*

$$M = M(V,\{0\}) = \mathrm{Conv}(V);$$

one can choose as V the set of all extreme points of M.

Vice versa – a polytope is a bounded and nonempty polyhedral set.

Proof. The first part of the statement – that a bounded nonempty polyhedral set is a polytope – is readily given by the Krein-Milman Theorem combined with Corollary B.2.1. Indeed, a polyhedral set always is closed (as a set given by nonstrict inequalities involving continuous functions) and convex; if it is also bounded and nonempty, it, by the Krein-Milman Theorem, is the convex hull of the set V of its extreme points; V is finite by Corollary B.2.1. □

Now let us prove the more difficult part of the statement – that a polytope is a bounded polyhedral set. The fact that a convex hull of a finite set is bounded is evident. Thus, all we need is to prove that the convex hull of finitely many points is a polyhedral set. To this end note that this convex hull clearly is polyhedrally representable:

$$\mathrm{Conv}\{v_1,...,v_N\} = \{x : \exists\lambda : \lambda \geq 0, \sum_i \lambda_i = 1, x = \sum_i \lambda_i v_i\}$$

and therefore is polyhedral by Theorem B.2.5. □

B.2.11.6 Structure of a general polyhedral set: completing the proof

Now let us prove the general Theorem B.2.13. The proof basically follows the lines of the one of Theorem B.2.15, but with one elaboration: now we cannot use the Krein-Milman Theorem to take upon itself part of our difficulties.

To simplify language let us call VR-sets ("V" from "vertex", "R" from rays) the sets of the form $M(V,R)$, and P-sets the nonempty polyhedral sets. We should prove that every P-set is a VR-set, and vice versa. We start with proving that every P-set is a VR-set.

P⇒VR:

P⇒VR, Step 1: reduction to the case when the P-set does not contain lines. Let M be a P-set, so that M is the set of all solutions to a solvable system of linear inequalities:

$$M = \{x \in \mathbf{R}^n : Ax \leq b\} \tag{B.2.19}$$

with $m \times n$ matrix A. Such a set may contain lines; if h is the direction of a line in M, then $A(x+th) \leq b$ for some x and all $t \in \mathbf{R}$, which is possible only if $Ah = 0$. Vice versa, if h is from the kernel of A, i.e., if $Ah = 0$, then the line $x + \mathbf{R}^h$ with $x \in M$ clearly is contained in M. Thus, we come to the following fact:

Lemma B.2.5 *Nonempty polyhedral set (B.2.19) contains lines iff the kernel of A is nontrivial, and the nonzero vectors from the kernel are exactly the directions of lines contained in M: if M contains a line with direction h, then $h \in \mathrm{Ker}\,A$, and vice versa: if $0 \neq h \in \mathrm{Ker}\,A$ and $x \in M$, then M contains the entire line $x + \mathbf{R}h$.*

Given a nonempty set (B.2.19), let us denote by L the kernel of A and by $L^\perp$ the orthogonal complement to the kernel, and let M' be the cross-section of M by $L^\perp$:

$$M' = \{x \in L^\perp : Ax \leq b\}.$$

The set M' clearly does not contain lines (since the direction of every line in M', on one hand, should belong to $L^\perp$ due to $M' \subset L^\perp$, and on the other hand – should belong to $L = \mathrm{Ker}\,A$, since a line in $M' \subset M$ is a line in M as well). The set M' is nonempty and, moreover, $M = M' + L$. Indeed, M' contains the orthogonal projections of all points from M onto $L^\perp$ (since to project a point onto $L^\perp$, you should move from this point along certain line with the direction in L, and all these movements, started in M, keep you in M by the Lemma) and therefore is nonempty, first, and is such that $M' + L \supset M$, second. On the other hand, $M' \subset M$ and $M + L = M$ by Lemma B.2.5, whence $M' + L \subset M$. Thus, $M' + L = M$.

Finally, M' is a polyhedral set together with M, since the inclusion $x \in L^\perp$ can be represented by $\dim L$ linear equations (i.e., by $2\dim L$ nonstrict linear inequalities): you should say that x is orthogonal to $\dim L$ somehow chosen vectors $a_1,...,a_{\dim L}$ forming a basis in L.

The results of our effort are as follows: given an arbitrary P-set M, we have represented is as the sum of a P-set M' not containing lines and a linear subspace L. With this decomposition in mind we see that in order to achieve our current goal – to prove that every P-set is a VR-set – it suffices to prove the same statement for P-sets not containing lines. Indeed, given that $M' = M(V,R')$ and denoting by R' a finite set such that $L = \mathrm{Cone}\,(R')$ (to get R', take the set of $2\dim L$ vectors $\pm a_i$, $i = 1,...,\dim L$, where $a_1,...,a_{\dim L}$ is a basis in L), we would obtain

$$\begin{array}{rcl} M &=& M' + L \\ &=& [\mathrm{Conv}(V) + \mathrm{Cone}\,(R)] + \mathrm{Cone}\,(R') \\ &=& \mathrm{Conv}(V) + [\mathrm{Cone}\,(R) + \mathrm{Cone}\,(R')] \\ &=& \mathrm{Conv}(V) + \mathrm{Cone}\,(R \cup R') \\ &=& M(V, R \cup R') \end{array}$$

We see that in order to establish that a P-set is a VR-set it suffices to prove the same statement for the case when the P-set in question does not contain lines.

P⇒VR, Step 2: the P-set does not contain lines. Our situation is as follows: we are given a not containing lines P-set in $\mathbf{R}^n$ and should prove that it is a VR-set. We shall prove this statement by induction on the dimension n of the space. The case of $n = 0$ is trivial. Now assume that the statement in question is valid for $n \leq k$, and let us prove that it is valid also for $n = k + 1$. Let M be a not containing lines P-set in $\mathbf{R}^{k+1}$:

$$M = \{x \in \mathbf{R}^{k+1} : a_i^T x \leq b_i, \, i = 1, ..., m\}. \tag{B.2.20}$$

Without loss of generality we may assume that all a_i are nonzero vectors (since M is nonempty, the inequalities with $a_i = 0$ are valid on the entire $\mathbf{R}^n$, and removing them from the system, we do not vary its solution set). Note that $m > 0$ – otherwise M would contain lines, since $k \geq 0$.

1^0. We may assume that M is unbounded – otherwise the desired result is given already by Theorem B.2.15. By Exercise B.18, there exists a recessive direction $r \neq 0$ of M Thus, M contains the ray $\{x + tr : t \geq 0\}$, whence, by Lemma B.2.3, $M + \text{Cone}\,(\{r\}) = M$. □

2^0. For every $i \leq m$, where m is the row size of the matrix A from (B.2.20), that is, the number of linear inequalities in the description of M, let us denote by M_i the corresponding "facet" of M – the polyhedral set given by the system of inequalities (B.2.20) with the inequality $a_i^T x \leq b_i$ replaced by the equality $a_i^T x = b_i$. Some of these "facets" can be empty; let I be the set of indices i of nonempty M_i's.

When $i \in I$, the set M_i is a nonempty polyhedral set – i.e., a P-set – which does not contain lines (since $M_i \subset M$ and M does not contain lines). Besides this, M_i belongs to the hyperplane $\{a_i^T x = b_i\}$, i.e., actually it is a P-set in $\mathbf{R}^k$. By the inductive hypothesis, we have representations

$$M_i = M(V_i, R_i), \; i \in I,$$

for properly chosen finite nonempty sets V_i and R_i. We claim that

$$M = M(\cup_{i \in I} V_i, \cup_{i \in I} R_i \cup \{r\}), \tag{B.2.21}$$

where r is a recessive direction of M found in 1^0; after the claim will be supported, our induction will be completed.

To prove (B.2.21), note, first of all, that the right hand side of this relation is contained in the left hand side one. Indeed, since $M_i \subset M$ and $V_i \subset M_i$, we have $V_i \subset M$, whence also $V = \cup_i V_i \subset M$; since M is convex, we have

$$\text{Conv}(V) \subset M. \tag{B.2.22}$$

Further, if $r' \in R_i$, then r' is a recessive direction of M_i; since $M_i \subset M$, r' is a recessive direction of M by Lemma B.2.3. Thus, every vector from $\cup_{i \in I} R_i$ is a recessive direction for M, same as r; thus, every vector from $R = \cup_{i \in I} R_i \cup \{r\}$ is a recessive direction of M, whence, again by Lemma B.2.3,

$$M + \text{Cone}\,(R) = M.$$

Combining this relation with (B.2.22), we get $M(V, R) \subset M$, as claimed.

It remains to prove that M is contained in the right hand side of (B.2.21). Let $x \in M$, and let us move from x along the direction $(-r)$, i.e., move along the ray $\{x - tr : t \geq 0\}$. After large enough step along this ray we leave M. (Indeed, otherwise the ray with the direction $-r$ started at x would be contained in M, while the opposite ray for sure is contained in M since r is a recessive direction of M; we would conclude that M contains a line, which is not the case by assumption.) Since the ray $\{x - tr : t \geq 0\}$ eventually leaves M and M is bounded, there exists the largest t, let it be called t^*, such that $x' = x - t^* r$ still belongs to M. It is clear that at x' one of the linear inequalities defining M becomes equality – otherwise we could slightly increase the parameter t^* still staying in M. Thus, $x' \in M_i$ for some $i \in I$. Consequently,

$$x' \in \text{Conv}(V_i) + \text{Cone}\,(R_i),$$

whence $x = x' + t^* r \in \text{Conv}(V_i) + \text{Cone}\,(R_i \cup \{r\}) \subset M(V, R)$, as claimed. □

VR⇒P: We already know that every P-set is a VR-set. Now we shall prove that every VR-set is a P-set, thus completing the proof of Theorem B.2.13. This is immediate: a VR-set is polyhedrally representable (why?) and thus is a P-set by Theorem B.2.5. □

Appendix C

Convex functions

C.1 Convex functions: first acquaintance

C.1.1 Definition and Examples

Definition C.1.1 [convex function] *A function $f : Q \to \mathbf{R}$ defined on a nonempty subset Q of $\mathbf{R}^n$ and taking real values is called convex, if*

- *the domain Q of the function is convex;*
- *for every $x, y \in Q$ and every $\lambda \in [0,1]$ one has*

$$f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda)f(y). \tag{C.1.1}$$

If the above inequality is strict whenever $x \neq y$ and $0 < \lambda < 1$, f is called strictly convex.

A function f such that $-f$ is convex is called *concave*; the domain Q of a concave function should be convex, and the function itself should satisfy the inequality opposite to (C.1.1):

$$f(\lambda x + (1-\lambda)y) \ge \lambda f(x) + (1-\lambda)f(y), \;\; x, y \in Q, \lambda \in [0,1].$$

The simplest example of a convex function is an *affine function*

$$f(x) = a^T x + b$$

– the sum of a linear form and a constant. This function clearly is convex on the entire space, and the "convexity inequality" for it is equality. An affine function is both convex and concave; it is easily seen that a function which is both convex and concave on the entire space is affine.

Here are several elementary examples of "nonlinear" convex functions of one variable:

- functions convex on the whole axis:
  x^{2p}, p is a positive integer;
  $\exp\{x\}$;
- functions convex on the nonnegative ray:
  x^p, $1 \le p$;
  $-x^p$, $0 \le p \le 1$;
  $x \ln x$;
- functions convex on the positive ray:
  $1/x^p$, $p > 0$;
  $-\ln x$.

To the moment it is not clear why these functions are convex; in the mean time we shall derive a simple analytic criterion for detecting convexity which immediately demonstrates that the above functions indeed are convex.

A very convenient equivalent definition of a convex function is in terms of its *epigraph*. Given a real-valued function f defined on a nonempty subset Q of $\mathbf{R}^n$, we define its epigraph as the set

$$\mathrm{Epic}(f) = \{(t,x) \in \mathbf{R}^{n+1} : x \in Q, t \geq f(x)\};$$

geometrically, to define the epigraph, you should take the *graph* of the function – the surface $\{t = f(x), x \in Q\}$ in $\mathbf{R}^{n+1}$ – and add to this surface all points which are "above" it. The equivalent, more geometrical, definition of a convex function is given by the following simple statement (prove it!):

Proposition C.1.1 [definition of convexity in terms of the epigraph]
A function f defined on a subset of $\mathbf{R}^n$ is convex if and only if its epigraph is a nonempty convex set in $\mathbf{R}^{n+1}$.

More examples of convex functions: norms. Equipped with Proposition C.1.1, we can extend our initial list of convex functions (several one-dimensional functions and affine ones) by more examples – *norms*. Let $\pi(x)$ be a norm on $\mathbf{R}^n$ (see Section B.1.2.2). To the moment we know three examples of norms – the Euclidean norm $\|x\|_2 = \sqrt{x^T x}$, the 1-norm $\|x\|_1 = \sum\limits_i |x_i|$ and the ∞-norm $\|x\|_\infty = \max\limits_i |x_i|$. It was also claimed (although not proved; for proof, see p. 486) that these are three members of an infinite family of norms

$$\|x\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}, \quad 1 \leq p \leq \infty$$

(the right hand side of the latter relation for $p = \infty$ is, by definition, $\max\limits_i |x_i|$).

We are about to prove that every norm is convex:

Proposition C.1.2 *Let $\pi(x)$ be a real-valued function on $\mathbf{R}^n$ which is positively homogeneous of degree 1:*

$$\pi(tx) = t\pi(x) \quad \forall x \in \mathbf{R}^n, t \geq 0.$$

π is convex iff it is subadditive:

$$\pi(x+y) \leq \pi(x) + \pi(y) \quad \forall x, y \in \mathbf{R}^n.$$

In particular, a norm (which by definition is positively homogeneous of degree 1 and is subadditive) is convex.

Proof is immediate: the epigraph of a positively homogeneous of degree 1 function π clearly is a conic set: $(t,x) \in \mathrm{Epic}(\pi) \Rightarrow \lambda(t,x) \in \mathrm{Epic}(\pi)$ whenever $\lambda \geq 0$. Now, by Proposition C.1.1 π is convex iff $\mathrm{Epic}(\pi)$ is convex. It is clear that a conic set is convex iff it contains the sum of every two its elements (why ?); this latter property is satisfied for the epigraph of a real-valued function iff the function is subadditive (evident). □

Conjugate norm. Given a norm $\|\cdot\|$ on Euclidean space E with inner product $\langle\cdot,\cdot\rangle$, one can associate with it the function

$$\|x\|_* = \max_y\{y^T x : \|y\| \leq 1\}.$$

On a closest inspection, $\|\cdot\|_*$ is a norm, called the norm *conjugate* to $\|\cdot\|$. From the definition of the conjugate norm it immediately follows that

$$|\langle x, y\rangle| \leq \|x\|\|y\|_* \ \forall x, y \in E \qquad (!)$$

It can be proved that for every fixed y, $\|y\|_*$ is the smallest constant which makes (!) valid for all $x \in E$, and that twice taken, conjugacy recovers the original norm:

$$(\|\cdot\|_*)_* \equiv \|\cdot\|.$$

When $E = \mathbf{R}^n$, the norm conjugate to $\|\cdot\|_p$ is the norm $\|\cdot\|_q$ with $\frac{1}{p} + \frac{1}{1} = 1$, see p. 486.

Frobenius and Spectral norms. There are several natural norms on the space $\mathbf{R}^{m\times n}$ of matrices, the most frequently used being

- Frobenius norm
$$\|A\|_F = \sqrt{\mathrm{Tr}(AA^T)} = \sqrt{\sum_{i,j} A_{ij}^2}.$$
– Euclidean norm induced by the Frobenius inner product,
- Spectral norm, denoted $\|\cdot\|_{2,2}$, or sometimes just $\|\cdot\|$:
$$\|A\|_{2,2} = \max_x \{\|Ax\|_2 : \|x\|_2 \leq 1\};$$
Equivalently, $\|A\|_{2,2}$ is the largest singular value of A.
- Shatten p-norm $\|A\|_{\mathrm{Sh},p}$ – the p-norm of the vector of singular values of A. It can be proved that when $1 \leq p \leq \infty$, the Shatten p-norm is indeed a norm, and the conjugate norm is the Shatten q-norm with $\frac{1}{p}+\frac{1}{q}=1$.

Note that $\|A\|_F = \|A\|_{\mathrm{Sh},2}$, and $\|A\|_{2,2} = \|A\|_{\mathrm{Sh},\infty}$. The Shatten 1-norm of a matrix A – the sum of singular values of the matrix – has a name, it is called the *nuclear norm* $\|\cdot\|_{\mathrm{UNC}}$. By the above, the spectral and the nuclear norms are conjugates of each other. When $m=n$, the restriction of the nuclear norm onto the subspace $\mathbf{S}^n$ of $\mathbf{R}^{n\times n}$, i.e., the sum of magnitudes of eigenvalues of a symmetric matrix, is called *trace* norm.

C.1.2 Elementary properties of convex functions

C.1.2.1 Jensen's inequality

The following elementary observation is, we believe, one of the most useful observations in the world:

Proposition C.1.3 [Jensen's inequality] *Let f be convex and Q be the domain of f. Then for every convex combination*
$$\sum_{i=1}^N \lambda_i x_i$$
of points from Q one has
$$f(\sum_{i=1}^N \lambda_i x_i) \leq \sum_{i=1}^N \lambda_i f(x_i).$$

The proof is immediate: the points $(f(x_i), x_i)$ clearly belong to the epigraph of f; since f is convex, its epigraph is a convex set, so that the convex combination
$$\sum_{i=1}^N \lambda_i (f(x_i), x_i) = (\sum_{i=1}^N \lambda_i f(x_i), \sum_{i=1}^N \lambda_i x_i)$$
of the points also belongs to Epic(f). By definition of the epigraph, the latter means exactly that $\sum\limits_{i=1}^N \lambda_i f(x_i) \geq f(\sum\limits_{i=1}^N \lambda_i x_i)$. □

Note that the definition of convexity of a function f is exactly the requirement on f to satisfy the Jensen inequality for the case of $N=2$; we see that to satisfy this inequality for $N=2$ is the same as to satisfy it for *all* N.

C.1.2.2 Convexity of level sets of a convex function

The following simple observation is also very useful:

Proposition C.1.4 [convexity of level sets] *Let f be a convex function with the domain Q. Then, for every real α, the set*
$$\mathrm{Levy}_\alpha(f) = \{x \in Q : f(x) \leq \alpha\}$$
– the level set of f – is convex.

The proof takes one line: if $x, y \in \mathrm{Levy}_\alpha(f)$ and $\lambda \in [0,1]$, then $f(\lambda x+(1-\lambda)y) \leq \lambda f(x)+(1-\lambda)f(y) \leq \lambda\alpha + (1-\lambda)\alpha = \alpha$, so that $\lambda x + (1-\lambda)y \in \mathrm{Levy}_\alpha(f)$.

Note that the convexity of level sets does *not* characterize convex functions; there are nonconvex functions which share this property (e.g., every monotone function on the axis). The "proper" characterization of convex functions in terms of convex sets is given by Proposition C.1.1 – convex functions are exactly the functions with convex epigraphs. Convexity of level sets specify a wider family of functions, the so called *quasiconvex* ones.

C.1.3 What is the value of a convex function outside its domain?

Literally, the question which entitles this subsection is senseless. Nevertheless, when speaking about *convex* functions, it is extremely convenient to think that the function outside its domain also has a value, namely, takes the value $+\infty$; with this convention, we can say that

a convex function f on $\mathbf{R}^n$ is a function taking values in the extended real axis $\mathbf{R} \cup \{+\infty\}$ such that the domain $\mathrm{Dom}\, f$ of the function – the set of those x's where $f(x)$ is finite – is nonempty, and for all $x, y \in \mathbf{R}^n$ and all $\lambda \in [0,1]$ one has

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y). \tag{C.1.2}$$

If the expression in the right hand side involves infinities, it is assigned the value according to the standard and reasonable conventions on what are arithmetic operations in the "extended real axis" $\mathbf{R}\cup\{+\infty\}\cup\{-\infty\}$:

- arithmetic operations with reals are understood in their usual sense;
- the sum of $+\infty$ and a real, same as the sum of $+\infty$ and $+\infty$ is $+\infty$; similarly, the sum of a real and $-\infty$, same as the sum of $-\infty$ and $-\infty$ is $-\infty$. The sum of $+\infty$ and $-\infty$ is undefined;
- the product of a real and $+\infty$ is $+\infty$, 0 or $-\infty$, depending on whether the real is positive, zero or negative, and similarly for the product of a real and $-\infty$. The product of two "infinities" is again infinity, with the usual rule for assigning the sign to the product.

Note that it is not clear in advance that our new definition of a convex function is equivalent to the initial one: initially we included into the definition requirement for the domain to be convex, and now we omit explicit indicating this requirement. In fact, of course, the definitions are equivalent: convexity of $\mathrm{Dom}\, f$ – i.e., the set where f is finite – is an immediate consequence of the "convexity inequality" (C.1.2).

It is convenient to think of a convex function as of something which is defined everywhere, since it saves a lot of words. E.g., with this convention we can write $f+g$ (f and g are convex functions on $\mathbf{R}^n$), and everybody will understand what is meant; without this convention, we were supposed to add to this expression the explanation as follows: "$f+g$ is a function with the domain being the intersection of those of f and g, and in this intersection it is defined as $(f+g)(x) = f(x)+g(x)$".

C.2 How to detect convexity

In an optimization problem

$$\min_x \{f(x) : g_j(x) \leq 0,\, j = 1, ..., m\}$$

convexity of the objective f and the constraints g_i is crucial: it turns out that problems with this property possess nice theoretical properties (e.g., the local *necessary* optimality conditions for these problems are *sufficient for global optimality*); and what is much more important, convex problems can be efficiently (both in theoretical and, to some extent, in the practical meaning of the word) solved numerically, which is not, unfortunately, the case for general nonconvex problems. This is why it is so important to know how one can detect convexity of a given function. This is the issue we are coming to.

The scheme of our investigation is typical for mathematics. Let me start with the example which you know from Analysis. How do you detect continuity of a function? Of course, there is a definition of continuity in terms of ϵ and δ, but it would be an actual disaster if each time we need to prove continuity of a function, we were supposed to write down the proof that "for every positive ϵ there exists positive δ such that ...". In fact we use another approach: we list once for ever a number of standard operations which preserve continuity, like addition, multiplication, taking superpositions, etc., and point out a number of standard examples of continuous functions – like the power function, the exponent, etc. To prove that the operations in the list preserve continuity, same as to prove that the standard functions are continuous, this takes certain effort and indeed is done in $\epsilon - \delta$ terms; but after this effort is once invested, we normally have no difficulties with proving continuity of a given function: it suffices to demonstrate that the function can be obtained, in finitely many steps, from our "raw materials" – the standard functions which are known to be continuous – by applying our machinery – the combination rules which preserve continuity. Normally this demonstration is given by a single word "evident" or even is understood by default.

This is exactly the case with convexity. Here we also should point out the list of operations which preserve convexity and a number of standard convex functions.

C.2.1 Operations preserving convexity of functions

These operations are as follows:

- [stability under taking weighted sums] if f, g are convex functions on $\mathbf{R}^n$, then their linear combination $\lambda f + \mu g$ with *nonnegative* coefficients again is convex, provided that it is finite at least at one point;

  [this is given by straightforward verification of the definition]

- [stability under affine substitutions of the argument] the superposition $f(Ax+b)$ of a convex function f on $\mathbf{R}^n$ and affine mapping $x \mapsto Ax+b$ from $\mathbf{R}^m$ into $\mathbf{R}^n$ is convex, provided that it is finite at least at one point.

  [you can prove it directly by verifying the definition or by noting that the epigraph of the superposition, if nonempty, is the inverse image of the epigraph of f under an affine mapping]

- [stability under taking pointwise sup] upper bound $\sup\limits_{\alpha} f_\alpha(\cdot)$ of every family of convex functions on $\mathbf{R}^n$ is convex, provided that this bound is finite at least at one point.

  [to understand it, note that the epigraph of the upper bound clearly is the intersection of epigraphs of the functions from the family; recall that the intersection of every family of convex sets is convex]

- ["Convex Monotone superposition"] Let $f(x) = (f_1(x), ..., f_k(x))$ be vector-function on $\mathbf{R}^n$ with convex components f_i, and assume that F is a convex function on $\mathbf{R}^k$ which is monotone, i.e., such that $z \le z'$ always implies that $F(z) \le F(z')$. Then the superposition

$$\phi(x) = F(f(x)) = F(f_1(x), ..., f_k(x))$$

  is convex on $\mathbf{R}^n$, provided that it is finite at least at one point.

  Remark C.2.1 *The expression $F(f_1(x), ..., f_k(x))$ makes no evident sense at a point x where some of f_i's are $+\infty$. By definition, we assign the superposition at such a point the value $+\infty$.*

  [To justify the rule, note that if $\lambda \in (0,1)$ and $x, x' \in \mathrm{Dom}\,\phi$, then $z = f(x), z' = f(x')$ are vectors from $\mathbf{R}^k$ which belong to $\mathrm{Dom}\,F$, and due to the convexity of the components of f we have

$$f(\lambda x + (1-\lambda)x') \le \lambda z + (1-\lambda)z';$$

  in particular, the left hand side is a vector from $\mathbf{R}^k$ – it has no "infinite entries", and we may further use the monotonicity of F:

$$\phi(\lambda x + (1-\lambda)x') = F(f(\lambda x + (1-\lambda)x')) \le F(\lambda z + (1-\lambda)z')$$

  and now use the convexity of F:

$$F(\lambda z + (1-\lambda)z') \le \lambda F(z) + (1-\lambda)F(z')$$

  to get the required relation

$$\phi(\lambda x + (1-\lambda)x') \le \lambda\phi(x) + (1-\lambda)\phi(x').$$

  ]

Imagine how many extra words would be necessary here if there were no convention on the value of a convex function outside its domain!

Two more rules are as follows:

- [stability under partial minimization] if $f(x,y) : \mathbf{R}^n_x \times \mathbf{R}^m_y$ is convex (as a function of $z = (x,y)$; this is called *joint convexity*) and the function

$$g(x) = \inf_y f(x,y)$$

  is *proper*, i.e., is $> -\infty$ everywhere and is finite at least at one point, then g is convex [this can be proved as follows. We should prove that if $x, x' \in \mathrm{Dom}\, g$ and $x'' = \lambda x + (1-\lambda)x'$ with $\lambda \in [0,1]$, then $x'' \in \mathrm{Dom}\, g$ and $g(x'') \leq \lambda g(x) + (1-\lambda)g(x')$. Given positive ϵ, we can find y and y' such that $(x,y) \in \mathrm{Dom}\, f$, $(x',y') \in \mathrm{Dom}\, f$ and $g(x) + \epsilon \geq f(x,y)$, $g(y') + \epsilon \geq f(x',y')$. Taking weighted sum of these two inequalities, we get

$$\begin{array}{l} \lambda g(x) + (1-\lambda)g(y) + \epsilon \geq \lambda f(x,y) + (1-\lambda)f(x',y') \\ \geq f(\lambda x + (1-\lambda)x', \lambda y + (1-\lambda)y') = f(x'', \lambda y + (1-\lambda)y') \end{array}$$

  (the last $\geq$ follows from the convexity of f). The concluding quantity in the chain is $\geq g(x'')$, and we get $g(x'') \leq \lambda g(x) + (1-\lambda)g(x') + \epsilon$. In particular, $x'' \in \mathrm{Dom}\, g$ (recall that g is assumed to take only the values from $\mathbf{R}$ and the value $+\infty$). Moreover, since the resulting inequality is valid for all $\epsilon > 0$, we come to $g(x'') \leq \lambda g(x) + (1-\lambda)g(x')$, as required.]

- the "conic transformation" of a convex function f on $\mathbf{R}^n$ – the function $g(y,x) = yf(x/y)$ – is convex in the half-space $y > 0$ in $\mathbf{R}^{n+1}$.

Now we know what are the basic operations preserving convexity. Let us look what are the standard functions these operations can be applied to. A number of examples was already given, but we still do not know why the functions in the examples are convex. The usual way to check convexity of a "simple" – given by a simple formula – function is based on *differential criteria of convexity*. Let us look what are these criteria.

C.2.2 Differential criteria of convexity

From the definition of convexity of a function if immediately follows that convexity is one-dimensional property: a proper (i.e., finite at least at one point) function f on $\mathbf{R}^n$ taking values in $\mathbf{R} \cup \{+\infty\}$ is convex if and only if its restriction on every line, i.e., every function of the type $g(t) = f(x+th)$ on the axis, is either convex, or is identically $+\infty$.

It follows that to detect convexity of a function, it, in principle, suffices to know how to detect convexity of functions of one variable. This latter question can be resolved by the standard Calculus tools. Namely, in the Calculus they prove the following simple

Proposition C.2.1 [Necessary and Sufficient Convexity Condition for smooth functions on the axis] *Let (a,b) be an interval in the axis (we do not exclude the case of $a = -\infty$ and/or $b = +\infty$). Then*

(i) *A differentiable everywhere on (a,b) function f is convex on (a,b) iff its derivative f' is monotonically nondecreasing on (a,b);*

(ii) *A twice differentiable everywhere on (a,b) function f is convex on (a,b) iff its second derivative f'' is nonnegative everywhere on (a,b).*

With the Proposition, you can immediately verify that the functions listed as examples of convex functions in Section C.1.1 indeed are convex. The only difficulty which you may meet is that some of these functions (e.g., x^p, $p \geq 1$, and $-x^p$, $0 \leq p \leq 1$, were claimed to be convex on the half-interval $[0, +\infty)$, while the Proposition speaks about convexity of functions on intervals. To overcome this difficulty, you may use the following simple

Proposition C.2.2 *Let M be a convex set and f be a function with $\mathrm{Dom}\, f = M$. Assume that f is convex on $\mathrm{ri}\, M$ and is continuous on M, i.e.,*

$$f(x_i) \to f(x),\ i \to \infty,$$

whenever $x_i, x \in M$ and $x_i \to x$ as $i \to \infty$. Then f is convex on M.

Proof of Proposition C.2.1:

(i), necessity. Assume that f is differentiable and convex on (a,b); we should prove that then f' is monotonically nondecreasing. Let $x < y$ be two points of (a,b), and let us prove that $f'(x) \leq f'(y)$. Indeed, let $z \in (x,y)$. We clearly have the following representation of z as a convex combination of x and y:

$$z = \frac{y-z}{y-x}x + \frac{x-z}{y-x}y,$$

whence, from convexity,

$$f(z) \leq \frac{y-z}{y-x}f(x) + \frac{x-z}{y-x}f(y),$$

whence

$$\frac{f(z)-f(x)}{x-z} \leq \frac{f(y)-f(z)}{y-z}.$$

Passing here to limit as $z \to x+0$, we get

$$f'(x) \leq \frac{(f(y)-f(x)}{y-x},$$

and passing in the same inequality to limit as $z \to y-0$, we get

$$f'(y) \geq \frac{(f(y)-f(x)}{y-x},$$

whence $f'(x) \leq f'(y)$, as claimed. □

(i), sufficiency. We should prove that if f is differentiable on (a,b) and f' is monotonically nondecreasing on (a,b), then f is convex on (a,b). It suffices to verify that if $x < y$, $x, y \in (a,b)$, and $z = \lambda x + (1-\lambda)y$ with $0 < \lambda < 1$, then

$$f(z) \leq \lambda f(x) + (1-\lambda)f(y),$$

or, which is the same (write $f(z)$ as $\lambda f(z) + (1-\lambda)f(z)$), that

$$\frac{f(z)-f(x)}{\lambda} \leq \frac{f(y)-f(z)}{1-\lambda}.$$

noticing that $z - x = \lambda(y-x)$ and $y - z = (1-\lambda)(y-x)$, we see that the inequality we should prove is equivalent to

$$\frac{f(z)-f(x)}{z-x} \leq \frac{f(y)-f(z)}{y-z}.$$

But in this equivalent form the inequality is evident: by the Lagrange Mean Value Theorem, its left hand side is $f'(\xi)$ with some $\xi \in (x,z)$, while the right hand one is $f'(\eta)$ with some $\eta \in (z,y)$. Since f' is nondecreasing and $\xi \leq z \leq \eta$, we have $f'(\xi) \leq f'(\eta)$, and the left hand side in the inequality we should prove indeed is $\leq$ the right hand one. □

(ii) is immediate consequence of (i), since, as we know from the very beginning of Calculus, a differentiable function – in the case in question, it is f' – is monotonically nondecreasing on an interval iff its derivative is nonnegative on this interval. □

In fact, for functions of one variable there is a differential criterion of convexity which does not assume any smoothness (we shall not prove this criterion):

Proposition C.2.3 [convexity criterion for univariate functions]

Let $g : \mathbf{R} \to \mathbf{R} \cup \{+\infty\}$ be a function. Let the domain $\Delta = \{t : g(t) < \infty\}$ of the function be a convex set which is not a singleton, i.e., let it be an interval (a,b) with possibly added one or both endpoints ($-\infty \leq a < b \leq \infty$). g is convex iff it satisfies the following 3 requirements:

1) g is continuous on (a,b);

2) g is differentiable everywhere on (a,b), excluding, possibly, a countable set of points, and the derivative $g'(t)$ is nondecreasing on its domain;

3) at each endpoint u of the interval (a,b) which belongs to Δ g is upper semicontinuous:

$$g(u) \geq \limsup_{t \in (a,b), t \to u} g(t).$$

Proof of Proposition C.2.2: Let $x, y \in M$ and $z = \lambda x + (1-\lambda)y$, $\lambda \in [0,1]$, and let us prove that

$$f(z) \leq \lambda f(x) + (1-\lambda)f(y).$$

As we know from Theorem B.1.1.(iii), there exist sequences $x_i \in \operatorname{ri} M$ and $y_i \in \operatorname{ri} M$ converging, respectively to x and to y. Then $z_i = \lambda x_i + (1-\lambda)y_i$ converges to z as $i \to \infty$, and since f is convex on $\operatorname{ri} M$, we have

$$f(z_i) \leq \lambda f(x_i) + (1-\lambda)f(y_i);$$

passing to limit and taking into account that f is continuous on M and x_i, y_i, z_i converge, as $i \to \infty$, to $x, y, z \in M$, respectively, we obtain the required inequality. □

From Propositions C.2.1.(ii) and C.2.2 we get the following convenient *necessary and sufficient* condition for convexity of a *smooth* function of n variables:

Corollary C.2.1 [convexity criterion for smooth functions on $\mathbf{R}^n$]

Let $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ be a function. Assume that the domain Q of f is a convex set with a nonempty interior and that f is

- *continuous on Q*

  and

- *twice differentiable on the interior of Q.*

Then f is convex iff its Hessian is positive semidefinite on the interior of Q:

$$h^T f''(x) h \geq 0 \;\; \forall x \in \operatorname{int} Q \; \forall h \in \mathbf{R}^n.$$

Proof.* The "only if" part is evident: if f is convex and $x \in Q' = \operatorname{int} Q$, then the function of one variable

$$g(t) = f(x + th)$$

(h is an arbitrary fixed direction in $\mathbf{R}^n$) is convex in certain neighbourhood of the point $t = 0$ on the axis (recall that affine substitutions of argument preserve convexity). Since f is twice differentiable in a neighbourhood of x, g is twice differentiable in a neighbourhood of $t = 0$, so that $g''(0) = h^T f''(x) h \geq 0$ by Proposition C.2.1. □

Now let us prove the "if" part, so that we are given that $h^T f''(x) h \geq 0$ for every $x \in \operatorname{int} Q$ and every $h \in \mathbf{R}^n$, and we should prove that f is convex.

Let us first prove that f is convex on the interior Q' of the domain Q. By Theorem B.1.1, Q' is a convex set. Since, as it was already explained, the convexity of a function on a convex set is one-dimensional fact, all we should prove is that every one-dimensional function

$$g(t) = f(x + t(y-x)), \;\; 0 \leq t \leq 1$$

(x and y are from Q') is convex on the segment $0 \leq t \leq 1$. Since f is continuous on $Q \supset Q'$, g is continuous on the segment; and since f is twice continuously differentiable on Q', g is continuously differentiable on $(0,1)$ with the second derivative

$$g''(t) = (y-x)^T f''(x + t(y-x))(y-x) \geq 0.$$

Consequently, g is convex on $[0,1]$ (Propositions C.2.1.(ii) and C.2.2). Thus, f is convex on Q'. It remains to note that f, being convex on Q' and continuous on Q, is convex on Q by Proposition C.2.2. □

Applying the combination rules preserving convexity to simple functions which pass the "infinitesimal' convexity tests, we can prove convexity of many complicated functions. Consider, e.g., an *exponential posynomial* – a function

$$f(x) = \sum_{i=1}^{N} c_i \exp\{a_i^T x\}$$

with positive coefficients c_i (this is why the function is called *pos*ynomial). How could we prove that the function is convex? This is immediate:

$\exp\{t\}$ is convex (since its second order derivative is positive and therefore the first derivative is monotone, as required by the infinitesimal convexity test for smooth functions of one variable);

consequently, all functions $\exp\{a_i^T x\}$ are convex (stability of convexity under affine substitutions of argument);

consequently, f is convex (stability of convexity under taking linear combinations with nonnegative coefficients).

And if we were supposed to prove that the maximum of three posynomials is convex? Ok, we could add to our three steps the fourth, which refers to stability of convexity under taking pointwise supremum.

C.3 Gradient inequality

An extremely important property of a convex function is given by the following

Proposition C.3.1 [Gradient inequality] *Let f be a function taking finite values and the value $+\infty$, x be an interior point of the domain of f and Q be a convex set containing x. Assume that*

- *f is convex on Q*

and

- *f is differentiable at x,*

and let $\nabla f(x)$ be the gradient of the function at x. Then the following inequality holds:

$$(\forall y \in Q): \quad f(y) \geq f(x) + (y-x)^T \nabla f(x). \tag{C.3.1}$$

Geometrically: the graph

$$\{(y,t) \in \mathbf{R}^{n+1} : y \in \mathrm{Dom}\, f \cap Q,\ t = f(y)\}$$

of the function f restricted onto the set Q is above the graph

$$\{(y,t) \in \mathbf{R}^{n+1} : t = f(x) + (y-x)^T \nabla f(x)\}$$

of the linear form tangent to f at x.

Proof. Let $y \in Q$. There is nothing to prove if $y \notin \mathrm{Dom}\, f$ (since there the right hand side in the gradient inequality is $+\infty$), same as there is nothing to prove when $y = x$. Thus, we can assume that $y \neq x$ and $y \in \mathrm{Dom}\, f$. Let us set

$$y_\tau = x + \tau(y-x), \ \ 0 < \tau \leq 1,$$

so that $y_1 = y$ and y_τ is an interior point of the segment $[x,y]$ for $0 < \tau < 1$. Now let us use the following extremely simple

Lemma C.3.1 *Let x, x', x'' be three distinct points with $x' \in [x, x'']$, and let f be convex and finite on $[x, x'']$. Then*

$$\frac{f(x') - f(x)}{\|x' - x\|_2} \leq \frac{f(x'') - f(x)}{\|x'' - x\|_2}. \tag{C.3.2}$$

Proof of the Lemma. We clearly have

$$x' = x + \lambda(x'' - x), \quad \lambda = \frac{\|x' - x\|_2}{\|x'' - x\|_2} \in (0,1)$$

or, which is the same,

$$x' = (1-\lambda)x + \lambda x''.$$

From the convexity inequality

$$f(x') \leq (1-\lambda) f(x) + \lambda f(x''),$$

or, which is the same,

$$f(x') - f(x) \leq \lambda(f(x'') - f(x')).$$

Dividing by λ and substituting the value of λ, we come to (C.3.2). □

Applying the Lemma to the triple $x, x' = y_\tau, x'' = y$, we get

$$\frac{f(x+\tau(y-x)) - f(x)}{\tau\|y-x\|_2} \le \frac{f(y)-f(x)}{\|y-x\|_2};$$

as $\tau \to +0$, the left hand side in this inequality, by the definition of the gradient, tends to $\|y-x\|_2^{-1}(y-x)^T\nabla f(x)$, and we get

$$\|y-x\|_2^{-1}(y-x)^T\nabla f(x) \le \|y-x\|_2^{-1}(f(y)-f(x)),$$

or, which is the same,

$$(y-x)^T\nabla f(x) \le f(y) - f(x);$$

this is exactly the inequality (C.3.1). □

It is worthy of mentioning that in the case when Q is convex set with a nonempty interior and f is continuous on Q and differentiable on $\operatorname{int} Q$, f is convex on Q iff the Gradient inequality (C.3.1) is valid for every pair $x \in \operatorname{int} Q$ and $y \in Q$.

Indeed, the "only if" part, i.e., the implication

convexity of f $\Rightarrow$ Gradient inequality for all $x \in \operatorname{int} Q$ and all $y \in Q$

is given by Proposition C.3.1. To prove the "if" part, i.e., to establish the implication inverse to the above, assume that f satisfies the Gradient inequality for all $x \in \operatorname{int} Q$ and all $y \in Q$, and let us verify that f is convex on Q. It suffices to prove that f is convex on the interior Q' of the set Q (see Proposition C.2.2; recall that by assumption f is continuous on Q and Q is convex). To prove that f is convex on Q', note that Q' is convex (Theorem B.1.1) and that, due to the Gradient inequality, on Q' f is the upper bound of the family of affine (and therefore convex) functions:

$$f(y) = \sup_{x\in Q'} f_x(y), \quad f_x(y) = f(x) + (y-x)^T\nabla f(x).$$ □

C.4 Boundedness and Lipschitz continuity of a convex function

Convex functions possess nice local properties.

Theorem C.4.1 [local boundedness and Lipschitz continuity of convex function]
Let f be a convex function and let K be a closed and bounded set contained in the relative interior of the domain $\operatorname{Dom} f$ of f. Then f is Lipschitz continuous on K – there exists constant L – the Lipschitz constant of f on K – such that

$$|f(x) - f(y)| \le L\|x-y\|_2 \quad \forall x, y \in K. \tag{C.4.1}$$

In particular, f is bounded on K.

Remark C.4.1 All three assumptions on K – (1) closedness, (2) boundedness, and (3) $K \subset \operatorname{ri} \operatorname{Dom} f$ – are essential, as it is seen from the following three examples:

- $f(x) = 1/x$, $\operatorname{Dom} F = (0, +\infty)$, $K = (0, 1]$. We have (2), (3) but not (1); f is neither bounded, nor Lipschitz continuous on K.
- $f(x) = x^2$, $\operatorname{Dom} f = \mathbf{R}$, $K = \mathbf{R}$. We have (1), (3) and not (2); f is neither bounded nor Lipschitz continuous on K.
- $f(x) = -\sqrt{x}$, $\operatorname{Dom} f = [0, +\infty)$, $K = [0, 1]$. We have (1), (2) and not (3); f is not Lipschitz continuous on K [1], although is bounded. With properly chosen convex function f of two variables and non-polyhedral compact domain (e.g., with $\operatorname{Dom} f$ being the unit circle), we could demonstrate also that lack of (3), even in presence of (1) and (2), may cause unboundedness of f at K as well.

[1] indeed, we have $\lim\limits_{t\to+0} \frac{f(0)-f(t)}{t} = \lim\limits_{t\to+0} t^{-1/2} = +\infty$, while for a Lipschitz continuous f the ratios $t^{-1}(f(0)-f(t))$ should be bounded

Remark C.4.2 *Theorem C.4.1 says that a convex function f is bounded on every compact (i.e., closed and bounded) subset of the relative interior of* Dom f*. In fact there is much stronger statement on the below boundedness of f: f is below bounded on any bounded subset of $\mathbf{R}^n$!.*

Proof of Theorem C.4.1. We shall start with the following local version of the Theorem.

Proposition C.4.1 *Let f be a convex function, and let $\bar{x}$ be a point from the relative interior of the domain* Dom f *of f. Then*

(i) *f is bounded at $\bar{x}$: there exists a positive r such that f is bounded in the r-neighbourhood $U_r(\bar{x})$ of $\bar{x}$ in the affine span of* Dom f*:*

$$\exists r > 0, C : \quad |f(x)| \le C \quad \forall x \in U_r(\bar{x}) = \{x \in \mathrm{Aff}(\mathrm{Dom}\, f) : \|x - \bar{x}\|_2 \le r\};$$

(ii) *f is Lipschitz continuous at $\bar{x}$, i.e., there exists a positive ρ and a constant L such that*

$$|f(x) - f(x')| \le L\|x - x'\|_2 \ \forall x, x' \in U_\rho(\bar{x}).$$

Implication "Proposition C.4.1 ⇒ Theorem C.4.1" is given by standard Analysis reasoning. All we need is to prove that if K is a bounded and closed (i.e., a compact) subset of ri Dom f, then f is Lipschitz continuous on K (the boundedness of f on K is an evident consequence of its Lipschitz continuity on K and boundedness of K). Assume, on contrary, that f is not Lipschitz continuous on K; then for every integer i there exists a pair of points $x_i, y_i \in K$ such that

$$f(x_i) - f(y_i) \ge i\|x_i - y_i\|_2. \tag{C.4.2}$$

Since K is compact, passing to a subsequence we can ensure that $x_i \to x \in K$ and $y_i \to y \in K$. By Proposition C.4.1 the case $x = y$ is impossible – by Proposition f is Lipschitz continuous in a neighbourhood B of $x = y$; since $x_i \to x, y_i \to y$, this neighbourhood should contain all x_i and y_i with large enough indices i; but then, from the Lipschitz continuity of f in B, the ratios $(f(x_i) - f(y_i))/\|x_i - y_i\|_2$ form a bounded sequence, which we know is not the case. Thus, the case $x = y$ is impossible. The case $x \neq y$ is "even less possible" – since, by Proposition, f is continuous on Dom f at both the points x and y (note that Lipschitz continuity at a point clearly implies the usual continuity at it), so that we would have $f(x_i) \to f(x)$ and $f(y_i) \to f(y)$ as $i \to \infty$. Thus, the left hand side in (C.4.2) remains bounded as $i \to \infty$. In the right hand side one factor – i – tends to ∞, and the other one has a nonzero limit $\|x - y\|$, so that the right hand side tends to ∞ as $i \to \infty$; this is the desired contradiction. □

Proof of Proposition C.4.1.

1^0. We start with proving the *above boundedness* of f in a neighbourhood of $\bar{x}$. This is immediate: we know that there exists a neighbourhood $U_{\bar{r}}(\bar{x})$ which is contained in Dom f (since, by assumption, $\bar{x}$ is a relative interior point of Dom f). Now, we can find a small simplex Δ of the dimension $m = \dim \mathrm{Aff}(\mathrm{Dom}\, f)$ with the vertices $x_0, ..., x_m$ in $U_{\bar{r}}(\bar{x})$ in such a way that $\bar{x}$ will be a convex combination of the vectors x_i with *positive* coefficients, even with the coefficients $1/(m+1)$:

$$\bar{x} = \sum_{i=0}^{m} \frac{1}{m+1} x_i \quad ^{2)}.$$

We know that $\bar{x}$ is the point from the relative interior of Δ (Exercise B.8); since Δ spans the same affine subspace as Dom f, it means that Δ contains $U_r(\bar{x})$ with certain $r > 0$. Now, we have

$$\Delta = \{\sum_{i=0}^{m} \lambda_i x_i : \lambda_i \ge 0, \sum_i \lambda_i = 1\}$$

[2] to see that the required Δ exists, let us act as follows: first, the case of Dom f being a singleton is evident, so that we can assume that Dom f is a convex set of dimension $m \ge 1$. Without loss of generality, we may assume that $\bar{x} = 0$, so that $0 \in$ Dom f and therefore Aff(Dom f) = Lin(Dom f). By Linear Algebra, we can find m vectors $y_1, ..., y_m$ in Dom f which form a basis in Lin(Dom f) = Aff(Dom f). Setting $y_0 = -\sum_{i=1}^{m} y_i$ and taking into account that $0 = \bar{x} \in$ ri Dom f, we can find $\epsilon > 0$ such that the vectors $x_i = \epsilon y_i$, $i = 0, ..., m$, belong to $U_{\bar{r}}(\bar{x})$. By construction, $\bar{x} = 0 = \frac{1}{m+1} \sum_{i=0}^{m} x_i$.

so that in Δ f is bounded from above by the quantity $\max\limits_{0\leq i\leq m} f(x_i)$ by Jensen's inequality:

$$f(\sum_{i=0}^{m}\lambda_i x_i) \leq \sum_{i=0}^{m}\lambda_i f(x_i) \leq \max_i f(x_i).$$

Consequently, f is bounded from above, by the same quantity, in $U_r(\bar{x})$.

2^0. Now let us prove that if f is above bounded, by some C, in $U_r = U_r(\bar{x})$, then it in fact is below bounded in this neighbourhood (and, consequently, is bounded in U_r). Indeed, let $x \in U_r$, so that $x \in \mathrm{Aff}(\mathrm{Dom}\, f)$ and $\|x-\bar{x}\|_2 \leq r$. Setting $x' = \bar{x} - [x-\bar{x}] = 2\bar{x} - x$, we get $x' \in \mathrm{Aff}(\mathrm{Dom}\, f)$ and $\|x'-\bar{x}\|_2 = \|x-\bar{x}\|_2 \leq r$, so that $x' \in U_r$. Since $\bar{x} = \frac{1}{2}[x+x']$, we have

$$2f(\bar{x}) \leq f(x) + f(x'),$$

whence

$$f(x) \geq 2f(\bar{x}) - f(x') \geq 2f(\bar{x}) - C, \;\; x \in U_r(\bar{x}),$$

and f is indeed below bounded in U_r.

(i) is proved.

3^0. (ii) is an immediate consequence of (i) and Lemma C.3.1. Indeed, let us prove that f is Lipschitz continuous in the neighbourhood $U_{r/2}(\bar{x})$, where $r > 0$ is such that f is bounded in $U_r(\bar{x})$ (we already know from (i) that the required r does exist). Let $|f| \leq C$ in U_r, and let $x, x' \in U_{r/2}$, $x \neq x'$. Let us extend the segment $[x, x']$ through the point x' until it reaches, at certain point x'', the (relative) boundary of U_r. We have

$$x' \in (x, x''); \quad \|x''-\bar{x}\|_2 = r.$$

From (C.3.2) we have

$$f(x') - f(x) \leq \|x'-x\|_2 \frac{f(x'')-f(x)}{\|x''-x\|_2}.$$

The second factor in the right hand side does not exceed the quantity $(2C)/(r/2) = 4C/r$; indeed, the numerator is, in absolute value, at most $2C$ (since $|f|$ is bounded by C in U_r and both x, x'' belong to U_r), and the denominator is at least $r/2$ (indeed, x is at the distance at most $r/2$ from $\bar{x}$, and x'' is at the distance exactly r from $\bar{x}$, so that the distance between x and x'', by the triangle inequality, is at least $r/2$). Thus, we have

$$f(x') - f(x) \leq (4C/r)\|x'-x\|_2, \;\; x, x' \in U_{r/2};$$

swapping x and x', we come to

$$f(x) - f(x') \leq (4C/r)\|x'-x\|_2,$$

whence

$$|f(x) - f(x')| \leq (4C/r)\|x-x'\|_2, \;\; x, x' \in U_{r/2},$$

as required in (ii). □

C.5 Maxima and minima of convex functions

As it was already mentioned, optimization problems involving convex functions possess nice theoretical properties. One of the most important of these properties is given by the following

Theorem C.5.1 ["Unimodality"] *Let f be a convex function on a convex set $Q \subset \mathbf{R}^n$, and let $x^* \in Q \cap \mathrm{Dom}\, f$ be a local minimizer of f on Q:*

$$(\exists r > 0): \quad f(y) \geq f(x^*) \quad \forall y \in Q, \; \|y-x\|_2 < r. \tag{C.5.1}$$

Then x^* is a global minimizer of f on Q:

$$f(y) \geq f(x^*) \quad \forall y \in Q. \tag{C.5.2}$$

Moreover, the set $\mathop{\mathrm{Argmin}}\limits_{Q} f$ of all local ($\equiv$ global) minimizers of f on Q is convex.

If f is strictly convex (i.e., the convexity inequality $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$ is strict whenever $x \neq y$ and $\lambda \in (0,1)$), then the above set is either empty or is a singleton.

Proof. 1) Let x^* be a local minimizer of f on Q and $y \in Q$, $y \neq x^*$; we should prove that $f(y) \geq f(x^*)$. There is nothing to prove if $f(y) = +\infty$, so that we may assume that $y \in \mathrm{Dom}\, f$. Note that also $x^* \in \mathrm{Dom}\, f$ for sure – by definition of a local minimizer.

For all $\tau \in (0,1)$ we have, by Lemma C.3.1,

$$\frac{f(x^* + \tau(y - x^*)) - f(x^*)}{\tau\|y - x^*\|_2} \leq \frac{f(y) - f(x^*)}{\|y - x^*\|_2}.$$

Since x^* is a local minimizer of f, the left hand side in this inequality is nonnegative for all small enough values of $\tau > 0$. We conclude that the right hand side is nonnegative, i.e., $f(y) \geq f(x^*)$. □

2) To prove convexity of $\operatorname*{Argmin}_Q f$, note that $\operatorname*{Argmin}_Q f$ is nothing but the level set $\mathrm{Levy}_\alpha(f)$ of f associated with the minimal value $\min_Q f$ of f on Q; as a level set of a convex function, this set is convex (Proposition C.1.4).

3) To prove that the set $\operatorname*{Argmin}_Q f$ associated with a strictly convex f is, if nonempty, a singleton, note that if there were two distinct minimizers x', x'', then, from strict convexity, we would have

$$f(\frac{1}{2}x' + \frac{1}{2}x'') < \frac{1}{2}[f(x') + f(x'') == \min_Q f,$$

which clearly is impossible - the argument in the left hand side is a point from Q! □

Another pleasant fact is that in the case of differentiable convex functions the known from Calculus necessary optimality condition (the Fermat rule) is sufficient for global optimality:

Theorem C.5.2 [Necessary and sufficient optimality condition for a differentiable convex function]

Let f be convex function on convex set $Q \subset \mathbf{R}^n$, and let x^* be an interior point of Q. Assume that f is differentiable at x^*. Then x^* is a minimizer of f on Q iff

$$\nabla f(x^*) = 0.$$

Proof. As a *necessary* condition for local optimality, the relation $\nabla f(x^*) = 0$ is known from Calculus; it has nothing in common with convexity. The essence of the matter is, of course, the *sufficiency* of the condition $\nabla f(x^*) = 0$ for *global optimality* of x^* in the case of *convex* f. This sufficiency is readily given by the Gradient inequality (C.3.1): by virtue of this inequality and due to $\nabla f(x^*) = 0$,

$$f(y) \geq f(x^*) + (y - x^*)\nabla f(x^*) = f(x^*)$$

for all $y \in Q$. □

A natural question is what happens if x^* in the above statement is not necessarily an interior point of Q. Thus, assume that x^* is an arbitrary point of a convex set Q and that f is convex on Q and differentiable at x^* (the latter means exactly that $\mathrm{Dom}\, f$ contains a neighbourhood of x^* and f possesses the first order derivative at x^*). Under these assumptions, when x^* is a minimizer of f on Q?

The answer is as follows: let

$$T_Q(x^*) = \{h \in \mathbf{R}^n : x^* + th \in Q \quad \forall \text{ small enough } t > 0\}$$

be the *radial cone* of Q at x^*; geometrically, this is the set of all directions leading from x^* inside Q, so that a small enough positive step from x^* along the direction keeps the point in Q. From the convexity of Q it immediately follows that the radial cone indeed is a convex cone (not necessary closed). E.g., when x^* is an interior point of Q, then the radial cone to Q at x^* clearly is the entire $\mathbf{R}^n$. A more interesting example is the radial cone to a polyhedral set

$$Q = \{x : a_i^T x \leq b_i,\ i = 1, ..., m\}; \tag{C.5.3}$$

for $x^* \in Q$ the corresponding radial cone clearly is the polyhedral cone

$$\{h : a_i^T h \leq 0 \quad \forall i : a_i^T x^* = b_i\} \tag{C.5.4}$$

corresponding to the *active* at x^* (i.e., satisfied at the point as equalities rather than as strict inequalities) constraints $a_i^T x \leq b_i$ from the description of Q.

Now, for the functions in question the necessary and sufficient condition for x^* to be a minimizer of f on Q is as follows:

Proposition C.5.1 *Let Q be a convex set, let $x^* \in Q$, and let f be a convex on Q function which is differentiable at x^*. The necessary and sufficient condition for x^* to be a minimizer of f on Q is that the derivative of f taken at x^* along every direction from $T_Q(x^*)$ should be nonnegative:*

$$h^T\nabla f(x^*) \geq 0 \quad \forall h \in T_Q(x^*).$$

Proof is immediate. The necessity is an evident fact which has nothing in common with convexity: assuming that x^* is a local minimizer of f on Q, we note that if there were $h \in T_Q(x^*)$ with $h^T\nabla f(x^*) < 0$, then we would have

$$f(x^* + th) < f(x^*)$$

for all small enough positive t. On the other hand, $x^* + th \in Q$ for all small enough positive t due to $h \in T_Q(x^*)$. Combining these observations, we conclude that in every neighbourhood of x^* there are points from Q with strictly better than the one at x^* values of f; this contradicts the assumption that x^* is a local minimizer of f on Q.

The sufficiency is given by the Gradient Inequality, exactly as in the case when x^* is an interior point of Q. □

Proposition C.5.1 says that whenever f is convex on Q and differentiable at $x^* \in Q$, the necessary and sufficient condition for x^* to be a minimizer of f on Q is that the linear form given by the gradient $\nabla f(x^*)$ of f at x^* should be nonnegative at all directions from the radial cone $T_Q(x^*)$. The linear forms nonnegative at all directions from the radial cone also form a cone; it is called the cone *normal* to Q at x^* and is denoted $N_Q(x^*)$. Thus, Proposition says that the necessary and sufficient condition for x^* to minimize f on Q is the inclusion $\nabla f(x^*) \in N_Q(x^*)$. What does this condition actually mean, it depends on what is the normal cone: whenever we have an explicit description of it, we have an explicit form of the optimality condition.

E.g., when $T_Q(x^*) = \mathbf{R}^n$ (it is the same as to say that x^* is an interior point of Q), then the normal cone is comprised of the linear forms nonnegative at the entire space, i.e., it is the trivial cone $\{0\}$; consequently, for the case in question the optimality condition becomes the Fermat rule $\nabla f(x^*) = 0$, as we already know.

When Q is the polyhedral set (C.5.3), the normal cone is the polyhedral cone (C.5.4); it is comprised of all directions which have nonpositive inner products with all a_i coming from the active, in the aforementioned sense, constraints. The normal cone is comprised of all vectors which have nonnegative inner products with all these directions, i.e., of vectors a such that the inequality $h^Ta \geq 0$ is a consequence of the inequalities $h^Ta_i \leq 0$, $i \in I(x^*) \equiv \{i : a_i^Tx^* = b_i\}$. From the Homogeneous Farkas Lemma we conclude that the normal cone is simply the conic hull of the vectors $-a_i$, $i \in I(x^*)$. Thus, in the case in question (*) reads:

$x^* \in Q$ is a minimizer of f on Q iff there exist nonnegative reals λ_i^* associated with "active" (those from $I(x^*)$) values of i such that

$$\nabla f(x^*) + \sum_{i\in I(x^*)} \lambda_i^* a_i = 0.$$

These are the famous *Karush-Kuhn-Tucker* optimality conditions; these conditions are necessary for optimality in an essentially wider situation.

The indicated results demonstrate that the fact that a point $x^* \in \operatorname{Dom} f$ is a global minimizer of a convex function f depends only on the local behaviour of f at x^*. This is not the case with maximizers of a convex function. First of all, such a maximizer, if exists, in all nontrivial cases should belong to the boundary of the domain of the function:

Theorem C.5.3 *Let f be convex, and let Q be the domain of f. Assume that f attains its maximum on Q at a point x^* from the relative interior of Q. Then f is constant on Q.*

Proof. Let $y \in Q$; we should prove that $f(y) = f(x^*)$. There is nothing to prove if $y = x^*$, so that we may assume that $y \neq x^*$. Since, by assumption, $x^* \in \operatorname{ri} Q$, we can extend the segment $[x^*, y]$ through the endpoint x^*, keeping the left endpoint of the segment in Q; in other words, there exists a point $y' \in Q$ such that x^* is an interior point of the segment $[y', y]$:

$$x^* = \lambda y' + (1-\lambda)y$$

for certain $\lambda \in (0, 1)$. From the definition of convexity

$$f(x^*) \leq \lambda f(y') + (1-\lambda)f(y).$$

Since both $f(y')$ and $f(y)$ do not exceed $f(x^*)$ (x^* is a maximizer of f on Q!) and both the weights λ and $1-\lambda$ are strictly positive, the indicated inequality can be valid only if $f(y') = f(y) = f(x^*)$. □

The next theorem gives further information on maxima of convex functions:

Theorem C.5.4 *Let f be a convex function on $\mathbf{R}^n$ and E be a subset of $\mathbf{R}^n$. Then*

$$\sup_{\mathrm{Conv}\,E} f = \sup_E f. \tag{C.5.5}$$

In particular, if $S \subset \mathbf{R}^n$ is convex and compact set, then the supremum of f on S is equal to the supremum of f on the set of extreme points of S:

$$\sup_S f = \sup_{\mathrm{Ext}(S)} f \tag{C.5.6}$$

Proof. To prove (C.5.5), let $x \in \mathrm{Conv}\,E$, so that x is a convex combination of points from E (Theorem B.1.4 on the structure of convex hull):

$$x = \sum_i \lambda_i x_i \quad [x_i \in E,\, \lambda_i \geq 0,\, \sum_i \lambda_i = 1].$$

Applying Jensen's inequality (Proposition C.1.3), we get

$$f(x) \leq \sum_i \lambda_i f(x_i) \leq \sum_i \lambda_i \sup_E f = \sup_E f,$$

so that the left hand side in (C.5.5) is $\leq$ the right hand one; the inverse inequality is evident, since $\mathrm{Conv}\,E \supset E$. □

To derive (C.5.6) from (C.5.5), it suffices to note that from the Krein-Milman Theorem (Theorem B.2.10) for a convex compact set S one has $S = \mathrm{Conv}\,\mathrm{Ext}(S)$. □

The last theorem on maxima of convex functions is as follows:

Theorem C.5.5 *Let f be a convex function such that the domain Q of f is closed and does not contain lines. Then*

(i) *If the set*

$$\mathop{\mathrm{Argal}}_Q f \equiv \{x \in Q : f(x) \geq f(y)\, \forall y \in Q\}$$

of global maximizers of f is nonempty, then it intersects the set $\mathrm{Ext}(Q)$ of the extreme points of Q, so that at least one of the maximizers of f is an extreme point of Q;

(ii) *If the set Q is polyhedral and f is above bounded on Q, then the maximum of f on Q is achieved: $\mathop{\mathrm{Argal}}_Q f \neq \emptyset$.*

Proof. Let us start with (i). We shall prove this statement by induction on the dimension of Q. The base $\dim Q = 0$, i.e., the case of a singleton Q, is trivial, since here $Q = \mathrm{Ext}Q = \mathop{\mathrm{Argal}}_Q f$. Now assume that the statement is valid for the case of $\dim Q \leq p$, and let us prove that it is valid also for the case of $\dim Q = p+1$. Let us first verify that the set $\mathop{\mathrm{Argal}}_Q f$ intersects with the (relative) boundary of Q. Indeed, let $x \in \mathop{\mathrm{Argal}}_Q f$. There is nothing to prove if x itself is a relative boundary point of Q; and if x is not a boundary point, then, by Theorem C.5.3, f is constant on Q, so that $\mathop{\mathrm{Argal}}_Q f = Q$; and since Q is closed, every relative boundary point of Q (such a point does exist, since Q does not contain lines and is of positive dimension) is a maximizer of f on Q, so that here again $\mathop{\mathrm{Argal}}_Q f$ intersects $\partial_{\mathrm{ri}}\, Q$.

Thus, among the maximizers of f there exists at least one, let it be x, which belongs to the relative boundary of Q. Let H be the hyperplane which supports Q at x (see Section B.2.8), and let $Q' = Q \cap H$. The set Q' is closed and convex (since Q and H are), nonempty (it contains x) and does not contain lines (since Q does not). We have $\max_Q f = f(x) = \max_{Q'} f$ (note that $Q' \subset Q$), whence

$$\emptyset \neq \mathop{\mathrm{Argal}}_{Q'} f \subset \mathop{\mathrm{Argal}}_Q f.$$

Same as in the proof of the Krein-Milman Theorem (Theorem B.2.10), we have dim $Q' <$ dim Q. In view of this inequality we can apply to f and Q' our inductive hypothesis to get

$$\mathrm{Ext}(Q') \cap \underset{Q'}{\mathrm{Argal}}\, f \neq \emptyset.$$

Since $\mathrm{Ext}(Q') \subset \mathrm{Ext}(Q)$ by Lemma B.2.4 and, as we just have seen, $\underset{Q'}{\mathrm{Argal}}\, f \subset \underset{Q}{\mathrm{Argal}}\, f$, we conclude that the set $\mathrm{Ext}(Q) \cap \underset{Q}{\mathrm{Argal}}\, f$ is not smaller than $\mathrm{Ext}(Q') \cap \underset{Q'}{\mathrm{Argal}}\, f$ and is therefore nonempty, as required. □

To prove (ii), let us use the known to us from Lecture 4 results on the structure of a polyhedral convex set:

$$Q = \mathrm{Conv}(V) + \mathrm{Cone}\,(R),$$

where V and R are finite sets. We are about to prove that the upper bound of f on Q is exactly the maximum of f on the finite set V:

$$\forall x \in Q: \quad f(x) \leq \max_{v \in V} f(v). \tag{C.5.7}$$

This will mean, in particular, that f attains its maximum on Q – e.g., at the point of V where f attains its maximum on V.

To prove the announced statement, we first claim that if f is above bounded on Q, then every direction $r \in \mathrm{Cone}\,(R)$ is *descent* for f, i.e., is such that every step in this direction taken from every point $x \in Q$ decreases f:

$$f(x + tr) \leq f(x) \quad \forall x \in Q \forall t \geq 0. \tag{C.5.8}$$

Indeed, if, on contrary, there were $x \in Q$, $r \in R$ and $t \geq 0$ such that $f(x + tr) > f(x)$, we would have $t > 0$ and, by Lemma C.3.1,

$$f(x + sr) \geq f(x) + \frac{s}{t}(f(x + tr) - f(x)), \;\; s \geq t.$$

Since $x \in Q$ and $r \in \mathrm{Cone}\,(R)$, $x + sr \in Q$ for all $s \geq 0$, and since f is above bounded on Q, the left hand side in the latter inequality is above bounded, while the right hand one, due to $f(x + tr) > f(x)$, goes to $+\infty$ as $s \to \infty$, which is the desired contradiction.

Now we are done: to prove (C.5.7), note that a generic point $x \in Q$ can be represented as

$$x = \sum_{v \in V} \lambda_v v + r \quad [r \in \mathrm{Cone}\,(R); \sum_v \lambda_v = 1, \lambda_v \geq 0],$$

and we have

$$\begin{array}{rcll} f(x) & = & f(\sum\limits_{v \in V} \lambda_v v + r) & \\ & \leq & f(\sum\limits_{v \in V} \lambda_v v) & [\text{by (C.5.8)}] \\ & \leq & \sum\limits_{v \in V} \lambda_v f(v) & [\text{Jensen's Inequality}] \\ & \leq & \max\limits_{v \in V} f(v) & \quad\quad \Box \end{array}$$

C.6 Subgradients and Legendre transformation

C.6.1 Proper functions and their representation

According to one of two equivalent definitions, a convex function f on $\mathbf{R}^n$ is a function taking values in $\mathbf{R} \cup \{+\infty\}$ such that the epigraph

$$\mathrm{Epic}(f) = \{(t, x) \in \mathbf{R}^{n+1} : t \geq f(x)\}$$

is a nonempty convex set. Thus, there is no essential difference between convex functions and convex sets: convex function generates a convex set – its epigraph – which of course remembers everything about the function. And the only specific property of the epigraph as a convex set is that it has a recessive direction – namely, $e = (1, 0)$ – such that the intersection of the epigraph with every line directed by h is either empty, or is a closed ray. Whenever a nonempty convex set possesses such a property with respect to certain direction, it can be represented, in properly chosen coordinates, as the epigraph of some convex function.

Thus, a convex function is, basically, nothing but a way to look, in the literal meaning of the latter verb, at a convex set.

Now, we know that "actually good" convex sets are closed ones: they possess a lot of important properties (e.g., admit a good outer description) which are not shared by arbitrary convex sets. It means that among convex functions there also are "actually good" ones – those with closed epigraphs. Closedness of the epigraph can be "translated" to the functional language and there becomes a special kind of continuity – *lower semicontinuity:*

Definition C.6.1 [Lower semicontinuity] *Let f be a function (not necessarily convex) defined on $\mathbf{R}^n$ and taking values in $\mathbf{R}\cup\{+\infty\}$. We say that f is lower semicontinuous at a point $\bar{x}$, if for every sequence of points $\{x_i\}$ converging to $\bar{x}$ one has*

$$f(\bar{x}) \leq \lim\inf_{i\to\infty} f(x_i)$$

(here, of course, $\lim\inf$ *of a sequence with all terms equal to $+\infty$ is $+\infty$).*

f is called lower semicontinuous, if it is lower semicontinuous at every point.

A trivial example of a lower semicontinuous function is a continuous one. Note, however, that a semicontinuous function is not obliged to be continuous; what it is obliged, is to make only "jumps down". E.g., the function

$$f(x) = \left\{\begin{array}{ll} 0, & x \neq 0 \\ a, & x = 0 \end{array}\right.$$

is lower semicontinuous if $a \leq 0$ ("jump down at $x = 0$ or no jump at all"), and is not lower semicontinuous if $a > 0$ ("jump up").

The following statement links lower semicontinuity with the geometry of the epigraph:

Proposition C.6.1 *A function f defined on $\mathbf{R}^n$ and taking values from $\mathbf{R}\cup\{+\infty\}$ is lower semicontinuous iff its epigraph is closed (e.g., due to its emptiness).*

We shall not prove this statement, same as most of other statements in this Section; the reader definitely is able to restore (very simple) proofs we are skipping.

An immediate consequence of the latter proposition is as follows:

Corollary C.6.1 *The upper bound*

$$f(x) = \sup_{\alpha\in\mathcal{A}} f_\alpha(x)$$

of arbitrary family of lower semicontinuous functions is lower semicontinuous.

[from now till the end of this Section, if the opposite is not explicitly stated, "a function" means "a function defined on the entire $\mathbf{R}^n$ and taking values in $\mathbf{R}\cup\{+\infty\}$"]

Indeed, the epigraph of the upper bound is the intersection of the epigraphs of the functions forming the bound, and the intersection of closed sets always is closed.

Now let us look at *convex* lower semicontinuous functions; according to our general convention, "convex" means "satisfying the convexity inequality and finite at least at one point", or, which is the same, "with convex nonempty epigraph"; and as we just have seen, "lower semicontinuous" means "with closed epigraph". Thus, we are interested in functions with closed convex nonempty epigraphs; to save words, let us call these functions proper.

What we are about to do is to translate to the functional language several constructions and results related to convex sets. In the usual life, a translation (e.g. of poetry) typically results in something less rich than the original; in contrast to this, in mathematics this is a powerful source of new ideas and constructions.

"Outer description" of a proper function. We know that a closed convex set is intersection of closed half-spaces. What does this fact imply when the set is the epigraph of a proper function f? First of all, note that the epigraph is not a completely arbitrary convex set: it has a recessive direction $e = (1, 0)$ – the basic orth of the t-axis in the space of variables $t \in \mathbf{R}, x \in \mathbf{R}^n$ where the epigraph lives. This direction, of course, should be recessive for every closed half-space

$$(*) \quad \Pi = \{(t, x) : \alpha t \geq d^T x - a\} \quad [|\alpha| + |d| > 0]$$

containing Epic(f) (note that what is written in the right hand side of the latter relation, is one of many universal forms of writing down a general nonstrict linear inequality in the space where the epigraph lives;

this is the form the most convenient for us now). Thus, e should be a recessive direction of $\Pi \supset \mathrm{Epic}(f)$; as it is immediately seen, recessivity of e for Π means exactly that $\alpha \geq 0$. Thus, speaking about closed half-spaces containing $\mathrm{Epic}(f)$, we in fact are considering some of the half-spaces (*) with $\alpha \geq 0$.

Now, there are two essentially different possibilities for α to be nonnegative – (A) to be positive, and (B) to be zero. In the case of (B) the boundary hyperplane of Π is "vertical" – it is parallel to e, and in fact it "bounds" only x – Π is comprised of all pairs (t, x) with x belonging to certain half-space in the x-subspace and t being arbitrary real. These "vertical" subspaces will be of no interest for us.

The half-spaces which indeed are of interest for us are the "nonvertical" ones: those given by the case (A), i.e., with $\alpha > 0$. For a non-vertical half-space Π, we always can divide the inequality defining Π by α and to make $\alpha = 1$. Thus, a "nonvertical" candidate to the role of a closed half-space containing $\mathrm{Epic}(f)$ always can be written down as

$$(**) \quad \Pi = \{(t, x) : t \geq d^T x - a\},$$

i.e., *can be represented as the epigraph of an affine function of x.*

Now, when such a candidate indeed is a half-space containing $\mathrm{Epic}(f)$? The answer is clear: it is the case *iff the affine function $d^T x - a$ everywhere in $\mathbf{R}^n$ is $\leq f(\cdot)$* – as we shall say, "is an affine minorant of f"; indeed, the smaller is the epigraph, the larger is the function. *If we knew that* $\mathrm{Epic}(f)$ – which definitely is the intersection of all closed half-spaces containing $\mathrm{Epic}(f)$ – *is in fact the intersection of already nonvertical closed half-spaces containing* $\mathrm{Epic}(f)$, or, which is the same, *the intersection of the epigraphs of all affine minorants of f*, we would be able to get a nice and nontrivial result:

(!) *a proper convex function is the upper bound of affine functions – all its affine minorants.*

(indeed, we already know that it is the same – to say that a function is an upper bound of certain family of functions, and to say that the epigraph of the function is the intersection of the epigraphs of the functions of the family).

(!) indeed is true:

Proposition C.6.2 *A proper convex function f is the upper bound of all its affine minorants. Moreover, at every point $\bar{x} \in \mathrm{ri}\,\mathrm{Dom}\, f$ from the relative interior of the domain f f is even not the upper bound, but simply the maximum of its minorants: there exists an affine function $f_{\bar{x}}(x)$ which is $\leq f(x)$ everywhere in $\mathbf{R}^n$ and is equal to f at $x = \bar{x}$.*

Proof. I. We start with the "Moreover" part of the statement; this is the key to the entire statement. Thus, we are about to prove that if $\bar{x} \in \mathrm{ri}\,\mathrm{Dom}\, f$, then there exists an affine function $f_{\bar{x}}(x)$ which is everywhere $\leq f(x)$, and at $x = \bar{x}$ the inequality becomes an equality.

I.1^0 First of all, we easily can reduce the situation to the one when $\mathrm{Dom}\, f$ is full-dimensional. Indeed, by shifting f we may make the affine span $\mathrm{Aff}(\mathrm{Dom}\, f)$ of the domain of f to be a linear subspace L in $\mathbf{R}^n$; restricting f onto this linear subspace, we clearly get a proper function on L. If we believe that our statement is true for the case when the domain of f is full-dimensional, we can conclude that there exists an affine function

$$d^T x - a \quad [x \in L]$$

on L ($d \in L$) such that

$$f(x) \geq d^T x - a \quad \forall x \in L; f(\bar{x}) = d^T \bar{x} - a.$$

The affine function we get clearly can be extended, by the same formula, from L on the entire $\mathbf{R}^n$ and is a minorant of f on the entire $\mathbf{R}^n$ – outside of $L \supset \mathrm{Dom}\, f$ f simply is $+\infty$! This minorant on $\mathbf{R}^n$ is exactly what we need.

I.2^0. Now let us prove that our statement is valid when $\mathrm{Dom}\, f$ is full-dimensional, so that $\bar{x}$ is an interior point of the domain of f. Let us look at the point $y = (f(\bar{x}), \bar{x})$. This is a point from the epigraph of f, and we claim that it is a point from the relative boundary of the epigraph. Indeed, if y were a relative interior point of $\mathrm{Epic}(f)$, then, taking $y' = y + e$, we would get a segment $[y', y]$ contained in $\mathrm{Epic}(f)$; since the endpoint y of the segment is assumed to be relative interior for $\mathrm{Epic}(f)$, we could extend this segment a little through this endpoint, not leaving $\mathrm{Epic}(f)$; but this clearly is impossible, since the t-coordinate of the new endpoint would be $< f(\bar{x})$, and the x-component of it still would be $\bar{x}$.

Thus, y is a point from the relative boundary of $\mathrm{Epic}(f)$. Now we claim that y' is an interior point of $\mathrm{Epic}(f)$. This is immediate: we know from Theorem C.4.1 that f is continuous at $\bar{x}$, so that there exists a neighbourhood U of $\bar{x}$ in $\mathrm{Aff}(\mathrm{Dom}\, f) = \mathbf{R}^n$ such that $f(x) \leq f(\bar{x} + 0.5)$ whenever $x \in U$, or, in other words, the set

$$V = \{(t, x) : x \in U, t > f(\bar{x}) + 0.5\}$$

is contained in Epic(f); but this set clearly contains a neighbourhood of y' in $\mathbf{R}^{n+1}$.

Now let us look at the supporting linear form to Epic(f) at the point y of the relative boundary of Epic(f). This form gives us a linear inequality on $\mathbf{R}^{n+1}$ which is satisfied everywhere on Epic(f) and becomes equality at y; besides this, the inequality is not equality identically on Epic(f), it is strict somewhere on Epic(f). Without loss of generality we may assume that the inequality is of the form

$$(+) \quad \alpha t \geq d^T x - a.$$

Now, since our inequality is satisfied at $y' = y + e$ and becomes equality at $(t, x) = y$, α should be ≥ 0; it cannot be 0, since in the latter case the inequality in question would be equality also at $y' \in \operatorname{int} \mathrm{Epic}(f)$. But a linear inequality which is satisfied at a convex set and is *equality* at an *interior* point of the set is trivial – coming from the zero linear form (this is exactly the statement that a linear form attaining its minimum on a convex set at a point from the relative interior of the set is constant on the set and on its affine hull).

Thus, inequality (+) which is satisfied on Epic(f) and becomes equality at y is an inequality with $\alpha > 0$. Let us divide both sides of the inequality by α; we get a new inequality of the form

$$(\&) \quad t \geq d^T x - a$$

(we keep the same notation for the right hand side coefficients – we never will come back to the old coefficients); this inequality is valid on Epic(f) and is equality at $y = (f(\bar{x}), \bar{x})$. Since the inequality is valid on Epic(f), it is valid at every pair (t, x) with $x \in \operatorname{Dom} f$ and $t = f(x)$:

$$(\#) \quad f(x) \geq d^T x - a \quad \forall x \in \operatorname{Dom} f;$$

so that the right hand side is an affine minorant of f on $\operatorname{Dom} f$ and therefore – on $\mathbf{R}^n$ ($f = +\infty$ outside $\operatorname{Dom} f$!). It remains to note that (#) is equality at $\bar{x}$, since (&) is equality at y. □

II. We have proved that if $\mathcal{F}$ if the set of all affine functions which are minorants of f, then the function

$$\bar{f}(x) = \sup_{\phi \in \mathcal{F}} \phi(x)$$

is equal to f on ri Dom f (and at x from the latter set in fact sup in the right hand side can be replaced with max); to complete the proof of the Proposition, we should prove that $\bar{f}$ is equal to f also outside ri Dom f.

II.1^0. Let us first prove that $\bar{f}$ is equal to f outside $\operatorname{cl} \operatorname{Dom} f$, or. which is the same, prove that $\bar{f}(x) = +\infty$ outside $\operatorname{cl} \operatorname{Dom} f$. This is easy: is $\bar{x}$ is a point outside $\operatorname{cl} \operatorname{Dom} f$, it can be strongly separated from $\operatorname{Dom} f$, see Separation Theorem (ii) (Theorem B.2.9). Thus, there exists $z \in \mathbf{R}^n$ such that

$$z^T \bar{x} \geq z^T x + \zeta \quad \forall x \in \operatorname{Dom} f \quad [\zeta > 0]. \tag{C.6.1}$$

Besides this, we already know that there exists at least one affine minorant of f, or, which is the same, there exist a and d such that

$$f(x) \geq d^T x - a \quad \forall x \in \operatorname{Dom} f. \tag{C.6.2}$$

Let us add to (C.6.2) inequality (C.6.1) multiplied by positive weight λ; we get

$$f(x) \geq \phi_\lambda(x) \equiv (d + \lambda z)^T x + [\lambda \zeta - a - \lambda z^T \bar{x}] \quad \forall x \in \operatorname{Dom} f.$$

This inequality clearly says that $\phi_\lambda(\cdot)$ is an affine minorant of f on $\mathbf{R}^n$ for every $\lambda > 0$. The value of this minorant at $x = \bar{x}$ is equal to $d^T \bar{x} - a + \lambda \zeta$ and therefore it goes to $+\infty$ as $\lambda \to +\infty$. We see that the upper bound of affine minorants of f at $\bar{x}$ indeed is $+\infty$, as claimed.

II.2^0. Thus, we know that the upper bound $\bar{f}$ of all affine minorants of f is equal to f everywhere on the relative interior of $\operatorname{Dom} f$ and everywhere outside the closure of $\operatorname{Dom} f$; all we should prove that this equality is also valid at the points of the relative boundary of $\operatorname{Dom} f$. Let $\bar{x}$ be such a point. There is nothing to prove if $\bar{f}(\bar{x}) = +\infty$, since by construction $\bar{f}$ is everywhere $\leq f$. Thus, we should prove that if $\bar{f}(\bar{x}) = c < \infty$, then $f(\bar{x}) = c$. Since $\bar{f} \leq f$ everywhere, to prove that $f(\bar{x}) = c$ is the same as to prove that $f(\bar{x}) \leq c$. This is immediately given by lower semicontinuity of f: let us choose $x' \in$ ri Dom f and look what happens along a sequence of points $x_i \in [x', \bar{x})$ converging to $\bar{x}$. All the points of this sequence are relative interior points of $\operatorname{Dom} f$ (Lemma B.1.1), and consequently

$$f(x_i) = \bar{f}(x_i).$$

Now, $x_i = (1-\lambda_i)\bar{x} + \lambda_i x'$ with $\lambda_i \to +0$ as $i \to \infty$; since $\bar{f}$ clearly is convex (as the upper bound of a family of affine and therefore convex functions), we have

$$\bar{f}(x_i) \le (1-\lambda_i)\bar{f}(\bar{x}) + \lambda_i \bar{f}(x').$$

Putting things together, we get

$$f(x_i) \le (1-\lambda_i)\bar{f}(\bar{x}) + \lambda_i f(x');$$

as $i \to \infty$, $x_i \to \bar{x}$, and the right hand side in our inequality converges to $\bar{f}(\bar{x}) = c$; since f is lower semicontinuous, we get $f(\bar{x}) \le c$. □

We see why "translation of mathematical facts from one mathematical language to another" – in our case, from the language of convex sets to the language of convex functions – may be fruitful: because we invest a lot into the process rather than run it mechanically.

Closure of a convex function. We got a nice result on the "outer description" of a *proper* convex function: it is the upper bound of a family of affine functions. Note that, vice versa, the upper bound of every family of affine functions is a proper function, provided that this upper bound is finite at least at one point (indeed, as we know from Section C.2.1, upper bound of every family of convex functions is convex, provided that it is finite at least at one point; and Corollary C.6.1 says that upper bound of lower semicontinuous functions (e.g., affine ones – they are even continuous) is lower semicontinuous).

Now, what to do with a convex function which is not lower semicontinuous? The similar question about convex sets – what to do with a convex set which is not closed – can be resolved very simply: we can pass from the set to its closure and thus get a "normal" object which is very "close" to the original one: the "main part" of the original set – its relative interior – remains unchanged, and the "correction" adds to the set something relatively small – the relative boundary. The same approach works for convex functions: if a convex function f is not proper (i.e., its epigraph, being convex and nonempty, is not closed), we can "correct" the function – replace it with a new function with the epigraph being the closure of $\mathrm{Epic}(f)$. To justify this approach, we, of course, should be sure that the closure of the epigraph of a convex function is also an epigraph of such a function. This indeed is the case, and to see it, it suffices to note that a set G in $\mathbf{R}^{n+1}$ is the epigraph of a function taking values in $\mathbf{R} \cup \{+\infty\}$ iff the intersection of G with every vertical line $\{x = \mathrm{const}, t \in \mathbf{R}\}$ is either empty, or is a closed ray of the form $\{x = \mathrm{const}, t \ge \bar{t} > -\infty\}$. Now, it is absolutely evident that if G is the closure of the epigraph of a function f, that its intersection with a vertical line is either empty, or is a closed ray, or is the entire line (the last case indeed can take place – look at the closure of the epigraph of the function equal to $-\frac{1}{x}$ for $x > 0$ and $+\infty$ for $x \le 0$). We see that in order to justify our idea of "proper correction" of a convex function we should prove that if f is convex, then the last of the indicated three cases – the intersection of $\mathrm{cl}\,\mathrm{Epic}(f)$ with a vertical line is the entire line – never occurs. This fact evidently is a corollary of the following simple

Proposition C.6.3 *A convex function is below bounded on every bounded subset of* $\mathbf{R}^n$.

Proof. Without loss of generality we may assume that the domain of the function f is full-dimensional and that 0 is the interior point of the domain. According to Theorem C.4.1, there exists a neighbourhood U of the origin – which can be thought of to be a centered at the origin ball of some radius $r > 0$ – where f is bounded from above by some C. Now, if $R > 0$ is arbitrary and x is an arbitrary point with $|x| \le R$, then the point

$$y = -\frac{r}{R}x$$

belongs to U, and we have

$$0 = \frac{r}{r+R}x + \frac{R}{r+R}y;$$

since f is convex, we conclude that

$$f(0) \le \frac{r}{r+R}f(x) + \frac{R}{r+R}f(y) \le \frac{r}{r+R}f(x) + \frac{R}{r+R}c,$$

and we get the lower bound

$$f(x) \ge \frac{r+R}{r}f(0) - \frac{r}{R}c$$

for the values of f in the centered at 0 ball of radius R. □

Thus, we conclude that the closure of the epigraph of a convex function f is the epigraph of certain function, let it be called *the closure* $\mathrm{cl}\, f$ *of* f. Of course, this latter function is convex (its epigraph is convex – it is the closure of a convex set), and since its epigraph is closed, $\mathrm{cl}\, f$ is proper. The following statement gives direct description of $\mathrm{cl}\, f$ in terms of f:

Proposition C.6.4 *Let f be a convex function and $\mathrm{cl}\, f$ be its closure. Then*

(i) *For every x one has*

$$\mathrm{cl}\, f(x) = \lim_{r\to+0} \inf_{x':\|x'-x\|_2\le r} f(x').$$

In particular,

$$f(x) \ge \mathrm{cl}\, f(x)$$

for all x, and

$$f(x) = \mathrm{cl}\, f(x)$$

whenever $x \in \mathrm{ri\ Dom}\, f$, same as whenever $x \notin \mathrm{cl\,Dom}\, f$.
Thus, the "correction" $f \mapsto \mathrm{cl}\, f$ may vary f only at the points from the relative boundary of $\mathrm{Dom}\, f$*,*

$$\mathrm{Dom}\, f \subset \mathrm{Dom\,cl}\, f \subset \mathrm{cl\,Dom}\, f,$$

whence also

$$\mathrm{ri\ Dom}\, f = \mathrm{ri\ Dom\,cl}\, f.$$

(ii) *The family of affine minorants of* $\mathrm{cl}\, f$ *is exactly the family of affine minorants of f, so that*

$$\mathrm{cl}\, f(x) = \sup\{\phi(x) : \phi \textit{ is an affine minorant of } f\},$$

and the sup *in the right hand side can be replaced with* max *whenever $x \in \mathrm{ri\ Dom\,cl}\, f = \mathrm{ri\ Dom}\, f$.*
["so that" comes from the fact that $\mathrm{cl}\, f$ is proper and is therefore the upper bound of its affine minorants]

C.6.2 Subgradients

Let f be a convex function, and let $x \in \mathrm{Dom}\, f$. It may happen that there exists an affine minorant $d^T x - a$ of f which coincides with f at x:

$$f(y) \ge d^T y - a \quad \forall y, \quad f(x) = d^T x - a.$$

From the equality in the latter relation we get $a = d^T x - f(x)$, and substituting this representation of a into the first inequality, we get

$$f(y) \ge f(x) + d^T(y - x) \quad \forall y. \tag{C.6.3}$$

Thus, if f admits an affine minorant which is exact at x, then there exists d which gives rise to inequality (C.6.3). Vice versa, if d is such that (C.6.3) takes place, then the right hand side of (C.6.3), regarded as a function of y, is an affine minorant of f which is exact at x.

Now note that (C.6.3) expresses certain property of a vector d. A vector satisfying, for a given x, this property – i.e., the slope of an exact at x affine minorant of f – is called a *subgradient* of f at x, and the set of all subgradients of f at x is denoted $\partial f(x)$.

Subgradients of convex functions play important role in the theory and numerical methods of Convex Programming – they are quite reasonable surrogates of gradients. The most elementary properties of the subgradients are summarized in the following statement:

Proposition C.6.5 *Let f be a convex function and x be a point from* $\mathrm{Dom}\, f$*. Then*

(i) *$\partial f(x)$ is a closed convex set which for sure is nonempty when $x \in \mathrm{ri\ Dom}\, f$*

(ii) *If $x \in \mathrm{int\,Dom}\, f$ and f is differentiable at x, then $\partial f(x)$ is the singleton comprised of the usual gradient of f at x.*

Proof. (i): Closedness and convexity of $\partial f(x)$ are evident – (C.6.3) is an infinite system of nonstrict linear inequalities with respect to d, the inequalities being indexed by $y \in \mathbf{R}^n$. Nonemptiness of $\partial f(x)$ for the case when $x \in \mathrm{ri\ Dom}\, f$ – this is the most important fact about the subgradients – is readily given by our preceding results. Indeed, we should prove that if $x \in \mathrm{ri\ Dom}\, f$, then there exists an affine minorant of f which is exact at x. But this is an immediate consequence of Proposition C.6.4: part (i) of the proposition says that there exists an affine minorant of f which is equal to $\mathrm{cl}\, f(x)$ at the point x, and part (i) says that $f(x) = \mathrm{cl}\, f(x)$.

(ii): If $x \in \mathrm{int}\, \mathrm{Dom}\, f$ and f is differentiable at x, then $\nabla f(x) \in \partial f(x)$ by the Gradient Inequality. To prove that in the case in question $\nabla f(x)$ is the only subgradient of f at x, note that if $d \in \partial f(x)$, then, by definition,

$$f(y) - f(x) \geq d^T(y - x) \quad \forall y$$

Substituting $y - x = th$, h being a fixed direction and t being > 0, dividing both sides of the resulting inequality by t and passing to limit as $t \to +0$, we get

$$h^T \nabla f(x) \geq h^T d.$$

This inequality should be valid for all h, which is possible if and only if $d = \nabla f(x)$. □

Proposition C.6.5 explains why subgradients are good surrogates of gradients: at a point where gradient exists, it is the only subgradient, but, in contrast to the gradient, a subgradient exists basically everywhere (for sure in the relative interior of the domain of the function). E.g., let us look at the simple function

$$f(x) = |x|$$

on the axis. It is, of course, convex (as maximum of two linear forms x and $-x$). Whenever $x \neq 0$, f is differentiable at x with the derivative $+1$ for $x > 0$ and -1 for $x < 0$. At the point $x = 0$ f is not differentiable; nevertheless, it must have subgradients at this point (since 0 is an interior point of the domain of the function). And indeed, it is immediately seen that the subgradients of $|x|$ at $x = 0$ are exactly the reals from the segment $[-1, 1]$. Thus,

$$\partial |x| = \left\{ \begin{array}{ll} \{-1\}, & x < 0 \\ [-1,1], & x = 0 \\ \{+1\}, & x > 0 \end{array} \right. .$$

Note also that if x is a relative boundary point of the domain of a convex function, even a "good" one, the set of subgradients of f at x may be empty, as it is the case with the function

$$f(y) = \left\{ \begin{array}{ll} -\sqrt{y}, & y \geq 0 \\ +\infty, & y < 0 \end{array} \right. ;$$

it is clear that there is no non-vertical supporting line to the epigraph of the function at the point $(0, f(0))$, and, consequently, there is no affine minorant of the function which is exact at $x = 0$.

A significant – and important – part of Convex Analysis deals with *subgradient calculus* – with the rules for computing subgradients of "composite" functions, like sums, superpositions, maxima, etc., given subgradients of the operands. These rules extend onto nonsmooth convex case the standard Calculus rules and are very nice and instructive; the related considerations, however, are beyond our scope.

C.6.3 Legendre transformation

Let f be a convex function. We know that f "basically" is the upper bound of all its affine minorants; this is exactly the case when f is proper, otherwise the corresponding equality takes place everywhere except, perhaps, some points from the relative boundary of $\mathrm{Dom}\, f$. Now, when an affine function $d^T x - a$ is an affine minorant of f? It is the case iff

$$f(x) \geq d^T x - a$$

for all x or, which is the same, iff

$$a \geq d^T x - f(x)$$

for all x. We see that if the slope d of an affine function $d^T x - a$ is fixed, then in order for the function to be a minorant of f we should have

$$a \geq \sup_{x \in \mathbf{R}^n} [d^T x - f(x)].$$

The supremum in the right hand side of the latter relation is certain function of d; this function is called the *Legendre transformation* of f and is denoted f^*:

$$f^*(d) = \sup_{x \in \mathbf{R}^n} [d^T x - f(x)].$$

Geometrically, the Legendre transformation answers the following question: given a slope d of an affine function, i.e., given the hyperplane $t = d^T x$ in $\mathbf{R}^{n+1}$, what is the minimal "shift down" of the hyperplane which places it below the graph of f?

From the definition of the Legendre transformation it follows that this is a proper function. Indeed, we loose nothing when replacing $\sup\limits_{x\in\mathbf{R}^n}[d^Tx - f(x)]$ by $\sup\limits_{x\in\mathrm{Dom}\, f}[d^Tx - f(x)]$, so that the Legendre transformation is the upper bound of a family of affine functions. Since this bound is finite at least at one point (namely, at every d coming form affine minorant of f; we know that such a minorant exists), it is a convex lower semicontinuous function, as claimed.

The most elementary (and the most fundamental) fact about the Legendre transformation is its symmetry:

Proposition C.6.6 *Let f be a convex function. Then twice taken Legendre transformation of f is the closure $\mathrm{cl}\, f$ of f:*

$$(f^*)^* = \mathrm{cl}\, f.$$

In particular, if f is proper, then it is the Legendre transformation of its Legendre transformation (which also is proper).

Proof is immediate. The Legendre transformation of f^* at the point x is, by definition,

$$\sup_{d\in\mathbf{R}^n}[x^Td - f^*(d)] = \sup_{d\in\mathbf{R}^n, a\geq f^*(d)}[d^Tx - a];$$

the second sup here is exactly the supremum of all affine minorants of f (this is the origin of the Legendre transformation: $a \geq f^*(d)$ iff the affine form $d^Tx - a$ is a minorant of f). And we already know that the upper bound of all affine minorants of f is the closure of f. □

The Legendre transformation is a very powerful tool – this is a "global" transformation, so that *local* properties of f^* correspond to *global* properties of f. E.g.,

- $d = 0$ belongs to the domain of f^* iff f is below bounded, and if it is the case, then $f^*(0) = -\inf f$;
- if f is proper, then the subgradients of f^* at $d = 0$ are exactly the minimizers of f on $\mathbf{R}^n$;
- $\mathrm{Dom}\, f^*$ is the entire $\mathbf{R}^n$ iff $f(x)$ grows, as $\|x\|_2 \to \infty$, faster than $\|x\|_2$: there exists a function $r(t) \to \infty$, as $t \to \infty$ such that
$$f(x) \geq r(\|x\|_2) \quad \forall x,$$

etc. Thus, whenever we can compute explicitly the Legendre transformation of f, we get a lot of "global" information on f. Unfortunately, the more detailed investigation of the properties of Legendre transformation is beyond our scope; I simply list several simple facts and examples:

- From the definition of Legendre transformation,
$$f(x) + f^*(d) \geq x^Td \quad \forall x, d.$$
Specifying here f and f^*, we get certain inequality, e.g., the following one:

  [Young's Inequality] *if p and q are positive reals such that $\frac{1}{p} + \frac{1}{q} = 1$, then*
$$\frac{|x|^p}{p} + \frac{|d|^q}{q} \geq xd \quad \forall x, d \in \mathbf{R}$$
(indeed, as it is immediately seen, the Legendre transformation of the function $|x|^p/p$ is $|d|^q/q$)

  Consequences. Very simple-looking Young's inequality gives rise to a very nice and useful *Hölder inequality:*

  Let $1 \leq p \leq \infty$ and let q be such $\frac{1}{p} + \frac{1}{q} = 1$ ($p = 1 \Rightarrow q = \infty$, $p = \infty \Rightarrow q = 1$). For every two vectors $x, y \in \mathbf{R}^n$ one has
$$\sum_{i=1}^n |x_i y_i| \leq \|x\|_p \|y\|_q \tag{C.6.4}$$

Indeed, there is nothing to prove if p or q is ∞ – if it is the case, the inequality becomes the evident relation

$$\sum_i |x_i y_i| \le (\max_i |x_i|)(\sum_i |y_i|).$$

Now let $1 < p < \infty$, so that also $1 < q < \infty$. In this case we should prove that

$$\sum_i |x_i y_i| \le (\sum_i |x_i|^p)^{1/p}(\sum_i |y_i|^q)^{1/q}.$$

There is nothing to prove if one of the factors in the right hand side vanishes; thus, we can assume that $x \neq 0$ and $y \neq 0$. Now, both sides of the inequality are of homogeneity degree 1 with respect to x (when we multiply x by t, both sides are multiplied by $|t|$), and similarly with respect to y. Multiplying x and y by appropriate reals, we can make both factors in the right hand side equal to 1: $\|x\|_p = \|y\|_p = 1$. Now we should prove that under this normalization the left hand side in the inequality is ≤ 1, which is immediately given by the Young inequality:

$$\sum_i |x_i y_i| \le \sum_i [|x_i|^p/p + |y_i|^q/q] = 1/p + 1/q = 1.$$

Note that the Hölder inequality says that

$$|x^T y| \le \|x\|_p \|y\|_q; \tag{C.6.5}$$

when $p = q = 2$, we get the Cauchy inequality. Now, inequality (C.6.5) is *exact* in the sense that for every x there exists y with $\|y\|_q = 1$ such that

$$x^T y = \|x\|_p \quad [= \|x\|_p \|y\|_q];$$

it suffices to take

$$y_i = \|x\|_p^{1-p} |x_i|^{p-1} \mathrm{sign}(x_i)$$

(here $x \neq 0$; the case of $x = 0$ is trivial – here y can be an arbitrary vector with $\|y\|_q = 1$).

Combining our observations, we come to an extremely important, although simple, fact:

$$\|x\|_p = \max\{y^T x : \|y\|_q \le 1\} \quad [\frac{1}{p} + \frac{1}{q} = 1]. \tag{C.6.6}$$

It follows, in particular, that $\|x\|_p$ is convex (as an upper bound of a family of linear forms), whence

$$\|x' + x''\|_p = 2\|\frac{1}{2}x' + \frac{1}{2}x''\|_p \le 2(\|x'\|_p/2 + \|x''\|_p/2) = \|x'\|_p + \|x''\|_p;$$

this is nothing but the triangle inequality. Thus, $\|x\|_p$ satisfies the triangle inequality; it clearly possesses two other characteristic properties of a norm – positivity and homogeneity. Consequently, $\|\cdot\|_p$ is a norm – the fact that we announced twice and have finally proven now.

- The Legendre transformation of the function

$$f(x) \equiv -a$$

is the function which is equal to a at the origin and is $+\infty$ outside the origin; similarly, the Legendre transformation of an affine function $\bar{d}^T x - a$ is equal to a at $d = \bar{d}$ and is $+\infty$ when $d \neq \bar{d}$;

- The Legendre transformation of the strictly convex quadratic form

$$f(x) = \frac{1}{2}x^T A x$$

(A is positive definite symmetric matrix) is the quadratic form

$$f^*(d) = \frac{1}{2}d^T A^{-1} d$$

- The Legendre transformation of the Euclidean norm

$$f(x) = \|x\|_2$$

is the function which is equal to 0 in the closed unit ball centered at the origin and is $+\infty$ outside the ball.

The latter example is a particular case of the following statement:

Let $\|x\|$ be a norm on $\mathbf{R}^n$, and let

$$\|d\|_* = \sup\{d^T x : \|x\| \leq 1\}$$

be the conjugate to $\| \cdot \|$ norm.

Exercise C.1 *Prove that $\| \cdot \|_*$ is a norm, and that the norm conjugate to $\| \cdot \|_*$ is the original norm $\| \cdot \|$.*

Hint: Observe that the unit ball of $\| \cdot \|_*$ is exactly the polar of the unit ball of $\| \cdot \|$.

The Legendre transformation of $\|x\|$ is the characteristic function of the unit ball of the conjugate norm, i.e., is the function of d equal to 0 when $\|d\|_* \leq 1$ and is $+\infty$ otherwise.

E.g., (C.6.6) says that the norm conjugate to $\| \cdot \|_p$, $1 \leq p \leq \infty$, is $\| \cdot \|_q$, $1/p + 1/q = 1$; consequently, the Legendre transformation of p-norm is the characteristic function of the unit $\| \cdot$ _q-ball.

Appendix D

Convex Programming, Lagrange Duality, Saddle Points

D.1 Mathematical Programming Program

A (constrained) Mathematical Programming program is a problem as follows:

$$\text{(P)} \quad \min\left\{f(x) : x \in X, \quad g(x) \equiv (g_1(x), ..., g_m(x)) \leq 0, \; h(x) \equiv (h_1(x), ..., h_k(x)) = 0\right\}. \tag{D.1.1}$$

The standard terminology related to (D.1.1) is:

- [domain] X is called the *domain* of the problem
- [objective] f is called the *objective*
- [constraints] g_i, $i = 1, ..., m$, are called the (functional) *inequality constraints*; h_j, $j = 1, ..., k$, are called the *equality constraints*[1)]

In the sequel, if the opposite is not explicitly stated, it always is assumed that the objective and the constraints are well-defined on X.

- [feasible solution] a point $x \in \mathbf{R}^n$ is called a *feasible solution* to (D.1.1), if $x \in X$, $g_i(x) \leq 0$, $i = 1, ..., m$, and $h_j(x) = 0$, $j = 1, ..., k$, i.e., if x satisfies all restrictions imposed by the formulation of the problem
  - [feasible set] the set of all feasible solutions is called the *feasible set* of the problem
  - [feasible problem] a problem with a nonempty feasible set (i.e., the one which admits feasible solutions) is called *feasible* (or consistent)
  - [active constraints] an inequality constraint $g_i(\cdot) \leq 0$ is called *active at a given feasible solution* x, if this constraint is satisfied at the point as an equality rather than strict inequality, i.e., if

    $$g_i(x) = 0.$$

    A equality constraint $h_i(x) = 0$ by definition is active at every feasible solution x.

- [optimal value] the quantity

  $$f^* = \begin{cases} \inf\limits_{x \in X : g(x) \leq 0, h(x) = 0} f(x), & \text{the problem is feasible} \\ +\infty, & \text{the problem is infeasible} \end{cases}$$

  is called *the optimal value* of the problem

  - [below boundedness] the problem is called *below bounded*, if its optimal value is $> -\infty$, i.e., if the objective is below bounded on the feasible set

[1)] rigorously speaking, the constraints are not the functions g_i, h_j, but the relations $g_i(x) \leq 0$, $h_j(x) = 0$; in fact the word "constraints" is used in both these senses, and it is always clear what is meant. E.g., saying that x satisfies the constraints, we mean the relations, and saying that the constraints are differentiable, we mean the functions

- [optimal solution] a point $x \in \mathbf{R}^n$ is called an *optimal solution* to (D.1.1), if x is feasible and $f(x) \leq f(x')$ for any other feasible solution, i.e., if

$$x \in \operatorname*{Argmin}_{x' \in X: g(x') \leq 0, h(x')=0} f(x')$$

  - [solvable problem] a problem is called *solvable*, if it admits optimal solutions
  - [optimal set] the set of all optimal solutions to a problem is called its *optimal set*

To solve the problem *exactly* means to find its optimal solution or to detect that no optimal solution exists.

D.2 Convex Programming program and Lagrange Duality Theorem

A Mathematical Programming program (P) is called *convex* (or *Convex Programming* program), if

- X is a *convex* subset of $\mathbf{R}^n$
- $f, g_1, ..., g_m$ are *real-valued convex* functions on X,

  and
- there are no equality constraints at all.

Note that instead of saying that there are no equality constraints, we could say that there are constraints of this type, but only *linear* ones; this latter case can be immediately reduced to the one without equality constraints by replacing $\mathbf{R}^n$ with the affine subspace given by the (linear) equality constraints.

D.2.1 Convex Theorem on Alternative

The simplest case of a convex program is, of course, a Linear Programming program – the one where $X = \mathbf{R}^n$ and the objective and all the constraints are linear. We already know what are optimality conditions for this particular case – they are given by the Linear Programming Duality Theorem. How did we get these conditions?

We started with the observation that the fact that a point x^* is an optimal solution can be expressed in terms of solvability/unsolvability of certain systems of inequalities: in our now terms, these systems are

$$x \in G,\ f(x) \leq c,\ g_j(x) \leq 0,\ j = 1, ..., m \tag{D.2.1}$$

and

$$x \in G,\ f(x) < c,\ g_j(x) \leq 0,\ j = 1, ..., m; \tag{D.2.2}$$

here c is a parameter. Optimality of x^* for the problem means exactly that for appropriately chosen c (this choice, of course, is $c = f(x^*)$) the first of these systems is solvable and x^* is its solution, while the second system is unsolvable. Given this trivial observation, we converted the "negative" part of it – the claim that (D.2.2) is unsolvable – into a positive statement, using the General Theorem on Alternative, and this gave us the LP Duality Theorem.

Now we are going to use the same approach. What we need is a "convex analogy" to the Theorem on Alternative – something like the latter statement, but for the case when the inequalities in question are given by convex functions rather than the linear ones (and, besides it, we have a "convex inclusion" $x \in X$).

It is easy to *guess* the result we need. How did we come to the formulation of the Theorem on Alternative? The question we were interested in was, basically, how to express in an affirmative manner the fact that a system of linear inequalities has no solutions; to this end we observed that if we can combine, in a linear fashion, the inequalities of the system and get an obviously false inequality like $0 \leq -1$, then the system is unsolvable; this condition is certain affirmative statement with respect to the weights with which we are combining the original inequalities.

Now, the scheme of the above reasoning has nothing in common with linearity (and even convexity) of the inequalities in question. Indeed, consider *an arbitrary* inequality system of the type (D.2.2):

$$\text{(I)}\qquad \begin{array}{rcl} f(x) & < & c \\ g_j(x) & \leq & 0,\ j=1,...,m \\ x & \in & X; \end{array}$$

all we assume is that X is a nonempty subset in $\mathbf{R}^n$ and $f, g_1, ..., g_m$ are real-valued functions on X. It is absolutely evident that

if there exist nonnegative $\lambda_1, ..., \lambda_m$ such that the inequality

$$f(x) + \sum_{j=1}^{m} \lambda_j g_j(x) < c \tag{D.2.3}$$

has no solutions in X, then (I) *also has no solutions.*

Indeed, a solution to (I) clearly is a solution to (D.2.3) – the latter inequality is nothing but a combination of the inequalities from (I) with the weights 1 (for the first inequality) and λ_j (for the remaining ones).

Now, what does it mean that (D.2.3) has no solutions? A necessary and sufficient condition for this is that the infimum of the left hand side of (D.2.3) in $x \in X$ is $\geq c$. Thus, we come to the following evident

Proposition D.2.1 [Sufficient condition for insolvability of (I)] *Consider a system* (I) *with arbitrary data and assume that the system*

$$\text{(II)}\qquad \begin{array}{rcl} \inf\limits_{x\in X}\left[f(x) + \sum\limits_{j=1}^{m}\lambda_j g_j(x)\right] & \geq & c \\ \lambda_j & \geq & 0,\ j=1,...,m \end{array}$$

with unknowns $\lambda_1, ..., \lambda_m$ has a solution. Then (I) *is infeasible.*

Let me stress that this result is completely general; it does not require any assumptions on the entities involved.

The result we have obtained, unfortunately, does not help us: the actual power of the Theorem on Alternative (and the fact used to prove the Linear Programming Duality Theorem) is not the *sufficiency* of the condition of Proposition for infeasibility of (I), but the *necessity* of this condition. Justification of necessity of the condition in question has nothing in common with the evident reasoning which gives the sufficiency. The necessity in the linear case ($X = \mathbf{R}^n$, f, $g_1, ..., g_m$ are linear) can be established via the Homogeneous Farkas Lemma. Now we shall prove the necessity of the condition for the *convex* case, and already here we need some additional, although minor, assumptions; and in the general nonconvex case the condition in question simply is *not* necessary for infeasibility of (I) [and this is very bad – this is the reason why there exist difficult optimization problems which we do not know how to solve efficiently].

The just presented "preface" explains what we should do; now let us carry out our plan. We start with the aforementioned "minor regularity assumptions".

Definition D.2.1 [Slater Condition] *Let $X \subset \mathbf{R}^n$ and $g_1, ..., g_m$ be real-valued functions on X. We say that these functions satisfy the Slater condition on X, if there exists $x \in X$ such that $g_j(x) < 0$, $j = 1, ..., m$.*

An inequality constrained program

$$\text{(IC)}\qquad \min\{f(x) : g_j(x) \leq 0,\ j = 1, ..., m,\ x \in X\}$$

($f, g_1, ..., g_m$ are real-valued functions on X) is called to satisfy the Slater condition, if $g_1, ..., g_m$ satisfy this condition on X.

We are about to establish the following fundamental fact:

Theorem D.2.1 [Convex Theorem on Alternative]
Let $X \subset \mathbf{R}^n$ be convex, let $f, g_1, ..., g_m$ be real-valued convex functions on X, and let $g_1, ..., g_m$ satisfy the Slater condition on X. Then system (I) *is solvable iff system* (II) *is unsolvable.*

Proof. The first part of the statement – "if (II) has a solution, then (I) has no solutions" – is given by Proposition D.2.1. What we need is to prove the inverse statement. Thus, let us assume that (I) has no solutions, and let us prove that then (II) has a solution.

Without loss of generality we may assume that X is full-dimensional: $\mathrm{ri}\, X = \mathrm{int}\, X$ (indeed, otherwise we could replace our "universe" $\mathbf{R}^n$ with the affine span of X).

1^0. Let us set

$$F(x) = \left[\begin{array}{c} f(x) \\ g_1(x) \\ ... \\ g_m(x) \end{array}\right]$$

and consider two sets in $\mathbf{R}^{m+1}$:

$$S = \{u = (u_0, ..., u_m) \mid \exists x \in X : F(x) \le u\}$$

and

$$T = \{(u_0, ..., u_m) \mid u_0 < c, u_1 \le 0, u_2 \le 0, ..., u_m \le 0\}.$$

We claim that

- (i) S and T are nonempty convex sets;
- (ii) S and T does not intersect.

Indeed, convexity and nonemptiness of T is evident, same as nonemptiness of S. Convexity of S is an immediate consequence of the fact that X and $f, g_1, ..., g_m$ are convex. Indeed, assuming that $u', u'' \in S$, we conclude that there exist $x', x'' \in X$ such that $F(x') \le u'$ and $F(x'') \le u''$, whence, for every $\lambda \in [0, 1]$.

$$\lambda F(x') + (1-\lambda)F(x'') \le \lambda u' + (1-\lambda)u''.$$

The left hand side in this inequality, due to convexity of X and $f, g_1, ..., g_m$, is $\ge F(y)$, $y = \lambda x' + (1-\lambda)x''$. Thus, for the point $v = \lambda u' + (1-\lambda)u''$ there exists $y \in X$ with $F(y) \le v$, whence $v \in S$. Thus, S is convex.

The fact that $S \cap T = \emptyset$ is an evident equivalent reformulation of the fact that (I) has no solutions.

2^0. Since S and T are nonempty convex sets with empty intersection, by Separation Theorem (Theorem B.2.9) they can be separated by a linear form: there exist $a = (a_0, ..., a_m) \ne 0$ such that

$$\inf_{u \in S} \sum_{j=0}^{m} a_j u_j \ge \sup_{u \in T} \sum_{j=0}^{m} a_j u_j. \tag{D.2.4}$$

3^0. Let us look what can be said about the vector a. We claim that, first,

$$a \ge 0 \tag{D.2.5}$$

and, second,

$$a_0 > 0. \tag{D.2.6}$$

Indeed, to prove (D.2.5) note that if some a_i were negative, then the right hand side in (D.2.4) would be $+\infty$ [2], which is forbidden by (D.2.4).

Thus, $a \ge 0$; with this in mind, we can immediately compute the right hand side of (D.2.4):

$$\sup_{u \in T} \sum_{j=0}^{m} a_j u_j = \sup_{u_0 < c, u_1, ..., u_m \le 0} \sum_{j=0}^{m} a_j u_j = a_0 c.$$

Since for every $x \in X$ the point $F(x)$ belongs to S, the left hand side in (D.2.4) is not greater than

$$\inf_{x \in X} \left[a_0 f(x) + \sum_{j=1}^{m} a_j g_j(x) \right];$$

[2] look what happens when all coordinates in u, except the ith one, are fixed at values allowed by the description of T and u_i is a large in absolute value negative real

combining our observations, we conclude that (D.2.4) implies the inequality

$$\inf_{x\in X}\left[a_0 f(x)+\sum_{j=1}^m a_j g_j(x)\right]\geq a_0 c. \tag{D.2.7}$$

Now let us prove that $a_0>0$. This crucial fact is an immediate consequence of the Slater condition. Indeed, let $\bar{x}\in X$ be the point given by this condition, so that $g_j(\bar{x})<0$. From (D.2.7) we conclude that

$$a_0 f(\bar{x})+\sum_{j=0}^m a_j g_j(\bar{x})\geq a_0 c.$$

If a_0 were 0, then the right hand side of this inequality would be 0, while the left one would be the combination $\sum\limits_{j=0}^m a_j g_j(\bar{x})$ of *negative* reals $g_j(\bar{x})$ with *nonnegative* coefficients a_j *not all equal to 0* [3], so that the left hand side is strictly negative, which is the desired contradiction.

4^0. Now we are done: since $a_0>0$, we are in our right to divide both sides of (D.2.7) by a_0 and thus get

$$\inf_{x\in X}\left[f_0(x)+\sum_{j=1}^m \lambda_j g_j(x)\right]\geq c, \tag{D.2.8}$$

where $\lambda_j=a_j/a_0\geq 0$. Thus, (II) has a solution. □

D.2.2 Lagrange Function and Lagrange Duality

D.2.2.1 Lagrange function

The result of Convex Theorem on Alternative brings to our attention the function

$$\underline{L}(\lambda)=\inf_{x\in X}\left[f(x)+\sum_{j=1}^m \lambda_j g_j(x)\right], \tag{D.2.9}$$

same as the aggregate

$$L(x,\lambda)=f(x)+\sum_{j=1}^m \lambda_j g_j(x) \tag{D.2.10}$$

from which this function comes. Aggregate (D.2.10) is called the *Lagrange function* of the inequality constrained optimization program

$$\text{(IC)}\qquad \min\left\{f(x): g_j(x)\leq 0,\ j=1,...,m,\ x\in X\right\}.$$

The Lagrange function of an optimization program is a very important entity: most of optimality conditions are expressed in terms of this function. Let us start with translating of what we already know to the language of the Lagrange function.

D.2.2.2 Convex Programming Duality Theorem

Theorem D.2.2 *Consider an arbitrary inequality constrained optimization program* (IC). *Then*

(i) *The infimum*

$$\underline{L}(\lambda)=\inf_{x\in X} L(x,\lambda)$$

of the Lagrange function in $x\in X$ is, for every $\lambda\geq 0$, a lower bound on the optimal value in (IC), *so that the optimal value in the optimization program*

$$\text{(IC}^*)\qquad \sup_{\lambda\geq 0}\underline{L}(\lambda)$$

also is a lower bound for the optimal value in (IC);

(ii) [Convex Duality Theorem] *If* (IC)

[3] indeed, from the very beginning we know that $a\neq 0$, so that if $a_0=0$, then not all a_j, $j\geq 1$, are zeros

- *is convex,*
- *is below bounded*

and

- *satisfies the Slater condition,*

then the optimal value in (IC*) *is attained and is equal to the optimal value in* (IC)*.*

Proof. (i) is nothing but Proposition D.2.1 (why?). It makes sense, however, to repeat here the corresponding one-line reasoning:

> Let $\lambda \geq 0$; in order to prove that
>
> $$\underline{L}(\lambda) \equiv \inf_{x \in X} L(x,\lambda) \leq c^* \quad [L(x,\lambda) = f(x) + \sum_{j=1}^m \lambda_j g_j(x)],$$
>
> where c^* is the optimal value in (IC), note that if x is feasible for (IC), then evidently $L(x,\lambda) \leq f(x)$, so that the infimum of L over $x \in X$ is $\leq$ the infimum c^* of f over the feasible set of (IC). □

(ii) is an immediate consequence of the Convex Theorem on Alternative. Indeed, let c^* be the optimal value in (IC). Then the system

$$f(x) < c^*, g_j(x) \leq 0,\ j = 1, ..., m$$

has no solutions in X, and by the above Theorem the system (II) associated with $c = c^*$ has a solution, i.e., there exists $\lambda^* \geq 0$ such that $\underline{L}(\lambda^*) \geq c^*$. But we know from (i) that the strict inequality here is impossible and, besides this, that $\underline{L}(\lambda) \leq c^*$ for every $\lambda \geq 0$. Thus, $\underline{L}(\lambda^*) = c^*$ and λ^* is a maximizer of $\underline{L}$ over $\lambda \geq 0$. □

D.2.2.3 Dual program

Theorem D.2.2 establishes certain connection between two optimization programs – the "primal" program

$$\text{(IC)} \qquad \min\{f(x) : g_j(x) \leq 0,\ j = 1, ..., m,\ x \in X\}$$

and its *Lagrange dual program*

$$\text{(IC}^*\text{)} \qquad \max\left\{\underline{L}(\lambda) \equiv \inf_{x \in X} L(x,\lambda) : \lambda \geq 0\right\}$$

(the variables λ of the dual problem are called the *Lagrange multipliers* of the primal problem). The Theorem says that the optimal value in the dual problem is $\leq$ the one in the primal, and under some favourable circumstances (the primal problem is convex below bounded and satisfies the Slater condition) the optimal values in the programs are equal to each other.

In our formulation there is some asymmetry between the primal and the dual programs. In fact both of the programs are related to the Lagrange function in a quite symmetric way. Indeed, consider the program

$$\min_{x \in X} \overline{L}(x), \quad \overline{L}(x) = \sup_{\lambda \geq 0} L(\lambda, x).$$

The objective in this program clearly is $+\infty$ at every point $x \in X$ which is not feasible for (IC) and is $f(x)$ on the feasible set of (IC), so that the program is equivalent to (IC). We see that both the primal and the dual programs come from the Lagrange function: in the primal problem, we minimize over X the result of maximization of $L(x,\lambda)$ in $\lambda \geq 0$, and in the dual program we maximize over $\lambda \geq 0$ the result of minimization of $L(x,\lambda)$ in $x \in X$. This is a particular (and the most important) example of a *zero sum two person game* – the issue we will speak about later.

We have seen that under certain convexity and regularity assumptions the optimal values in (IC) and (IC*) are equal to each. There is also another way to say when these optimal values are equal – this is always the case when the Lagrange function possesses a saddle point, i.e., there exists a pair $x^* \in X, \lambda^* \geq 0$ such that at the pair $L(x,\lambda)$ attains its minimum as a function of $x \in X$ and attains its maximum as a function of $\lambda \geq 0$:

$$L(x,\lambda^*) \geq L(x^*,\lambda^*) \geq L(x^*,\lambda) \quad \forall x \in X, \lambda \geq 0.$$

It can be easily demonstrated (do it by yourself or look at Theorem D.4.1) that

Proposition D.2.2 *(x^*,λ^*) is a saddle point of the Lagrange function L of* (IC) *iff x^* is an optimal solution to* (IC)*, λ^* is an optimal solution to* (IC*) *and the optimal values in the indicated problems are equal to each other.*

Our current goal is to extract from what we already know optimality conditions for convex programs.

D.2.3 Optimality Conditions in Convex Programming

D.2.3.1 Saddle point form of optimality conditions

Theorem D.2.3 [Saddle Point formulation of Optimality Conditions in Convex Programming] *Let* (IC) *be an optimization program, $L(x,\lambda)$ be its Lagrange function, and let $x^* \in X$. Then*

(i) *A sufficient condition for x^* to be an optimal solution to* (IC) *is the existence of the vector of Lagrange multipliers $\lambda^* \geq 0$ such that (x^*,λ^*) is a saddle point of the Lagrange function $L(x,\lambda)$, i.e., a point where $L(x,\lambda)$ attains its minimum as a function of $x \in X$ and attains its maximum as a function of $\lambda \geq 0$:*

$$L(x,\lambda^*) \geq L(x^*,\lambda^*) \geq L(x^*,\lambda) \quad \forall x \in X, \lambda \geq 0. \tag{D.2.11}$$

(ii) *if the problem* (IC) *is convex and satisfies the Slater condition, then the above condition is necessary for optimality of x^*: if x^* is optimal for* (IC)*, then there exists $\lambda^* \geq 0$ such that (x^*,λ^*) is a saddle point of the Lagrange function.*

Proof. (i): assume that for a given $x^* \in X$ there exists $\lambda^* \geq 0$ such that (D.2.11) is satisfied, and let us prove that then x^* is optimal for (IC). First of all, x^* is feasible: indeed, if $g_j(x^*) > 0$ for some j, then, of course, $\sup\limits_{\lambda \geq 0} L(x^*,\lambda) = +\infty$ (look what happens when all λ's, except λ_j, are fixed, and $\lambda_j \to +\infty$); but $\sup\limits_{\lambda \geq 0} L(x^*,\lambda) = +\infty$ is forbidden by the second inequality in (D.2.11).

Since x^* is feasible, $\sup\limits_{\lambda \geq 0} L(x^*,\lambda) = f(x^*)$, and we conclude from the second inequality in (D.2.11) that $L(x^*,\lambda^*) = f(x^*)$. Now the first inequality in (D.2.11) reads

$$f(x) + \sum_{j=1}^m \lambda_j^* g_j(x) \geq f(x^*) \quad \forall x \in X.$$

This inequality immediately implies that x^* is optimal: indeed, if x is feasible for (IC), then the left hand side in the latter inequality is $\leq f(x)$ (recall that $\lambda^* \geq 0$), and the inequality implies that $f(x) \geq f(x^*)$. □

(ii): Assume that (IC) is a convex program, x^* is its optimal solution and the problem satisfies the Slater condition; we should prove that then there exists $\lambda^* \geq 0$ such that (x^*,λ^*) is a saddle point of the Lagrange function, i.e., that (D.2.11) is satisfied. As we know from the Convex Programming Duality Theorem (Theorem D.2.2.(ii)), the dual problem (IC*) has a solution $\lambda^* \geq 0$ and the optimal value of the dual problem is equal to the optimal value in the primal one, i.e., to $f(x^*)$:

$$f(x^*) = \underline{L}(\lambda^*) \equiv \inf_{x \in X} L(x,\lambda^*). \tag{D.2.12}$$

We immediately conclude that

$$\lambda_j^* > 0 \Rightarrow g_j(x^*) = 0$$

(this is called *complementary slackness*: positive Lagrange multipliers can be associated only with active (satisfied at x^* as equalities) constraints. Indeed, from (D.2.12) it for sure follows that

$$f(x^*) \leq L(x^*,\lambda^*) = f(x^*) + \sum_{j=1}^m \lambda_j^* g_j(x^*);$$

the terms in the $\sum\limits_j$ in the right hand side are nonpositive (since x^* is feasible for (IC)), and the sum itself is nonnegative due to our inequality; it is possible if and only if all the terms in the sum are zero, and this is exactly the complementary slackness.

From the complementary slackness we immediately conclude that $f(x^*) = L(x^*,\lambda^*)$, so that (D.2.12) results in

$$L(x^*,\lambda^*) = f(x^*) = \inf_{x \in X} L(x,\lambda^*).$$

On the other hand, since x^* is feasible for (IC), we have $L(x^*,\lambda) \leq f(x^*)$ whenever $\lambda \geq 0$. Combining our observations, we conclude that

$$L(x^*,\lambda) \leq L(x^*,\lambda^*) \leq L(x,\lambda^*)$$

for all $x \in X$ and all $\lambda \geq 0$. □

Note that (i) is valid for an arbitrary inequality constrained optimization program, not necessarily convex. However, in the nonconvex case the *sufficient* condition for optimality given by (i) is extremely far from being necessary and is "almost never" satisfied. In contrast to this, in the convex case the condition in question is not only sufficient, but also "nearly necessary" – it for sure is necessary when (IC) is a convex program satisfying the Slater condition.

We are about to prove a modification of Theorem D.2.3, where we slightly relax the Slater condition.

Theorem D.2.4 *Consider a convex problem* (IC), *and let x^* be a feasible solution of the problem. Assume that the functions $g_1,...,g_k$ are affine, while the functions f, $g_{k+1},...,g_m$ are differentiable at x. Finally, assume the restricted Slater condition: there exists $\bar{x} \in \operatorname{ri} X$ such that $g_i(\bar{x}) \leq 0$ for $i \leq k$ and $g_i(\bar{x}) < 0$ for $i > k$. Then x_* is an optimal solution to* (IC) *if and only if there exists $\lambda^* \geq 0$ such that (x^*,λ^*) is a saddle point of $L(x,\lambda)$ on $X \times \{\lambda \geq 0\}$.*

Proof. The "if" part of the statement is given by Theorem D.2.3.(i). Let us focus on the "only if" part. Thus, assume that x^* is an optimal solution of (IC), and let us prove the existence of required λ_*. As always, we may assume without loss of generality that $\operatorname{int} X \neq \emptyset$. Let $I(x^*)$ be the set of indices of the constraints which are active at x^*. Consider the radial cone of X at x^*:

$$M_1 = \{h : \exists t > 0 : x^* + th \in X\}$$

along with the polyhedral cone

$$M_2 = \{h : (\nabla g_j^T(x^*)h \leq 0 \;\forall j \in I(x^*)\}.$$

We claim that

(I): M_2 is a closed cone which has a nonempty intersection with the interior of the convex cone M_1;

(II): the vector $\nabla f(x^*)$ belongs to the cone dual to the cone $M = M_1 \cap M_2$.

Postponing for the time being the proofs, let us derive from (I), (II) the existence of the required vector of Lagrange multipliers. Applying the Dubovitski-Milutin Lemma (Theorem B.2.7), which is legitimate due to (I), (II), we conclude that there exists a representation

$$\nabla f(x^*) = u + v, \quad u \in M_1', \, v \in M_2',$$

where M_i' is the cone dual to the cone M_i. By the Homogeneous Farkas Lemma, we have

$$v = -\sum_{j \in I(x^*)} \lambda_j^* \nabla g_j(x^*),$$

where $\lambda_j^* \geq 0$. Setting $\lambda_j^* = 0$ for $j \notin I(x^*)$, we get a vector $\lambda^* \geq 0$ such that

$$\begin{array}{c} \nabla_x \Big|_{x=x^*} L(x,\lambda^*) = \nabla f(x^*) + \sum\limits_j \lambda_j^* \nabla g_j(x^*) = \nabla f(x^*) - v = u, \\ \lambda_j^* g_j(x^*) = 0, \, j = 1,...,m. \end{array} \tag{D.2.13}$$

Since the function $L(x,\lambda^*)$ is convex in $x \in X$ and differentiable at x^*, the first relation in (D.2.13) combines with the inclusion $u \in M_1'$ and Proposition C.5.1 to imply that x^* is a minimizer of $L(x,\lambda^*)$ over $x \in X$. The second relation in (D.2.13) is the complementary slackness which, as we remember from the proof of Theorem D.2.3.(ii), combines with the feasibility of x^* to imply that λ^* is a maximizer of $L(x^*,\lambda)$ over $\lambda \geq 0$. Thus, (x^*,λ^*) is a saddle point of the Lagrange function, as claimed.

It remains to verify (I) and (II).

(I): the fact that M_2 is a closed cone is evident (M_2 is a polyhedral cone). The fact that M_1 is a convex cone with a nonempty interior is an immediate consequence of the convexity of X and the relation $\operatorname{int} X \neq \emptyset$. By assumption, there exists a point $\bar{x} \in \operatorname{int} X$ such that $g_j(\bar{x} \leq 0$ for all j. Since $\bar{x} \in \operatorname{int} X$,the vector $h = \bar{x} - x^*$ clearly belongs to $\mathrm{int}\, M_1$; since $g_j(x^*) = 0$, $j \in I(x^*)$, and $g_j(\bar{x}) \leq 0$, from Gradient Inequality it follows that $h^T \nabla g_j(x^*) \leq g_j(\bar{x}) - g_j(x^*) \leq 0$ for $j \in I(x^*)$, so that $h \in M_1$. Thus, h is the intersection of $\mathrm{int}\, M_1$ and M_2, so that this intersection is nonempty. □

(II): Assume, on the contrary to what should be proven, that there exists a vector $d \in M_1 \cap M_2$ such that $d^T \nabla f(x^*) < 0$. Let h be the same vector as in the proof of (I). Since $d^T \nabla f(x^*) < 0$, we can choose $\epsilon > 0$ such that with $d_\epsilon = d + \epsilon h$ one has $d_\epsilon^T \nabla f(x^*) < 0$. Since both d and h belong to M_1, there exists $\delta > 0$ such that $x_t = x^* + t d_\epsilon \in X$ for $0 \leq t \leq \delta$; since $d_\epsilon^T \nabla f(x^*) < 0$, we may further assume that $f(x_t) < f(x^*)$ when $0 < t \leq \delta$. Let us verify that for every $j \leq m$ one has

$(*_j)$: *There exists $\delta_j > 0$ such that $g_j(x_t) \leq 0$ for $0 \leq t \leq \delta_j$.*

This will yield the desired contradiction, since, setting $t = \min[\delta, \min\limits_j \delta_j]$, we would have $x_t \in X$, $g_j(x_t) \leq 0$, $j = 1, ..., m$, $f(x_t) < f(x^*)$, which is impossible, since x^* is an optimal solution of (IC).

To prove $(*_j)$, consider the following three possibilities:

$j \notin I(x^*)$: here $(*_j)$ is evident, since $g_j(x)$ is negative at x^* and is continuous in $x \in X$ at the point x^* (recall that all g_j are assumed even to be differentiable at x^*).

$\jmath \in I(x^*)$ and $j \leq k$: For j in question, the function $g_j(x)$ is affine and vanishes at x^*, while $\nabla g_j(x^*)$ has nonpositive inner products with both d (due to $d \in M_2$) and h (due to $g_j(x^*) = 0$, $g_j(x^* + h) = g_j(\bar{x}) \leq 0$); it follows that $\nabla g_j(x^*)$ has nonpositive inner product with d_ϵ, and since the function is affine, we arrive at $g_j(x^* + t d_\epsilon) \leq g_j(x^*) = 0$ for $t \geq 0$.

$j \in I(x^*)$ and $j > k$: In this case, the function $\gamma_j(t) = g_j(x^* + t_\epsilon)$ vanishes at $t = 0$ and is differentiable at $t = 0$ with the derivative $\gamma_j'(0) = (\epsilon h + d)^T \nabla g_j(x^*)$. This derivative is negative, since $d^T \nabla g_j(x^*) \leq 0$ due to $d \in M_2$ and $j \in I(x^*)$, while by the Gradient Inequality $h^T \nabla g_j(x^*) \leq g_j(x^* + h) - g_j(x^*) = g_j(\bar{x}) - g_j(x^*) \leq g_j(\bar{x}) < 0$. Since $\gamma_j(0) = 0$, $\gamma_j'(0) < 0$, $\gamma_j(t)$ is negative for all small enough positive t, as required in $(*_j)$. □

D.2.3.2 Karush-Kuhn-Tucker form of optimality conditions

Theorems D.2.3, D.2.4 express, basically, the strongest optimality conditions for a Convex Programming program. These conditions, however, are "implicit" – they are expressed in terms of saddle point of the Lagrange function, and it is unclear how to verify that something is or is not the saddle point of the Lagrange function. Fortunately, the proof of Theorem D.2.4 yields more or less explicit optimality conditions as follows:

Theorem D.2.5 [Karush-Kuhn-Tucker Optimality Conditions in Convex Programming] *Let* (IC) *be a convex program, let x^* be its feasible solution, and let the functions f, g_1,...,g_m be differentiable at x^*. Then*

(i) [Sufficiency] *The Karush-Kuhn-Tucker condition:*

There exist nonnegative Lagrange multipliers λ_j^*, $j = 1, ..., m$, such that

$$\lambda_j^* g_j(x^*) = 0,\ j = 1, ..., m \quad \text{[complementary slackness]} \tag{D.2.14}$$

and

$$\nabla f(x^*) + \sum_{j=1}^m \lambda_j^* \nabla g_j(x^*) \in N_X(x^*) \tag{D.2.15}$$

(that is, $(x - x^*)^T \nabla f(x^*) + \sum\limits_{j=1}^m \lambda_j^* \nabla g_j(x^*) \geq 0$ for all $x \in X$)

is sufficient for x^* to be optimal solution to (IC).

(ii) [Necessity and sufficiency] *If, in addition to the premise, the "restricted Slater assumption" holds, that is, there exists $\bar{x} \in X$ such that at $\bar{x}$ the nonlinear g_j are strictly negative, and linear g_j are nonpositive, then the Karush-Kuhn-Tucker condition from* (i) *is necessary and sufficient for x^* to be optimal solution to* (IC).

Proof. (i) is readily given by Theorem D.2.3.(ii); indeed, it is immediately seen that under the premise of Theorem D.2.5 the Karush-Kuhn-Tucker condition is sufficient for $x^*, \lambda^*)$ to be a saddle point of the Lagrange function.

(ii) is contained in the proof of Theorem D.2.4. □

Note that the optimality conditions stated in Theorem C.5.2 and Proposition C.5.1 are particular cases of the above Theorem corresponding to $m = 0$.

D.3 Duality in Linear and Convex Quadratic Programming

The fundamental role of the Lagrange function and Lagrange Duality in Optimization is clear already from the Optimality Conditions given by Theorem D.2.3, but this role is not restricted by this Theorem only. There are several cases when we can explicitly write down the Lagrange dual, and whenever it is the case, we get a pair of explicitly formulated and closely related to each other optimization programs – the *primal-dual pair*; analyzing the problems simultaneously, we get more information about their properties (and get a possibility to solve the problems numerically in a more efficient way) than it is possible when we restrict ourselves with only one problem of the pair. The detailed investigation of Duality in "well-structured" Convex Programming – in the cases when we can explicitly write down both the primal and the dual problems – goes beyond the scope of our course (mainly because the Lagrange duality is not the best possible approach here; the best approach is given by the *Fenchel Duality* – something similar, but not identical). There are, however, simple cases when already the Lagrange duality is quite appropriate. Let us look at two of these particular cases.

D.3.1 Linear Programming Duality

Let us start with some general observation. Note that the Karush-Kuhn-Tucker condition under the assumption of the Theorem ((IC) is convex, x^* is an interior point of X, $f, g_1, ..., g_m$ are differentiable at x^*) is exactly the condition that $(x^*, \lambda^* = (\lambda_1^*, ..., \lambda_m^*))$ is a saddle point of the Lagrange function

$$L(x, \lambda) = f(x) + \sum_{j=1}^{m} \lambda_j g_j(x) : \tag{D.3.1}$$

(D.2.14) says that $L(x^*, \lambda)$ attains its maximum in $\lambda \geq 0$, and (D.2.15) says that $L(x, \lambda^*)$ attains its at λ^* minimum in x at $x = x^*$.

Now consider the particular case of (IC) where $X = \mathbf{R}^n$ is the entire space, the objective f is convex and everywhere differentiable and the constraints $g_1, ..., g_m$ are linear. For this case, Theorem D.2.5 says to us that the KKT (Karush-Kuhn-Tucker) condition is necessary and sufficient for optimality of x^*; as we just have explained, this is the same as to say that the necessary and sufficient condition of optimality for x^* is that x^* along with certain $\lambda^* \geq 0$ form a saddle point of the Lagrange function. Combining these observations with Proposition D.2.2, we get the following simple result:

Proposition D.3.1 *Let* (IC) *be a convex program with* $X = \mathbf{R}^n$, *everywhere differentiable objective* f *and linear constraints* $g_1, ..., g_m$. *Then* x^* *is optimal solution to* (IC) *iff there exists* $\lambda^* \geq 0$ *such that* (x^*, λ^*) *is a saddle point of the Lagrange function (D.3.1) (regarded as a function of* $x \in \mathbf{R}^n$ *and* $\lambda \geq 0$*). In particular,* (IC) *is solvable iff* L *has saddle points, and if it is the case, then both* (IC) *and its Lagrange dual*

$$(\mathrm{IC}^*): \quad \max_{\lambda} \{\underline{L}(\lambda) : \lambda \geq 0\}$$

are solvable with equal optimal values.

Let us look what this proposition says in the Linear Programming case, i.e., when (IC) is the program

$$(P) \quad \min_{x} \left\{ f(x) = c^T x : g_j(x) \equiv b_j - a_j^T x \leq 0, \, j = 1, ..., m \right\}.$$

In order to get the Lagrange dual, we should form the Lagrange function

$$L(x, \lambda) = f(x) + \sum_{j=1}^{m} \lambda_j g_j(x) = [c - \sum_{j=1}^{m} \lambda_j a_j]^T x + \sum_{j=1}^{m} \lambda_j b_j$$

of (IC) and to minimize it in $x \in \mathbf{R}^n$; this will give us the dual objective. In our case the minimization in x is immediate: the minimal value is $-\infty$, if $c - \sum\limits_{j=1}^{m} \lambda_j a_j \neq 0$, and is $\sum\limits_{j=1}^{m} \lambda_j b_j$ otherwise. We see that the Lagrange dual is

$$(D) \quad \max_{\lambda} \left\{ b^T \lambda : \sum_{j=1}^{m} \lambda_j a_j = c, \, \lambda \geq 0 \right\}.$$

The problem we get is the usual LP dual to (P), and Proposition D.3.1 is one of the equivalent forms of the Linear Programming Duality Theorem which we already know.

D.3.2 Quadratic Programming Duality

Now consider the case when the original problem is linearly constrained convex quadratic program

$$(P) \quad \min_x \left\{ f(x) = \frac{1}{2}x^T Dx + c^T x : g_j(x) \equiv b_j - a_j^T x \le 0,\ j = 1, ..., m \right\},$$

where the objective is a strictly convex quadratic form, so that $D = D^T$ is positive definite matrix: $x^T Dx > 0$ whenever $x \neq 0$. It is convenient to rewrite the constraints in the vector-matrix form

$$g(x) = b - Ax \le 0,\ b = \begin{bmatrix} b_1 \\ ... \\ b_m \end{bmatrix},\ A = \begin{bmatrix} a_1^T \\ ... \\ a_m^T \end{bmatrix}.$$

In order to form the Lagrange dual to (P) program, we write down the Lagrange function

$$\begin{array}{rcl} L(x, \lambda) & = & f(x) + \sum\limits_{j=1}^{m} \lambda_j g_j(x) \\ & = & c^T x + \lambda^T (b - Ax) + \frac{1}{2}x^T Dx \\ & = & \frac{1}{2}x^T Dx - [A^T\lambda - c]^T x + b^T\lambda \end{array}$$

and minimize it in x. Since the function is convex and differentiable in x, the minimum, if exists, is given by the Fermat rule

$$\nabla_x L(x, \lambda) = 0,$$

which in our situation becomes

$$Dx = [A^T\lambda - c].$$

Since D is positive definite, it is nonsingular, so that the Fermat equation has a unique solution which is the minimizer of $L(\cdot, \lambda)$; this solution is

$$x = D^{-1}[A^T\lambda - c].$$

Substituting the value of x into the expression for the Lagrange function, we get the dual objective:

$$\underline{L}(\lambda) = -\frac{1}{2}[A^T\lambda - c]^T D^{-1}[A^T\lambda - c] + b^T\lambda,$$

and the dual problem is to maximize this objective over the nonnegative orthant. Usually people rewrite this dual problem equivalently by introducing additional variables

$$t = -D^{-1}[A^T\lambda - c] \quad [[A^T\lambda - c]^T D^{-1}[A^T\lambda - c] = t^T Dt];$$

with this substitution, the dual problem becomes

$$(D) \quad \max_{\lambda, t} \left\{ -\frac{1}{2}t^T Dt + b^T\lambda : A^T\lambda + Dt = c,\ \lambda \ge 0 \right\}.$$

We see that the dual problem also turns out to be linearly constrained convex quadratic program.

Note also that in the case in question feasible problem (P) automatically is solvable[4)]

With this observation, we get from Proposition D.3.1 the following

Theorem D.3.1 [Duality Theorem in Quadratic Programming]
Let (P) be feasible quadratic program with positive definite symmetric matrix D in the objective. Then both (P) and (D) are solvable, and the optimal values in the problems are equal to each other.

The pair $(x; (\lambda, t))$ of feasible solutions to the problems is comprised of the optimal solutions to them

(i) *iff the primal objective at x is equal to the dual objective at (λ, t)* ["zero duality gap" optimality condition]

same as

(ii) *iff*

$$\lambda_i(Ax - b)_i = 0,\ i = 1, ..., m, \quad \textit{and} \quad t = -x. \tag{D.3.2}$$

[4)] since its objective, due to positive definiteness of D, goes to infinity as $|x| \to \infty$, and due to the following general fact:
Let (IC) *be a feasible program with closed domain X, continuous on X objective and constraints and such that $f(x) \to \infty$ as $x \in X$ "goes to infinity" (i.e., $|x| \to \infty$). Then* (IC) *is solvable.*
You are welcome to prove this simple statement (it is among the exercises accompanying the Lecture)

Proof. (i): we know from Proposition D.3.1 that the optimal value in minimization problem (P) is equal to the optimal value in the maximization problem (D). It follows that the value of the primal objective at any primal feasible solution is $\geq$ the value of the dual objective at any dual feasible solution, and equality is possible if and only if these values coincide with the optimal values in the problems, as claimed in (i).

(ii): Let us compute the difference Δ between the values of the primal objective at primal feasible solution x and the dual objective at dual feasible solution (λ, t):

$$\begin{array}{rcl}\Delta & = & c^T x + \frac{1}{2}x^T Dx - [b^T\lambda - \frac{1}{2}t^T Dt] \\ & = & [A^T\lambda + Dt]^T x + \frac{1}{2}x^T Dx + \frac{1}{2}t^T Dt - b^T\lambda \\ & & [\text{since } A^T\lambda + Dt = c] \\ & = & \lambda^T[Ax - b] + \frac{1}{2}[x+t]^T D[x+t]\end{array}$$

Since $Ax - b \geq 0$ and $\lambda \geq 0$ due to primal feasibility of x and dual feasibility of (λ, t), both terms in the resulting expression for Δ are nonnegative. Thus, $\Delta = 0$ (which, by (i). is equivalent to optimality of x for (P) and optimality of (λ, t) for (D)) iff $\sum\limits_{j=1}^{m} \lambda_j (Ax - b)_j = 0$ and $(x+t)^T D(x+t) = 0$. The first of these equalities, due to $\lambda \geq 0$ and $Ax \geq b$, is equivalent to $\lambda_j(Ax - b)_j = 0$, $j = 1, ..., m$; the second, due to positive definiteness of D, is equivalent to $x + t = 0$. □

D.4 Saddle Points

D.4.1 Definition and Game Theory interpretation

When speaking about the "saddle point" formulation of optimality conditions in Convex Programming, we touched a very interesting in its own right topic of Saddle Points. This notion is related to the situation as follows. Let $X \subset \mathbf{R}^n$ and $\Lambda \in \mathbf{R}^m$ be two nonempty sets, and let

$$L(x, \lambda) : X \times \Lambda \to \mathbf{R}$$

be a real-valued function of $x \in X$ and $\lambda \in \Lambda$. We say that a point $(x^*, \lambda^*) \in X \times \Lambda$ is a *saddle point* of L on $X \times \Lambda$, if L attains in this point its maximum in $\lambda \in \Lambda$ and attains at the point its minimum in $x \in X$:

$$L(x, \lambda^*) \geq L(x^*, \lambda^*) \geq L(x^*, \lambda) \quad \forall (x, \lambda) \in X \times \Lambda. \tag{D.4.1}$$

The notion of a saddle point admits natural interpretation in *game terms*. Consider what is called a *two person zero sum game* where player I chooses $x \in X$ and player II chooses $\lambda \in \Lambda$; after the players have chosen their decisions, player I pays to player II the sum $L(x, \lambda)$. Of course, I is interested to minimize his payment, while II is interested to maximize his income. What is the natural notion of the equilibrium in such a game – what are the choices (x, λ) of the players I and II such that every one of the players is not interested to vary his choice independently on whether he knows the choice of his opponent? It is immediately seen that the equilibria are exactly the saddle points of the cost function L. Indeed, if (x^*, λ^*) is such a point, than the player I is not interested to pass from x to another choice, given that II keeps his choice λ fixed: the first inequality in (D.4.1) shows that such a choice cannot decrease the payment of I. Similarly, player II is not interested to choose something different from λ^*, given that I keeps his choice x^* – such an action cannot increase the income of II. On the other hand, if (x^*, λ^*) is not a saddle point, then either the player I can decrease his payment passing from x^* to another choice, given that II keeps his choice at λ^* – this is the case when the first inequality in (D.4.1) is violated, or similarly for the player II; thus, equilibria are exactly the saddle points.

The game interpretation of the notion of a saddle point motivates deep insight into the structure of the set of saddle points. Consider the following two situations:

(A) player I makes his choice first, and player II makes his choice already knowing the choice of I;

(B) vice versa, player II chooses first, and I makes his choice already knowing the choice of II.

In the case (A) the reasoning of I is: If I choose some x, then II of course will choose λ which maximizes, for my x, my payment $L(x, \lambda)$, so that I shall pay the sum

$$\overline{L}(x) = \sup_{\lambda \in \Lambda} L(x, \lambda);$$

Consequently, my policy should be to choose x which minimizes my *loss function* L, i.e., the one which solves the optimization problem

$$\text{(I)} \qquad \min_{x\in X} \overline{L}(x);$$

with this policy my anticipated payment will be

$$\inf_{x\in X} \overline{L}(x) = \inf_{x\in X} \sup_{\lambda\in\Lambda} L(x,\lambda).$$

In the case (B), similar reasoning of II enforces him to choose λ maximizing his *profit function*

$$\underline{L}(\lambda) = \inf_{x\in X} L(x,\lambda),$$

i.e., the one which solves the optimization problem

$$\text{(II)} \qquad \max_{\lambda\in\Lambda} \underline{L}(\lambda);$$

with this policy, the anticipated profit of II is

$$\sup_{\lambda\in\Lambda} \underline{L}(\lambda) = \sup_{\lambda\in\Lambda} \inf_{x\in X} L(x,\lambda).$$

Note that these two reasonings relate to two *different* games: the one with priority of II (when making his decision, II already knows the choice of I), and the one with similar priority of I. Therefore we should not, generally speaking, expect that the anticipated loss of I in (A) is equal to the anticipated profit of II in (B). What can be guessed is that the anticipated loss of I in (B) is *less than or equal to* the anticipated profit of II in (A), since the conditions of the game (B) are better for I than those of (A). Thus, we may guess that independently of the structure of the function $L(x,\lambda)$, there is the inequality

$$\sup_{\lambda\in\Lambda} \inf_{x\in X} L(x,\lambda) \le \inf_{x\in X} \sup_{\lambda\in\Lambda} L(x,\lambda). \tag{D.4.2}$$

This inequality indeed is true; which is seen from the following reasoning:

$$\begin{array}{l} \forall y\in X: \quad \inf\limits_{x\in X} L(x,\lambda) \le L(y,\lambda) \Rightarrow \\ \forall y\in X: \sup\limits_{\lambda\in\Lambda} \inf\limits_{x\in X} L(x,\lambda) \le \sup\limits_{\lambda\in\Lambda} L(y,\lambda) \equiv \underline{L}(y); \end{array}$$

consequently, the quantity $\sup\limits_{\lambda\in\Lambda} \inf\limits_{x\in X} L(x,\lambda)$ is a lower bound for the function $\underline{L}(y)$, $y\in X$, and is therefore a lower bound for the infimum of the latter function over $y\in X$, i.e., is a lower bound for $\inf\limits_{y\in X} \sup\limits_{\lambda\in\Lambda} L(y,\lambda)$.

Now let us look what happens when the game in question has a saddle point (x^*,λ^*), so that

$$L(x,\lambda^*) \ge L(x^*,\lambda^*) \ge L(x^*,\lambda) \quad \forall (x,\lambda)\in X\times\Lambda. \tag{D.4.3}$$

We claim that if it is the case, then

(*) x^* *is an optimal solution to* (I), λ^* *is an optimal solution to* (II) *and the optimal values in these two optimization problems are equal to each other* (and are equal to the quantity $L(x^*,\lambda^*)$).

Indeed, from (D.4.3) it follows that

$$\underline{L}(\lambda^*) \ge L(x^*,\lambda^*) \ge \overline{L}(x^*),$$

whence, of course,

$$\sup_{\lambda\in\Lambda} \underline{L}(\lambda) \ge \underline{L}(\lambda^*) \ge L(x^*,\lambda^*) \ge \overline{L}(x^*) \ge \inf_{x\in X} \overline{L}(x).$$

the very first quantity in the latter chain is $\le$ the very last quantity by (D.4.2), which is possible iff all the inequalities in the chain are equalities, which is exactly what is said by (A) and (B).

Thus, if (x^*,λ^*) is a saddle point of L, then (*) takes place. We are about to demonstrate that the inverse also is true:

Theorem D.4.1 [Structure of the saddle point set] *Let $L : X \times Y \to \mathbf{R}$ be a function. The set of saddle points of the function is nonempty if and only if the related optimization problems* (I) *and* (II) *are solvable and the optimal values in the problems are equal to each other. If it is the case, then the saddle points of L are exactly all pairs (x^*, λ^*) with x^* being an optimal solution to* (I) *and λ^* being an optimal solution to* (II)*, and the value of the cost function $L(\cdot,\cdot)$ at every one of these points is equal to the common optimal value in* (I) *and* (II)*.*

Proof. We already have established "half" of the theorem: if there are saddle points of L, then their components are optimal solutions to (I), respectively, (II), and the optimal values in these two problems are equal to each other and to the value of L at the saddle point in question. To complete the proof, we should demonstrate that if x^* is an optimal solution to (I), λ^* is an optimal solution to (II) and the optimal values in the problems are equal to each other, then (x^*, λ^*) is a saddle point of L. This is immediate: we have

$$\begin{array}{rcll} L(x,\lambda^*) & \geq & \underline{L}(\lambda^*) & [\text{ definition of } \underline{L}] \\ & = & \overline{L}(x^*) & [\text{by assumption}] \\ & \geq & L(x^*,\lambda) & [\text{definition of } \overline{L}] \end{array}$$

whence

$$L(x,\lambda^*) \geq L(x^*,\lambda) \quad \forall x \in X, \lambda \in \Lambda;$$

substituting $\lambda = \lambda^*$ in the right hand side of this inequality, we get $L(x,\lambda^*) \geq L(x^*,\lambda^*)$, and substituting $x = x^*$ in the right hand side of our inequality, we get $L(x^*,\lambda^*) \geq L(x^*,\lambda)$; thus, (x^*,λ^*) indeed is a saddle point of L. □

D.4.2 Existence of Saddle Points

It is easily seen that a "quite respectable" cost function may have no saddle points, e.g., the function $L(x,\lambda) = (x-\lambda)^2$ on the unit square $[0,1] \times [0,1]$. Indeed, here

$$\underline{L}(x) = \sup_{\lambda \in [0,1]} (x-\lambda)^2 = \max\{x^2, (1-x)^2\},$$

$$\overline{L}(\lambda) = \inf_{x \in [0,1]} (x-\lambda)^2 = 0, \;\; \lambda \in [0,1],$$

so that the optimal value in (I) is $\frac{1}{4}$, and the optimal value in (II) is 0; according to Theorem D.4.1 it means that L has no saddle points.

On the other hand, there are generic cases when L has a saddle point, e.g., when

$$L(x,\lambda) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) : X \times \mathbf{R}^m_+ \to \mathbf{R}$$

is the Lagrange function of a solvable convex program satisfying the Slater condition. Note that in this case L is convex in x for every $\lambda \in \Lambda \equiv \mathbf{R}^m_+$ and is linear (and therefore concave) in λ for every fixed X. As we shall see in a while, these are the structural properties of L which take upon themselves the "main responsibility" for the fact that in the case in question the saddle points exist. Namely, there exists the following

Theorem D.4.2 [Existence of saddle points of a convex-concave function (Sion-Kakutani, Bounded case)] *Let X and Λ be convex compact sets in $\mathbf{R}^n$ and $\mathbf{R}^m$, respectively, and let*

$$L(x,\lambda) : X \times \Lambda \to \mathbf{R}$$

be a continuous convex-concave (that is, convex in $x \in X$ for every fixed $\lambda \in \Lambda$ and concave in $\lambda \in \Lambda$ for every fixed $x \in X$) function. Then L has saddle points on $X \times \Lambda$.

Proof. According to Theorem D.4.1, we should prove that

- (i) Optimization problems (I) and (II) are solvable
- (ii) the optimal values in (I) and (II) are equal to each other.

(i) is valid independently of convexity-concavity of L and is given by the following routine reasoning from the Analysis:

Since X and Λ are compact sets and L is continuous on $X \times \Lambda$, due to the well-known Analysis theorem L is uniformly continuous on $X \times \Lambda$: for every $\epsilon > 0$ there exists $\delta(\epsilon) > 0$ such that

$$|x - x'| + |\lambda - \lambda'| \leq \delta(\epsilon) \Rightarrow |L(x,\lambda) - L(x',\lambda')| \leq \epsilon \quad {}^{5)} \tag{D.4.4}$$

In particular,

$$|x - x'| \leq \delta(\epsilon) \Rightarrow |L(x,\lambda) - L(x'\lambda)| \leq \epsilon,$$

whence, of course, also

$$|x - x'| \leq \delta(\epsilon) \Rightarrow |\overline{L}(x) - \overline{L}(x')| \leq \epsilon,$$

so that the function $\overline{L}$ is continuous on X. Similarly, $\underline{L}$ is continuous on Λ. Taking in account that X and Λ are compact sets, we conclude that the problems (I) and (II) are solvable.

(ii) is the essence of the matter; here, of course, the entire construction heavily exploits convexity-concavity of L.

0^0. To prove (ii), we first establish the following statement, which is important by its own right:

Lemma D.4.1 [Minmax Lemma] *Let X be a convex compact set and $f_0, ..., f_N$ be a collection of $N+1$ convex and continuous functions on X. Then the minmax*

$$\min_{x \in X} \max_{i=0,...,N} f_i(x) \tag{D.4.5}$$

of the collection is equal to the minimum in $x \in X$ of certain convex combination of the functions: there exist nonnegative μ_i, $i = 0, ..., N$, with unit sum such that

$$\min_{x \in X} \max_{i=0,...,N} f_i(x) = \min_{x \in X} \sum_{i=0}^{N} \mu_i f_i(x)$$

Remark D.4.1 Minimum of *every* convex combination of a collection of *arbitrary* functions is $\leq$ the minmax of the collection; this evident fact can be also obtained from (D.4.2) as applied to the function

$$M(x,\mu) = \sum_{i=0}^{N} \mu_i f_i(x)$$

on the direct product of X and the standard simplex

$$\Delta = \{\mu \in \mathbf{R}^{N+1} \mid \mu \geq 0, \sum_i \mu_i = 1\}.$$

The Minmax Lemma says that if f_i are convex and continuous on a convex compact set X, then the indicated inequality is in fact equality; you can easily verify that this is nothing but the claim that the function M possesses a saddle point. Thus, the Minmax Lemma is in fact a particular case of the Sion-Kakutani Theorem; we are about to give a direct proof of this particular case of the Theorem and then to derive the general case from this particular one.

Proof of the Minmax Lemma. Consider the optimization program

$$(S) \quad \min_{t,x} \{t : f_0(x) - t \leq 0,\ f_1(x) - t \leq 0, ..., f_N(x) - t \leq 0,\ x \in X\}.$$

This clearly is a convex program with the optimal value

$$t^* = \min_{x \in X} \max_{i=0,...,N} f_i(x)$$

[5)] for those not too familiar with Analysis, we wish to stress the difference between the usual continuity and the uniform continuity: continuity of L means that given $\epsilon > 0$ *and a point* (x,λ), it is possible to choose $\delta > 0$ such that (D.4.4) is valid; the corresponding δ may depend on (x,λ), not only on ϵ. Uniform continuity means that this positive δ may be chosen as a function of ϵ only. The fact that a continuous on a compact set function automatically is uniformly continuous on the set is one of the most useful features of compact sets (note that (t,x) is feasible solution for (S) iff $x \in X$ and $t \geq \max\limits_{i=0,...,N} f_i(x)$). The problem clearly satisfies the Slater condition and is solvable (since X is compact set and f_i, $i = 0,...,N$, are continuous on X; therefore their maximum also is continuous on X and thus attains its minimum on the compact set X). Let (t^*, x^*) be an optimal solution to the problem. According to Theorem D.2.3, there exists $\lambda^* \geq 0$ such that $((t^*, x^*), \lambda^*)$ is a saddle point of the corresponding Lagrange function

$$L(t, x; \lambda) = t + \sum_{i=0}^{N} \lambda_i (f_i(x) - t) = t(1 - \sum_{i=0}^{N} \lambda_i) + \sum_{i=0}^{N} \lambda_i f_i(x),$$

and the value of this function at $((t^*, x^*), \lambda^*)$ is equal to the optimal value in (S), i.e., to t^*.

Now, since $L(t, x; \lambda^*)$ attains its minimum in (t, x) over the set $\{t \in \mathbf{R}, x \in X\}$ at (t^*, x^*), we should have

$$\sum_{i=0}^{N} \lambda_i^* = 1$$

(otherwise the minimum of L in (t, x) would be $-\infty$). Thus,

$$[\min_{x \in X} \max_{i=0,...,N} f_i(x) =] \quad t^* = \min_{t \in \mathbf{R}, x \in X} \left[t \times 0 + \sum_{i=0}^{N} \lambda_i^* f_i(x) \right],$$

so that

$$\min_{x \in X} \max_{i=0,...,N} f_i(x) = \min_{x \in X} \sum_{i=0}^{N} \lambda_i^* f_i(x)$$

with some $\lambda_i^* \geq 0$, $\sum\limits_{i=0}^{N} \lambda_i^* = 1$, as claimed. □

From the Minmax Lemma to the Sion-Kakutani Theorem. We should prove that the optimal values in (I) and (II) (which, by (i), are well defined reals) are equal to each other, i.e., that

$$\inf_{x \in X} \sup_{\lambda \in \Lambda} L(x, \lambda) = \sup_{\lambda \in \Lambda} \inf_{x \in X} L(x, \lambda).$$

We know from (D.4.4) that the first of these two quantities is greater than or equal to the second, so that all we need is to prove the inverse inequality. For me it is convenient to assume that the right quantity (the optimal value in (II)) is 0, which, of course, does not restrict generality; and all we need to prove is that the left quantity – the optimal value in (I) – cannot be positive.

1^0. What does it mean that the optimal value in (II) is zero? When it is zero, then the function $\underline{L}(\lambda)$ is nonpositive for every λ, or, which is the same, the convex continuous function of $x \in X$ – the function $L(x, \lambda)$ – has nonpositive minimal value over $x \in X$. Since X is compact, this minimal value is achieved, so that the set

$$X(\lambda) = \{x \in X \mid L(x, \lambda) \leq 0\}$$

is nonempty; and since X is convex and L is convex in $x \in X$, the set $X(\lambda)$ is convex (as a level set of a convex function, Proposition C.1.4). Note also that the set is closed (since X is closed and $L(x, \lambda)$ is continuous in $x \in X$).

2^0. Thus, if the optimal value in (II) is zero, then the set $X(\lambda)$ is a nonempty convex compact set for every $\lambda \in \Lambda$. And what does it mean that the optimal value in (I) is nonpositive? It means exactly that there is a point $x \in X$ where the function $\overline{L}$ is nonpositive, i.e., the point $x \in X$ where $L(x, \lambda) \leq 0$ for all $\lambda \in \Lambda$. In other words, to prove that the optimal value in (I) is nonpositive is the same as to prove that *the sets $X(\lambda)$, $\lambda \in \Lambda$, have a point in common.*

3^0. With the above observations we see that the situation is as follows: we are given a family of closed nonempty convex subsets $X(\lambda)$, $\lambda \in \Lambda$, of a compact set X, and we should prove that these sets have a point in common. To this end, in turn, it suffices to prove that every *finite* number of sets from our family have a point in common (to justify this claim, we can refer to the Helley Theorem II, which gives us much stronger result: to prove that all $X(\lambda)$ have a point in common, it suffices to prove that every $(n+1)$ sets of this family, n being the affine dimension of X, have a point in common). Let $X(\lambda_0), ..., X(\lambda_N)$ be $N+1$ sets from our family; we should prove that the sets have a point in common. In other words, let

$$f_i(x) = L(x, \lambda_i),\ i = 0, ..., N;$$

all we should prove is that there exists a point x where all our functions are nonpositive, or, which is the same, that the minmax of our collection of functions – the quantity

$$\alpha \equiv \min_{x \in X} \max_{i=1,...,N} f_i(x)$$

– is nonpositive.

The proof of the inequality $\alpha \leq 0$ is as follows. According to the Minmax Lemma (which can be applied in our situation – since L is convex and continuous in x, all f_i are convex and continuous, and X is compact), α is the minimum in $x \in X$ of certain convex combination $\phi(x) = \sum\limits_{i=0}^{N} \nu_i f_i(x)$ of the functions $f_i(x)$. We have

$$\phi(x) = \sum_{i=0}^{N} \nu_i f_i(x) \equiv \sum_{i=0}^{N} \nu_i L(x, \lambda_i) \leq L(x, \sum_{i=0}^{N} \nu_i \lambda_i)$$

(the last inequality follows from concavity of L in λ; this is the only – and crucial – point where we use this assumption). We see that $\phi(\cdot)$ is majorated by $L(\cdot, \lambda)$ for a properly chosen λ; it follows that the minimum of ϕ in $x \in X$ – and we already know that this minimum is exactly α – is nonpositive (recall that the minimum of L in x is nonpositive for every λ). □

The next theorem lifts the assumption of boundedness of X and Λ in Theorem D.4.2 – now only one of these sets should be bounded – at the price of some weakening of the conclusion.

Theorem D.4.3 [Swapping min and max in convex-concave saddle point problem (Sion-Kakutani)] *Let X and Λ be convex sets in $\mathbf{R}^n$ and $\mathbf{R}^m$, respectively, with X being compact, and let*

$$L(x, \lambda) : X \times \Lambda \to \mathbf{R}$$

be a continuous function which is convex in $x \in X$ for every fixed $\lambda \in \Lambda$ and is concave in $\lambda \in \Lambda$ for every fixed $x \in X$. Then

$$\inf_{x \in X} \sup_{\lambda \in \Lambda} L(x, \lambda) = \sup_{\lambda \in \Lambda} \inf_{x \in X} L(x, \lambda). \tag{D.4.6}$$

Proof. By general theory, in (D.4.6) the left hand side is $\geq$ the right hand side, so that there is nothing to prove when the right had side is $+\infty$. Assume that this is not the case. Since X is compact and L is continuous in $x \in X$, $\inf_{x \in X} L(x, \lambda) > -\infty$ for every $\lambda \in \Lambda$, so that the left hand side in (D.4.6) cannot be $-\infty$; since it is not $+\infty$ as well, it is a real, and by shift, we can assume w.l.o.g. that this real is 0:

$$\sup_{\lambda \in \Lambda} \inf_{x \in X} L(x, \lambda) = 0.$$

All we need to prove now is that the left hand side in (D.4.6) is nonpositive. Assume, on the contrary, that it is positive and thus is $> c$ with some $c > 0$. Then for every $x \in X$ there exists $\lambda_x \in \Lambda$ such that $L(x, \lambda_x) > c$. By continuity, there exists a neighborhood V_x of x on X such that $L(x', \lambda_x) \geq c$ for all $x' \in V_x$. Since X is compact, we can find finitely many points $x_1, ..., x_n$ in X such that the union, over $1 \leq i \leq n$, of V_{x_i} is exactly X, implying that $\max_{1 \leq i \leq n} L(x, \lambda_{x_i}) \geq c$ for every $x \in X$. Now let $\bar{\Lambda}$ be the convex hull of $\{\lambda_{x_1}, ..., \lambda_{x_n}\}$, so that $\max_{\lambda \in \bar{\Lambda}} L(x, \lambda) \geq c$ for every $x \in X$. Applying to L and the convex *compact* sets $X, \bar{\Lambda}$ Theorem D.4.2, we get the equality in the following chain:

$$c \leq \min_{x \in X} \max_{\lambda \in \bar{\Lambda}} L(x, \lambda) = \max_{\lambda \in \bar{\Lambda}} \min_{x \in X} L(x, \lambda) \leq \sup_{\lambda \in \Lambda} \min_{x \in X} L(x, \lambda) = 0,$$

which is a desired contradiction (recall that $c > 0$). □

Slightly strengthening the premise of Theorem D.4.3, we can replace (D.4.6) with existence of a saddle point:

Theorem D.4.4 [Existence of saddle point in convex-concave saddle point problem (Sion-Kakutani, Semi-Bounded case)] *Let X and Λ be closed convex sets in $\mathbf{R}^n$ and $\mathbf{R}^m$, respectively, with X being compact, and let*

$$L(x,\lambda) : X \times \Lambda \to \mathbf{R}$$

be a continuous function which is convex in $x \in X$ for every fixed $\lambda \in \Lambda$ and is concave in $\lambda \in \Lambda$ for every fixed $x \in X$. Assume that for every $a \in \mathbf{R}$, there exist a collection $x_1^a, ..., x_{n_a}^a \in X$ such that the set

$$\{\lambda \in \Lambda : L(x_i^a, \lambda) \geq a \, 1 \leq i \leq n_a\}$$

is bounded[6]*. Then L has saddle points on $X \times \Lambda$.*

Proof. Since X is compact and L is continuous, the function $\underline{L}(\lambda) = \min_{x \in X} L(x,\lambda)$ is real-valued and continuous on Λ. Further, for every $a \in \mathbf{R}$, the set $\{\lambda \in \Lambda : \underline{L}(\lambda) \geq a\}$ clearly is contained in the set $\{\lambda : L(x_i^a, \lambda) \geq a, \, 1 \leq i \leq n_a\}$ and thus is bounded. Thus, $\underline{L}(\lambda)$ is a continuous function on a closed set Λ, and the level sets $\{\lambda \in \Lambda : \underline{L}(\lambda) \geq a\}$ are bounded, implying that $\underline{L}$ attains its maximum on Λ. Invoking Theorem D.4.3, it follows that $\inf_{x \in X} [\overline{L}(x) := \sup_{\lambda \in \Lambda} L(x,\lambda)]$ is finite, implying that the function $\overline{L}(\cdot)$ is not $+\infty$ identically in $x \in X$. Since L is continuous, $\overline{L}$ is lower semicontinuous. Thus, $\overline{L} : X \to \mathbf{R} \cup \{+\infty\}$ is a lower semicontinuous proper (i.e., not identically $+\infty$) function on X; since X is compact, $\overline{L}$ attains its minimum on X. Thus, both problems $\max_{\lambda \in \Lambda} \underline{L}(\lambda)$ and $\min_{x \in X} \overline{L}(x)$ are solvable, and the optimal values in the problem are equal by Theorem D.4.3. Invoking Theorem D.4.1, L has a saddle point. □

[6] this definitely is the case true when $L(\bar{x}, \lambda)$ is coercive in λ for some $\bar{x} \in X$, meaning that the sets $\{\lambda \in \Lambda : L(\bar{x}, \lambda) \geq a\}$ are bounded for every $a \in \mathbf{R}$, or, equivalently, whenever $\lambda_i \in \Lambda$ and $\|\lambda_i\|_2 \to \infty$ as $i \to \infty$, we have $L(\bar{x}, \lambda_i) \to -\infty$ as $i \to \infty$.

Appendix E

Conic Programming

E.1 Preliminaries

The material of this Section is contained in Appendix A; we present it here to allow for ‘standalone” reading of Appendix E.

E.1.1 Euclidean spaces

A *Euclidean space* is a finite dimensional linear space over reals equipped with an *inner product* $\langle x,y\rangle_E$ real-valued function of $x,y\in E$ which is

- symmetric ($\langle x,y\rangle_E\equiv\langle y,x\rangle_E$),
- bilinear ($\langle\lambda u+\mu v,y\rangle_E=\lambda\langle u,y\rangle_E+\mu\langle v,y\rangle_E$, and similarly w.r.t. the second argument) and
- positive definite ($\langle x,x\rangle_E>0$ whenever $x\neq 0$).

In the sequel, we usually shorten $\langle x,y\rangle_E$ to $\langle x,y\rangle$, provided that E is fixed by the context.

Example: The standard Euclidean space $\mathbf{R}^n$. This space is comprised of n-dimensional real column vectors with the standard coordinate-wise linear operations and the inner product $\langle x,y\rangle_{\mathbf{R}^n}=x^Ty$. $\mathbf{R}^n$ is a universal example of an Euclidean space: for every Euclidean n-dimensional space $(E,\langle\cdot,\cdot\rangle_E)$ there exists a one-to-one linear mapping $x\mapsto Ax:\mathbf{R}^n\to E$ such that $x^Ty\equiv\langle Ax,Ay\rangle_E$. All we need in order to build such a mapping, is to find an *orthonormal basis* $e_1,...,e_n$, $n=\dim E$, in E, that is, a basis such that $\langle e_i,e_j\rangle_E=\delta_{ij}\equiv\left\{\begin{array}{ll}1, & i=j\\ 0, & i\neq j\end{array}\right.$; such a basis always exists. Given an orthonormal basis $\{e_i\}_{i=1}^n$, a one-to-one mapping $A:\mathbf{R}^n\to E$ preserving the inner product is given by $Ax=\sum_{i=1}^n x_ie_i$.

Example: The space $\mathbf{R}^{m\times n}$ of $m\times n$ real matrices with the Frobenius inner product. The elements of this space are $m\times n$ real matrices with the standard linear operations and the inner product $\langle A,B\rangle_F=\mathrm{Tr}(AB^T)=\sum_{i,j}A_{ij}B_{ij}$.

Example: The space $\mathbf{S}^n$ of $n\times n$ real symmetric matrices with the Frobenius inner product. This is the subspace of $\mathbf{R}^{n\times n}$ comprised of all symmetric $n\times n$ matrices; the inner product is inherited from the embedding space. Of course, for symmetric matrices, this product can be written down without transposition:

$$A,B\in\mathbf{S}^n\Rightarrow\langle A,B\rangle_F=\mathrm{Tr}(AB)=\sum_{i,j}A_{ij}B_{ij}.$$

The last example explains why we need Euclidean spaces instead of sticking to $\mathbf{R}^n$ with the standard inner product: we intend in the future to works also with the Euclidean space $\mathbf{S}^n,\langle\cdot,\cdot\rangle_F)$; while it is possible to identify it with $\mathbf{R}^N$, $N=\frac{n(n+1)}{2}$, equipped with the standard inner product, it would be complete disaster to work with “vector representations” of matrices from $\mathbf{S}^n$ instead of working with these matrices directly.

E.1.2 Linear forms on Euclidean spaces

Every homogeneous linear form $f(x)$ on a Euclidean space $(E,\langle\cdot,\cdot\rangle_E)$ can be represented in the form $f(x)=\langle e_f,x\rangle_E$ for certain vector $e_f\in E$ uniquely defined by $f(\cdot)$. The mapping $f\mapsto e_f$ is a one-to-one linear mapping of the space of linear forms on E onto E.

E.1.3 Conjugate mapping

Let $(E, \langle\cdot,\cdot\rangle_E)$ and $(F, \langle\cdot,\cdot\rangle_F)$ be Euclidean spaces. For a linear mapping $A : E \to F$ and every $f \in F$, the function $\langle Ae, f\rangle_F$ is a linear function of $e \in E$ and as such it is representable as $\langle e, A^* f\rangle_E$ for certain uniquely defined vector $A^* f \in E$. It is immediately seen that the mapping $f \mapsto A^* f$ is a linear mapping of F into E; the characteristic identity specifying this mapping is

$$\langle Ae, f\rangle_F = \langle e, A^* f\rangle_E \ \forall (e \in E, f \in F).$$

The mapping A^* is called *conjugate* to A. It is immediately seen that the conjugation is a linear operation with the properties $(A^*)^* = A$, $(AB)^* = B^* A^*$. If $\{e_j\}_{j=1}^m$ and $\{f_i\}_{i=1}^n$ are orthonormal bases in E, F, then every linear mapping $A : E \to F$ can be associated with the matrix $[a_{ij}]$ ("matrix of the mapping in the pair of bases in question") according to the identity

$$A \sum_{j=1}^m x_j e_j = \sum_i \left[\sum_j a_{ij} x_j\right] f_i$$

(in other words, a_{ij} is the i-th coordinate of the vector Ae_j in the basis $f_1, ..., f_n$). With this representation of linear mappings by matrices, the matrix representing A^* in the pair of bases $\{f_i\}$ in the argument and $\{e_j\}$ in the image spaces of A^* is the transpose of the matrix representing A in the pair of bases $\{e_j\}$, $\{f_i\}$.

E.1.4 Cones in Euclidean spaces

A nonempty subset $\mathbf{K}$ of a Euclidean space $(E, \langle\cdot,\cdot\rangle_E)$ is called a cone, if it is a convex set comprised of rays emanating from the origin, or, equivalently, whenever $t_1, t_2 \geq 0$ and $x_1, x_2 \in \mathbf{K}$, we have $t_1 x_1 + t_2 x_2 \in \mathbf{K}$.

A cone $\mathbf{K}$ is called *regular*, if it is closed, possesses a nonempty interior and is *pointed* — does not contain lines, or, which is the same, is such that $a \in \mathbf{K}$, $-a \in \mathbf{K}$ implies that $a = 0$.

Dual cone. If $\mathbf{K}$ is a cone in a Euclidean space $(E, \langle\cdot,\cdot\rangle_E)$, then the set

$$\mathbf{K}^* = \{e \in E : \langle e, h\rangle_E \geq 0 \, \forall h \in \mathbf{K}\}$$

also is a cone called the cone *dual* to $\mathbf{K}$. The dual cone always is closed. The cone dual to dual is the closure of the original cone: $(\mathbf{K}^*)^* = \mathrm{cl}\, \mathbf{K}$; in particular, $(\mathbf{K}^*)^* = \mathbf{K}$ for every closed cone $\mathbf{K}$. For a closed cone $\mathbf{K}$, the cone $\mathbf{K}^*$ possesses a nonempty interior iff $\mathbf{K}$ is pointed, and $\mathbf{K}^*$ is pointed iff $\mathbf{K}$ possesses a nonempty interior; in particular, $\mathbf{K}$ is regular iff $\mathbf{K}^*$ is so.

Example: Nonnegative ray and nonnegative orthants. The simplest one-dimensional cone is the nonnegative ray $\mathbf{R}_+ = \{t \geq 0\}$ on the real line $\mathbf{R}^1$. The simplest cone in $\mathbf{R}^n$ is the *nonnegative orthant* $\mathbf{R}^n_+ = \{x \in \mathbf{R}^n : x_i \geq 0, 1 \leq i \leq n\}$. This cone is regular and self-dual: $(\mathbf{R}^n_+)^* = \mathbf{R}^n_+$.

Example: Lorentz cone $\mathbf{L}^n$. The cone $\mathbf{L}^n$ "lives" in $\mathbf{R}^n$ and is comprised of all vectors $x = [x_1; ...; x_n] \in \mathbf{R}^n$ such that $x_n \geq \sqrt{\sum_{j=1}^{n-1} x_j^2}$; same as $\mathbf{R}^n_+$, the Lorentz cone is regular and self-dual.

By definition, $\mathbf{L}^1 = \mathbf{R}_+$ is the nonnegative orthant; this is in full accordance with the "general" definition of a Lorentz cone combined with the standard convention "a sum over an empty set of indices is 0."

Example: Semidefinite cone $\mathbf{S}^n_+$. The cone $\mathbf{S}^n_+$ "lives" in the Euclidean space $\mathbf{S}^n$ of $n \times n$ symmetric matrices equipped with the Frobenius inner product. The cone is comprised of all $n \times n$ symmetric *positive semidefinite* matrices A, i.e., matrices $A \in \mathbf{S}^n$ such that $x^T A x \geq 0$ for all $x \in \mathbf{R}^n$, or, equivalently, such that all eigenvalues of A are nonnegative. Same as $\mathbf{R}^n_+$ and $\mathbf{L}^n$, the cone $\mathbf{S}^n_+$ is regular and self-dual.

Finally, we remark that the direct product of regular cones is regular, and the dual of this product is the direct product of the duals of the original cones.

> When checking this absolutely evident statement, you should take into account how we take the direct product of Euclidean spaces, since without a Euclidean structure on the product of the Euclidean spaces embedding the cones we are multiplying, the claim about the dual of a direct product of cones becomes senseless. The Euclidean structure on the direct product $E = E_1 \times ... \times E_m$ of Euclidean spaces is defined as follows: vectors from E, by the definition of direct product, are ordered tuples $(x^1, ..., x^m)$ with $x^i \in E_i$, and we set
>
> $$\langle (x^1, ..., x^m), (y^1, ..., y^m)\rangle_E = \sum_{i=1}^m \langle x^i, y^i\rangle_{E_i}.$$

with this definition, a direct product of the spaces $\mathbf{R}^{n_1}$,...,$\mathbf{R}^{n_m}$ equipped with the standard inner products is $\mathbf{R}^{n_1+\cdots+n_m}$, also equipped with the standard inner product, and the direct product of the spaces $\mathbf{S}^{n_1}$,...,$\mathbf{S}^{n_m}$ equipped with the Frobenius inner products can be viewed as the space $\mathbf{S}^{n_1,\dots,n_m}$ of *block-diagonal* symmetric matrices with m diagonal blocks of sizes $n_1, ..., n_m$, equipped with the Frobenius inner product.

We have made several not self-evident claims, and here are their proofs (we slightly alter the order of claims and, aside of the latter item, assume w.l.o.g. that the Euclidean space in question is $\mathbf{R}^n$ with the standard inner product).

- *For every cone K, one has K^* is a closed cone, and $(K^*)^* = \mathrm{cl}\, K$.* The closedness of a dual cone is evident, same as the facts that $\mathrm{cl}\, K$ is a closed cone such that $(\mathrm{cl}\, K)^* = K^*$. Besides this, we clearly have $(K^*)^* \supset \mathrm{cl}\, K$. To prove that the latter $\supset$ is in fact $=$, assume that this is not the case, so that $(K^*)^*$ contains a vector $x \notin \mathcal{L}K$. By Separation Theorem for Convex Sets (Theorem B.2.9), there exists a linear form $e^T w$ such that $e^T x < \inf\limits_{y \in \mathrm{cl}\, K} e^T y = \inf_{yzK} e^T y$. But the infimum of a linear form $e^T y$ on a cone K is either $-\infty$ (this is the case when e has negative inner product with certain vector from K, i.e., when $e \in K^*$), or is 0 (this is the case when $e \in K^*$). We are in the case when the infimum $\inf_{y \in K} e^T y$ is $> e^T x$ and thus is finite, whence $e \in K^*$, the infimum is 0 and thus and $e^T x < 0$, which impossible due to $x \in (K^*)^*$. □

- *For every cone $\mathbf{K}$, $\mathbf{K}^*$ is pointed iff $\mathrm{int}\, \mathbf{K} \neq \emptyset$.* Indeed, if $\mathrm{int}\, \mathbf{K}$ is nonempty and thus contains a ball B of radius $r > 0$, and $h, -h \in \mathbf{K}^*$, then the linear form $h^T x$ should be both nonnegative and nonpositive on B; but a vector h can be orthogonal to all vectors from a ball of positive radius iff $h = 0$. Thus, $\mathbf{K}^*$ is pointed. On the other hand, if $\mathrm{int}\, \mathbf{K} = \emptyset$, then $\mathrm{Aff}(\mathbf{K}) \neq \mathbf{R}^n$ due to Theorem B.1.1. Since $0 \in \mathbf{K}$, $\mathrm{Aff}(\mathbf{K})$ is a linear subspace in $\mathbf{R}^n$, and since it differs from $\mathbf{R}^n$, its orthogonal complement does not reduce to $\{0\}$. In other words, there exists $h \neq 0$ which is orthogonal to $\mathrm{Aff}(\mathbf{K})$, whence $\pm h \in \mathbf{K}^*$, and the latter cone is not pointed. □

- *For a closed cone $\mathbf{K}$, $\mathbf{K}^*$ has a nonempty interior iff $\mathbf{K}$ is pointed.* This is readily given by the previous item due to $\mathbf{K} = (\mathbf{K}^*)^*$.

- *The nonnegative orthant $\mathbf{R}^n_+$ is regular and self-dual.* This is evident.

- *The Lorentz cone $\mathbf{L}^n$ is regular and self-dual.* Regularity is evident. To prove self-duality, we should verify that given $[u; t]$ with $u \in \mathbf{R}^{n-1}$, the relation $[u; t]^T [v; \tau] \geq 0$ holds true for all $[v; \tau]$ with $\tau \geq \|v\|_2$ iff $t \geq \|u\|_2$, or, which is the same, to verify that for every vector $[u; t]$ one has $\inf\limits_{[v;\tau]:\|v\|_2 \leq \tau} [u; t]^T [v; \tau] \geq 0$ iff $t \geq \|u\|_2$. This is immediate, since
$$\inf_{[v;\tau]:\|v\|_2 \leq \tau} [u; t]^T [v; \tau] = \inf_{\tau \geq 0} \left[t\tau + \inf_{v:\|v\|_2 \leq \tau} u^T v \right] = \inf_{\tau \geq 0} \tau [t - \|u\|_2]. \qquad \square$$

- *The cone $\mathbf{S}^n_+$ of positive semidefinite matrices in the space of $\mathbf{S}^n$ of symmetric $n \times n$ matrices equipped with the Frobenius inner product is regular and self-dual.* Regularity if evident. To prove self-duality we should verify that if $B \in \mathbf{S}^n$, then $\mathrm{Tr}(BX) \equiv \langle B, X \langle_{\mathbf{S}^n} \geq 0$ for all $X \in \mathbf{S}^n_+$ iff $B \in \mathbf{S}^n_+$. In one direction: setting $X = xx^T$ with $x \in \mathbf{R}^n$, we get $X \succeq 0$. Thus, if $B \in (\mathbf{S}^n_+)^*$, then $\mathrm{Tr}(xx^T B) = \mathrm{Tr}(x^T Bx) = x^T BX \geq 0$ for all x, and thus $B \in \mathbf{S}^n_+$ [1]. In the opposite direction: When $B \in \mathbf{S}^n_+$, then, by the Eigenvalue Decomposition Theorem (Theorem A.7.2), $B = \sum_{i=1}^n \lambda_i e_i e_i^T$ with orthonormal $e_1, ..., e_n$ and nonnegative λ_i (the latter in fact are eigenvalues of B). It follows that when $X \in \mathbf{S}^n_+$, then $\mathrm{Tr}(BX) = \sum_i \lambda_i \mathrm{Tr}(e_i e_i^T X) = \sum_i \lambda_i e_i^T X e_i$; when $X \in \mathbf{S}^n_+$, and all terms in the resulting sum are nonnegative, and thus $\mathrm{Tr}(BX) \geq 0$ whenever $X \in \mathbf{S}^n_+$, that is, $B \in (\mathbf{S}^n_+)^*$.

[1] We have used a simple and useful identity: *when P and Q are matrices such that PQ makes sense and is a square matrix, so that* $\mathrm{Tr}(PQ)$ *makes sense, then* $\mathrm{Tr}(PQ) = \mathrm{Tr}(QP)$ (why?).

E.2 Conic Problems

A *conic program* is an optimization program of the form

$$\mathrm{Opt}(P) = \min_x \left\{ \langle c, x\rangle_E : \begin{array}{l} A_i x - b_i \in \mathbf{K}_i,\ i = 1, ..., m, \\ Rx = r \end{array} \right\} \tag{P}$$

where

- $(E, \langle\cdot,\cdot\rangle_E)$ is a Euclidean space of *decision vectors* x and $c \in E$ is the *objective*;
- A_i, $1 \leq i \leq m$, are linear maps from E into Euclidean spaces $(F_i, \langle\cdot,\cdot\rangle_{F_i})$, $b_i \in F_i$ and $\mathbf{K}_i \subset F_i$ are regular cones;
- R is a linear mapping from E into a Euclidean space $(F, \langle\cdot,\cdot\rangle_F)$ and $r \in F$.

A relation $a - b \in \mathcal{K}$, where $\mathcal{K}$ is a regular cone, is often called *conic inequality* between a and b and is denoted $a \geq_{\mathbf{K}} b$; such a relation indeed preserves the major properties of the usual coordinate-wise vector inequality $\geq$. While in the sequel we do not use the notation $A \geq_{\mathbf{K}} b$, we do call a constraint of the form $Ax - b \in \mathbf{K}$ a *conic inequality constraint* or simply *conic constraint*.

Note that we can rewrite (P) equivalently as a conic program involving a *single* cone $\mathbf{K} = \mathbf{K}_1 \times ... \times \mathbf{K}_m$, specifically, as

$$\min_x \left\{ \langle c, x\rangle_E : \begin{array}{l} Ax - b \in \mathbf{K} = \mathbf{K}_1 \times ... \times \mathbf{K}_m, \\ Rx = r \end{array} \right\}, \quad Ax - b = \left[\begin{array}{c} A_1 x - b_1 \\ \vdots \\ A_m x - b_m \end{array} \right]; \tag{P'}$$

Since the direct product of several regular cones clearly is regular as well, (P') indeed is a legitimate "single cone" conic program.

Examples: Linear, Conic Quadratic and Semidefinite Programming. We will be especially interested in the three generic conic problems as follows:

- *Linear Programming*, or *Linear Optimization:* this is the family of all conic programs associated with nonnegative orthants $\mathbf{R}^m_+$, that is, the family of all usual LPs $\min_x\{c^T x : Ax - b \geq 0\}$;
- *Conic Quadratic Programming*, or *Conic Quadratic Optimization*, or *Second Order Cone Programming:* this is the family of all conic programs associated with the cones that are *finite direct products* of Lorentz cones, that is, the conic programs of the form

  $$\min_x \left\{ c^T x : [A_1; ...; A_m]x - [b_1; ...; b_m] \in \mathbf{L}^{k_1} \times ... \times \mathbf{L}^{k_m} \right\}$$

  where A_i are $k_i \times \dim x$ matrices and $b_i \in \mathbf{R}^{k_i}$. The "Mathematical Programming" form of such a program is

  $$\min_x \left\{ c^T x : \|\bar{A}_i x - \bar{b}_i\|_2 \leq \alpha_i^T x - \beta_i,\ 1 \leq i \leq m \right\},$$

  where $A_i = [\bar{A}_i; \alpha_i^T]$ and $b_i = [\bar{b}_i; \beta_i]$, so that α_i is the last row of A_i, and β_i is the last entry of b_i;
- *Semidefinite Programming*, or *Semidefinite Optimization:* this is the family of all conic programs associated with the cones that are *finite direct products* of Semidefinite cones, that is, the conic programs of the form

  $$\min_x \left\{ c^T x : A_i^0 + \sum_{j=1}^{\dim x} x_j A_i^j \succeq 0,\ 1 \leq i \leq m \right\}, \tag{$*$}$$

  where A_i^j are symmetric matrices of appropriate sizes.

E.3 Conic Duality

E.3.1 Conic duality — derivation

The origin of conic duality is the desire to find a systematic way to bound from below the optimal value in a conic program (P). This way is based on *linear aggregation* of the constraints of (P), namely, as follows. Let $y_i \in \mathbf{K}_i^*$ and $z \in F$. By the definition of the dual cone, for every x feasible for (P) we have

$$\langle A_i^* y_i, x\rangle_E - \langle y_i, b_i\rangle_{F_i} \equiv \langle y_i, Ax_i - b_i\rangle_{F_i} \geq 0,\ 1 \leq i \leq m,$$

and of course

$$\langle R^* z, x\rangle_E - \langle z, r\rangle_F = \langle z, Rx - r\rangle_F = 0.$$

Summing up the resulting inequalities, we get

$$\langle R^* z + \sum_i A_i^* y_i, x\rangle_E \geq \langle z, r\rangle_F + \sum_i \langle y_i, b_i\rangle_{F_i}. \tag{C}$$

By its origin, this scalar linear inequality on x is a consequence of the constraints of (P), that is, it is valid for all feasible solutions x to (P). It may happen that the left hand side in this inequality is, identically in $x \in E$, equal to the objective $\langle c, x\rangle_E$; this happens iff

$$R^* z + \sum_i A_i^* y_i = c.$$

Whenever it is the case, the right hand side of (C) is a valid lower bound on the optimal value of (P). The dual program is nothing but the program

$$\mathrm{Opt}(D) = \max_{z,\{y_i\}} \left\{ \langle z, r\rangle_F + \sum_i \langle y_i, b_i\rangle_{F_i} : \begin{array}{c} y_i \in \mathbf{K}_i^*,\ 1 \leq i \leq m, \\ R^* z + \sum_i A_i^* y_i = c \end{array} \right\} \tag{D}$$

of maximizing this lower bound.

Remark: Note that the construction we have presented is completely similar to the one we used in Section B.2.7.1 to derive the LP dual of a given LP program. The latter is the particular case of (P) where all $\mathbf{K}_i$ are nonnegative orthants of various dimensions, or, which is the same the cone $\mathbf{K}$ in (P') is a nonnegative orthant. The only minor differences stem from the facts that now it is slightly more convenient to write the primal program as a minimization one, while in LP we preferred to write down the primal program as a maximization one. Modulo this absolutely unessential difference, our derivation of the dual of an LP program is nothing but our present construction as applied to the case when all $\mathbf{K}_i$ are nonnegative rays. In fact, a reader will see that all Conic Duality constructions and results we are about to present mirror already known to us constructions and results of LP Duality.

Remark:

Coming back to conic dual of a conic program, observe that by the origin of the dual we have

Weak Duality: *One has* $\mathrm{Opt}(D) \leq \mathrm{Opt}(P)$.

Besides this, we see that (D) is a conic program. A nice and important fact is that *conic duality is symmetric.*

Symmetry of Duality: *The conic dual to* (D) *is (equivalent to)* (P).

Proof:

In order to apply to (D) the outlined recipe for building the conic dual, we should rewrite (D) as a *minimization* program

$$-\mathrm{Opt}(D) = \min_{z,\{y_i\}} \left\{ \langle z, -r\rangle_F + \sum_i \langle y_i, -b_i\rangle_{F_i} : \begin{array}{c} y_i \in \mathbf{K}_i^*, 1 \leq i \leq m \\ R^* z + \sum_i A_i^* y_i = c \end{array} \right\}; \tag{D'}$$

the corresponding space of decision vectors is the direct product $F \times F_1 \times \ldots \times F_m$ of Euclidean spaces equipped with the inner product

$$\langle [z; y_1, \ldots, y_m], [z'; y_1', \ldots, y_m'] \rangle = \langle z, z'\rangle_F + \sum_i \langle y_i, y_i'\rangle_{F_i}.$$

The above "duality recipe" as applied to (D') reads as follows: pick weights $\eta_i \in (\mathbf{K}_i^*)^* = \mathbf{K}_i$ and $\zeta \in E$, so that the scalar inequality

$$\underbrace{\langle \zeta, R^* z + \sum_i A_i^* y_i \rangle_E + \sum_i \langle \eta_i, y_i \rangle_{F_i}}_{=\langle R\zeta, z\rangle_F + \sum_i \langle A_i\zeta + \eta_i, y_i\rangle_{F_i}} \geq \langle \zeta, c \rangle_E \qquad (C')$$

in variables z, $\{y_i\}$ is a consequence of the constraints of (D'), and impose on the "aggregation weights" $\zeta, \{\eta_i \in \mathbf{K}_i\}$ an additional restriction that the left hand side in this inequality is, identically in $z, \{y_i\}$, equal to the objective of (D'), that is, the restriction that

$$R\zeta = -r,\ A_i\zeta + \eta_i = -b_i,\ 1 \leq i \leq m,$$

and maximize under this restriction the right hand side in (C'), thus arriving at the program

$$\max_{\zeta, \{\eta_i\}} \left\{ \langle c, \zeta \rangle_E : \begin{array}{l} \mathbf{K}_i \ni \eta_i = A_i[-\zeta] - b_i, 1 \leq i \leq m \\ R[-\zeta] = r \end{array} \right\}.$$

Substituting $x = -\zeta$, the resulting program, after eliminating η_i variables, is nothing but

$$\max_x \left\{ -\langle c, x \rangle_E : \begin{array}{l} A_i x - b_i \in \mathbf{K}_i,\ 1 \leq i \leq m \\ Rx = r \end{array} \right\},$$

which is equivalent to (P). □

E.3.2 Conic Duality Theorem

A conic program (P) is called *strictly feasible*, if it admits a *strictly feasible* solution, that is, a feasible solution $\bar{x}$ such that $A_i\bar{x} - b_i \in \operatorname{int} \mathbf{K}_i$, $i = 1, ..., m$.

Conic Duality Theorem is the following statement resembling very much the Linear Programming Duality Theorem:

Theorem E.3.1 [Conic Duality Theorem] *Consider a primal-dual pair of conic programs (P), (D). Then*

(i) [Weak Duality] *One has* $\mathrm{Opt}(D) \leq \mathrm{Opt}(P)$.

(ii) [Symmetry] *The duality is symmetric: (D) is a conic program, and the program dual to (D) is (equivalent to) (P).*

(iii) [Strong Duality] *If one of the programs (P), (D) is strictly feasible and bounded, then the other program is solvable, and* $\mathrm{Opt}(P) = \mathrm{Opt}(D)$.

If both the programs are strictly feasible, then both are solvable with equal optimal values.

Proof:

We have already verified Weak Duality and Symmetry. Let us prove the first claim in Strong Duality. By Symmetry, we can restrict ourselves to the case when the strictly feasible and bounded program is (P).

Consider the following two sets in the Euclidean space $G = \mathbf{R} \times F \times F_1 \times ... \times F_m$:

$$\begin{array}{rcl} T & = & \{[t; z; y_1; ...; y_m] : \exists x : t = \langle c, x \rangle_E; y_i = A_i x - b_i, 1 \leq i \leq m; \\ & & z = Rx - r\}, \\ S & = & \{[t; z; y_1; ...; y_m] : t < \mathrm{Opt}(P), y_1 \in \mathbf{K}_1, ..., y_m \in \mathbf{K}_m, z = 0\}. \end{array}$$

The sets T and S clearly are convex and nonempty; observe that they do not intersect. Indeed, assuming that $[t; z; y_1; ...; y_m] \in S \cap T$, we should have $t < \mathrm{Opt}(P)$, and $y_i \in \mathbf{K}_i$, $z = 0$ (since the point is in S), and at the same time for certain $x \in E$ we should have $t = \langle c, x \rangle_E$ and $A_i x - b_i = y_i \in \mathbf{K}_i$, $Rx - r = z = 0$, meaning that there exists a feasible solution to (P) with the value of the objective $< \mathrm{Opt}(P)$, which is impossible. Since the convex and nonempty sets S and T do not intersect, they can be separated by a linear form (Theorem B.2.9): there exists $[\tau; \zeta; \eta_1; ...; \eta_m] \in G = \mathbf{R} \times F \times F_1 \times ... \times F_m$ such that

$$\begin{array}{ll} (a) & \sup\limits_{[t;z;y_1;...;y_m] \in S} \langle [\tau; \zeta; \eta_1; ...; \eta_m], [t; z; y_1; ...; y_m] \rangle_G \\ & \leq \inf\limits_{[t;z;y_1;...;y_m] \in T} \langle [\tau; \zeta; \eta_1; ...; \eta_m], [t; z; y_1; ...; y_m] \rangle_G, \\ (b) & \inf\limits_{[t;z;y_1;...;y_m] \in S} \langle [\tau; \zeta; \eta_1; ...; \eta_m], [t; z; y_1; ...; y_m] \rangle_G \\ & < \sup\limits_{[t;z;y_1;...;y_m] \in T} \langle [\tau; \zeta; \eta_1; ...; \eta_m], [t; z; y_1; ...; y_m] \rangle_G, \end{array}$$

or, which is the same,

$$\begin{array}{ll}(a) & \sup\limits_{t<\mathrm{Opt}(P),y_i\in\mathbf{K}_i} [\tau t+\sum_i\langle\eta_i,y_i\rangle_{F_i}] \\ & \leq \inf\limits_{x\in E}[\tau\langle c,x\rangle_E+\langle\zeta,Rx-r\rangle_F+\sum_i\langle\eta_i,A_ix-b_i\rangle_{F_i}], \\ (b) & \inf\limits_{t<\mathrm{Opt}(P),y_i\in\mathbf{K}_i} [\tau t+\sum_i\langle\eta_i,y_i\rangle_{F_i}] \\ & < \sup\limits_{x\in E}[\tau\langle c,x\rangle+\langle\zeta,Rx-r\rangle_F+\sum_i\langle\eta_i,A_ix-b_i\rangle_{F_i}]. \end{array} \tag{E.3.1}$$

Since the left hand side in (E.3.1.a) is finite, we have

$$\tau\geq 0,\ -\eta_i\in\mathbf{K}_i^*, 1\leq i\leq m, \tag{E.3.2}$$

whence the left hand side in (E.3.1.a) is equal to $\tau\mathrm{Opt}(P)$. Since the right hand side in (E.3.1.a) is finite, we have

$$R^*\zeta+\sum_i A_i^*\eta_i+\tau c=0 \tag{E.3.3}$$

and the right hand side in (a) is $\langle -\zeta,r\rangle_F-\sum_i\langle\eta_i,b_i\rangle_{F_i}$, so that (E.3.1.$a$) reads

$$\tau\mathrm{Opt}(P)\leq\langle -\zeta,r\rangle_F-\sum_i\langle\eta_i,b_i\rangle_{F_i}. \tag{E.3.4}$$

We claim that $\tau>0$. Believing in our claim, let us extract from it Strong Duality. Indeed, setting $y_i=-\eta_i/\tau$, $z=-\zeta/\tau$, (E.3.2), (E.3.3) say that $z,\{y_i\}$ is a feasible solution for (D), and by (E.3.4) the value of the dual objective at this dual feasible solution is $\geq\mathrm{Opt}(P)$. By Weak Duality, this value cannot be larger than $\mathrm{Opt}(P)$, and we conclude that our solution to the dual is in fact an optimal one, and that $\mathrm{Opt}(P)=\mathrm{Opt}(D)$, as claimed.

It remains to prove that $\tau>0$. Assume this is not the case; then $\tau=0$ by (E.3.2). Now let $\bar{x}$ be a strictly feasible solution to (P). Taking inner product of both sides in (E.3.3) with $\bar{x}$, we have

$$\langle\zeta,R\bar{x}\rangle_F+\sum_i\langle\eta_i,A_i\bar{x}\rangle_{F_i}=0,$$

while (E.3.4) reads

$$-\langle\zeta,r\rangle_F-\sum_i\langle\eta_i,b_i\rangle_{F_i}\geq 0.$$

Summing up the resulting inequalities and taking into account that $\bar{x}$ is feasible for (P), we get

$$\sum_i\langle\eta_i,A_i\bar{x}-b_i\rangle\geq 0.$$

Since $A_i\bar{x}-b_i\in\mathrm{int}\,\mathbf{K}_i$ and $\eta_i\in-\mathbf{K}_i^*$, the inner products in the left hand side of the latter inequality are nonpositive, and i-th of them is zero iff $\eta_i=0$; thus, the inequality says that $\eta_i=0$ for all i. Adding this observation to $\tau=0$ and looking at (E.3.3), we see that $R^*\zeta=0$, whence $\langle\zeta,Rx\rangle_F=0$ for all x and, in particular, $\langle\zeta,r\rangle_F=0$ due to $r=R\bar{x}$. The bottom line is that $\langle\zeta,Rx-r\rangle_F=0$ for all x. Now let us look at (E.3.1.b). Since $\tau=0$, $\eta_i=0$ for all i and $\langle\zeta,Rx-r\rangle_F=0$ for all x, both sides in this inequality are equal to 0, which is impossible. We arrive at a desired contradiction.

We have proved the first claim in Strong Duality. The second claim there is immediate: if both (P), (D) are strictly feasible, then both programs are bounded as well by Weak Duality, and thus are solvable with equal optimal values by the already proved part of Strong Duality. □

Remark: The Conic Duality Theorem is a bit weaker than its LP counterpart: where in the LP case plain feasibility was enough, now strong feasibility is required. It can be easily demonstrated by examples that this difference stems from the essence of the matter rather than being a shortcoming of our proofs. Indeed, it can be easily demonstrated by examples that in the case of non-polyhedral cones various "pathologies" can take place, e.g.

- (P) can be strictly feasible and below bounded while being unsolvable;
- both (P) and (D) can be solvable, but with different optimal values, etc.

Importance of strong feasibility is the main reason for our chosen way to represent constraints of a conic program as conic inequality/inequalities augmented by a system of linear equality constraints. In principle, we could write a conic problem (P) without equality constrains, namely, as

$$\min_x \left\{c^T x : A_i x - b_i \in \mathbf{K}_i, 1 \leq i \leq m, Rx - r \in \mathbf{R}^k_+,\, r - Rx \in \mathbf{R}^k_+\right\} \qquad [k = \dim r]$$

— the possibility we used, to save notation, in LP. Now it would be unwise to treat equality constraints via pairs of opposite inequalities – the resulting problem would definitely be *not* strictly feasible[2].

E.3.3 Refinement

We can slightly refine the Conic Duality Theorem, extending the "special treatment" from linear equality constraints to *scalar* linear inequalities. Specifically, consider problem (P) and assume that one of the conic constraints in the problem, say, the first one, is just $A_1 x - b_1 \geq 0$, that is, $F_1 = \mathbf{R}^\mu$ with the standard inner product, and $\mathbf{K}_1$ is the corresponding nonnegative orthant. Thus, our primal problem is

$$\mathrm{Opt}(P) = \min_x \left\{ \langle c, x\rangle_E : \begin{array}{ll} A_1 x - b_1 \geq 0 & (a) \\ A_i x - b_i \in \mathbf{K}_i,\ 2 \leq i \leq m, & (b) \\ Rx = r & (c) \end{array} \right\} \tag{P}$$

so that the dual (D) is

$$\mathrm{Opt}(D) = \max_{z, \{y_i\}_{i=1}^m} \left\{ \langle r, z\rangle_F + b_1^T y_1 + \sum_{i=2}^m \langle b_i, y_i\rangle_{F_i} : \begin{array}{l} y_1 \geq 0, \\ y_i \in \mathbf{K}_i^*,\ 2 \leq i \leq m, \\ R^* z + \sum_{i=1}^m A_i^* y_i = c \end{array} \right\} \tag{D}$$

Essentially strict feasibility. Note that the structure of problem (D) is completely similar to the one of (P) – the variables, let them be called ξ, are subject to finitely many *scalar* linear equalities and inequalities and, on the top of it, finitely many conic inequalities $P_i\xi - p_i \in \mathbf{L}_i$, $i \in I$, where $\mathbf{L}_i$ are regular cones in Euclidean spaces. Let us call such a conic problem *essentially strictly feasible*, if it admits a feasible solution $\bar\xi$ at which the conic inequalities are satisfied strictly: $P_i\bar\xi - p_i \in \operatorname{int} \mathbf{L}_i$, $i \in I$. It turns out that the Conic Duality Theorem E.3.1 remains valid when one replaces in it "strict feasibility" with "essentially strict feasibility," which is some progress: a strictly feasible conic problem clearly is essentially strictly feasible, but not necessarily vice versa. Thus, we intend to prove

Theorem E.3.2 [Refined Conic Duality Theorem] *Consider a primal-dual pair of conic programs (P), (D). Then*

(i) [Weak Duality] *One has* $\mathrm{Opt}(D) \leq \mathrm{Opt}(P)$.

(ii) [Symmetry] *The duality is symmetric: (D) is a conic program, and the program dual to (D) is (equivalent to) (P).*

(iii) [Refined Strong Duality] *If one of the programs (P), (D) is essentially strictly feasible and bounded, then the other program is solvable, and* $\mathrm{Opt}(P) = \mathrm{Opt}(D)$.

If both the programs are essentially strictly feasible, then both are solvable with equal optimal values.

Note that the Refined Conic Duality Theorem covers the usual Linear Programming Duality Theorem: the latter is the particular case $m = 1$ of the former.

Proof of Theorem E.3.2. With the Conic Duality Theorem at our disposal, all we should take care of now is the refined strong duality. In other words, invoking primal-dual symmetry, all we need is to prove that

> (!) *If (P) is essentially strictly feasible and bounded, then (D) is solvable, and* $\mathrm{Opt}(P) = \mathrm{Opt}(D)$.

Thus, assume that (P) is essentially strictly feasible and bounded.

[2] Another way to eliminate equality constraints, which is free of the outlined shortcoming, could be to use the equality constraints to express part of the variables as linear functions of the remaining variables, thus reducing the design dimension of the problem and getting rid of equality constraints.

1^0. Let

$$X = \{x : A_1x \geq b_1, Rx = r\}.$$

This set is nonempty. Let $\Pi = \{x : Rx = r, Sx = s\}$ be the affine span of X. We claim that *Either*

(A) $X = \Pi = \{x : Rx = r, Sx = s\}$, or

(B) $X = \{x : Rx = r, Sx = s, C_1x \geq d_1\}$ with properly selected C_1, d_1 such that there exists $x' \in X$ satisfying $C_1x' > d_1$.

Indeed, assume that (A) is not the case, and let $L = \{x : Rx = 0, Sx = 0\}$ be the linear subspace to which Π is parallel. Let α_i^T, $1 \leq i \leq \mu$, be the rows of A_1, β_i be the entries in b_1, and σ_i be the orthogonal projections of α_i onto L, so that

$$L \ni \sigma_i = \alpha_i - R^Tu_i - S^Tv_i$$

for properly selected vectors u_i, v_i. We clearly have

$$X = \{x : Rx = r, Sx = s, \sigma_i^T x \geq \delta_i := \beta_i - r^Tu_i - s^Tv_i, 1 \leq i \leq \mu\}$$

Let $I = \{i : \sigma_i \neq 0\}$. For every $i \notin I$, the inequality $[0^Tx =]\sigma_i^Tx \geq \delta_i$ holds true for some x (namely, for every $x \in X$) and thus is identically true. It follows that

$$X = \{x : Rx = r, Sx = s, C_1x \geq d_1\}, \tag{E.3.5}$$

where the rows of C_1 are the vectors σ_i^T, $i \in I$, and the entries in d_1 are the respective quantities δ_i. Note that $I \neq \emptyset$ since $X \neq \Pi$. To complete the justification of our claim, it remains to note that every point x' from the relative interior of X (this set is nonempty!) satisfies $x' \in \Pi$, $C_1x' > d_1$. Indeed, X contains a set of the form $U(x') = x' + \{h \in L : \|h\|_2 \leq \delta(x')\}$ with some $\delta(x') > 0$; since the relation $C_1x \geq d_1$ should be valid on $U(x')$ and the rows of C_1 are (transposes of) nonzero vectors from L, this implies $C_1x' > d_1$.

2^0. Consider the case of (B), and let us pass from the original problem (P) to the *equivalent* problem

$$\mathrm{Opt}(P) = \min_x \left\{ \langle c, x\rangle_E : \begin{array}{ll} C_1x - d_1 \geq 0 & (a) \\ A_ix - b_i \in \mathbf{K}_i,\ i = 2, 3, ..., m, & (b) \\ \left\{ \begin{array}{lcl} Rx & = & r \\ Sx & = & s \end{array} \right. & (c) \end{array} \right\} \tag{$\bar{P}$}$$

The equivalence of (P) and $(\bar{P})$ is an immediate corollary of the fact that the set X of x's satisfying $(P.a)$ and $(P.c)$ is, by construction, exactly the same as the set of x's satisfying $(\bar{P}.a)$ and $(\bar{P}.c)$. Our next observation is that $(\bar{P})$ is strictly feasible. Indeed, let $\bar{x}$ be a feasible solution to (P) such that $A_i\bar{x} - b_i \in \mathrm{int}\,\mathbf{K}_i$, $i = 2, ..., m$; existence of such a solution is given by essentially strict feasibility of (P). Now let $x' \in X$ be such that $C_1x' > d_1$ (we have seen in the previous item that such x' exists). Then for every $\lambda \in (0, 1)$, the point $x_\lambda = (1 - \lambda)\bar{x} + \lambda x'$ belongs to X and thus satisfies $(P.c)$, same as satisfies the strict version $C_1x_\lambda > d_1$ of $(\bar{P}.a)$. For small positive λ x_λ clearly satisfies also the inclusions $A_ix_\lambda - b_i \in \mathrm{int}\,\mathbf{K}_i$, $i = 2, 3, ..., m$ and therefore is a strictly feasible solution to $(\bar{P})$. Thus, $(\bar{P})$ is strictly feasible (and bounded along with (P)), so that by the Conic Duality Theorem the dual to $(\bar{P})$ – the problem

$$\mathrm{Opt}(\bar{D}) = \max_{z,w,\eta,\{y_i\}_{i=2}^m} \left\{ \langle r, z\rangle_F + s^Tw + d_1^T\eta + \sum\nolimits_{i=2}^m \langle y_i, b_i\rangle_{F_i} : \begin{array}{c} \eta \geq 0 \\ y_i \in \mathbf{K}_i^*,\ 2 \leq i \leq m \\ R^*z + S^Tw + C_1^*\eta + \sum_{i=2}^m A_i^*y_i = c \end{array} \right\} \tag{$\bar{D}$}$$

is solvable with the optimal value $\mathrm{Opt}(P)$. Let $z^*, w^*, \eta^*, y_2^*, ..., y_m^*$ be an optimal solution to $(\bar{D})$. All we need to prove is that this solution can be converted to a feasible solution to (D) with the value of the objective of (D) at this feasible solution at least $\mathrm{Opt}(\bar{D})$ (since we have seen that $\mathrm{Opt}(\bar{D}) = \mathrm{Opt}(P)$, by Weak duality the resulting solution will be optimal for (D) with the value of the objective equal to $\mathrm{Opt}(P)$, which is all we need).

To convert $(z^*, w^*, \eta^*, y_2^*, ..., y_m^*)$ into a feasible solution to (D), let us act as follows. To simplify notation, we may assume w.l.o.g. that $w^* \geq 0$. Indeed, we can multiply by (-1) the equations in the system $Sx = s$ which correspond to negative entries in w^*, replacing simultaneously these entries with their magnitudes; in our context, this clearly changes nothing.

Now, the system of linear inequalities and equations

$$C_1x \geq d_1 \ \& \ Rx = r \ \& \ Sx \geq s$$

is satisfied everywhere on the *nonempty* solution set X of the system

$$A_1x \geq b_1 \ \& \ Rx = r.$$

Consequently, by Inhomogeneous Farkas Lemma, there exist entrywise nonnegative matrices G, H and matrices U, V of appropriate sizes such that

$$\begin{array}{llcl} (a_1) & C_1 & = & GA_1 + UR, \\ (a.2) & Gb_1 + Ur & \geq & d_1; \\ (b.1) & S & = & HA_1 + VR, \\ (b.2) & Hb_1 + Vr & \geq & s. \end{array} \tag{E.3.6}$$

Now consider the candidate solution $\bar{z}, \bar{y}_1, ..., \bar{y}_m$ to (D) as follows:

$$\begin{array}{rcl} \bar{z} & = & z^* + U^*\eta^* + V^*w^* \\ \bar{y}_1 & = & G^*\eta^* + H^*w^* \\ \bar{y}_i & = & y_i^*, \ i = 2, ..., m. \end{array}$$

This indeed is a feasible solution to (D); all we need to verify is that $\bar{y}_1 \geq 0$ (this is true due to $\eta^* \geq 0$, $w^* \geq 0$ and to entrywise nonnegativity of G, H) and that $R^*\bar{z} + \sum_{i=1}^m A_i^*\bar{y}_i = c$. The latter is immediate:

$$\begin{array}{rcl} c & = & R^*z^* + S^*w^* + C_1^*\eta^* + \sum_{i=2}^m A_i^*y_i^* = R^*[z^* + U^*\eta^* + V^*w^*] + A_1^*[G^*\eta^* + H^*w^*] + \sum_{i=2}^m A_i^*y_i^* \\ & = & R^*\bar{z} + A_1^*\bar{y}_1 + \sum_{i=2}^m A_i^*\bar{y}_i^*, \end{array}$$

where the equalities follow from (a) the fact that $(z^*, w^*, \eta^*, y_2^*, ..., y_m^*)$ is feasible for $(\bar{D})$, (b) (E.3.6.a.1,b.1), and (c) the definition of $\bar{z}, \bar{y}_i$.

Further, we have

$$\begin{array}{rcl} [\mathrm{Opt}(P) =] \, \mathrm{Opt}(\bar{D}) & = & \langle r, z^* \rangle_F + s^Tw^* + d_1^T\eta^* + \sum_{i=2}^m \langle y_i^*, b_i \rangle_{F_i} \\ & \leq & \langle r, z^* \rangle_F + [Hb_1 + Vr]^Tw^* + [Gb_1 + Ur]^T\eta^* + \sum_{i=2}^m \langle y_i^*, b_i \rangle_{F_i} \\ & = & \langle r, z^* + U^*\eta^* + V^*w^* \rangle_F + b_1^T[H^*w^* + G^*\eta^*] + \sum_{i=2}^m \langle y_i^*, b_i \rangle_{F_i} \\ & = & \langle r, \bar{z} \rangle_F + \sum_{i=1}^m \langle \bar{y}_i, b_i \rangle_{F_i}, \end{array}$$

where the first inequality is due to $w^* \geq 0, \eta^* \geq 0$ and (E.3.6.a.2,b.2), and the last equality is due to the definition of $\bar{z}, \bar{y}_1, ..., \bar{y}_m$. The resulting inequality, as it was already explained, implies that $\bar{z}, \bar{y}_1, ..., \bar{y}_m$ form an optimal solution to (D) and that $\mathrm{Opt}(P) = \mathrm{Opt}(D)$, which is all we need.

3^0. We have verified (!) in the case of (B). In the case of (A) the verification is completely similar, up to straightforward simplifications, with the equivalent reformulation

$$\mathrm{Opt}(P) = \min_x \left\{ \langle c, x \rangle_E : \begin{array}{ll} A_ix - b_i \in \mathbf{K}_i, \ i = 2, 3, ..., m, & (b) \\ \left\{ \begin{array}{lcl} Rx & = & r \\ Sx & = & s \end{array} \right. & (c) \end{array} \right\} \tag{$\widetilde{P}$}$$

of (P) in the role of $(\bar{P})$ and the relations (E.3.6.b) in the role of (E.3.6.a,b). The detailed proof is left to the reader.

Another option is to note that in the case of (A) we still have

$$X = \{x : Rx = R, Sx = s, C_1x \geq d_1\}$$

with essentially strictly feasible system of constraints (i.e., $C_1x > d_1$ for some $x \in X$); indeed, it suffices to set $C_1 = [0, ..., 0]$, $d_1 = -1$. This allows to reduce the case of (A) to the case of (B).[3] □

[3] The second option follows the old joke: Given an egg, an empty pan, a stove and a source of water, you need to boil the egg. How should you act? – Well, you pour water into the pan, put the egg there and place the pan on the stove to boil the water. Now you need to solve a close problem: all as before, except that now the pan contains the water from the very beginning. How should you act? – You pour the water out of the pan and reduce the problem to the previous one...

E.3.4 Consequences of Conic Duality Theorem

E.3.4.1 Optimality Conditions in Conic Programming

Optimality conditions in Conic Programming are given by the following statement:

Theorem E.3.3 *Consider a primal-dual pair (P), (D) of conic programs, and let both programs be essentially strictly feasible. A pair $(x, \xi \equiv [z; y_1; ...; y_m])$ of feasible solutions to (P) and (D) is comprised of optimal solutions to the respective programs iff*

(i) [Zero duality gap] *One has*

$$\begin{array}{rcl}\mathrm{DualityGap}(x;\xi) & := & \langle c, x\rangle_E - [\langle z, r\rangle_F + \sum_i \langle b_i, y_i\rangle_{F_i}] \\ & = & 0,\end{array}$$

same as iff

(ii) [Complementary slackness]

$$\forall i : \langle y_i, A_i x_i - b_i\rangle_{F_i} = 0.$$

Proof:

By Refined Conic Duality Theorem, we are in the situation when $\mathrm{Opt}(P) = \mathrm{Opt}(D)$. Therefore

$$\begin{array}{rl}\mathrm{DualityGap}(x;\xi) = & \underbrace{[\langle c, x\rangle_E - \mathrm{Opt}(P)]}_{a} \\ & + \underbrace{\left[\mathrm{Opt}(D) - \left[\langle z, b\rangle_F + \sum_i \langle b_i, y_i\rangle_{F_i}\right]\right]}_{b}\end{array}$$

Since x and ξ are feasible for the respective programs, the duality gap is nonnegative and it can vanish iff $a = b = 0$, that is, iff x and ξ are optimal solutions to the respective programs, as claimed in (i). To prove (ii), note that since x is feasible, we have

$$Rx = r,\ A_i x - b_i \in \mathbf{K}_i,\ c = A^* z + \sum_i A_i^* y_i, y_i \in \mathbf{K}_i^*,$$

whence

$$\begin{array}{l}\mathrm{DualityGap}(x;\xi) = \langle c, x\rangle_E - [\langle z, r\rangle_F + \sum_i \langle b_i, y_i\rangle_{F_i}] \\ = \langle R^* z + \sum_i A_i^* y_i, x\rangle_E - [\langle z, r\rangle_F + \sum_i \langle b_i, y_i\rangle_{F_i}] \\ = \underbrace{\langle z, Rx - r\rangle_F}_{=0} + \sum_i \underbrace{\langle y_i, A_i x - b_i\rangle_{F_i}}_{\geq 0},\end{array}$$

where the nonnegativity of the terms in the last $\sum_i$ follows from $y_i \in \mathbf{K}_i^*, A_i x_i - b_i \in \mathbf{K}_i$. We see that the duality gap, as evaluated at a pair of primal-dual feasible solutions, vanishes iff the complementary slackness holds true, and thus (ii) is readily given by (i). □

E.3.4.2 A Surrogate of GTA

The following statement is a slightly weakened forms of the Inhomogeneous Farkas Lemma (which is equivalent to GTA):

Proposition E.3.1 [Conic Inhomogeneous Farkas Lemma] *Let $\mathbf{K}$ be a regular cone. A scalar linear inequality*

$$p^T x \geq q \tag{$*$}$$

is a consequence of essentially strictly feasible system

$$Ax - b \in \mathbf{K}, Rx = r \tag{\#}$$

comprised of a conic inequality and a system of linear equations[4] *iff $(*)$ is "linear consequence" of $(\#)$, i.e., iff there exists λ, μ such that*

$$\lambda \in \mathbf{K}^*,\ \ A^*\lambda + R^*\mu = p,\ \langle b, \lambda\rangle + \langle r, \mu\rangle \geq q. \tag{E.3.7}$$

[4] Essentially strict feasibility of (#) is defined completely similarly to essentially strict feasibility of a conic problem; it means that $\mathbf{K}$ is the direct product of a nonnegative orthant (perhaps of dimension 0) and a regular cone $\mathbf{K}'$, and (#) has a feasible solution $\bar{x}$ with the $\mathbf{K}'$-component belonging to the interior of $\mathbf{K}'$.

Proof. Let (*) be a consequence of (#). Then the (essentially strictly feasible!) conic program

$$\min_x \{p^T x : Ax - b \in \mathbf{K}, Rx = r\}$$

is below bounded with optimal value $\geq q$. Applying the Refined Conic Duality Theorem, the dual program has a feasible solution with the value of the dual objective $\geq q$, which is nothing but the solvability of (E.3.7) (look at the dual!). Vice versa, if λ, μ solve (E.3.7) and x solves (#), then

$$0 \leq \langle \lambda, Ax - b\rangle + \langle \mu, Rx - r\rangle = [A^*\lambda + R^*\mu]^T x - \langle b, \lambda\rangle - \langle r, \mu\rangle = p^T x - \langle b, \lambda\rangle - \langle r, \mu\rangle \leq p^T x - q,$$

so that $(*)$ indeed is a consequence of (#); note that to get the latter conclusion, no assumption of essentially strict feasibility (and even feasibility) of (#) is needed. □

E.3.4.3 *Certificates of robust solvability status

In LP, we know complete certification schemes for the basic components of the "solvability status" of an LP program, that is, we know how to certify that the program is feasible/infeasible, feasible and bounded or feasible and unbounded, same as we know that a program is solvable iff it is feasible and bounded; all these certification schemes stem from LP Duality Theorem (Theorem B.2.7), see [127, Section 3.1]. In the conic case, we have a slightly weaker version of the Duality Theorem, and a result, the question of what are the complete certification schemes for feasibility, boundedness and solvability becomes much more difficult. It, however, admits simple answers when we ask about *robust presence* of a particular property rather than of its "plain" presence. Specifically, consider a conic problem in the single-cone form:

$$\min_x \left\{c^T x : Ax - b \in \mathbf{K}, Rx = r\right\} \tag{$\mathcal{P}$}$$

along with its dual problem

$$\max_{y,z} \{\langle b, y\rangle + \langle r, z\rangle : y \in \mathbf{K}^*, A^*y + R^*z = c\} \tag{$\mathcal{D}$}$$

and assume that the systems of linear equality constraints in $(\mathcal{P})$ and in $(\mathcal{D})$ are feasible[5]. Now imagine that we fix once for ever part of the data, namely, R, r, $A, \mathbf{K}$, but allow to perturb slightly the objective c, keeping it all the time in the image C of the linear map $(y, z) \mapsto A^*y + R^*z$, same as allow to perturb slightly the primal right hand side vector b. It may happen that arbitrarily small perturbations allow to change a particular component of the solvability status of $(\mathcal{P})$, say, convert a feasible program $(\mathcal{P})$ into an infeasible one; in this case, we say that $(\mathcal{P})$ possesses the property in a *non-robust* fashion. It may happen also that there exist $\delta > 0$ such that whenever b is replaced with b' such that $\|b - b'\| \leq \epsilon$, and c is replaced with $c' \in C$ such that $\|c - c'\| \leq \epsilon$, the property in question remains intact. In this case, we say that $(\mathcal{P})$ possess the property in a *robust* fashion. Specifically, we say that program $(\mathcal{P})$ is

- *robustly feasible*, if it is feasible and remains to be so when we replace b with a b' which is close enough to b (specifically, "if there exists $\epsilon > 0$ such that whenever $\|b' - b\| \leq \epsilon$...");
- *robustly infeasible*, if it is infeasible and remains to be so when we replace b with a b' close enough to b;
- *robustly bounded*, if it is robustly feasible and bounded, and remains bounded when we replace c with $c' \in C$ close enough to c;
- *robustly unbounded*, if it is robustly feasible and unbounded, and remains unbounded when we replace c with $c' \in C$ close enough to c;
- *robustly solvable*, if it is solvable and remains so when we replace b and c with b', $c' \in C$ close enough to b, c, respectively;

[5] This assumption, which can be easily checked by elementary Linear Algebra tools, is quite natural: when the system of linear inequalities $RX = r$ in $(\mathcal{P})$ is infeasible, $(\mathcal{P})$ is infeasible. When the system of linear inequalities $A^*y + R^*z = c$ in $(\mathcal{D})$ is unsolvable, c has a negative inner product with certain vector h such that $Ah = 0$ and $Rh = 0$ ("kernel of a linear mapping is the orthogonal complement to the image space of the conjugate mapping"), meaning that $(\mathcal{P})$ is either infeasible, or unbounded (why?). In all these cases $(\mathcal{P})$ is "bad."

- *robustly unsolvable*, if it is unsolvable and remains so when we replace b and c with b', $c' \in C$ close enough to b, c, respectively.

In the sequel, we refer to the problem $(\mathcal{P})$ with vectors b, c replaced with b', c' as to $(\mathcal{P}[b', c'])$, so that $(\mathcal{P})$ is the same as $(\mathcal{P}[b, c])$.
Warning! *In the above pairs of properties, one member is not the negation of the other member.* E.g., a program which is not robustly feasible, not necessary is robustly infeasible! In fact, it can be even feasible, but "at the border of feasibility:" – by appropriate arbitrarily small perturbation in b we can make it infeasible. This is like colors in gray scale: an object can be black, it can be white, and it can be gray – neither black nor white.
Why should we bother on robustness? The point is that if a conic program $(\mathcal{P})$ possesses certain solvability-related feature – say, feasibility – in a non-robust fashion, then, by definition, we can change this feature by appropriate *arbitrary small* perturbation of the part of the data, namely, b and c. In contrast to this, the property to possess *robustly* certain feature "is itself robust" – it is preserved by small enough perturbations in b and c. Theoretically speaking, in LP with rational data we can use, e.g., Khachiyan's algorithm to recover in CT-polynomial time the "true" solvability status of an LP instance, along with building the corresponding certificates. Similar possibility exists in LP with real data in the Real Arithmetic model of computations — it suffices to solve the LP program in question by the Simplex method. Note, however, that the resulting procedure is *not* polynomial in the Real Arithmetic Complexity Model. In "non-polyhedral" conic optimization, even in the Real Arithmetics model of computations, there is no theoretical possibility to recover in finite time the true solvability status of an instance (or at least we do not know how to do it). As about real life finite precision computations, the situation is even worse. When solving a problem with non-robust (or even "robust, but poorly so") solvability status, rounding errors can lead to completely wrong conclusions on this status, and this happens in the LP case as well[6].

Motivated by the above discussion, we are about to understand what robustness amounts to and how to it can be certified.

A. Robust feasibility. We claim that $(\mathcal{P})$ *is robust feasible iff* $(\mathcal{P})$ *is strictly feasible.* As a result, *to certify robust feasibility, it suffices to point out a vector x such that $Ax - b \in \operatorname{int} \mathbf{K}$, and this certification scheme is complete.*

The second claim is an immediate corollary of the first one, and the latter can be certified as follows. If $(\mathcal{P})$ is strictly feasible, this program clearly is robust feasible (why?). Now assume that the program is not strictly feasible, and let us prove that then it is not robust feasible as well. To this end, let $\Delta \in \operatorname{int} \mathbf{K}$; given $t > 0$, consider the program $(\mathcal{P}[b', c])$ with $b' = b - t\Delta$. We claim that this problem is infeasible. Indeed, assuming that $\bar{x}$ is a feasible solution to $(\mathcal{P}[b', c])$, we would get $R\bar{x} = r$ and $A\bar{x} - b = A\bar{x} - b' + t\Delta$; the latter vector clearly belongs to $\operatorname{int} \mathbf{K}$ (since as the sum of the vector $A\bar{x} - b'$ from $\mathbf{K}$ and the vector $t\Delta \in \operatorname{int} \mathbf{K}$), meaning that $(\mathcal{P})$ is strictly feasible, which was assumed not to be the case. When t is close to 0, $t\Delta$ is small, and we see that by arbitrarily small perturbations of $(\mathcal{P})$ we can make the problem infeasible, so that $(\mathcal{P})$ is not robustly feasible, as claimed. □

B. Robust infeasibility. We claim that $(\mathcal{P})$ *is robustly infeasible iff the constraints of the problem can be led to a contradiction by linear aggregation, that is, iff there exists* $[y; z]$ *such that*

$$y \in \mathbf{K}^*, \ A^*y + R^*z = 0, \ \langle b, y \rangle + \langle r, z \rangle > 0. \tag{E.3.8}$$

As a result, to certify robust infeasibility of $(\mathcal{P})$, *it suffices to point out a solution* $[y; z]$ *to* (E.3.8), *and this certification scheme is complete.*

Here again it suffices to prove the first claim only. Assume, first, that (E.3.8) has a solution $[y; z]$, and let us prove than $(\mathcal{P})$ is robustly infeasible. Indeed, since $\langle b, y \rangle + \langle z, r \rangle$, there exists a small enough centered at b ball B of positive radius such that $\langle b', y \rangle + \langle z, t \rangle > 0$ whenever $b' \in B$. Let us prove that every problem $(\mathcal{P}[b', c])$ $b' \in B$ is infeasible (this would mean that $(\mathcal{P})$ is robustly infeasible, as desired). Indeed, assuming

[6] Everybody with even small experience of solving LP's with commercial software knows a lot of examples when definitely solvable problems were claimed infeasible or unbounded; one of the authors remembers his shock when a quite respectable code managed to solve his problem to optimality before removed one of the constraints and claimed it infeasible after the constraint was removed. Good LP solvers at least report that in course of computations, some numerical difficulties were met, so that the results should be accepted "with a grain of salt;" not so good solvers do not care to report on difficulties...

that x is feasible for $(\mathcal{P}[b',c])$ and $b'\in B$, we would get $\langle y,Ax-b'\rangle\geq 0$ (since $y\in\mathbf{K}^*$ and $Ax-b'\in\mathbf{K}$) and $\langle z,Rx-r\rangle=0$; summing up these inequalities, we get $\langle A^*y+R^*z,x\rangle-\langle y,b'\rangle-\langle z,r\rangle\geq 0$, that is, $-\langle y,b'\rangle-\langle z,r\rangle\geq 0$, which is impossible due to $b'\in B$. Thus, $(\mathcal{P}[b',c])$ is infeasible when $b'\in B$, as claimed.

Now let us verify that if $(\mathcal{P})$ is robust infeasible, then (E.3.8) is solvable. To this end, let us choose $\Delta\in\operatorname{int}\mathbf{K}$, and consider the conic program

$$\min_{t,x}\left\{t: Ax+t\Delta-b\in\mathbf{K}, Rx=r\right\}. \tag{\$}$$

We claim that this problem is feasible with strictly positive optimal value. Strict feasibility is evident: take a whatever x satisfying $Rx=r$; then for all large enough values of t we have $t^{-1}[Ax-b]+\Delta\in\operatorname{int}\mathbf{K}$ due to $\Delta\in\operatorname{int}\mathbf{K}$, whence $Ax-b+t\Delta\in\operatorname{int}\mathbf{K}$ as well. The fact that the optimal value is positive stems from the fact that otherwise (\$) would have feasible solutions with t arbitrarily close to 0, that is, program $(\mathcal{P}[b',c])$ with $b'=b-t\Delta$, would be feasible for all close to 0 values of t, which contradicts the robust infeasibility of $(\mathcal{P})$. Since (\$) is strictly feasible with positive value of the objective, by Conic Duality Theorem the program dual to (\$), that is, the program

$$\max_{y,z}\left\{\langle b,y\rangle+\langle r,z\rangle: y\in\mathbf{K}^*, A^*y+R^*z=0, \langle y,\Delta\rangle=1\right\}$$

has a feasible solution with positive values of the objective, meaning that (E.3.8) is solvable. □

C. Robust boundedness. We claim that $(\mathcal{P})$ *is robustly bounded iff either* $(\mathcal{P})$ *is robustly infeasible, or* $(\mathcal{D})$ *is strictly feasible.* As a result, *in order to certify robust boundedness of* $(\mathcal{P})$*, it suffices either to certify robust infeasibility of* $(\mathcal{P})$ (which we already know how to do), *or to point out a strictly feasible solution to* $(\mathcal{D})$*, and this certification scheme is complete.*

We need to prove the first claim only. In one direction: Assume that $(\mathcal{P})$ is robustly bounded. We should prove that if, in addition, $(\mathcal{P})$ is not robustly infeasible, then $(\mathcal{D})$ is strictly feasible. We should, therefore, lead to a contradiction the assumption that $(\mathcal{P})$ is robustly bounded, $(\mathcal{P})$ is not robustly infeasible, and $(\mathcal{D})$ is not strictly feasible. Assume that all these properties take place. Since $(\mathcal{P})$ is robustly bounded, there exists $\epsilon>0$ such that whenever $\|b'-b\|\leq\epsilon$ and $c'\in C$, $\|c'-c\|\leq\epsilon$, the problem $(\mathcal{P}[b',c'])$ is bounded. Since $(\mathcal{P})$ is not robustly infeasible, we can find b'', $\|b''-b\|\leq\epsilon/2$, such that $(\mathcal{P}[b'',c])$ is feasible; therefore, setting $b'=b''-\Delta$ with small in norm $\Delta\in\operatorname{int}\mathbf{K}$, we get a *strictly* feasible problem $(\mathcal{P}[b',c])$ and ensure that $\|b-b'\|\leq\epsilon$. Now, the dual to $(\mathcal{P}[b',c'])$ is the problem $(\mathcal{D}[b',p'])$ obtained from $(\mathcal{D})$ by replacing b with b' and c with c'. Since $(\mathcal{P}[b',c])$ is strictly feasible and bounded, its dual $(\mathcal{D}[b',c])$ is solvable and thus feasible. But the feasible set of $(\mathcal{D}[b',c])$ is the same as the feasible set of $(\mathcal{D})$, so that the latter is nonempty. Now let $[y;z]$ be a feasible solution to $(\mathcal{D})$, and let $\Delta\in\operatorname{int}\mathbf{K}^*$. Setting $c(t)=A^*[y-t\Delta]+R^*z$, we see that $c(t)\in C$, $c(0)=c$ and the program $(\mathcal{D}[b,c(t)])$ is infeasible for every $t>0$ (since if $[\tilde{y};\tilde{z}]$ were a feasible solution to the latter program, $[\tilde{y}+t\Delta;\tilde{z}]$ would be a strictly feasible solution to $(\mathcal{D})$, and we are in the case when $(\mathcal{D})$ is not strictly feasible). Choosing small enough $t>0$ and setting $c'=c(t)$, we ensure that $\|c'-c\|\leq\epsilon$ and the problem $(\mathcal{D}[b,c'])$ is infeasible, meaning that the problem $(\mathcal{D}[b',c'])$ also is infeasible. We arrive at a contradiction: since $\|b'-b\|\leq\epsilon$, $\|c'-c\|\leq\epsilon$ and $c'\in C$, program $(\mathcal{P}[b',c'])$ should be either infeasible, or feasible and bounded. The former option is impossible, since the feasible set of the program is the same as for $(\mathcal{P}[b',c])$, that is, the program is even strictly feasible. We conclude that $(\mathcal{P}[b',c'])$ is strictly feasible and bounded, whence, by Conic Duality Theorem, $(D[b',c'])$ is feasible, which by construction of c' is not the case. We have arrived at the desired contradiction, thus proving that *if* $(\mathcal{P})$ is robustly bounded, *then* the program either is robustly infeasible, or $(\mathcal{D})$ is strictly feasible.

In the opposite direction: We should prove that *if* $(\mathcal{P})$ is either robustly infeasible, or $(\mathcal{D})$ is strictly feasible, *then* $(\mathcal{P})$ is robustly bounded. If $(\mathcal{P})$ is robustly infeasible, then of course it is robustly bounded. Now let $(\mathcal{D})$ be strictly feasible, and let $[\bar{y};\bar{z}]$ be a strictly feasible solution to the problem, so that a ball Y of a positive radius r centered at $\bar{y}$ is contained in $\mathbf{K}^*$. By elementary linear algebra, we can find *continuous in* $c'\in C$ functions $y(c')$ and $z(c')$ such that $y(c)=\bar{y}$ and $A^*y(c')+R^*z(c')=c'$ for all $c'\in C$. In particular, we can find $\epsilon>0$ such that whenever $c'\in C$ and $\|c'-c\|\leq\epsilon$, we have $\|y(c')-\bar{y}\|\leq r$, meaning that $y(c')\in\mathbf{K}^*$ and thus $[y(c');z'(c')]$ is a feasible solution to $(\mathcal{D}[b',c'])$ for all b'. By Weak Duality it follows that $(\mathcal{P}[b',c'])$ is either infeasible, or bounded for every b' and every $c'\in C$ such that $\|c'-c\|\leq\epsilon$, meaning that $(\mathcal{P})$ is robustly bounded. □

B. Robust unboundedness. We claim that $(\mathcal{P})$ *is robustly unbounded iff* $(\mathcal{P})$ *is robustly feasible and there exists h such that*

$$Ah\in\mathbf{K}, Rh=0, c^Th<0. \tag{E.3.9}$$

Consequently, *to certify robust unboundedness of* $(\mathcal{P})$*, it suffices to point out a certificate of robust feasibility* (that is, a strictly feasible solution to $(\mathcal{P})$, see above) *and a vector* h *satisfying* (E.3.9), *and this certification scheme is complete.*
As always, it suffices to prove the first claim. In one direction: assume that $(\mathcal{P}')$ is robustly feasible and a direction h satisfying (E.3.9) does exist. Then there exists a ball B of positive radius centered at c such that $[c']^T h < 0$ for all $c' \in B$. It follows that there exists $\epsilon > 0$ such that whenever $\|b' - b\| \le \epsilon$ and $\|c' - c\| \le \epsilon$, the problem $(\mathcal{P}[b', c'])$ is feasible and $[c']^T h < 0$. With b', c' as above, h clearly is a recessive direction of the feasible set of $(\mathcal{P}[b', c'])$, and along this direction the objective of $(\mathcal{P}[b', c'])$ strictly decreases, meaning that $(\mathcal{P}[b', c'])$ is unbounded.

Now assume that $(\mathcal{P})$ is robustly unbounded, and let us verify that then $(\mathcal{P})$ is robustly feasible (this is evident is evident) and (E.3.9) has a solution. It may happen that the linear mapping $x \mapsto \mathcal{A}x = \left[\begin{array}{c} Ax \\ Rx \end{array}\right]$ has a nontrivial kernel L. Note that in this case the image C of the mapping $[y; z] \mapsto A^* y + R^* x$ is exactly $L^\perp$, so that restricting x and h to reside in $C = L^\perp$ does not affect neither the robust unboundedness of $(\mathcal{P})$, nor the solvability status of (E.3.9). Thus, we can assume w.l.o.g. that $C = \mathbf{R}^n$ is the entire space of x's, and the linear mapping $\mathcal{A}$ has a trivial kernel. Now let us prove the existence of h satisfying (E.3.9).

Let $H = \{h \in \mathbf{R}^n : Ah \in \mathbf{K}, Rh = 0\}$. H clearly is a closed convex cone in $\mathbf{R}^n$. We claim that this cone is pointed. Indeed, if h is such that $\pm h \in H$, then $Rh = 0$ and $\pm Ah \in \mathbf{K}$; since $\mathbf{K}$ is pointed, it follows that $Ah = 0$. Thus, $Ah = 0$ and $Rh = 0$, whence $h = 0$ (recall that we are in the situation when the mapping $x \mapsto (Ax, Rx)$ has the trivial kernel). Now let us use the following simple and important

Lemma E.3.1 *If* M *is a closed pointed cone in* $\mathbf{R}^n$*, then there exists* $f \in H^*$ *and a constant* $\theta > 0$ *such* $\|h\|_2 \le \theta f^T h$ *for all* $h \in M$*. Specifically, one can take as* f *any vector from the nonempty set* $\mathrm{int}\, H^*$*.*

Proof of Lemma. Since M is a closed pointed cone, its dual cone M^* has a nonempty interior. Let $f \in \mathrm{int}\, M^*$, so that there exists r such that $f + e \in M^*$ whenever $\|e\|_2 \le r$. It follows that when $h \in M$, we have $0 \le \min\limits_{e:\|e\|_2 \le r} [f + e]^T h = f^T h - r\|h\|_2$, so that $r\|h\|_2 \le f^T h$ for all $h \in M$; it remains to take $\theta = r^{-1}$. □

Now we are ready to prove that (E.3.9) has a solution. Applying Lemma to the closed pointed cone H, we see that there exists a vector $f \in \mathbf{R}^n$ and $\theta > 0$ satisfying $\theta f^T h \|h\|_2$ for all $h \in H$. Since $(\mathcal{P})$ is robustly unbounded and we are in the situation $C = \mathbf{R}^n$, there exists $\epsilon > 0$ such that with $c' = c + \epsilon f$, the program $(\mathcal{P}[b, c'])$ is unbounded, that is, there exists a sequence $\{h_t\}_{t=1}^\infty$ of feasible solutions to this problem such that $[c']^T h_t \to -\infty$ as $t \to \infty$. This relation is possible only if $\rho_t = \|h_t\| \to \infty$ as $t \to \infty$. Now let $e_t = \rho_t^{-1} h_t$. Passing to a subsequence $t_i \to \infty$, $i \to \infty$ we can assume that the unit vectors e_{t_i} converge, as $i \to \infty$, to a unit vector e. We have

$$Re_t = \rho_t^{-1} Rh_t = \rho^{-1} r \to 0,\ t \to \infty \Rightarrow Re = 0$$

and

$$Ae_t - \rho_{t-1} b = \rho_t^{-1}[Ah_t - b] \in \mathbf{K},$$

and since $\mathbf{K}$ is closed and $\rho_t^{-1} b \to 0$, $t \to \infty$, we see that $Ae \in \mathbf{K}$. The bottom line is that the unit vector e belongs to H. Finally, we have $[c']^T h_t \to -\infty$ as $t \to \infty$, whence $[c']^T e = \lim_{i\to\infty} \rho_{t_i}^{-1} [c']^T h_{t_i} \le 0$. Recalling what c' is, we see that $c^T e + \epsilon f^T e \le 0$, and since $e \in H$, we have $\theta f^T e \ge \|e\|_2 = 1$, that is, $\epsilon f^T e \ge \epsilon \theta^{-1}$. The bottom line is that $c^T e + \epsilon \theta^{-1} \le 0$, whence $c^T e < 0$. Since $e \in H$, (E.3.9) holds true with $e = h$. □

E. Robust solvability. We claim that $(\mathcal{P})$ *is robustly solvable iff both* $(\mathcal{P})$ *and* $(\mathcal{D})$ *are strictly feasible.* As a result *to certify robust solvability of* $(\mathcal{P})$*, it suffices to point out strictly feasible solutions to* $(\mathcal{P})$ *and to* $(\mathcal{D})$*, and this certification scheme is complete.*

Indeed, assume that $(\mathcal{P})$ is robustly solvable. Then $(\mathcal{P})$ clearly is robustly feasible and robustly bounded. The latter fact, in view of item C, implies that $(\mathcal{D})$ is strictly feasible, while strict feasibility of $P')$, by item A, implies that $(\mathcal{P})$ is strictly feasible. Thus, if $(\mathcal{P})$ is robustly solvable, both $(\mathcal{P})$ and $(\mathcal{D})$ are strictly feasible.

To prove the inverse statement, assume that $(\mathcal{P})$ and $(\mathcal{D})$ are strictly feasible, and let $\bar{y}, \bar{z}$ be a strictly feasible solution to $(\mathcal{D})$. By the argument from item C, we can find continuous functions $y(c')$, $z(c')$ of $c' \in C$and a neighborhood U of c in C such that for all $c' \in U$ the pair $y(c'), z(c')$ is a strictly feasible solution to $(\mathcal{D}([b, c'])$, and thus for $(\mathcal{D}[b', c'])$ for all b'. Besides this, by item A $(\mathcal{P}[b', c])$ is feasible for all b' close enough to b, meaning that for these b' the program $(\mathcal{P}[b', c'])$ is feasible for all c'. The bottom line is that for all b' close enough to b and all $c' \in C$ close enough to c problem $(\mathcal{P}[b', c'])$ is feasible (and thus $(\mathcal{D}[b', c'])$ is bounded by Weak duality), and $(\mathcal{D}[b', c'])$ is strictly feasible. Since $(d'[b', c'])$ is strictly feasible and bounded for indicated b', c', $(\mathcal{P}[b', c'])$ is solvable (Conic Duality Theorem). Thus, $(\mathcal{P}[b', c'])$ is solvable for all b' and $c' \in C$ close enough to b, c, that is, $(\mathcal{P})$ is robustly solvable. $\square$

Remark: It should be added that $(\mathcal{P})$ *is robustly solvable iff* $(\mathcal{P})$ *is robustly feasible and robustly bounded.* Indeed, robustly solvable program clearly is robustly feasible and robustly bounded. To see that inverse also is true, note that if $(\mathcal{P})$ is robustly feasible, then $(\mathcal{P})$ is strictly feasible by A and $(\mathcal{D})$ is strictly feasible by C, whence $(\mathcal{P})$ is robustly solvable by F.

Remark: From the definitions of robustness it follows that, say, the property to be, say, robust feasible itself is "robust:" if $(\mathcal{P})$ possesses this property, so are all problems obtained from $(\mathcal{P})$ by small enough perturbations of b and c, and similarly for all other properties we have considered. Sometimes we can say more. For example, assume that C is the entire x-space (that is, that the mapping $(y, z) \mapsto A^*y + R^*z$ is an onto mapping, or, equivalently, the mapping $x \mapsto (Ax, Rx)$ is with trivial kernel. Then robust solvability of (P) is preserved by small enough perturbations of *all* the data, including A and R (why?).

F. Robust insolvability. We claim that $(\mathcal{P})$ *is robustly unsolvable iff the program is either robustly infeasible, or there exists h satisfying* (E.3.9), *or both.* Since we know how to certify robust infeasibility, this claim yields a complete certification scheme for robust insolvability.

Let us prove our claim. In one direction this is easy: when $(\mathcal{P})$ is robustly infeasible, then of course $(\mathcal{P})$ is robustly unsolvable. If there exists h satisfying (E.3.9), then $(\mathcal{P})$ is robustly unsolvable as well. Indeed, looking at (E.3.9), we see that h satisfying (E.3.9) satisfies also

$$Ah \in \mathbf{K}, Rh = 0, [c']^T h < 0 \tag{$*$}$$

for all $c' \in V$, where V is a small enough neighborhood of c in C. Now let us prove that $(\mathcal{P}[b', c'])$ is unsolvable, specifically, is either infeasible, or unbounded, for all b' and all $c' \in V$ (and thus $(\mathcal{P})$ is robustly unsolvable). Indeed, when $(\mathcal{P}[b', c'])$, $c' \in V$, is feasible, h from $(*)$ clearly is a recessive direction of the feasible set of the program such that the objective of the program strictly decreases along this ray, meaning that the program is unbounded.

Now assume that $(\mathcal{P})$ is robustly unsolvable, and let us prove that the program is either robust infeasible, or (E.3.9) takes place. It suffices to verify that if $(\mathcal{P})$ is robustly unsolvable and is *not* robustly feasible, then $(\mathcal{P})$ is robustly unbounded. Thus, let $(\mathcal{P})$ be robustly unsolvable and not robustly infeasible. Since $(\mathcal{P})$ is robustly unsolvable, there exists $\epsilon > 0$ such that whenever $b' \in U = \{b' : \|b - b'\| < \epsilon\}$ and $c' \in V = \{c' \in C : \|c' - c\| < \epsilon\}$, the program $(\mathcal{P}[b', c'])$ is unsolvable. Since $(\mathcal{P})$ is not robustly infeasible, there exists $\bar{b}$, $\|\bar{b} - b\| \leq \epsilon/2$, such that $(\mathcal{P}[\bar{b}, c'])$ is feasible for all c'. Setting $\tilde{b} = \bar{b} - \Delta$, where $\Delta \in \operatorname{int} \mathbf{K}$ is of norm $\leq \epsilon/3$, we ensure that $\tilde{b} \in U$ and $(\mathcal{P}[\tilde{b}, c'])$ is strictly feasible. We claim that the program $(\mathcal{P}') = (\mathcal{P}[\tilde{b}, c])$ is robustly unbounded. Indeed, since $(\mathcal{P}[\tilde{b}, c])$ is strictly feasible, there exists a neighborhood $W \subset U$ of $\tilde{b}$ such that all programs $(\mathcal{P}[b', c'])$ with $b' \in U$ are strictly feasible. Assuming that $(\mathcal{P}[\tilde{b}, c]$ is not robustly unbounded, we can find $b' \in W$ and $c' \in V$ such that $(\mathcal{P}[b', c'])$ is bounded; since the latter program is strictly feasible due to $b' \in W$, the problem $(\mathcal{D}[b', c'])$ is solvable and thus feasible. If y, z is a feasible solution to $(\mathcal{D}[b', c'])$, then, for every $\Delta \in \operatorname{int} \mathbf{K}^*$, the pair $y_\Delta = y + \Delta, z$ is a strictly feasible solution to $(\mathcal{D}[b', c_\Delta])$ with $c_\Delta = c' + A^*\Delta$. Choosing $\Delta \in \mathbf{K}^*$ to have a small enough norm, we can ensure that $x'' = x_\Delta \in V$. Thus, $b' \in U$, $c'' \in V$, problem $(\mathcal{P}[b', c''])$ is feasible, and problem $(\mathcal{D}[b', c''])$ is strictly feasible. By Weak duality, the latter program is not only strictly feasible, but also bounded, which, by Conic Duality Theorem, implies that $(\mathcal{P}[b', c''])$ is solvable, which contradicts the origin of $U \ni b'$ and $V \ni c''$. This contradiction proves that program $(\mathcal{P}[\tilde{b}, c])$ is robustly unbounded. Invoking item D, it follows that either $(\mathcal{P}[\tilde{b}, c])$ is robust infeasible, or there exists an h satisfying (E.3.9). By choice of $\tilde{b}$, the first option is impossible – problem $(\mathcal{P}[\tilde{b}, c])$ is feasible by construction. We conclude that (E.3.9) has a solution . $\square$

How "rare" are primal-dual strictly feasible programs? Here we intend to demonstrate that strict primal-dual feasibility (or, which is the same by E, robust solvability) is not too rare commodity: whenever problem $(\mathcal{P})$ is feasible and bounded, properly chosen arbitrarily small perturbations of b and c make it strictly primal-dual feasible. Specifically,

Proposition E.3.2 *Let $(\mathcal{P})$ be feasible and bounded. Whenever $\Delta b \in \operatorname{int} \mathbf{K}$ and $\Delta y \in \operatorname{int} \mathbf{K}^*$, the problem $(\mathcal{P}[b - \Delta b, c + A^*\Delta y])$ is strictly primal-dual feasible.*

Proof. Let $\bar{b} = b - \Delta b$ and $\bar{x} = x + A^*\Delta y$. Every feasible solution to $(\mathcal{P})$ clearly is a strictly feasible solution to $(\mathcal{P}[\bar{b}, c'])$ for all c', so that $(P[\bar{b}, c'])$ is strictly feasible for all c' (recall that $(\mathcal{P})$ is feasible). We claim that the problem $(\mathcal{P}[\bar{b}, c]$ cannot be robustly unbounded. Indeed, assume that $(P[\bar{b}, c'])$ is robustly unbounded. Then, by item D, (E.3.9) has a solution h. But such an h is a recessive direction of the (nonempty!) feasible domain of $(\mathcal{P})$, and along this direction the objective of $(\mathcal{P})$ strictly decreases, meaning that $(\mathcal{P})$ is unbounded, which in fact is not the case. Thus, $(\mathcal{P}[\bar{b},c])$ is not robustly unbounded. Further, $c\in C$. indeed, otherwise, by Linear Algebra, there would exist h such that $Ah=0$, $Rh=0$ and $c^Th<0$, which, as we just have seen, is impossible. The system of linear equations $A^*y+B^*r=d$ is solvable for all $d\in C$; by Linear Algebra, it admits a linear in $d\in C$ solution $(Y(d),R(d))$. Now let $r>0$ be such that $\Delta y-e\in\operatorname{int}\mathbf{K}^*$ whenever $\|e\|\leq r$. Recalling that $(P[\bar{b},c])$ is strictly feasible and $y(d)$ is linear, there exists $\epsilon>0$ such that $(P[b',c])$ is strictly feasible whenever $\|b'-\bar{b}\|\leq\epsilon$ and $\|Y(d)\|\leq r$ whenever $d\in C$ and $\|d\|_2\leq\epsilon$. Let us prove that Since $(\mathcal{P}[\bar{b},c]$ is not robustly unbounded, there exist b', $\|b'-\bar{b}\|\leq\epsilon$, and $c'\in C$ with $\|c'-c\|_2\leq\epsilon$ such that the program $(\mathcal{P}[b',c'])$ is not unbounded; since this program is strictly feasible, it should be bounded which, by Conic Duality Theorem, implies that $(\mathcal{D}[b',c'])$ is feasible. Thus, there exist $\bar{y}$ and $\bar{r}$ satisfying

$$\bar{y}\in\mathbf{K}^*, A^*\bar{y}+R^*\bar{z}=c'.$$

Now let us set

$$y^+=\bar{y}+\Delta y+Y(c-c'),\ z^+=\bar{z}+Z(c-c').$$

Observe that $\|Y(c-c')\|\leq r$ due to $\|c-c'\|\leq\epsilon$, whence $\Delta y+Y(c-c')\in\operatorname{int}\mathbf{K}^*$. Since $\bar{y}\in\mathbf{K}^*$, we conclude that $y^+\in\operatorname{int}\mathbf{K}^*$. At the same time

$$\begin{array}{l}A^*y^++R^*z^+=[A^*\bar{y}+R^*\bar{z}]+A\Delta y+[A^*Y(c-c')+R^*Z(c-c')]\\ =c'+(c-c')+A\Delta y=c'+[c-c']+A^*\Delta y=c+A^*\Delta y=\bar{c}.\end{array}$$

We see that (y^+,z^+) is a strictly feasible solution to $(\mathcal{D}[\bar{b},\bar{c}])$. Since by construction $(\mathcal{P}[\bar{b},\bar{c}])$ also is strictly feasible, $(\mathcal{P}[\bar{b},\bar{c}])$ is strictly primal-dual feasible. □

E.3.5 Sensitivity Analysis

The results we are about to present resemble those of Sensitivity Analysis for LP. Consider a primal-dual pair of cone program in the single-cone form:

$$\min_x\left\{c^Tx: Ax-b\in\mathbf{K}, Rx=r\right\}\qquad(\mathcal{P})$$

along with its dual problem

$$\max_{y,z}\left\{\langle b,y\rangle+\langle r,z\rangle: y\in\mathbf{K}^*, A^*y+R^*z=c\right\}\qquad(\mathcal{D})$$

In what follows, we treat the part of the data $A,b,\mathbf{K}$, R as fixed, and b,r,c – as varying, so that it makes sense to refer to $(\mathcal{P})$ as $(\mathcal{P}[b,r;c])$, and to its dual $(\mathcal{D})$ as to $(\mathcal{D}[b,r;c])$, and to denote the optimal value of $(\mathcal{P}[b,r;c])$ as $\mathrm{Opt}(b,r;c)$. Our goal is to explore the structure of the *cost function* $\mathrm{Opt}(b,r;c)$ as a function of (b,r), c being fixed, and of c, b,r being fixed.

E.3.5.1 The cost function as a function of c

Let $[b;r]$ be fixed at certain value $(\bar{b},\bar{r})$ such that $(\mathcal{P}[\bar{b},\bar{r};c])$ is feasible (this fact is independent of the value of c). An immediate observation is that in this case, the function $\mathrm{Opt}_{b,r}(c)=\mathrm{Opt}(b,r;c)$ is a concave function of c. Indeed, this is the infimum of the nonempty family $\{f_x(c)=c^Tx: Ax-b\in\mathbf{K}, Rx=r\}$ of linear (and thus concave) functions of c. A less trivial observation is as follows:

Proposition E.3.3 *Let $\bar{c}$ be such that $(\mathcal{P}[\bar{b},\bar{r},\bar{c})$ is solvable, and $\bar{x}$ be the corresponding optimal solution. Then $\bar{x}$ is a supergradient of $\mathrm{Opt}_{\bar{b},\bar{r}}(\cdot)$ at $\bar{c}$, meaning that*

$$\forall c:\mathrm{Opt}_{\bar{b},\bar{r}}(c)\leq\mathrm{Opt}_{\bar{b},\bar{r}}(\bar{c})+\bar{x}^T(c-\bar{c}).$$

Geometrically: the graph of $\mathrm{Opt}_{\bar{b},\bar{r}}(c)$ never goes above the graph of the affine function $\ell(c)=\mathrm{Opt}_{\bar{b},\bar{r}}(\bar{c})+\bar{x}^T(c-\bar{c})$ and touches this graph at the point $[\bar{c};\mathrm{Opt}_{\bar{b},\bar{r}}(\bar{c})]$ (and perhaps at other points as well).

Proof is immediate: since $\bar{x}$ is a feasible solution of $(\mathcal{P}[\bar{b},\bar{r};c])$ for every c, we have

$$\mathrm{Opt}_{\bar{b},\bar{r}}(c)\leq c^T\bar{x}=(c-\bar{c})^T\bar{x}+\bar{c}^T\bar{x}=\mathrm{Opt}_{\bar{b},\bar{r}}(\bar{c})+\bar{x}^T(c-\bar{c}).\qquad\square$$

Remark: The fact that a function is convex (or concave) implies, in particular, that the function possesses certain "regularity." e.g., the following is true:

Let f be a convex (or concave) function and X be a closed and bounded set belonging to the relative interior of function's domain. Then f is Lipschitz continuous on X: there exists $L < \infty$ such that

$$\forall (x, y \in X) : |f(x) - f(y)| \le L\|x - y\|.$$

E.3.5.2 The cost function as a function of (b, r)

Let now c be fixed at certain value $\bar{c}$, and assume that there exists $[b'; r']$ such that the problem $(\mathcal{P}[b', r'; \bar{c}])$ is strictly feasible and bounded. Then the dual problem $(\mathcal{D}[b', r'; \bar{c})$ is feasible (and even solvable) by Conic Duality Theorem, meaning that the duals to *all* problems $(\mathcal{D}[b, r; \bar{c})$ are feasible (since the feasible set of the dual is independent of b, r). By Weak duality, it follows that problems $(P[b, r; \bar{c}])$ are bounded, and thus the cost function $\mathrm{Opt}_{\bar{c}}(b, r) = \mathrm{Opt}(b, r; \bar{c})$ takes only real values and the value $+\infty$. It is easily seen that this function is convex.

Indeed, denoting for short $q = (b, r)$ and suppressing temporarily the subscript $_{\bar{c}}$, we should prove that $\mathrm{Opt}((1-\lambda)q + \lambda q') \le (1-\lambda)\mathrm{Opt}(q) + \lambda \mathrm{Opt}(q')$ for all q, q' and all $\lambda \in [0, 1]$. There is nothing to prove when $\lambda = 0$ or $\lambda = 1$; when $0 < \lambda < 1$, there is nothing to prove when $\mathrm{Opt}(q)$ or $\mathrm{Opt}(q')$ are infinite. Thus, we can restrict ourselves with the case $q = (b, r) \in \mathrm{dom}\mathrm{Opt}$, $q' = (b', r') \in \mathrm{dom}\mathrm{Opt}$ and $0 < \lambda < 1$. Given $\epsilon > 0$, we can find x and x' such that

$$\begin{array}{l} Ax - b \in \mathbf{K}, Rx = r, \bar{c}^T x \le \mathrm{Opt}(q) + \epsilon, \\ Ax' - b; \in \mathbf{K}, Rx' = r, \bar{c}^T x' \le \mathrm{Opt}(q') + \epsilon. \end{array}$$

Setting $\tilde{x} = (1-\lambda)x + \lambda x'$, $\tilde{q} = (1-\lambda)q + \lambda q'$, we have $A\tilde{x} - \tilde{q} = (1-\lambda)[Ax - b] + \lambda[Ax' - b'] \in \mathbf{K}$, where the inclusion follows from the fact that $\mathbf{K}$ is a cone; this, $\tilde{q}$ is a feasible solution for $(\mathcal{P}[\tilde{q}, \bar{c}])$. We also have

$$\begin{array}{l} \bar{c}^T \tilde{q} = (1-\lambda)\bar{c}^T x + \lambda \bar{c}^T y \le (1-\lambda)[\mathrm{Opt}(q) + \epsilon] + \lambda[\mathrm{Opt}(q') + \epsilon] \\ = (1-\lambda)\mathrm{Opt}(q) + \lambda \mathrm{Opt}(q') + \epsilon]. \end{array}$$

Since $\tilde{q}$ is feasible for $(\mathcal{P}[\tilde{q}, \bar{c}])$, we have

$$\mathrm{Opt}(\tilde{q}) \le \bar{c}^T \tilde{q} \le (1-\lambda)\mathrm{Opt}(q) + \lambda \mathrm{Opt}(q') + \epsilon$$

The resulting inequality holds true got every $\epsilon > 0$, whence

$$\mathrm{Opt}(\tilde{q}) \le (1-\lambda)\mathrm{Opt}(q) + \lambda \mathrm{Opt}(q'),$$

which completes the proof of convexity of $\mathrm{Opt}(\cdot)$.

We have the following analogy of Proposition E.3.3:

Proposition E.3.4 *Let $\bar{b}, \bar{r}$ be such that $(\mathcal{P}[\bar{b}, \bar{r}, \bar{c}])$ is strictly feasible. Then the dual problem $(\mathcal{D}[\bar{b}, \bar{r}; \bar{c}])$ is solvable, and every optimal solution $(\bar{y}, \bar{r})$ to the latter program is a subgradient of the convex function $\mathrm{Opt}_{\bar{c}}(\cdot)$ at the point $(\bar{b}, \bar{r})$, meaning that*

$$\forall (b, r) : \mathrm{Opt}_{\bar{c}}(b, r) \ge \mathrm{Opt}_{\bar{c}}(\bar{b}, \bar{r}) + \langle \bar{y}, b - \bar{b}\rangle + \langle \bar{z}, r - \bar{r}\rangle.$$

Geometrically: the graph of $\mathrm{Opt}_{\bar{c}}(b, r)$ never goes below the graph of the affine function $\ell(b, r) = \mathrm{Opt}_{\bar{c}}(\bar{b}, \bar{r}) + \langle \bar{y}, b - \bar{b}\rangle + \langle \bar{z}, r - \bar{r}\rangle$ and touches this graph at the point $(\bar{b}, \bar{r}; \mathrm{Opt}_{\bar{c}}(\bar{b}, \bar{r}))$ (and perhaps at other points as well).

Proof is immediate. As we have already mentioned, our choice of $\bar{c}$ ensures that $(\mathcal{D}[\bar{b}, \bar{r}; \bar{c}])$ is feasible, and thus the program $(\mathcal{P}[\bar{b}, \bar{r}, \bar{c})$ is bounded; since by assumption the latter program is strictly feasible, the Conic Duality Theorem says that $(\mathcal{D}[\bar{b}, \bar{r}; \bar{c}])$ is solvable, which is the first claim in the Proposition. Now let $(\bar{y}, \bar{r})$ be an optimal solution to $(\mathcal{D}[\bar{b}, \bar{r}; \bar{c}])$. Then for every feasible solution x to $(\mathcal{P}[b, r; \bar{c}])$ we have

$$\begin{array}{l} \bar{c}^T x = [A^* \bar{y} + R^* \bar{z}]^T x = \langle \bar{y}, Ax - b\rangle + \langle \bar{z}, Rx\rangle + \langle \bar{y}, b\rangle \ge \langle \bar{y}, b\rangle + \langle \bar{z}, r\rangle \\ = \langle \bar{y}, b - \bar{b}\rangle + \langle \bar{z}, r - \bar{r}\rangle + \underbrace{\langle \bar{y}, \bar{b}\rangle + \langle \bar{z}, \bar{r}\rangle}_{=\mathrm{Opt}_{\bar{c}}(\bar{b}, \bar{r})}; \end{array}$$

Since the resulting inequality is valid for all feasible solutions x to $(\mathcal{P}[b, r; \bar{c}])$, we conclude that

$$\mathrm{Opt}_{\bar{c}}(b, r) \ge \mathrm{Opt}_{\bar{c}}(\bar{b}, \bar{r}) + \langle \bar{y}, b - \bar{b}\rangle + \langle \bar{z}, r - \bar{r}\rangle. \qquad \square$$

E.3.6 Geometry of Primal-Dual Pair of Conic Problems

We are about to derive geometric interpretation of a primal-dual pair (P), (D) of conic programs. As was explained in the beginning of Section E.2, we lose nothing when assuming that the primal program is a single-cone one and that the space E of the primal decision vectors is $\mathbf{R}^n$, so that the primal program reads:

$$\mathrm{Opt}(P)=\min_x\left\{c^Tx : Ax-b\in\mathbf{K}, Rx=r\right\} \tag{P}$$

where $x\mapsto Rx$ is a linear mapping from $\mathbf{R}^n$ to Euclidean space F, and $\mathbf{K}$ is a regular cone in a Euclidean space F_1 which, for aesthetical reasons (we do not need index anymore!), we now denote H. The dual program now reads

$$\mathrm{Opt}(D)=\max_{z,y}\left\{\langle r,z\rangle+\langle b,y\rangle : y\in\mathbf{K}^*, A^*y+R^*z=c\right\} \tag{D}$$

(to save notation, we skip the indices in $\langle\cdot,\cdot\rangle$). Assume that the systems of linear equality constraints in (P) and in (D) are solvable, and let $\bar x$ and $[\bar y;\bar z]$ be solutions to these systems:

$$\begin{array}{ll}(a) & R\bar x=r\\ (b) & A^*\bar y+R^*\bar z=c.\end{array} \tag{E.3.10}$$

Let us express (P) in terms of the primal slack $\xi=Ax-b\in H$. The constraints of (P) say that this vector should belong to the intersection of $\mathbf{K}$ and the *primal feasible plane* $\mathcal{M}_P$ which is the image of the affine plane $\{x:Rx=r\}$ in the x-space under the affine mapping $x\mapsto Ax-b$. The linear subspace $\mathcal{L}_P$ in E which is parallel to $\mathcal{M}_P$ is $\mathcal{L}_P=\{\xi=Ax:Rx=0\}$, and we can take the point $A\bar x-b:=-\bar\xi$ as the shift vector for $\mathcal{M}_P$. Thus,

$$\mathcal{M}_P=\mathcal{L}_P-\bar\xi,\ \bar\xi=b-A\bar x,\ \mathcal{L}_P=\{\xi=Ax:Rx=0\}. \tag{E.3.11}$$

Now let us express the primal objective in terms of the primal slack. Given x satisfying the equality constraints in (P), we have (F is the destination space of the mapping $x\mapsto Rx$):

$$\begin{array}{l}c^Tx=[A^*\bar y+R^*\bar z]^Tx=[A^*\bar y]^Tx+[R^*\bar z]^Tx=\langle\bar y,Ax\rangle+\langle Rx,\bar z\langle\\ =\langle\bar y,\underbrace{Ax-b}_{\xi}\rangle+\mathrm{const}_P,\ \mathrm{const}_P=\langle\bar y,b\rangle+\langle r,\bar z\rangle.\end{array}$$

We have arrived at the following intermediate conclusion:

Program (P) can be reduced to the program

$$\begin{array}{c}\mathrm{Opt}(P)=\min\limits_{\xi\in H}\left\{\langle\bar y,\xi\rangle:\xi\in\mathbf{K}\cap\mathcal{M}_P\right\}\\ \left[\begin{array}{rcl}\mathcal{M}_P&=&\mathcal{L}_P-\bar\xi:=b-A\bar x\\ \mathcal{L}_P&=&\{\xi=Ax:Rx=0\}\\ \mathrm{Opt}(P)&=&\mathrm{Opt}(P)+\langle\bar y,b\rangle+\langle r,\bar z\rangle.\end{array}\right]\end{array} \tag{P}$$

Now let us process in a similar fashion the dual program (D), specifically, express it in terms of the vector y. The constraints of (D) say that this vector should belong to the intersection of $\mathbf{K}^*$ and the *dual feasible plane* $\mathcal{M}_D=\{y:\exists z:A^*y+R^*z=c\}$. This plane is parallel to the linear subspace $\mathcal{L}_D=\{y:\exists z:A^*y+R^*z=0\}$, and as a shift vector for $\mathcal{M}_D$ we can take $\bar y\in\mathcal{M}_D$. It remains to express the dual objective in terms of y. To this end note that if $[y;z]$ satisfies the linear equality constraints of (D), then

$$\begin{array}{l}\langle r,z\rangle+\langle b,y\rangle=\langle R\bar x,z\langle+\langle b,y\rangle=\bar x^T[R^*z]+\langle b,y\rangle\\ =\bar x^T[c-A^*y]+\langle b,y\rangle=\bar x^Tc+\langle b-A\bar x,y\rangle=\langle\bar\xi,y\rangle+\mathrm{const}_D,\ \mathrm{const}_D=c^T\bar x.\end{array}$$

We have arrived at the following conclusion:

Program (D) can be reduced to the program

$$\begin{array}{c}\mathrm{Opt}(D)=\max\limits_{y\in H}\left\{\langle\bar{xi},y\rangle:y\in\mathbf{K}_*\cap\mathcal{M}_D\right\}\\ \left[\begin{array}{rcl}\mathcal{M}_D&=&\mathcal{L}_D+\bar y:=b-A\bar x\\ \mathcal{L}_D&=&\{y:\exists z:A^*y+R^*z=0\}\\ \mathrm{Opt}(D)&=&\mathrm{Opt}(D)+c^T\bar x.\end{array}\right]\end{array} \tag{D}$$

Now, same as in the LP case, $\mathcal{L}_D$ is just the orthogonal complement of $\mathcal{L}_P$. Indeed, $h \in (\mathcal{L}_P)^\perp$ iff $\langle h, Ax\rangle = 0$ whenever $Rx = 0$, that is, iff the linear equation $x^T[A^*h] = 0$ in variables x is a consequence of the linear system $Rx = 0$, which is the case iff $A^*h = R^*w$ for some w, which, after substitution $w = -z$, is nothing but the characterization of $\mathcal{L}_D$.

Finally, let us compute the duality gap at a pair $(x, [y, z])$ of candidate solutions satisfying the equality constraints in (P) and (D):

$$\begin{array}{l} c^Tx - \langle b, y\rangle - \langle r, z\rangle = [A^*y + R^*z]^Tx - \langle b, y\rangle - \langle r, z\rangle \\ = \langle Ax - b, y\rangle + \langle Rx - r, z\rangle = \langle Ax - b, y\rangle. \end{array}$$

Putting things together, we arrive at a perfectly symmetric purely geometric description of (P), (D):

> *Assume that the systems of linear constraints in (P) and (D) are solvable. Then the primal-dual pair (P), (D) of conic problems reduces to the following geometric problem. We are given*
> - *two dual to each other cones $\mathbf{K}$, $\mathbf{K}_*$ in a Euclidean space H,*
> - *a pair of linear subspaces $\mathcal{L}_P$, cL_D in H which are orthogonal complements to each other, and*
> - *a pair of shift vectors $\bar{\xi}, \bar{y}$ in H.*
>
> *These geometric data define affine subspaces $\mathcal{M}_P = \mathcal{L}_P - \bar{\xi}$, $\mathcal{M}_D = \mathcal{L}_D + \bar{y}$.*
>
> *The primal problem (P) reduces to minimizing the linear form $\langle \bar{y}, \cdot\rangle$ over the intersection of the primal feasible plane $\mathcal{M}_P$ and the cone $\mathbf{K}$, which is the primal feasible set; the dual problem (D) reduces to maximizing the linear form $\langle \bar{\xi}, \cdot\rangle$ over the intersection of the dual feasible plane $\mathcal{M}_D$ and the cone $\mathcal{K}^*$, which is the dual feasible set. Given feasible solutions ξ, y to these geometric problems, the corresponding duality gap is the inner product of the solutions.*
>
> *Strict feasibility of a problem from our primal-dual pair means that the corresponding feasible plane intersects the interior of the corresponding cone. Whenever both problems are strictly feasible, the minimal value of the duality gap is zero, and the duality gap, as evaluated at a pair of primal and dual feasible solutions, is the sum of their non-optimalities, in terms of the objectives of the respective problems. Under the same assumption of primal-dual strict feasibility, pairs of optimal solutions to the respective problems are exactly the pairs of orthogonal to each other primal and dual feasible solutions, and these pairs do exist.*

We see that geometrically, a primal-dual pair of conic problems looks completely similar to a pair of primal-dual LP programs: in both situations (in the second — under additional assumption that both problems are strictly feasible) we are looking for pairs of orthogonal to each other vectors with one member of the pair belonging to the intersection of "primal" affine plane and "primal" cone, and the other member belonging to the intersection of the "dual" affine plane and the "dual" cone.

The pair of primal and dual affine planes cannot be arbitrary: they should be shifts of linear subspaces which are orthogonal complements to each other. Similarly, the pair of cones in question are "rigidly connected" to each other — they are duals of each other. In LP, the underlying cone is the nonnegative orthant and thus is *self-dual*, this is why in the LP case we do not see *two* cones, just one of them.

We complete this Section by mentioning that the choice of the shift vectors for $\mathcal{M}_P$, $\mathcal{M}_D$ (or, which is the same, the objectives in (P) and (D)) is immaterial: when replacing the above $\bar{\xi}$ with any other vector from the minus primal feasible plane $[-\mathcal{M}_P]$, the primal problem (P) clearly remains intact, and the dual objective $\langle \bar{\xi}^T, \cdot\rangle$, *restricted on the dual feasible plane*, changes by an additive constant, which affects nothing but the optimal value $\mathrm{Opt}(D)$. By similar reasons, replacing $\bar{y}$ with any other vector from $\mathcal{M}_D$ keeps (D) intact and changes by additive constant the restriction of the primal objective $\langle \bar{y}, \cdot\rangle$ on the primal feasible plane.

E.4 Conic Representations of Sets and Functions

It is easily seen that every convex program $\min\limits_{x\in X} f(x)$ ($f : \mathbf{R}^n \to \mathbf{R}$ is convex, $X \subset \mathbf{R}^n$ is convex and closed) can be equivalently reformulated as a conic program. This fact is of no actual use, since a general-type cone is not simpler than a general-type closed convex set. What indeed is important, is to recognize when a given convex program can be posed as a conic program *from a given family*, primarily — when it can be posed as an LP/CQP/SDP program. To this end we can develop an approach completely similar to the one we used in Section B.2.4.4. Specifically, assume we are given a family $\mathcal{K}$ of regular cones, every one of them "living" in its own Euclidean space. It makes sense to assume also that the family contains the nonnegative ray and is closed w.r.t. taking finite direct products and to passing from a cone to its dual cone. The most important for us examples are:

- the family $\mathcal{LP}$ of nonnegative orthants; this family underlies LP.
- the family $\mathcal{CQP}$ of finite direct products of Lorentz cones; this family underlies CQP (Conic Quadratic Programming).
  Note that $\mathcal{CQP}$ contains $\mathbf{R}_+ = \mathbf{L}^1$; the fact that all other requirements are satisfied is evident (recall that the Lorentz cones are self-dual).
- the family $\mathcal{SDP}$ of finite direct products of semidefinite cones; this family underlies SDP (Semidefinite Programming).
  Note that $\mathcal{SDP}$ contains $\mathbf{R}_+ = \mathbf{S}^1_+$, and satisfies all other requirements by exactly the same reasons as $\mathcal{CQP}$.

Now, given a family $\mathcal{K}$, we call a set $X \subset \mathbf{R}^n$ *$\mathcal{K}$-representable*, if it can be represented in the form

$$X = \{x \in \mathbf{R}^n : \exists w : Px + Qw + r \in \mathbf{K}\} \tag{†}$$

where $\mathbf{K}$ is a cone from $\mathcal{K}$; corresponding data $(P, Q, r, \mathbf{K})$, same as the representation itself, are called *$\mathcal{K}$-representation* of X ($\mathcal{K}$-r. of X for short). note that this definition mirrors the definition of a polyhedral representation of a set (which in our now language becomes $\mathcal{LP}$-representation). Completely similar to the polyhedral case, given a $\mathcal{K}$-representation (†) of X, we can immediately rewrite the problem of minimizing a linear objective c^Tx over X as a conic program on the cone from the family $\mathcal{K}$, specifically, the program

$$\min_{x,w} \left\{c^Tx : Px + Qw + r \in \mathbf{K}\right\}.$$

Bearing in mind this observation, we understand why it is important to build a calculus of $\mathcal{K}$-representable sets *and functions*. A $\mathcal{K}$-r. of a function f is, by definition, the same as $\mathcal{K}$-r. of its epigraph, and a function is called $\mathcal{K}$-representable, if it admits a $\mathcal{K}$-r. Same as in the polyhedral case, a $\mathcal{K}$-r.

$$\{[x;\tau] : \tau \geq f(x)\} = \{[x;\tau] : \exists w : Px + \tau p + Qw + r \in \mathbf{K}\}$$

of a function f implies $\mathcal{K}$-r.'s of the level sets of the function:

$$\{x : a \geq f(x)\} = \{x : \exists w : Px + ap + Qw + r \in \mathbf{K}\}.$$

The basic calculus rules from Section B.2.4.4 extend word by word from the polyhedral representability to $\mathcal{K}$-representability; there are just two facts which, on a close inspection, are responsible for the validity of the calculus rules:

• the fact that, given two systems of linear inequalities (that is, two component-wise vector inequalities $Ax \leq b$ and $Cx \leq d$), we can put them together, thus getting a single vector inequality $[A; B]x \leq [c; d]$. What underlies this fact, is the closedness of the associated family of cones (in the case of polyhedral representability, the family of nonnegative orthants) w.r.t. taking direct products. Since we have required from $\mathcal{K}$ to possess the latter property, putting together two conic inequalities $Ax - b \in \mathbf{K}$ and $A'x - b' \in \mathbf{K}'$ with $\mathbf{K}, \mathbf{K}' \in \mathcal{K}$ results in a single conic inequality $[Ax; A'x] - [b; b'] \in \mathbf{K} \times \mathbf{K}'$ involving a cone from $\mathcal{K}$.

• the fact that the feasible sets of finite systems of linear inequalities/equations are polyhedrally representable. Since we require from $\mathcal{K}$ to contain rays and to be closed w.r.t. taking direct products, these sets and $\mathcal{K}$-representable as well.

Extending more advanced rules of calculus of polyhedral representability to the case of $\mathcal{K}$-representability requires certain care. For example,

1. Assume that an LP program $\min\limits_x\{c^Tx : Ax \geq b, Rx = r\}$ is feasible and bounded for some value of $[b; r]$; under this assumption the function $\mathrm{Opt}([b; r]) = \min\limits_x\{c^Tx : Ax \geq b, Rx = r\}$ is convex and polyhedrally representable:

$$\{[b; r; \tau] : \tau \geq \mathrm{Opt}([b; r])\} = \left\{[b; r; \tau] : \exists x : Ax - b \geq 0, Bx = r, c^Tx \leq \tau\right\}. \tag{‡}$$

Now let us pass from the optimal value in a LP program to the one in a conic program: Passing to the optimal value of a $\mathcal{K}$-conic problem

$$\mathrm{Opt}([b;r]) = \inf\left\{c^T x : Ax - b \in \mathbf{K}, Bx = r\right\}. \tag{E.4.1}$$

Assume that the program is strictly feasible and bounded for some value of $[b;r]$. Then the dual program is feasible, and since the latter fact is independent of $[b;r]$, we conclude from Weak duality that (E.4.1) is bounded for all values of $[b;r]$, so that the cost function $\mathrm{Opt}([b;r])$ takes only real values and perhaps the value $+\infty$. You can easily verify that the cost function is convex. Now, the "literal analogy" of (‡) would be

$$\{[b;r;\tau] : \tau \geq \mathrm{Opt}([b;r])\} = \left\{[b;r;\tau] : \exists x : Ax - b \in \mathbf{K}, Bx = r, c^T x \leq \tau\right\},$$

but this relation is not necessarily true: $\mathrm{Opt}([b;r])$ can be finite and non-achievable, meaning that the fact that the point $[b;r;\tau]$ with $\tau = \mathrm{Opt}([b;r])$ is in the epigraph of Opt cannot be "certified" by any x.

The correct version of (‡) is as follows:

> *Let* (E.4.1) *be bounded and strictly feasible for some value of* $[b;r]$. *Then the cost function* $\mathrm{Opt}([b;r])$ *is a convex function, and the* $\mathcal{K}$*-r. set*
>
> $$\mathcal{G} = \left\{[b;r;\tau] : \exists x : Ax - b \in \mathbf{K}, Rx = r, c^T x \leq \tau\right\}.$$
>
> *is in-between the epigraph*
>
> $$\mathrm{Epic}(\mathrm{Opt}(\cdot)) = \{[b;r;\tau] : \tau \geq \mathrm{Opt}([b;r])\}$$
>
> *of the cost function and the "strictly upper part"*
>
> $$\mathrm{Epic}^+(\mathrm{Opt}(\cdot)) = \{[b;r;\tau] : \tau > \mathrm{Opt}([b;r])\}$$
>
> *of this epigraph:*
>
> $$\mathrm{Epic}^+(\mathrm{Opt}(\cdot)) \subset \mathcal{G} \subset \mathrm{Epic}(\mathrm{Opt}(\cdot)).$$

2. The support function of a nonempty polyhedral set is polyhedrally representable, meaning that its epigraph admits a p.r. readily given by a p.r. of the set (see [127, Section 3.3]). To get similar result in the general conic case, we need strict feasibility of the representation of the set. The precise statement reads:

> *Let* X *be a nonempty* $\mathcal{K}$*-representable set given by a* $\mathcal{K}$*-representation*
>
> $$X = \{x : \exists w : Px + Qw - b \in \mathbf{K},\ Rx + Sw = r\} \qquad [\mathbf{K} \in \mathcal{K}]$$
>
> *which is strictly feasible, meaning that there exists* $\bar{x}, \bar{w}$ *such that*
>
> $$P\bar{x} + Q\bar{w} - b \in \mathrm{int}\,\mathbf{K},\ R\bar{x} + S\bar{w} = r.$$
>
> *Then the support function*
>
> $$\mathrm{Supp}(\xi) = \sup_{x \in X} \xi^T x$$
>
> *of the set* X *admits the explicit* $\mathcal{K}$*-representation*
>
> $$\begin{array}{l}\{[\xi;\tau] : \tau \geq \mathrm{Supp}(\xi)\} \\ = \left\{[\xi;\tau] : \exists \lambda, \mu : \begin{array}{l}\lambda \in \mathbf{K}^*, P^*\lambda + R^*\mu + \xi = 0, Q^*\lambda + S^*\mu = 0, \\ \langle b, \lambda\rangle + \langle r, \mu\rangle + \tau \geq 0\end{array}\right\}.\end{array} \tag{E.4.2}$$
>
> Note that (E.4.2) indeed is a $\mathcal{K}$-representation of the support function, since $\mathcal{K}$ is closed w.r.t. passing from a cone to its dual.

Indeed, $-\mathrm{Supp}(\xi)$ is the optimal value in the $\mathcal{K}$-conic program

$$\min_{x,w}\left\{-\xi^T x : Px + Qw - b \in \mathbf{K},\ Rx + Sw = r\right\}. \tag{$*$}$$

The latter problem is strictly feasible; thus, for a real τ, we have $\tau \geq \mathrm{Supp}(\xi)$ iff the latter problem is bounded with the optimal value $\geq -\tau$, which, by Conic Duality Theorem, is the case iff the conic dual of $(*)$ admits a feasible solution with the optimal value $\geq -\tau$, that is, iff

$$\exists \lambda, \mu : \lambda \in \mathbf{K}^*, P^*\lambda + R^*\mu = -\xi, Q^*\lambda + S^*\mu = 0, \langle b, \lambda\rangle + \langle r, \mu\rangle \geq -\tau,$$

and (E.4.2) follows.

E.4.1 Expressive abilities of $\mathcal{CQP}$ and $\mathcal{SDP}$

We have seen that the "rule" part of the calculus of $\mathcal{K}$-representable sets and functions remains intact (and in its "advanced" parts is even slightly weaker) than the calculus of polyhedral representability, see Section B.2.4.4 . What extends *dramatically* when passing from LP to CQP and especially SDP, is the spectrum of "raw materials" of the calculus, that is, "elementary" $\mathcal{CQP}$- and $\mathcal{SDP}$-representable functions and sets. With slight exaggeration, one can say that "for all practical purposes," *all* computationally tractable convex sets and functions arising in applications are $\mathcal{SDP}$-representable, so that all "real life" convex problems are within the grasp of SDP[7] We omit the list of "raw materials" for the calculus of $\mathcal{CQP}$- and $\mathcal{SDP}$-representable functions/sets; such a list can be found in [14]. Here we restrict ourselves with a single "advertising example:" the messy and highly nonlinear optimization program

	$\text{minimize} \sum\limits_{\ell=1}^{n} x_\ell^2$
(a)	$x \geq 0;$
(b)	$a_\ell^T x \leq b_\ell,\ \ell = 1, ..., n;$
(c)	$\Vert Px - p\Vert_2 \leq c^T x + d;$
(d)	$x_\ell^{\frac{\ell+1}{\ell}} \leq e_\ell^T x + f_\ell,\ \ell = 1, ..., n;$
(e)	$x_\ell^{\frac{\ell}{\ell+3}} x_{\ell+1}^{\frac{1}{\ell+3}} \geq g_\ell^T x + h_\ell,\ \ell = 1, ..., n-1;$
(f)	$\text{Det} \begin{bmatrix} x_1 & x_2 & x_3 & \cdots & x_n \\ x_2 & x_1 & x_2 & \cdots & x_{n-1} \\ x_3 & x_2 & x_1 & \cdots & x_{n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_n & x_{n-1} & x_{n-2} & \cdots & x_1 \end{bmatrix} \geq 1;$
(g)	$1 \leq \sum\limits_{\ell=1}^{n} x_\ell \cos(\ell\omega) \leq 1 + \sin^2(5\omega)\ \forall \omega \in \left[-\frac{\pi}{7}, 1.3\right]$

can be converted *in a systematic way* into a semidefinite program; omitting the constraints (f) and (g), the problem can be systematically converted into a CQP (and thus solving it within accuracy ϵ can be reduced in polynomial time to a similar problem for an LP program, see [127, Section 1.3]).

Remark: In the case of polyhedral representability, ignoring the "compactness" of a representation, we always can avoid slack variables: if a set in $\mathbf{R}^n$ admits a polyhedral representation, it always is polyhedral, i.e., can be represented as the solution set of a system of linear inequalities in the "original" variables – the coordinates of a vector sunning through $\mathbf{R}^n$. In the non-polyhedral case, using slack variables in representations of sets and functions is a must. For example, take the epigraph of the univariate function x^4; this set is conic quadratic representable:

$$G := \{[x, \tau] \in \mathbf{R}^2 : \tau \geq x^4\} = \{[x; \tau] : \exists w \in \mathbf{R} : \underbrace{\|[2x; w-1]\|_2 \leq w+1}_{(a)}, \underbrace{\|[2w; \tau-1]\|_2 \leq \tau+1}_{(b)}\}.$$

Indeed, (a) says that $w \geq x^2$, and (b) says that $\tau \geq w^2$; what these inequalities say about τ, x is exactly $\tau \geq x^4$. On the other hand, assume that we managed to find a conic quadratic representation of the same set without slack variables:

$$G = \{[x; \tau] : \|xa_i + \tau b_i + c_i\|_2 \leq \alpha_i x + \beta_i \tau + \gamma_i,\ 1 \leq i \leq m, xp + \tau q = r\}. \qquad (\flat)$$

Observe, first, that the system of linear equations should be trivial: $p = q = r = 0$. Indeed, otherwise these equations would cut off the 2D plane of x and τ a line or a point containing G, which clearly is impossible. Now, the sets of the form $\{[x; \tau] : \|xa_i + \tau b_i + c_i\|_2 \leq \alpha_i x + \beta_i \tau + \gamma_i\}$ are convex sets representable as intersections of solutions sets of quadratic inequalities

$$\|xa_i + \tau b_i + \gamma_i\|_2^2 \leq [\alpha_i x + \beta_i \tau + \gamma_i]^2$$

[7] Of course, whatever be a family $\mathcal{K}$ of cones, the $\mathcal{K}$-representable sets and functions are convex (why?), so that Conic Programming stays within the boundaries of Convex Optimization.

with half-planes (or the entire 2D planes) $\{[x;\tau] : \alpha_i x + \beta_i \tau + \gamma_i \geq 0\}$. A set of this type is bounded by finitely many "arcs," every one of them being either a line segment (including rays and entire lines), or parts of ellipses/parabolas/hyperbolas, and thus the right hand side set in ($\flat$) is bounded by finitely many arcs of the same type. But such an arc, as it is easily seen, can intersect the true boundary of G – the curve given by $t = x^4$ – only in finitely many points, so that a finite number of the arcs cannot cover the curve. The conclusion is that G cannot represent by conic quadratic inequalities in variables x, τ only.

Relations between LP, CQP and SDP. Clearly, polyhedral representations of sets and functions are their $\mathcal{CQP}$- and $\mathcal{SDP}$-representations as well — recall that the nonnegative ray is the same as the one-dimensional Lorentz and one-dimensional semidefinite cones, so that nonnegative orthants "sit" in $\mathcal{CQP}$ and $\mathcal{SDP}$, as a result, an LP program can be straightforwardly converted into a conic quadratic and into a semidefinite program. For example, the "single-cone" semidefinite reformulation of LP program

$$\min_x \left\{ c^T x : Ax \geq b, Rx = r \right\} \tag{$*$}$$

is as follows: keep the objective and the linear equality constraints as they are, and put the entries of the m-dimensional vector $Ax - b$ on the diagonal of a diagonal $m \times m$ matrix $\mathcal{A}(x)$ which, of course, will depend affinely on x. Since a diagonal matrix is symmetric and is positive semidefinite iff its diagonal entries are nonnegative, $(*)$ is equivalent to the SDP program

$$\min_x \left\{ c^T x : \mathcal{A}(x) \succeq 0, Rx = r \right\}.$$

A less trivial, but still simple, observation is that *conic quadratic representable sets/functions are semidefinite representable as well, with semidefinite representations readily given by conic quadratic ones.* The reason is that a Lorentz cone $\mathcal{L}^n$ is $\mathcal{SDP}$-representable – it is just the intersection of the semidefinite cone $\mathbf{S}^n_+$ and an appropriate linear subspace of $\mathbf{S}^n$ (this is completely similar to the fact that the nonnegative orthant $\mathbf{R}^n_+$ is the intersection of $\mathbf{S}^n_+$ and the subspace of $n \times n$ diagonal matrices). Specifically, given a vector $x \in \mathbf{R}^n$, let us build the $n \times n$ symmetric matrix

$$\mathrm{Arrow}(x) = \left[\begin{array}{c|cccc} x_n & x_1 & x_2 & \cdots & x_n \\ \hline x_2 & x_1 & & & \\ x_3 & & x_1 & & \\ \vdots & & & \ddots & \\ x_n & & & & x_1 \end{array}\right]$$

(blanks are filled with zeros).

Lemma E.4.1 *Let $x \in \mathbf{R}^n$. The matrix* $\mathrm{Arrow}(x)$ *is positive semidefinite iff $x \in \mathbf{L}^n$. As a result, a conic quadratic representation of a set*

$$X = \{x : \exists w : A_i x + B_i w + b_i \in \mathcal{L}^{n_i}, \, 1 \leq i \leq m\}$$

can be converted into a semidefinite representation of the same set, specifically, the representation

$$X = \{x : \exists w : \mathrm{Arrow}(A_i x + B_i w + b_i) \succeq 0, \, 1 \leq i \leq m\}.$$

Proof. All we need is to prove the equivalence $x \in \mathbf{L}^n \Leftrightarrow \mathrm{Arrow}(x) \succeq 0$. In one direction: assume that $x \in \mathbf{L}^n$, and let us prove that $\mathrm{Arrow}(x) \succeq 0$. The symmetry of $\mathrm{Arrow}(x)$ is evident. To verify positive semidefiniteness, we should prove that $h^T \mathrm{Arrow}(x) h \geq 0$ for all $h \in \mathbf{R}^n$. Partitioning $h = [g; t]$ with scalar t, and denoting $y = [x_1; ...; x_{n-1}]$, we have $x_n \geq \|y\|_2$ due to $x \in \mathbf{L}^n$, whence

$$\begin{array}{l} h^T \mathrm{Arrow}(x) h = x_n(t^2 + g^T g) + 2t y^T g \geq x_n(t^2 + g^T g) - 2|t| \|y\|_2 \|g\|_2 \geq x_n[t^2 + g^T g - 2|t| \|g\|_2] \\ = x_n(|t| - \|g\|_2)^2 \geq 0. \end{array}$$

In the opposite direction: let $x \in \mathbf{R}^n$ and $\mathrm{Arrow}(x) \succeq 0$, and let us prove that $x \in \mathbf{L}^n$. The statement is evident when $n = 1$, so let $n > 1$. Setting $x = [y; x_n]$, let $h = [g; 1]$, where g is the unit vector such that $g^T y = -\|y\|_2$. Then

$$0 \leq h^T \mathrm{Arrow}(x) h = x_n \|h\|_2^2 + 2 \cdot 1 \cdot g^T y = 2x_n - 2\|y\|_2 = 2[x_n - \|y\|_2].$$

We see that $x_n \geq \|y\|_2$, meaning that $x \in \mathbf{L}^n$. □

Remark. The possibility to convert straightforwardly LP and CQP to Semidefinite Programming does not mean that this is the best way to solve LP's and CQP's in actual computations. Two former problems are somehow simpler than the latter one, and dedicated LP and CQP solvers available today in commercial packages can solve linear and conic quadratic programs much faster, and in a much wider range of sizes, than "universal" SDP solvers. This being said, when solving "moderate size" LP's and CQP's (what is "moderate," it depends on "fine structure" of a program being solved and may vary from few hundreds to few thousands of variables), it is very attractive to reduce everything to SDP and thus to use a single solver. This idea is implemented in the `cvx` package[8] which uses the calculus of the semidefinite representable sets and functions to convert the input "high level" description of a problem into its "inner" SDP-reformulation which then is forwarded to an SDP solver. The input description of a problem utilizes full capabilities of MATLAB and thus is incredibly transparent and easy to use, making `cvx` an ideal tool for a classroom (and not only for it).

[8] "CVX: Matlab Software for Disciplined Convex Programming," Michael Grant and Stephen Boyd, http://www.stanford.edu/~boyd/cvx/

Index