STATISTICAL INFERENCE VIA CONVEX OPTIMIZATION
STATISTICAL INFERENCE VIA CONVEX OPTIMIZATION

Anatoli Juditsky
Université Grenoble Alpes

Arkadi Nemirovski
Georgia Institute of Technology
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of Figures</td>
<td>xiii</td>
</tr>
<tr>
<td>Preface</td>
<td>xv</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>xix</td>
</tr>
<tr>
<td>Notational conventions</td>
<td>xxi</td>
</tr>
<tr>
<td>About proofs</td>
<td>xxiii</td>
</tr>
<tr>
<td>On computational tractability</td>
<td>xxiii</td>
</tr>
<tr>
<td><strong>1 Sparse Recovery via $\ell_1$ Minimization</strong></td>
<td>1</td>
</tr>
<tr>
<td>1.1 Compressed Sensing: What it is about?</td>
<td>1</td>
</tr>
<tr>
<td>1.1.1 Signal Recovery Problem</td>
<td>1</td>
</tr>
<tr>
<td>1.1.2 Signal Recovery: parametric and non-parametric cases</td>
<td>2</td>
</tr>
<tr>
<td>1.1.3 Compressed Sensing via $\ell_1$ minimization: Motivation</td>
<td>6</td>
</tr>
<tr>
<td>1.1.3.1 Preliminaries</td>
<td>6</td>
</tr>
<tr>
<td>1.2 Validity of sparse signal recovery via $\ell_1$ minimization</td>
<td>8</td>
</tr>
<tr>
<td>1.2.1 Validity of $\ell_1$ minimization in the noiseless case</td>
<td>8</td>
</tr>
<tr>
<td>1.2.1.1 Notational convention</td>
<td>8</td>
</tr>
<tr>
<td>1.2.1.2 $s$-Goodness</td>
<td>9</td>
</tr>
<tr>
<td>1.2.1.3 Nullspace property</td>
<td>10</td>
</tr>
<tr>
<td>1.2.2 Imperfect $\ell_1$ minimization</td>
<td>11</td>
</tr>
<tr>
<td>1.2.2.1 Contrast matrices and quantifications of Nullspace property</td>
<td>11</td>
</tr>
<tr>
<td>1.2.3 Regular $\ell_1$ recovery</td>
<td>13</td>
</tr>
<tr>
<td>1.2.4 Penalized $\ell_1$ recovery</td>
<td>13</td>
</tr>
<tr>
<td>1.2.5 Discussion</td>
<td>14</td>
</tr>
<tr>
<td>1.3 Verifiability and tractability issues</td>
<td>17</td>
</tr>
<tr>
<td>1.3.1 Restricted Isometry Property and $s$-goodness of random matrices</td>
<td>20</td>
</tr>
<tr>
<td>1.3.2 Verifiable sufficient conditions for $Q_q(s, \kappa)$</td>
<td>20</td>
</tr>
<tr>
<td>1.3.3 Tractability of $Q_\infty(s, \kappa)$</td>
<td>22</td>
</tr>
<tr>
<td>1.3.3.1 Mutual Incoherence</td>
<td>23</td>
</tr>
<tr>
<td>1.3.3.2 From RIP to conditions $Q_q(\gamma, \kappa)$</td>
<td>24</td>
</tr>
<tr>
<td>1.3.3.3 Limits of performance of verifiable sufficient conditions for goodness</td>
<td>24</td>
</tr>
<tr>
<td>1.4 Exercises for Chapter 1</td>
<td>26</td>
</tr>
<tr>
<td>1.5 Proofs</td>
<td>30</td>
</tr>
<tr>
<td>1.5.1 Proofs of Theorem 1.3, 1.4</td>
<td>30</td>
</tr>
<tr>
<td>1.5.2 Proof of Theorem 1.5</td>
<td>31</td>
</tr>
<tr>
<td>1.5.3 Proof of Proposition 1.7</td>
<td>33</td>
</tr>
</tbody>
</table>
1.5.4 Proof of Propositions 1.8 and 1.12 ........................................ 36
1.5.5 Proof of Proposition 1.10 .................................................. 37
1.5.6 Proof of Proposition 1.13 .................................................. 39

2 Hypothesis Testing .............................................................. 41
  2.1 Preliminaries from Statistics: Hypotheses, Tests, Risks .......... 41
    2.1.1 Hypothesis Testing Problem ...................................... 41
    2.1.2 Tests ..................................................................... 42
    2.1.3 Testing from repeated observations ............................. 42
      2.1.3.1 Stationary $K$-repeated observations ...................... 43
      2.1.3.2 Semi-stationary $K$-repeated observations ............... 43
      2.1.3.3 Quasi-stationary $K$-repeated observations ............. 44
    2.1.4 Risk of a simple test ................................................ 45
    2.1.5 Two-point lower risk bound ...................................... 46
  2.2 Hypothesis Testing via Euclidean Separation ..................... 49
    2.2.1 Situation ............................................................... 49
    2.2.2 Pairwise Hypothesis Testing via Euclidean Separation .... 50
      2.2.2.1 The simplest case ............................................. 50
      2.2.2.2 Extension ....................................................... 51
      2.2.2.3 Further extensions: spherical families of distributions 53
    2.2.3 Euclidean Separation, Repeated Observations, and Majority Tests ....................................................... 55
      2.2.3.1 Preliminaries: Repeated observations in “signal plus noise” observation model ....................... 55
      2.2.3.2 Majority Test ................................................... 56
    2.2.4 From Pairwise to Multiple Hypotheses Testing .............. 58
      2.2.4.1 Situation ........................................................ 58
      2.2.4.2 Closeness relation and “up to closeness” risks ........ 59
      2.2.4.3 Multiple Hypothesis Testing via pairwise tests ...... 60
      2.2.4.4 Testing Multiple Hypotheses via Euclidean separation ....................................................... 61
  2.3 Detectors and Detector-Based Tests .................................... 64
    2.3.1 Detectors and their risks .......................................... 64
    2.3.2 Detector-based tests ................................................ 65
      2.3.2.1 Structural properties of risks ................................ 65
      2.3.2.2 Renormalization .............................................. 66
      2.3.2.3 Detector-based testing from repeated observations 66
      2.3.2.4 Limits of performance of detector-based tests ...... 70
  2.4 Simple observation schemes .............................................. 72
    2.4.1 Simple observation schemes – Motivation .................... 72
    2.4.2 Simple observation schemes – the definition ................ 73
    2.4.3 Simple observation schemes – examples ........................ 74
      2.4.3.1 Gaussian observation scheme ............................... 74
      2.4.3.2 Poisson observation scheme ............................... 75
      2.4.3.3 Discrete observation scheme ............................... 77
      2.4.3.4 Direct products of simple observation schemes ....... 78
    2.4.4 Simple observation schemes – Main result .................... 79
      2.4.4.1 Executive summary of convex-concave saddle point problems ....................................................... 79
      2.4.4.2 Main Result ................................................... 81
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.4.5</td>
<td>Simple observation schemes – Examples of optimal detectors</td>
<td>83</td>
</tr>
<tr>
<td>2.4.5.1</td>
<td>Gaussian o.s.</td>
<td>84</td>
</tr>
<tr>
<td>2.4.5.2</td>
<td>Poisson o.s.</td>
<td>85</td>
</tr>
<tr>
<td>2.4.5.3</td>
<td>Discrete o.s.</td>
<td>85</td>
</tr>
<tr>
<td>2.4.5.4</td>
<td>$K$-th power of simple o.s.</td>
<td>86</td>
</tr>
<tr>
<td>2.5</td>
<td>Testing multiple hypotheses</td>
<td>88</td>
</tr>
<tr>
<td>2.5.1</td>
<td>Testing unions</td>
<td>88</td>
</tr>
<tr>
<td>2.5.1.1</td>
<td>Situation and goal</td>
<td>88</td>
</tr>
<tr>
<td>2.5.1.2</td>
<td>The construction</td>
<td>88</td>
</tr>
<tr>
<td>2.5.2</td>
<td>Testing multiple hypotheses “up to closeness”</td>
<td>92</td>
</tr>
<tr>
<td>2.5.2.1</td>
<td>Situation and goal</td>
<td>92</td>
</tr>
<tr>
<td>2.5.2.2</td>
<td>“Building blocks” and construction</td>
<td>93</td>
</tr>
<tr>
<td>2.5.2.3</td>
<td>Testing multiple hypotheses via repeated observations</td>
<td>95</td>
</tr>
<tr>
<td>2.5.2.4</td>
<td>Consistency and near-optimality</td>
<td>96</td>
</tr>
<tr>
<td>2.5.3</td>
<td>Illustration: Selecting the best among a family of estimates</td>
<td>97</td>
</tr>
<tr>
<td>2.5.3.1</td>
<td>The construction</td>
<td>99</td>
</tr>
<tr>
<td>2.5.3.2</td>
<td>A modification</td>
<td>100</td>
</tr>
<tr>
<td>2.5.3.3</td>
<td>“Near-optimality”</td>
<td>103</td>
</tr>
<tr>
<td>2.5.3.4</td>
<td>Numerical illustration</td>
<td>104</td>
</tr>
<tr>
<td>2.6</td>
<td>Sequential Hypothesis Testing</td>
<td>105</td>
</tr>
<tr>
<td>2.6.1</td>
<td>Motivation: Election Polls</td>
<td>105</td>
</tr>
<tr>
<td>2.6.2</td>
<td>Sequential hypothesis testing</td>
<td>109</td>
</tr>
<tr>
<td>2.6.3</td>
<td>Concluding remarks</td>
<td>113</td>
</tr>
<tr>
<td>2.7</td>
<td>Measurement Design in simple observation schemes</td>
<td>114</td>
</tr>
<tr>
<td>2.7.1</td>
<td>Motivation: Opinion polls revisited</td>
<td>114</td>
</tr>
<tr>
<td>2.7.2</td>
<td>Measurement Design: SetUp</td>
<td>115</td>
</tr>
<tr>
<td>2.7.3</td>
<td>Formulating the MD problem</td>
<td>116</td>
</tr>
<tr>
<td>2.7.3.1</td>
<td>Simple case, Discrete o.s.</td>
<td>118</td>
</tr>
<tr>
<td>2.7.3.2</td>
<td>Simple case, Poisson o.s.</td>
<td>121</td>
</tr>
<tr>
<td>2.7.3.3</td>
<td>Simple case, Gaussian o.s.</td>
<td>122</td>
</tr>
<tr>
<td>2.8</td>
<td>Affine detectors beyond simple observation schemes</td>
<td>123</td>
</tr>
<tr>
<td>2.8.1</td>
<td>Situation</td>
<td>124</td>
</tr>
<tr>
<td>2.8.1.1</td>
<td>Preliminaries: Regular data and associated families of distributions</td>
<td>124</td>
</tr>
<tr>
<td>2.8.1.2</td>
<td>Basic examples of simple families of probability distributions</td>
<td>125</td>
</tr>
<tr>
<td>2.8.1.3</td>
<td>Calculus of regular and simple families of probability distributions</td>
<td>127</td>
</tr>
<tr>
<td>2.8.2</td>
<td>Main result</td>
<td>132</td>
</tr>
<tr>
<td>2.8.2.1</td>
<td>Situation &amp; Construction</td>
<td>132</td>
</tr>
<tr>
<td>2.8.2.2</td>
<td>Main Result</td>
<td>133</td>
</tr>
<tr>
<td>2.8.2.3</td>
<td>Sub-Gaussian and Gaussian cases</td>
<td>137</td>
</tr>
<tr>
<td>2.9</td>
<td>Beyond the scope of affine detectors: lifting the observations</td>
<td>139</td>
</tr>
<tr>
<td>2.9.1</td>
<td>Motivation</td>
<td>139</td>
</tr>
<tr>
<td>2.9.2</td>
<td>Quadratic lifting: Gaussian case</td>
<td>140</td>
</tr>
<tr>
<td>2.9.3</td>
<td>Quadratic lifting – does it help?</td>
<td>143</td>
</tr>
<tr>
<td>2.9.4</td>
<td>Quadratic lifting: sub-Gaussian case</td>
<td>145</td>
</tr>
<tr>
<td>2.9.5</td>
<td>Generic application: quadratically constrained hypotheses</td>
<td>148</td>
</tr>
<tr>
<td>2.9.5.1</td>
<td>Simple change detection</td>
<td>150</td>
</tr>
</tbody>
</table>
3.4.1.5 A modification .......................... 227
3.4.2 Estimating quadratic form, sub-Gaussian case .......................... 227
  3.4.2.1 Situation ................................ 227
  3.4.2.2 Construction and result ....................... 227
  3.4.2.3 Numerical illustration, direct observations .......... 231
  3.4.2.4 Numerical illustration, indirect observations .......... 234

3.5 Exercises for Chapter 3 ........................................ 237
3.6 Proofs .......................................................... 249
  3.6.1 Proof of Proposition 3.3 ................................. 249
  3.6.2 Verifying 1-convexity of the conditional quantile .......... 252
  3.6.3 Proof of Proposition 3.4 ................................ 253
  3.6.3.1 Proof of Proposition 3.4.i ................................. 253
  3.6.3.2 Proof of Proposition 3.4.ii ............................... 255
  3.6.4 Proof of Proposition 3.14 ................................ 257

4 Signal Recovery by Linear Estimation .................................. 260
Overview ............................................................... 260
4.1 Preliminaries: executive summary on Conic Programming .......... 262
  4.1.1 Cones ...................................................... 262
  4.1.2 Conic problems and their duals ............................ 263
  4.1.3 Schur Complement Lemma .................................. 265
4.2 Near-optimal linear estimation from Gaussian observations .......... 265
  4.2.1 Situation and goal ........................................ 265
  4.2.1.1 Ellitopes ............................................. 265
  4.2.1.2 Estimates and their risks ............................... 266
  4.2.1.3 Main goal ............................................. 267
  4.2.2 Building linear estimate .................................... 267
  4.2.2.1 Illustration: recovering temperature distribution .......... 269
  4.2.2.2 Near-optimality of \( \hat{x}_{H^*} \) .................. 271
  4.2.2.3 Relaxing the symmetry requirement .................... 272
  4.2.2.4 Comments ............................................... 272
  4.2.2.5 How near is “near-optimal:” numerical illustration .......... 273
  4.2.3 Byproduct on semidefinite relaxation .......................... 273

4.3 From ellitopes to spectratopes .................................... 274
  4.3.1 Spectratopes: definition and examples ........................ 275
  4.3.1.1 Examples of spectratopes ................................. 276
  4.3.2 Semidefinite relaxation on spectratopes ....................... 277
  4.3.3 Linear estimates beyond ellitopic signal sets and \( \| \cdot \|_2 \)-risk ............... 278
    4.3.3.1 Situation and goal .................................... 278
    4.3.3.2 Assumptions ........................................ 279
    4.3.3.3 Building linear estimate ................................ 281
    4.3.3.4 Upper-bounding \( \| \cdot \|_{x,H^*} \) ................... 281
    4.3.3.5 Upper-bounding \( \Psi_H(\cdot) \) ........................... 282
    4.3.3.6 Putting things together ................................ 284
    4.3.3.7 Illustration: covariance matrix estimation ............... 284
    4.3.3.8 Estimation from repeated observations .................. 289
    4.3.3.9 Near-optimality in Gaussian case .......................... 290

4.4 Linear estimates of stochastic signals ................................ 291
  4.4.1 Minimizing Euclidean risk .................................. 293
  4.4.2 Minimizing \( \| \cdot \|_\cdot \)-risk ................................. 294
5.1.4 Efficient upper-bounding of $\mathcal{R}[H]$ and contrast design, I... 392
  5.1.4.1 Assumptions ........................................... 392
  5.1.4.2 Simple observation .................................. 392
  5.1.4.3 Specifying contrasts .................................. 394
  5.1.4.4 Illustration: diagonal case ........................... 397
5.1.5 Efficient upper-bounding of $\mathcal{R}[H]$ and contrast design, II... 399
  5.1.5.1 Outline ................................................. 399
  5.1.5.2 Specifying cones $H$ .................................... 402
  5.1.5.3 Specifying functions $M$ ................................. 404
  5.1.5.4 Putting things together ................................. 405
  5.1.5.5 Compatibility: basic examples and calculus .......... 406
  5.1.5.6 Near-optimality of polyhedral estimate in the spectratopic sub-Gaussian case ........................................... 409
5.1.6 Assembling estimates: contrast aggregation .................. 411
5.1.7 Numerical illustration ..................................... 413
5.1.8 Calculus of compatibility ................................... 413
5.2 Recovering signals from nonlinear observations by Stochastic Optimization 415
  5.2.1 Problem’s setting ........................................... 415
  5.2.2 Assumptions ............................................... 417
  5.2.3 Estimating via Sample Average Approximation .............. 420
  5.2.4 Stochastic Approximation estimate ......................... 423
  5.2.5 Numerical illustration ..................................... 425
  5.2.6 "Single-observation" case .................................. 427
5.3 Exercises for Chapter 5 .................................... 430
  5.3.1 Estimation by Stochastic Optimization ...................... 430
5.4 Proofs .................................................. 440
  5.4.1 Proof of (5.8) ............................................ 440
  5.4.2 Proof of Lemma 5.6 ....................................... 441
  5.4.3 Verification of (5.44) ...................................... 442
  5.4.4 Proof of Proposition 5.10 .................................. 443

Appendix: Executive Summary on Efficient Solvability of Convex Optimization Problems 447

Bibliography .................................................. 452

Index .......................................................... 467
List of Figures

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 in the figure. ................................................. 5
1.2 Single-pixel camera ................................................................. 6
1.3 Regular and penalized ℓ1 recovery of nearly s-sparse signals. o: true signals, +: recoveries (to make the plots readable, one per eight consecutive vector’s entries is shown). Problem’s sizes are m = 256 and n = 2m = 512, noise level is σ = 0.01, deviation from s-sparsity is ∥x − x∗∥1 = 1, contrast pair is (H = √n/mA, ∥·∥∞). In penalized recovery, λ = 2s, parameter ρ of regular recovery is set to ErfcInv(0.005/n). .................................................. 18
1.4 Erroneous ℓ1 recovery of a 25-sparse signal, no observation noise. Top: frequency domain, o – true signal, + – recovery. Bottom: time domain. .................................................. 25

2.1 “Gaussian Separation” (Example 2.2): Optimal test deciding on whether the mean of Gaussian r.v. belongs to the domain A (H1) or to the domain B (H2). Hyperplane o-o separates the acceptance domains for H1 (“left” half-space) and for H2 (“right” half-space). ................................. 48
2.2 Drawing for Proposition 2.4 ...................................................... 52
2.3 Positron emission Tomography (PET) ........................................ 76
2.4 9 hypotheses on the location of the mean µ of observation ω ∼ N(µ, I2), each stating that µ belongs to specific polygon. ........................................ 93
2.5 Signal (top, bumpy) and candidate estimates (top, smooth). Bottom: the primitive of the signal. .................................................. 106
2.6 3-candidate hypotheses in probabilistic simplex Δ3: ...................... 110
2.7 PET scanner .......................................................... 122
2.8 Frames from a “movie” .................................................... 151

3.1 Boxplot of empirical distributions, over 20 random estimation problems, of the upper 0.01-risk bounds max1≤i,j≤100 ρij (as in (3.15)) for different observation sample sizes K. ........................................... 193
3.2 Bisection via Hypothesis Testing ............................................... 194
3.3 A circuit with 9 nodes and 16 arcs. a is the arc of interest; the current is measured in arcs b; external current and voltage are measured at the input node c. .................................................. 208
3.4 Histograms of recovery errors in experiments, 1000 simulations per experiment. .......................................................... 237

4.1 True distribution of temperature $U_\ast = B(x)$ at time $t_0 = 0.01$ (left) along with its recovery $\hat{U}$ via optimal linear estimate (center) and the “naive” recovery $\tilde{U}$ (right). ............................................. 271

5.1 Recovery errors for near-optimal linear estimate (left) and for polyhedral estimates yielded by Proposition 5.7 (PolyI, middle) and by the construction from Section 5.1.4 (PolyII, right), 20 simulations per each value of $\sigma$. ......................................................... 414

5.2 Left: functions $h$; right: moduli of strong monotonicity of the operators $F(\cdot)$ in $\{z : \|z\|_2 \leq R\}$ as functions of $R$. Dashed lines – case A (logistic sigmoid), solid lines – case B (linear regression), dash-dotted lines – case C (hinge function), dotted line – case D (ramp sigmoid). 426

5.3 Mean errors and CPU times for SA (solid lines) and SAA estimates (dashed lines) as functions of the number of observations $K$. o – case A (logistic link), x – case B (linear link), + – case C (hinge function), □ – case D (ramp sigmoid). ......................................................... 427

5.4 Solid curve: $M_{\omega,K}(z)$, dashed curve: $H_{\omega,K}(z)$. True signal $x$ (solid vertical line): 0.081; SAA estimate (unique minimizer of $H_{\omega,K}$, dashed vertical line): -0.252; ML estimate (global minimizer of $M_{\omega,K}$ on $[-20, 20]$): -20.00, closest to $x$ local minimizer of $M_{\omega,K}$ (dotted vertical line): -0.363. ......................................................... 428

5.5 Mean errors and CPU times for standard deviation $\lambda = 1$ (dashed line) and $\lambda = 0.1$ (solid line). ......................................................... 431
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, 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 this book is to present another type of links between Optimization and Statistics, those which have little 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. Therefore, 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:

A. Sparsity-oriented Compressive Sensing. Here the role of Convex Optimization theory as creative tool motivating the construction of inference procedures is relatively less important than in two other topics. This being said, its role is by far non-negligible in the analysis of Compressive Sensing routines (it allows, e.g., to derive from “first principles” the necessary and sufficient conditions for the validity of \( \ell_1 \) recovery). On account of this, and also due to its popularity and the fact that now it is one of the major “customers” of advanced convex optimization algorithms, we believe that Compressive Sensing is worthy of being considered.
B. Pairwise and Multiple Hypothesis Testing, including sequential tests, estimation of linear functionals, and some rudimentary design of experiments.
C. Recovery of signals from noisy observations of their linear images.

B and C are the topics where, as of now, the approaches we present in this book appear the most successful.

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 basic Calculus, Probability, and Linear Algebra.

Selection and treatment of our topics are inspired by a kind of “philosophy”
which can be explained to an expert as follows. Compare two well known results of nonparametric statistics (“⟨…⟩” marks fragments irrelevant to the discussion to follow):

**Theorem A** [I. Ibragimov & R. Khas’minskii [122], 1979] Given \( \alpha, L, k \), let \( \mathcal{X} \) be the set of all functions \( f : [0,1] \rightarrow \mathbb{R} \) with \((\alpha, L)\)-Hölder continuous \( k \)-th derivative. For a given \( t \), the minimax risk of estimating \( f(t), f \in \mathcal{X} \), from noisy observations \( y = f|_{\Gamma_n} + \xi, \xi \sim \mathcal{N}(0; I_n) \) taken along \( n \)-point equidistant grid \( \Gamma_n \), up to a factor \( C(\beta) = ⟨…⟩, \beta := k + \alpha \), is \( (Ln^{-\beta})^{1/(2\beta+1)} \), and the upper risk bound is attained at the affine in \( y \) estimate explicitly given by ⟨…⟩.

**Theorem B** [D. Donoho [65], 1994] Let \( \mathcal{X} \subseteq \mathbb{R}^N \) be a convex compact set, \( A \) be an \( n \times N \) matrix, and \( g(\cdot) \) be a linear form on \( \mathcal{X} \). The minimax, over \( f \in \mathcal{X} \), risk of recovering \( g(f) \) from the noisy observations \( y = Af + \xi, \xi \sim \mathcal{N}(0, I_n) \), within factor 1.2 is attained at an affine in \( y \) estimate which, along with its risk, can be built efficiently by solving convex optimization problem ⟨…⟩.

In many respects, **A** and **B** are similar: both are theorems on minimax optimal estimation of a given linear form of an unknown “signal” \( f \) known to belong to a given convex set \( \mathcal{X} \) from observations, corrupted by Gaussian noise, of the image of \( f \) under linear mapping, and both are associated with efficiently computable near-optimal, in a minimax sense, estimators which happen to be affine in observations. There is, however, a significant structural difference: **A** gives an explicit “closed form” analytic description of the minimax risk as a function of \( n \) and smoothness parameters of \( f \), along with explicit description of the near-optimal estimator. Numerous results of this type, let us call them **descriptive**, form the backbone of deep and rich theory of Nonparametric Statistics. This being said, strong “explanation power” of descriptive results has its price: we need to impose assumptions, sometimes quite restrictive, on the entities involved. For example, **A** says nothing on what happens with the minimax risk/estimate when in addition to smoothness other a priori information on \( f \), like monotonicity or convexity, is available, and/or when “direct” observations of \( f|_{\Gamma_n} \) are replaced with observations of a linear image of \( f \) (say, convolution of \( f \) with a given kernel; more often than not, this is what happens in applications), and descriptive answers to the just posed questions require a dedicated (and sometimes quite problematic) investigation more or less “from scratch.” In contrast, the explanation power of **B** is basically nonexistent: the statement presents no “closed form” expressions neither for the near-optimal estimate, nor for its worst-case risk. As a compensation, **B** makes only (relatively) mild general structural assumptions about the model (convexity and compactness of \( \mathcal{X} \), linear dependence of \( y \) on \( f \)), and all the rest – the near-optimal estimate and its risk – can be found by efficient computation. Moreover, we know in advance that the risk, whatever it happens to be, is within 20% of the actual minimax risk achievable under the circumstances. In this respect, **B** is an **operational**, rather than a descriptive, result: it explains **how to act** to achieve the (nearly) best possible performance, with no a priori prediction of what this performance will be. The latter hardly is a “big issue” in applications – with huge computational power readily available, efficient computability is, basically, as good as a “simple explicit

---

\(^1\text{Infinite dimensionality of } \mathcal{X} \text{ in } \text{A} \text{ is of no importance – nothing changes when replacing the original } \mathcal{X} \text{ with its } n \text{-dimensional image under the mapping } f \mapsto f|_{\Gamma_n}.\)
formula.” We strongly believe that as far as applications of high-dimensional statistics are concerned, operational results, possessing much broader scope than their descriptive counterparts, are of significant importance and potential. Our main motivation when writing this book was to contribute to the body of operational results in Statistics, and this is what Chapters 2 – 5 to follow are about.

Anatoli Juditsky & Arkadi Nemirovski
March 6, 2019
ACKNOWLEDGEMENTS

We are greatly indebted to Prof. H. Edwin Romeijn who initiated creating the Ph.D. course “Topics in Data Science;” Lecture Notes for this course form the seed of the book to follow. We gratefully acknowledge support from NSF Grant CC-1523768 Statistical Inference via Convex Optimization; this research project is the source of basically all novel results presented in Chapters 2 – 5. Our deepest gratitude goes to Prof. Stephen Boyd who many years ago taught one of the authors “operational philosophy” motivating the research we are presenting. Needless to say, responsibility for all drawbacks of the book is ours.

A. J. & A. N.
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 \(A, B, ..., Z\) of the same “width” \([A; B; C; ...; D]\) (vectors/matrices \(A, B, C, ..., Z\) of the same “height,”) is the matrix obtained by vertical (horizontal) concatenation of \(A, B, C, \ldots\). 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

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

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

\[
\begin{bmatrix}
1 & 2 \\
3 & 4 \\
5 & 6
\end{bmatrix} = \begin{bmatrix}
1 & 0 & 0 \\
2 & 0 & 0 \\
3 & 4 & 5
\end{bmatrix}.
\]

\(\text{Diag}\{A_1, A_2, ..., A_k\}\) stands for block-diagonal matrix with diagonal blocks \(A_1, A_2, ..., A_k\). For example,

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

For an \(m \times n\) matrix \(A\), \(\text{dg}(A)\) 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 book are \(\mathbb{R}^n\) (the space of \(n\)-dimensional column vectors), \(\mathbb{R}^{m \times n}\) (the space of \(m \times n\) real matrices), and \(\mathbb{S}^n\) (the space of \(n \times n\) real symmetric matrices). All these linear spaces are equipped with the standard inner product:

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

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 \(\text{Tr}(AB^T)\).

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

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

\[
\langle A(x), y \rangle = \langle x, A^*(y) \rangle \quad \forall (x \in E_x, y \in E_y).
\]

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

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

\( Z^n \) is the set of \( n \)-dimensional integer vectors.

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

\[
\|x\|_p = \begin{cases} 
(\sum_{i=1}^{n} |x_i|^p)^{1/p} & , 1 \leq p < \infty \\
\max_i |x_i| = \lim_{p' \to \infty} \|x\|_{p'} & , p = \infty
\end{cases}
\]

The spectral norm (the largest singular value) of a matrix \( A \) is denoted \( \|A\|_{2,2} \); notation for other norms of matrices is specified when used.

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

**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 \( A > B \) (\( \Leftrightarrow B < A \)) 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 \( x \in \mathbb{R}^n \), LMI reads

\[
A_0 + \sum_{i} x_i A_i \succeq 0 \quad [A_i \in \mathbb{S}^n, 0 \leq i \leq n]
\]

- \( \mathcal{N}(\mu, \Theta) \) stands for the Gaussian distribution with mean \( \mu \) and covariance matrix \( \Theta \). Poisson(\( \mu \)) denotes Poisson distribution with parameter \( \mu \in \mathbb{R}_+ \), i.e., the distribution of random variable taking values \( i = 0, 1, 2, ... \) with probabilities \( \frac{\mu^i}{i!} e^{-\mu} \). Uniform([\( a, b \)]) is the uniform distribution on segment \([ a, b ]\).
- For a probability distribution \( P \),
  - \( \xi \sim P \) means that \( \xi \) 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);
  - \( \mathbb{E}_{\xi \sim P}\{f(\xi)\} \) is the expectation of \( f(\xi) \), \( \xi \sim P \); when \( P \) is clear from the context, this notation can be shortened to \( \mathbb{E}_{\xi}\{f(\xi)\} \), or \( \mathbb{E}_{P}\{f(\xi)\} \), or even \( \mathbb{E}\{f(\xi)\} \). Similarly, \( \text{Prob}_{\xi \sim P}\{\cdot\} \), \( \text{Prob}_{\xi}\{\cdot\} \), \( \text{Prob}_P\{\cdot\} \), \( \text{Prob}\{\cdot\} \) denote the \( P \)-probability of the event specified inside the braces.
- \( O(1) \)'s stand for positive absolute constants – positive reals with numerical values (completely independent of parameters of the situation at hand) 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 \( \Pi \) over domain \( \Omega \), of function \( f \).
ABOUT PROOFS

The book is 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 ...” at the end of each chapter, and this is where a reader should look for “missing” proofs.

ON COMPUTATIONAL TRACTABILITY

In the main body of the book, one can frequently meet sentences like “Φ(·) is an efficiently computable convex function,” or “X is a computationally tractable convex set,” or “(P) is an explicit, and therefore efficiently solvable, convex optimization problem.” For “executive summary” on what these words actually mean, we refer the reader to Appendix.
Chapter One

Sparse Recovery via \( \ell_1 \) Minimization

In this chapter, we overview basic results of Compressed Sensing – a relatively new and rapidly developing area in Statistics and Signal Processing dealing with recovering signals (vectors \( x \) from some \( \mathbb{R}^n \)) from their noisy observations \( Ax + \eta \) (\( A \) is a given \( m \times n \) sensing matrix, \( \eta \) 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 signal. This setup leads to deep, elegant and highly innovative theory and possesses quite significant application 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 and restrict ourselves with 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 \( x \in \mathbb{R}^n \) from noisy observations

\[
y = Ax + \eta
\]  

(1.1)

of linear image of the signal under a given sensing mapping \( x \mapsto Ax : \mathbb{R}^n \rightarrow \mathbb{R}^m \); in (1.1), \( \eta \) is the observation error. Matrix \( A \) in (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 transmitter, \( y \) is the signal recorded by receiver, \( A \) represents the communication channel (reflecting, e.g., dependencies of decays in signals’ amplitude on the transmitter-receiver distances); \( \eta \) here typically is modeled as the standard (zero mean, unit covariance matrix) \( m \)-dimensional Gaussian noise;\(^1\)

\(^1\)While the “physical” noise indeed is often Gaussian with zero mean, its covariance matrix is not necessarily the unit matrix. Note, however, that a zero mean Gaussian noise \( \eta \) always can be represented as \( Q\xi \) with standard Gaussian \( \xi \). Assuming that \( Q \) is known and is nonsingular (which indeed is so when the covariance matrix of \( \eta \) is positive definite), we can rewrite (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; the new observation
• **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 $\eta$ 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 $a^i \in \mathbb{R}^n$” to a “black box” — output $y^i \in \mathbb{R}$ 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 $y = [y_1; ...; y_m] \in \mathbb{R}^m$, we arrive at the problem of recovering vector $x$ from noisy observations of $Ax$. Here again the most popular model for $\eta$ 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 arises only when $m \geq n$ (“more observations than 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:** $m \gg n$, $\eta$ is nontrivial noise with zero mean, say, standard Gaussian one. This is the classical statistical setup with the emphasis on how to use numerous available observations in order to suppress in the recovery, to the extent possible, the influence of observation noise.

**Nonparametric case:** $m \ll n$. If addressed literally, this case seems to be senseless: when the number of observations is less that the number of unknown parameters, even in the noiseless case we arrive at the necessity to solve an underdetermined (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) some a priori information about the signal. In traditional Nonparametric Statistics, this additional information is summarized in a given to us in advance bounded convex set $X \subset \mathbb{R}^n$ 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 a noise $\xi$ 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.

Of course, this is a blatant simplification – the nonparametric case covers also a variety of important and by far nontrivial situations in which $m$ is comparable to $n$ or larger than $n$ (or even $\gg n$). However, this simplification is very convenient, and we will use it in this introduction.
properly selected and known to us in advance 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 $[53, 85, 122, 110, 205]$. 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 $s$ $[85, 122, 175, 219, 220, 226, 234]$. Of course, this outline of traditional 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 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.

Compressed Sensing. The situation changed dramatically around Year 2000 as a consequence of important theoretical breakthroughs due to D. Donoho, T. Tao, J. Romberg, E. Candes, J.-J. Fuchs among many other researchers $[44, 45, 46, 47, 48, 50, 68, 69, 70, 63, 92, 93]$; as a result of these breakthroughs, novel and rich 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 approach to the $m \ll n$ case, it is assumed that after passing to an appropriate basis, the signal to be recovered is $s$-sparse (has $\leq s$ nonzero entries with $s \ll m$), or is well approximated by an $s$-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 assume that the true vector of parameters $x$ 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 $j = 1, 2, \ldots, n$, then the $m$-dimensional vector $y$ of measurements of the $m$ receivers is $y =$
\[ Ax + \eta, \text{ where } \eta \text{ is the observation noise. Given } y, \text{ we intend to recover } x. \]

Now, if the receivers are, say, hydrophones registering noises emitted by submarines in a 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) may be in the range of tens of thousands, if not tens of millions. At the same time, presumably, there is only a handful of “submarines” (i.e., nonzero entries in \( x \)) in the area.

To “see” sparsity in everyday life, look at the 256 \( \times \) 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). At the bottom of Figure 1.1 we see what happens when we zero out all but a small percentage of the largest in magnitude wavelet coefficients and replace the true image by its sparse, in the wavelet basis, approximations.

This simple 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, 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 (to some extent, already actualized) 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 an analog 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” [83], see Figure 1.2) as follows: in reality, the digital 256 \( \times \) 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 light’s intensities 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 intensities in each of the 65,536 pixels, we compute (using an analog device) the integral, taken over the field of view, of the product of light intensity and an analogously generated “mask.” We repeat 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
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 in the figure.
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 [95, 172].

1.1.3 Compressed Sensing via $\ell_1$ 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 $\text{Col}_j[A]$ 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 $\sum_{\ell=1}^{s} x_{j_\ell} \text{Col}_{j_\ell}[A] = y$ 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

$$\min_z \{ \|z\|_0 : Az = y \}, \quad (1.2)$$

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 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 (1.2).

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 $s$-sparse, we could solve the combinatorial optimization problem

$$\min_z \{ \| z \|_0 : \| Az - y \| \leq \epsilon \}.$$  \hspace{1cm} (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 $\hat{x}$ of (1.3) are close to each other within accuracy of order of $\epsilon$: $\| x - \hat{x} \|_2 \leq 2C\epsilon$. 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: (1.2) is a difficult combinatorial problem. As far as known theoretical complexity guarantees are concerned, they are not better than for the “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., 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 current option. 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.\footnote{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 $\approx 7.53e7$ — much, but perhaps doable. When $n = 200$ and $s = 20$, the number of systems to be processed jumps to $\approx 1.61e27$, 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.}

A partial remedy is as follows. Well, if we do not know how to minimize under linear constraints, as in (1.2), the “bad” objective $\| z \|_0$, let us “approximate” this objective with one which we do know how to minimize. The true objective is separable: $\| z \| = \sum_{i=1}^n \xi(z_i)$, 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, (1.2) converts into the $\ell_1$ minimization problem

$$\min_z \{ \| z \|_1 := \sum_{i=1}^n |z_i| : Az = y \}.$$  \hspace{1cm} (1.4)
and (1.3) becomes the convex optimization problem
\[
\min \{ \| z \|_1 : \| Az - y \| \leq \epsilon \}. \tag{1.5}
\]
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 $\ell_1$ recovery – its 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 $\ell_1$ Minimization

What follows is based on the standard basic results of Compressed Sensing theory originating from [19, 44, 47, 45, 48, 49, 51, 68, 70, 63, 92, 93, 228] and augmented by the results of [128, 129, 131, 132].

1.2.1 Validity of $\ell_1$ minimization in the noiseless case

The minimal requirement on sensing matrix $A$ which makes $\ell_1$ 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 $x \in \mathbb{R}^n$

- $I_x = \{ j : x_j \neq 0 \}$ stands for the support of $x$; we also set
  \[
  I_x^+ = \{ j : x_j > 0 \}, \quad I_x^- = \{ j : x_j < 0 \}
  \Rightarrow I_x = I_x^+ \cup I_x^-
  \]
- for a subset $I$ of the index set $\{1, \ldots, n\}$, $x_I$ stands for the vector obtained from $x$ by zeroing out entries with indices not in $I$, and $I^o$ for the complement of $I$:
  \[
  I^o = \{ i \in \{1, \ldots, n\} : i \notin I \};
  \]
- for $s \leq n$, $x^s$ stands for the vector obtained from $x$ by zeroing out all but the $s$ largest in magnitude entries.\footnote{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 [129] to the case of plain sparsity.} Note that $x^s$ is the best $s$-sparse approximation of $x$ in all $\ell_p$ norms, $1 \leq p \leq \infty$;
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.

### 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 $\ell_1$ minimization. In other words, $A$ is $s$-good, if whenever in $y$ in (1.4) is of the form $y = Ax$ with $s$-sparse $x$, $x$ is the unique optimal solution to (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 originating from [63]. 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 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

$$\min \{\|z\|_1 : Az = Ax\}, \quad (P[x])$$

we refer to the latter property of $A$ as weak $s$-goodness. Our first observation is as follows:

**Proposition 1.1.** If $A$ is weakly $s$-good, then the following condition holds true: whenever $I$ is a subset of $\{1, \ldots, n\}$ of cardinality $\leq s$, we have

$$\forall w \in \text{Ker}A \quad \|w_I\|_1 \leq \|w_I^*\|_1. \quad (1.6)$$

**Proof** is immediate. Assume $A$ is weakly $s$-good, and let us verify (1.6). Let $I$ be an $s$-element subset of $\{1, \ldots, n\}$, and $x$ be $s$-sparse vector with support $I$. Since $A$ is weakly $s$-good, $x$ is an optimal solution to $(P[x])$. Rewriting the latter problem in the form of LP, that is, as

$$\min \{\sum_j t_j : t_j + z_j \geq 0, t_j - z_j \geq 0, Az = Ax\},$$

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 $\lambda_j^+, \lambda_j^-$, $\mu \in \mathbb{R}^m$ (Lagrange multipliers for the constraints $t_j - z_j \geq 0$, $t_j + z_j \geq 0$, and
Indeed, we already know that if necessary from what we already know it immediately follows that a
weak In fact, it can be shown that (1.6) is not only necessary, but also sufficient condition
1.2.1.3 Nullspace property
of signs of an for all was to guess A clearly is every nonzero w ∈ Ker A, not 1) such that
if and only if there exists vector µ ∈ R^m such that the j-th entry of A^T µ is −1, if x_j > 0, +1, if x_j < 0, and a real from
[−1,1], if x_j = 0.
Now let w ∈ Ker A be a vector with the same signs of entries w_i, i ∈ I, as these of the
entries in x. Then
0 = µ^T Aw = [A^T µ]^T w = \sum_j [A^T µ]_j w_j
⇒ \sum_{j \in I} |w_j| = \sum_{j \in I} |[A^T µ]_j| w_j = − \sum_{j \notin I} |[A^T µ]_j| w_j ≤ \sum_{j \notin I} |w_j|
(we have used the fact that |[A^T µ]_j| = \text{sign} x_j = \text{sign} w_j for j ∈ I_x and |[A^T µ]_j| ≤ 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, (1.6) holds true.
□

1.2.1.3 Nullspace property
In fact, it can be shown that (1.6) is not only necessary, but also sufficient condition for weak s-goodness of A; we, however, skip this verification, since our goal so far was to guess condition for s-goodness, and this goal has already been achieved – from what we already know it immediately follows that a necessary condition for s-goodness is for the inequality in (1.6) to be strict whenever w ∈ Ker A is nonzero.
Indeed, we already know that if A is s-good, then for every I of cardinality s and every nonzero w ∈ Ker A it holds
\|w_I\|_1 ≤ \|w_{I^c}\|_1.
If the latter inequality for some I and w in question holds true as equality, then A clearly is not s-good, since the s-sparse signal x = w_I is not the unique optimal solution to (P[x]) – the vector −w_{I^c} 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
\forall(0 \neq w \in Ker A, I, \text{Card}(I) \leq s) : \|w_I\|_1 < \|w_{I^c}\|_1.
By the standard compactness argument, this is the same as the existence of γ ∈ (0,1) such that
\forall(w \in Ker A, I, \text{Card}(I) \leq s) : \|w_I\|_1 ≤ γ\|w_{I^c}\|_1,
or, which is the same, existence of \( \kappa \in (0, 1/2) \) such that
\[
\forall (w \in \text{Ker} A, I, \text{Card}(I) \leq s) : \|w_I\|_1 \leq \kappa \|w\|_1.
\]
Finally, the supremum of \( \|w_I\|_1 \) over \( I \) of cardinality \( s \) is the norm \( \|w\|_{s,1} \) (the sum of \( s \) largest magnitudes of entries) of \( w \), so that the condition we are processing finally can be formulated as
\[
\exists \kappa \in (0, 1/2) : \|w\|_{s,1} \leq \kappa \|w\|_1 \forall w \in \text{Ker} A. \quad (1.8)
\]
The resulting nullspace condition in fact is necessary and sufficient for \( A \) to be \( s \)-good:

**Proposition 1.2.** Condition (1.8) 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, and \( y \) be an optimal solution to \( (P[x]) \); all we need is to prove that \( y = x \). Let \( I \) be the support of \( x \), and \( w = y - x \), so that \( w \in \text{Ker} A \). By the nullspace property we have
\[
\|w_I\|_1 \leq \kappa \|w\|_1 = \kappa (\|w_I\|_1 + \|w_{\bar{I}}\|_1) = \kappa (\|w_I\|_1 + \|y_I - y_{\bar{I}}\|_1)
\]
\[
\Rightarrow \|w_I\|_1 \leq \frac{\kappa}{1 - \kappa} \|y_{\bar{I}}\|_1
\]
\[
\Rightarrow \|x\|_1 = \|x_I\|_1 = \|y_I - y_{\bar{I}}\|_1 \leq \|y_I\|_1 + \frac{\kappa}{1 - \kappa} \|y_{\bar{I}}\|_1 \leq \|y_I\|_1 + \|y_{\bar{I}}\|_1 = \|y\|_1
\]
where the concluding \( \leq \) is due to \( \kappa \in [0, 1/2) \). Since \( x \) is a feasible, and \( y \) is an optimal solution to \( (P[x]) \), the resulting inequality \( \|x\|_1 \leq \|y\|_1 \) must be equality, which, again due to \( \kappa \in [0, 1/2) \), is possible only when \( y_{\bar{I}} = 0 \). Thus, \( y \) has the same support \( I \) as \( x \), and \( w = x - y \in \text{Ker} A \) is supported on \( s \)-element set \( I \); by nullspace property, we should have \( \|w_I\|_1 \leq \kappa \|w\|_1 = \kappa \|w_I\|_1 \), which is possible only when \( w = 0 \). \( \square \)

### 1.2.2 Imperfect \( \ell_1 \) minimization

We have found a necessary and sufficient condition for \( \ell_1 \) minimization to recover exactly \( s \)-sparse signals in the noiseless case. More often than not, both these assumptions 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 the 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
\[ \|w\|_{s,1} \leq \kappa \|w\|_1 \forall w \in \text{Ker } A \]  
(1.9)
is satisfied if and only if for a properly selected constant \( C \) one has\(^6\)
\[ \|w\|_{s,1} \leq C \|Aw\|_2 + \kappa \|w\|_1 \forall w. \]  
(1.10)
Indeed, (1.10) clearly implies (1.9); to get the inverse implication, note that for every \( h \) orthogonal to \( \text{Ker } A \) it holds
\[ \|Ah\|_2 \geq \sigma \|h\|_2, \]
where \( \sigma > 0 \) is the minimal positive singular value of \( A \). Now, given \( w \in \mathbb{R}^n \), we can decompose \( w \) into the sum of \( \tilde{w} \in \text{Ker } A \) and \( h \in (\text{Ker } A)^\perp \), so that
\[ \|w\|_{s,1} \leq \|\tilde{w}\|_{s,1} + \|h\|_{s,1} \leq \kappa (\|\tilde{w}\|_1 + \sqrt{\sigma}\|\tilde{h}\|_2) \leq \kappa (\|\tilde{w}\|_1 + \|\tilde{h}\|_1 + \sqrt{\sigma}\|\tilde{h}\|_2)
\]
\[ \leq \kappa \|\tilde{w}\|_1 + \kappa \sqrt{n} \sqrt{\sigma} \|h\|_2 \leq \sigma^{-1} \kappa \sqrt{n} \sqrt{\sigma} \|Ah\|_2 + \kappa \|w\|_1, \]
as required in (1.10).

**Condition \( Q_1(s, \kappa) \).** For our purposes, it is convenient to present the condition (1.10) in the following flexible form:
\[ \|w\|_{s,1} \leq s \|H^T Aw\| + \kappa \|w\|_1, \]  
(1.11)
where \( H \) is an \( m \times N \) contrast matrix and \( \| \cdot \| \) is some norm on \( \mathbb{R}^N \). Whenever a pair \((H, \| \cdot \|)\), called contrast pair, satisfies (1.11), we say that \((H, \| \cdot \|)\) satisfies condition \( Q_1(s, \kappa) \). From what we have seen, \( \text{If } A \) 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 \( Q_1(s, \kappa) \), e.g., to take \( H = CI_m \) with appropriately large \( C \) and \( \| \cdot \| = \| \cdot \|_2 \).

**Conditions \( Q_q(s, \kappa) \).** As we will see in a while, it makes sense to embed the condition \( Q_1(s, \kappa) \) into a parametric family of conditions \( Q_q(s, \kappa) \), 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 \( \mathbb{R}^N \) satisfy condition \( Q_q(s, \kappa) \), if
\[ \|w\|_{s,q} \leq s^{\frac{1}{q}} \|H^T Aw\| + \kappa s^{\frac{1}{q} - 1} \|w\|_1 \forall w \in \mathbb{R}^n. \]  
(1.12)
Let us make two immediate observations on relations between the conditions:

**A.** When a pair \((H, \| \cdot \|)\) satisfies condition \( Q_q(s, \kappa) \), the pair satisfies also all conditions \( Q_{q'}(s, \kappa) \) with \( 1 \leq q' \leq q \). Indeed in the situation in question for \( 1 \leq q' \leq q \) it holds
\[ \|w\|_{s,q'} \leq s^{\frac{1}{q'}} \|w\|_{s,q} \leq s^{\frac{1}{q'}} \left[ s^{\frac{1}{q}} \|H^T Aw\| + \kappa s^{\frac{1}{q'} - 1} \|w\|_1 \right]
\]
\[ = s^{\frac{1}{q'}} \|H^T Aw\| + \kappa s^{\frac{1}{q'} - 1} \|w\|_1, \]
\[^6\text{Note that (1.9) is exactly the } \phi^2(s, \kappa)-\text{Compatibility condition of [227] with } \phi(s, \kappa) = C/\sqrt{s}; \text{ see also [228] for the analysis of relationships of this condition with other assumptions (e.g., a similar Restricted Eigenvalue assumption of [20]) used to analyze } \ell_1-\text{minimization procedures.} \]
where the first inequality is the standard inequality between $\ell_p$-norms of the $s$-dimensional vector $w$.

**B. When a pair $(H, \| \cdot \|)$ satisfies condition $Q_s(s, \kappa)$ and $1 \leq s' \leq s$, the pair $((s/s')^{\frac{1}{2}} H, \| \cdot \|)$ satisfies the condition $Q_s(s', \kappa)$.**

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

$$
\|w\|_{s', q} \leq \|w\|_{s, q} \leq (s')^{\frac{1}{2}} \left\| (s/s')^{\frac{1}{2}} H \right\| A w \| + \kappa \frac{(s')^{\frac{1}{2}} - 1}{\kappa^{\frac{1}{2}} - 1} \|w\|_1.
$$

### 1.2.3 Regular $\ell_1$ recovery

Given the observation scheme (1.1) with an $m \times n$ sensing matrix $A$, we define the regular $\ell_1$ recovery of $x$ via observation $y$ as

$$
\hat{x}_{\text{reg}}(y) \in \text{Argmin} \left\{ \|u\|_1 : \|H^T(Au - y)\| \leq \rho \right\}, \quad (1.13)
$$

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

The role of $Q$-conditions we have introduced is clear from the following

**Theorem 1.3.** Let $s$ be a positive integer, $q \in [1, \infty]$ and $\kappa \in (0, 1/2)$. Assume that a pair $(H, \| \cdot \|)$ satisfies the condition $Q_s(s, \kappa)$ associated with $A$, and let

$$
\Xi_\rho = \{ \eta : \|H^T\eta\| \leq \rho \}. \quad (1.14)
$$

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

$$
\|\hat{x}_{\text{reg}}(Ax + \eta) - x\|_p \leq \frac{4(2s)^{\frac{1}{2}}}{1 - 2\kappa} \left[ \rho + \frac{\|x - x^*\|_1}{2s} \right], \quad 1 \leq p \leq q. \quad (1.15)
$$

The above result can be slightly strengthened by replacing the assumption that $(H, \| \cdot \|)$ satisfies $Q_s(s, \kappa)$ with some $\kappa < 1/2$, with a weaker, by observation $A$ from Section 1.2.2.1, assumption that $(H, \| \cdot \|)$ satisfies $Q_1(s, \kappa)$ with $\kappa < 1/2$ and satisfies $Q_s(s, \kappa)$ with some (perhaps large) $\kappa$:

**Theorem 1.4.** Given $A$, integer $s > 0$ and $q \in [1, \infty]$, assume that $(H, \| \cdot \|)$ satisfies the condition $Q_1(s, \kappa)$ with $\kappa < 1/2$ and the condition $Q_s(s, \kappa)$ with some $\kappa \geq \kappa$, and let $\Xi_\rho$ be given by (1.14). Then for all $x \in \mathbb{R}^n$ and $\eta \in \Xi_\rho$ it holds:

$$
\|\hat{x}_{\text{reg}}(Ax + \eta) - x\|_p \leq \frac{4(2s)^{\frac{1}{2}}}{1 - 2\kappa} \left[ \rho + \frac{\|x - x^*\|_1}{2s} \right], \quad 1 \leq p \leq q. \quad (1.16)
$$

For proofs of Theorems 1.3, 1.4, see Section 1.5.1.

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

### 1.2.4 Penalized $\ell_1$ recovery

Penalized $\ell_1$ recovery of signal $x$ from its observation (1.1) is

$$
\hat{x}_{\text{pen}}(y) \in \text{Argmin} \left\{ \|u\|_1 + \lambda \|H^T(Au - y)\| \right\}, \quad (1.17)
$$
where $H \in \mathbb{R}^{m \times N}$, a norm $\| \cdot \|$ on $\mathbb{R}^N$ and a positive real $\lambda$ are parameters of the construction.

**Theorem 1.5.** Given $A$, positive integer $s$, and $q \in [1, \infty]$, assume that $(H, \| \cdot \|)$ satisfies the conditions $Q_q(s, \kappa)$ and $Q_1(s, \kappa)$ with $\kappa < 1/2$ and $\kappa \geq \kappa$. Then

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

$$
\| \hat{x}_{\text{pen}}(y) - x \|_p \leq \frac{\lambda}{2s} \left[ 1 + \frac{\lambda}{2s} - \kappa \right] \frac{q(p-1)}{(q-1)^2} \left[ \| H^T (Ax - y) \| + \frac{\|x - x^*\|_1}{2s} \right], 1 \leq p \leq q.
$$

(1.18)

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

$$
\| \hat{x}_{\text{pen}}(y) - x \|_p \leq \frac{4}{(2s)^{1/2}} \left[ 1 + \kappa - \kappa \right] \frac{q(p-1)}{(q-1)^2} \left[ \| H^T (Ax - y) \| + \frac{\|x - x^*\|_1}{2s} \right], 1 \leq p \leq q.
$$

(1.19)

(ii) Let $\rho \geq 0$, and let $\Xi_{\rho}$ be given by (1.14). Then for all $x \in \mathbb{R}^n$ and all $\eta \in \Xi_{\rho}$ one has:

$$
\| \hat{x}_{\text{pen}}(Ax + \eta) - x \|_p \leq \frac{\lambda}{2s} \left[ 1 + \frac{\lambda}{2s} - \kappa \right] \frac{q(p-1)}{(q-1)^2} \left[ \rho + \frac{\|x - x^*\|_1}{2s} \right], 1 \leq p \leq q.
$$

(1.19)

For proof, see Section 1.5.2.

### 1.2.5 Discussion

Some remarks are in order.

**A.** Qualitatively speaking, Theorems 1.3, 1.4, 1.5 say the same: when $Q_q$-conditions are satisfied, the regular, resp., penalized recoveries reproduce the true signal exactly when there is no observation noise and the signal is $s$-sparse. In the presence of observation error $\eta$ 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 $\|x - x^*\|_1$ 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 recovery — by an a priori upper bound $\rho$ on $\| H^T \eta \|$: when $\rho \geq \| H^T \eta \|$, $\eta$ belongs to $\Xi_{\rho}$ and thus the bounds (1.15), (1.16) are applicable to the actual observation error $\eta$. Clearly, in qualitative terms, an 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 (1.15), (1.16), (1.20), up to factors $C$ depending solely on $\kappa$ and $\kappa$, are the same, specifically,

$$
\| \hat{x} - x \|_p \leq Cs^{1/p} \| H^T \eta \| + \|x - x^*\|_1/s, 1 \leq p \leq q.
$$

(1)

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: $A = aI$. Let us assume in addition that $x$ is exactly $s$-sparse. In this case, the simplest way to ensure condition $Q_q(s, \kappa)$, even with $\kappa = 0$, is to
take $\| \cdot \| = \| \cdot \|_{s,q}$ and $H = s^{-1/q} \alpha^{-1} I$, so that (1) becomes

$$\| \hat{x} - x \|_p \leq C \alpha^{-1} s^{1/p - 1/q} \| \eta \|_{s,q}, \quad 1 \leq p \leq q.$$  (!!)

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 largest in magnitude entries of $x$, we would be unable to recover $x$ is $q$-norm with error $\leq \alpha^{-1} \| \eta \|_{s,q}$. In addition, with the equal to each other $s$ largest magnitudes of entries in $\eta$, 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 $Q_1(s, \kappa)$ with some $\kappa < 1/2$, the larger is $q$ such that the pair $(H, \| \cdot \|)$ obeys the condition $Q_2(s, \kappa)$ with a given $\kappa \geq \kappa$ (recall that $\kappa$ can be $\geq 1/2$) and $s$, the larger is the range $p \leq q$ of values of $p$ where the error bounds (1.16) and (1.20) are applicable. This is in full accordance with the fact that if a pair $(H, \| \cdot \|)$ obeys condition $Q_2(s, \kappa)$, it obeys also all conditions $Q_{q'}(s, \kappa)$ 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 $\eta$ is “uncertain but bounded,” say, when all we know is that $\| \eta \|_2 \leq \delta$ with some given $\delta$, all what matters (on the top of the requirement for $(H, \| \cdot \|)$ to obey $Q$-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 $(H, \| \cdot \|)$ obeys $Q$-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:

$$y = Ax + \eta, \quad \eta \sim \mathcal{N}(0, \sigma^2 I_m)$$  (1.21)

looking at the spectral norm of $H$, with $\| \cdot \|_2$ in the role of $\| \cdot \|$, is counter-productive, since a typical realization of $\eta$ 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 $\eta$ is completely misleading – in typical cases, this product will grow rapidly with the number of observations $m$, completely ignoring the fact that $\eta$ is random with zero mean.\footnote{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 $\sigma \sqrt{m}$. This is in sharp contrast to the fact that typical values of $h^T \eta$ are of order of $\sigma \| h \|_2$, independently of what $m$ is!}

What is much better suited for the case of Gaussian noise, is $\| \cdot \|_\infty$ norm in the role of $\| \cdot \|$ and the norm of $H$ which is “the maximum of $\| \cdot \|_2$-norms of the columns in $H$,” denoted by $\| H \|_{1,2}$. 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,

$$\Pr\{\|H^T \eta\|_\infty \geq \rho\} \leq 2N \text{Erfc}\left(\frac{\rho}{\sigma \|H\|_{1,2}}\right) \leq N e^{-\frac{\rho^2 \|H\|_{1,2}^2}{2}}, \quad (1.22)$$

where

$$\text{Erfc}(s) = \Pr_{\xi \sim \mathcal{N}(0,1)}\{\xi \geq s\} = \frac{1}{\sqrt{2\pi}} \int_s^\infty e^{-t^2/2} dt$$

is the (slightly rescaled) complementary 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 $\sigma \sqrt{\ln(N)} \|H\|_{1,2}$. In applications we consider in this chapter, we have $N = O(m)$, so that with $\sigma$ and $\|H\|_{1,2}$ given, typical values $\|H^T \eta\|_\infty$ are nearly independent of $m$. The bottom line is that $\ell_1$ 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 $\ell_1$ recoveries with the same pair $(H, \|\cdot\|)$ is concerned, the situation is as follows. Assume for the sake of simplicity that $(H, \|\cdot\|)$ satisfies $Q_{q}(s, \kappa)$ with some $s$ and $\kappa < 1/2$, and let the observation error be random. Given $\epsilon \in (0,1)$, let

$$\rho_\epsilon[H, \|\cdot\|] = \min \{\rho : \Pr\{\eta : \|H^T \eta\| \leq \rho\} \geq 1 - \epsilon\}; \quad (1.23)$$

this is nothing but the smallest $\rho$ such that

$$\Pr\{\eta \in \Xi_\rho\} \geq 1 - \epsilon \quad (1.24)$$

(see (1.14)) and thus – the smallest $\rho$ for which the error bound (1.15) for the regular $\ell_1$ recovery holds true with probability $1 - \epsilon$ (or at least the smallest $\rho$ for which the latter claim is supported by Theorem 1.3). With $\rho = \rho_\epsilon[H, \|\cdot\|]$, the regular $\ell_1$ recovery guarantees (and that is the best guarantee one can extract from Theorem 1.3) that

(\#) For some set $\Xi$, $\Pr\{\eta \in \Xi\} \geq 1 - \epsilon$, of “good” realizations of $\eta \sim \mathcal{N}(0,\sigma^2 I_m)$, one has

$$\|\hat{x}(Ax + \eta) - x\|_p \leq \frac{4(2s)^{\frac{p}{2}}}{1 - 2\kappa} \left[\rho_\epsilon[H, \|\cdot\|] + \frac{\|x - x^*\|_1}{2s}\right], \quad 1 \leq p \leq q, \quad (1.25)$$

whenever $x \in \mathbb{R}^n$ and $\eta \in \Xi_\rho$.

The error bound (1.19) (where we set $\kappa = \kappa$) says that (\#) holds true for the penalized $\ell_1$ recovery with $\lambda = 2s$. The latter observation suggests that the penalized $\ell_1$ 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 $\rho$ 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 $\ell_1$ 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
(1.25) is no more than the “worst, with confidence 1 − ϵ, case,” while the typical values of the quantity ∥Hᵀη∥ which indeed participates in the error bound (1.18) may be essentially smaller than ρ(H,∥·∥). Numerical experience fully supports the above claim: 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 λ ≥ 2s, so that a rough upper bound on s will do. However, that bound (1.18) deteriorates as λ grows.

Finally, we remark that when H is m × N and η ∼ N(0,σ²Iₘ), we have
\[ ρ[H,∥·∥ₘ] ≤ σ\text{ErfcInv}(\frac{ϵ}{2N})\|H\|₁,₂ ≤ σ\sqrt{2 \ln(N/ϵ)}\|H\|₁,₂ \]
(see (1.26)); here ErfcInv(δ) is the inverse complementary error function:
\[ \text{Erfc}(\text{ErfcInv}(δ)) = δ, \quad 0 < δ < 1. \]

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 = (z₁,...,zₙ) and want to recover this series under the assumption that the series is “nearly s-sparse in frequency domain,” that is, that
z = Fx with ∥x - xₘ∥₁ ≤ δ,
where F is the matrix of n × n Inverse Discrete Cosine Transform, xₘ 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 × n submatrix of F corresponding to the time instants t where zₜ is observed, our observation becomes
\[ y = Ax + σξ, \]
where ξ is the standard Gaussian noise. After the signal in frequency domain, that is, x, is recovered by ℓ₁ minimization, let the recovery be ˆx, we recover the signal in the time domain as ˆz = F ˆx. In Figure 1.3, we present four test signals, of different (near) sparsity, along with their regular and penalized ℓ₁ recoveries. The data in Figure 1.3 clearly show how the quality of ℓ₁ 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.

1.3 VERIFIABILITY AND TRACTABILITY ISSUES

Good news about ℓ₁ recovery stated in Theorems 1.3, 1.4, 1.5 are “conditional” – we assume that we are smart enough to point out a pair (H,∥·∥) satisfying
Figure 1.3: Regular and penalized $\ell_1$ recovery of nearly $s$-sparse signals. $\circ$: true signals, $+$: recoveries (to make the plots readable, one per eight consecutive vector’s entries is shown). Problem’s sizes are $m = 256$ and $n = 2m = 512$, noise level is $\sigma = 0.01$, deviation from $s$-sparsity is $\|x - x^*\|_1 = 1$, contrast pair is $(H = \sqrt{n/mA}, \|\cdot\|_{\infty})$. In penalized recovery, $\lambda = 2s$, parameter $\rho$ of regular recovery is set to $\text{ErfcInv}(0.005/n)$. 

<table>
<thead>
<tr>
<th>$s$</th>
<th>$|x - x^*|_2$</th>
<th>$|x - x^*|_{\infty}$</th>
<th>$|x - x^*|_{\infty}$</th>
<th>$|x - x^*|_2$</th>
<th>$|x - x^*|_{\infty}$</th>
<th>$|x - x^*|_{\infty}$</th>
<th>$|x - x^*|_2$</th>
<th>$|x - x^*|_{\infty}$</th>
<th>$|x - x^*|_{\infty}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>0.170</td>
<td>0.0239</td>
<td>0.0079</td>
<td>0.0679</td>
<td>0.0069</td>
<td>0.0010</td>
<td>0.0143</td>
<td>0.0107</td>
<td>0.0143</td>
</tr>
<tr>
<td>32</td>
<td>0.220</td>
<td>0.0323</td>
<td>0.0123</td>
<td>0.0673</td>
<td>0.0067</td>
<td>0.0011</td>
<td>0.0145</td>
<td>0.0109</td>
<td>0.0145</td>
</tr>
<tr>
<td>64</td>
<td>0.365</td>
<td>0.0608</td>
<td>0.0245</td>
<td>0.0812</td>
<td>0.0068</td>
<td>0.0012</td>
<td>0.0187</td>
<td>0.0112</td>
<td>0.0187</td>
</tr>
<tr>
<td>128</td>
<td>0.772</td>
<td>0.1229</td>
<td>0.0492</td>
<td>3.665</td>
<td>0.0160</td>
<td>0.0014</td>
<td>0.0283</td>
<td>0.0143</td>
<td>0.0283</td>
</tr>
</tbody>
</table>
condition $Q_1(s, \kappa)$ with $\kappa < 1/2$ (and condition $Q_2(s, \kappa)$ with a “moderate” $\kappa$ \footnote{Always satisfied with “large enough” $\kappa$, e.g., $\kappa = s$, but such values of $\kappa$ are 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 $Q_1(s, \kappa)$.}). 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 $Q_1(s, \kappa)$, $\kappa < 1/2$, nullspace property can be satisfied; and without the nullspace property, $\ell_1$ 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 $Q_2(s, \kappa)$ with given parameters.

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

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 $s \leq O(1) \ln(n/m)$ \footnote{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 with specific numbers; following the standard mathematical practice, we do not do it, partly from laziness, partly because particular values of these numbers in our context are irrelevant.}. 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 Gaussian or Rademacher distributions (i.e., when filling the matrix with independent realizations of a random variable which is either a standard (zero mean, unit variance) Gaussian one, or takes values $\pm 1$ 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 $Q_2(s, \kappa)$ with, say, $\kappa = 1/4$.

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 $Q_1(s, \kappa)$ with $\kappa < 1/2$) 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.\footnote{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 still do not know how to check $s$-goodness efficiently.}

The “common life” analogy of this 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.

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 $Q_2(s, \kappa)$. A bad news is 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 that 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 $m \leq n/2$.

### 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 1.6.** 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 $\delta$ and $k$, $\text{RIP}(\delta, k)$, if for every $k$-sparse $x \in \mathbb{R}^n$ one has

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

(1.27)

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

**Proposition 1.7.** 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

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

(1.28)

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

For proof, see Section 1.5.3.

**Proposition 1.8.** Let $A \in \mathbb{R}^{m \times n}$ satisfy $\text{RIP}(\delta, 2s)$ for some $\delta < 1/3$ and positive integer $s$. Then

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

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

For proof, see Section 1.5.4.

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

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

**Proposition 1.9.** 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

$$
\nu(s, q)[H] = \max_{j \leq n} \|\text{Col}_j[I - H^TA]\|_{s, q},
$$

(1.29)
where Col\(_j[C]\) is \(j\)-th column of matrix \(C\). Then
\[
\|w\|_{s,q} \leq s^{1/q}\|H^T A w\|_\infty + \nu_{s,q}[H]\|w\|_1 \; \forall w \in \mathbb{R}^n, \tag{1.30}
\]

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

**Proof** is immediate. Setting \(V = I - H^T A\), we have
\[
\|w\|_{s,q} = \|H^T A + V|w|\|_{s,q} \leq \|H^T A w\|_{s,q} + \|V w\|_{s,q} \\
\leq s^{1/q}\|H^T A w\|_\infty + \sum_j |w_j|\|\text{Col}_j[V]\|_{s,q} \leq s^{1/q}\|H^T A\|_\infty + \nu_{s,q}[H]\|w\|_1. \quad \square
\]

Observe that the function \(\nu_{s,q}[H]\) is an efficiently computable convex function of \(H\), so that the set
\[
\mathcal{H}^c_{s,q} = \{H \in \mathbb{R}^{m \times n} : \nu_{s,q}[H] \leq s^{\frac{1}{q}-1}\kappa\} \tag{1.31}
\]
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 \(Q_q(s, \kappa)\), that is, we can find contrast matrices making \(\ell_1\) minimization valid. Moreover, we can design contrast matrix, e.g., by minimizing over \(\mathcal{H}^c_{s,q}\) the function \(|H|_{1,2}\), thus optimizing the sensitivity of the corresponding \(\ell_1\) 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 1.9 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 \(\mathbb{R}^n\), and \(X \subset \mathbb{R}^n\) be a nonempty bounded polytope represented as
\[
X = \{x \in \text{Conv}\{g_1, \ldots, g_N\} : Ax = 0\},
\]
where \(\text{Conv}\{g_1, \ldots, g_N\} = \{\sum \lambda_i g_i : \lambda \geq 0, \sum \lambda_i = 1\}\) is the convex hull of vectors \(g_1, \ldots, g_N\). Our goal is to upper-bound the maximum \(\text{Opt} = \max_{x \in X} f(x)\); this is a meaningful problem, since precise maximizing a convex function over a polyhedron 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 \([I - H^T A]x = x\). As a result,
\[
\text{Opt} := \max_{x \in X} f(x) = \max_{x \in X} f([I - H^T A]x) \\
\leq \max_{x \in \text{Conv}\{g_1, \ldots, g_N\}} f([I - H^T A]x) \\
= \max_{j \leq N} f([I - H^T A]g_j).
\]

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

The result of Proposition 1.9 is inspired by this construction as applied to the nullspace property: given \(m \times n\) sensing matrix \(A\) and setting
\[
X = \{x \in \mathbb{R}^n : \|x\|_1 \leq 1, Ax = 0\} = \{x \in \text{Conv}\{\pm e_1, \ldots, \pm e_n\} : Ax = 0\}\]
(\(e_i\) are the basic orths in \(\mathbb{R}^n\)), \(A\) is \(s\)-good if and only if

\[
\text{Opt}_s := \max_{x \in X} \{ f(x) := \|x\|_{s,1} \} < 1/2.
\]

A verifiable sufficient condition for this, as yielded by the above construction, is the existence of an \(m \times n\) matrix \(H\) such that

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

or, which is the same,

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

This observation brings 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 \(Q_{\infty}(s, \kappa)\)

As we have already mentioned, the conditions \(Q_q(s, \kappa)\) 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 \(q = \infty\). Namely,

**Proposition 1.10.** 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 \(Q_{\infty}(s, \kappa)\), there exists an \(m \times n\) matrix \(H\) such that

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

(so that \((H, \|\cdot\|)\) satisfies \(Q_{\infty}(s, \kappa)\) by Proposition 1.9), and also

\[
\|H^T \eta\|_{\infty} \leq \|\bar{H}^T \eta\| \forall \eta \in \mathbb{R}^m.
\]

(1.32)

In addition, \(m \times n\) contrast matrix \(H\) such that the pair \((H, \|\cdot\|)\) satisfies the condition \(Q_{\infty}(s, \kappa)\) with as small \(\kappa\) as possible can be found as follows. Consider \(n\) LP programs

\[
\text{Opt}_i = \min_{\nu, h} \nu : \|A^T h - e^i\|_{\infty} \leq \nu
\]

where \(e^i\) is \(i\)-th basic orth of \(\mathbb{R}^n\). Let \(\text{Opt}_i, h_i, i = 1, \ldots, n\) be optimal solutions to these problems, we set \(H = [h_1, \ldots, h_n]\); the corresponding value of \(\kappa\) is

\[
\kappa_s = s \max_i \text{Opt}_i.
\]

Besides this, there exists a transparent alternative description of the quantities \(\text{Opt}_i\) (and thus – of \(\kappa_s\)); specifically,

\[
\text{Opt}_i = \max_x \{ x_i : \|x\|_1 \leq 1, Ax = 0 \}.
\]

(1.33)

For proof, see Section 1.5.5.
Taken along with (1.32) and error bounds of Theorems 1.3, 1.4, 1.5, Proposition 1.10 says that

As far as the condition $Q_\infty(s, \kappa)$ is concerned, we lose nothing when restricting ourselves with pairs $(H \in \mathbb{R}^{m \times n}, \| \cdot \|_\infty)$ and contrast matrices $H$ satisfying the condition

$$\| I_n - H^T A \|_{1,ij} \leq s^{-1} \kappa$$

(1.34)

implying that $(H, \| \cdot \|_\infty)$ satisfies $Q_\infty(s, \kappa)$.

A good news is that (1.34) is an explicit convex constraint on $H$ (in fact, even on $H$ and $\kappa$), 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 $Q_\infty(s, \kappa)$ (and, perhaps, additional convex constraints on $H$ and $\kappa$).

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

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

(1.35)

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, $\mu(A)$ is small when the columns of $A$ are nearly mutually 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 (1.34):

**Proposition 1.11.** Let $A$ be an $m \times n$ matrix, and let the columns of an $m \times n$ matrix $H$ be given by

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

Then

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

(1.36)

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 $I - H^T A$ 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 $I - H^T A$ are $\leq \frac{\mu(A)}{1 + \mu(A)}$ as well, implying (1.36). The "in particular" claim

---

\footnote{As far as $\ell_1$ minimization is concerned, this normalization is non-restrictive: we always can enforce it by diagonal scaling of the signal underlying observations (1.1), and $\ell_1$ minimization in scaled variables is the same as weighted $\ell_1$ minimization in original variables.}
is given by (1.36) combined with Proposition 1.9.

1.3.3.2 From RIP to conditions $Q_q(\cdot, \kappa)$

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

**Proposition 1.12.** 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

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

the pair $(H = t^{\frac{1}{2}} I_m, \| \cdot \|_2)$ satisfies $Q_q(t, \kappa)$;

(ii) Whenever a positive integer $t$ satisfies (1.37), the pair $(H = \frac{s^{\frac{1}{2}} t^{\frac{1}{2}}}{1 + s} A, \| \cdot \|_\infty)$ satisfies $Q_q(t, \kappa)$.

For proof, see Section 1.5.4.

The most important consequence of Proposition 1.12 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 $Q_\infty(t, \frac{1}{3})$ – there exists a pair $(H, \| \cdot \|_\infty)$ satisfying this condition. The point is that by Proposition 1.10, if the condition $Q_\infty(t, \frac{1}{3})$ 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 $Q_\infty$ 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 1.13.** 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

$$\|\text{Col}_j[I_n - H^T A]\|_{s,q} < \frac{1}{2} s^{\frac{1}{s^{q-1}}}, 1 \leq j \leq n,$$

one has

$$s \leq \sqrt{m}.$$  \hspace{1cm} (1.38)

As a result, sufficient condition for the validity of $Q_q(s, \kappa)$ with $\kappa < 1/2$ from Proposition 1.9 can never be satisfied when $s > \sqrt{m}$. Similarly, the verifiable sufficient condition $Q_\infty(s, \kappa)$, $\kappa < 1/2$ for $s$-goodness of $A$ cannot be satisfied when $s > \sqrt{m}$.

For proof, see Section 1.5.6.

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
SPARSE RECOVERY VIA $\ell_1$ MINIMIZATION

Figure 1.4: Erroneous $\ell_1$ recovery of a 25-sparse signal, no observation noise. Top: frequency domain, o – true signal, + – recovery. Bottom: time domain.

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 $256 \times 512$ randomly selected submatrix $A$ of the matrix of $512 \times 512$ Inverse Discrete Cosine Transform which we used in experiments reported in Figure 1.3. These experiments exhibit nice performance of $\ell_1$ 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 $w_2$ 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 $\|w_3\|_{s,1} \geq \|w_2\|_{s,1}$, etc. 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. We 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 gets stuck — the observed progress in the objective falls below a given threshold, say, 1.e-6. When it happens we can restart the process from a new randomly selected in $W$ starting point, after getting stuck, restart again, 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 it may happen that previous experiments with recovering 64-sparse signals went fine, when in fact some 25-sparse signals cannot be recovered by $\ell_1$ minimization even in the ideal noiseless case. The answer is simple: in our experiments, we dealt with randomly selected signals, and typical randomly selected data are much nicer, whatever be the purpose of a numerical experiment, than the worst-case data.

It is interesting to understand also which goodness we can certify using our verifiable sufficient conditions. Computations show that the fully verifiable (and strongest in our scale of sufficient conditions for $s$-goodness) condition $Q_\infty(s, \kappa)$ can be satisfied with $\kappa < 1/2$ when $s$ is as large as 7 and $\kappa = 0.4887$, and cannot be satisfied with $\kappa < 1/2$ when $s = 8$. As about Mutual Incoherence, it can only justify 3-goodness, no more. We can hardly be happy with the resulting bounds — goodness at least 7 and at most 24; however, it could be worse...

1.4 EXERCISES FOR CHAPTER 1

Exercise 1.1.

$k$-th Hadamard matrix, $H_k$ (here $k$ is nonnegative integer) is the $n_k \times n_k$ matrix, $n_k = 2^k$, given by the recurrence

$$H_0 = [1]; H_{k+1} = \begin{bmatrix} H_k & H_k \\ \overline{H_k} & -H_k \end{bmatrix}$$

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

1. Check that $H_k$ is symmetric matrix with entries $\pm 1$, and columns of the matrix are mutually orthogonal, so that $H_k/\sqrt{m_k}$ is an orthogonal matrix.
2. Check that when $k > 0$, $H_k$ has just two distinct eigenvalues, $\sqrt{m_k}$ and $-\sqrt{m_k}$, each of multiplicity $m_k := 2^{k-1} = n_k/2$.
3. Prove that whenever $f$ is an eigenvector of $H_k$, one has

$$\|f\|_\infty \leq \|f\|_1/\sqrt{m_k}.$$  

Derive from this observation the conclusion as follows:

Let $a_1, \ldots, a_{m_k} \in \mathbb{R}^{n_k}$ be orthogonal to each other unit vectors which are eigenvectors of $H_k$ with eigenvalues $\sqrt{m_k}$ (by the above, the dimension of the eigenspace of $H_k$ associated with the eigenvalue $\sqrt{m_k}$ is $m_k$, so that the required $a_1, \ldots, a_{m_k}$ do exist), and let $A$ be the $m_k \times n_k$ matrix with
SPARSE RECOVERY VIA $\ell_1$ MINIMIZATION

the rows $a_1^T, \ldots, a_m^T$. For every $x \in \text{Ker } A$ it holds

$$\|x\|_\infty \leq \frac{1}{\sqrt{n_k}} \|x\|_1,$$

whence $A$ satisfies the nullspace property whenever the sparsity $s$ satisfies $2s < \sqrt{m_k} = \sqrt{2m_k}$. 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 $Q_\infty(s, \kappa_s = s/\sqrt{n_k}, < 1/2$,

and the $\| \cdot \|_2$-norms of columns of $H_k$ do not exceed $\sqrt{2 \sqrt{n_k} + 1}/\sqrt{n_k}$.

Note that the above conclusion yields a sequence of individual $(m_k = 2^{k-1}) \times (n_k = 2^k)$ sensing matrices, $k = 1, 2, \ldots$, with “size ratio” $n_k/m_k = 2$, which make an efficiently verifiable condition for $s$-goodness, say, $Q_\infty(s, \frac{1}{3})$ satisfiable in basically the entire range of values of $s$ allowed by Proposition 1.13. 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 ($m = 512$) \times $(n = 1024)$ sensing matrix $\bar{A}$ such that the efficiently verifiable condition $Q_\infty(15, \frac{15}{32})$ 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 are the certified by these incoherences goodness levels. What happens when the matrices are the Gaussian (independent $\mathcal{N}(0,1)$ entries) and the Rademacher ones (independent entries taking values $\pm 1$ with probabilities $1/2$)?

2. Generate at random a sample of $m \times n$ matrices with independent $\mathcal{N}(0,1/m)$ entries. Proposition 1.7 suggests that a sampled matrix $A$ has good chances to satisfy RIP($\delta, k$) with some $\delta < 1/3$ and some $k$, and thus to be $s$-good (and even more than this, see Proposition 1.8) for every $s \leq k/2$. Of course, given $A$ we cannot check whether the matrix indeed satisfies RIP($\delta, k$) with given $\delta, k$; what we can try to do is to certify that RIP($\delta, k$) does not 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 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 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 $\omega_i$ is an integer from the range \{0, 1, ..., $\nu_k - 1$\}; the group operation, denoted by $\oplus$, is
\[
(\omega_1, ..., \omega_k) \oplus (\omega'_1, ..., \omega'_k) = ((\omega_1 + \omega'_1) \mod \nu_1, ..., (\omega_k + \omega'_k) \mod \nu_k),
\]
where $a \mod b$ is the remainder, taking values in \{0, 1, ..., $b - 1$\}, in the division of an integer $a$ by positive integer $b$; say, $5 \mod 3 = 2$, and $6 \mod 3 = 0$. 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 $\chi(\omega \odot \omega') = \chi(\omega)\chi(\omega')$ 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 $\mu_i$ are restricted to be roots of degree $\nu_i$ 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 $F$ of size $n_k \times n_k$: the columns in the matrix are indexed by the elements $\omega$ of $G$, the rows – by the characters $\chi \in G_*$ of $G$, and the element in cell $(\chi, \omega)$ is $\chi(\omega)$. The standard example here corresponds to $k = 1$, in which case $F$ clearly is the $\nu_1 \times \nu_1$ matrix of Discrete Fourier Transform.

Now goes the exercise:

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

2. Let $\tilde{\omega}, \tilde{\omega}'$ be two elements of $G$. Prove that there exists a permutation $\Pi$ of elements of $G$ which maps $\tilde{\omega}$ into $\tilde{\omega}'$ and is such that
\[
\text{Col}_{\Pi(\omega)}[F] = D\text{Col}_{\omega}[F] \forall \omega \in G,
\]
where $D$ is diagonal matrix with diagonal entries $\chi(\tilde{\omega}')/\chi(\tilde{\omega})$, $\chi \in G_*$. Note

3. Consider the special case of the above construction where $\nu_1 = \nu_2 = ... = \nu_k = 2$. Verify that in this case $F$, 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 $H_k$.

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 $H_k$ 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 $Q_\infty(s, \kappa)$ with as small $\kappa$ as possible; by Proposition 1.10, to this end we should solve $n$ LP programs
\[
\text{Opt}_i = \min_h \|e^i - A^T h\|_\infty,
\]
where $e^i$ is $i$-th basic orth in $\mathbb{R}^n$. Prove that with $A$ coming from $H_k$, 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 1.13 states that the verifiable condition $Q_\infty(s, \kappa)$ 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 $Q_\infty(s, \kappa)$ with $\kappa < 1/2$ is satisfiable when $s$ is as large as $O(1)n/\sqrt{m^o}$.

1. Let $n = 2^kp$ 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 $\{1, \ldots, n = 2^kp\}$ into $2^k$ consecutive groups $I_1, \ldots, I_{2^k}$, $p$ 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 $\mathbb{R}^{2m^o}$ comprising of all eigenvectors, with eigenvalue $\sqrt{2^{k^o}}$, of the Hadamard matrix $H_k$, see Exercise 1.11, so that the dimension of $U$ is $2^{k-1} = m^o$, and let $L$ be given by
   \[
   L = \{u^+: u \in U\} \subset \mathbb{R}^n.
   \]
   Clearly, $L$ is a linear subspace in $\mathbb{R}^n$ 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 $\ker A = L$, then the verifiable sufficient condition $Q_\infty(s, \kappa)$ does certify $s$-goodness of $A$ whenever
   \[
   1 \leq s < \frac{n}{2\sqrt{2m^o}}.
   \]

2. Let $L$ be an $m^o$-dimensional subspace in $\mathbb{R}^n$. Prove that $L$ contains a nonzero vector $x$ with
   \[
   \|x\|_\infty \geq \frac{\sqrt{m^o}}{n}\|x\|_1,
   \]
   so that the condition $Q_\infty(s, \kappa)$ 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 $H_k$ 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 $Q_\infty(s, \kappa)$; call $s(m)$ the resulting value of $s$. 

\[\text{StatOptNoSolutions July 2, 2019 7x10}\]
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 RIP$(1/3, 2s)$; call $\hat{s}(m)$ 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 $\hat{s}(m)$; in other words, we again want to understand how “theoretically perfect” RIP compares to “conservative restricted scope” condition $Q_\infty$.

### 1.5 PROOFS

#### 1.5.1 Proofs of Theorem 1.3, 1.4

All we need is to prove Theorem 1.4, since Theorem 1.3 is the particular case $\kappa = n < 1/2$ of Theorem 1.4.

Let us fix $x \in \mathbb{R}^n$ and $\eta \in \Xi_\rho$, and let us set $\hat{x} = \hat{x}_{\text{reg}}(Ax + \eta)$. 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 $w \in \mathbb{R}^n$, $w_I$ 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 $z = \hat{x} - x$.

1°. By the definition of $\Xi_\rho$ and due to $\eta \in \Xi_\rho$, we have

$$
\|H^T([Ax + \eta] - Ax)\| \leq \rho, \quad (1.40)
$$

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

$$
\|\hat{x}_{I^o}\|_1 = \|\hat{x}\|_1 - \|\hat{x}_I\|_1 \leq \|x\|_1 - \|\hat{x}_I\|_1 = \|x_I\|_1 + \|x_{I^o}\|_1 - \|\hat{x}_I\|_1 \\
\leq \|z_I\|_1 + \|x_{I^o}\|_1, 
$$

(1.41)

and therefore

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

It follows that

$$
\|z\|_1 = \|z_I\|_1 + \|z_{I^o}\|_1 \leq 2\|z_I\|_1 + 2\|x_{I^o}\|_1. \quad (1.42)
$$

Further, by definition of $\hat{x}$ we have $\|H^T([Ax + \eta] - A\hat{x})\| \leq \rho$, which combines with (1.40) to imply that

$$
\|H^T A(\hat{x} - x)\| \leq 2\rho. \quad (1.43)
$$

2°. Since $(H, \| \cdot \|)$ satisfies $Q_1(s, \kappa)$, we have

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

By (1.43), it follows that $\|z\|_{s,1} \leq 2s\rho + \kappa\|z\|_1$, which combines with the evident inequality $\|z_I\|_1 \leq \|z\|_{s,1}$ (recall that $\text{Card}(I) = s$) and with (1.42) to imply that

$$
\|z_I\|_1 \leq 2s\rho + \kappa\|z\|_1 \leq 2s\rho + 2\kappa\|z_I\|_1 + 2\kappa\|x_{I^o}\|_1,
$$
whence
\[ \|z_I\|_1 \leq \frac{2s\rho + 2\kappa\|x_I\|_1}{1 - 2\kappa}. \]

Invoking (1.42), we conclude that
\[ \|z\|_1 \leq \frac{4s}{1 - 2\kappa} \left( \rho + \|x_I\|_1 \right). \]  \hspace{1cm} (1.44)

3°. Since \((H, \|\cdot\|)\) satisfies \(Q_{q}(s, \kappa)\), we have
\[ \|z\|_{s,q} \leq \frac{s^\frac{1}{q}}{\|H^T A z\|} + \kappa s^\frac{1}{q} \|z\|_1, \]
which combines with (1.44) and (1.43) to imply that
\[ \|z\|_{s,q} \leq \frac{s^\frac{1}{q}}{\|H^T A z\|} + \kappa s^\frac{1}{q} \|z\|_1 \leq \frac{4s^\frac{1}{q} \|H^T A z\|}{1 - 2\kappa} \left( \rho + \frac{\|x_I\|_1}{2s} \right). \]  \hspace{1cm} (1.45)

(we have taken into account that \(\kappa < 1/2\) and \(\kappa \geq \kappa\)).

Let \(\theta\) be the \((s+1)\)st largest magnitude of entries in \(z\), and let \(w = z - z^s\). Now (1.45) implies that
\[ \theta \leq \|z\|_{s,q} \leq \frac{4[1 + \kappa - \xi]}{1 - 2\kappa} \left( \rho + \frac{\|x_I\|_1}{2s} \right). \]

Hence invoking (1.44) we have
\[ \|w\|_q \leq \left( \|w\|_{\infty} \right)^\frac{1}{q} \leq \left( \theta^\frac{1}{q} \right)^\frac{1}{q} \leq \left( \rho + \frac{\|x_I\|_1}{2s} \right)^\frac{1}{q} \leq \frac{4s^\frac{1}{q} \|H^T A z\|}{1 - 2\kappa} \left( \rho + \frac{\|x_I\|_1}{2s} \right)^\frac{1}{q}. \]

Taking into account (1.45) and the fact that the supports of \(z^s\) and \(w\) do not intersect, we get
\[ \|z\|_q \leq 2^\frac{1}{q} \max \{\|z^s\|_q, \|w\|_q\} = 2^\frac{1}{q} \max \{\|z\|_{s,q}, \|w\|_q\} \leq \frac{4(2s)^\frac{1}{q} \|H^T A z\|}{1 + \kappa - \xi} \left( \rho + \frac{\|x_I\|_1}{2s} \right)^\frac{1}{q}. \]

This bound combines with (1.44), the Moment inequality,\(^{12}\) and with the relation \(\|x_I\|_1 = \|x - x^s\|_1\) to imply (1.16).

\(\square\)

1.5.2 Proof of Theorem 1.5

Let us prove (i). Let us fix \(x \in \mathbb{R}^n\) and \(\eta\), and let us set \(\hat{x} = \hat{x}_{\text{pen}}(Ax + \eta)\). 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 \(w \in \mathbb{R}^n\) let \(w_I, w_{I^o}\) be the vectors

\(^{12}\)Moment inequality states that if \((\Omega, \mu)\) is a space with measure and \(f\) is a \(\mu\)-measurable real-valued function on \(\Omega\), then \(\phi(\rho) = \ln \left( \int_{\Omega} |f(\omega)|^\frac{1}{p} \mu(d\omega) \right)^p\) is convex function of \(\rho\) on every segment \(\Delta \subset [0, 1]\) such that \(\phi(\cdot)\) is well defined at the endpoints of \(\Delta\). As a corollary, when \(x \in \mathbb{R}^n\) and \(1 \leq p \leq q \leq \infty\), one has \(\|x\|_p \leq \|x\|_1^\frac{q-p}{q} \|x\|_q^\frac{p}{q}\).
obtained from \( w \) by zeroing out all entries with indexes not in \( I \), respectively, not in \( I^o \). Finally, let \( z = \hat{x} - x \) and \( \nu = \|H^T \eta\|\).

1°. We have

\[
\|\hat{x}\|_1 + \lambda \|H^T (A\hat{x} - Ax - \eta)\| \leq \|x\|_1 + \lambda \|H^T \eta\|
\]

and

\[
\|H^T (A\hat{x} - Ax - \eta)\| = \|H^T (Az - \eta)\| \geq \|H^T Az\| - \|H^T \eta\|
\]

whence

\[
\|\hat{x}\|_1 + \lambda \|H^T Az\| \leq \|x\|_1 + 2\lambda \|H^T \eta\| = \|x\|_1 + 2\lambda \nu.
\] (1.46)

We have

\[
\|\hat{x}\|_1 = \|x + z\|_1 = \|x_I + z_I\|_1 + \|x_{I^o} + z_{I^o}\|_1 \\
\geq \|x_I\|_1 - \|z_I\|_1 + \|z_{I^o}\|_1 - \|x_{I^o}\|_1,
\]

which combines with (1.46) to imply that

\[
\|x_I\|_1 + \|z_{I^o}\|_1 + \lambda \|H^T Az\| \leq \|x\|_1 + 2\lambda \nu,
\]

or, which is the same,

\[
\|z_{I^o}\|_1 - \|z_I\|_1 + \lambda \|H^T Az\| \leq 2\|x_{I^o}\|_1 + 2\lambda \nu.
\] (1.47)

Since \((H, \|\cdot\|)\) satisfies \(Q_s(\kappa)\), we have

\[
\|z_I\|_1 \leq \|z\|_{s,1} \leq s \|H^T Az\| + \kappa \|z\|_1,
\]

so that

\[
(1 - \kappa)\|z_I\|_1 - \kappa \|z_{I^o}\|_1 - s \|H^T Az\| \leq 0.
\] (1.48)

Taking weighted sum of (1.47) and (1.48), the weights being 1, 2, respectively, we get

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

whence, due to \( \lambda \geq 2s\),

\[
\|z\|_1 \leq \frac{2\lambda \nu + 2\|x_{I^o}\|_1}{1 - 2\kappa} \leq \frac{2\lambda}{1 - 2\kappa} \left[\nu + \frac{\|x_{I^o}\|_1}{2s}\right].
\] (1.49)

Further, by (1.46) we have

\[
\lambda \|H^T Az\| \leq \|x\|_1 - \|\hat{x}\|_1 + 2\lambda \nu \leq \|z\|_1 + 2\lambda \nu,
\]

which combines with (1.49) to imply that

\[
\lambda \|HA^T z\| \leq \frac{2\lambda \nu + 2\|x_{I^o}\|_1}{1 - 2\kappa} + 2\lambda \nu = \frac{2\lambda \nu(2 - 2\kappa) + 2\|x_{I^o}\|_1}{1 - 2\kappa}.
\] (1.50)

From \(Q_q(s, \kappa)\) it follows that

\[
\|z\|_{s,q} \leq s^{\frac{1}{q}} \|H^T Az\| + \kappa s^{\frac{1}{q} - 1} \|z\|_1,
\]
which combines with (1.50) and (1.49) to imply that
\[
\|z\|_{s,q} \leq s^{-\frac{1}{q}} \left[ \|H^TAz\| + \kappa \|z\|_1 \right] \leq s^{-\frac{1}{q}} \left[ \frac{4s(1-\kappa)\|x_p\|_1}{1-2\kappa} + \frac{\kappa(2\lambda s + \frac{1}{2}\|x_p\|_1)}{1-2\kappa} \right]
= s^{-\frac{1}{q}} \left[ \frac{4(1-\kappa)+2s\lambda - \kappa}{1-2\kappa} \|x_p\|_1 \right] \leq 4s^{-\frac{1}{q}} \left[ 1 + \frac{\lambda s}{2\kappa} - \kappa \right] \left[ \nu + \frac{\|x_p\|_1}{2s} \right]
\]
(recall that \(\lambda \geq 2s, \kappa \geq \varkappa, \) and \(\varkappa < 1/2\). It remains to repeat the reasoning following (1.45) in item 3° of the proof of Theorem 1.4. Specifically, denoting by \(\theta\) the \((s+1)\)-st largest magnitude of entries in \(z\), (1.51) implies that
\[
\theta \leq s^{-1/q} \|z\|_{s,q} \leq \frac{4}{1-2\varkappa} [1 + \kappa \frac{\lambda}{2s} - \varkappa] \left[ \nu + \frac{\|x_p\|_1}{2s} \right],
\]
so that for the vector \(w = z - z^*\) one has
\[
\|w\|_q \leq \theta^{1/2} \|w\|_1^\frac{1}{q} \leq \frac{4(\lambda/2)^\frac{1}{q}}{1-2\varkappa} [1 + \kappa \frac{\lambda}{2s} - \varkappa] \left[ \nu + \frac{\|x_p\|_1}{2s} \right]
\]
(we have used (1.52) and (1.49)). Hence, taking into account that \(z^*\) and \(w\) have non-intersecting supports,
\[
\|z\|_q \leq 2^{\frac{1}{2}} \max[\|z^*\|_q, \|w\|_q] = 2^{\frac{1}{2}} \max[\|z\|_{s,q}, \|w\|_q] \leq \frac{4(\lambda/2)^\frac{1}{q}}{1-2\varkappa} [1 + \kappa \frac{\lambda}{2s} - \varkappa] \left[ \nu + \frac{\|x_p\|_1}{2s} \right]
\]
(we have used (1.51) along with \(\lambda \geq 2s\) and \(\kappa \geq \varkappa\). This combines with (1.49) and Moment inequality to imply (1.18). All remaining claims of Theorem 1.5 are immediate corollaries of (1.18).

1.5.3 Proof of Proposition 1.7

1°. Assuming \(k \leq m\) and selecting a set \(I\) of \(k\) distinct from each other indexes from \(\{1, \ldots, n\}\), 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 \(\mathbb{R}^k\). 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)
\[
\ln \left( \mathbb{E} \{ \exp(\gamma \zeta) \} \right) = m \ln \left( \frac{1}{\sqrt{2\pi}} \int e^{\gamma s^2 - \frac{1}{2} s^2} ds \right) = -\frac{m}{2} \ln(1 - 2\gamma).
\]
Given \(\alpha \in (0, 0.1]\) and selecting \(\gamma\) in such a way that \(1 - 2\gamma = \frac{1}{1+\alpha}\), we get \(0 < \gamma < 1/2\) and therefore
\[
\text{Prob}\{\zeta_u > m(1+\alpha)\} \leq \mathbb{E} \{ \exp(\gamma \zeta_u) \} \exp(-m\gamma(1+\alpha))
= \exp\left(-\frac{m}{2} \ln(1 - 2\gamma) - m\gamma(1+\alpha)\right)
= \exp\left(-\frac{m}{2} \ln(1 + \alpha) - \alpha\right) \leq \exp\left(-\frac{m}{2} \alpha^2\right),
\]
and similarly, selecting \(\gamma\) in such a way that \(1 - 2\gamma = \frac{1}{1-\alpha}\), we get \(-1/2 < \gamma < 0\).
and therefore
\[
\text{Prob}\{\zeta_u < m(1 - \alpha)\} \leq E\{\exp\{\gamma \zeta_u\}\} \exp\{-m\gamma(1 - \alpha)\}
\]
\[
= \exp\{-\frac{m}{2} \ln(1 - 2\gamma) - m\gamma(1 - \alpha)\}
\]
\[
= \exp\{\frac{m}{2} \ln(1 - \alpha) + \alpha\} \leq \exp\{-\frac{m^2}{2}\alpha^2\},
\]
and we end up with
\[
u \in \mathbb{R}^k, \|u\|_2 = 1 \Rightarrow \begin{cases}
\text{Prob}\{A : \|A_I u\|_2^2 > 1 + \alpha\} \leq \exp\{-\frac{m^2}{2}\alpha^2\}, \\
\text{Prob}\{A : \|A_I u\|_2^2 < 1 - \alpha\} \leq \exp\{-\frac{m}{2}\alpha^2\}.
\end{cases} \tag{1.53}
\]

2°. 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 \(\epsilon\)-net on the unit sphere \(S\) in \(\mathbb{R}^k\) as follows. We start with a point \(u_1 \in S\); after \(\{u_1, \ldots, u_t\} \subset S\) 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 \(\{u_1, \ldots, u_t\}\). By compactness of \(S\) and due to \(\epsilon > 0\), this process eventually terminates; upon termination, we have at our disposal collection \(\{u_1, \ldots, u_N\}\) 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 \(\epsilon\) from some point of the collection. We claim that the cardinality \(N\) of the resulting set can be bounded as
\[
N \leq \left[\frac{2 + \epsilon}{\epsilon}\right]^k = \left[\frac{4 + 9\alpha}{\alpha}\right]^k \leq \left(\frac{5}{\alpha}\right)^k. \tag{1.54}
\]
Indeed, the interiors of the \(\|\cdot\|_2\)-balls of radius \(\epsilon/2\) centered at the points \(u_1, \ldots, u_N\) 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 (1.54).

3°. Consider event \(E\) comprised of all realizations of \(A\) such that for all \(k\)-element subsets \(I\) of \(\{1, \ldots, n\}\) and all \(t \leq n\) it holds
\[
1 - \alpha \leq \|A_I u_t\|_2^2 \leq 1 + \alpha. \tag{1.55}
\]
By (1.53) and the union bound,
\[
\text{Prob}\{A \notin E\} \leq 2N\binom{n}{k} \exp\{-\frac{m}{5}\alpha^2\}. \tag{1.56}
\]
We claim that
\[
A \in E \Rightarrow (1 - 2\alpha) \leq \|A_I u\|_2^2 \leq 1 + 2\alpha \forall \left(\begin{array}{c}
I \subset \{1, \ldots, n\} : \text{Card}(I) = k \\
u \in \mathbb{R}^k : \|u\|_2 = 1
\end{array}\right). \tag{1.57}
\]
Indeed, let \(A \in E\), let us fix \(I \in \{1, \ldots, n\}\), \(\text{Card}(I) = k\), 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 \(\mathbb{R}^k\). 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 $\epsilon$. 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 (1.57). Now let $u$ be unit vector in $\mathbb{R}^k$, and $u_s$ be a point in the net at the $\| \cdot \|_2$-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 (1.57).

The bottom line is:

$$\delta \in (0, 0.2], 1 \leq k \leq n \Rightarrow \text{Prob}\{A : A \text{ does not satisfy RIP}(\delta, k)\} \leq 2\left(\frac{10}{\delta}\right)^k \binom{n}{k} \exp\{-\frac{m\delta^2}{20}\}. \quad (1.58)$$

Indeed, setting $\alpha = \delta/2$, we have seen that whenever $A \notin E$, we have $(1 - \delta) \leq \|Au\|_2^2 \leq (1 + \delta)$ for all $k$-sparse $u$, which is nothing but RIP($\delta, k$); with this in mind, (1.58) follows from (1.56) and (1.54).

4$^\circ$. It remains to verify that with properly selected, depending solely on $\delta$, positive quantities $c, d, f$, for every $k \geq 1$ satisfying (1.28) the right hand side in (1.58) is at most $\exp\{-fm\}$. Passing to logarithms, our goal is to ensure the relation

$$G := a(\delta)m - b(\delta)k - \ln \binom{n}{k} \geq mf(\delta) > 0$$

provided that $k \geq 1$ satisfies (1.28).

Let $k$ satisfy (1.28) 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

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

whence

$$C \leq n \left[\frac{m}{n} y \ln \left(\frac{n}{my}\right) + \frac{n-k}{n} \ln \left(\frac{n}{n-k}\right)\right] \leq \frac{m}{ny} \leq \frac{m}{ny} \leq \frac{1}{ny} \leq \frac{2m}{ny} \leq \frac{2my}{ny}.$$
(recall that \(n \geq m\) and \(y \leq 1/3\)). It follows that

\[
G = a(\delta)m - b(\delta)k - C \geq a(\delta)m - b(\delta)ym - 2my \ln\left(\frac{n}{my}\right)
\]

and all we need is to select \(c, d\) in such a way that (1.28) 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 \(2y \ln(1/y) + b(\delta)y \leq \frac{1}{3}u(\delta)\); selecting \(d(\delta) \geq 3\) large enough, (1.28) would imply \(y \leq u(\delta)\), and thus would imply

\[
H \geq \frac{2}{3}a(\delta) - 2y \ln\left(\frac{n}{m}\right).
\]

Now we can select \(c(\delta)\) large enough for (1.28) to ensure that \(2y \ln(\frac{n}{m}) \leq \frac{1}{3}u(\delta)\). With just specified \(c, d\), (1.28) implies that \(H \geq \frac{1}{3}a(\delta)\), and we can take the latter quantity as \(f(\delta)\). \(\square\)

### 1.5.4 Proof of Propositions 1.8 and 1.12

Let \(x \in \mathbb{R}^n\), and let \(x^1, \ldots, x^q\) be 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 \(x = x^1 + \ldots + x^q\). 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

\[
\|u^T A^T A u\| \leq \delta\|u\|_2\|v\|_2.
\]

(\(*\)

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 \(v^T A^T A u = v^T [A_I^T A_I] u_I\). 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 \(u, z\) of their coordinates in the orthonormal eigenbasis of \(Q\), we get \(|v^T A^T A u| = |\sum_1^s \lambda_i w_i z_i| = |\sum_1^s w_i z_i + \sum_1^s \mu_i w_i z_i| \leq |w^T z| + \delta\|w\|_2\|z\|_2\). It remains to note that \(w^T z = u^T e_I = 0\) and \(\|w\|_2 = \|u\|_2\), \(\|z\|_2 = \|v\|_2\).

We have

\[
\|A x^1\|_2\|A x\|_2 \geq \|x^1\|_2 A^T A x = \|A x^1\|_2^2 + \sum_{j=2}^q \|x^j\|_2\|A x^j\|_2 \text{ by (*)}
\]

\[
\geq \|A x^1\|_2^2 - \delta \sum_{j=2}^q \|x^j\|_2 \|A x^j\|_2 \text{ by (*)}
\]

\[
\geq \|A x^1\|_2^2 - \delta s^{-1/2}\|x^1\|_2 \sum_{j=2}^q \|x^j-1\|_1 \geq \|A x^1\|_2^2 - \delta s^{-1/2}\|x^1\|_2 \|x\|_1
\]

\[
\Rightarrow \|A x^1\|_2 \leq \|A x^1\|_2 \|A x\|_2 + \delta s^{-1/2}\|x\|_2 \|x\|_1
\]

\[
\Rightarrow \|x\|_2 = \frac{\|x\|_2}{\|A x\|_2} \|A x\|_2 \leq \frac{\|x\|_2}{\|A x\|_2} \|A x\|_2 + \delta s^{-1/2} \left(\frac{\|x\|_2}{\|A x\|_2}\right)^2 \|x\|_1
\]

\[
\Rightarrow \|x\|_s,2 = \|x\|_2 \leq \frac{1}{\sqrt{1-\delta}} \|A x\|_2 + \frac{\delta s^{-1/2}}{1-\delta} \|x\|_1
\]

(\(!\)

[by RIP(\(\delta, 2s\))]

\]
and we see that the pair \( (H = \frac{t^{-1/2}}{\sqrt{1/\delta}} I_m, \| \cdot \|) \) satisfies \( Q_2(s, \frac{\delta}{1-\delta}) \), as claimed in Proposition 1.8.i. Moreover, when \( q \geq 2, \kappa > 0 \) and integer \( t \geq 1 \) satisfy \( t \leq s \) and \( \kappa t^{1/q-1} \geq \frac{\delta s^{-1/2}}{1-\delta} \), by (!) 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,
\[
1 \leq t \leq \min \left[ \left[ \frac{\kappa(1-\delta)}{\delta} \right]^{\frac{q}{2-\delta}}, s \frac{s^{-2}}{2} \right] s \frac{s^{-2}}{2}
\]
\[
\Rightarrow (H = \frac{t^{-1/2}}{\sqrt{1/\delta}} I_m, \| \cdot \|) \text{ satisfies } Q_q(t, \kappa),
\]
as required in Proposition 1.12.i.

Next, we have
\[
\|x\|_1 \| A^T Ax \|_\infty \geq [x^1]^T A^T Ax = \|Ax\|_2 + \sum_{j=2}^q [x^1]^T A^T Ax^j \geq \|Ax\|_2^2 - \delta s^{-1/2} \|x^1\|_2 \|x\|_1 \quad \text{[exactly as above]}
\]
\[
\Rightarrow \|Ax\|_2^2 \leq \|x^1\|_1 \| A^T Ax \|_\infty + \delta s^{-1/2} \|x^1\|_2 \|x\|_1
\]
\[
\Rightarrow (1-\delta) \|x^1\|_2^2 \leq \|x^1\|_1 \| A^T Ax \|_\infty + \delta s^{-1/2} \|x^1\|_2 \|x\|_1 \quad \text{[by RIP(\delta, 2s)]}
\]
\[
\leq s^{1/2} \|x^1\|_2 \| A^T Ax \|_\infty + \delta s^{-1/2} \|x^1\|_2 \|x\|_1 \leq \|x\|_{s,2} = \|x^1\|_2 \leq \frac{s^{1/2}}{\sqrt{1-\delta}} \| A^T Ax \|_\infty + \frac{\delta}{1-\delta} s^{-1/2} \|x\|_1
\]

and we see that the pair \( (H = \frac{t^{-1/2}}{\sqrt{1/\delta}} A, \| \cdot \| \) \) satisfies the condition \( Q_2 \left( s, \frac{\delta}{1-\delta} \right) \), as required in Proposition 1.8.ii. 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 by (!!)
\[
\|x\|_{t,q} \leq \|x\|_{s,q} \leq \|x\|_{s,2} \leq \frac{1}{1-\delta} s^{1/2} \| A^T Ax \|_\infty + \kappa t^{1/q-1} \|x\|_1,
\]
or, equivalently,
\[
1 \leq t \leq \min \left[ \left[ \frac{\kappa(1-\delta)}{\delta} \right]^{\frac{q}{2-\delta}}, s \frac{s^{-2}}{2} \right] s \frac{s^{-2}}{2}
\]
\[
\Rightarrow (H = \frac{t^{-1/2}}{\sqrt{1/\delta}} A, \| \cdot \| \) \text{ satisfies } Q_q(t, \kappa),
\]
as required in Proposition 1.12.ii. \( \square \)

### 1.5.5 Proof of Proposition 1.10

(i): Let \( \tilde{H} \in \mathbb{R}^{m \times N} \) and \( \| \cdot \| \) satisfy \( Q_\infty(s, \kappa) \). Then for every \( k \leq n \) we have
\[
|x_k| \leq \| \tilde{H}^T Ax \| + s^{-1} \kappa \|x\|_1,
\]
or, which is the same by homogeneity,
\[
\min_x \{ \| \tilde{H}^T Ax - x_k : \|x\|_1 \leq 1 \} \geq -s^{-1} \kappa.
\]
In other words, the optimal value $\text{Opt}_k$ of the conic optimization problem\textsuperscript{13}

$$\text{Opt}_k = \min_{x,i} \left\{ t - [e^k]^T x : \|H^T Ax\| \leq t, \|x\|_1 \leq 1 \right\},$$

where $e^k \in \mathbb{R}^n$ 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

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

where $\| \cdot \|_*$ is the norm conjugate to $\| \cdot \|_1$:

$$\|u\|_* = \max_{\|h\|_1 \leq 1} h^T u$$

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

\[(a) : e^k = A^T h^k + g^k, \quad (b) : h^k := \bar{H} \eta^k, \|\eta^k\|_* \leq 1, \quad (c) : \|g^k\|_\infty \leq s^{-1} \kappa.\]

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

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

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

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

whence, by Proposition 1.9, $(H, \| \cdot \|_\infty)$ satisfies $Q_{\infty}(s, \kappa)$. Moreover, for every $\eta \in \mathbb{R}^m$ 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

$$\|I_m - H^T A\|_{i,j} = \|[e^i]^T - h^T_i A\|_{i,j} \leq \|[e^i]^T - h^T_i A\|_\infty = \|e^i - A^T h_i\|_\infty \leq \text{Opt}_i,$$

implying by Proposition 1.9 that $(H, \| \cdot \|_\infty)$ does satisfy the condition $Q_{\infty}(s, \kappa_*)$ with $\kappa_* = s \max_i \text{Opt}_i$. Now assume that there exists a matrix $H'$ which, taken along with some norm $\| \cdot \|$, satisfies the condition $Q_{\infty}(s, \kappa)$ with $\kappa < \kappa_*$, and let us lead this assumption to a contradiction. By the already proved first part of Proposition 1.10, our assumption implies that there exists $m \times n$ matrix $\bar{H} = [\bar{h}_1, ..., \bar{h}_n]$ such that $\|\text{Col}_j[I_m - \bar{H}^T A]\|_\infty \leq s^{-1} \kappa$ for all $j \leq n$, implying that $\|[e^i]^T - \bar{h}_i^T A\| \leq s^{-1} \kappa$ 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 $\text{Opt}_i$, we have $\text{Opt}_i \leq \|e^i - A^T \bar{h}_i\|_\infty$ for all $i$.

\textsuperscript{13}For summary on conic programming, see Section 4.1.
and we arrive at $s^{-1}\kappa_s = \max_i \text{Opt}_i \leq s^{-1}\kappa$, that is, $\kappa_s \leq \kappa$, which is a desired contradiction.

It remains to prove (1.33), which is just an exercise on LP duality: denoting by $e$ an $n$-dimensional all-ones vector, we have

$$\text{Opt}_i := \min_h \|e^i - A^T h\|\infty = \min_{h, t} \{ t : e^i - A^T h \leq te, A^T h - e^i \leq te \}$$

$$= \max_{\lambda, \mu} \{ \lambda_i - \mu_i : \lambda, \mu \geq 0, A[\lambda - \mu] = 0, \sum_i \lambda_i + \sum_i \mu_i = 1 \} \quad \text{[LP duality]}$$

$$= \max_{x, \lambda, \mu} \{ x_i : Ax = 0, \|x\|_1 \leq 1 \}$$

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 $\|x\|_1 \leq 1$. □

1.5.6 Proof of Proposition 1.13

Let $H$ satisfy (1.38). Since $\|v\|_{s,1} \leq s^{1-1/q}\|v\|_{s,q}$, it follows that $H$ satisfies for some $\alpha < 1/2$ the condition

$$\|\text{Col}_j[I_n - H^T A]\|_{s,1} \leq \alpha, \quad 1 \leq j \leq n. \quad (1.61)$$

whence, as we know from Proposition 1.9,

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

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 $H$ and $\bar{A}$ be the $m \times \bar{n}$ matrices comprised of the first $2m$ columns of $H$, respectively, $A$. Relation (1.61) implies that the matrix $V = I_n - H^T \bar{A}$ satisfies

$$\|\text{Col}_j[V]\|_{s,1} \leq \alpha < 1/2, \quad 1 \leq j \leq \bar{n}. \quad (1.62)$$

Now, since the rank of $H^T \bar{A}$ is $\leq m$, at least $\bar{n} - m$ singular values of $V$ are $\geq 1$, 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

$$\|f\|_2^2 \leq \max \left[ \frac{\bar{n}}{s^2}, 1 \right] \|f\|_{s,1}^2. \quad (1.63)$$

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

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

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,

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

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

\[
\|V\|_F^2 = \sum_{j=1}^{2m} \|\text{Col}_j[V]\|^2 \leq \max\left[1, \frac{2m}{s^2}\right] \sum_{j=1}^{2m} \|\text{Col}_j[V]\|^2_{s,1} \leq 2\alpha^2 m \max\left[1, \frac{2m}{s^2}\right].
\]

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

\[
m \leq 2\alpha^2 m \max\left[1, \frac{2m}{s^2}\right].
\]

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

It remains to prove that when \( m \leq n/2 \), the condition \( Q_\infty(s, \kappa) \) with \( \kappa < 1/2 \) can be satisfied only when \( s \leq \sqrt{m} \). This is immediate: by Proposition 1.10, assuming \( Q_\infty(s, \kappa) \) satisfiable, there exists \( m \times n \) contrast matrix \( H \) such that \( |[I_n - H^T A]_{ij}| \leq \kappa/s \) for all \( i, j \), which, by the already proved part of Proposition 1.13, is impossible when \( s > \sqrt{m} \). \( \Box \)
Chapter Two

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 $\sigma$-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 $\sigma$-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 $\eta$ be a deterministic transformation of random variable $\xi$” should be read as “let $\eta = f(\xi)$ for some Borel function $f$”, and the sentence “Consider a test $\mathcal{T}$ deciding on hypotheses $H_1, \ldots, H_L$ via observation $\omega \in \Omega$” should be read as “Consider a Borel function $\mathcal{T}$ on Polish space $\Omega$, the values of the function being subsets of the set $\{1, \ldots, L\}$.” 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 fundamental problems of Statistics. Informally, this is the problem where one is given an observation – a realization of random variable with unknown (at least partly) 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** $\Omega$, where the observed random variable (r.v.) takes its values;
- **$L$ families** $\mathcal{P}_\ell$ of probability distributions on $\Omega$. We associate with these families $L$ hypotheses $H_1, ..., H_L$, with $H_\ell$ 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 $H_\ell$.

Hypothesis $H_\ell$ is called **simple**, if $\mathcal{P}_\ell$ is a singleton, and is called **composite** otherwise.

Our goal is, given an observation – a realization $\omega$ of the r.v. in question – to decide which 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 $T(\omega)$ of $\omega \in \Omega$; the value $T(\omega)$ of this function at a point $\omega$ is some subset of the set $\{1, ..., L\}$:

$$T(\omega) \subseteq \{1, ..., L\}.$$  

Given observation $\omega$, the test accepts all hypotheses $H_\ell$ with $\ell \in T(\omega)$ and rejects all hypotheses $H_\ell$ with $\ell \notin T(\omega)$. We call a test **simple**, if $T(\omega)$ is a singleton for every $\omega$, 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 $\omega$ and of a realization $\theta$ of independent of $\omega$ 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 $\omega$ and the outcome $\theta$ 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 $\omega$ to new observation $\omega_+ = (\omega, \theta)$, where $\theta \sim \text{Uniform}[0, 1]$ is independent of $\omega$; the $\omega$-component of our new observation $\omega_+$ is, as before, generated by “the nature,” and the $\theta$-component is generated by ourselves. Now, given families $\mathcal{P}_\ell$, $1 \leq \ell \leq L$, of probability distributions on the original observation space $\Omega$, 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 $\omega$ is the same as to decide on the hypotheses associated with the families $\mathcal{P}_{\ell,+}$ of our new observation $\omega_+$, 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 the
observation space $\Omega$ 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 $\omega^K$ “is positioned” w.r.t. the families $\mathcal{P}_\ell, 1 \leq \ell \leq L$.

We will be 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 $\omega^K$, 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

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

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

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

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 $\Omega$, and then draws, *independently across* $k$, observations $\omega_k, k = 1, ..., K$, from these distributions:

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

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 $\Omega^K$. 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 $\mathcal{P}_\ell$. 

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 that of the direct product

$$\mathcal{P}^\oplus_K = \bigoplus_{k=1}^K \mathcal{P}_{t,k}$$

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

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

2.1.3.3 Quasi-stationary $K$-repeated observations

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

“In the nature” there exists random sequence $\zeta^K = (\zeta_1, \ldots, \zeta_K)$ of “driving factors” (or states) such that for every $k$, $\omega_k$ is a deterministic function of $\zeta_1, \ldots, \zeta_k$:

$$\omega_k = \theta_k(\zeta_1, \ldots, \zeta_k)$$

and the conditional, $\zeta_1, \ldots, \zeta_{k-1}$ given, distribution $P_{\omega_k|\zeta_1,\ldots,\zeta_{k-1}}$ of $\omega_k$ always (i.e., for all $\zeta_1, \ldots, \zeta_{k-1}$) belongs to $\mathcal{P}$.

With the above mechanism, the collection $\omega^K = (\omega_1, \ldots, \omega_K)$ has some distribution $P^K$ which depends on the distribution of driving factors and functions $\theta_k(.)$. 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 $\omega^K$ with distribution $P^K$ the just defined 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 $\mathcal{P}_\ell$, $\ell = 1, \ldots, L$, on $\Omega$ and an observation $\omega^K \in \Omega^K$, decide on the hypotheses

$$H^\otimes\ell, K = \{P^K \in \mathcal{P}^\otimes_K, 1 \leq \ell \leq K\}$$

on the distribution $P^K$ of the observation $\omega^K$.

A related notion is that of quasi-direct product

$$\mathcal{P}^\otimes_K = \bigotimes_{k=1}^K \mathcal{P}_{t,k}$$

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

The description of quasi-stationary $K$-repeated observations seems to be too complicated. However, this is exactly what happens in some important applications, e.g., in hidden Markov chains. Suppose that $\Omega = \{1, \ldots, d\}$ is a finite set, and that $\omega_k \in \Omega$, $k = 1, 2, \ldots$, are generated as follows: “in the nature there” exists a Markov chain with $D$-element state space $S$ split into $d$ non-overlapping bins, and $\omega_k$ is the
index $\beta(\eta)$ of the bin to which the state $\eta_k$ of the chain belongs. Every column $Q^j$ of the transition matrix $Q$ of the chain (this column is a probability distribution on $\{1, \ldots, D\}$) generates a probability distribution $P_j$ on $\Omega$, specifically, the distribution of $\beta(\eta)$, $\eta \sim Q^j$. Now, a family $\mathcal{P}$ of distributions on $\Omega$ induces a family $Q[\mathcal{P}]$ of all $D \times D$ stochastic matrices $Q$ for which all $D$ distributions $P^j$, $j = 1, \ldots, D$, belong to $\mathcal{P}$. When $Q \in Q[\mathcal{P}]$, observations $\omega_k$, $k = 1, 2, \ldots$ clearly are given by the above “quasi-stationary mechanism” with $\eta_k$ 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 $\mathcal{P}_\ell$, $\ell = 1, \ldots, L$ of probability distributions on $S$, deciding on hypotheses $Q \in Q[\mathcal{P}_\ell]$, $\ell = 1, \ldots, L$, 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 $\mathcal{P}_\ell$, $\ell = 1, \ldots, L$, be families of probability distributions on observation space $\Omega$; these families give rise to hypotheses

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

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

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

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

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

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

**Total risk** $\text{Risk}_{\text{tot}}(T|H_1, \ldots, H_L)$ is the sum of all $L$ partial risks:

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

**Risk** $\text{Risk}(T|H_1, \ldots, H_L)$ is the maximum of all $L$ partial risks:

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

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 (\( \omega^K \) in the role of \( \omega \) and, say, \( \mathcal{P}_{\ell}^{\oplus,K} = \bigoplus_{k=1}^{K} \mathcal{P}_{\ell} \) 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 fashion: as the maximal, 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 any of the \( L \) hypotheses.

Remark 2.1. “The smaller are the hypotheses, the less are the risks.” Specifically, given families of probability distributions \( \mathcal{P}_{\ell} \subset \mathcal{P}_{\ell}', \ell = 1, \ldots, L \), on observation space \( \Omega \), 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 \( T \) deciding on the “larger” hypotheses \( H_{1}', \ldots, H_{L}' \) can be considered as a test deciding on smaller hypotheses \( H_{1}, \ldots, H_{L} \) as well, and the risks of the test when passing from larger hypotheses to smaller ones can only drop down:

\[
\mathcal{P}_{\ell} \subset \mathcal{P}_{\ell}', 1 \leq \ell \leq L \Rightarrow \text{Risk}(T|H_{1}, \ldots, H_{L}) \leq \text{Risk}(T|H_{1}', \ldots, H_{L}').
\]

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

\[
H_{\ell}^{\oplus,K} : P^K \in \mathcal{P}_{\ell}^{\oplus,K}, \quad H_{\ell}^{\oplus,K} : P^K \in \mathcal{P}_{\ell}^{\oplus,K} = \bigoplus_{k=1}^{K} \mathcal{P}_{\ell},
\]

\[
H_{\ell}^{\otimes,K} : P^K \in \mathcal{P}_{\ell}^{\otimes,K} = \bigotimes_{k=1}^{K} \mathcal{P}_{\ell}, 1 \leq \ell \leq L,
\]

(see Section 2.1.3), and clearly

\[
\mathcal{P}_{\ell}^{\oplus,K} \subset \mathcal{P}_{\ell}^{\oplus,K} \subset \mathcal{P}_{\ell}^{\oplus,K}.
\]

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 well known [158, 160] observation is nearly evident:

**Proposition 2.2.** 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 \( \Pi \) on \( \Omega \). \(^1\) Then for any simple test \( T \) deciding

\(^1\) This assumption is w.l.o.g. – we can take, as \( \Pi \), the sum of the measures \( P_{1} \) and \( P_{2} \).
HYPOTHESIS TESTING

on $H_1, H_2$ it holds

$$\text{Risk}_{\text{tot}}(T|H_1, H_2) \geq \int_{\Omega} \min[p_1(\omega), p_2(\omega)]\Pi(d\omega).$$  \hfill (2.1)

Note that the right hand side in this relation is independent of how $\Pi$ is selected.

Proof. Consider a simple test $T$, 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 $\omega$. Since the test is simple, the probability for $T$ to accept $H_2$ and to reject $H_1$, observation being $\omega$, is $1 - \pi(\omega)$. Consequently,

$$\text{Risk}_1(T|H_1, H_2) = \int_{\Omega} (1 - \pi(\omega))p_1(\omega)\Pi(d\omega),$$

$$\text{Risk}_2(T|H_1, H_2) = \int_{\Omega} \pi(\omega)p_2(\omega)\Pi(d\omega),$$

whence

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

Remark 2.3. Note that the lower risk bound (2.1) is achievable; given an observation $\omega$, the corresponding test $T$ 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 a tie (i.e., $\pi(\omega) = 1/2$ when $p_1(\omega) = p_2(\omega)$). This is nothing but the likelihood ratio test naturally adjusted to account for ties.

Example 2.1. Let $\Omega = \mathbb{R}^d$, let the reference measure $\Pi$ be the Lebesgue measure on $\mathbb{R}^d$, and let $p_\chi(\cdot) = N(\mu_\chi, I_d)$, be the Gaussian densities on $\mathbb{R}^d$ with unit covariance and means $\mu_\chi$, $\chi = 1, 2$. In this case, assuming $\mu_1 \neq \mu_2$, the recipe from Remark 2.3 reduces to the following:

Let

$$\phi_{1,2}(\omega) = \frac{1}{2}[\mu_1 - \mu_2]^T[\omega - w], \quad w = \frac{1}{2}[\mu_1 + \mu_2].$$  \hfill (2.2)

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

$$\text{Risk}_1(T|H_1, H_2) = \text{Risk}_2(T|H_1, H_2) = \text{Risk}(T|H_1, H_2) = \frac{1}{2}\text{Risk}_{\text{tot}}(T|H_1, H_2) = \text{Erfc}(\frac{1}{2}\|\mu_1 - \mu_2\|_2)$$ \hfill (2.3)

(see (1.22) for the definition of Erfc), and the test is optimal in terms of its risk and its total risk.

Note that optimality of $T$ in terms of total risk is given by Proposition 2.2 and Remark 2.3; optimality in terms of risk is ensured by optimality in terms of total risk combined with the first equality in (2.3).

Example 2.1 admits an immediate and useful extension [36, 37, 127]:

Example 2.2. Let $\Omega = \mathbb{R}^d$, let the reference measure $\Pi$ be the Lebesgue measure on $\mathbb{R}^d$, and let $M_1$ and $M_2$ be two nonempty closed convex sets in $\mathbb{R}^d$ with empty
Figure 2.1: “Gaussian Separation” (Example 2.2): Optimal test deciding on whether the mean of Gaussian r.v. belongs to the domain $A$ ($H_1$) or to the domain $B$ ($H_2$). Hyperplane o-o separates the acceptance domains for $H_1$ (“left” half-space) and for $H_2$ (“right” half-space).

Intersection and such that the convex optimization program

$$
\min_{\mu_1, \mu_2} \{ \|\mu_1 - \mu_2\|_2 : \mu_\chi \in M_\chi, \chi = 1, 2 \} \quad (\star)
$$

has an optimal solution $\mu^*_1, \mu^*_2$ (this definitely is the case when at least one of the sets $M_1, M_2$ is bounded). Let

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

and let the simple test $T$ 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 $\omega$, $T$ accepts $H_1$ (and rejects $H_2$) when $\phi_{1,2}(\omega) \geq 0$, and accepts $H_2$ (and rejects $H_1$) otherwise. Then

$$
\text{Risk}(T|H_1, H_2) = \text{Risk}_2(T|H_1, H_2) = \text{Risk}(T|H_1, H_2) = \frac{1}{2}\text{Risk}_{\text{tot}}(T|H_1, H_2) = \text{Erfc}(\frac{1}{2}\|\mu_1^* - \mu_2^*\|_2),
$$

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 in the direction of $\mu_1^* - \mu_2^*$, and let $\xi[\omega] = e^T(\omega - w)$. From optimality conditions for $(\star)$ 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 $\omega$ is $p_{\mu} = \mathcal{N}(\mu, I_d)$, the random variable $\xi[\omega]$ is a scalar Gaussian random variable with unit variance and expectation $\geq \delta := \frac{1}{2}\|\mu_1^* - \mu_2^*\|_2$, implying that $p_{\mu}$-probability for $\xi[\omega]$ to be negative (which is exactly the same as the $p_{\mu}$-probability for $T$ to reject $H_1$ and accept $H_2$) is at most Erfc$(\delta)$. Similarly, when $\mu \in M_2$ and the density of $\omega$ is $p_{\mu} = \mathcal{N}(\mu, I_d)$, $\xi[\omega]$ is a scalar Gaussian random variable with unit variance and expectation $\leq -\delta$, implying that $p_{\mu}$-probability for $\xi[\omega]$ to be positive (which is exactly the same as the $p_{\mu}$-probability for $T$ to reject $H_2$ and accept $H_1$) is at most Erfc($\delta$).
implying that the $p_\mu$-probability for $\xi[\omega]$ to be nonnegative (which is exactly the same as the probability for $T$ to reject $H_2$ and accept $H_1$) is at most $\text{Erfc}(\delta)$. These observations imply the validity of (2.5). The test 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 $\omega$ can be only larger than the risks of a simple test deciding on two simple hypotheses $p = p_\mu^1$ and $p = p_\mu^2$. In other words, the quantity $\text{Erfc}(1/2\|\mu_1^* - \mu_2^*\|_2)$, 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 $T$ in terms of risks are immediate consequences of (2.5).

We remark that the (nearly self-evident) result stated in Example 2.2 seems first been noticed in [36]. 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 the recent paper [108]. A less straightforward and, we believe, more instructive extensions, originating from [100], 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

$$H_\ell : P \in \mathcal{P}_\ell, \ell = 1, \ldots, L \tag{2.6}$$

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

- A family $\mathcal{P}$ of probability distributions on $\Omega = \mathbb{R}^d$ such that all distributions from $\mathcal{P}$ possess densities with respect to the Lebesgue measure on $\mathbb{R}^n$, and these densities are even functions on $\mathbb{R}^d$;
- A collection $X_1, \ldots, X_L$ of nonempty closed and convex subsets of $\mathbb{R}^d$, with at most one of the sets unbounded.

These data specify $L$ families $\mathcal{P}_\ell$ of distributions on $\mathbb{R}^d$: $\mathcal{P}_\ell$ is comprised of distributions of random vectors of the form $x + \xi$, where $x \in X_\ell$ is deterministic, and $\xi$ is random with distribution from $\mathcal{P}$. Note that with this setup, deciding upon hypotheses (2.6) via observation $\omega \sim P$ is exactly the same as to decide, given observation

$$\omega = x + \xi, \tag{2.7}$$

\footnote{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 an abuse of notation), where $p(\cdot)$ is the density of the probability distribution $P$, to express the fact that $P \in \mathcal{P}$.}
where $x$ is a deterministic “signal” and $\xi$ is random noise with distribution $P$ known to belong to $P$, on the “position” of $x$ w.r.t. $X_1, \ldots, X_L$, the $\ell$-th hypothesis $H_\ell$ saying that $x \in X_\ell$. The latter allows us to write down the $\ell$-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 where $L = 2$, $X_1 = \{x^1\}$ and $X_2 = \{x^2\}$, $x^1 \neq x^2$, are singletons, and $P$ also is a singleton. Let the probability density of the only distribution from $P$ be of the form

$$p(u) = f(\|u\|_2), \quad f(\cdot) \text{ is a strictly monotonically decreasing function on the nonnegative ray.} \quad (2.8)$$

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

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

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 (2.7). Let us set

$$\delta = \frac{1}{2}\|x^1 - x^2\|_2, \quad e = \frac{x^1 - x^2}{\|x^1 - x^2\|_2}, \quad \phi(\omega) = e^T \omega - \frac{1}{2} e^T [x^1 + x^2], \quad (2.9)$$

and consider the test $T$ 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)

$$\text{Risk}_1(T|H_1, H_2) = \int_{\omega: \phi(\omega) < 0} p_1(\omega) d\omega = \int_{u: e^T u < -\delta} f(\|u\|_2) du = \int_{u: e^T u \geq \delta} f(\|u\|_2) du = \int_{\omega: \phi(\omega) \geq 0} p_2(\omega) d\omega = \text{Risk}_2(T|H_1, H_2).$$
Since \( p(u) \) is strictly decreasing function of \( \|u\|_2 \), we have also
\[
\min[p_1(u), p_2(u)] = \begin{cases} 
  p_1(u), & \phi(u) < 0 \\
  p_2(u), & \phi(u) \geq 0 
\end{cases}
\]
whence
\[
\begin{align*}
\text{Risk}_1(T|H_1, H_2) + \text{Risk}_2(T|H_1, H_2) &= \int_{\omega: \phi(\omega) < 0} p_1(\omega) d\omega + \int_{\omega: \phi(\omega) \geq 0} p_2(\omega) d\omega \\
&= \int_{u} \min[p_1(u), p_2(u)] du
\end{align*}
\]

Invoking Proposition 2.2, we conclude that the test \( T \) is the minimum risk simple test deciding on \( H_1, H_2 \), and the risk of this test is
\[
\text{Risk}(T|H_1, H_2) = \int_{u: e^T u \geq \delta} f(\|u\|_2) du.
\] (2.10)

### 2.2.2.2 Extension

Now consider a slightly more complicated case of the situation from Section 2.2.1 with \( L = 2 \) so that \( X_1 \) and \( X_2 \) are nonempty and nonintersecting closed convex sets, 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 (2.8). Our now situation is an extension of that in Example 2.2. For exactly the same reasons as in the latter Example, with \( X_1, X_2 \) as above, the convex minimization problem
\[
\text{Opt} = \min_{x^1 \in X_1, x^2 \in X_2} \frac{1}{2}\|x^1 - x^2\|_2
\] (2.11)
is solvable, and denoting by \((x^1_*, x^2_*)\) an optimal solution and setting
\[
\phi(\omega) = e^T \omega - c, \quad e = \frac{x^1_* - x^2_*}{\|x^1_* - x^2_*\|_2}, \quad c = \frac{1}{2} e^T [x^1_* + x^2_*]
\] (2.12)
the stripe \( \{\omega: -\text{Opt} \leq \phi(\omega) \leq \text{Opt}\} \) separates \( X_1 \) and \( X_2 \):
\[
\phi(x^1) \geq \phi(x^1_*) = \text{Opt} \ \forall x^1 \in X_1 \ \& \ \phi(x^2) \leq \phi(x^2_*) = -\text{Opt} \ \forall x^2 \in X_2
\] (2.13)

**Proposition 2.4.** Let \( X_1, X_2 \) be nonempty and nonintersecting closed convex sets in \( \mathbb{R}^d \), one of the sets being bounded. With \( \text{Opt} \) and \( \phi(\cdot) \) given by (2.11) and (2.12), let us split the width \( 2\text{Opt} \) of the stripe \( \{\omega: -\text{Opt} \leq \phi(\omega) \leq \text{Opt}\} \) separating \( X_1 \) and \( X_2 \) into two nonnegative parts:
\[
\delta_1 \geq 0, \delta_2 \geq 0, \delta_1 + \delta_2 = 2\text{Opt}
\] (2.14)
and consider the simple test \( T \) which decides on the hypotheses \( H_1: x \in X_1 \) and
\( H_2 : x \in X_2 \) via observation (2.7) accepting \( H_1 \) when
\[
\phi(\omega) \geq \frac{1}{2} [\delta_2 - \delta_1]
\]
and accepting \( H_2 \) otherwise. Then
\[
\text{Risk}_\chi(T|H_1, H_2) \leq \int_{\delta_\chi}^\infty \gamma(s)ds, \; \chi = 1, 2,
\]
where \( \gamma(\cdot) \) is the univariate marginal density of \( \xi \), that is, probability density of the scalar random variable \( h^T \xi \), where \( ||h||_2 = 1 \) (note that due to (2.8), \( \gamma(\cdot) \) is independent of how we select \( h \) with \( ||h||_2 = 1 \)).

In addition, when \( \delta_1 = \delta_2 = \text{Opt} \), \( T \) is the minimum risk test deciding on \( H_1, H_2 \). The risk of this test is
\[
\text{Risk}(T|H_1, H_2) = \int_{\text{Opt}}^\infty \gamma(s)ds.
\]

**Proof.** By (2.8) and (2.13), for \( x \in X_1 \) we have (see Figure 2.2):
\[
\text{Prob}_{\xi \sim p(\cdot)} \{ \phi(x + \xi) < \frac{1}{2} [\delta_2 - \delta_1] \} \leq \text{Prob}_{\xi \sim p(\cdot)} \{ [-e]^T \xi \geq \delta_1 \} = \int_{\delta_1}^\infty \gamma(s)ds.
\]
By the “symmetric” reasoning, for \( x \in X_2 \) we have
\[
\text{Prob}_{\xi \sim p(\cdot)} \{ \phi(x + \xi) \geq \frac{1}{2} [\delta_2 - \delta_1] \} \leq \text{Prob}_{\xi \sim p(\cdot)} \{ e^T \xi \geq \delta_2 \} = \int_{\delta_2}^\infty \gamma(s)ds,
\]
and we arrive at (2.15). The fact that in the case of \( \delta_1 = \delta_2 = \text{Opt} \) our test \( T \) becomes the minimum risk test deciding on composite hypotheses \( H_1, H_2 \) is readily given by the analysis in Section 2.2.2.1: the minimal, over all possible tests, risk of
deciding on two simple hypotheses $H'_1 : x = x_1^1$, $H'_2 : x = x_2^2$ is given by (2.10), i.e., it is equal to $\int_{x_0}^{\infty} \gamma(s)ds$. In the case of $\delta_1 = \delta_2 = \text{Opt}$ this is exactly the upper bound (2.16) on the risk of the test $T$ deciding on the larger than $H'_x$ composite hypotheses $H_x, \chi = 1, 2$.

\[ \square \]

2.2.2.3 Further extensions: spherical families of distributions

Same as in Section 2.2.2.2, we continue to 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. Our next goal 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 (2.15) in Proposition 2.4 used the only fact that whenever $h, u \in \mathbb{R}^d$ is a $\| \cdot \|_2$-unit vector and $\delta \geq 0$, we have $\int_{h^T u \geq \delta} p(u)du \leq \int_{h^T}^{\infty} \gamma(s)ds$, 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.

2.2.2.3.A. Spherical families 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 $\gamma$ a spherical family of densities $\mathcal{P}_\gamma$ comprised of all probability densities $p(\cdot)$ on $\mathbb{R}^d$ such that

A. $p(\cdot)$ is even
B. Whenever $e \in \mathbb{R}^d$, $\|e\|_2 = 1$, and $\delta \geq 0$, we have

\[
\text{Prob}_{\xi \sim p}\{\xi : e^T \xi \geq \delta\} \leq P_\gamma(\delta) := \int_{\delta}^{\infty} \gamma(s)ds. \tag{2.17}
\]

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

Note that density (2.8) belongs to the family $\mathcal{P}_\gamma$ with $\gamma(\cdot)$ defined in Proposition 2.4; the resulting $\gamma$, in addition to being an even density, is strictly monotonically decreasing on the nonnegative ray. When speaking about general-type spherical families $\mathcal{P}_\gamma$, we do not impose monotonicity requirements on $\gamma(\cdot)$. If a spherical family $\mathcal{P}_\gamma$ includes a density $p(\cdot)$ of the form (2.8) such that $\gamma(\cdot)$ is the induced by $p(\cdot)$ univariate marginal density, as in Proposition 2.4, we say that $\mathcal{P}_\gamma$ has a cap, and this cap is $p(\cdot)$.

2.2.2.3.B. Example: Gaussian mixtures. Let $\eta \sim \mathcal{N}(0, \Theta)$, where the $d \times d$ covariance matrix $\Theta$ satisfies $\Theta \preceq I_d$, and let $Z$ be an independent of $\eta$ positive scalar random variable. Gaussian mixture of $Z$ and $\eta$ (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 [88, 149] include

- Gaussian distribution $\mathcal{N}(0, \Theta)$ (take $Z$ identically equal to 1),
- multidimensional Student’s $t$-distribution with $\nu \in \{1, 2, \ldots\}$ degrees of freedom and “covariance structure” $\Theta$; here $Z$ is given by the requirement that $\nu/Z$ has $\chi^2$-distribution with $\nu$ degrees of freedom.
An immediate observation (see Exercise 2.2) is that with \( \gamma \) given by the distribution \( P_Z \) of \( Z \) according to

\[
\gamma_Z(s) = \int_{z>0} \frac{1}{\sqrt{2\pi z}} e^{-\frac{s^2}{2z}} P_Z(dz),
\]

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

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

\[
\gamma_G(s) = \frac{1}{\sqrt{2\pi}} e^{-\frac{s^2}{2}};
\]

This family has a cap, specifically, the standard Gaussian \( d \)-dimensional distribution \( N(0, I_d) \).

**2.2.2.3.C. Main result.** Observing the proof of Proposition 2.4, we arrive at the following

**Proposition 2.5.** Let \( X_1 \) and \( X_2 \) be nonempty and nonintersecting closed convex sets in \( \mathbb{R}^d \), 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 (2.11) – (2.12), let us split the width 2Opt of the stripe \( \{ \omega : -\text{Opt} \leq \phi(\omega) \leq \text{Opt} \} \) separating \( X_1 \) and \( X_2 \) into two nonnegative parts:

\[
\delta_1 \geq 0, \delta_2 \geq 0, \delta_1 + \delta_2 = 2\text{Opt}.
\]

Let us consider a simple test \( T \) deciding on the hypotheses \( H_1 : x \in X_1, H_2 : x \in X_2 \) via observation (2.7) accepting \( H_1 \) when

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

and accepting \( H_2 \) otherwise. Then

\[
\text{Risk}_\chi(T|H_1, H_2) \leq \int_{s_\chi}^\infty \gamma(s) ds, \quad \chi = 1, 2.
\]

In addition, when \( \delta_1 = \delta_2 = \text{Opt} \) and \( \mathcal{P}^d_{\gamma} \) has a cap, \( T \) is the minimum risk test deciding on \( H_1, H_2 \). The risk of this test is given by

\[
\text{Risk}(T|H_1, H_2) = P'_\gamma(\text{Opt}) := \int_{0}^{\infty} \gamma(s) ds.
\]

To illustrate the power of Proposition 2.5, consider the case when \( \gamma \) is the function (2.18) stemming from Student’s \( t \)-distribution on \( \mathbb{R}^d \) with \( \nu \) degrees of freedom. It is known that in this case \( \gamma \) is the density of univariate Student’s
$t$-distribution with $\nu$ degrees of freedom [149]:

$$
\gamma(s) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu}}\left(1+s^2/\nu\right)^{-\frac{\nu+1}{2}},
$$

where $\Gamma(\cdot)$ is Euler’s Gamma function. When $\nu = 1$, $\gamma(\cdot)$ 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 $\nu$) and as such are of interest in statistical application in Finance.

### 2.2.3 Euclidean Separation, Repeated Observations, and Majority Tests

Assume that $X_1, X_2$ and $\mathcal{P}_d^\gamma$ are as in the premise of Proposition 2.5 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 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 $\mathcal{P}_d^\gamma$ with the family of product distributions $P \times \cdots \times P$ stemming from $P \in \mathcal{P}_d^\gamma$, and it is unclear how to apply to the resulting single-observation testing problem our machinery based on Euclidean separation. Instead, we will use the $K$-step majority test.

#### 2.2.3.1 Preliminaries: Repeated observations in “signal plus noise” observation model

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

$$
\omega^K = (\omega_1, \omega_2, \ldots, \omega_K),
$$

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

- **S** [stationary $K$-repeated observations, cf. Section 2.1.3.1]: $\omega_1, \ldots, \omega_K$ are drawn independently of each other from the same distribution $Q$, that is, $Q^K$ is the product distribution $Q \times \cdots \times Q$. 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 $\xi$ belongs to the family $\mathcal{P}_d^\gamma$.

- **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 $\{P_k \in \mathcal{P}_d^\gamma\}_{k=1}^K$, 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 $X_\chi$.

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

2.2.3.2.A. The construction of $K$-observation majority test is very natural. We use Euclidean separation to build simple single-observation test $T$ our actions is as follows: construction from Proposition 2.5 applied with $\delta$ deterministic, and the distribution of noise $\xi$ belongs to $\mathcal{P}_\gamma^d$. $H$ on hypotheses $X$.

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

Proposition 2.6. With $X_1, X_2, \mathcal{P}_\gamma^d$ obeying the premise of Proposition 2.5, in the case QS of quasi-stationary observations the risk of $K$-observation Majority test $T_K^{maj}$ can be bounded as

$$
\text{Risk}(T_K^{maj} | H_1, H_2) \leq \epsilon_K \equiv \sum_{K/2 \leq k \leq K} \binom{K}{k} \epsilon_k^k (1 - \epsilon_*)^{K-k}, \quad \epsilon_* = \int_{\text{Opt}} \gamma(s) ds. \quad (2.23)
$$

Proof. For the sake of clarity, 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, the proposition is proved in Section 2.11.2.

Assume that $H_1$ 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 $P_k \in \mathcal{P}_\gamma^d$. Let us fix $\{x_k \in X_1\}_{k=1}^K$ and $\{P_k \in \mathcal{P}_\gamma^d\}_{k=1}^K$. Then the random reals $v_k = \phi(\omega_k = x_k + \xi_k)$ are independent across $k$, and so are the Boolean random variables

$$
\chi_k = \begin{cases} 
1, & v_i < 0, \\
0, & v_i \geq 0;
\end{cases}
$$

$\chi_k = 1$ if and only if test $T$, as applied to observation $\omega_k$, rejects hypothesis $H_1: x_k \in X_1$. By Proposition 2.5, $P_k$-probability $p_k$ of the event $\chi_k = 1$ is at most $\epsilon_*$. Further, by construction of the Majority test, if $T_K^{maj}$ rejects the true
HYPOTHESIS TESTING

hypothesis \( H_1 \), then the number of \( k \)'s with \( \chi_k = 1 \) is \( \geq K/2 \). Thus, with \( x_k \in X_1 \) and \( P_k \in \mathcal{P}_d^{\gamma} \), \( 1 \leq k \leq K \), the probability to reject \( H_1 \) 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 (2.23), implying that \( \text{Risk}_1(\mathcal{T}_{\text{maj}}^\text{max}|H_1, H_2) \leq \epsilon_K \). A “symmetric” reasoning yields

\[
\text{Risk}_2(\mathcal{T}_{\text{maj}}^\text{max}|H_1, H_2) \leq \epsilon_K,
\]

completing the proof of (2.23). \( \square \)

**Corollary 2.7.** Under the premise of Proposition 2.6, the upper bounds \( \epsilon_K \) 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 \( \epsilon_* < \frac{1}{4} \).

**Remark 2.8.** When proving (SS-version of) Proposition 2.6, we have used an “evident” observation as follows:

(\#) Let \( \chi_1, \ldots, \chi_K \) be independent random variables taking values 0 and 1, and let the probabilities \( p_k \) for \( \chi_k \) to take value 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 \( \chi_1, \ldots, \chi_K \) are equal to 1” is upper-bounded by the probability \( \sum_{M \leq k \leq K} (\frac{K}{k})^k (1-\epsilon)^{K-k} \) 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, it requires a proof; this proof (finally, not completely evident) can be found in Section 2.11.2.

**2.2.3.2.C. Near-optimality.** We are about to show that under appropriate assumptions, the majority test \( \mathcal{T}_{\text{maj}}^\text{max} \) is near-optimal. The precise statement is as follows:

**Proposition 2.9.** Let \( X_1, X_2, \mathcal{P}_\gamma^d \) obey the premise of Proposition 2.5. Assume that the spherical family \( \mathcal{P}_\gamma \) and positive reals \( D, \alpha, \beta \) are such that

\[
\beta D \leq \frac{1}{4}, \tag{2.24}
\]

\[
\int_0^\delta \gamma(s)ds \geq \beta \delta, \ 0 \leq \delta \leq D, \tag{2.25}
\]

and \( \mathcal{P}_\gamma \) contains a density \( q(\cdot) \) such that

\[
\int_{\mathbb{R}^n} \sqrt{q(\xi - e)q(\xi + e)}d\xi \geq \exp\{-\alpha e^T e\} \ \forall (e: ||e||_2 \leq D). \tag{2.26}
\]

\( \beta \)Recall that we have assumed from the very beginning that \( \gamma \) is an even probability density on the axis, and there is no neighbourhood of the origin where \( \gamma = 0 \) a.s.
Let also the sets $X_1$ and $X_2$ be such that $\text{Opt}$ as given by (2.11) satisfies the relation

$$\text{Opt} \leq D. \tag{2.27}$$

Given tolerance $\epsilon \in (0, 1/5)$, the risk of $K$-observation majority test $T_{K}^{\text{maj}}$ utilizing $QS$ observations ensures the relation

$$K \geq K^* := \left\lfloor \frac{\ln(1/\epsilon)}{2\beta^2 \text{Opt}^2} \right\rfloor \Rightarrow \text{Risk}(T_{K}^{\text{maj}}|H_1, H_2) \leq \epsilon \tag{2.28}$$

(here $\lfloor x \rfloor$ stands for the smallest integer $\geq x \in \mathbb{R}$). In addition, for every $K$-observation test $T_K$ utilizing stationary repeated observations and satisfying

$$\text{Risk}(T_K|H_1, H_2) \leq \epsilon \tag{2.29}$$

it holds

$$K \geq K_* := \frac{\ln(\frac{1}{\epsilon})}{2\alpha \text{Opt}^2}. \tag{2.29}$$

As a result, the majority test $T_{K}^{\text{maj}}$ (which by (2.28) has risk $\leq \epsilon$) is near-optimal in terms of the required number of observations among all tests with risk $\leq \epsilon$: the number $K$ of observations in such a test satisfies the relation

$$K^*/K \leq \theta := K^*/K_* = O(1) \frac{\alpha}{\beta^2}. \tag{2.30}$$

Proof of the proposition is the subject of Exercise 2.3.

**Illustration.** Given $\nu \geq 1$, consider the case when $\mathcal{P} = \mathcal{P}_\gamma$ is the spherical family with $n$-variate (spherical) Student’s distribution in the role of the cap, so that

$$\gamma(s) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right) (\pi \nu)^{1/2}} \left[1 + s^2/\nu\right]^{-(\nu+1)/2} \tag{2.30}$$

It is easily seen (cf. Exercise 2.3) that $\mathcal{P}$ contains the $\mathcal{N}(0, I_n)$ density $q(\cdot)$, implying that setting

$$D = 1, \ \alpha = 1, \ \beta = \frac{1}{4},$$

one ensures relations (2.24), (2.25) and (2.27). As a result, when $\text{Opt}$ as yielded by (2.11) is $\leq 1$, the non-optimality factor $\theta$ of the majority test $T_{K}^{\text{maj}}$ as defined in Proposition 2.9 is $O(1)$.

### 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 $\Omega$, and observe a realization of random variable $\omega \sim P$ taking values in $\Omega$. Given $\omega$, we want to decide on the $L$ hypotheses

$$H_{\ell}: P \in \mathcal{P}_\ell, \ 1 \leq \ell \leq L. \tag{2.31}$$
Our *ideal goal* would be to find a low-risk simple test deciding on the hypotheses. However, it may happen that this “ideal goal” cannot be achieved, 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 $< 1/2$.

**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 rectangles on the plane:

![3 rectangles on the plane](image)

and 3 hypotheses, with $H_\ell, \ell \in \{A, B, C\}$, stating that our observation is $\omega = x + \xi$ with deterministic “signal” $x$ belonging to rectangle $\ell$ and $\xi \sim \mathcal{N}(0, \sigma^2 I_2)$. Whatever small $\sigma$ is, no test can decide on the 3 hypotheses with risk $< 1/2$; e.g., there is no way to decide reliably on $H_A$ vs. $H_B$. However, we may hope that when $\sigma$ is small (or when repeated observations are allowed), observations allow us to discard reliably at least some of the hypotheses. For instance, when the signal belongs to rectangle $A$ (i.e., $H_A$ holds true), we hardly can discard reliably the hypothesis $H_B$ stating that the signal belongs to rectangle $B$, but hopefully can reliably discard $H_C$ (that is, infer that the signal is *not* in rectangle $C$).

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* $C$ on a collection of $L$ hypotheses $H_1, \ldots, H_L$ is defined as some set of pairs $(\ell, \ell')$ with $1 \leq \ell, \ell' \leq L$. We interpret the relation $(\ell, \ell') \in C$ as the fact that the hypotheses $H_\ell$ and $H_{\ell'}$ are close to each other. Sometimes we shall use the words “$\ell$ and $\ell'$ are/are not $C$-close to each other” as an equivalent form of “hypotheses $H_\ell, H_{\ell'}$ are/are not $C$-close to each other.”

We always assume that

- $C$ contains all “diagonal pairs” $(\ell, \ell)$, $1 \leq \ell \leq L$ (“every hypothesis is close to itself”);
- $(\ell, \ell') \in C$ is and only if $(\ell', \ell) \in C$ (“closeness is a symmetric relation”).

Note that by symmetry of $C$, the relation $(\ell, \ell') \in T$ is in fact a property of unordered pair $(\ell, \ell')$.

**“Up to closeness” risks.** Let $T$ be a test deciding on $L$ hypotheses $H_1, \ldots, H_L$, see (2.31); given observation $\omega$, $T$ accepts all hypotheses $H_\ell$ with indexes $\ell \in T(\omega)$ and rejects all other hypotheses. We say that the $\ell$-th partial $C$-risk of test $T$ is $\leq \epsilon$ if whenever $H_\ell$ is true: $\omega \sim P \in \mathcal{P}_\ell$, the $P$-probability of the event
\textbf{CHAPTER 2}

\begin{itemize}
  \item $T$ accepts $H_\ell$: $\ell \in T(\omega)$
  \item all hypotheses $H_{\ell'}$ accepted by $T$ are $C$-close to $H_\ell$: $(\ell, \ell') \in C, \forall \ell' \in T(\omega)$
\end{itemize}

is at least $1 - \epsilon$.

The $\ell$-th partial $C$-risk $\text{Risk}_C^\ell(T|H_1, \ldots, H_L)$ of $T$ is the smallest $\epsilon$ with the outlined property, or, equivalently,

$$\text{Risk}_C^\ell(T|H_1, \ldots, H_L) = \sup_{P_\omega \sim P} \text{Prob}_\omega \left\{ (\ell \not\in T(\omega)) \right\}$$

$C$-risk $\text{Risk}_C^C(T|H_1, \ldots, H_L)$ of $T$ is the largest of the partial $C$-risks of the test:

$$\text{Risk}_C^C(T|H_1, \ldots, H_L) = \max_{1 \leq \ell \leq L} \text{Risk}_C^\ell(T|H_1, \ldots, H_L).$$

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

\subsection{Multiple Hypothesis Testing via pairwise tests}

Assume that for every unordered pair $\{\ell, \ell'\}$ with $(\ell, \ell') \not\in C$ we are given a simple test $T_{(\ell, \ell')}$ deciding on $H_\ell$ vs. $H_{\ell'}$ via observation $\omega$.

Our goal is to “assemble” the tests $T_{(\ell, \ell')}$, $(\ell, \ell') \not\in C$, into a test $T$ deciding on $H_1, \ldots, H_L$ up to closeness $C$.

\textbf{The construction} we intend to use is as follows:

\begin{itemize}
  \item For $1 \leq \ell, \ell' \leq L$, we define functions $T_{\ell\ell'}(\omega)$ as follows:
    \begin{itemize}
      \item when $(\ell, \ell') \in C$, we set $T_{\ell\ell'}(\cdot) = 0$.
      \item when $(\ell, \ell') \not\in C$, so that $\ell \neq \ell'$, we set
    
    $$T_{\ell\ell'}(\omega) = \begin{cases} 1, & T_{(\ell, \ell')} (\omega) = \{\ell\} \\ -1, & T_{(\ell, \ell')} (\omega) = \{\ell'\} \end{cases}. \tag{2.32}$$
    \end{itemize}

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

    Note that by construction and since $C$ is symmetric, we have

    $$T_{\ell\ell'}(\omega) = -T_{\ell'\ell}(\omega), \ 1 \leq \ell, \ell' \leq L. \tag{2.33}$$

  \item The test $T$ is as follows: \textit{given observation $\omega$, we build the $L \times L$ matrix $T(\omega) = [T_{\ell\ell'}(\omega)]$ and accept exactly those of the hypotheses $H_\ell$ for which the $\ell$-th row in $T(\omega)$ is nonnegative.}
\end{itemize}
Observation 2.10. When $T$ accepts a hypothesis $H_\ell$, all hypotheses accepted by $T$ are $C$-close to $H_\ell$.

Indeed, if $\omega$ is such that $\ell \in T(\omega)$, then the $\ell$-th row in $T(\omega)$ is nonnegative. If now $\ell'$ is not $C$-close to $\ell$, we have $T_{\ell\ell'}(\omega) \geq 0$ and $T_{\ell\ell'}(\omega) \in \{-1, 1\}$, whence $T_{\ell\ell'}(\omega) = 1$. Consequently, by (2.33) it holds $T_{\ell\ell'}(\omega) = -1$, implying that $\ell'$-th row in $T(\omega)$ is not nonnegative, and thus $\ell' \notin T(\omega)$. \qed

Risk analysis. For $(\ell, \ell') \notin C$, let

$$
\epsilon_{\ell \ell'} = \text{Risk}_1(T_{\ell\ell'}, H_{\ell'}, H_{\ell}) = \sup_{P \in \mathcal{P}_\ell} \text{Prob}_{\omega \sim P}\{\ell \notin T_{\ell \ell'}(\omega)\} = \sup_{P \in \mathcal{P}_\ell} \text{Prob}_{\omega \sim P}\{T_{\ell\ell'}(\omega) = 1\} = \sup_{P \in \mathcal{P}_\ell} \text{Prob}_{\omega \sim P}\{\ell' \notin T_{\ell \ell'}(\omega)\} = \text{Risk}_2(T_{\ell\ell'}, H_{\ell'}, H_{\ell}).
$$

(2.34)

Proposition 2.11. For the just defined test $T$ it holds

$$
\forall \ell \leq L : \text{Risk}_1^\ell(T|H_1, \ldots, H_L) \leq \epsilon_\ell := \sum_{\ell' : (\ell, \ell') \notin C} \epsilon_{\ell \ell'}.
$$

(2.35)

Proof. Let us fix $\ell$, let $H_\ell$ be true, and let $P \in \mathcal{P}_\ell$ be the distribution of observation $\omega$. Set $I = \{\ell' \leq L : (\ell, \ell') \notin C\}$. For $\ell' \in I$, let $E_{\ell'}$ be the event

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

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

$$
\text{Prob}_{\omega \sim P}(\bigcup_{\ell' \in I} E_{\ell'}) \leq \epsilon_\ell.
$$

When the event $E$ does not take place, we have $T_{\ell\ell'}(\omega) = 1$ for all $\ell' \in I$, so that $T_{\ell\ell'}(\omega) \geq 0$ for all $\ell'$, $1 \leq \ell' \leq L$, whence $\ell \in T(\omega)$. By Observation 2.10, the latter inclusion implies that

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

Invoking the definition of partial $C$-risk, we get

$$
\text{Risk}_1^\ell(T|H_1, \ldots, H_L) \leq \text{Prob}_{\omega \sim P}(E) \leq \epsilon_\ell.
$$

\qed

2.2.4.4 Testing Multiple Hypotheses via Euclidean separation

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

Our goal is to decide on the hypotheses $\mathcal{H}_1, \ldots, \mathcal{H}_L$ up to closeness $C$ via $K$-repeated observation $\omega^K$. 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 $C$ is the only meaningful closeness on a two-hypotheses set: $(\ell, \ell') \in C$ if and only if $\ell = \ell'$).

**Standing Assumption** which we assume to hold by default everywhere in this section is:

*Whenever $\ell, \ell'$ are not $C$-close: $(\ell, \ell') \notin C$, the sets $X_\ell, X_{\ell'}$ 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 constructed for the pairs $\mathcal{H}_\ell, \mathcal{H}_{\ell'}$ of hypotheses with not close to each other $\ell$ and $\ell'$, that is, with $(\ell, \ell') \notin C$. These tests are built as explained in Section 2.2.3.2; for reader's convenience, here is the construction. For a pair $(\ell, \ell') \notin C$, we

1. Find the optimal value $\text{Opt}_{\ell\ell'}$ and an optimal solution $(u_{\ell\ell'}, v_{\ell\ell'})$ to the convex optimization problem

   \[ \text{Opt}_{\ell\ell'} = \min_{u \in X_\ell, v \in X_{\ell'}} \frac{1}{2} \| u - v \|_2. \]  

   (2.36)

   The latter problem is solvable, since we have assumed from the very beginning that $X_\ell, X_{\ell'}$ are nonempty, closed and convex, and that at least one of these sets is bounded;

2. Set

   \[ e_{\ell\ell'} = \frac{u_{\ell\ell'} - v_{\ell\ell'}}{\| u_{\ell\ell'} - v_{\ell\ell'} \|_2}, \quad c_{\ell\ell'} = \frac{1}{2} e_{\ell\ell'}^T [u_{\ell\ell'} + v_{\ell\ell'}], \quad \phi_{\ell\ell'}(\omega) = e_{\ell\ell'}^T \omega - c_{\ell\ell'}. \]

   Note that the construction makes sense, since by our Standing Assumption for $\ell, \ell'$ in question $X_\ell$ and $X_{\ell'}$ do not intersect. Further, $e_{\ell\ell'}$ and $c_{\ell\ell'}$ clearly depend solely on $(\ell, \ell')$, but not on how we select an optimal solution $(u_{\ell\ell'}, v_{\ell\ell'})$ to (2.36). Finally, we have

   \[ e_{\ell\ell'} = -e_{\ell'\ell}, \quad c_{\ell\ell'} = -c_{\ell'\ell}, \quad \phi_{\ell\ell'}(\cdot) \equiv \phi_{\ell'\ell}(\cdot). \]

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

   \[ \delta_{\ell\ell'} + \delta_{\ell'\ell} = 2\text{Opt}_{\ell\ell'} \]  

   (2.37)

   and specify the single-observation simple test $T_{\ell\ell'}$ deciding on the hypotheses $\mathcal{H}_\ell, \mathcal{H}_{\ell'}$ according to

   \[ T_{\ell\ell'}(\omega) = \begin{cases} \{ \ell \}, & \text{if } \phi_{\ell\ell'}(\omega) \geq \frac{1}{2} [\delta_{\ell\ell'} - \delta_{\ell'\ell}], \\ \{ \ell' \}, & \text{otherwise}. \end{cases} \]
Note that by Proposition 2.5, setting
\[ P_\gamma(\delta) = \int_{\delta}^{\infty} \gamma(s)ds, \]  
we have
\[ \text{Risk}_1(T_{\ell\ell'} | H_\ell, H_{\ell'}) \leq P_\gamma(\delta_{\ell\ell'}), \]
\[ \text{Risk}_2(T_{\ell\ell'} | H_\ell, H_{\ell'}) \leq P_\gamma(\delta_{\ell\ell'}). \]  

(2.39)

b) when \( K > 1 \), we specify the \( K \)-observation simple test \( T_{\ell\ell'} \) deciding on \( H_\ell, H_{\ell'} \) according to
\[ T_{\ell\ell'}(\omega^k = (\omega_1, ..., \omega_k)) = \begin{cases} \{ \ell \}, & \text{Card} \{ k \leq K : \phi_{\ell\ell'} \geq 0 \} \geq K/2, \\ \{ \ell' \}, & \text{otherwise} \end{cases} \]

Note that by Proposition 2.6 we have
\[ \text{Risk}(T_{\ell\ell'} | H_\ell, H_{\ell'}) \leq \epsilon_{\ell\ell'} := \sum_{K/2 \leq k \leq K} \binom{K}{k} \epsilon^k_{\ell\ell'} (1 - \epsilon^k_{\ell\ell'})^{K-k}, \]
\[ \epsilon^*_{\ell\ell'} = P_\gamma(\text{Opt}_{\ell\ell'}) = \epsilon^*_{\ell\ell}. \]

Assembling the building blocks, case of \( K = 1 \). In the case of \( K = 1 \), we specify the simple pairwise tests \( T_{\ell,\ell'} \), \( (\ell, \ell') \notin \mathcal{C} \), participating in the construction of the multi-hypothesis test presented in Section 2.2.4.3, as follows. Given unordered pair \( \{ \ell, \ell' \} \) with \( (\ell, \ell') \notin \mathcal{C} \) (which is exactly the same as \( (\ell', \ell) \notin \mathcal{C} \)), we arrange \( \ell, \ell' \) in an ascending order, thus arriving at ordered pair \( \ell, \ell' \), and set
\[ T_{\ell,\ell'}(\cdot) = T_{\ell\ell'}(\cdot), \]
with the right hand side tests defined as explained above. We then assemble, as explained in Section 2.2.4.3, the tests \( T_{\ell,\ell'} \) into a single-observation test \( T_1 \) deciding on hypotheses \( H_1, ..., H_L \). From (2.34) and (2.39) we conclude that for the just defined tests \( T_{\ell,\ell'} \) and the associated with the tests \( T_{\ell,\ell'} \), via (2.34), quantities \( \epsilon_{\ell\ell'} \) it holds
\[ (\ell, \ell') \notin \mathcal{C} \Rightarrow \epsilon_{\ell\ell'} \leq P_\gamma(\delta_{\ell\ell'}). \]  

(2.40)

Invoking Proposition 2.11, we get

**Proposition 2.12.** 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 \( T_1 \), whatever the choice of nonnegative \( \delta_{\ell\ell'} \), \( (\ell, \ell') \notin \mathcal{C} \), satisfying (2.37), can be upper-bounded as
\[ \text{Risk}^C(T_1 | H_1, ..., H_L) \leq \sum_{\ell' : (\ell, \ell') \notin \mathcal{C}} P_\gamma(\delta_{\ell\ell'}). \]  

(2.41)

with \( P_\gamma(\cdot) \) given by (2.38).

Case of \( K = 1 \) (continued): Optimizing the construction. We can try to
optimize the risk bounds (2.41) over the parameters $\delta_{\ell\ell'}$ of the construction. The first question to be addressed here is what to minimize — we have defined several risks. A natural model here may be 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

$$\min_{t, (\delta_{\ell\ell'}: (\ell, \ell') \notin C)} \left\{ t : \frac{W \cdot \left[ \sum_{\ell': (\ell, \ell') \notin C} P_\gamma(\delta_{\ell\ell'}) \right]}{\delta_{\ell\ell} \geq 0, \delta_{\ell\ell'} + \delta_{\ell'\ell} = 2\text{Opt}_{\ell\ell'}, (\ell, \ell') \notin C} \leq tw \right\}. \quad (2.42)$$

For instance, when $M = 1$ and $w = 1$, we minimize a weighted sum of (upper bounds on) partial $C$-risks of our test; when $W$ is a diagonal matrix with positive diagonal entries and $w$ is the all-ones vector, we minimize the largest of scaled partial risks. Note that when $P_\gamma(\cdot)$ is convex on $\mathbb{R}_+$, or, which is the same, $\gamma(\cdot)$ is nonincreasing in $\mathbb{R}_+$, (2.42) 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 $T_{(\ell, \ell')}$, $(\ell, \ell') \notin C$, we have already specified, to tests $T_{(\ell, \ell')} = T_{(\bar{\ell}, \bar{\ell'})}$, 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 $T_K$. By Proposition 2.6, the quantities $\epsilon_{\ell\ell'}$ associated with the tests $T_{(\ell, \ell')}$ via (2.34) satisfy the relation

$$(\ell, \ell') \notin C \Rightarrow \epsilon_{\ell\ell'} \leq \sum_{K/2 \leq k \leq K} \binom{K}{k} [P_\gamma(\text{Opt}_{\ell\ell'})]^k [1 - P_\gamma(\text{Opt}_{\ell\ell'})]^{K-k}, \quad (2.43)$$

which combines with Proposition 2.11 to imply

**Proposition 2.13.** Consider the situation described in the beginning of Section 2.2.4.4, and let $K > 1$. Under Standing Assumption, the $C$-risks of the just defined test $T_K$ can be upper-bounded as

$$\text{Risk}_C^\bar{c}(T_K|H_1, ..., H_L) \leq \sum_{(\ell', (\ell', (\ell, (\ell') \notin C} \sum_{K/2 \leq k \leq K} \binom{K}{k} [P_\gamma(\text{Opt}_{\ell\ell'})]^k [1 - P_\gamma(\text{Opt}_{\ell\ell'})]^{K-k}, \quad (2.44)$$

with $P_\gamma(\cdot)$ given by (2.38) and $\text{Opt}_{\ell\ell'}$ given by (2.36).

Note that by Standing Assumption the quantities $P_\gamma(\text{Opt}_{\ell\ell'})$ for $(\ell, \ell') \notin C$ are $< 1/2$, so that the risks $\text{Risk}_C^\bar{c}(T_K|H_1, ..., H_L)$ go to 0 exponentially fast as $K \to \infty$.

## 2.3 DETECTORS AND DETECTOR-BASED TESTS

### 2.3.1 Detectors and their risks

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

defined as follows:

\[
\begin{align*}
(a) \quad \text{Risk}_-[\phi|\mathcal{P}_1] &= \sup_{P \in \mathcal{P}_1} \int_\Omega \exp\{-\phi(\omega)\} P(d\omega) \\
(b) \quad \text{Risk}_+[\phi|\mathcal{P}_2] &= \sup_{P \in \mathcal{P}_2} \int_\Omega \exp\{\phi(\omega)\} P(d\omega) \\
(c) \quad \text{Risk}[\phi|\mathcal{P}_1, \mathcal{P}_2] &= \max\{\text{Risk}_-[\phi|\mathcal{P}_1], \text{Risk}_+[\phi|\mathcal{P}_2]\}
\end{align*}
\]

Given a detector \( \phi \), we can associate with it a simple test \( T_\phi \) deciding via observation \( \omega \sim P \) on the hypotheses

\[
H_1 : P \in \mathcal{P}_1, \quad H_2 : P \in \mathcal{P}_2.
\]

Namely, given observation \( \omega \in \Omega \), the test \( T_\phi \) accepts \( H_1 \) (and rejects \( H_2 \)) whenever \( \phi(\omega) \geq 0 \), and accepts \( H_2 \) and rejects \( H_1 \) otherwise.

Let us make the following immediate observation:

**Proposition 2.14.** Let \( \Omega \) be an observation space, \( \mathcal{P}_\chi, \chi = 1, 2 \), be two families of probability distributions on \( \Omega \), and \( \phi \) be a detector. The risks of the test \( T_\phi \) associated with this detector satisfy

\[
\begin{align*}
\text{Risk}_1(T_\phi|H_1, H_2) &\leq \text{Risk}_-[\phi|\mathcal{P}_1], \\
\text{Risk}_2(T_\phi|H_1, H_2) &\leq \text{Risk}_+[\phi|\mathcal{P}_2].
\end{align*}
\]

**Proof.** Let \( \omega \sim P \in \mathcal{P}_1 \). Then the \( P \)-probability of the event \( \{\omega : \phi(\omega) < 0\} \) does not exceed \( \text{Risk}_-[\phi|\mathcal{P}_1] \), since on the set \( \{\omega : \phi(\omega) < 0\} \) the integrand in (2.45.a) is

> 1, and this integrand is nonnegative everywhere, so that the integral in (2.45.a) is \( \geq P\{\omega : \phi(\omega) < 0\} \). Recalling what \( T_\phi \) is, we see that the \( P \)-probability to reject \( H_1 \) is at most \( \text{Risk}_-[\phi|\mathcal{P}_1] \), implying the first relation in (2.47). By the similar argument, with (2.45.b) in the role of (2.45.a), when \( \omega \sim P \in \mathcal{P}_2 \), the \( P \)-probability of the event \( \{\omega : \phi(\omega) \geq 0\} \) is upper-bounded by \( \text{Risk}_+[\phi|\mathcal{P}_2] \), implying the second relation in (2.47).

\[\square\]

### 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 \( \epsilon_1 \) and \( \epsilon_2 \) are upper bounds on the risks of a detector are expressed by system of convex constraints

\[
\begin{align*}
\sup_{P \in \mathcal{P}_1} \int_\Omega \exp\{-\phi(\omega)\} P(d\omega) &\leq \epsilon_1 \\
\sup_{P \in \mathcal{P}_2} \int_\Omega \exp\{\phi(\omega)\} P(d\omega) &\leq \epsilon_2
\end{align*}
\]

on \( \epsilon_1, \epsilon_2 \) and \( \phi(\cdot) \). This observation is interesting, but not very 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 an 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 \( \Omega, \mathcal{P}_1 \) and \( \mathcal{P}_2 \) be the same as in Section 2.3.1, and let \( \phi \) be a detector. When shifting this detector by a real \( a \) — passing from \( \phi \) to the detector \( \phi_a(\omega) = \phi(\omega) - a \)

— the risks clearly update according to:

\[
\begin{align*}
\text{Risk}_-[\phi_a|\mathcal{P}_1] &= e^a \text{Risk}_-[\phi|\mathcal{P}_1], \\
\text{Risk}_+[\phi_a|\mathcal{P}_2] &= e^{-a} \text{Risk}_+[\phi|\mathcal{P}_2].
\end{align*}
\]

(2.48)

We see that

When speaking about risks of a detector, what matters is the product

\[
\text{Risk}_-[\phi|\mathcal{P}_1]\text{Risk}_+[\phi|\mathcal{P}_2]
\]

of the risks, not these risks individually: by shifting the detector, we can re-distribute 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

\[
\text{Risk}_-[\phi|\mathcal{P}_1] = \text{Risk}_+[\phi|\mathcal{P}_2] = \text{Risk}[\phi|\mathcal{P}_1, \mathcal{P}_2].
\]

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 \( T_\phi \) associated with a balanced detector \( \phi \) is bounded by the risk \( \text{Risk}[\phi|\mathcal{P}_1, \mathcal{P}_2] \) of the detector:

\[
\text{Risk}(T_\phi|H_1, H_2) := \max [\text{Risk}_1(T_\phi|H_1, H_2), \text{Risk}_2(T_\phi|H_1, H_2)] \\
\leq \text{Risk}[\phi|\mathcal{P}_1, \mathcal{P}_2].
\]

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 a single observation to those based on repeated observations.

Given \( K \) observation spaces \( \Omega_k, 1 \leq k \leq K \), each equipped with a pair \( \mathcal{P}_{k,1}, \mathcal{P}_{k,2} \) of families of probability distributions, we can build a new observation space

\[ \Omega^K = \Omega_1 \times ... \times \Omega_K = \{\omega^K = (\omega_1, ..., \omega_K) : \omega_k \in \Omega_k, k \leq K\} \]

and equip it with two families \( \mathcal{P}^\chi, \chi = 1, 2, \) of probability distributions; distributions from \( \mathcal{P}^\chi \) are exactly the product-type distributions \( P = P_1 \times ... \times P_K \) with all factors \( P_k \) taken from \( \mathcal{P}_{k,\chi} \). Observations \( \omega^K = (\omega_1, ..., \omega_K) \) from \( \Omega^K \) drawn from a distribution \( P = P_1 \times ... \times P_K \in \mathcal{P}^\chi \) are nothing but collections of observations \( \omega_k, k = 1, ..., K \), drawn, independently of each other, from distributions \( P_k \). Now,
given detectors $\phi_k(\cdot)$ on observation spaces $\Omega_k$ and setting

$$\phi^{(K)}(\omega^K) = \sum_{k=1}^{K} \phi_k(\omega_k) : \Omega^K \to \mathbb{R},$$

we clearly have

$$\text{Risk}_-[\phi^{(K)}|\mathcal{P}_1^K] = \prod_{k=1}^{K} \text{Risk}_-[\phi_k|\mathcal{P}_{k,1}],$$

$$\text{Risk}_+[\phi^{(K)}|\mathcal{P}_2^K] = \prod_{k=1}^{K} \text{Risk}_+[\phi_k|\mathcal{P}_{k,2}].$$

(2.49)

Let us look at some useful consequences of (2.49).

**Stationary $K$-repeated observations.** Consider the case of Section 2.1.3.1: we are given an observation space $\Omega$ and a positive integer $K$, and what we observe, is a sample $\omega^K = (\omega_1, ..., \omega_K)$ with $\omega_1, ..., \omega_K$ drawn, independently of each other, from some distribution $P$ on $\Omega$. Let now $\mathcal{P}_1$, $\mathcal{P}_2$, be two families of probability distributions on $\Omega$; we can associate with these families two hypotheses, $H_1^{\otimes,K}$, $H_2^{\otimes,K}$, on the distribution of $K$-repeated observation $\omega^K = (\omega_1, ..., \omega_K)$, with $H_1^{\otimes,K}$ stating that $\omega_1, ..., \omega_K$ are drawn, independently of each other, from a distribution $P \in \mathcal{P}_1$. Given a detector $\phi$ on $\Omega$, we can associate with it the detector

$$\phi^{(K)}(\omega^K) = \sum_{k=1}^{K} \phi(\omega_k)$$
on $\Omega^K := \Omega \times \ldots \times \Omega$. Combining (2.49) and Proposition 2.14, we arrive at the following nice result:

**Proposition 2.15.** Consider the simple test $T_{\phi^{(K)}}$ deciding, given $K$-repeated observation $\omega^K = (\omega_1, ..., \omega_K)$ on the hypotheses

$$H_1^{\otimes,K} : \omega_k, k \leq K, \text{ are independently of each other drawn from } P \in \mathcal{P}_1,$$

$$H_2^{\otimes,K} : \omega_k, k \leq K, \text{ are independently of each other drawn from } P \in \mathcal{P}_2$$

according to the rule

$$\phi^{(K)}(\omega^K) := \sum_{k=1}^{K} \phi(\omega_k) \begin{cases} 
\geq 0 & \Rightarrow \text{ accept } H_1^{\otimes,K} \\
< 0 & \Rightarrow \text{ accept } H_2^{\otimes,K}
\end{cases}$$

The risks of $T_{\phi^{(K)}}$ admit the upper bounds

$$\text{Risk}_1(T_{\phi^{(K)}}|H_1^{\otimes,K}, H_2^{\otimes,K}) \leq (\text{Risk}_-[\phi|\mathcal{P}_1])^K$$

$$\text{Risk}_2(T_{\phi^{(K)}}|H_1^{\otimes,K}, H_2^{\otimes,K}) \leq (\text{Risk}_+[\phi|\mathcal{P}_2])^K$$

**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 \( \Omega \) were defined in Sections 2.1.3.2 and 2.1.3.3, respectively. It turns out that Proposition 2.15 extends to quasi-stationary \( K \)-repeated observations:

**Proposition 2.16.** Let \( \Omega \) be an observation space, \( \mathcal{P}_\chi, \chi = 1, 2 \) be families of probability distributions on \( \Omega \), \( \phi: \Omega \to \mathbb{R} \) 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 \( \omega^K \):

\[
H^{\otimes^K}_\chi: P^K \in \mathcal{P}^{\otimes^K}_\chi, \chi = 1, 2
\]

(see Section 2.1.3.3), and \( \phi \) 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 \( \mathcal{P}^{\otimes^K}_\chi, \chi = 1, 2 \), can be upper-bounded as follows:

\[
\begin{align*}
\text{Risk}_- [\phi^K | P^{\otimes^K}_1] & \leq (\text{Risk}_- [\phi | P_1])^K, \\
\text{Risk}_+ [\phi^K | P^{\otimes^K}_2] & \leq (\text{Risk}_+ [\phi | P_2])^K.
\end{align*}
\]

Furthermore, the detector \( \phi^K \) induces simple test \( T_{\phi^K} \) deciding on \( H^{\otimes^K}_\chi, \chi = 1, 2 \) as follows: given \( \omega^K \), the test accepts \( H^{\otimes^K}_1 \) when \( \phi^K(\omega^K) \geq 0 \), and accepts \( H^{\otimes^K}_2 \) otherwise. The risks of this test can be upper-bounded as

\[
\begin{align*}
\text{Risk}_1 (T_{\phi^K} | H^{\otimes^K}_1, H^{\otimes^K}_2) & \leq (\text{Risk}_- [\phi | P_1])^K, \\
\text{Risk}_2 (T_{\phi^K} | H^{\otimes^K}_1, H^{\otimes^K}_2) & \leq (\text{Risk}_+ [\phi | P_2])^K.
\end{align*}
\]

Finally, the above results remain intact when passing from quasi-stationary to semi-stationary \( K \)-repeated observations (that is, when replacing \( \mathcal{P}^{\otimes^K}_\chi \) with \( \mathcal{P}^{\oplus^K}_\chi = \bigoplus_{k=1}^K \mathcal{P}_\chi \) and \( H^{\otimes^K}_\chi \) with the hypotheses \( H^{\oplus^K}_\chi \) stating that the distribution of \( \omega^K \) belongs to \( \mathcal{P}^{\oplus^K}_\chi, \chi = 1, 2 \)).

**Proof.** All we need is to verify (2.50) – in view of Proposition 2.14, all other claims in Proposition 2.16 are immediate consequences of (2.50) and the inclusions \( \mathcal{P}^{\oplus^K}_\chi \subset \mathcal{P}^{\otimes^K}_\chi, \chi = 1, 2 \). Verification of (2.50) is as follows. Let \( P^K \in \mathcal{P}^{\otimes^K}_1 \), so that by definition of \( \mathcal{P}^{\otimes^K}_1, P^K \) is the distribution of random sequence \( \omega^K = (\omega_1, ..., \omega_K) \) such that there exists a random sequence of driving factors \( \zeta_1, ..., \zeta_K \) such that \( \omega_k \) is a deterministic function of \( \zeta_k \):

\[
\omega_k = \theta_k(\zeta_1, ..., \zeta_k),
\]

and the conditional, given \( \zeta_1, ..., \zeta_{k-1} \), distribution \( P_{\omega_k|\zeta_{k-1}} \) belongs to \( P_1 \). Let \( P_{k} \) be the distribution of the first \( k \) driving factors, and \( P_{k|\zeta_{k-1}} \) be the conditional,
given \( \zeta_1, \ldots, \zeta_{k-1} \), distribution of \( \zeta_k \). Let us put

\[
\psi^{(k)}(\zeta_1, \ldots, \zeta_k) = \sum_{\ell=1}^{k} \phi(\theta_{\ell}(\zeta_1, \ldots, \zeta_{\ell})),
\]

so that

\[
\int_{\Omega^K} \exp\{-\phi^{(K)}(\omega^k)\} P^K(d\omega^K) = \int \exp\{-\psi^{(K)}(\zeta^k)\} P_{\zeta^k} K(d\zeta^k). \tag{2.51}
\]

On the other hand, denoting \( C_0 = 1 \), we have

\[
C_k := \int e^{-\psi^{(k)}(\zeta^k)} P_{\zeta^k}(d\zeta^k) = \int \exp\{-\psi^{(k-1)}(\zeta^{k-1}) - \phi(\theta_k(\zeta^k))\} P_{\zeta^k}(d\zeta^k)
\]

\[
= \int e^{-\psi^{(k-1)}(\zeta^{k-1})} \left[ \int e^{-\phi(\theta_k(\zeta^k))} P_{\zeta_k|\zeta^{k-1}}(d\zeta^k) \right] P_{\zeta^{k-1}}(d\zeta^{k-1})
\]

\[
\leq \operatorname{Risk}_-[\phi|\mathcal{P}_1] \int e^{-\phi^{(k-1)}(\zeta^{k-1})} P_{\zeta^{k-1}}(d\zeta^{k-1}) = \operatorname{Risk}_-[\phi|\mathcal{P}_1] C_{k-1}
\]

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

\[
C_K \leq (\operatorname{Risk}_-[\phi|\mathcal{P}_1])^K,
\]

which combines with (2.51) to imply that

\[
\int_{\Omega^K} e^{-\phi^{(K)}(\omega^k)} P^K(d\omega^K) \leq (\operatorname{Risk}_-[\phi|\mathcal{P}_1])^K.
\]

The latter inequality holds true for every distribution \( P^K \in \mathcal{P}_\chi^K \), and the first inequality in (2.50) follows. The second inequality in (2.50) is given by a completely similar reasoning, with \( \mathcal{P}_2 \) in the role of \( \mathcal{P}_1 \), and \(-\phi, -\phi^{(K)}\) in the roles of \( \phi, \phi^{(K)} \), respectively.

The fact that observations \( \omega_k \) under hypotheses \( H^\otimes_{\chi} \), \( \ell = 1, 2 \) are related to “constant in time” families \( \mathcal{P}_t \) has no importance here, and in fact the proof of Proposition 2.16 after absolutely evident modifications of wording allows to justify the following “non-stationary” version of the proposition:

**Proposition 2.17.** For \( k = 1, \ldots, K \), let \( \Omega_k \) be observation spaces, \( \mathcal{P}_{\chi,k} \), \( \chi = 1, 2 \) be families of probability distributions on \( \Omega_k \), and \( \phi_k : \Omega_k \rightarrow \mathbb{R} \) be detectors.

Families \( \mathcal{P}_{\chi,k} \), \( \chi = 1, 2 \), give rise to quasi-direct products (see Section 2.1.3.3)

\[
\mathcal{P}_\chi^K = \bigotimes_{k=1}^{K} \mathcal{P}_{\chi,k}
\]

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 \( \omega^K = (\omega_1, \ldots, \omega_K) \in \Omega^K = \Omega_1 \times \ldots \times \Omega_K \):

\[
H^\otimes_{\chi} : P^K \in \mathcal{P}_\chi^K, \chi = 1, 2.
\]
Detectors $\phi_k$, $1 \leq k \leq K$, induce the detector

$$\phi^K(\omega^K) := \sum_{k=1}^{K} \phi_k(\omega_k).$$

The risks of the detector $\phi^K$ on the families $\mathcal{P}_x^{\otimes,K}$, $x = 1, 2$, can be upper-bounded as follows:

$$\text{Risk}_-\left[\phi^K|\mathcal{P}_1^{\otimes,K}\right] \leq \prod_{k=1}^{K} \text{Risk}_-\left[\phi_k|\mathcal{P}_{1,k}\right],$$

$$\text{Risk}_+\left[\phi^K|\mathcal{P}_2^{\otimes,K}\right] \leq \prod_{k=1}^{K} \text{Risk}_+\left[\phi_k|\mathcal{P}_{2,k}\right].$$

Further, the detector $\phi^K$ induces simple test $T_{\phi^K}$ deciding on $H_x^{\otimes,K}$, $x = 1, 2$ as follows: given $\omega^K$, 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

$$\text{Risk}_1\left(T_{\phi^K}|H_1^{\otimes,K},H_2^{\otimes,K}\right) \leq \prod_{k=1}^{K} \text{Risk}_-\left[\phi_k|\mathcal{P}_{1,k}\right],$$

$$\text{Risk}_2\left(T_{\phi^K}|H_1^{\otimes,K},H_2^{\otimes,K}\right) \leq \prod_{k=1}^{K} \text{Risk}_+\left[\phi_k|\mathcal{P}_{2,k}\right].$$

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 $\mathcal{P}_x^{\otimes,K}$ with $\mathcal{P}_x^{\oplus,K}$ and $H_x^{\otimes,K}$ with the hypotheses $H_x^{\oplus,K}$ stating that the distribution of $\omega^K$ belongs to $\mathcal{P}_x^{\oplus,K}$, $x = 1, 2$).

### 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 2.18.** Let $\Omega$ be an observation space, and $\mathcal{P}_x$, $x = 1, 2$, be families of probability distributions on $\Omega$. Assume that for some $\epsilon \in (0, 1/2)$ “in the nature” there exists a simple test (deterministic or randomized) deciding on the hypotheses

$$H_1 : P \in \mathcal{P}_1, \quad H_2 : P \in \mathcal{P}_2$$

on the distribution $P$ of observation $\omega$ with risks $\leq \epsilon$:

$$\text{Risk}_1(T|H_1, H_2) \leq \epsilon \quad \& \quad \text{Risk}_2(T|H_1, H_2) \leq \epsilon.$$

Then there exists a detector-based test $T_{\phi}$ deciding on the same pair of hypotheses with the risk “comparable” with $\epsilon$:

$$\text{Risk}_1(T_{\phi}|H_1, H_2) \leq \epsilon^+ \quad \& \quad \text{Risk}_2(T_{\phi}|H_1, H_2) \leq \epsilon^+, \quad \epsilon^+ = 2\sqrt{\epsilon(1-\epsilon)}. \quad (2.52)$$

**Proof.** Let us prove the claim in the case when the test $T$ is deterministic; the case when this test is randomized is the subject of Exercise 2.11.

For $x = 1, 2$, let $\Omega_x$ be the set of $\omega \in \Omega$ such that $T$ if “fed” with observation $\omega$ accepts $H_x$. Since $T$ is simple, $\Omega_1, \Omega_2$ split $\Omega$ into two non-overlapping parts, and
since the risks of $\mathcal{T}$ are $\leq \epsilon$, we have

$$
\epsilon_2(P) := P\{\Omega_2\} \leq \epsilon \quad \forall P \in \mathcal{P}_1,
$$

$$
\epsilon_1(P) := P\{\Omega_1\} \leq \epsilon \quad \forall P \in \mathcal{P}_2.
$$

Let $\delta = \sqrt{(1-\epsilon)/\epsilon}$, so that $\delta \geq 1$ due to $0 < \epsilon \leq 1/2$, and let

$$
\psi(\omega) = \begin{cases} 
\delta, & \omega \in \Omega_1 \\
1/\delta, & \omega \in \Omega_2
\end{cases},
\phi(\omega) = \ln(\psi(\omega)).
$$

When $P \in \mathcal{P}_1$ we have

$$
\int_{\Omega} \exp\{-\phi(\omega)\} P(d\omega) = \frac{1}{\delta} P\{\Omega_1\} + \delta P\{\Omega_2\} = \frac{1}{\delta} + \left[\delta - \frac{1}{\delta}\right] \epsilon_2(P) \leq \epsilon + \left[\delta - \frac{1}{\delta}\right] \epsilon = \epsilon^+, \\
\text{whence Risk}_{-}[\phi|\mathcal{P}_1] \leq \epsilon^+.
$$

Similarly, when $P \in \mathcal{P}_2$ we have

$$
\int_{\Omega} \exp\{\phi(\omega)\} P(d\omega) = \delta P\{\Omega_1\} + \frac{1}{\delta} P\{\Omega_2\} = \left[\delta - \frac{1}{\delta}\right] \epsilon_1(P) + \frac{1}{\delta} \leq \left[\delta - \frac{1}{\delta}\right] \epsilon + \frac{1}{\delta} = \epsilon^+, \\
\text{whence Risk}_{+}[\phi|\mathcal{P}_2] \leq \epsilon^+.
$$

**Discussion.** Proposition 2.18 states that we can restrict ourselves with detector-based tests at the price of passing from risk $\epsilon$ 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 $\epsilon$ and $\epsilon^+$ “comparable” is too optimistic. 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 $\mathcal{T}_\delta$ 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 $\mathcal{T}_\delta$ 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 2.2.3.2.\(^4\) Computation shows that with $\epsilon_* = 0.01$, to make the risk of the majority test $\leq 0.0001$ takes 5 observations, which is only marginally better than the 6 observations needed in the detector-based construction.

\(^4\)In Section 2.2.3.2, we dealt with “signal plus noise” observations and with specific test $\mathcal{T}$ given by Euclidean separation. However, a straightforward inspection of the construction and the proof of Proposition 2.6 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 (2.23), with the risk of $\mathcal{T}$ in the role of $\epsilon_*$. 
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 \( \Omega \) and two families, \( P_1 \) and \( P_2 \), of probability distributions on \( \Omega \), solve the optimization problem

\[
\text{Opt} = \min_{\phi: \Omega \to \mathbb{R}} \max_{P \in P_1} \int_{\Omega} e^{-\phi(\omega)} P(d\omega), \quad \max_{P \in P_2} \int_{\Omega} e^{\phi(\omega)} P(d\omega).
\]  

(2.53)

While being convex, problem (2.53) is typically 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 \( P_1 \) and/or \( P_2 \) 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 (2.53) is efficiently solvable.

To arrive at the notion of a simple observation scheme, consider the case when all distributions from \( P_1 \), \( P_2 \) admit densities taken w.r.t. some reference measure \( \Pi \) on \( \Omega \), and these densities are parameterized by a “parameter” \( \mu \) running through some parameter space \( M \). In other words, \( P_1 \) is comprised of all distributions with densities \( p_\mu(\cdot) \) and \( \mu \) belonging to some subset \( M_1 \) of \( M \), while \( P_2 \) is comprised of distributions with densities \( p_\mu(\cdot) \) and \( \mu \) belonging to another subset, \( M_2 \), of \( M \). To save words, we shall identify distributions with their densities taken w.r.t. \( \Pi \), so that

\[
P_\chi = \{ p_\mu : \mu \in M_\chi \}, \quad \chi = 1, 2,
\]

where \( \{ p_\mu(\cdot) : \mu \in M \} \) is a given “parametric” family of probability densities. Quotation marks in “parametric” reflect the fact that at this point in time, the “parameter” \( \mu \) can be infinite-dimensional (e.g., we can parameterize a density by itself), so that assuming “parametric” representation of the distributions from \( P_1 \), \( P_2 \) in fact does not restrict the generality.

Our first observation is that in our “parametric” setup, we can rewrite problem (2.53) equivalently as

\[
\ln(\text{Opt}) = \min_{\phi: \Pi \to \mathbb{R}} \sup_{\mu \in M_1, \nu \in M_2} \frac{1}{2} \left[ \ln \left( \int_{\Omega} e^{-\phi(\omega)} p_\mu(\omega) \Pi(d\omega) \right) + \ln \left( \int_{\Omega} e^{\phi(\omega)} p_\nu(\omega) \Pi(d\omega) \right) \right].
\]  

(2.54)

Indeed, when shifting \( \phi \) by a constant: \( \phi(\cdot) \mapsto \phi(\cdot) - a \), the positive quantities \( F[\phi] \) and \( G[\phi] \) participating in (2.53) are multiplied by \( e^a \) and \( e^{-a} \), respectively, and their product remains intact. It follows that to minimize over \( \phi \) the maximum of \( F[\phi] \) and \( G[\phi] \) (this is what (2.53) wants of us) is exactly the same as to minimize over \( \phi \) the quantity \( H[\phi] := \sqrt{F[\phi]G[\phi]} \).

Indeed, a candidate solution \( \phi \) 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
As a result, minimizing $H$ over all $\phi$ is the same as minimizing $H$ over balanced $\phi$, and the latter problem clearly is equivalent to (2.53). It remains to note that (2.54) is nothing but the problem of minimizing $\ln(H[\phi])$.

Now, (2.54) 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 $\Psi$ is convex in the minimization argument (this indeed is the case for (2.54)), and, moreover, the domain of the maximization argument is convex, and the cost function is concave in this argument (this not necessarily is the case for (2.54)). 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 – the definition

Consider the situation in which we are given

1. A Polish (complete separable metric) observation space $\Omega$ equipped with $\sigma$-finite $\sigma$-additive Borel reference measure $\Pi$ such that the support of $\Pi$ is the entire $\Omega$. 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:
   - $\Omega = \mathbb{R}^d$ equipped with the Lebesgue measure $\Pi$, and
   - a finite or countable set $\Omega$ which is discrete (distances between distinct points are equal to 1) and is equipped with the counting measure $\Pi$.

2. A parametric family $\{p_\mu(\cdot) : \mu \in \mathcal{M}\}$ of probability densities, taken w.r.t. $\Pi$, such that
   - the space $\mathcal{M}$ of parameters is a convex set in some $\mathbb{R}^n$ which coincides with its relative interior,
   - the function $p_\mu(\cdot) : \mathcal{M} \times \Omega \rightarrow \mathbb{R}$ is continuous in $(\mu, \omega)$ and positive everywhere.

3. A finite-dimensional linear subspace $\mathcal{F}$ of the space of continuous functions on $\Omega$ 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
     $$\ln \left( \int_{\Omega} e^{\phi(\omega)} p_\mu(\omega) \Pi(d\omega) \right)$$
     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 the basic examples of s.o.s.’s.

**2.4.3.1 Gaussian observation scheme**

In Gaussian o.s.,

- the observation space $(\Omega, \Pi)$ is the space $\mathbb{R}^d$ 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 $\Theta$, distributions from the family are parameterized by their expectations $\mu$. Thus, $\mathcal{M} = \mathbb{R}^d$, $p_\mu(\omega) = \frac{1}{(2\pi)^{d/2} \sqrt{\text{Det}(\Theta)}} \exp\left\{ -\frac{1}{2} (\omega - \mu)^T \Theta^{-1} (\omega - \mu) \right\}$;
- the family $\mathcal{F}$ is the family of all affine functions on $\mathbb{R}^d$.

It is immediately seen that Gaussian o.s. meets all requirements imposed on a simple o.s. For example,

$$\ln(p_\mu(\omega)/p_\nu(\omega)) = (\nu - \mu)^T \Theta^{-1} \omega + \frac{1}{2} [\nu^T \Theta^{-1} \nu - \mu^T \Theta^{-1} \mu]$$

is an affine function of $\omega$ 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

$$f(\mu) := \ln \left( \int_{\mathbb{R}^d} e^{\phi(\omega)} p_\mu(\omega) d\omega \right) = \ln \left( E_{\xi \sim \mathcal{N}(0, I_d)} \left\{ \exp\left\{ a^T (\Theta^{1/2} \xi + \mu) + b \right\} \right\} \right)$$

is affine (and thus concave) function of $\mu$.

As we remember from Chapter 1, Gaussian o.s. is responsible for the standard signal processing model where one is given a noisy observation

$$\omega = Ax + \xi$$

of the image $Ax$ of unknown signal $x \in \mathbb{R}^n$ 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 $\omega$ is drawn from Gaussian distribution with known covariance matrix $\Theta$ and expectation known to belong to the set $M = \{\mu = Ax : x \in X\}$. Therefore, deciding upon various hypotheses on where $x$ is located
HYPOTHESIS TESTING

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 $\Omega$ is the set $\mathbb{Z}_+^d$ 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, i.e.
  \[ \mathcal{M} = \{\mu \in \mathbb{R}^d : \mu > 0\}, \quad p_\mu(\omega) = \frac{\mu_{\omega_1}^{\omega_1} \mu_{\omega_2}^{\omega_2} \cdots \mu_{\omega_d}^{\omega_d}}{\omega_1! \omega_2! \cdots \omega_d!} e^{-\mu_1 - \mu_2 - \cdots - \mu_d}, \quad \omega \in \mathbb{Z}_+^d. \]

In other words, random variable $\omega \sim p_\mu$, $\mu \in \mathcal{M}$, is $d$-dimensional vector with independent random entries, with the $i$-th entry $\omega_i \sim \text{Poisson}(\mu_i)$;
- the space $\mathcal{F}$ is comprised of affine functions on $\mathbb{Z}_+^d$.

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 an affine function of $\omega$ 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

$$
f(\mu) := \ln \left( \int_{\Omega} e^{\phi(\omega)} p_\mu(\omega) \Pi(d\omega) \right) = \ln \left( \sum_{\omega \in \mathbb{Z}_+^d} e^{a^T\omega + b} \prod_{i=1}^d \frac{\mu_i^{\omega_i} e^{-\mu_i}}{\omega_i!} \right)
= b + \ln \left( \prod_{i=1}^d \left( e^{-\mu_i} \sum_{s=0}^\infty \frac{[e^{a_i}\mu_i^s]}{s!} \right) \right) = b + \sum_{i=1}^d \ln(\exp\{e^{a_i}\mu_i - \mu_i\})
= \sum_i [e^{a_i} - 1]\mu_i + b
$$

is an affine (and thus concave) function of $\mu$.

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 $\lambda_j$, distribution, so that the number of customers arriving at source $j$ in a unit time interval is Poisson random variable with parameter $\lambda_j$. 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. The dispatches are independent of each other and of the arrival processes. What we observe is the vector $\omega = (\omega_1, ..., \omega_d)$, where $\omega_i$ 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 $\omega_i$ in $\omega$ 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 \( \omega \) 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” \( \lambda \).

Poisson imaging has several important applications,\(^5\) for example, in Positron Emission Tomography (PET). In PET (see Figure 2.3), 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, giving rise to two \( \gamma \)-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 \( \gamma \)-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 each bin. When discretizing the field of view (patient’s body) into small 3D cubes (voxels) we arrive at an accurate enough model of the data which is a realization \( \omega \) of a random vector with independent

\(^5\)In all these applications, the signal \( \lambda \) we ultimately are interested in is an image, this is where “Imaging” comes from.
Poisson entries $\omega_i \sim \text{Poisson}(\mu_i)$, with $\mu_i$ given by

$$
\mu_i = \sum_{j=1}^{n} p_{ij} \lambda_j
$$

where $\lambda_j$ is proportional to the amount of the 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 to concentrate 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 $\omega$ 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 $\lambda_j$ 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 worth noting that there are two other real life applications of Poisson Imaging: Large Binocular Telescope and Nanoscale Fluorescent Microscopy.\(^6\)

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 $\Omega$, that is,

$$
\mathcal{M} = \left\{ \mu \in \mathbb{R}^d : \mu > 0, \sum_{\omega \in \Omega} \mu_\omega = 1, p_\mu(\omega) = \mu_\omega, \ \omega \in \Omega \right\};
$$

- $\mathcal{F}$ is the space of all real-valued functions on the finite set $\Omega$.

Clearly, Discrete o.s. is simple; the function

$$
f(\mu) := \ln \left( \int_{\Omega} e^{\phi(\omega)} p_\mu(\omega) \Pi(d\omega) \right) = \ln \left( \sum_{\omega \in \Omega} e^{\phi(\omega)} \mu_\omega \right)
$$

indeed is concave in $\mu \in \mathcal{M}$.

\(^6\)Large Binocular Telescope [16, 17] 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 triggered by the advent of techniques [18, 111, 115, 208] (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.
2.4.3.4 Direct products of simple observation schemes

Given $K$ simple observation schemes

$$\mathcal{O}_k = (\Omega_k, \Pi_k; \{p_{\mu,k}(\cdot) : \mu \in M_k\}; \mathcal{F}_k), \quad 1 \leq k \leq K,$$

we can define their direct product

$$\mathcal{O}^K = \prod_{k=1}^K \mathcal{O}_k = (\Omega^K, \Pi^K; \{p_\mu : \mu \in M^K\}; \mathcal{F}^K)$$

by modeling the situation where our observation is a tuple $\omega^K = (\omega_1, \ldots, \omega_K)$ with components $\omega_k$ yielded, independently of each other, by observation schemes $\mathcal{O}_k$, namely, as follows:

- The observation space $\Omega^K$ is the direct product of observations spaces $\Omega_1, \ldots, \Omega_K$, and the reference measure $\Pi^K$ is the product of the measures $\Pi_1, \ldots, \Pi_K$;
- The parameter space $M^K$ is the direct product of partial parameter spaces $M_1, \ldots, M_K$, and the distribution $p_\mu(\omega^K)$ associated with parameter $\mu = (\mu_1, \mu_2, \ldots, \mu_K) \in M^K = M_1 \times \ldots \times M_K$ is the probability distribution on $\Omega^K$ with the density

$$p_\mu(\omega^K) = \prod_{k=1}^K p_{\mu,k}(\omega_k)$$

w.r.t. $\Pi^K$. In other words, random observation $\omega^K \sim p_\mu$ is a sample of observations $\omega_1, \ldots, \omega_K$, drawn, independently of each other, from the distributions $p_{\mu_1,1}, p_{\mu_2,2}, \ldots, p_{\mu_K,K}$;
- 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 simple o.s.

$$\mathcal{O} = (\Omega, \Pi; \{p_\mu : \mu \in M\}; \mathcal{F}),$$

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

$$[\mathcal{O}]^K = (\Omega^K, \Pi^K; \{p_\mu^K : \mu \in M\}; \mathcal{F}^{(K)})$$

and defined as follows.
HYPOTHESIS TESTING

- $\Omega^K$ and $\Pi^K$ are exactly the same as in the direct product:
  $$\Omega^K = \Omega \times \ldots \times \Omega, \quad \Pi^K = \Pi \times \ldots \times \Pi;$$

- the parameter space is $\mathcal{M}$ rather than the direct product of $K$ copies of $\mathcal{M}$, and the densities are
  $$p^{(K)}(\omega^K = (\omega_1, \ldots, \omega_K)) = \prod_{k=1}^{K} p_{\mu}(\omega_k).$$
  In other words, random observations $\omega^K \sim p^{(K)}$ are $K$-element samples with components drawn, independently of each other, from $p_{\mu}$;

- 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 2.19.** 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 an “executive summary” of 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, see, e.g., [218] or [15, Section D.4].

Let $U$ and $V$ be nonempty sets, and let $\Phi : U \times V \to \mathbb{R}$ 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 $U, V, \Phi$ 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

\[ \text{Opt}(P) = \inf_{u \in U} \Phi(u) = \inf_{u \in U} \sup_{v \in V} \Phi(u, v) \quad (P) \]

associated with the data \( U, V, \Phi \).

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

\[ \Phi(v) = \inf_{u \in U} \Phi(u, v). \]

In this situation, a risk-averse player II would select \( v \) by maximizing the above worst-case payment, by solving the dual problem

\[ \text{Opt}(D) = \sup_{v \in V} \Phi(v) = \sup_{v \in V} \inf_{u \in U} \Phi(u, v) \quad (D) \]

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, \( \text{Opt}(P) \), should be \( \geq \) his guaranteed payment, \( \text{Opt}(D) \), in the second situation:

\[ \text{Opt}(P) := \inf_{u \in U} \sup_{v \in V} \Phi(u, v) \geq \sup_{v \in V} \inf_{u \in U} \Phi(u, v) =: \text{Opt}(D). \]

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 \( \Phi \) has a saddle point on \( U \times V \).

**Definition 2.20.** A point \((u_*, v_*) \in U \times V\) is called a saddle point \(^7\) of function \( \Phi(u, v) : U \times V \to \mathbb{R} \), if \( \Phi \) 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

\[ \Phi(u, v_*) \geq \Phi(u, v) \geq \Phi(u_*, v) \quad \forall (u \in U, v \in V). \]

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. Indeed, 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 \( \Phi \) 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 \( \Phi \) can have many saddle points, the values of \( \Phi \) at all these points are equal to each other, we denote this common 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 \( u \in U \),

\(^7\)more precisely, “saddle point \((\min_{u \in U} \max_{v \in V} \Phi(u, v))\);” 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.
his worst-case, over the selections of player II, loss can be only $\geq \text{SadVal}$. Similarly, when player II selects $v = v_*$, his worst-case, over the selections of player I, gain is $\text{SadVal}$, and if player II selects a whatever $v \in V$, his worst-case, over the selections of player I, gain can be only $\leq \text{SadVal}$.

Existence of saddle points of $\Phi$ (min in $u \in U$, max in $v \in V$) can be expressed in terms of the primal problem ($P$) and the dual problem ($D$):

**Proposition 2.21.** $\Phi$ 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:

$$
\text{Opt}(P) := \inf_{u \in U} \sup_{v \in V} \Phi(u, v) = \sup_{v \in V} \inf_{u \in U} \Phi(u, v) =: \text{Opt}(D). \quad (2.55)
$$

Whenever this is the case, the saddle points of $\Phi$ are exactly the pairs $(u_*, v_*)$ comprised of optimal solutions to problems ($P$) and ($D$), and the value of $\Phi$ at every one of these points is the common value $\text{SadVal}$ of $\text{Opt}(P)$ and $\text{Opt}(D)$.

Existence of a saddle point of a function is a “rare commodity,” an the standard sufficient condition for it is convexity-concavity of $\Phi$ coupled with convexity of $U$ and $V$. The precise statement is as follows:

**Theorem 2.22.** [Sion-Kakutani, see, e.g., [218] or [15, Theorems D.4.3, D.4.4]]

Let $U \subset \mathbb{R}^m, V \subset \mathbb{R}^n$ be nonempty closed convex sets, with $V$ bounded, and let $\Phi : U \times V \to \mathbb{R}$ 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 (2.55) holds true (although it may happen that $\text{Opt}(P) = \text{Opt}(D) = -\infty$).

If, in addition, $\Phi$ is coercive in $u$, meaning that the level sets

$$
\{u \in U : \Phi(u, v) \leq a\}
$$

are bounded for every $a \in \mathbb{R}$ 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 $\infty$: $\|u_i\| \to \infty$ as $i \to \infty$), then $\Phi$ 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 2.22; the latter, however, covers all our related needs.

### 2.4.4.2 Main Result

**Theorem 2.23.** 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

$$
\Phi(\phi, [\mu; \nu]) = \frac{1}{2} \left[ \ln \left( \int_{\Omega} e^{-\phi(\omega)} p_\mu(\omega) \Pi(d\omega) \right) + \ln \left( \int_{\Omega} e^{\phi(\omega)} p_\nu(\omega) \Pi(d\omega) \right) \right] : \mathcal{F} \times (M_1 \times M_2) \to \mathbb{R} \quad (2.56)
$$

is continuous on its domain, 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 $\mathcal{F} \times (M_1 \times M_2)$. W.l.o.g. $\phi_*$ can be assumed to satisfy the relation\(^8\)
\begin{equation}
\int_\Omega \exp\{-\phi_*(\omega)\} p_{\mu_*}(\omega) \Pi(d\omega) = \int_\Omega \exp\{\phi_*(\omega)\} p_{\nu_*}(\omega) \Pi(d\omega).
\end{equation}

Denoting the common value of the two quantities in (2.57) by $\varepsilon_*$, the saddle point value
\[\min_{\phi \in \mathcal{F}} \max_{[\mu,\nu] \in M_1 \times M_2} \Phi(\phi, [\mu; \nu])\]
is $\ln(\varepsilon_*)$. Besides this, setting $\phi_*^a(\cdot) = \phi_*(\cdot) - a$, one has
\begin{align}
(a) & \quad \int_\Omega \exp\{-\phi_*^a(\omega)\} p_{\mu}(\omega) \Pi(d\omega) \leq \exp\{a\} \varepsilon_* \ \forall \mu \in M_1, \\
(b) & \quad \int_\Omega \exp\{\phi_*^a(\omega)\} p_{\nu}(\omega) \Pi(d\omega) \leq \exp\{-a\} \varepsilon_* \ \forall \nu \in M_2.
\end{align}

In view of Proposition 2.14 this implies 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 $T_{\phi_*^a}$ based on the detector $\phi_*^a$ can be upper-bounded as follows:
\[
\text{Risk}_1(T_{\phi_*^a}|H_1, H_2) \leq \exp\{a\} \varepsilon_*, \quad \text{Risk}_2(T_{\phi_*^a}|H_1, H_2) \leq \exp\{-a\} \varepsilon_*.
\]

Moreover, $\phi_*, \varepsilon_*$ form an optimal solution to the optimization problem
\begin{equation}
\min_{\phi, \varepsilon} \left\{ \varepsilon : \int_\Omega e^{-\phi(\omega)} p_{\mu}(\omega) \Pi(d\omega) \leq \varepsilon \ \forall \mu \in M_1 \right\}
\end{equation}
(the minimum in (2.59) is taken over all $\varepsilon > 0$ and all $\Pi$-measurable functions $\phi(\cdot)$, not just over $\phi \in \mathcal{F}$).

(ii) The dual problem associated with the saddle point data $\Phi$, $\mathcal{F}$, $M_1 \times M_2$ is
\[
\max_{\mu \in M_1, \nu \in M_2} \left\{ \Phi(\mu, \nu) := \inf_{\phi \in \mathcal{F}} \Phi(\phi; [\mu; \nu]) \right\}.
\]

The objective in this problem is in fact the logarithm of Hellinger affinity of $p_\mu$ and $p_\nu$:
\begin{equation}
\Phi(\mu, \nu) = \ln \left( \int_\Omega \sqrt{p_{\mu}(\omega)p_{\nu}(\omega)} \Pi(d\omega) \right),
\end{equation}
and this objective is concave and continuous on $M_1 \times M_2$.

The $(\mu, \nu)$-components of saddle points of $\Phi$ are exactly the maximizers $(\mu_*, \nu_*)$ of the concave function $\Phi$ on $M_1 \times M_2$. Given such a maximizer $[\mu_*; \nu_*]$ and setting
\begin{equation}
\phi_*(\omega) = \frac{1}{2} \ln(p_{\mu_*}(\omega)/p_{\nu_*}(\omega))
\end{equation}
we get a saddle point $(\phi_*, [\mu_*; \nu_*])$ of $\Phi$ satisfying (2.57).

(iii) Let $[\mu_*; \nu_*]$ be a maximizer of $\Phi$ over $M_1 \times M_2$. Let, further, $\varepsilon \in [0, 1/2]$ be such that there exists a (whatever, perhaps randomized) test for deciding via

\(^8\)Note that $\mathcal{F}$ contains constants, and shifting by a constant the $\phi$-component of a saddle point of $\Phi$ and keeping its $[\mu; \nu]$-component intact, we clearly get another saddle point of $\Phi$.\]
HYPOTHESIS TESTING

observation \( \omega \in \Omega \) on two simple hypotheses

\[ (A) : \omega \sim p(\cdot) := p_\mu(\cdot), \quad (B) : \omega \sim q(\cdot) := p_\nu(\cdot) \]  

(2.62)

with total risk \( \leq 2\epsilon \). Then

\[ \epsilon_* \leq 2\sqrt{\epsilon(1-\epsilon)}. \]

In other words, if the simple hypotheses \((A), (B)\) can be decided, by a whatever test, with total risk \(2\epsilon\), the risks of the simple test with detector \(\phi_*\) given by (2.61) on the composite hypotheses \(H_1, H_2\) do not exceed \(2\sqrt{\epsilon(1-\epsilon)}\).

For proof, see Section 2.11.3.

Remark 2.24. Assume that we are under the premise of Theorem 2.23 and that the simple o.s. in question is nondegenerate (see Section 2.4.2). Then \(\epsilon_* < 1\) if and only if the sets \(M_1\) and \(M_2\) do not intersect.

Indeed, by Theorem 2.23.i, \(\ln(\epsilon_*)\) 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 2.23.ii, the maximum of the function (2.60) 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 (2.60) clearly is \(\leq 0\), we conclude that \(\epsilon_* \leq 1\) 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 \(\mu\) and \(\nu\). Since \(p_\mu(\cdot)\) and \(p_\nu(\cdot)\) are continuous and the support of \(\Pi\) is the entire \(\Omega\), the latter can happen if and only if \(p_\mu = p_\nu\) for our \(\mu, \nu\), or, by nondegeneracy of \(\mathcal{O}\), if and only if \(M_1 \cap M_2 \neq \emptyset\). \(\square\)

2.4.5 Simple observation schemes – Examples of optimal detectors

Theorem 2.23.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 \(M_\chi\) are convex and compact, building optimal, in terms of its risk, detector \(\phi_*\) – that is, solving (in general, semi-infinite and infinite-dimensional) optimization problem (2.59) – reduces to solving a finite-dimensional convex problem. Specifically, an optimal solution \((\phi_*, \epsilon_*)\) can be built as follows:

1. We solve optimization problem

\[ \text{Opt} = \max_{\mu \in M_1, \nu \in M_2} \Phi(\mu, \nu) := \ln \left( \int_\Omega \sqrt{p_\mu(\omega)p_\nu(\omega)} \Pi(d\omega) \right) \]  

(2.63)

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 \([\mu_\ast; \nu_\ast]\) to (2.63) gives rise to an optimal detector \(\phi_\ast\) and its risk \(\varepsilon_\ast\), according to
\[
\phi_\ast(\omega) = \frac{1}{2} \ln \left( \frac{p_{\mu_\ast}(\omega)}{p_{\nu_\ast}(\omega)} \right), \quad \varepsilon_\ast = \exp\{\text{Opt}\}.
\]
The risks of the simple test \(T_{\phi_\ast}\) associated with the above detector and deciding on \(H_1, H_2\), satisfy the bounds
\[
\max \left\{ \text{Risk}_1(\phi_\ast|H_1, H_2), \text{Risk}_2(\phi_\ast|H_1, H_2) \right\} \leq \varepsilon_\ast,
\]
and the test is near-optimal, meaning that whenever the hypotheses \(H_1, H_2\) (and in fact – even two simple hypotheses stating that \(\omega \sim p_{\mu_\ast}\) and \(\omega \sim p_{\nu_\ast}\), respectively) can be decided upon by a test with total risk \(\leq 2\epsilon \leq 1\), \(T_{\phi_\ast}\) exhibits a “comparable” risk:
\[
\varepsilon_\ast \leq 2\sqrt{\epsilon(1-\epsilon)}.
\]
(2.64)

The test \(T_{\phi_\ast}\) is just the maximum likelihood test induced by the probability densities \(p_{\mu_\ast}\) and \(p_{\nu_\ast}\).

Note that after we know that \((\phi_\ast, \varepsilon_\ast)\) form an optimal solution to (2.59), some kind of near-optimality of the test \(T_{\phi_\ast}\) is guaranteed already by Proposition 2.18. By this proposition, whenever in the nature there exists a simple test \(T\) which decides on \(H_1, H_2\) with risks \(\text{Risk}_1, \text{Risk}_2\) bounded by some \(\epsilon \leq 1/2\), the upper bound \(\varepsilon_\ast\) on the risks of \(T_{\phi_\ast}\) can be bounded according to (2.64). Our now near-optimality statement is slightly stronger: first, we allow \(T\) to have the total risk \(\leq 2\epsilon\), which is weaker than to have both risks \(\leq \epsilon\); second, and more important, now \(2\epsilon\) should upper-bound the total risk of \(T\) on a pair of simple hypotheses “embedded” into the hypotheses \(H_1, H_2\); both these modifications extend the family of tests \(T\) to which we compare the test \(T_{\phi_\ast}\), 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 \(O\) is a Gaussian o.s., that is, \(\{p_\mu : \mu \in M\}\) are Gaussian densities with expectations \(\mu \in M = \mathbb{R}^d\) and common positive definite covariance matrix \(\Theta\), and \(\mathcal{F}\) is the family of affine functions on \(\Omega = \mathbb{R}^d\),

- \(M_1, M_2\) can be arbitrary nonempty convex compact subsets of \(\mathbb{R}^d\),
- problem (2.63) becomes the convex optimization problem
  \[
  \text{Opt} = - \min_{\mu \in M_1, \nu \in M_2} \frac{1}{2}(\mu - \nu)^T \Theta^{-1}(\mu - \nu)
  \]
  (2.65)
- the optimal detector \(\phi_\ast\) and the upper bound \(\varepsilon_\ast\) on its risks given by an optimal solution \((\mu_\ast, \nu_\ast)\) to (2.65) are
  \[
  \phi_\ast(\omega) = \frac{1}{2} [\mu_\ast - \nu_\ast]^T \Theta^{-1}[\omega - w], \quad w = \frac{1}{2}[\mu_\ast + \nu_\ast]
  \]
  \[
  \varepsilon_\ast = \exp\left\{-\frac{1}{2}[\mu_\ast - \nu_\ast]^T \Theta^{-1}[\mu_\ast - \nu_\ast]\right\}
  \]
(2.66)

Note that when \(\Theta = I_d\), the test \(T_{\phi_\ast}\) 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 \(\text{Erfc}(\frac{1}{2}\|\mu_\ast - \nu_\ast\|_2)\)) is slightly better than the bound
HYPOTHESIS TESTING

\[ \varepsilon^* = \exp\{-\|\mu^* - \nu^*\|^2/8\} \] given by (2.66) 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 \( \leq \varepsilon \ll 1 \), the results of Example 2.1 and (2.66) say nearly the same: Example 2.1 says that \( \delta \) should be \( \geq 2 \text{ErfcInv}(\varepsilon) \), with ErfcInv defined in (1.26), and (2.66) says that \( \delta \) should be \( \geq 2\sqrt{2 \ln(1/\varepsilon)} \). When \( \varepsilon \to +0 \), the ratio of these two lower bounds on \( \delta \) 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 \( O \) is a Poisson o.s., that is, \( \mathcal{M} = \mathbb{R}_d^{++} \) is the interior of nonnegative orthant in \( \mathbb{R}^d \), and \( p_\mu, \mu \in \mathcal{M} \), is the density

\[ p_\mu(\omega) = \prod_i \left( \frac{\mu_i^{\omega_i}}{\omega_i!} e^{-\mu_i} \right), \quad \omega = (\omega_1, ..., \omega_d) \in \mathbb{Z}_d^+ \]

taken w.r.t. the counting measure \( \Pi \) on \( \Omega = \mathbb{Z}_d^+ \), and \( \mathcal{F} \) is the family of affine functions on \( \Omega \), 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 (2.63) becomes the convex optimization problem

\[ \text{Opt} = - \min_{\mu \in M_1, \nu \in M_2} \frac{1}{2} \sum_{i=1}^d (\sqrt{\mu_i} - \sqrt{\nu_i})^2; \quad (2.67) \]

- the optimal detector \( \phi_* \) and the upper bound \( \varepsilon_* \) on its risks given by an optimal solution \((\mu^*, \nu^*)\) to (2.67) are

\[ \phi_*(\omega) = \frac{1}{2} \sum_{i=1}^d \ln \left( \frac{\mu_i^*}{\nu_i^*} \right) \omega_i + \frac{1}{2} \sum_{i=1}^d [\nu_i^* - \mu_i^*], \quad \varepsilon_* = e^{\text{Opt}} \]

2.4.5.3 Discrete o.s.

When \( O \) is a Discrete o.s., that is, \( \Omega = \{1, ..., d\} \), \( \Pi \) is a counting measure on \( \Omega \), \( \mathcal{M} = \{\mu \in \mathbb{R}_d^d : \mu > 0, \sum_i \mu_i = 1\} \) and

\[ p_\mu(\omega) = \mu_\omega, \quad \omega = 1, ..., d, \mu \in \mathcal{M}, \]

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,

---

\[ \text{It should be mentioned that the results of this section as applied to Discrete observation scheme are a simple particular case – that of finite } \Omega – \text{ of the results of [21, 22, 25] on distinguishing convex sets of probability distributions.} \]
• problem (2.63) is equivalent to the convex program

\[ \varepsilon_* = \max_{\mu \in M_1, \nu \in M_2} \sum_{i=1}^{d} \sqrt{\mu_i \nu_i}; \]  

(2.68)

• the optimal detector \( \phi_* \) given by an optimal solution \( (\mu^*, \nu^*) \) to (2.67) is

\[ \phi_* (\omega) = \frac{1}{2} \ln \left( \frac{\mu^{*2}}{\nu^{*2}} \right), \]  

(2.69)

and the upper bound \( \varepsilon_* \) on the risks of this detector is given by (2.68).

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 \( \Omega^K \) is the direct product of \( K \) copies of \( \Omega \), \( \Pi^K \) is the product of \( K \) copies of \( \Pi \), the densities \( p_{\mu}^{(K)} \) are product densities induced by \( K \) copies of the density \( p_\mu \), \( \mu \in \mathcal{M} \):

\[ p_{\mu}^{(K)} (\omega^K) = \prod_{k=1}^{K} p_\mu (\omega_k), \]

and \( \mathcal{F}^{(K)} \) is comprised of functions

\[ \phi^{(K)} (\omega^K) = \sum_{k=1}^{K} \phi (\omega_k) \]

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 \( \omega_k \) 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, \( H^K_1, \chi = 1, 2 \), state that the components \( \omega_k \) of observation \( \omega^K = (\omega_1, ..., \omega_K) \) are independently of each other drawn from distribution \( p_\mu \) with \( \mu \in M_1 \) (hypothesis \( H^K_1 \)) or \( \mu \in M_2 \) (hypothesis \( H^K_2 \)).

• problem (2.63) is the convex program

\[ \text{Opt}(K) = \max_{\mu \in M_1, \nu \in M_2} \ln \left( \int_{\Omega^K} \sqrt{p_{\mu}^{(K)} (\omega^K) p_{\nu}^{(K)} (\omega^K) \Pi^K (d\Omega)} \right) \]

(\( D^K \))

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 \( (D_K) \) asso-
HYPOTHESIS TESTING

associated with $M_1$, $M_2$, and $\text{Opt}(K) = K\text{Opt}(1)$;
• the optimal detector $\phi_\ast(K)$ given by an optimal solution $(\mu_\ast, \nu_\ast)$ to $(D_1)$ (this solution is optimal for $(D_K)$ as well) is

$$
\phi_\ast(K)(\omega^K) = \sum_{k=1}^{K} \phi_\ast(\omega_k), \quad \phi_\ast(\omega) = \frac{1}{2} \ln \left( \frac{p_{\mu_\ast}(\omega)}{p_{\nu_\ast}(\omega)} \right), \quad (2.70)
$$

and the upper bound $\varepsilon_\ast(K)$ on the risks of the detector $\phi_\ast(K)$ on the pair of families of distributions obeying hypotheses $H^K_1$, resp., $H^K_2$, is

$$
\varepsilon_\ast(K) = e^{\text{Opt}(K)} = e^{K\text{Opt}(1)} = \left[ \varepsilon_\ast(1) \right]^K. \quad (2.71)
$$

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.

**Proposition 2.25.** 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 $(\mu_\ast, \nu_\ast)$ be an optimal solution to the convex optimization problem (cf. Theorem 2.23)

$$
\text{Opt} = \max_{\mu \in M_1, \nu \in M_2} \ln \left( \int_{\Omega} \sqrt{p_\mu(\omega)p_\nu(\omega)}\Pi(\omega) d\omega \right).
$$

Let $\phi_\ast$ and $\phi^K_\ast$ be single- and $K$-observation detectors induced by $(\mu_\ast, \nu_\ast)$ via (2.70).

Let $\epsilon \in (0, 1/2]$, and assume that for some positive integer $K$ in the nature exists a simple test $T^K$ 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

$$
H^K_{\chi} : \mu \in M_\chi, \chi = 1, 2,
$$

with risks $\text{Risk}_1$, $\text{Risk}_2$ not exceeding $\epsilon$. Then setting

$$
K_+ = \left\lfloor \frac{2}{1 - \ln(4(1 - \epsilon))/\ln(1/\epsilon)} K \right\rfloor,
$$

the simple test $T^{(K_+)}_{\phi_\ast}$ utilizing $K_+$ i.i.d. observations decides on $H^K_{1(K_+)}$, $H^K_{2(K_+)}$ with risks $\leq \epsilon$. Note that $K_+$ “is of the order of $K$.” $K_+/K \to 2$ as $\epsilon \to +0$.

**Proof.** Applying item (iii) of Theorem 2.23 to the simple o.s. $[\mathcal{O}]^K$, we see that what above was called $\varepsilon_\ast(K)$ satisfies

$$
\varepsilon_\ast(K) \leq 2\sqrt{\epsilon(1 - \epsilon)}.
$$

By (2.71), we conclude that $\varepsilon_\ast(1) \leq \left( 2\sqrt{\epsilon(1 - \epsilon)} \right)^{1/K}$, whence, by the same (2.71), $\varepsilon_\ast(T) \leq \left( 2\sqrt{\epsilon(1 - \epsilon)} \right)^{T/K}$, $T = 1, 2, ...$. When plugging in this bound $T = K_+$, we get the inequality $\varepsilon_\ast(K_+) \leq \epsilon$. It remains to recall that $\varepsilon_\ast(K_+)$ upper-bounds the risks of the test $T^{(K_+)}_{\phi_\ast}$ when deciding on $H^K_{1(K_+)}$ vs. $H^K_{2(K_+)}$. \(\square\)
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.

\[ \mathcal{O} = (\Omega, \Pi; \{p_{\mu} : \mu \in M\}; \mathcal{F}), \]

(2.72)

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 \( X \subset M \).

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 \( \Omega \) be an observation space, and assume we are given two finite collections of families of probability distributions on \( \Omega \): 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,

\[ R_i : P \in \mathcal{R}_i \text{ (red hypotheses)} \quad \text{and} \quad B_j : P \in \mathcal{B}_j \text{ (blue hypotheses)}. \]

Assume that for every \( i \leq r, j \leq b \) we have at our disposal a simple detector-based test \( T_{ij} \) capable to decide on \( R_i \) vs. \( B_j \). What we want is to assemble these tests into a test \( T \) deciding on the union \( R \) of red hypotheses vs. the union \( B \) of blue ones:

\[ R : P \in \mathcal{R} := \bigcup_{i=1}^{r} \mathcal{R}_i, \quad B : P \in \mathcal{B} := \bigcup_{j=1}^{b} \mathcal{B}_j. \]

Here \( P \), 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. (2.72):

\[ \mathcal{R}_i = \{p_{\mu} : \mu \in M_i\}, \quad \mathcal{B}_j = \{p_{\mu} : \mu \in N_j\}, \]

where \( M_i \) and \( N_j \) are convex compact subsets of \( M \). 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

\[ R : P \in \mathcal{R} = \{p_{\mu} : \mu \in X\}, \quad X = \bigcup_i M_i, \]
\[ B : P \in \mathcal{B} = \{p_{\mu} : \mu \in Y\}, \quad Y = \bigcup_j N_j. \]

While the structure of \( R, B \) is similar to that 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 \( \phi_{ij} \) be the detectors underlying the tests \( T_{ij} \); w.l.o.g., we can assume these detectors balanced (see Section 2.3.2.2) with some
HYPOTHESIS TESTING

89

risks $\epsilon_{ij}$:

\[
\int_\Omega e^{-\phi_{ij}(\omega)} P(d\omega) \leq \epsilon_{ij} \quad \forall P \in \mathcal{R}_i \\
\int_\Omega e^{\phi_{ij}(\omega)} P(d\omega) \leq \epsilon_{ij} \quad \forall P \in \mathcal{B}_j
\]

\[1 \leq i \leq r, 1 \leq j \leq b. \tag{2.73}\]

Let us assemble the detectors $\phi_{ij}$ into a detector for $R$, $B$ as follows:

\[
\phi(\omega) = \max_{1 \leq i \leq r} \min_{1 \leq j \leq b} [\phi_{ij}(\omega) - \alpha_{ij}],
\]

where the shifts $\alpha_{ij}$ are parameters of the construction.

**Proposition 2.26.** The risks of $\phi$ on $R$, $B$ can be bounded as

\[
\forall P \in \mathcal{R} : \quad \int_\Omega e^{-\phi(\omega)} P(d\omega) \leq \max_{1 \leq i \leq r} \left[ \sum_{j=1}^{b} \epsilon_{ij} e^{\alpha_{ij}} \right],
\]

\[
\forall P \in \mathcal{B} : \quad \int_\Omega e^{\phi(\omega)} P(d\omega) \leq \max_{1 \leq j \leq b} \left[ \sum_{i=1}^{r} \epsilon_{ij} e^{-\alpha_{ij}} \right]. \tag{2.75}
\]

Therefore, the risks of $\phi$ on $R$, $B$ are upper-bounded by the quantity

\[
\varepsilon_* = \max_{1 \leq i \leq r} \left[ \sum_{j=1}^{b} \epsilon_{ij} e^{\alpha_{ij}} \right], \max_{1 \leq j \leq b} \left[ \sum_{i=1}^{r} \epsilon_{ij} e^{-\alpha_{ij}} \right]. \tag{2.76}
\]

whence the risks of the based on the detector $\phi$ simple test $T_\phi$ deciding on $R$, $B$ are upper-bounded by $\varepsilon_*$. 

**Proof.** Let $P \in \mathcal{R}$, so that $P \in \mathcal{R}_{i_*}$ for some $i_* \leq r$. Then

\[
\int_\Omega e^{-\phi(\omega)} P(d\omega) = \int_\Omega e^{\min_{1 \leq i \leq r} \max_{1 \leq j \leq b} [\phi_{ij}(\omega) + \alpha_{ij}]} \leq \max_{1 \leq j \leq b} \left[ \sum_{i=1}^{r} \epsilon_{ij} e^{\alpha_{ij}} \right].
\]

Now let $P \in \mathcal{B}$, so that $P \in \mathcal{B}_{j_*}$ for some $j_*$. We have

\[
\int_\Omega e^{\phi(\omega)} P(d\omega) = \int_\Omega e^{\max_{1 \leq i \leq r} \min_{1 \leq j \leq b} [\phi_{ij}(\omega) - \alpha_{ij}]} \leq \max_{1 \leq i \leq r} \left[ \sum_{j=1}^{b} \epsilon_{ij} e^{-\alpha_{ij}} \right].
\]

(2.75) is proved. The remaining claims of the proposition are readily given by (2.75) combined with Proposition 2.14.

**Optimal choice of shift parameters.** The detector and the test considered in Proposition 2.26, same as the resulting risk bound $\varepsilon_*$, depend on the shifts $\alpha_{ij}$.

Let us optimize the risk bound w.r.t. these shifts. To this end, consider the $r \times b$ matrix

\[
E = [\epsilon_{ij}]_{r \times b}.
\]
and the symmetric \((r+b) \times (r+b)\) matrix
\[
\mathcal{E} = \begin{bmatrix} E \end{bmatrix}.
\]
As it is well known, the eigenvalues of the symmetric matrix \(\mathcal{E}\) are comprised of the pairs \((\sigma, -\sigma)\), where \(\sigma\) are the singular values of \(E\), and several zeros; in particular, the leading eigenvalue of \(\mathcal{E}\) is the spectral norm \(\|E\|_2\) (the largest singular value) of matrix \(E\). Further, \(E\) is a matrix with positive entries, so that \(E\) 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 \(\rho = \|E\|_2\), we have
\[
\rho g = Eh \quad \rho h = E^T g \quad (2.77)
\]
Observe that \(\rho > 0\) (evident), whence both \(g\) and \(h\) are nonzero (since otherwise (2.77) 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, (2.77) says that \(g\) and \(h\) are strictly positive vectors. The latter allows to define shifts \(\alpha_{ij}\) according to
\[
\alpha_{ij} = \ln(h_j/g_i). \quad (2.78)
\]
With these shifts, we get
\[
\max_{i \leq r} \left[ \sum_{j=1}^b \epsilon_{ij} e^{\alpha_{ij}} \right] = \max_{i \leq r} \sum_{j=1}^b \epsilon_{ij} h_j/g_i = \max_{i \leq r} (Eh)_i/g_i = \max_{i \leq r} \rho = \rho
\]
(we have used the first relation in (2.77)), and
\[
\max_{j \leq b} \left[ \sum_{i=1}^r \epsilon_{ij} e^{-\alpha_{ij}} \right] = \max_{j \leq b} \sum_{i=1}^r \epsilon_{ij} g_i/h_j = \max_{j \leq b} |E^T g|_j/h_j = \max_{j \leq b} \rho = \rho
\]
(we have used the second relation in (2.77)). The bottom line is as follows:

**Proposition 2.27.** In the situation and the notation of Section 2.5.1.1, the risks of the detector (2.74) with shifts (2.77), (2.78) on the families \(\mathcal{R}, \mathcal{B}\) do not exceed the quantity
\[
\|E := [\epsilon_{ij}]_{i \leq r, j \leq b}\|_2.
\]
As a result, the risks of the simple test \(T_\phi\) deciding on the hypotheses \(\mathcal{R}, \mathcal{B}\), does not exceed \(\|E\|_2\) 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 2.28.** Let \(\mathcal{E} = [\epsilon_{ij}]\) be nonzero entrywise nonnegative \(n \times n\) symmetric matrix. Then the optimal value in the optimization problem
\[
\text{Opt} = \min_{\alpha_{ij}} \left\{ \max_{i \leq n} \sum_{j=1}^n \epsilon_{ij} e^{\alpha_{ij}} : \alpha_{ij} = -\alpha_{ji} \right\}
\]
(∗)
is equal to $\|E\|_{2,2}$. When the Perron-Frobenius eigenvector $f$ of $E$ can be selected positive, the problem is solvable, and an optimal solution is given by

$$\alpha_{ij} = \ln(f_j/f_i), \quad 1 \leq i, j \leq n.$$  \hfill (2.79)

**Proof.** Let us prove that $\text{Opt} \leq \rho := \|E\|_{2,2}$. Given $\epsilon > 0$, we clearly can find an entrywise nonnegative symmetric matrix $E'$ with entries $e'_{ij}$ in-between $e_{ij}$ and $e_{ij} + \epsilon$ such that the Perron-Frobenius eigenvector $f$ of $E'$ can be selected positive (it suffices, e.g., to set $e'_{ij} = e_{ij} + \epsilon$). Selecting $\alpha_{ij}$ according to (2.79), we get a feasible solution to $(\ast)$ such that

$$\forall i : \sum_j e_{ij} e^{\alpha_{ij}} \leq \sum_j e'_{ij} f_j/f_i = \|E'\|_{2,2},$$

implying that $\text{Opt} \leq \|E'\|_{2,2}$. Passing to limit as $\epsilon \to +0$, we get $\text{Opt} \leq \|E\|_{2,2}$. As a byproduct of our reasoning, if $E$ admits a positive Perron-Frobenius eigenvector $f$, then (2.79) yields a feasible solution to $(\ast)$ with the value of the objective equal to $\|E\|_{2,2}$.

It remain to prove that $\text{Opt} \geq \|E\|_{2,2}$. Assume that this is not the case, so that $(\ast)$ admits a feasible solution $\hat{\alpha}_{ij}$ such that

$$\hat{\rho} := \max_i \sum_j e_{ij} e^{\hat{\alpha}_{ij}} < \rho := \|E\|_{2,2}.$$

By a arbitrarily small perturbation of $E$, we can make this matrix symmetric and entrywise positive, and still satisfying the above strict inequality; to save notation, assume that already the original $E$ is entrywise positive. Let $f$ be a positive Perron-Frobenius eigenvector of $E$, and let, as above, $\alpha_{ij} = \ln(f_j/f_i)$, so that

$$\sum_j e_{ij} e^{\alpha_{ij}} = \sum_j e_{ij} f_j/f_i = \rho \forall i.$$

Setting $\delta_{ij} = \hat{\alpha}_{ij} - \alpha_{ij}$, we conclude that the convex functions

$$\theta_i(t) = \sum_j e_{ij} e^{\alpha_{ij} + \delta_{ij}}$$

all are equal to $\rho$ as $t = 0$, and all are $\leq \hat{\rho} < \rho$ 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

$$\theta_i'(0) = \sum_j e_{ij} e^{\alpha_{ij}} \delta_{ij} = \sum_j e_{ij} (f_j/f_i) \delta_{ij} < 0 \forall i.$$

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. \hfill $\Box$
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). 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 $\Omega$ be an observation space, and let a collection $P_1, \ldots, P_L$ of families of probability distributions on $\Omega$ be given. As always, families $P_\ell$ give rise to hypotheses $H_\ell: P \in P_\ell$ on the distribution $P$ of observation $\omega \in \Omega$. Assume also that we are given a closeness relation $C$ on $\{1, \ldots, L\}$. Recall that, formally, a closeness relation is some set of pairs of indices $(\ell, \ell') \in \{1, \ldots, L\}$; we interpret the inclusion $(\ell, \ell') \in C$ as the fact that hypothesis $H_\ell$ “is close” to hypothesis $H_{\ell'}$. When $(\ell, \ell') \in C$, we say that $\ell'$ is close (or $C$-close) to $\ell$. We always assume that

- $C$ contains the diagonal: $(\ell, \ell) \in C$ for every $\ell \leq L$ (“each hypothesis is close to itself”), and
- $C$ is symmetric: whenever $(\ell, \ell') \in C$, we have also $(\ell', \ell) \in C$ (“if $\ell$-th hypothesis is close to $\ell'$-th one, then $\ell'$-th hypothesis is close to $\ell$-th one”).

Recall that a test $T$ deciding on the hypotheses $H_1, \ldots, H_L$ via observation $\omega \in \Omega$ is a procedure which, given on input $\omega \in \Omega$, builds some set $T(\omega) \subset \{1, \ldots, L\}$, accepts all hypotheses $H_\ell$ with $\ell \in T(\omega)$, and rejects all other hypotheses.

**Risks of an “up to closeness” test.** The notion of $C$-risk of a test was introduced in Section 2.2.4.2, we reproduce it here for reader’s convenience. Given closeness $C$ and a test $T$, we define the $C$-risk

$$\text{Risk}^C(T|H_1, \ldots, H_L)$$

of $T$ as the smallest $\epsilon \geq 0$ such that

Whenever an observation $\omega$ is drawn from a distribution $P \in \bigcup_\ell P_\ell$, and $\ell_*$ is such that $P \in P_{\ell_*}$ (i.e., hypothesis $H_{\ell_*}$ is true), the $P$-probability of the event “$\ell_* \notin T(\omega)$ (“true hypothesis $H_{\ell_*}$ is not accepted”) or there exists $\ell'$ not close to $\ell$ such that $H_{\ell'}$ is accepted” is at most $\epsilon$.

Equivalently:

$$\text{Risk}^C(T|H_1, \ldots, H_L) \leq \epsilon$$

if and only if the following takes place:

---

10Strictly 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. However, for now, the only situation where we indeed have at our disposal good pairwise detectors for red and blue components is that in which these components are convex hypotheses in a good o.s.
Whenever an observation $\omega$ is drawn from a distribution $P \in \bigcup_{\ell} \mathcal{P}_\ell$, and $\ell_* \in \mathcal{T}(\omega)$ ("the true hypothesis $H_{\ell_*}$ is true") and $\ell' \in \mathcal{T}(\omega)$ implies that $(\ell, \ell') \in \mathcal{C}$ ("all accepted hypotheses are $\mathcal{C}$-close to the true hypothesis $H_{\ell_*}$") is at least $1 - \epsilon$.

For example, consider 9 polygons presented on Figure 2.4 and associate with them 9 hypotheses on 2D "signal plus noise" observation $\omega \sim \mathcal{N}(\mu, I_2)$, with $\ell$-th hypothesis stating that $x$ belongs to $\ell$-th polygon. We define closeness $\mathcal{C}$ on the collection of hypotheses presented on Figure 2.4 as "two hypotheses are close if and only if the corresponding polygons intersect," like A and B, or A and E. Now the fact that a test $\mathcal{T}$ has $\mathcal{C}$-risk $\leq 0.01$ would imply, in particular, that if the probability distribution $P$ underlying the observed $\omega$ obeys hypothesis A (i.e., the mean of $P$ belongs to the polygon A), then with $P$-probability at least 0.99 the list of accepted hypotheses includes hypothesis A, and the only other hypotheses in this list are among hypotheses B, D and E.

2.5.2.2 “Building blocks” and construction

The construction we are about to present is, essentially, that used in Section 2.2.4.3 as applied to detector-generated tests. This being said, the presentation to follow is self-contained.

The building blocks of our construction are pairwise detectors $\phi_{\ell\ell'}(\omega)$, $1 \leq \ell < \ell' \leq L$, for pairs $\mathcal{P}_\ell, \mathcal{P}_{\ell'}$ along with (upper bounds on) the risks $\epsilon_{\ell\ell'}$ of these detectors:

$$\forall (P \in \mathcal{P}_\ell) : \int_\Omega e^{-\phi_{\ell\ell'}(\omega)} P(d\omega) \leq \epsilon_{\ell\ell'}$$

$$\forall (P \in \mathcal{P}_{\ell'}) : \int_\Omega e^{\phi_{\ell\ell'}(\omega)} P(d\omega) \leq \epsilon_{\ell'\ell'}$$

Setting

$$\phi_{\ell'\ell}(\omega) = -\phi_{\ell\ell'}(\omega), \epsilon_{\ell'\ell} = \epsilon_{\ell\ell'}, 1 \leq \ell < \ell' \leq L,$$

we get what we refer to as balanced system of detectors $\phi_{\ell\ell'}$ and risks $\epsilon_{\ell\ell'}$, $1 \leq \ell < \ell' \leq L$.
\(\ell, \ell' \leq L\), for the collection \(P_1, \ldots, P_L\), meaning that
\[
\phi_{\ell\ell'}(\omega) + \phi_{\ell'\ell}(\omega) \equiv 0, \quad \epsilon_{\ell\ell'} = \epsilon_{\ell'\ell}, \quad 1 \leq \ell, \ell' \leq L,
\]
(2.80)
Given closeness \(\mathcal{C}\), we associate with it the symmetric \(L \times L\) matrix \(C\) given by
\[
C_{\ell\ell'} = \begin{cases} 0, & (\ell, \ell') \in \mathcal{C} \\ 1, & (\ell, \ell') \not\in \mathcal{C} \end{cases}
\]
(2.81)

**Test** \(T_C\). Let a collection of shifts \(\alpha_{\ell\ell'} \in \mathbb{R}\) satisfying the relation
\[
\alpha_{\ell\ell'} = -\alpha_{\ell'\ell}, \quad 1 \leq \ell, \ell' \leq L
\]
(2.82)
be given. The detectors \(\phi_{\ell\ell'}\) and the shifts \(\alpha_{\ell\ell'}\) specify a test \(T_C\) deciding on hypotheses \(H_1, \ldots, H_L\). Precisely, given an observation \(\omega\), the test \(T_C\) accepts exactly those hypotheses \(H_\ell\) for which \(\phi_{\ell\ell'}(\omega) - \alpha_{\ell\ell'} > 0\) whenever \(\ell'\) is not \(\mathcal{C}\)-close to \(\ell\):
\[
T_C(\omega) = \{ \ell : \phi_{\ell\ell'}(\omega) - \alpha_{\ell\ell'} > 0 \forall (\ell' : (\ell, \ell') \not\in \mathcal{C}) \}.
\]

**Proposition 2.29.** (i) The \(\mathcal{C}\)-risk of the just defined test \(T_C\) is upper-bounded by the quantity
\[
\varepsilon[\alpha] = \max_{\ell \leq L} \sum_{\ell' = 1}^{L} \epsilon_{\ell\ell'} C_{\ell\ell'} e^{\alpha_{\ell\ell'}}
\]
with \(C\) given by (2.81).
(ii) The infimum, over shifts \(\alpha\) satisfying (2.82), of the risk bound \(\varepsilon[\alpha]\) is the quantity
\[
\varepsilon_* = \|E\|_{2,2},
\]
where the \(L \times L\) symmetric entrywise nonnegative matrix \(E\) is given by
\[
E = [e_{\ell\ell'} := \epsilon_{\ell\ell'} C_{\ell\ell'}]_{\ell, \ell' \leq L}.
\]
Assuming \(E\) admits a strictly positive Perron-Frobenius vector \(f\), an optimal choice of the shifts is
\[
\alpha_{\ell\ell'} = \ln(f_{\ell'})/f_{\ell}, \quad 1 \leq \ell, \ell' \leq L,
\]
resulting in \(\varepsilon[\alpha] = \varepsilon_* = \|E\|_{2,2}\).

**Proof.** (i): Setting
\[
\tilde{\phi}_{\ell\ell'}(\omega) = \phi_{\ell\ell'}(\omega) - \alpha_{\ell\ell'}, \quad \tilde{\epsilon}_{\ell\ell'} = \epsilon_{\ell\ell'} e^{\alpha_{\ell\ell'}},
\]
(2.80) and (2.82) imply that
\[
(a) \quad \tilde{\phi}_{\ell\ell'}(\omega) + \tilde{\phi}_{\ell'\ell}(\omega) \equiv 0, \quad 1 \leq \ell, \ell' \leq L \\
(b) \quad \forall P \in P_\ell: \int e^{-\tilde{\phi}_{\ell\ell'}(\omega)} P(d\omega) \leq \tilde{\epsilon}_{\ell\ell'}, \quad 1 \leq \ell, \ell' \leq L.
\]
(2.83)
Now let \(\ell_*\) be such that the distribution \(P\) of observation \(\omega\) belongs to \(P_{\ell_*}\). Then
By Proposition 2.16 and (2.80), we arrive at the following analog of (2.80):

\[ E_* = \{ \omega : \exists \ell' : (\ell, \ell') \notin C \& \bar{\phi}_{\ell, \ell'}(\omega) \leq 0 \} \]

is upper-bounded by

\[ \sum_{\ell' : (\ell, \ell') \notin C} \varepsilon_{\ell, \ell'} = \sum_{\ell' = 1}^{L} C_{\ell, \ell'} \varepsilon_{\ell, \ell'} e^{\alpha_{\ell, \ell'}} \leq \varepsilon[\alpha]. \]

Assume that \( E_* \) does not take place (as we have seen, this indeed is so with \( P \)-probability \( \geq 1 - \varepsilon[\alpha] \)). Then \( \bar{\phi}_{\ell, \ell'}(\omega) > 0 \) for all \( \ell' \) such that \( (\ell, \ell') \notin C \), implying, first, that \( H_{\ell} \) is accepted by our test. Second, \( \bar{\phi}_{\ell, \ell'}(\omega) = -\bar{\phi}_{\ell, \ell'}(\omega) < 0 \) whenever \( (\ell, \ell') \notin C \), or, due to the symmetry of closeness, whenever \( (\ell', \ell) \notin C \), implying that the test \( \mathcal{T}_C \) rejects the hypothesis \( H_{\ell'} \) when \( \ell' \) is not \( C \)-close to \( \ell_* \). Thus, the \( P \)-probability of the event “\( H_{\ell} \) is accepted, and all accepted hypotheses are \( C \)-close to \( H_{\ell'} \)” is at least \( 1 - \varepsilon[\alpha] \). We conclude that the \( C \)-risk \( \text{Risk}(\mathcal{T}_C|H_1, ..., H_L) \) of the test \( \mathcal{T}_C \) is at most \( \varepsilon[\alpha] \). (i) is proved. (ii) is readily given by Proposition 2.28.

\[ \square \]

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 \( \varepsilon_{\ell, \ell'} \), \( 1 \leq \ell, \ell' \leq L \), for the collection \( \mathcal{P}_1, ..., \mathcal{P}_L \) (see (2.80)) and a positive integer \( K \), we can

- pass from detectors \( \phi_{\ell, \ell'} \) and risks \( \varepsilon_{\ell, \ell'} \) to the entities
  \[ \phi_{\ell, \ell'}^{(K)}(\omega^K = (\omega_1, ..., \omega_K)) = \sum_{k=1}^{K} \phi_{\ell, \ell'}(\omega_k), \quad \varepsilon_{\ell, \ell'}^{(K)} = \varepsilon_{\ell, \ell'}, \quad 1 \leq \ell, \ell' \leq L; \]

- 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 \( H_{\ell}^{K} = H_{\ell}^{0,K} \) stating that the distribution \( P^K \) of \( K \)-repeated observation \( \omega^K = (\omega_1, ..., \omega_K) \), \( \omega_k \in \Omega \), belongs to the family \( \mathcal{P}_{\ell}^{0,K} = \bigotimes_{k=1}^{K} \mathcal{P}_\ell \), associated with \( \mathcal{P}_\ell \), see Section 2.1.3.3.

By Proposition 2.16 and (2.80), we arrive at the following analog of (2.80):

\[ \phi_{\ell, \ell'}^{(K)}(\omega^K) + \phi_{\ell', \ell}^{(K)}(\omega^K) \equiv 0, \quad \varepsilon_{\ell, \ell'}^{(K)} = \varepsilon_{\ell', \ell}^{(K)}, \quad 1 \leq \ell, \ell' \leq L \]
\[ \forall P^K \in \mathcal{P}_{\ell}^{(K)} : \int_{\Omega^K} e^{-\phi_{\ell, \ell'}^{(K)}(\omega^K)} P^K(d\omega^K) \leq \varepsilon_{\ell, \ell'}^{(K)}, \quad 1 \leq \ell, \ell' \leq L. \]

Given shifts \( \alpha_{\ell, \ell'} \) satisfying (2.82) and applying the construction from Section 2.5.2.2 to these shifts and our newly constructed detectors and risks, we arrive at the test \( \mathcal{T}_C^K \) deciding on hypotheses \( H_{\ell}^{K_1}, ..., H_{\ell}^{K_L} \) via \( K \)-repeated observation \( \omega^K \). Specifically, given an observation \( \omega^K \), the test \( \mathcal{T}_C^K \) accepts exactly those hypotheses \( H_{\ell}^{K} \) for which \( \phi_{\ell, \ell'}^{(K)}(\omega^K) - \alpha_{\ell, \ell'} > 0 \) whenever \( \ell' \) is not \( C \)-close to \( \ell \):

\[ \mathcal{T}_C^K(\omega^K) = \{ \ell : \phi_{\ell, \ell'}^{(K)}(\omega^K) - \alpha_{\ell, \ell'} > 0 \ \forall (\ell' : (\ell, \ell') \notin C) \} , \]
Invoking Proposition 2.29, we arrive at

**Proposition 2.30.** (i) The C-risk of the just defined test $T_c^K$ is upper-bounded by the quantity

$$
\varepsilon[\alpha, K] = \max_{\ell \leq L} \sum_{\ell' = 1}^L \epsilon_{\ell\ell'}^K C_{\ell\ell'} e^{\alpha \epsilon_{\ell\ell'}}.
$$

(ii) The infimum, over shifts $\alpha$ satisfying (2.82), of the risk bound $\varepsilon[\alpha, K]$ is the quantity

$$
\varepsilon_*(K) = \|\mathcal{E}^{(K)}\|_{2,2},
$$

where the $L \times L$ symmetric entrywise nonnegative matrix $\mathcal{E}^{(K)}$ is given by

$$
\mathcal{E}^{(K)} = \left[\epsilon_{\ell\ell'}^K := e^{\alpha \epsilon_{\ell\ell'}} C_{\ell\ell'}\right]_{\ell,\ell' \leq L}.
$$

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 $\varepsilon[\alpha, K] = \varepsilon_*(K) = \|\mathcal{E}^{(K)}\|_{2,2}.$

### 2.5.2.4 Consistency and near-optimality

Observe that when closeness $C$ is such that $\epsilon_{\ell\ell'} < 1$ whenever $\ell$, $\ell'$ are not $C$-close to each other, the entries on the matrix $\mathcal{E}^{(K)}$ go to 0 as $K \to \infty$ exponentially fast, whence the $C$-risk of test $T_c^K$ also goes to 0 as $K \to \infty$, meaning that test $T_c^K$ is consistent. When, in addition, $\mathcal{P}_\ell$ correspond to convex hypotheses in a simple o.s., the test $T_c^K$ possesses the property of near-optimality similar to that stated in Proposition 2.25.

**Proposition 2.31.** 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 $\mathcal{P}_\ell = \{p_\mu : \mu \in \mathcal{N}_\ell\}$, where $\mathcal{N}_\ell$, $1 \leq \ell \leq L$, 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 2.23 to the pairs $\mathcal{N}_\ell$, $\mathcal{N}_{\ell'}$, so that for $1 \leq \ell < \ell' \leq L$ one has

$$
\phi_{\ell\ell'}(\omega) = \frac{1}{2} \ln(p_{\mu_{\ell\ell'}}(\omega)/p_{\mu_{\ell'\ell'}}(\omega)), \quad \epsilon_{\ell\ell'} = \exp\{\text{Opt}_{\ell\ell'}\}
$$

where

$$
\text{Opt}_{\ell\ell'} = \min_{\mu \in \mathcal{N}_\ell, \nu \in \mathcal{N}_{\ell'}} \ln\left(\int_{\Omega} \frac{p_\mu(\omega)}{p_\nu(\omega)} d\Pi(\omega)\right),
$$

and $(\mu_{\ell\ell'}, \nu_{\ell\ell'})$ form an optimal solution to the optimization problem in the right hand side.

Assume that for some positive integer $K_*$ in the nature there exists a test $T_0^{K_*}$, which decides with $C$-risk $\varepsilon \in (0, 1/2)$, via stationary $K_*$-repeated observation $\omega^{K_*}$, on the hypotheses $H_\ell^{(K_*)}$, stating that the components in $\omega^{K_*}$ are drawn, independently of each other, from a distribution $P \in \mathcal{P}_\ell$, $\ell = 1, ..., L$, and let

$$
K = 2 \left[\frac{1 + \ln(L - 1)/\ln(1/\varepsilon)}{1 - \ln(4(1 - \varepsilon))/\ln(1/\varepsilon)}\right] K_*.
$$

(2.84)
Then the test \( T^K \) 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 \( H^K_\ell \), see Section 2.5.2.3, via quasi-stationary \( K \)-repeated observations \( \omega^K \), with \( C \)-risk \( \leq \epsilon \).

Note that \( K/K^* \to 2 \) as \( \epsilon \to +0 \).

Proof. Let

\[
\bar{\epsilon} = \max_{\epsilon, \ell, \ell'} \{ \epsilon_{\ell\ell'} : \ell < \ell', \ell, \ell' \text{ are not } C \text{-close to each other} \}.
\]

Denoting by \( (\ell_*, \ell'_*) \) the corresponding maximizer, note that \( T^{K_*} \) induces a simple test \( T \) able to decide via stationary \( K_* \)-repeated observations \( \omega^{K_*} \) on the pair of hypotheses \( H^{(K_*)}_\ell, H^{(K_*)}_{\ell'} \) with risks \( \leq \epsilon \) (it suffices to make \( T \) to accept the first of the hypotheses in the pair and reject the second one whenever \( T^{K_*} \) on the same observation accepts \( H^{(K_*)}_\ell \), otherwise \( T \) rejects the first hypothesis in the pair and accepts the second one). This observation, by the same argument as in the proof of Proposition 2.25, implies that \( \bar{\epsilon}^{K_*} \leq 2 \sqrt{\epsilon (1 - \epsilon)} < 1 \), whence all entries in the matrix \( E^{(K)} \) do not exceed \( \bar{\epsilon}_{K/K^*} \), implying by Proposition 2.29 that the \( C \)-risk of the test \( T^K \) does not exceed

\[
\epsilon(K) := (L - 1)[2 \sqrt{\epsilon (1 - \epsilon)}]^{K/K^*}.
\]

It remains to note that for \( K \) given by (2.84) one has \( \epsilon(K) \leq \epsilon \).

Remark 2.32. Note that tests \( T_C \) and \( T^K_C \) 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 \( C \)-closeness always can be modified to ensure that a hypothesis always is accepted. To this end, it suffices, for instance, that the modified test accepts exactly those hypotheses, if any, which are accepted by our original test, and accepts, say, hypothesis \# 1 when the original test accepts no hypotheses. It is immediate to see that the \( C \)-risk of the modified test cannot be larger than the risk 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 \( \mathbb{R}^n \),
- a convex compact set \( X \subset \mathbb{R}^n \) along with a collection of \( M \) points \( x_i \in \mathbb{R}^n \), \( 1 \leq i \leq M \) and a positive \( D \) such that the \( \| \cdot \| \)-diameter of the set

\[ A \text{ seminorm on } \mathbb{R}^n \text{ is defined by exactly the same requirements as a norm, except that now we allow zero seminorms for some nonzero vectors. Thus, a seminorm on } \mathbb{R}^n \text{ is a nonnegative function } \| \cdot \| \text{ which is even and homogeneous: } \| \lambda x \| = |\lambda| \| x \| \text{ and satisfies the triangle inequality } \| x + y \| \leq \| x \| + \| y \| \text{. A universal example is } \| x \| = \| B x \|_\infty, \text{ where } \| \cdot \|_\infty \text{ is a norm on some } \mathbb{R}^m \text{ and } B \text{ is an } m \times n \text{ matrix; whenever this matrix has a nontrivial kernel, } \| \cdot \| \text{ is a seminorm rather than a norm.} \]
\[ X^+ = X \cup \{ x_i : 1 \leq i \leq M \} \] is at most \( D \):
\[ \| x - x' \| \leq D \quad \forall (x, x' \in X^+), \]

- an affine mapping \( x \mapsto A(x) \) from \( \mathbb{R}^n \) into the embedding space of \( M \) such that \( A(x) \in M \) for all \( x \in X \),
- a tolerance \( \epsilon \in (0, 1) \).

We observe a \( K \)-element sample \( \omega^K = (\omega_1, ..., \omega_K) \) of independent across \( k \) observations
\[ \omega_k \sim p_{A(x_k)}, 1 \leq k \leq K, \quad (2.85) \]
where \( x_s \in \mathbb{R}^n \) is unknown signal known to belong to \( X \). Our “ideal goal” is to use \( \omega^K \) in order to identify, with probability \( \geq 1 - \epsilon \), the \( \| \cdot \| \)-closest to \( x^* \) point among the points \( x_1, ..., x_M \).

The just outlined goal may be 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 the design parameters \( \alpha \geq 1 \) and \( \beta \geq 0 \), 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
\[ \forall (x_s \in X) : \text{Prob}\{|\| x_s - x_{i_{\alpha, \beta}(\omega^K)} \| > \alpha \rho(x_s) + \beta \| \geq 1 - \epsilon. \quad (2.86) \]

Note that when passing from the “ideal” to the relaxed goal, the simplification is twofold: first, instead of the precise distance \( \min_{1 \leq i \leq M} \| x - x_i \| \) from \( x^* \) to \( \{ x_1, ..., x_M \} \) we look at the best upper bound \( \rho(x_s) \) on this distance from the grid \( \Gamma \); second, we allow factor \( \alpha \) and additive term \( \beta \) in mimicking the (discretized) distance \( \rho(x_s) \) by \( \| x_s - x_{i_{\alpha, \beta}(\omega^K)} \| \).

The problem we have posed is quite popular in Statistics and originates from the estimate aggregation problem \([182, 225, 99]\) as follows: let \( x_i \) be 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 \( \| \cdot \|_{\circ} \), approximation of \( Bx^* \), utilizing additional observations \( \omega^K \) of the signal. To bring this problem into our framework, it suffices to specify the seminorm as \( \| x \| = \| Bx \|_{\circ} \). We shall see in the mean time that in the context of this problem, the “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 \( \alpha \) and additive term \( \beta \), they indeed could be “expensive in terms of applications,” but the “nearly ideal” goal of making \( \alpha \) close to 1 and \( \beta \) close to 0 may be unattainable.
2.5.3.1 The construction

Let us associate with $i \leq M$ and $j$, $0 \leq j \leq N$, the hypothesis $H_{ij}$ stating that the independent across $k$ observations $\omega_k$, see (2.85), 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 $i \in \{1, \ldots, M\}$, $j \in \{0, 1, \ldots, N\}$, and $X_{ij} \neq \emptyset$. Further, we define closeness $C_{\alpha, \beta}$ on the set of hypotheses $H_{ij}$, $(i, j) \in \mathcal{J}$, as follows:

$$(ij, i'j') \in C_{\alpha, \beta} \text{ if and only if }$$

$$\|x_i - x_{i'}\| \leq \alpha(r_j + r_{j'}) + \beta, \quad \alpha = \frac{\alpha - 1}{2}. \quad (2.87)$$

(here and in what follows, $k\ell$ denotes the ordered pair $(k, \ell)$).

Applying Theorem 2.23, we can build, in a computation-friendly fashion, the system $\phi_{ij, i'j'}(\omega), i, j, i', j' \in \mathcal{J}$, of optimal balanced detectors for the hypotheses $H_{ij}$ along with the risks of these detectors, so that

$$\phi_{ij, i'j'}(\omega) = -\phi_{ij', i'j}(\omega)$$

$$\int_{\Omega} e^{-\phi_{ij, i'j'}(\omega)} p_{A(x)(\omega)\Pi(d\omega)} \leq \epsilon_{ij, i'j'}, \quad \forall (ij, i'j') \in \mathcal{J}, x \in X_{ij}.$$ 

Let us say that a pair $(\alpha, \beta)$ is admissible, if $\alpha \geq 1, \beta \geq 0$ and

$$\forall (i, j) \in \mathcal{J}, (i', j') \in \mathcal{J}, (ij, i'j') \not\in C_{\alpha, \beta} : A(X_{ij}) \cap A(X_{i'j'}) = \emptyset.$$ 

Note that checking admissibility of a given pair $(\alpha, \beta)$ is a computationally tractable task.

Given an admissible pair $(\alpha, \beta)$, we associate with it a positive integer $K = K(\alpha, \beta)$ and inference $w^K \mapsto i_{\alpha, \beta}(w^K)$ as follows:

1. $K = K(\alpha, \beta)$ is the smallest integer such that the detector-based test $T^K_{\epsilon_{ij, i'j}}$ yielded by the machinery of Section 2.5.2.3 decides on the hypotheses $H_{ij}$, $ij \in \mathcal{J}$, with $C_{\alpha, \beta}$-risk not exceeding $\epsilon$. Note that by admissibility, $\epsilon_{ij, i'j'} < 1$ whenever $(ij, i'j') \not\in C_{\alpha, \beta}$, so that $K(\alpha, \beta)$ is well defined.

2. Given observation $w^K$, $K = K(\alpha, \beta)$, we define $i_{\alpha, \beta}(w^K)$ as follows:
   a) We apply to $w^K$ the test $T^K_{\epsilon_{ij, i'j}}$. If the test accepts no hypothesis (case A), $i_{\alpha, \beta}(w^K)$ is undefined. The observations $w^K$ resulting in case A comprise some set, which we denote by $B$; given $w^K$, we can recognize whether or not $w^K \in B$.
   b) When $w^K \not\in B$, the test $T^K_{\epsilon_{ij, i'j}}$ accepts some of the hypotheses $H_{ij}$, let the set of their indices $ij$ be $J(\omega^K)$; we select from the pairs $ij \in J(\omega^K)$ one with the largest $j$, and set $i_{\alpha, \beta}(w^K)$ to be equal to the first component, and $j_{\alpha, \beta}(w^K)$ to be equal to the second component of the selected pair.

We have the following

**Proposition 2.33.** Assuming $(\alpha, \beta)$ admissible, for the just defined inference $w^K \mapsto i_{\alpha, \beta}(w^K)$ and for every $x_* \in X$, denoting by $P^K_{x_*}$ the distribution of stationary $K$-
repeated observation $\omega^K$ stemming from $x_*$ one has
\[ \|x_* - x_{i_\alpha,\beta}(\omega^K)\| \leq \alpha\rho(x_*) + \beta. \] (2.88)
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 $r_0 = D \geq \|x_* - x_1\|$, and that
\[ r_{j_*} = \rho(x_*). \]
We specify $i_* = i_*(x_*) \leq M$ in such a way that
\[ \|x_* - x_{i_*}\| \leq r_{j_*}. \] (2.89)
Note that $i_*$ is well defined and that observations (2.85) stemming from $x_*$ obey the hypothesis $H_{i_*,j_*}$.

Let $\mathcal{E}$ be the set of those $\omega^K$ for which the predicate $P^K_{\omega^K}$: As applied to observation $\omega^K$, the test $T^K_{\omega^K}$ accepts $H_{i_*,j_*}$, and all hypotheses accepted by the test are $C_{\alpha,\beta}$-close to $H_{i_*,j_*}$ holds true. Taking into account that the $C_{\alpha,\beta}$-risk of $T^K_{\omega^K}$ does not exceed $\epsilon$ 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 $\omega^K$ satisfy $\omega^K \in \mathcal{E}$. (2.90)

Then

1. The test $T^K_{\omega^K}$ accepts the hypothesis $H_{i_*,j_*}$, that is, $\omega^K \notin \mathcal{B}$. By construction of $i_\alpha,\beta(\omega^K)j_{\alpha,\beta}(\omega^K)$ (see the rule 2b above) and due to the fact that $T^K_{\omega^K}$ 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 $C_{\alpha,\beta}$-close to $H_{i_*,j_*}$, so that
\[ \|x_{i_*} - x_{i_\alpha,\beta}(\omega^K)\| \leq \alpha(r_{j_*} + r_{j_{\alpha,\beta}(\omega^K)}) + \beta \leq 2\alpha r_{j_*} + \beta = 2\alpha\rho(x_*) + \beta, \]
where the concluding inequality is due to the fact that, as we have already seen, $j_{\alpha,\beta}(\omega^K) \geq j_*$ when (2.90) takes place.

Invoking (2.89), we conclude that with $P^K_{x_*}$-probability at least $1 - \epsilon$ it holds
\[ \|x_* - x_{i_\alpha,\beta}(\omega^K)\| \leq (2\alpha + 1)\rho(x_*) + \beta = \alpha\rho(x_*) + \beta. \]

2.5.3.2 A modification

From the computational viewpoint, an obvious shortcoming of the construction presented in the previous section is the necessity to operate with $M(N+1)$ hypotheses, which might 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
HYPOTHESIS TESTING

is to replace simultaneous processing of all hypotheses \( H_{ij}, \ i \in J \), with processing them in stages \( j = 0, 1, \ldots \), with stage \( j \) operating only with the hypotheses \( H_{ij}, \ i = 1, \ldots, M \).

The implementation of this idea is as follows. In the situation of Section 2.5.3, given the same entities \( \Gamma, (\alpha, \beta), H_{ij}, X_{ij}, \ i, j \in J \), as in the beginning of Section 2.5.3.1 and specifying closeness \( C_{\alpha,\beta} \) according to (2.87), we now act as follows.

Preprocessing. For \( j = 0, 1, \ldots, N \)
1. we identify the set \( I_j = \{ i \leq M : X_{ij} \neq \emptyset \} \) and stop if this set is empty. If this set is nonempty,
2. we specify the closeness \( C^j_{\alpha,\beta} \) on the set of hypotheses \( H_{ij}, \ i \in I_j \) as a “slice” of the closeness \( C_{\alpha,\beta} \):

   \[
   H_{ij} \text{ and } H_{ij'} \text{ (equivalently, } i \text{ and } i' \text{) are } C^j_{\alpha,\beta}-\text{close to each other if } (ij, i'j) \text{ are } C_{\alpha,\beta}-\text{close, that is,}
   \]

   \[
   \|x_i - x_{i'}\| \leq 2\bar{\alpha}r_j + \beta, \ \bar{\alpha} = \frac{\alpha - 1}{2}.
   \]

   3. We build the optimal detectors \( \phi_{ij,i'j} \), along with their risks \( \epsilon_{ij,i'j} \), for all \( i, i' \in I_j \) such that \((i, i') \notin C^j_{\alpha,\beta} \).

   If \( \epsilon_{ij,i'j} = 1 \) for a pair \( i, i' \) of the latter type, that is, \( A(X_{ij}) \cap A(X_{i'j}) \neq \emptyset \), we claim that \((\alpha, \beta)\) is inadmissible and stop. Otherwise we find the smallest \( K = K_j \) such that the spectral norm of the symmetric \( M \times M \) matrix \( E_{ij}^{K_j} \) with the entries

   \[
   E_{ij}^{K_j} = \begin{cases} \epsilon_{ij,i'j}^K, & i \in I_j, i' \in I_j, (i, i') \notin C^j_{\alpha,\beta} \\ 0, & \text{otherwise} \end{cases}
   \]

   does not exceed \( \bar{\epsilon} = \epsilon/(N + 1) \). We then use the machinery of Section 2.5.2.3 to build detector-based test \( T_{C^j_{\alpha,\beta}}^{K_j} \) which decides on the hypotheses \( H_{ij}, i \in I_j \), with \( C^j_{\alpha,\beta}-\text{risk not exceeding } \bar{\epsilon} \).

   It may happen that the outlined process stops when processing some value \( \tilde{j} \) of \( j \); if this does not happen, we set \( \tilde{j} = N + 1 \). Now, if the process does stop, and stops with the claim that \((\alpha, \beta)\) is inadmissible, we call \((\alpha, \beta)\) inadmissible and terminate – in this case we fail to produce a desired inference; note that if this is the case, \((\alpha, \beta)\) is inadmissible in the sense of Section 2.5.3.1 as well. When we do not stop with inadmissibility claim, we call \((\alpha, \beta)\) admissible, and in this case we do produce an inference, specifically, as follows.

Processing observations:

1. We set \( \bar{J} = \{ 0, 1, \ldots, \tilde{j} = \tilde{j} - 1 \} \), \( K = K(\alpha, \beta) = \max_{0 \leq j \leq \tilde{j}} K_j \). Note that \( \bar{J} \) is nonempty due to \( \tilde{j} > 0 \).  \(^{12}\)

   \(^{12}\)All the sets \( X_{ij} \) contain \( X \) and thus are nonempty, so that \( I_0 = \{ 1, \ldots, M \} \neq \emptyset \), and thus we cannot stop at step \( j = 0 \) due to \( I_0 = \emptyset \); other possibility to stop at step \( j = 0 \) is ruled out by the fact that we are in the case when \((\alpha, \beta)\) is admissible.
2. Let \( \omega^K = (\omega_1, \ldots, \omega_K) \) with independent across \( k \) components stemming from unknown signal \( x_\ast \in X \) according to (2.85). We put \( \hat{I}_{-1}(\omega^K) = \{1, \ldots, M\} = \hbar_0 \\
\)
a) For \( j = 0, 1, \ldots, \hat{j} \), we act as follows. When processing \( j \), we have at our disposal subsets \( \hat{I}_k(\omega^K) \subset \{1, \ldots, M\}, -1 \leq k < j \). To build the set \( \hat{I}_j(\omega^K) \)
\begin{enumerate}
    \item we apply the test \( T^K_{\alpha, \beta} \) to the initial \( K_j \) components of the observation \( \omega^K \). Let \( \hat{I}_j^+(\omega^K) \) be the set of hypotheses \( H_{ij} \), \( i \in \hat{I}_j \), accepted by the test;
    \item it may happen that \( \hat{I}_j^+(\omega^K) = \emptyset \); if it is so, we terminate;
    \item if \( \hat{I}_j^+(\omega^K) \) is nonempty, we look, one after one, at indices \( i \in \hat{I}_j^+(\omega^K) \) and call the index \( i \) good if for every \( \ell \in \{-1, 0, \ldots, j - 1\}, i \in \hat{I}_\ell(\omega^K) \);
    \item we define \( \hat{I}_j(\omega^K) \) as the set of good indices of \( \hat{I}_j^+(\omega^K) \) if this set is not empty and proceed to the next value of \( j \) (if \( j < \hat{j} \)), or terminate (if \( j = \hat{j} \)).
\end{enumerate}

We terminate if there are no good indices in \( \hat{I}_j^+(\omega^K) \).

b) Upon termination, we have at our disposal a collection \( \hat{I}_j(\omega^K), 0 \leq j \leq \hat{j}(\omega^K) \), of all sets \( \hat{I}_j(\omega^K) \) we have built (this collection can be empty, which we encode by setting \( \hat{j}(\omega^K) = -1 \)). When \( \hat{j}(\omega^K) = -1 \), our inference remains undefined. Otherwise we select from the set \( \hat{I}_{\hat{j}}(\omega^K) \) an index \( i_{\alpha, \beta}(\omega^K) \), say, the smallest one, and claim that the point \( x_{\alpha, \beta}(\omega^K) \) is the “nearly closest” to \( x_\ast \) point among \( x_1, \ldots, x_M \).

We have the following analog of Proposition 2.33:

**Proposition 2.34.** Assuming \((\alpha, \beta)\) admissible, for the just defined inference \( \omega^K \rightarrow i_{\alpha, \beta}(\omega^K) \) and for every \( x_\ast \in X \), denoting by \( P^K_{x_\ast} \) the distribution of stationary \( K \)-repeated observation \( \omega^K \) stemming from \( x_\ast \) one has

\[
P^K_{x_\ast}\{\omega^K : i_{\alpha, \beta}(\omega^K) \text{ is well defined and } \|x_\ast - x_{i_{\alpha, \beta}(\omega^K)}\| \leq \alpha \rho(x_\ast) + \beta \} \geq 1 - \epsilon.
\]

**Proof.** Let us fix the signal \( x_\ast \in X \) underlying observations \( \omega^K \). Same as in the proof of Proposition 2.33, let \( j_\ast \) be such that \( \rho(x_\ast) = r_{j_\ast} \), and let \( i_\ast \leq M \) be such that \( x_\ast \in X_{i_\ast, j_\ast} \). Clearly, \( i_\ast \) and \( j_\ast \) are well defined, and the hypotheses \( H_{i_\ast, j_\ast} \), \( 0 \leq j \leq j_\ast \), are true. In particular, \( X_{i, j} \neq \emptyset \) when \( j \leq j_\ast \), implying that \( i_\ast \in \hat{I}_j \), \( 0 \leq j \leq j_\ast \), whence also \( \hat{j} \geq j_\ast \).

For \( 0 \leq j \leq j_\ast \), let \( \mathcal{E}_j \) be the set of all realizations of \( \omega^K \) such that

\[
i_\ast \in \hat{I}_j^+(\omega^K) \& \{ (i, i) \in C^j_{i, \beta, \delta} \forall i \in \hat{I}_j^+(\omega^K) \}.
\]

Since \( C^j_{i, \beta, \delta} \)-risk of the test \( T^K_{C^j_{i, \beta, \delta}} \) is \( \leq \epsilon \), we conclude that the \( P^K_{x_\ast} \) probability of \( \mathcal{E}_j \) is at least \( 1 - \epsilon \), whence the \( P^K_{x_\ast} \) probability of the event

\[
\mathcal{E} = \bigcap_{j=0}^{j_\ast} \mathcal{E}_j
\]

is at least \( 1 - (N + 1)\epsilon = 1 - \epsilon \).

Now let

\[
\omega^K \in \mathcal{E}.
\]
Then,

- By the definition of $E_j$, when $j \leq j_*$, we have $i_* \in \mathcal{I}_j^+(\omega^K)$, whence, by evident induction in $j$, $i_* \in \mathcal{I}_j(\omega^K)$ for all $j \leq j_*$.
- We conclude from the above that $\tilde{j}(\omega^K) \geq j_*$. In particular, $i := i_{\alpha,\beta}(\omega^K)$ is well defined and turned out to be good at step $\tilde{j} \geq j_*$, implying that $i \in \mathcal{I}_{\tilde{j}_*}(\omega^K) \subset \mathcal{I}_{j_*}^+(\omega^K)$.

Thus, $i \in \mathcal{I}_{j_*}^+(\omega^K)$, which combines with the definition of $E_{j_*}$ to imply that $i$ and $i_*$ are $C_{\alpha,\beta}$-close to each other, whence

$$\|x_{i_{(\alpha,\beta)}(\omega^K)} - x_*\| \leq 2\alpha r_j + \beta = 2\alpha \rho(x_*) + \beta,$$

resulting in the desired relation

$$\|x_{i_{(\alpha,\beta)}(\omega^K)} - x_*\| \leq 2\alpha \rho(x_*) + \beta + \|x_{i_*} - x_*\| \leq [2\alpha + 1] \rho(x_*) + \beta = \alpha \rho(x_*) + \beta. \quad \square$$

### 2.5.3.3 “Near-optimality”

We augment the above constructions with the following

**Proposition 2.35.** Let in the nature for some positive integer $K$, $\epsilon \in (0,1/2)$ and a pair $(a,b) \geq 0$ there exists an inference $\omega^K \mapsto i(\omega^K) \in \{1,\ldots,M\}$ such that whenever $x_* \in X$, we have

$$\Pr_{\omega^K \sim P_{x_*}^K} \{ \|x_* - x_{i(\omega^K)}\| \leq a \rho(x_*) + b \} \geq 1 - \epsilon.$$

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 the both constructions in Sections 2.5.3.1 and 2.5.3.2 one has

$$K(\alpha, \beta) \leq \text{Ceil} \left( \frac{2 + \ln(M(N+1))/\ln(1/\epsilon)}{1 - \ln(1-\epsilon)/\ln(1/\epsilon)} \right); \quad (2.91)$$

**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 way). Observe that with $\alpha, \beta$ as above, there exists a simple test deciding on a pair of hypotheses $H_{ij}, H_{ij'}$ which are not $C_{\alpha,\beta}$-close to each other via stationary $K$-repeated observation $\omega^K$ with risk $\leq \epsilon$. Indeed, the desired test $T$ is as follows: given $ij \in J$, $ij' \in J$, and observation $\omega^K$, we compute $i(\omega^K)$ and accept $H_{ij}$ if and only if $\|x_{i(\omega^K)} - x_i\| \leq (a+1)r_j + b$, and accept $H_{ij'}$ otherwise. Let us check that the risk of this test indeed is at most $\epsilon$. Assume, first, that $H_{ij}$ takes place. The $P_{x_*}^K$-probability of the event

$$\mathcal{E} : \|x_{i(\omega^K)} - x_*\| \leq a \rho(x_*) + b$$

is at least $1 - \epsilon$ 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

$$\|x_{i(\omega^K)} - x_i\| \leq \|x_{i(\omega^K)} - x_*\| + \|x_i - x_*\| \leq a \rho(x_*) + b + r_j \leq (a+1)r_j + b.$$
We conclude that if \( H_{ij} \) is true and \( \omega^K \in \mathcal{E} \), then the test \( T \) accepts \( H_{ij} \), and thus the \( P^K_{x_i} \)-probability for the simple test \( T \) 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 \( \omega^K \in \mathcal{E} \), which happens with the \( P^K_{x_i} \)-probability at least \( 1 - \epsilon \), for the same reasons as above, we have

\[
\|x_i(\omega^K) - x_{i'}\| \leq (a + 1)r_j + b.
\]

It follows that when \( H_{i'j'} \) takes place and \( \omega^K \in \mathcal{E} \), we have

\[
\|x_i(\omega^K) - x_{i'}\| > (a + 1)r_j + b,
\]

since otherwise we would have

\[
\|x_i - x_{i'}\| \leq \|x_i(\omega^K) - x_{i'}\| \leq (a + 1)r_j + b + (a + 1)r_{j'} + b
\]

what contradicts the fact that \( ij \) and \( i'j' \) are not \( C_{\alpha,\beta} \)-close. Thus, whenever \( H_{i'j'} \) holds true and \( \mathcal{E} \) takes place, we have \( \|x_i(\omega^K) - x_{i'}\| > (a + 1)r_j + b \), implying by the definition of \( T \) that \( T \) accepts \( H_{i'j'} \). Thus, the \( P^K_{x_i} \)-probability not to accept \( H_{i'j'} \) when the hypotheses is true is at most \( \epsilon \).

From the just established fact that whenever \( (ij, i'j') \notin C_{\alpha,\beta} \), the hypotheses \( H_{ij} \), \( H_{i'j'} \) can be decided upon, via \( K \) 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 \( (\alpha, \beta) \) is an admissible pair.

Same as in the proof of Proposition 2.31, by basic properties of simple observation schemes, the fact that the hypotheses \( H_{ij} \), \( H_{i'j'} \) with \( (ij, i'j') \notin C_{\alpha,\beta} \) can be decided upon via \( K \)-repeated observations (2.85) with risk \( \leq \epsilon < 1/2 \) implies that \( \epsilon_{ij,i'j'} \leq [2\sqrt{\epsilon(1-\epsilon)}]^{1/K} \), whence, again by basic results on simple observation scheme (look once again at the proof of Proposition 2.31), the \( C_{\alpha,\beta} \)-risk of \( K \)-observation detector-based test \( T_K \) deciding on the hypotheses \( H_{ij} \), \( ij \in \mathcal{J} \), up to closeness \( C_{\alpha,\beta} \) does not exceed \( \text{Card}(\mathcal{J})[2\sqrt{\epsilon(1-\epsilon)}]^{K/K} \leq M(N + 1)[2\sqrt{\epsilon(1-\epsilon)}]^{K/K} \), and (2.91) follows.

\[\square\]

**Comment.** Proposition 2.35 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 \( 10,000 \). (2.91) reads \( K = \text{Ceil}(10.7K) \) not a disaster! The actual statistical toll of 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

As an illustration of the approach presented in this section consider the following (toy) problem:

A signal \( x_* \in \mathbb{R}^n \) (one may 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

\[
\omega = Ax_* + \xi, \; \xi \sim \mathcal{N}(0, \sigma^2I_n),
\]

(2.92)
where $A$ is “discretized integration”:

$$(Ax)_s = \frac{1}{n} \sum_{j=1}^{s} x_s, \quad s = 1, \ldots, n.$$ 

We want to approximate $x$ in the discrete version of $L_1$-norm

$$\|y\| = \frac{1}{n} \sum_{s=1}^{n} |y_s|, \quad y \in \mathbb{R}^n$$ 

by a low order polynomial.

In order to build the approximation, we use a single observation $\omega$ as in (2.92), to build 5 candidate estimates $x_i, \ i = 1, \ldots, 5,$ of $x_\ast$. Specifically, $x_i$ is the Least Squares polynomial, of degree $\leq i - 1$, approximation of $x$:

$$x_i = \arg\min_{y \in P_{i-1}} \|Ay - \omega\|_2^2,$$

where $P_\kappa$ is the linear space of algebraic polynomials, of degree $\leq \kappa$, of discrete argument $s$ varying in $\{1, 2, \ldots, n\}$. After the candidate estimates are built, we use additional $K$ observations (2.92) “to select the model” – to select among our estimates the $\|\cdot\|$-closest to $x_\ast$.

In the experiment reported below we use $n = 128$ and $\sigma = 0.01$. The true signal $x_\ast$ is a discretization of a piece-wise linear function of continuous argument $t \in [0, 1]$, with slope 2 to the left of $t = 0.5$, and with slope $-2$ to the right of $t = 0.5$; at $t = 0.5$, the function has a jump. The a priori information on the true signal is that it belongs to the box $\{x \in \mathbb{R}^n : \|x\|_{\infty} \leq 1\}$. The signal and sample polynomial approximations $x_i$ of $x_\ast$, $1 \leq i \leq 5$, are presented on the top plot in Figure 2.5; their actual $\|\cdot\|$-distances to $x_\ast$ are as follows:

<table>
<thead>
<tr>
<th>$i$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$|x_i - x_\ast|$</td>
<td>0.534</td>
<td>0.364</td>
<td>0.233</td>
<td>0.161</td>
<td>0.172</td>
</tr>
</tbody>
</table>

Setting $\epsilon = 0.01, \ N = 22$ and $\theta = 2^{1/4}, \ \alpha = 3$ and $\beta = 0.05$ resulted in $K = 3$. In a series of 1000 simulations of the resulting inference, all 1000 results correctly identified the $\|\cdot\|$-closest to $x_\ast$ candidate estimate $x_4$, in spite of the factor $\alpha = 3$ in (2.88). 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 guarantees deteriorate: with $K = 1$ and $K = 2$, the theory guarantees the validity of (2.88) with probabilities 0.77 and 0.97, respectively.

### 2.6 SEQUENTIAL HYPOTHESIS TESTING

#### 2.6.1 Motivation: Election Polls

Let us consider the following “practical” question.

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
candidate. 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 $\mathbb{R}^L$ with unit entry in a position $\ell$ meaning that the voter is about to vote for the $\ell$-th candidate. The entries $\mu_\ell$ in the average $\mu$, over the population, of these vectors are the fractions of votes in favor of the $\ell$-th candidate, and the elected candidate is the one “indexing” the largest of $\mu_\ell$’s. Now assume that we select at random, from the uniform distribution, a member of the population and observe his preference vector. Our observation $\omega$ is a realization of discrete random variable taking values in the set $\Omega = \{e_1, ..., e_L\}$ of basic orths in $\mathbb{R}^L$, and $\mu$ is the distribution of $\omega$ (technically, the density of this distribution w.r.t. the counting measure $\Pi$ on $\Omega$). Selecting a small threshold $\delta$ and assuming that the true – unknown to us – $\mu$ is such that the largest entry in $\mu$ is at least by $\delta$ larger than every other entry and that $\mu_\ell \geq \frac{1}{N}$ for all $\ell$, $N$ being the population size, we can model the population preference for the $\ell$-th candidate with

$$
\mu \in M_\ell = \{\mu \in \mathbb{R}^d : \mu_i \geq \frac{1}{N}, \sum_i \mu_i = 1, \mu_\ell \geq \mu_i + \delta \forall (i \neq \ell)\} \\
\subseteq M = \{\mu \in \mathbb{R}^d : \mu > 0, \sum_i \mu_i = 1\}.
$$

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 $\mu$ on $\Omega$, with $\mu$ known to belong to $\bigcup_{\ell=1}^L M_\ell$. Therefore, to predict the winner is the same as to decide on $L$ convex hypotheses, $H_1, ..., H_L$, in the Discrete o.s., with $H_\ell$ stating that $\omega_1, ..., \omega_K$.

---

13With the size $N$ of population in the range of tens of thousands and $\delta$ like $1/N$, both these assumptions seem to be quite realistic.
are drawn, independently of each other, from a distribution $\mu \in M_\ell$. What we end up with, is the problem of deciding on $L$ convex hypotheses in the Discrete o.s. with $L$-element $\Omega$ via stationary $K$-repeated observations.

**Illustration.** Consider two-candidate elections; now the goal of a poll is, given $K$ independent of each other realizations $\omega_1, \ldots, \omega_K$ of random variable $\omega$ taking value $\chi = 1, 2$ with probability $\mu_\chi$, $\mu_1 + \mu_2 = 1$, to decide what is larger, $\mu_1$ or $\mu_2$. As explained above, we select somehow a threshold $\delta$ and impose on the unknown $\mu$ a priori assumption that the gap between the largest and the next largest (in our case – just the smallest) entry of $\mu$ is at least $\delta$, 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

$$H_1 : \mu \in M_1 = \{ \mu : \mu_1 \geq \frac{1+\delta}{2}, \mu_2 \geq 0, \mu_1 + \mu_2 = 1 \},$$

$$H_2 : \mu \in M_2 = \{ \mu : \mu_2 \geq \frac{1+\delta}{2}, \mu_1 \geq 0, \mu_1 + \mu_2 = 1 \}.$$

We now want to decide on these two hypotheses via stationary $K$-repeated observation. We are in the case of simple (specifically, Discrete) o.s.; the optimal detector explained above, we select somehow a threshold $\delta$ and impose on the unknown $\mu$ a priori assumption that the gap between the largest and the next largest (in our case – just the smallest) entry of $\mu$ is at least $\delta$, thus arriving at two hypotheses:

$$\mu_1 \geq \mu_2 + \delta, \quad \mu_2 \geq \mu_1 + \delta,$$

which is the same as

$$H_1 : \mu \in M_1 = \{ \mu : \mu_1 \geq \frac{1+\delta}{2}, \mu_2 \geq 0, \mu_1 + \mu_2 = 1 \},$$

$$H_2 : \mu \in M_2 = \{ \mu : \mu_2 \geq \frac{1+\delta}{2}, \mu_1 \geq 0, \mu_1 + \mu_2 = 1 \}.$$

We now want to decide on these two hypotheses via stationary $K$-repeated observation. We are in the case of simple (specifically, Discrete) o.s.; the optimal detector explained above, we select somehow a threshold $\delta$ and impose on the unknown $\mu$ a priori assumption that the gap between the largest and the next largest (in our case – just the smallest) entry of $\mu$ is at least $\delta$, 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

$$H_1 : \mu \in M_1 = \{ \mu : \mu_1 \geq \frac{1+\delta}{2}, \mu_2 \geq 0, \mu_1 + \mu_2 = 1 \},$$

$$H_2 : \mu \in M_2 = \{ \mu : \mu_2 \geq \frac{1+\delta}{2}, \mu_1 \geq 0, \mu_1 + \mu_2 = 1 \}.$$

We now want to decide on these two hypotheses via stationary $K$-repeated observation. We are in the case of simple (specifically, Discrete) o.s.; the optimal detector explained above, we select somehow a threshold $\delta$ and impose on the unknown $\mu$ a priori assumption that the gap between the largest and the next largest (in our case – just the smallest) entry of $\mu$ is at least $\delta$, 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

$$H_1 : \mu \in M_1 = \{ \mu : \mu_1 \geq \frac{1+\delta}{2}, \mu_2 \geq 0, \mu_1 + \mu_2 = 1 \},$$

$$H_2 : \mu \in M_2 = \{ \mu : \mu_2 \geq \frac{1+\delta}{2}, \mu_1 \geq 0, \mu_1 + \mu_2 = 1 \}.$$
Table 2.1: Sample of values of poll size $K_{0.01}(\delta)$ as a function of $\delta$ for 2-candidate $(L = 2)$ and 5-candidate $(L = 5)$ elections. Values of $\delta$ form a geometric progression with ratio $10^{-1/4}$.

<table>
<thead>
<tr>
<th>$\delta$</th>
<th>$K_{0.01}(\delta)$, $L = 2$</th>
<th>$K_{0.01}(\delta)$, $L = 5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5623</td>
<td>25</td>
<td>32</td>
</tr>
<tr>
<td>0.3162</td>
<td>88</td>
<td>114</td>
</tr>
<tr>
<td>0.1778</td>
<td>287</td>
<td>373</td>
</tr>
<tr>
<td>0.1000</td>
<td>917</td>
<td>1198</td>
</tr>
<tr>
<td>0.0562</td>
<td>2908</td>
<td>3788</td>
</tr>
<tr>
<td>0.0316</td>
<td>9209</td>
<td>92084</td>
</tr>
<tr>
<td>0.0177</td>
<td>29118</td>
<td>92098</td>
</tr>
</tbody>
</table>

Reality is captured by our model). Indeed, from Proposition 2.25 we know that our poll size is within an explicit factor, depending solely on $\epsilon$, 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 easily construct a more accurate “numerical” lower bounds on the sizes of ideal polls, specifically, as follows. When computing the optimal detector $\phi_*$, we get, as a byproduct, two distributions, $\mu^*$, $\nu^*$ obeying $H_1$, $H_2$, respectively. Denoting by $\mu^*_K$, $\nu^*_K$ the distributions of $K$-element i.i.d. samples drawn from $\mu^*$ and $\nu^*$, the risk of deciding on two simple hypotheses on the distribution of $\omega^K$, stating that this distribution is $\mu^*_K$, respectively, $\nu^*_K$ 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 2.2, namely,

$$\sum_{i_1, \ldots, i_K \in \{1, 2\}} \min \left[ \prod_\ell (\mu^*_{i_\ell}), \prod_\ell (\nu^*_{i_\ell}) \right] = \mathbb{E}(i_1, \ldots, i_K) \left\{ \min \left[ \prod_\ell (2\mu^*_{i_\ell}), \prod_\ell (2\nu^*_{i_\ell}) \right] \right\},$$

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 $\leq 0.01$. Here are the resulting lower bounds (along with upper bounds from Table 2.1):

<table>
<thead>
<tr>
<th>$\delta$</th>
<th>0.5623</th>
<th>0.3162</th>
<th>0.1778</th>
<th>0.1000</th>
<th>0.0562</th>
<th>0.0316</th>
<th>0.0177</th>
<th>0.0100</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K / K$</td>
<td>14.12</td>
<td>38.01</td>
<td>106.16</td>
<td>206.29</td>
<td>344.34</td>
<td>617.17</td>
<td>1099.69</td>
<td>2012.52</td>
</tr>
</tbody>
</table>

Lower ($\overline{K}$) and upper ($\underline{K}$) 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 $L \geq 2$. In our model of the corresponding problem we decide, via stationary $K$-repeated observations drawn from unknown probability distribution

---

14In actual opinion polls, additional information is used. For instance, 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.
µ on L-element set, on L hypotheses

\[ H_\ell : \mu \in M_\ell = \left\{ \mu \in \mathbb{R}^d : \mu_i \geq \frac{1}{N}, 1 \leq i \leq L, \sum_i \mu_i = 1, \mu_\ell \geq \mu_{\ell'} + \delta \forall (\ell' \neq \ell) \right\}, \ell \leq L. \]  

(2.94)

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 – \( H_\ell \) is close to \( H_{\ell'} \) if and only if \( \ell = \ell' \), predicting the outcome of elections with risk \( \epsilon \) becomes the problem of deciding upon our multiple hypotheses with \( \mathcal{C} \)-risk \( \leq \epsilon \). Thus, we can use pairwise detectors yielded by Theorem 2.23 to identify the smallest possible \( K = K_\epsilon \) such that the test \( T^K_\mathcal{C} \) 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 \( \epsilon \) is set to 0.01).

### 2.6.2 Sequential hypothesis testing

In view of the above analysis, when predicting outcomes of “close run” elections, huge poll sizes are necessary. 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 [232] is to pass to sequential tests where the observations are processed one by one, and at every instant 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 our machinery can be utilized to conceive a sequential test for the problem of predicting the outcome of L-candidate elections. Thus, our goal is, given a small threshold \( \delta \), to decide upon L hypotheses (2.94). 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 \( \delta \), and let the number of the thresholds be \( S \).

2. We split somehow (e.g., equally) the risk \( \epsilon \) which we want to guarantee into \( S \) portions \( \epsilon_s, 1 \leq s \leq S \), so that \( \epsilon_s \) are positive and

\[ \sum_{s=1}^{S} \epsilon_s = \epsilon. \]

3. For \( s \in \{1, 2, \ldots, S\} \), we define, along with the hypotheses \( H_\ell \), the hypotheses

\[ H^*_s : \mu \in M^*_\ell = \left\{ \mu \in M_\ell : \mu_\ell \geq \mu_{\ell'} + \delta_s, \forall (\ell' \neq \ell) \right\}, \ell = 1, \ldots, L, \]

see Figure 2.6, and introduce \( 2L \) hypotheses \( G^*_{2\ell-1} = H_\ell \), and \( G^*_{2\ell} = H^*_\ell \), \( 1 \leq \ell \leq L \). It is convenient to color these hypotheses in \( L \) colors, with \( G^*_{2\ell-1} = H_\ell \) and \( G^*_{2\ell} = H^*_\ell \) assigned color \( \ell \). We define also s-th closeness \( \mathcal{C}_s \) as follows:

When \( s < S \), hypotheses \( G^*_s \) and \( G^*_1 \) 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
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 $T_s$ yielded by this construction as applied to a stationary $K$-repeated observation allows to decide on the hypotheses $G_1^s, ..., G_{2K}^s$, with $C_s$-risk $\leq \epsilon_s$. The required $K$ exists due to the already mentioned separation of members in a pair of not $C_s$-close hypotheses $G_i^s, G_j^s$. It is easily seen that $K(1) \leq K(2) \leq ... \leq K(S-1)$. However, it may happen that

have odd indices (that is, one of them is $H_\ell$, and another one is $H_{\ell'}$ with $\ell \neq \ell'$).

When $s = S$ (in this case $G_{2S-1}^S = H_\ell = G_{2\ell}^S$), hypotheses $G_\ell^S$ and $G_{\ell'}^S$ are $C_S$-close to each other if and only if they are of the same color, i.e., both coincide with the same hypothesis $H_\ell$.

Observe that $G_i^s$ is a convex hypothesis:

$$G_i^s : \mu \in Y_i^s \quad \text{[} Y_{2\ell-1}^s = M_\ell, Y_{2\ell}^s = M_{\ell'} \text{]}$$

The key observation is that when $G_i^s$ and $G_j^s$ are not $C_s$-close, sets $Y_i^s$ and $Y_j^s$ are “separated” by at least $\delta_s$, meaning that for some vector $e \in \mathbb{R}^L$ with just two nonvanishing entries, equal to 1 and $-1$ we have

$$\min_{\mu \in Y_i^s} e^T \mu \geq \delta_s + \max_{\mu \in Y_j^s} e^T \mu. \quad (2.95)$$

Indeed, let $G_i^s$ and $G_j^s$ be not $C_s$-close to each other. That means that the hypotheses are of different colors, say, $\ell$ 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 $G_{\ell'}^s$, so that

$$Y_j^s \subset \{ \mu : \mu_\ell - \mu_{\ell'} \geq \delta_s \},$$

while $Y_i^s$ is contained in the set $\{ \mu : \mu_\ell \geq \mu_{\ell'} \}$. Specifying $e$ as the vector with just two nonzero entries, $\ell$-th equal to 1 and $\ell'$-th equal to $-1$, we ensure (2.95).

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 $T_s$ yielded by this construction as applied to a stationary $K$-repeated observation allows to decide on the hypotheses $G_1^s, ..., G_{2K}^s$, with $C_s$-risk $\leq \epsilon_s$. The required $K$ exists due to the already mentioned separation of members in a pair of not $C_s$-close hypotheses $G_i^s, G_j^s$. It is easily seen that $K(1) \leq K(2) \leq ... \leq K(S-1)$. However, it may happen that
HYPOTHESIS TESTING

\(K(S - 1) > K(S)\), the reason being that \(C_S\) is defined differently than \(C_s\) with \(s < S\). We set

\[S = \{s \leq S : K(s) \leq K(S)\}.\]

For example, here is what we get in \(L\)-candidate Opinion Poll problem when \(S = 8\), \(\delta = \delta_S = 0.01\), and for properly selected \(\epsilon_s\) with \(\sum_{s=1}^{S} \epsilon_s = 0.01:\)

<table>
<thead>
<tr>
<th>(L)</th>
<th>(K(1))</th>
<th>(K(2))</th>
<th>(K(3))</th>
<th>(K(4))</th>
<th>(K(5))</th>
<th>(K(6))</th>
<th>(K(7))</th>
<th>(K(8))</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>177</td>
<td>617</td>
<td>1829</td>
<td>5099</td>
<td>15704</td>
<td>49699</td>
<td>153299</td>
<td>160118</td>
</tr>
<tr>
<td>5</td>
<td>208</td>
<td>723</td>
<td>2175</td>
<td>6204</td>
<td>19205</td>
<td>60781</td>
<td>188203</td>
<td>187718</td>
</tr>
</tbody>
</table>

\(S = 8\), \(\delta_S = 10^{-s/4}\),

\(S = \{1, 2, \ldots, 8\}\) when \(L = 2\) and \(S = \{1, 2, \ldots, 6\} \cup \{8\}\) when \(L = 5\).

5. Our sequential test \(T_{\text{seq}}\) works in attempts (stages) \(s \in S\) – it tries to make conclusions after observing \(K(s)\), \(s \in S\), realizations \(\omega_k\) of \(\omega\). At the \(s\)-th attempt, we apply the test \(T_s\) to the collection \(\omega^K(s)\) of observations obtained so far to decide on hypotheses \(G_1^s, \ldots, G_{2L}^s\). If \(T_s\) accepts some of these hypotheses and all accepted hypotheses are of the same color, let it be \(\ell\), the sequential test accepts the hypothesis \(H_\ell\) and terminates; otherwise we continue to observe the realizations of \(\omega\) (if \(s < S\)) or terminate with no hypotheses accepted/rejected (if \(s = S\)).

It is easily seen that the risk of the outlined sequential test \(T_{\text{seq}}\) does not exceed \(\epsilon\), meaning that whatever be a distribution \(\mu \in \bigcup_{\ell=1}^{L} M_{\ell}\) underlying observations \(\omega_1, \omega_2, \ldots, \omega_{K(S)}\) and \(\epsilon_s\) such that \(\mu \in M_{\ell_s}\), the \(\mu\)-probability of the event

\[\text{\(T_{\text{seq}}\) accepts exactly one hypothesis, namely, \(H_{\ell_s}\),}\]

is at least \(1 - \epsilon\).

Indeed, observe, first, that the sequential test always accepts at most one of the hypotheses \(H_1, \ldots, H_L\). Second, let \(\omega_k \sim \mu\) with \(\mu\) obeying \(H_{\ell_s}\). Consider events \(E_s\), \(s \in S\), defined as follows:

- when \(s < S\), \(E_s\) is the event “the test \(T_s\) as applied to observation \(\omega^K(s)\) does not accept the true hypothesis \(G_{2\ell_s - 1}^s = H_{\ell_s}\);”
- \(E_S\) is the event “as applied to observation \(\omega^K(S)\), the test \(T_S\) does not accept the true hypothesis \(G_S^S = H_{\ell_s}\) or accepts a hypothesis not \(C_S\)-close to \(G_S^S\).”

Note that by our selection of \(K(s)\)'s, the \(\mu\)-probability of \(E_s\) does not exceed \(\epsilon_s\), so that the \(\mu\)-probability of none of the events \(E_s\), \(s \in S\), taking place is at least \(1 - \epsilon\). To justify the above claim on the risk of the sequential test, all we need to verify is that when none of the events \(E_s\), \(s \in S\), takes place, the sequential test accepts the true hypothesis \(H_{\ell_s}\). Verification is immediate: let the observations be such that none of \(E_s\) takes place. We claim that in this case

(a) The sequential test does accept a hypothesis – if this does not happen at the \(s\)-th attempt with some \(s < S\), it definitely happens at the \(S\)-th attempt. Indeed, since \(E_S\) does not take place, \(T_S\) accepts \(G_{2\ell_s - 1}^S\) and all other hypotheses, if any, accepted by \(T_S\) are \(C_S\)-close to \(G_{2\ell_s - 1}^S\), implying by construction of \(C_S\) that \(T_S\) 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, \(H_{\ell'}, \ell' \neq \ell_s\), and it happens at the \(s\)-th attempt, and let us lead this assumption to a contradiction. Observe that under our assumption the test \(T_s\) as applied to observation \(\omega^K(s)\) does accept some hypothesis \(G_1^s\), but does not accept the true hypothesis
SL

It can be easily seen that when

\[ G_{2k-1}^* = H_{\ell_*} \]

(indeed, assuming \( G_{2k-1}^* \) to be accepted, its color, which is \( \ell_* \), should be the same as the color \( \ell' \) of \( G_1^* \) – 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, \( T_n \) does accept the true hypothesis \( G_{2k-1}^* \), and we arrive at the desired contradiction.

(a) and (b) provide us with the 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 \( \epsilon \). At the same time it is easily seen that when the distribution \( \mu \) of our observation is “deeply inside” some set \( M_\ell \), specifically, \( \mu \in M_\ell^s \) for some \( s \in S \), \( s < S \), then the \( \mu \)-probability to terminate not later than after just \( K(s) \) realizations \( \omega_k \) 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 \( s \in S \) and \( \omega_k \sim \mu \in M_\ell^s \), so that \( \mu \) obeys the hypothesis \( G_{2k}^* \). 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 \( H_\ell \), wrong hypothesis \( H_{\ell'} \), \( \ell' \neq \ell \). Note that \( E_t \) can take place only when \( T_t \) does not accept the true hypothesis \( G_{2k-1}^* = H_\ell \), and \( \mu \)-probability of this outcome is \( \leq \epsilon_t \).
- \( E_s \) occurs when \( T_s \) does not accept the true hypothesis \( G_{2k}^* \) or accepts it along with some hypothesis \( G_j^*, 1 \leq j \leq 2L \), of color different from \( \ell \). Note that we are in the situation where the hypothesis \( G_{2k}^* \) is true, and, by construction of \( C_\ell \), all hypotheses \( C_\ell \)-close to \( G_{2k}^* \) are of the same color \( \ell \) as \( G_{2k}^* \). Recalling what \( C_\ell \)-risk is and that the \( C_\ell \)-risk of \( T_s \) is \( \leq \epsilon_s \), we conclude that the \( \mu \)-probability of \( E_s \) is at most \( \epsilon_s \).

The bottom line is that \( \mu \)-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 \( H_\ell \). 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 \( \delta \); 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)

<table>
<thead>
<tr>
<th>( S / K(S) )</th>
<th>( K(L = 2) )</th>
<th>( K(L = 5) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( K(L = 2) )</td>
<td>25</td>
<td>32</td>
</tr>
<tr>
<td>( K(L = 5) )</td>
<td>32</td>
<td>42</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( S / K(S) )</th>
<th>0.5623</th>
<th>0.3162</th>
<th>0.1778</th>
<th>0.1000</th>
<th>0.0562</th>
<th>0.0316</th>
<th>0.0177</th>
<th>0.0100</th>
</tr>
</thead>
<tbody>
<tr>
<td>( K(L = 2) )</td>
<td>25</td>
<td>28</td>
<td>917</td>
<td>2908</td>
<td>9206</td>
<td>29118</td>
<td>92098</td>
<td></td>
</tr>
<tr>
<td>( K(L = 5) )</td>
<td>32</td>
<td>37.3</td>
<td>1193</td>
<td>3184</td>
<td>11976</td>
<td>37885</td>
<td>119745</td>
<td></td>
</tr>
</tbody>
</table>

Volume \( K \) of non-sequential test, number \( S \) of stages and worst-case volume \( K(S) \) of sequential test as functions of threshold \( \delta = \delta_S \). Risk \( \epsilon \) is set to 0.01.

As it should be, the worst-case volume of sequential test is significantly larger than
the volume of the non-sequential test.\footnote{\textsuperscript{15}} This being said, look at what happens in the “average,” rather than the worst, case, specifically, let us look at the empirical distribution of the volume when the distribution $\mu$ of observations is selected in the $L$-dimensional probabilistic simplex $\Delta_L = \{ \mu \in \mathbb{R}^L : \mu \geq 0, \sum_{\ell} \mu_{\ell} = 1 \}$ at random. Here is the empirical statistics of test volume obtained when drawing $\mu$ from the uniform distribution on $\bigcup_{\ell \leq L} M_\ell$ and running the sequential test\footnote{\textsuperscript{16}corresponding to $\delta = 0.01$, $\theta = 10^{-1/4}$ and $\epsilon = 0.01$} on observations drawn from the selected $\mu$:

<table>
<thead>
<tr>
<th>$L$</th>
<th>risk</th>
<th>median</th>
<th>mean</th>
<th>60%</th>
<th>65%</th>
<th>70%</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.0010</td>
<td>177</td>
<td>9182</td>
<td>177</td>
<td>397</td>
<td>617</td>
</tr>
<tr>
<td>5</td>
<td>0.0040</td>
<td>1449</td>
<td>18564</td>
<td>2175</td>
<td>4189</td>
<td>6204</td>
</tr>
<tr>
<td>$L$</td>
<td>75%</td>
<td>80%</td>
<td>85%</td>
<td>90%</td>
<td>95%</td>
<td>100%</td>
</tr>
<tr>
<td>2</td>
<td>617</td>
<td>1223</td>
<td>1829</td>
<td>8766</td>
<td>87911</td>
<td>160118</td>
</tr>
<tr>
<td>5</td>
<td>12704</td>
<td>19205</td>
<td>39993</td>
<td>60781</td>
<td>124249</td>
<td>187718</td>
</tr>
</tbody>
</table>

Parameters (columns “median, mean”) and quantiles (columns “60%,...100%”) of the sample distribution of the observation volume of Sequential test for a given empirical risk (column ”risk”).

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 $\mu$’s (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 $O$ 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 the risk level $\epsilon$, we want to decide $(1 - \epsilon)$-reliably on the color of the distribution underlying observations (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 related constructions and results, an interested reader is referred to \cite{133}.\footnote{\textsuperscript{15}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 $\epsilon$ is now distributed among several tests, so that each of them should be more reliable than the non-sequential test with risk $\epsilon$.}
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 an 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 the fraction \( \theta_i \) 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 \( i \) (\( \ell \)-th entry \( p_{i\ell} \) in \( p_i \) is the fraction of voters of group \( i \) voting for candidate \( \ell \)). For instance, 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 \( q_1, \ldots, q_I \), 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 \( \theta_i \), specifically, to make the ratios \( q_i / \theta_i \) “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 in which 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 \( p^1, \ldots, p^I \) 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 the two-stage selection of respondents, the outcome of a particular interview becomes a pair \((i, \ell)\), with \( i \) identifying the group to which the respondent belongs, and \( \ell \) identifying the candidate preferred by this respondent. In subsequent interviews, the pairs \((i, \ell)\) – these are our observations – are drawn, independently of each other, from the probability distribution on the pairs \((i, \ell), i \leq I, \ell \leq L\), with the probability of an outcome \((i, \ell)\) 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 \( \Omega \) of cardinality \( IL \); the distribution from which the observations are drawn is a probabilistic vector \( \mu \) of the form

\[
\mu = Ax,
\]

where

- \( x = [p^1; \ldots; p^I] \) 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 $p^1, \ldots, p^I$:

$$\mathcal{X} = \{x = [x^1; \ldots; x^I] : x^i \in \Pi_i, 1 \leq i \leq I\},$$

(2.96)

where $\Pi_i$ 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 $\Pi_i$ are convex compact subsets of the positive part $\Delta^+_L = \{p \in \mathbb{R}^L : p > 0, \sum_{\ell} p_{\ell} = 1\}$ of the $L$-dimensional probabilistic simplex:

- $A$ is a “sensing matrix” which, to some extent, is under our control; specifically,

$$A[x^1; \ldots; x^I] = [q_1 x^1; q_2 x^2; \ldots; q_I x^I],$$

(2.97)

with $q = [q_1; \ldots; q_I]$ 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 $\mu$ on $\Omega$, with $\mu$ belonging to the image of some convex compact set $X_\ell \subset \mathcal{X}$ under the mapping $x \mapsto \mu = Ax$. For example, when $\nu = \sum_i \theta_i x^i$, the hypotheses

$$H_\ell : \nu \in M_\ell = \left\{ \nu \in \mathbb{R}^L : \sum_j \nu_j = 1, \nu_j \geq \frac{1}{N}, \nu_\ell \geq \nu_\ell' + \delta, \ell' \neq \ell \right\}, 1 \leq \ell \leq L,$$

considered in Section 2.6.1 can be expressed in terms of the signal $x = [x^1; \ldots; x^I]$:

$$H_\ell : \mu = Ax, x \in X_\ell = \left\{ x = [x^1; \ldots; x^I] : x^i \geq 0, \sum_{\ell} x^i_{\ell} = 1 \forall i \leq I, \sum_i \theta_i x^i_{\ell} \geq \sum_{\ell'} \theta_i x^i_{\ell'} + \delta \forall (\ell' \neq \ell), \sum_i \theta_i x^i_j \geq \frac{1}{N}, \forall j \right\}.$$

(2.98)

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 is called “design of experiments” in Statistics.)

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 Opinion Poll Design (OPD) problem).

Given are

- simple observation scheme $\mathcal{O} = (\Omega, \Pi; \{p_\mu : \mu \in \mathcal{M}\}; F)$, specifically, Gaussian, Poisson or Discrete one, with $\mathcal{M} \subset \mathbb{R}^d$. 


In OPD, \( Q \) is the Discrete o.s. with \( \Omega = \{(i, \ell) : 1 \leq i \leq I, 1 \leq \ell \leq L\} \), that is, points of \( \Omega \) are the potential outcomes “reference group, preferred candidate” of individual interviews.

- a nonempty closed convex signal space \( \mathcal{X} \subset \mathbb{R}^n \), along with \( L \) nonempty convex compact subsets \( X_\ell \) of \( \mathcal{X} \), \( \ell = 1, ..., L \).

In OPD, \( \mathcal{X} \) is the set (2.96) comprised by tuples of allowed distributions of voters’ preferences from various groups, and \( X_\ell \) are the sets (2.98) of signals associated with the hypotheses \( H_q \) we intend to decide upon.

- a nonempty convex compact set \( \mathcal{Q} \) in some \( \mathbb{R}^N \) along with a continuous mapping \( q \mapsto A_q \) from \( \mathcal{Q} \) into the space of \( d \times n \) matrices such that

\[
\forall (x \in \mathcal{X}, q \in \mathcal{Q}) : A_q x \in \mathcal{M}.
\] (2.99)

In OPD, \( \mathcal{Q} \) is the set of probabilistic vectors \( q = [q_1; ...; q_L] \) specifying our measurement design, and \( A_q \) is the matrix of the mapping (2.97).

- a closeness \( C \) on the set \( \{1, ..., L\} \) (that is, a set \( C \) of pairs \( (i, j) \) with \( 1 \leq i, j \leq L \) such that \( (i, i) \in C \) for all \( i \leq L \) and \( (j, i) \in C \) whenever \( (i, j) \in C \)), and a positive integer \( K \).

In OPD, the closeness \( C \) is as strict as it could be – \( i \) is close to \( j \) if and only if \( i = j \). 17 and \( K \) is the total number of interviews in the poll.

We associate with \( q \in \mathcal{Q} \) and \( X_\ell, \ell \leq L \), nonempty convex compact sets \( M_q^\ell \) in the space \( \mathcal{M} \):

\[
M_q^\ell = \{A_q x : x \in X_\ell\}
\]

and hypotheses \( H^q_\ell \) on \( K \)-repeated stationary observations \( \omega^K = (\omega_1, ..., \omega_K) \), \( H^q_\ell \) stating that \( \omega_k, k = 1, ..., K \), are drawn, independently of each other, from a distribution \( \mu \in M_q^\ell, \ell = 1, ..., L \). Closeness \( C \) can be thought of as closeness on the collection of hypotheses \( H^q_1, H^q_2, ..., H^q_L \). Given \( q \in \mathcal{Q} \), we can use the construction from Section 2.5.2 in order to build the test \( \mathcal{T}^{K}_\phi \), deciding on the hypotheses \( H^q_\ell \) up to closeness \( C \), the \( C \)-risk of the test being the smallest allowed by the construction. Note that this \( C \)-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 \( C \)-risk of the associated test \( \mathcal{T}^{K}_\phi \).

### 2.7.3 Formulating the MD problem

By Proposition 2.30, the \( C \)-risk of the test \( \mathcal{T}^{K}_\phi \) 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 if \( (\ell, \ell') \in C \). For \( (\ell, \ell') \not\in C \) 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,

---

17 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 Gaussian observation scheme (see Section 2.4.5.1), restriction (2.99) does not restrain the dependence $A_q$ on $q$ at all (modulo the default constraint that $A_q$ is a continuous in $q \in Q$ $d \times n$ matrix), and

$$
\epsilon_{\ell \ell'}(q) = \exp\{K_{\text{Opt}_{\ell \ell'}}(q)\}
$$

where

$$
\text{Opt}_{\ell \ell'}(q) = \max_{x \in X_{\ell}, y \in X_{\ell'}} -\frac{1}{2}[A_q(x-y)]^T \Theta^{-1}[A_q(x-y)] (G_q)
$$

and $\Theta$ 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 (2.99) requires from $A_q x$ to be positive vector whenever $q \in Q$ and $x \in \mathcal{X}$, and

$$
\epsilon_{\ell \ell'}(q) = \exp\{K_{\text{Opt}_{\ell \ell'}}(q)\},
$$

where

$$
\text{Opt}_{\ell \ell'}(q) = \max_{x \in X_{\ell}, y \in X_{\ell'}} \sum_i \sqrt{|A_q x_i| |A_q y_i|} - \frac{1}{2} |A_q x_i| - \frac{1}{2} |A_q y_i| (P_q)
$$

- In the case of Discrete o.s. (see Section 2.4.5.3), restriction (2.99) requires from $A_q x$ to be a positive probabilistic vector whenever $q \in Q$ and $x \in \mathcal{X}$, and

$$
\epsilon_{\ell \ell'}(q) = [\text{Opt}_{\ell \ell'}(q)]^K (D_q)
$$

where

$$
\text{Opt}_{\ell \ell'}(q) = \max_{x \in X_{\ell}, y \in X_{\ell'}} \sum_i \sqrt{|A_q x_i| |A_q y_i|}.
$$

The summary of above observations is as follows. The norm $\|E^K\|_{2,2}$ – the quantity we are interested to minimize in $q \in Q$ – as a function of $q \in Q$ is of the form

$$
\Psi(q) = \psi\left(\frac{\{\text{Opt}_{\ell \ell'}(q) : (\ell, \ell') \notin \mathcal{C}\}}{\text{Opt}(q)}\right) \tag{2.100}
$$

where the outer function $\psi$ is an explicitly given real-valued function on $\mathbb{R}^N$ ($N$ is the cardinality of the set of pairs $(\ell, \ell')$, $1 \leq \ell, \ell' \leq L$, with $(\ell, \ell') \notin \mathcal{C}$) which is convex and nondecreasing in each argument. Indeed, denoting by $\Gamma(S)$ the spectral norm of $d \times d$ matrix $S$, note that $\Gamma$ 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.\footnote{The monotonicity follows from the fact that for an entrywise nonnegative $S$, we have $\|S\|_{2,2} = \max_{x, y} x^T S y : \|x\|_2 \leq 1, \|y\|_2 \leq 1 = \max_{x, y} x^T S y : \|x\|_2 \leq 1, \|y\|_2 \leq 1, x \geq 0, y \geq 0$.} $\psi(\cdot)$ is obtained from $\Gamma(S)$ by substituting for the entries $S_{\ell \ell'}$ of $S$, certain explicit everywhere convex, nonnegative and nondecreasing functions of variables $z = \{z_{\ell \ell'} : (\ell, \ell') \notin \mathcal{C}, 1 \leq \ell, \ell' \leq L\}$. Namely,

- when $(\ell, \ell') \in \mathcal{C}$, we set $S_{\ell \ell'}$ to zero;
- when $(\ell, \ell') \notin \mathcal{C}$, we set $S_{\ell \ell'} = \exp\{K z_{\ell \ell'}\}$ in the case of Gaussian and Poisson
o.s.’s, and set \( S_{\ell'\ell} = \max[0, z_{\ell'\ell}]^\mathcal{K} \), in the case of Discrete o.s.

As a result, we indeed get a convex and nondecreasing in every argument function \( \psi \) of \( z \in \mathbb{R}^N \).

Now, the Measurement Design problem we want to solve reads

\[
\operatorname{Opt} = \min_{q \in \mathcal{Q}} \psi(\overline{\operatorname{Opt}(q)}); \quad (2.101)
\]

As we remember, the entries in the inner function \( \overline{\operatorname{Opt}(q)} \) 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 (2.101), due to the already established convexity and monotonicity properties of \( \psi \), is a convex function of \( q \), meaning that (2.101) is a convex and thus efficiently solvable problem. On the other hand, when some of the entries in \( \overline{\operatorname{Opt}(q)} \) are nonconvex in \( q \), we can hardly expect (2.101) to be easy to solve. Unfortunately, convexity of the entries in \( \overline{\operatorname{Opt}(q)} \) 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 functions of \( A_q \) (not of \( q! \)) are concave rather than convex. Thus, the optimal values in the problems, as 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 can hardly be convex as functions of \( A_q \). And indeed, as a matter of fact, the MD problem usually is nonconvex and difficult to solve. We intend to consider a “Simple case” where this difficulty does not arise, i.e., the case where the objectives of the optimization problems specifying \( \operatorname{Opt}_{\ell'\ell}(q) \) are affine in \( q \). In this case, \( \operatorname{Opt}_{\ell'\ell}(q) \) 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

\[
A_q = \operatorname{Diag}\{Bq\}A, \quad (2.102)
\]

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. The Opinion Poll Design problem of Section 2.7.1 provides an instructive example of the Simple case of Measurement Design in Discrete o.s.: recall that in this problem the voting population is split into \( I \) groups, with \( i \)-th group constituting fraction \( \theta_i \) of the entire population. The distribution of voters’ preferences in the \( i \)-th group is represented by unknown \( L \)-dimensional probabilistic vector \( x_i = [x_{i1}; \ldots; x_{iL}] \) (\( L \) is the number of candidates, \( x_{i\ell} \) is the fraction of voters in \( i \)-th group intending to vote for \( \ell \)-th candidate), known to belong to a given con-
HYPOTHESIS TESTING

wey compact subset Π of the “positive part” \( \Delta_L^q = \{ x \in \mathbb{R}^L : x > 0, \sum x_i = 1 \} \) of the \( L \)-dimensional probabilistic simplex. We are given threshold \( \delta > 0 \) and want to decide on \( L \) hypotheses \( H_1, ..., H_L \), with \( H_\ell \) stating that the population-wide vector \( y = \sum_{i=1}^I \theta_i x^i \) of voters’ preferences belongs to the closed convex set

\[
Y_\ell = \left\{ y = \sum_{i=1}^I \theta_i x^i : x^i \in \Pi_i, 1 \leq i \leq I, y_\ell \geq y_{\ell'} + \delta, \forall (\ell' \neq \ell) \right\}.
\]

Note that \( Y_\ell \) 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

\[
\mathcal{X} = \{ x = [x^1; ...; x^I] : x^i \in \Pi_i, 1 \leq i \leq I \}.
\]

The \( k \)-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 \( \omega_k \) – is the pair \((i, \ell)\), where \( i \) is the group to which the respondent belongs, and \( \ell \) is the candidate preferred by the respondent.

This results in a sensing matrix \( A_q \), see (2.97), which is in the form of (2.102), namely,

\[
A_q = \text{Diag}\{q_1 I_L, q_2 I_L, ..., q_I I_L\}, \quad [q \in \Delta_I]
\]

and 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 \( \epsilon \) on our \( L \) hypotheses. Whether this goal is or is not achievable, depends on \( K \) and on \( A_q \). What we want, is to find \( q \) for which the above goal can be attained with as small \( K \) as possible; in the case in question, this reduces to solving, for various trial values of \( K \), problem (2.101), which under the circumstances is an explicit \textit{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 use \( \delta = 0.05 \), \( \epsilon = 0.01 \) and equal fractions \( \theta_i = I^{-1} \) for all groups. The sets \( \Pi_i, 1 \leq i \leq I \), are generated as follows: we pick at random a probabilistic vector \( \tilde{p} \) of dimension \( L \), and define \( \Pi_i \) as the intersection of the box \( \{ p : \tilde{p}_i - u_i \leq p_i \leq \tilde{p}_i + u_i \} \) centered at \( \tilde{p} \) with the probabilistic simplex \( \Delta_L \), 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:
Effect of measurement design: poll sizes required for 0.99-reliable winner prediction when \( q = \theta \) (column \( K_{\text{ini}} \)) and \( q = q_{\text{opt}} \) (column \( K_{\text{opt}} \)).

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 designs.\(^{19}\) On the other hand, the difference between \( K_{\text{ini}} \) and \( K_{\text{opt}} \) is fully due to the measurement design.

**Comparative drug study.** A related to OPD and perhaps more interesting Simple case of the Measurement Design in Discrete o.s. is as follows. Suppose that now, instead of \( L \) competing candidates running for an office we have \( L \) competing drugs, and population of patients the drugs are aimed at rather than the 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 categorical outcomes, like “strong positive effect,” “weak positive effect,” “negative effect,” and alike). Our goal is to organize a clinical study in order to decide 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 may only administer to a participant of the study a single drug of our choice and look at the result.

As in the OPD problem, we assume that the population of patients is split into \( I \) groups, with the \( i \)-th group comprising a fraction \( \theta_i \) of the entire population.

We model the situation as follows. We associate with a patient a Boolean vector of dimension \( 2L \), with the \( \ell \)-th entry in the vector equal to 1 or 0 depending on whether drug \( \ell \) has effect on the patient, and the \( (L+\ell) \)-th entry complementing the \( \ell \)-th one to 1 (that is, if \( \ell \)-th entry is \( \chi \), 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 \([x^1;\ldots;x^I]\) and assume that our a priori information allows to localize \( x \) in a closed convex subset \( X \) of the set

\[
\mathcal{Y} = \{ x = [x^1;\ldots;x^I] : x^i \geq 0, x^i + x^i_{L+\ell} = 1, 1 \leq i \leq I, 1 \leq \ell \leq L \}
\]

To which all our signals belong by construction. Note that the vector

\[
y = Bx = \sum_i \theta_i x^i
\]

\(^{19}\)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.
can be treated as “population-wise distribution of drug effects”: \( y_\ell, \ell \leq L \), is the fraction, in the entire population of patients, of those patients on whom drug \( \ell \) has effect, and \( y_{L+\ell} = 1 - y_\ell \). As a result, typical hypotheses related to comparison of the drugs, like “drug \( \ell \) has effect on a larger, at least by margin \( \delta \), fraction of patients than drug \( \ell' \),” 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 \( q_{d\ell} \), a pair \((i, \ell)\) 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 to the patient the drug \( \ell \) and check whether the drug has effect. Thus, a single observation is a triple \((i, \ell, \chi)\), where \( \chi = 0 \) if the administered drug has no effect on the patient, and \( \chi = 1 \) otherwise. The probability to get observation \((i, \ell, 1)\) is \( q_{d\ell} x^I_\ell \), and the probability to get observation \((i, \ell, 0)\) is \( q_{d\ell} x^I_{L+\ell} \). Thus, we arrive at the Discrete o.s. where the distribution \( \mu \) of observations is of the form \( \mu = A_q x \), with the rows in \( A_q \) indexed by triples \( \omega = (i, \ell, \chi) \in \Omega := \{1, 2, \ldots, I\} \times \{1, 2, \ldots, L\} \times \{0, 1\} \) and given by

\[
(A_q[x^1; \ldots; x^I])_{i,\ell,\chi} = \begin{cases} 
q_{d\ell} x^I_\ell & \chi = 1, \\
q_{d\ell} x^I_{L+\ell} & \chi = 0.
\end{cases}
\]

Specifying the set \( Q \) of admissible measurement designs as a closed convex subset of the set of all non-vanishing discrete probability distributions on the set \( \{1, 2, \ldots, I\} \times \{1, 2, \ldots, L\} \), we find ourselves in the Simple case of Discrete o.s., as defined by (2.102), and \( A_q x \) is a probabilistic vector whenever \( q \in Q \) 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

\[
A_q = \text{Diag}\{Bq\} A,
\]

where \( A \) is some fixed \( d \times n \) matrix, and \( B \) is some fixed \( d \times (\dim q) \) matrix such that \( B_q \) is positive whenever \( q \in Q \). On the top of this, we should ensure that when \( q \in Q \) and \( x \in \mathcal{X} \), \( A_q x \) is a positive vector; this amounts to some restrictions linking \( Q, \mathcal{X}, A, \) 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 \( \omega \in \mathbb{R}^d \) with independent entries \( \omega_i \sim \text{Poisson}(\mu_i) \), \( 1 \leq i \leq d \), where the vector of parameters \( \mu = [\mu_1; \ldots; \mu_d] \) of the Poisson distributions is the linear image \( \mu = A \lambda \) of unknown “signal” (tracer’s density in patient’s body) \( \lambda \) belonging to some known subset \( \Lambda \) of \( \mathbb{R}_0^d \), with entrywise nonnegative matrix \( A \). Our goal is to make inferences about \( \lambda \). 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 Figure 2.7, 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 \( \omega \) can be split into blocks \( \omega^i, i = 1, \ldots, I, \) of data acquired during the \( i \)-th period, \( 1 \leq i \leq I \). On closer
inspection, the corresponding block $\mu^i$ in $\mu$ 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 the $i$-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, 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 $\Theta$ is diagonal, and

$$A_q = \text{Diag}\{\sqrt{q_1},...,\sqrt{q_d}\} A$$

(2.103)

where $A$ is a fixed $d \times n$ matrix, and $q$ runs through a convex compact subset of $\mathbb{R}_+^d$.

It turns out that there are situations where assumption (2.103) makes perfect sense. Let us start with preamble. In Gaussian o.s.

$$\omega = A x + \xi$$

(2.104)

$[A \in \mathbb{R}^{d \times n}, \xi \sim \mathcal{N}(0, \Sigma), \Sigma = \text{Diag}\{\sigma_1^2,...,\sigma_d^2\}]$

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 the $i$-th sensor on time interval $\Delta_i$. The deterministic component of the measurement registered by sensor $i$ is the integral of the corresponding input taken over $\Delta_i$, 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, what matters is

---

$^{20}$T cannot be too large: aside from other considerations, the tracer disintegrates, and its density can be considered as nearly constant only on a properly restricted time horizon.
that when the white noise affecting $i$-th sensor is integrated over a time interval $\Delta_i$, the result is a Gaussian random variable with zero mean and variance $\sigma_i^2|\Delta_i|$ (here $|\Delta_i|$ is the length of $\Delta_i$), 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 (2.104). 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 (2.104) 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

$$z_i = \sigma_i \sqrt{q_i} \xi_i + q_i \sum_j a_{ij} x_j, \quad i = 1, ..., d,$$

with independent of each other standard (zero mean, unit variance) Gaussian noises $\xi_i$. Now, since we know $q_i$, we can scale the latter observations by making the standard deviation of the noisy component the same $\sigma_i$ as in the reference case. Specifically, we lose nothing when assuming that our observations are

$$\omega_i = z_i/\sqrt{q_i} = \frac{\sigma_i \xi_i}{\sqrt{q_i}} + \sqrt{q_i} \sum_j a_{ij} x_j,$$

or, equivalently,

$$\omega = \xi + \text{Diag}\{\sqrt{q_1}, ..., \sqrt{q_d}\} A x, \quad \xi \sim \mathcal{N}(0, \text{Diag}\{\sigma_1^2, ..., \sigma_d^2\}) \quad [A = [a_{ij}]]$$

where $q$ runs through a convex compact subset $Q$ of the simplex $\{q \in \mathbb{R}^d_+ : \sum_i q_i = d\}$. 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 (2.103), and, as a result, end up with an 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 $F$ is comprised of affine functions on the corresponding observation space $\Omega$ ($\mathbb{R}^d$ for Gaussian o.s., and $\mathbb{Z}_+^d \subset \mathbb{R}^d$ for Poisson o.s.), but is not so for the Discrete o.s. – in the latter case, $\Omega = \{1, \ldots, d\}$, and $F$ is comprised of all functions on $\Omega$, while “affine functions on $\Omega = \{1, \ldots, d\}$” merely make no sense. Note, however, that we can encode (and from now on this is what we do) the points $i = 1, \ldots, d$ of a $d$-element set by basic orths $e_i = [0; \ldots; 0; 1; 0; \ldots; 0] \in \mathbb{R}^d$, thus making our observation space $\Omega$ a subset of $\mathbb{R}^d$. With this encoding, every real valued function on $\{1, \ldots, d\}$ becomes restriction on $\Omega$ 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, in our context the following two properties of simple o.s.’s are essential:

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 requisite.

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, cannot claim anymore, is that the tests associated with such 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 [134, 135].

\subsection{Situation}

In what follows, we fix an observation space $\Omega = \mathbb{R}^d$, and let $\mathcal{P}_j$, $1 \leq j \leq J$, be given families of probability distributions on $\Omega$. Put broadly, our goal still is, given a random observation $\omega \sim P$, where $P \in \bigcup_{j \leq J} \mathcal{P}_j$, to decide upon the hypotheses $H_j : P \in \mathcal{P}_j$, $j = 1, \ldots, J$. 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.

\subsubsection{Preliminaries: Regular data and associated families of distributions}

\begin{definition} A. Regular data is as a triple $\mathcal{H}, \mathcal{M}, \Phi(\cdot, \cdot)$, where

\begin{itemize}
  \item $\mathcal{H}$ is a nonempty closed convex set in $\Omega = \mathbb{R}^d$ symmetric w.r.t. the origin,
  \item $\mathcal{M}$ is a closed convex set in some $\mathbb{R}^n$,
  \item $\Phi(h; \mu) : \mathcal{H} \times \mathcal{M} \to \mathbb{R}$ is a continuous function convex in $h \in \mathcal{H}$ and concave
\end{itemize}
\end{definition}
in $\mu \in \mathcal{M}$.

**B. Regular data** $\mathcal{H}, \mathcal{M}, \Phi(\cdot, \cdot)$ define two families of probability distributions on $\Omega$:
- the family of **regular** distributions
  \[ \mathcal{R} = \mathcal{R}[\mathcal{H}, \mathcal{M}, \Phi] \]
  comprised of all probability distributions $P$ on $\Omega$ such that
  \[ \forall h \in \mathcal{H}, \exists \mu \in \mathcal{M} : \ln \left( \int_{\Omega} \exp\{h^T \omega\} P(d\omega) \right) \leq \Phi(h; \mu). \]
- the family of **simple** distributions
  \[ \mathcal{S} = \mathcal{S}[\mathcal{H}, \mathcal{M}, \Phi] \]
  comprised of probability distributions $P$ on $\Omega$ such that
  \[ \exists \mu \in \mathcal{M} : \forall h \in \mathcal{H} : \ln \left( \int_{\Omega} \exp\{h^T \omega\} P(d\omega) \right) \leq \Phi(h; \mu). \quad (2.105) \]

For a probability distribution $P \in \mathcal{S}[\mathcal{H}, \mathcal{M}, \Phi]$, every $\mu \in \mathcal{M}$ satisfying (2.105) is referred to as a parameter of $P$ w.r.t. $\mathcal{S}$. Note that a distribution may have many different from each other parameters.

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 $\omega_1, \ldots, \omega_t$ 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 $\mathcal{H} = \Omega = \mathbb{R}^d$, $\mathcal{M}$ be a closed convex subset of the set $\mathcal{G}_d = \{ \mu = (\theta, \Theta) : \theta \in \mathbb{R}^d, \Theta \in \mathbb{S}_+^d \}$, where $\mathbb{S}_+^d$ is cone of positive semidefinite matrices in the space $\mathbb{S}^d$ of symmetric $d \times d$ matrices, and let
\[ \Phi(h; \theta, \Theta) = \theta^T h + \frac{1}{2} h^T \Theta h. \]

Recall that a distribution $P$ on $\Omega = \mathbb{R}^d$ is called sub-Gaussian with sub-Gaussianity parameters $\theta \in \mathbb{R}^d$ and $\Theta \in \mathbb{S}_+^d$, if
\[ \mathbf{E}_{\omega \sim P}\{\exp\{h^T \omega\}\} \leq \exp\{\theta^T h + \frac{1}{2} h^T \Theta h\} \quad \forall h \in \mathbb{R}^d. \quad (2.106) \]
Whenever this is the case, $\theta$ is the expected value of $P$. We shall use the notation $\xi \sim \mathcal{S}\mathcal{G}(\theta, \Theta)$ as a shortcut for the sentence “random vector $\xi$ is sub-Gaussian with parameters $\theta, \Theta$.” It is immediately seen that when
ξ ∼ \mathcal{N}(\theta, \Theta), we also have ξ ∼ \mathcal{SG}(\theta, \Theta), and (2.106) in this case is an identity rather than inequality.

With Φ as above, \mathcal{S}[\mathcal{H}, \mathcal{M}, \Phi] clearly contains every sub-Gaussian distribution \mathcal{P} on \mathbb{R}^d with sub-Gaussianity parameters (forming a parameter of \mathcal{P} w.r.t. \mathcal{S}) from \mathcal{M}. In particular, \mathcal{S}[\mathcal{H}, \mathcal{M}, \Phi] contains all Gaussian distributions \mathcal{N}(\theta, \Theta) with (\theta, \Theta) ∈ \mathcal{M}.

2.8.1.2.B. Poisson distributions: Let \mathcal{H} = \Omega = \mathbb{R}^d, let \mathcal{M} be a closed convex subset of \mathbb{R}^d, and let

\[ \Phi(\mathbf{h} = [h_1; \ldots; h_d]; \mu = [\mu_1; \ldots; \mu_d]) = \sum_{i=1}^{d} \mu_i [\exp\{h_i\} - 1] : \mathcal{H} \times \mathcal{M} \to \mathbb{R}. \]

The family \mathcal{S} = \mathcal{S}[\mathcal{H}, \mathcal{M}, \Phi] contains all Poisson distributions \text{Poisson}[\mu] with vectors \mu of parameters belonging to \mathcal{M}; here \text{Poisson}[\mu] is the distribution of random \text{d}-dimensional vector with independent of each other entries, \text{i}-th entry being Poisson random variable with parameter \mu_i. \mu is a parameter of \text{Poisson}[\mu] w.r.t. \mathcal{S}.

2.8.1.2.C. Discrete distributions. Consider a discrete random variable taking values in \text{d}-element set \{1, 2, ..., \text{d}\}, and let us think of such a variable as of random variable taking values \mathbf{e}_i ∈ \mathbb{R}^d, i = 1, ..., \text{d}, where \mathbf{e}_i = [0; \ldots; 0; 1; 0; \ldots; 0] (1 in position i) are standard basic orths in \mathbb{R}^d. Probability distribution of such a variable can be identified with a point \mu = [\mu_1; \ldots; \mu_d] from the \text{d}-dimensional probabilistic simplex

\[ \Delta_d = \{ \nu ∈ \mathbb{R}_+^d : \sum_{i=1}^{d} \nu_i = 1 \}, \]

where \mu_i is the probability for the variable to take value \mathbf{e}_i. With these identifications, setting \mathcal{H} = \mathbb{R}^d, specifying \mathcal{M} as a closed convex subset of \Delta_d and setting

\[ \Phi(\mathbf{h} = [h_1; \ldots; h_d]; \mu = [\mu_1; \ldots; \mu_d]) = \ln \left( \sum_{i=1}^{d} \mu_i \exp\{h_i\} \right), \]

the family \mathcal{S} = \mathcal{S}[\mathcal{H}, \mathcal{M}, \Phi] contains distributions of all discrete random variables taking values in \{1, ..., \text{d}\} with probabilities \mu_1, ..., \mu_\text{d} comprising a vector from \mathcal{M}. This vector is a parameter of the corresponding distribution w.r.t. \mathcal{S}.

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 \text{X} ⊂ \Omega = \mathbb{R}^d, and let

\[ \phi_X(\mathbf{h}) = \max_{x ∈ \text{X}} h^T x \]

be the support function of \text{X}. We have the following result (to be refined in Section 2.8.1.3):

**Proposition 2.37.** For every \mathcal{P} ∈ \mathcal{P}[X] it holds

\[ \forall \mathbf{h} ∈ \mathbb{R}^d : \ln \left( \int_{\mathbb{R}^d} \exp\{h^T \omega\} \mathcal{P}(d\omega) \right) ≤ h^T e[\mathcal{P}] + \frac{1}{\phi} [\phi_X(\mathbf{h}) + \phi_X(-\mathbf{h})]^2, \quad (2.107) \]

where \( e[P] = \int_{\mathbb{R}^d} \omega P(d\omega) \) is the expectation of \( P \), and the function in the right hand side of (2.107) is convex. As a result, setting

\[
\mathcal{H} = \mathbb{R}^d, \quad \mathcal{M} = X, \quad \Phi(h; \mu) = h^T \mu + \frac{1}{8} [\phi_X(h) + \phi_X(-h)]^2
\]

we obtain a regular data such that \( \mathcal{P}[X] \subset S[\mathcal{H}, \mathcal{M}, \Phi], e[p] \) being a parameter of a distribution \( P \in \mathcal{P}[X] \) w.r.t. \( S \).

For proof, see Section 2.11.4

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 \( \mathcal{H}_\ell \subset \Omega_\ell = \mathbb{R}^{d_\ell}, \mathcal{M}_\ell \subset \mathbb{R}^{n_\ell}, \Phi_\ell(h; \mu_\ell) : \mathcal{H}_\ell \times \mathcal{M}_\ell \to \mathbb{R} \) be given. Let us set

\[
\Omega = \Omega_1 \times \ldots \times \Omega_L = \mathbb{R}^d, \quad d = d_1 + \ldots + d_L,
\]

\[
\mathcal{H} = \mathcal{H}_1 \times \ldots \times \mathcal{H}_L = \{h = [h^1; \ldots; h^L] : h^\ell \in \mathcal{H}_\ell, \ell \leq L\},
\]

\[
\mathcal{M} = \mathcal{M}_1 \times \ldots \times \mathcal{M}_L = \{\mu = [\mu^1; \ldots; \mu^L] : \mu^\ell \in \mathcal{M}^\ell, \ell \leq L\} \subset \mathbb{R}^n, \quad n = n_1 + \ldots + n_L
\]

\[
\Phi(h = [h^1; \ldots; h^L]; \mu = [\mu^1; \ldots; \mu^L]) = \sum_{\ell=1}^L \Phi_\ell(h^\ell; \mu^\ell) : \mathcal{H} \times \mathcal{M} \to \mathbb{R}.
\]

Then \( \mathcal{H} \) is a symmetric w.r.t. the origin closed convex set in \( \Omega = \mathbb{R}^d, \mathcal{M} \) is a nonempty closed convex set in \( \mathbb{R}^n, \Phi : \mathcal{H} \times \mathcal{M} \to \mathbb{R} \) is a continuous convex-concave function, and clearly

- the family \( \mathcal{R}[\mathcal{H}, \mathcal{M}, \Phi] \) contains all product-type distributions \( P = P_1 \times \ldots \times P_L \) on \( \Omega = \Omega_1 \times \ldots \times \Omega_L \) with \( P_\ell \in \mathcal{R}[\mathcal{H}_\ell, \mathcal{M}_\ell, \Phi_\ell], \mathcal{R} \leq \mathcal{L} \);
- the family \( \mathcal{S} = \mathcal{S}[\mathcal{H}, \mathcal{M}, \Phi] \) contains all product-type distributions \( P = P_1 \times \ldots \times P_L \) on \( \Omega = \Omega_1 \times \ldots \times \Omega_L \) with \( P_\ell \in \mathcal{S}_\ell = \mathcal{S}[\mathcal{H}_\ell, \mathcal{M}_\ell, \Phi_\ell], 1 \leq \ell \leq L \), a parameter of \( P \) w.r.t. \( \mathcal{S} \) being the vector of parameters of \( P_\ell \) w.r.t. \( \mathcal{S}_\ell \).

2.8.1.3.B. Mixing. For \( 1 \leq \ell \leq L \), let regular data \( \mathcal{H}_\ell \subset \Omega = \mathbb{R}^{d_\ell}, \mathcal{M}_\ell \subset \mathbb{R}^{n_\ell}, \Phi_\ell(h; \mu_\ell) : \mathcal{H}_\ell \times \mathcal{M}_\ell \to \mathbb{R} \) be given, with compact \( \mathcal{M}_\ell \). Let also \( \nu_\ell = [\nu^1; \ldots; \nu^L] \) 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 \( \nu \)-mixture of distributions \( P_1, \ldots, P_L \) defined as the distribution of random vector \( \omega \sim \Omega \) generated as follows: we draw at random, from probability distribution \( \nu \) on \( \{1, \ldots, L\} \), index \( \ell \), and then draw \( \omega \) at random from the distribution \( P_\ell \). Finally, let \( \mathcal{P} \) be the set of all probability distributions on \( \Omega \) which can be obtained as \( \Pi[P^L, \nu] \) from the outlined tuples \( P^L \) and vectors \( \nu \) running through the probabilistic simplex \( \Delta_L = \{\mu \in \mathbb{R}^L : \nu \geq 0, \sum_\ell \nu^\ell = 1\} \).

Let us set

\[
\mathcal{H} = \bigcap_{\ell=1}^L \mathcal{H}_\ell, \quad \Psi_\ell(h) = \max_{\mu_\ell \in \mathcal{M}_\ell} \Phi_\ell(h; \mu_\ell) : \mathcal{H}_\ell \to \mathbb{R}, \quad \Phi(h; \nu) = \ln \left( \sum_{\ell=1}^L \nu^\ell \exp(\Psi_\ell(h)) \right) : \mathcal{H} \times \Delta_L \to \mathbb{R}.
\]
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

$$P = \Pi[P^L, \nu] \Rightarrow \ln \left( \int_\Omega e^{h^T \omega} P(d\omega) \right) \leq \Phi(h; \nu) \quad \forall h \in \mathcal{H},$$

(2.108)

implying that $P \subset \mathcal{S}[\mathcal{H}, \mathbf{\Delta}_L, \Phi], \nu$ being a parameter of $P = \Pi[P^L, \nu] \in \mathcal{P}$.

Indeed, (2.108) is readily given by the fact that for $P = \Pi[P^L, \nu] \in \mathcal{P}$ and $h \in \mathcal{H}$ it holds

$$\ln \left( E_{\omega \sim P} \left\{ e^{h^T \omega} \right\} \right) = \ln \left( \sum_{\ell=1}^L \nu_\ell E_{\omega \sim P_\ell} \left\{ e^{h^T \omega} \right\} \right) \leq \ln \left( \sum_{\ell=1}^L \nu_\ell \exp \{ \Psi_\ell(h) \} \right) = \Phi(h; \nu),$$

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 built 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 $\nu$, while the parameters of $P_\ell$’s disappear. In some situations, this makes the 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

$$\Phi(h; \mu_1, ..., \mu_L) = \sum_{\ell=1}^L \bar{\nu}_\ell \Phi_\ell(h; \mu_\ell) : \mathcal{H} \times (\mathcal{M}_1 \times ... \times \mathcal{M}_L) \to \mathbb{R}.$$ 

Let $d \times d$ matrix $Q \succeq 0$ satisfy

$$\left( \Phi_\ell(h; \mu_\ell) - \Phi(h; \mu_1, ..., \mu_L) \right)^2 \leq h^T Q h \quad \forall (h \in \mathcal{H}, \ell \leq L, \mu \in \mathcal{M}_1 \times ... \times \mathcal{M}_L),$$

(2.109)

and let

$$\Phi(h; \mu_1, ..., \mu_L) = \frac{3}{4} h^T Q h + \Phi(h; \mu_1, ..., \mu_L) : \mathcal{H} \times (\mathcal{M}_1 \times ... \times \mathcal{M}_L) \to \mathbb{R}. $$

(2.110)

$\Phi$ 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 a regular data.

**Proposition 2.38.** In the just defined situation, denoting by $\mathcal{P}$ the family of all probability distributions $P = \Pi[P^L, \bar{\nu}], \Pi$ stemming from tuples

$$P^L = \left\{ P_\ell \in \mathcal{S}[\mathcal{H}_\ell, \mathcal{M}_\ell, \Phi_\ell] \right\}_{\ell=1}^L,$$

(2.111)

one has

$$\mathcal{P}_\bar{\nu} \subset \mathcal{S}[\mathcal{H}, \mathcal{M}_1 \times ... \times \mathcal{M}_L, \Phi].$$

As a parameter of distribution $P = \Pi[P^L, \bar{\nu}] \in \mathcal{P}_\bar{\nu}$ with $P^L$ as in (2.111), one can take $\mu^L = [\mu_1; ..., \mu_L]$. 
Proof. It is easily seen that
\[ e^a \leq a + e^{\frac{a^2}{2}}, \forall a. \]
As a result, when \( a_\ell, \ell = 1, ..., L, \) satisfy \( \sum_\ell \bar{\nu}_\ell a_\ell = 0, \) we have
\[ \sum_\ell \bar{\nu}_\ell e^{a_\ell} \leq \sum_\ell \bar{\nu}_\ell a_\ell + \sum_\ell \bar{\nu}_\ell e^{\frac{a_\ell^2}{2}} \leq e^{\frac{a^2}{2}} \max_\ell a_\ell^2. \] (2.112)

Now let \( P^L \) be as in (2.111), and let \( h \in \mathcal{H} = \bigcap_\ell \mathcal{H}_\ell. \) Setting \( P = \Pi[P^L, \bar{\nu}], \) we have
\[
\ln \left( \int_{\Omega} e^{h^T \omega} P(d\omega) \right) = \ln \left( \sum_\ell \bar{\nu}_\ell \int_{\Omega} e^{h^T \omega} P_\ell(d\omega) \right) = \ln \left( \sum_\ell \bar{\nu}_\ell \exp \{ \Phi_\ell(h, \mu_\ell) \} \right)
\leq \Phi(h; \mu_1, ..., \mu_L) + \frac{3}{2} \max_\ell [\Phi_\ell(h, \mu_\ell) - \Phi(h; \mu_1, ..., \mu_L)]^2 \leq \Phi(h; \mu_1, ..., \mu_L),
\]
where \( a \) is given by (2.112) as applied to \( a_\ell = \Phi_\ell(h, \mu_\ell) - \bar{\Phi}(h; \mu_1, ..., \mu_L), \) and \( b \) is due to (2.109), (2.110). The resulting inequality, which holds true for all \( h \in \mathcal{H}, \) is all we need. \( \square \)

2.8.1.3.C. i.i.d. summation. Let \( \Omega = \mathbb{R}^d \) be an observation space, \((\mathcal{H}, \mathcal{M}, \Phi)\) be a 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 \( \Omega \) by setting
\[
\mathcal{H}_\lambda = \{ h \in \Omega : ||\lambda||_\infty h \in \mathcal{H} \}, \quad \Phi_\lambda(h; \mu) = \sum_{\ell=1}^L \Phi(\lambda_\ell h; \mu) : \mathcal{H}_\lambda \times \mathcal{M} \rightarrow \mathbb{R}.
\]
Now, given a probability distribution \( P \) on \( \Omega, \) we can associate with it and with the above \( \lambda \) a new probability distribution \( P^\lambda \) on \( \Omega \) as follows: \( P^\lambda \) is the distribution of \( \sum_\ell \lambda_\ell \omega_\ell, \) where \( \omega_1, \omega_2, ..., \omega_L \) are drawn, independently of each other, from \( P, \) each with parameter \( \mu. \) An immediate observation is that the data \((\mathcal{H}_\lambda, \mathcal{M}, \Phi_\lambda)\) is regular, and
\[
\text{whenever a probability distribution } P \text{ belongs to } \mathcal{S}[\mathcal{H}, \mathcal{M}, \Phi], \text{ the distribution } P^\lambda \text{ belongs to } \mathcal{S}[\mathcal{H}_\lambda, \mathcal{M}, \Phi_\lambda], \text{ and every parameter of } P \text{ is a parameter of } P^\lambda. \]
In particular, when \( \omega \sim P \in \mathcal{S}[\mathcal{H}, \mathcal{M}, \Phi] \) the distribution \( P^L \) of the sum of \( L \) independent copies of \( \omega \) belongs to \( \mathcal{S}[\mathcal{H}, \mathcal{M}, \mathcal{L}\Phi]. \)

2.8.1.3.D. Semi-direct summation. For \( 1 \leq \ell \leq L, \) let regular data \( \mathcal{H}_\ell \subset \Omega_\ell = \mathbb{R}^{d_\ell}, \mathcal{M}_\ell, \Phi_\ell \) be given. To avoid complications, we assume that for every \( \ell, \)
- \( \mathcal{H}_\ell = \Omega_\ell, \)
- \( \mathcal{M}_\ell \) is bounded.
Let also an \( \epsilon > 0 \) be given. We assume that \( \epsilon \) is small, namely, \( L\epsilon < 1. \) Let us aggregate the given regular data into a new one by setting
\[
\mathcal{H} = \Omega := \Omega_1 \times ... \times \Omega_L = \mathbb{R}^d, \quad d = d_1 + ... + d_L, \quad \mathcal{M} = \mathcal{M}_1 \times ... \times \mathcal{M}_L.
\]
and let us define function \( \Phi(h; \mu) : \Omega^d \times \mathcal{M} \to \mathbb{R} \) as follows:

\[
\Phi(h = [h^1; \ldots; h^L]; \mu = [\mu^1; \ldots; \mu^L]) = \inf_{\lambda \in \Delta^c} \sum_{\ell=1}^L \lambda^\ell \Phi^\ell(h^\ell/\lambda^\ell; \mu^\ell),
\]

\[
\Delta^c = \{ \lambda \in \mathbb{R}^d : \lambda_\ell \geq \epsilon \forall \ell \land \sum_{\ell=1}^L \lambda^\ell = 1 \}. 
\]  

(2.113)

For evident reasons, the infimum in the description of \( \Phi \) is achieved, and \( \Phi \) is continuous. In addition, \( \Phi \) is convex in \( h \in \mathbb{R}^d \) and concave in \( \mu \in \mathcal{M} \). Postponing for a moment verification, the consequences are that \( \mathcal{H} = \Omega = \mathbb{R}^d \), \( \mathcal{M} \) and \( \Phi \) form a regular data. We claim that

Whenever \( \omega = [\omega^1; \ldots; \omega^L] \) is a random variable taking values in \( \Omega = \mathbb{R}^{d_1} \times \ldots \times \mathbb{R}^{d_L} \), and the marginal distributions \( P_\ell \), \( 1 \leq \ell \leq L \), of \( \omega \) belong to the families \( \mathcal{S}_\ell = \mathcal{S}^{\mathbb{R}^{d_\ell}, \mathcal{M}_\ell, \Phi_\ell} \) for all \( 1 \leq \ell \leq L \), the distribution \( P \) of \( \omega \) belongs to \( \mathcal{S} = \mathcal{S}^{\mathbb{R}^d, \mathcal{M}, \Phi} \), a parameter of \( P \) w.r.t. \( \mathcal{S} \) being the vector comprised of parameters of \( P_\ell \) w.r.t. \( \mathcal{S}_\ell \).

Indeed, since \( P_\ell \in \mathcal{S}[\mathbb{R}^{d_\ell}, \mathcal{M}_\ell, \Phi_\ell] \), there exists \( \hat{\mu}^\ell \in \mathcal{M}_\ell \) such that

\[
\ln(\mathbb{E}_{\omega \sim P_\ell} \{ \exp(g^T \omega^\ell) \}) \leq \Phi_\ell(g; \hat{\mu}^\ell) \ \forall g \in \mathbb{R}^{d_\ell}.
\]

Let us set \( \hat{\mu} = [\hat{\mu}^1; \ldots; \hat{\mu}^L] \), and let \( h = [h^1; \ldots; h^L] \in \Omega \) be given. We can find \( \lambda \in \Delta^c \) such that

\[
\Phi(h; \hat{\mu}) = \sum_{\ell=1}^L \lambda^\ell \Phi^\ell(h^\ell/\lambda^\ell; \hat{\mu}^\ell).
\]

Applying the Hölder inequality, we get

\[
\mathbb{E}_{[\omega^1; \ldots; \omega^L] \sim P} \left\{ \exp \left( \sum_{\ell} h^\ell T^\ell \omega^\ell \right) \right\} \leq \prod_{\ell=1}^L \left( \mathbb{E}_{[\omega^1; \ldots; \omega^L] \sim P} \left\{ |h^\ell|^T \omega^\ell / \lambda^\ell \right\} \right)^{\lambda^\ell},
\]

whence

\[
\ln \left( \mathbb{E}_{[\omega^1; \ldots; \omega^L] \sim P} \left\{ \exp \left( \sum_{\ell} |h^\ell|^T \omega^\ell \right) \right\} \right) \leq \sum_{\ell=1}^L \lambda^\ell \Phi^\ell(h^\ell/\lambda^\ell; \hat{\mu}^\ell) = \Phi(h; \hat{\mu}).
\]

We see that

\[
\ln \left( \mathbb{E}_{[\omega^1; \ldots; \omega^L] \sim P} \left\{ \exp \left( \sum_{\ell} |h^\ell|^T \omega^\ell \right) \right\} \right) \leq \Phi(h; \hat{\mu}) \ \forall h \in \mathcal{H} = \mathbb{R}^d,
\]

and thus \( P \in \mathcal{S}[\mathbb{R}^d, \mathcal{M}, \Phi] \), as claimed.

It remains to verify that the function \( \Phi \) defined by (2.113) indeed is convex in \( h \in \mathbb{R}^d \) and concave in \( \mu \in \mathcal{M} \). Concavity in \( \mu \) is evident. Further, functions \( \lambda^\ell \Phi^\ell(h^\ell/\lambda^\ell; \mu^\ell) \) (as perspective transformation of convex functions \( \Phi^\ell(\cdot; \mu^\ell) \)) are jointly convex in \( \lambda \) and \( h^\ell \), and so is \( \Psi(\lambda, h; \mu) = \sum_{\ell=1}^L \lambda^\ell \Phi^\ell(h^\ell/\lambda^\ell, \mu^\ell) \). Thus \( \Phi(\cdot; \mu) \), obtained by partial minimization of \( \Psi \) in \( \lambda \), indeed is convex.

**2.8.1.3.4. Affine image.** Let \( \mathcal{H}, \mathcal{M}, \Phi \) be a regular data, \( \Omega \) be the embedding space of \( \mathcal{H} \), and \( x \mapsto Ax + a \) be an affine mapping from \( \Omega \) to \( \Omega = \mathbb{R}^d \), and let us
set
\[ \mathcal{H} = \{ h \in \mathbb{R}^d : A^T h \in \mathcal{H} \}, \quad \mathcal{M} = \mathcal{M}, \quad \tilde{\Phi}(h; \mu) = \Phi(A^T h; \mu) + a^T h : \mathcal{H} \times \mathcal{M} \to \mathbb{R}. \]

Note that \( \mathcal{H}, \mathcal{M} \) and \( \tilde{\Phi} \) is a regular data. It is immediately seen that

\[ \text{Whenever the probability distribution } P \text{ of a random variable } \omega \text{ belongs to } \mathcal{R}[\mathcal{H}, \mathcal{M}, \tilde{\Phi}] \text{ (or belongs to } \mathcal{S}[\mathcal{H}, \mathcal{M}, \tilde{\Phi}], \text{ the distribution } \tilde{P}[P] \text{ of the random variable } \tilde{\omega} = A\omega + a \text{ belongs to } \mathcal{R}[\mathcal{H}, \mathcal{M}, \tilde{\Phi}] \text{ (respectively, belongs to } \mathcal{S}[\mathcal{H}, \mathcal{M}, \tilde{\Phi}], \text{ and every parameter of } P \text{ is a parameter of } \tilde{P}[P]. \]

2.8.1.3.F. Incorporating support information. Consider the situation as follows. We are given a regular data \( \mathcal{H} \subset \Omega = \mathbb{R}^d, \mathcal{M}, \tilde{\Phi} \) and are interested in a family \( \mathcal{P} \) of distributions known to belong to \( \mathcal{S}[\mathcal{H}, \mathcal{M}, \tilde{\Phi}] \). In addition, we know that all distributions \( P \) from \( \mathcal{P} \) are supported on a given closed convex set \( X \subset \mathbb{R}^d \). How could we incorporate this domain information to pass from the family \( \mathcal{S}[\mathcal{H}, \mathcal{M}, \tilde{\Phi}] \) 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 \( \mathcal{H} = \Omega \). When denoting by \( \phi_X(\cdot) \) the support function of \( X \) and selecting somehow a closed convex set \( G \subset \mathbb{R}^d \) containing the origin, let us set

\[ \tilde{\Phi}(h; \mu) = \inf_{g \in G} \left[ \Phi^+(h, g; \mu) := \Phi(h - g; \mu) + \phi_X(g) \right], \]

where \( \Phi(h; \mu) : \mathbb{R}^d \times \mathcal{M} \to \mathbb{R} \) is the continuous convex-concave function participating in the original regular data. Assuming that \( \tilde{\Phi} \) is real-valued and continuous on the domain \( \mathbb{R}^d \times \mathcal{M} \) (which definitely is the case when \( G \) is a compact set such that \( \phi_X \) is finite and continuous on \( G \)), note that \( \tilde{\Phi} \) is convex-concave on this domain, so that \( \mathbb{R}^d, \mathcal{M}, \tilde{\Phi} \) is a regular data. We claim that

\[ \text{The family } \mathcal{S}[\mathbb{R}^d, \mathcal{M}, \tilde{\Phi}] \text{ contains } \mathcal{P}, \text{ provided the family } \mathcal{S}[\mathbb{R}^d, \mathcal{M}, \Phi] \text{ does so, and the first of these two families is smaller than the second one.} \]

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 \( c \in \mathbb{R}^d \) it holds
\[ \ln \left( \int_{\mathbb{R}^d} \exp\{c^T \omega\} P(d\omega) \right) \leq \Phi(c; \mu_P). \]

On the other hand, for every \( g \in G \) we have \( \phi_X(g) - g^T \omega \geq 0 \) on the support of \( P \), whence for every \( h \in \mathbb{R}^d \) one has
\[ \ln \left( \int_{\mathbb{R}^d} \exp\{h^T \omega\} P(d\omega) \right) \leq \ln \left( \int_{\mathbb{R}^d} \exp\{h^T \omega + \phi_X(g) - g^T \omega\} P(d\omega) \right) \]
\[ \leq \phi_X(g) + \Phi(h - g; \mu_P). \]

Since the resulting inequality holds true for all \( g \in G \), we get
\[ \ln \left( \int_{\mathbb{R}^d} \exp\{h^T \omega\} P(d\omega) \right) \leq \tilde{\Phi}(h; \mu_P) \quad \forall h \in \mathbb{R}^d, \]

implying that \( P \in \mathcal{S}[\mathbb{R}^d, \mathcal{M}, \tilde{\Phi}] \); because \( P \in \mathcal{P} \) is arbitrary, the first part of the claim is justified. The inclusion \( \mathcal{S}[\mathbb{R}^d, \mathcal{M}, \tilde{\Phi}] \subset \mathcal{S}[\mathbb{R}^d, \mathcal{M}, \Phi] \) is readily given by the
inequality $\hat{\Phi} \leq \Phi$, and the latter is due to $\hat{\Phi}(h, \mu) \leq \Phi(h - 0, \mu) + \phi_X(0)$.

Illustration: distributions with bounded support revisited. In Section 2.8.1.2, given a convex compact set $X \subset \mathbb{R}^d$ with support function $\phi_X$, we checked that the data $\mathcal{H} = \mathbb{R}^d$, $\mathcal{M} = X$, $\Phi(h, \mu) = h^T\mu + \frac{1}{2}||\phi_X(h) + \phi_X(-h)||^2$ is regular and the family $\mathcal{S}[\mathbb{R}^d, \mathcal{M}, \Phi]$ contains the family $\mathcal{P}[X]$ of all probability distributions supported on $X$. Moreover, for every $\mu \in \mathcal{M} = X$, the family $\mathcal{S}[\mathbb{R}^d, \{\mu\}, \Phi_{\mathbb{R}^d \times \{\mu\}}]$ 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 $F_P(h) = \ln \left(\int_{\mathbb{R}^d} e^{h^T\omega} P(d\omega)\right)$ of the moment-generating function of $P \in \mathcal{P}[X]$ when $h$ is small (indeed, $F_P(h) = h^T e[P] + O(\|h\|^2)$ as $h \to 0$), and by far overestimates $F_P(h)$ when $h$ is large. Utilizing the above construction, we replace $\Phi$ with the real-valued, convex-concave and continuous on $\mathbb{R}^d \times \mathcal{M} = \mathbb{R}^d \times X$ (see Exercise 2.22) function

$$\hat{\Phi}(h; \mu) = \inf_g \left[\hat{\Psi}(h; g; \mu) := (h - g)^T \mu + \frac{1}{2}||\phi_X(h - g) + \phi_X(-h + g)||^2 + \phi_X(g)\right] \leq \Phi(h; \mu). \quad (2.114)$$

It is easy to see that $\hat{\Phi}(\cdot; \cdot)$ still ensures the inclusion $P \in \mathcal{S}[\mathbb{R}^d, \{e[P]\}, \hat{\Phi}_{\mathbb{R}^d \times \{e[P]\}}]$ for every distribution $P \in \mathcal{P}[X]$ and “reproduces $F_P(h)$ reasonably well” for both small and large $h$. Indeed, since $F_P(h) \leq \hat{\Phi}(h; e[P]) \leq \Phi(h; e[P])$, for small $h$ $\hat{\Phi}(h; e[P])$ reproduces $F_P(h)$ even better than $\Phi(h; e[P])$, and we clearly have

$$\hat{\Phi}(h; \mu) \leq [(h - h)^T \mu + \frac{1}{2}||\phi_X(h - h) + \phi_X(-h + h)||^2 + \phi_X(h)] = \phi_X(h) \forall \mu,$$

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 $\Omega = \mathbb{R}^d$ 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

$$\Psi(h; \mu_1, \mu_2) = \frac{1}{2} \left[\Phi_1(-h; \mu_1) + \Phi_2(h; \mu_2)\right]: \mathcal{H} \times (\mathcal{M}_1 \times \mathcal{M}_2) \to \mathbb{R} \quad (2.115)$$

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 (2.115) is

Condition A: The sets $\mathcal{M}_1$ and $\mathcal{M}_2$ are compact, and the function

$$\overline{\Phi}(h) = \max_{\mu_1 \in \mathcal{M}_1, \mu_2 \in \mathcal{M}_2} \Phi(h; \mu_1, \mu_2)$$

where
is coercive on $\mathcal{H}$, meaning that $\Phi(h_i) \to \infty$ 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 2.22) it holds

$$\text{SadVal}[\Phi] := \inf_{h \in \mathcal{H}} \max_{\mu_1 \in M_1, \mu_2 \in M_2} \Phi(h; \mu_1, \mu_2) = \sup_{\mu_1 \in M_1, \mu_2 \in M_2} \inf_{h \in \mathcal{H}} \Phi(h; \mu_1, \mu_2),$$

so that the optimization problems

$$(P) \quad \text{Opt}(P) = \min_{h \in \mathcal{H}} \Phi(h)$$

$$(D) \quad \text{Opt}(D) = \max_{\mu_1 \in M_1, \mu_2 \in M_2} \Phi(\mu_1, \mu_2)$$

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, Opt$(P) = \text{Opt}(D)$ is a real. Problem $(D)$ clearly is the problem of maximizing over a compact set an upper semi-continuous (since $\Phi$ is continuous) function taking real values and, perhaps, value $-\infty$, and not identically equal to $-\infty$ (since Opt$(D)$ is a real), and thus $(D)$ is solvable. As a result, $(P)$ and $(D)$ are solvable with common optimal values, and therefore $\Phi$ has a saddle point.

2.8.2.2 Main Result

An immediate (and essential) observation is as follows:

**Proposition 2.39.** In the situation of Section 2.8.2.1, let $h \in \mathcal{H}$ be such that the quantities

$$\Psi_1(h) = \sup_{\mu_1 \in M_1} \Phi_1(-h; \mu_1), \quad \Psi_2(h) = \sup_{\mu_2 \in M_2} \Phi_2(h; \mu_2)$$

are finite. Consider the affine detector

$$\phi_h(\omega) = h^T \omega + \frac{1}{2} \left[ \Psi_1(h) - \Psi_2(h) \right].$$

Then

$$\text{Risk}[\phi_h|\mathcal{R}[\mathcal{H}, M_1, \Phi_1], \mathcal{R}[\mathcal{H}, M_2, \Phi_2]] \leq \exp\{\frac{1}{2} \left[ \Psi_1(h) + \Psi_2(h) \right]\}.$$ 

**Proof.** Let $h$ satisfy the premise of the proposition. For every $\mu_1 \in M_1$, we have $\Phi_1(-h; \mu_1) \leq \Psi_1(h)$, and for every $P \in \mathcal{R}[\mathcal{H}, M_1, \Phi_1]$ we have

$$\int_{\Omega} \exp\{-h^T \omega\} P(d\omega) \leq \exp\{\Phi_1(-h; \mu_1)\}$$

for properly selected $\mu_1 \in M_1$. Thus,

$$\int_{\Omega} \exp\{-h^T \omega\} P(d\omega) \leq \exp\{\Psi_1(h)\} \quad \forall P \in \mathcal{R}[\mathcal{H}, M_1, \Phi_1],$$
whence also
\[
\int_{\Omega} \exp\{-h^T\omega - x\} P(d\omega) \leq \exp\{\Psi_1(h) - x\} = \exp\{\frac{1}{2}[\Psi_1(h) + \Psi_2(h)]\} \forall P \in \mathcal{R}[\mathcal{H}, \mathcal{M}_1, \Phi_1].
\]
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,
\[
\int_{\Omega} \exp\{h^T\omega\} P(d\omega) \leq \exp\{\Psi_2(h)\} \forall P \in \mathcal{R}[\mathcal{H}, \mathcal{M}_2, \Phi_2],
\]
and
\[
\int_{\Omega} \exp\{h^T\omega + x\} P(d\omega) \leq \exp\{\Psi_2(h) + x\} = \exp\{\frac{1}{2}[\Psi_1(h) + \Psi_2(h)]\} \forall P \in \mathcal{R}[\mathcal{H}, \mathcal{M}_2, \Phi_2].
\]
An immediate corollary is as follows:

**Proposition 2.40.** In the situation of Section 2.8.2.1 and under Assumption I, let us associate with a saddle point \((h_\ast; \mu_1^\ast, \mu_2^\ast)\) of the convex-concave function (2.115) the following entities:

- **the risk**
  \[
  \epsilon_\ast := \exp\{\Psi(h_\ast; \mu_1^\ast, \mu_2^\ast)\};
  \]
  this quantity is uniquely defined by the saddle point value of \(\Psi\) and thus is independent of how we select a saddle point;
- **the detector** \(\phi_\ast(\omega)\) - the affine function of \(\omega \in \mathbb{R}^d\) given by
  \[
  \phi_\ast(\omega) = h_\ast^T\omega + a, \quad a = \frac{1}{2} [\Phi_1(-h_\ast; \mu_1^\ast) - \Phi_2(h_\ast; \mu_2^\ast)].
  \]

Then
\[
\text{Risk}[\phi_\ast|\mathcal{R}[\mathcal{H}, \mathcal{M}_1, \Phi_1], \mathcal{R}[\mathcal{H}, \mathcal{M}_2, \Phi_2]] \leq \epsilon_\ast.
\]

**Consequences.** Assume we are given \(L\) collections \((\mathcal{H}, \mathcal{M}_\ell, \Phi_\ell)\) of regular data on a common observation space \(\Omega = \mathbb{R}^d\) 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 \((\ell, \ell')\), \(1 \leq \ell < \ell' \leq L\), the pair \((\mathcal{H}, \mathcal{M}_\ell, \Phi_\ell), (\mathcal{H}, \mathcal{M}_{\ell'}, \Phi_{\ell'})\) of regular data satisfies Assumption I, so that the convex-concave functions
\[
\Psi_{\ell\ell'}(h; \mu_\ell, \mu_{\ell'}) = \frac{1}{2} [\Phi_\ell(-h; \mu_\ell) + \Phi_{\ell'}(h; \mu_{\ell'})] : \mathcal{H} \times (\mathcal{M}_\ell \times \mathcal{M}_{\ell'}) \to \mathbb{R}
\]
[\(1 \leq \ell < \ell' \leq L\)] have saddle points \((h^\ast_{\ell\ell'}; (\mu^\ast_\ell, \mu^\ast_{\ell'}))\) (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
\[
\phi_{\ell\ell'}(\omega) = [h^\ast_{\ell\ell'}]^T\omega + \frac{1}{2} [\Phi_\ell(-h^\ast_{\ell\ell'}; \mu^\ast_\ell) - \Phi_{\ell'}(h^\ast_{\ell\ell'}; \mu^\ast_{\ell'})]
\]
[\(1 \leq \ell < \ell' \leq L\)]
and the quantities
\[ \epsilon_{\ell\ell'} = \exp \left\{ \frac{1}{2} \left[ \Phi_{\ell}(h_{\ell\ell'}^*; \mu_{\ell}) + \Phi_{\ell'}(h_{\ell\ell'}^*; \mu_{\ell'}) \right] \right\} ; \quad [1 \leq \ell < \ell' \leq L] \]
by Proposition 2.40, \( \epsilon_{\ell\ell'} \) are upper bounds on the risks, taken w.r.t. \( P_{\ell}, P_{\ell'} \), of the detectors \( \phi_{\ell\ell'} \):

\[ \int_{\Omega} e^{-\phi_{\ell\ell'}(\omega)} P(d\omega) \leq \epsilon_{\ell\ell'} \forall P \in P_{\ell} \& \int_{\Omega} e^{\phi_{\ell\ell'}(\omega)} P(d\omega) \leq \epsilon_{\ell\ell'} \forall P \in P_{\ell'} . \]

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 (2.80) 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 2.40 extends the grasp of our computation-friendly machinery of test design consider a toy problem as follows:

We are given an observation
\[ \omega = Ax + \sigma A \text{Diag} \{\sqrt{x_1}, \ldots, \sqrt{x_n}\} \xi, \quad (2.118) \]
where
- unknown signal \( x \) is known to belong to a given convex compact subset \( M \) of the interior of \( \mathbb{R}^n_+ \);
- \( A \) 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 a \( K \)-repeated version of observations (2.118) 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
\[ \Theta(x) = \sigma^2 A \text{Diag}\{x\} A^T \]
depending on the true signal – the larger the signal, the greater the noise.

We can easily process the situation in question utilizing the machinery developed in this section. Namely, let us set
\[ \mathcal{H}_\chi = \mathbb{R}^n, \mathcal{M}_\chi = \{(x, \text{Diag}\{x\}) : x \in X_\chi\} \subset \mathbb{R}^n_+ \times S^n_+, \Phi_\chi(h; x, \Xi) = h^T A^T x + \frac{\sigma^2}{2} h^T [A \Xi A^T] h : \mathcal{M}_\chi \to \mathbb{R}. \quad [\chi = 1, 2] \]
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 (2.118) stemming from a signal \( x \in X_\chi \) belongs to \( S[\mathcal{H}, \mathcal{M}_\chi, \Phi_\chi] \), so that we can use Proposition 2.40 to build an affine detector for the families \( P_\chi, \chi = 1, 2 \), of distributions of observations (2.118) 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

\[
\Psi(h; x, y) = \frac{1}{2} \left[ h^T A (y - x) + \frac{\sigma^2}{2} h^T A \text{Diag} \{x + y\} A^T h \right]
\]

(min in \(h \in \mathbb{R}^n\), 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

\[
\phi_*(\omega) = h_*^T \omega + \frac{\sigma^2}{2} h_*^T A \text{Diag} \{x_* - y_*\} A^T h_* - \frac{\sigma}{2} h_*^T A \{x_* + y_*\}
\]

such that

\[
\text{Risk}[\phi_*|\mathcal{P}_1, \mathcal{P}_2] \leq \exp \left\{ \frac{1}{2} \left[ h_*^T A [y_* - x_*] + \frac{\sigma^2}{2} h_*^T A \text{Diag} \{x_* + y_*\} A^T h_* \right] \right\}. \tag{2.119}
\]

Note that we could also process the situation when defining the regular data as \(\mathcal{H}, \mathcal{M}^\chi_\chi = X_\chi, \Phi^\chi_\chi, \chi = 1, 2\), where

\[
\Phi^\chi_\chi(h; x) = h^T A x + \frac{\sigma^2}{2} h^T A A^T h \quad \text{with} \quad |\theta = \max_{x \in X_1 \cup X_2} \|x\|_\infty|
\]

which, basically, means passing from our actual observations (2.118) to the “more noisy” observations given by the Gaussian o.s.

\[
\omega = A x + \eta, \eta \sim \mathcal{N}(0, \sigma^2 AA^T). \tag{2.120}
\]

It is easily seen that the risk \(\text{Risk}[\phi_*|\mathcal{P}_1, \mathcal{P}_2]\) of the optimal, for this Gaussian o.s., detector \(\phi_*\), can be upper-bounded by the known to us risk \(\text{Risk}[\phi_*|\mathcal{P}^\chi_\chi^+, \mathcal{P}^\chi_\chi^+]\), where \(\mathcal{P}^\chi_\chi^+\) is the family of distributions of observations (2.120) induced by signals \(x \in X_\chi\). Note that \(\text{Risk}[\phi_*|\mathcal{P}^\chi_\chi^+, \mathcal{P}^\chi_\chi^+]\) is seemingly the best risk bound available for us “within the realm of detector-based tests in simple o.s.’s.” The goal of the small numerical experiment we are about to report on is to understand how our new risk bound (2.119) compares to the “old” bound \(\text{Risk}[\phi_*|\mathcal{P}^\chi_\chi^+, \mathcal{P}^\chi_\chi^+]\). We use

\[
n = 16, \ \ \ \ X_1 = \left\{ x \in \mathbb{R}^{16} : \begin{array}{l}
0.001 \leq x_1 \leq \delta \\
0.001 \leq x_i \leq 1, \ 2 \leq i \leq 16
\end{array} \right\},
\]

\[
X_2 = \left\{ x \in \mathbb{R}^{16} : \begin{array}{l}
2 \delta \leq x_1 \leq 1 \\
0.001 \leq x_i \leq 1, \ 2 \leq i \leq 16
\end{array} \right\}
\]

and \(\sigma = 0.1\). The “separation parameter” \(\delta\) is set to 0.1. Finally, \(16 \times 16\) matrix \(A\) has condition number 100 (singular values 0.01\((i-1)/15\), \(1 \leq i \leq 16\)) and randomly oriented systems of left- and right singular vectors. With this setup, a typical numerical result is as follows:

- the right hand side in (2.119) is 0.4346, implying that with detector \(\phi_*\), 6-repeated observation is sufficient to decide on our two hypotheses with risk \(\leq 0.01\);
- the quantity \(\text{Risk}[\phi_*|\mathcal{P}^\chi_\chi^+, \mathcal{P}^\chi_\chi^+]\) is 0.8825, meaning that with detector \(\phi_*\), we need at least 37-repeated observation to guarantee risk \(\leq 0.01\).

When the separation parameter \(\delta\) participating in the descriptions of \(X_1, X_2\) is reduced to 0.01, the risks in question grow to 0.9201 and 0.9988, respectively (56-
repeated observation to decide on the hypotheses with risk 0.01 when \( \phi_* \) is used vs. 3685-repeated observation needed when \( \phi_\# \) is used. The bottom line is that the new developments can indeed improve significantly the performance of our inferences.

### 2.8.2.3 Sub-Gaussian and Gaussian cases

For \( \chi = 1, 2 \), let \( U_\chi \) be nonempty closed convex set in \( \mathbb{R}^d \), and \( \mathcal{V}_\chi \) be a compact convex subset of the interior of the positive semidefinite cone \( \mathcal{S}_+^d \). We assume that \( U_1 \) is compact. Setting

\[
\mathcal{H}_\chi = \Omega = \mathbb{R}^d, \quad \mathcal{M}_\chi = U_\chi \times \mathcal{V}_\chi, \\
\Phi_{\chi}(h; \theta, \Theta) = \theta^T h + \frac{1}{2} h^T \Theta h : \mathcal{H}_\chi \times \mathcal{M}_\chi \to \mathbb{R}, \chi = 1, 2,
\]

(2.121)

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 \( \mathcal{S}[\mathbb{R}^d, \mathcal{M}_\chi, \Phi_{\chi}] \) contain the families \( \mathcal{S}[U_\chi, \mathcal{V}_\chi] \) of sub-Gaussian distributions on \( \mathbb{R}^d \) with sub-Gaussianity parameters \((\theta, \Theta) \in U_\chi \times \mathcal{V}_\chi \) (see (2.106)), as well as families \( \mathcal{G}[U_\chi, \mathcal{V}_\chi] \) of Gaussian distributions on \( \mathbb{R}^d \) with parameters \((\theta, \Theta) \) (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 \( \mathcal{M}_\chi^* \) given by the above construction as applied to the collections of regular data (2.121) is well defined and allows to decide on hypotheses

\[
H_\chi : P \in \mathcal{R}[\mathbb{R}^d, U_\chi, \mathcal{V}_\chi], \chi = 1, 2,
\]

on the distribution \( P \) underlying \( K \)-repeated observation \( \omega^K \). The same test can be also used to decide on stricter hypotheses \( H_\chi^G \), \( \chi = 1, 2 \), 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 \( \mathcal{M}_\chi^* \) when the Gaussian hypotheses \( H_\chi^G \) are considered and the situation is symmetric, that is, when \( \mathcal{V}_1 = \mathcal{V}_2 \).

Observe, first, that the convex-concave function \( \Psi \) from (2.115) in the current setting becomes

\[
\Psi(h; \theta_1, \Theta_1, \theta_2, \Theta_2) = \frac{1}{2} h^T [\theta_2 - \theta_1] + \frac{1}{2} h^T \Theta_1 h + \frac{1}{2} h^T \Theta_2 h.
\]

(2.122)

We are interested in solutions to the saddle point problem

\[
\min_{h \in \mathbb{R}^d} \max_{\theta_1 \in U_1, \theta_2 \in U_2} \max_{\Theta_1 \in \mathcal{V}_1, \Theta_2 \in \mathcal{V}_2} \Psi(h; \theta_1, \Theta_1, \theta_2, \Theta_2)
\]

(2.123)

associated with the function (2.122). From the structure of \( \Psi \) 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

\[
(a) \quad h_* = [\Theta_*^1 + \Theta_*^2]^{-1}[\theta_*^1 - \theta_*^2], \\
(b) \quad h_*^T (\theta_1 - \theta_2) \geq 0 \forall \theta_1 \in U_1, \quad h_*^T (\theta_2 - \theta_2) \geq 0 \forall \theta_2 \in U_2,
\]

\[
(c) \quad h_*^T \Theta_1 h_* \leq h_*^T \Theta_2 h_* \forall \Theta_1 \in \mathcal{V}_1, \quad h_*^T \Theta_2 h_* \leq h_*^T \Theta_2 h_* \forall \Theta_2 \in \mathcal{V}_2.
\]

(2.124)
From (2.124a) it immediately follows that the affine detector \( \phi_*(\cdot) \) and risk \( \epsilon_* \), as given by (2.116) and (2.117), are

\[
\begin{align*}
\phi_*(\omega) &= h_*^T[\omega - w_*], \quad w_* = \frac{1}{2}[\theta_1^* + \theta_2^*]; \\
\epsilon_* &= \exp\{-\frac{1}{4}[\theta_1^* - \theta_2^*]^T[\Theta_1^* + \Theta_2^*]^{-1}[\theta_1^* - \theta_2^*]\};
\end{align*}
\]

(2.125)

Note that in the symmetric case (where \( \mathcal{V}_1 = \mathcal{V}_2 \)), there always exists a saddle point of \( \Psi \) with \( \Theta_1^* = \Theta_2^* \), and the test \( \mathcal{T}_1^* \) associated with such saddle point is quite transparent: it is the maximum likelihood test for two Gaussian distributions, \( \mathcal{N}(\theta_1^*, \Theta_1^*), \mathcal{N}(\theta_2^*, \Theta_2^*) \), where \( \Theta_* \) is the common value of \( \Theta_1^* \) and \( \Theta_2^* \). The bound \( \epsilon_* \) on the risk of the test is nothing but the Hellinger affinity of these two Gaussian distributions, or, equivalently,

\[
\epsilon_* = \exp\left\{-\frac{1}{8}[\theta_1^* - \theta_2^*]^T\Theta_*^{-1}[\theta_1^* - \theta_2^*]\right\}.
\]

We arrive at the following result:

**Proposition 2.41.** In the symmetric sub-Gaussian case (i.e., in the case of (2.121) with \( \mathcal{V}_1 = \mathcal{V}_2 \)), saddle point problem (2.122), (2.123) admits a saddle point of the form \( (h_*; \theta_1^*, \Theta_1^*, \theta_2^*, \Theta_2^*) \), and the associated affine detector and its risk are given by

\[
\begin{align*}
\phi_*(\omega) &= h_*^T[\omega - w_*], \quad w_* = \frac{1}{2}[\theta_1^* + \theta_2^*]; \\
\epsilon_* &= \exp\{-\frac{1}{8}[\theta_1^* - \theta_2^*]^T\Theta_*^{-1}[\theta_1^* - \theta_2^*]\};
\end{align*}
\]

As a result, when deciding, via \( \omega^K \), on “sub-Gaussian hypotheses” \( H_\chi, \chi = 1,2 \), the risk of the test \( \mathcal{T}_*^K \) associated with \( \phi_*^{(K)}(\omega^K) := \sum_{i=1}^K \phi_*(\omega_i) \) is at most \( \epsilon^K_* \).

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 \( \omega = \omega_1 \) in order to decide on the hypotheses \( H_{\chi}' \), \( \chi = 1,2 \), the above risk bound can be improved:

**Proposition 2.42.** 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 \( \Psi \) given by (2.122), and let \( \phi_* \) be the affine detector given by (2.124) and (2.125):

\[
\phi_*(\omega) = h_*^T[\omega - w_*], \quad h_* = \frac{1}{2}\Theta_*^{-1}[\theta_1^* - \theta_2^*], \quad w_* = \frac{1}{2}[\theta_1^* + \theta_2^*].
\]

Let also

\[
\delta = \sqrt{h_*^T\Theta_*h_*} = \frac{1}{2}\sqrt{[\theta_1^* - \theta_2^*]^T\Theta_*^{-1}[\theta_1^* - \theta_2^*]},
\]

so that

\[
\delta^2 = h_*^T[\theta_1^* - w_*] = h_*^T[w_* - \theta_2^*] \quad \text{and} \quad \epsilon_* = \exp\{-\frac{1}{2}\delta^2\}.
\]

\[\text{21}\]Indeed, from (2.122) 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. \( \Theta_1, \Theta_2 \), implying similar symmetry of the function \( \Psi(\theta_1, \Theta_1, \theta_2, \Theta_2) = \min_{h \in \mathcal{H}} \Psi(h; \theta_1, \Theta_1, \theta_2, \Theta_2) \). Since \( \Psi \) is concave, the set \( M \) of its maximizers over \( \mathcal{H}_1 \times \mathcal{H}_2 \) (which, as we know, is nonempty) is symmetric w.r.t. the swap of \( \Theta_1 \) and \( \Theta_2 \) 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 \( \Psi \) with \( \Theta_1 = \Theta_2 \).
Let, further, \( \alpha \leq \delta^2, \beta \leq \delta^2 \). Then

\[
\begin{align*}
\forall (\theta \in U_1, \Theta \in \mathcal{V}) : \text{Prob}_{\omega \sim \mathcal{N}(\theta, \Theta)} \{ \phi_*(\omega) \leq \alpha \} & \leq \text{Erfc}(\delta - \alpha/\delta), \\
\forall (\theta \in U_2, \Theta \in \mathcal{V}) : \text{Prob}_{\omega \sim \mathcal{N}(\theta, \Theta)} \{ \phi_*(\omega) \geq -\beta \} & \leq \text{Erfc}(\delta - \beta/\delta).
\end{align*}
\]

(2.127)

In particular, when deciding, via a single observation \( \omega \), on Gaussian hypotheses \( H^G_\chi, \chi = 1, 2 \), with \( H^G_\chi \) stating that \( \omega \sim \mathcal{N}(\theta, \Theta) \) with \( (\theta, \Theta) \in U_\chi \times \mathcal{V} \), the risk of the test \( T_*^1 \) associated with \( \phi_* \) is at most \( \text{Erfc}(\delta) \).

**Proof.** Let us prove (a) (the proof of (b) is completely similar). For \( \theta \in U_1, \Theta \in \mathcal{V} \) we have

\[
\text{Prob}_{\omega \sim \mathcal{N}(\theta, \Theta)} \{ \phi_*(\omega) \leq \alpha \} = \text{Prob}_{\xi \sim \mathcal{N}(0, I)} \{ h_*^T [\theta + \Theta^{1/2} \xi - w_*] \leq \alpha \} = \text{Prob}_{\xi \sim \mathcal{N}(0, I)} \{ [\Theta^{1/2} h_*]^T \xi \leq \alpha - h_*^T [\theta - w_*] \} \leq \text{Erfc}(\delta^2 - \alpha)/\| \Theta^{1/2} h_* \|_2
\]

\[
= \text{Erfc}(\delta^2 - \alpha)/\| \Theta^{1/2} h_* \|_2
\]

[due to \( \delta^2 - \alpha \geq 0 \) and \( h_*^T \Theta h_* \leq h_*^T \Theta h_* \) by (2.124.c)]

\[
= \text{Erfc}(\delta^2 - \alpha)/\delta.
\]

The “in particular” part of Proposition is readily given by (2.127) 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 and forever fixed. Now neither the mean nor the covariance matrix of the observed Gaussian random are known in advance. In this setting, the mean is running through a closed convex set (depending on the hypothesis), and the covariance is running, 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.

### 2.9 BEYOND THE SCOPE OF AFFINE DETECTORS: LIFTING THE OBSERVATIONS

#### 2.9.1 Motivation

The detectors considered in Section 2.8 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

\[ \zeta = A[u; 1] + \xi \in \mathbb{R}^n, \]

where \( u \) is an unknown signal known to belong to a given set \( U \subset \mathbb{R}^n \), \( u \mapsto A[u; 1] \) is a given affine mapping from \( \mathbb{R}^n \) into the observation space \( \mathbb{R}^d \), and \( \xi \) is zero mean Gaussian observation noise with covariance matrix \( \Theta \) known to belong to a given convex compact subset \( \mathcal{V} \) of the interior of the positive semidefinite
cone $\mathbb{S}^d_+$. 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 $\zeta$ under a whatever deterministic mapping, e.g., the “quadratic lifting” $\zeta \mapsto (\zeta, \zeta^T)$. A detector affine in the new observation is quadratic in $u$ and $\xi$ – we get access to a wider set of detectors as compared to those affine in $\zeta$! At first glance, applying our “affine detectors” machinery to appropriate “nonlinear liftings” of actual observations we can handle quite complicated detectors, e.g., polynomial, of arbitrary degree, in $\zeta$. The bottleneck here stems from the fact that in general it is difficult to “cover” the distribution of “nonlinearly lifted” observation $\zeta$ (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 takes place in the case of the quadratic lifting $\zeta \mapsto (\zeta, \zeta^T)$ of (sub)Gaussian observation $\zeta$, and the resulting quadratic detectors allow to handle some important inference problems which are far beyond the grasp of “genuinely affine” detectors.

2.9.2 Quadratic lifting: Gaussian case

Given positive integer $d$, we define $\mathcal{E}^d$ as the linear space $\mathbb{R}^d \times \mathbb{S}^d$ equipped with the inner product

$$
\langle (z, S), (z', S') \rangle = s^T z' + \frac{1}{2} \text{Tr}(SS').
$$

Note that the quadratic lifting $z \mapsto (z, zz^T)$ maps the space $\mathbb{R}^d$ into $\mathcal{E}^d$.

In the sequel, an instrumental role is played by the following result.

**Proposition 2.43.**

(i) Assume we are given

- a nonempty and bounded subset $U$ of $\mathbb{R}^n$;
- a convex compact set $V$ contained in the interior of the cone $\mathbb{S}^d_+$ of positive semidefinite $d \times d$ matrices;
- a $d \times (n + 1)$ matrix $A$.

These data specify the family $\mathcal{G}_A[U, V]$ of distributions of quadratic liftings $(\zeta, \zeta^T)$ of Gaussian random vectors $\zeta \sim \mathcal{N}(A[u; 1], \Theta)$ stemming from $u \in U$ and $\Theta \in V$.

Let us select somehow

1. $\gamma \in (0, 1)$,
2. convex compact subset $Z$ of the set $\mathcal{Z}^n = \{Z \in \mathbb{S}^{n+1} : Z \succeq 0, Z_{n+1,n+1} = 1\}$ such that

$$
Z(u) := [u; 1][u; 1]^T \in Z \quad \forall u \in U, \quad (2.128)
$$

3. positive definite $d \times d$ matrix $\Theta_\ast \in \mathbb{S}^d_+$ and $\delta \in [0, 2]$ such that

$$
\Theta_\ast \succeq \Theta \quad \forall \Theta \in \mathcal{V} \quad \text{and} \quad \|\Theta^{1/2} \Theta_\ast^{1/2} - I_d\| \leq \delta \quad \forall \Theta \in \mathcal{V}, \quad (2.129)
$$

where $\|\cdot\|$ is the spectral norm.$^{22}$

---

$^{22}$It is easily seen that with $\delta = 2$, the second relation in (2.129) is satisfied for all $\Theta$ such that $0 \preceq \Theta \preceq \Theta_\ast$, so that the restriction $\delta \leq 2$ is w.l.o.g.
and set
\[ \mathcal{H} = \mathcal{H}^\dagger := \{(h, H) \in \mathbb{R}^d \times S^d : -\gamma \Theta^{-1}_* \leq H \leq \gamma \Theta^{-1}_*\} , \]

\[ \Phi_{A,Z}(h, H; \Theta) = -\frac{1}{2} \ln \det(I - \Theta^{1/2}_h H \Theta^{1/2}_H) + \frac{1}{2} \text{Tr}((\Theta - \Theta_*)H) + \frac{1}{2(1 - \|\Theta^{1/2}_h H \Theta^{1/2}_H\|)} \|\Theta^{1/2}_H H \Theta^{1/2}_H\|_F^2 + \frac{1}{2} \phi_Z \begin{bmatrix} H \chi_h \end{bmatrix} + [H, h]^T [\Theta^{-1}_* - H]^{-1} [H, h]B \] \quad \mathcal{H} \times \mathcal{V} \to \mathbb{R}, \tag{2.130}

where \( B \) is given by
\[ B = \begin{bmatrix} A \\ [0, \ldots, 0, 1] \end{bmatrix}, \tag{2.131} \]
the function
\[ \phi_Z(Y) := \max_{Z \in \mathcal{Z}} \text{Tr}(ZY) \tag{2.132} \]
is the support function of \( \mathcal{Z} \), and \( \| \cdot \|_F \) is the Frobenius norm.

Function \( \Phi_{A,Z} \) is continuous on its domain, convex in \((h, H) \in \mathcal{H} \) and concave in \( \Theta \in \mathcal{V} \), so that \((\mathcal{H}, \mathcal{V}, \Phi_{A,Z})\) is a regular data. Besides this,

(\#) Whenever \( u \in \mathbb{R}^n \) 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
\[ \forall (h, H) \in \mathcal{H} : \ln \left( \text{E}_{\zeta \sim \mathcal{N}(A[u; 1], \Theta)} \left\{ e^{\frac{1}{2} \zeta^T [H\zeta + H^T \zeta]} \right\} \right) \leq \Phi_{A,Z}(h, H; \Theta). \tag{2.133} \]
The latter relation combines with (2.128) to imply that
\[ G_A[U, V] \subset \mathcal{S}[\mathcal{H}, \mathcal{V}, \Phi_{A,Z}]. \]

In addition, \( \Phi_{A,Z} \) is coercive in \((h, H) : \Phi_{A,Z}(h_i, H_i; \Theta) \to +\infty \) as \( i \to \infty \) whenever \( \Theta \in \mathcal{V}, (h, H) \in \mathcal{H} \) and \( \|(h_i, H_i)\| \to \infty, i \to \infty \).

(ii) Let two collections of entities from (i), \((V_\chi, \Theta_\chi^{(x)}, \delta_\chi, \gamma_\chi, A_\chi, Z_\chi)\), \( \chi = 1, 2 \), with common \( d \) be given, giving rise to the sets \( \mathcal{H}_\chi \), matrices \( B_\chi \), and functions \( \Phi_{A,Z}(h, H; \Theta) \), \( \chi = 1, 2 \). These collections specify the families of normal distributions
\[ G_\chi = \{ \mathcal{N}(v, \Theta) : \Theta \in \mathcal{V}_\chi \land \exists u \in U : v = A_\chi[u; 1] \}, \chi = 1, 2. \]

Consider the convex-concave saddle point problem
\[ \begin{aligned} \mathcal{S}V &= \min_{(h, H) \in \mathcal{H}_1 \cap \mathcal{H}_2} \max_{\Theta_1, \Theta_2 \in \mathcal{V}_2 \cap \mathcal{V}_1} \left\{ \frac{1}{2} \left[ \Phi_{A_1,Z_1}(-h_1, H_1; \Theta_1) + \Phi_{A_2,Z_2}(h_2, H_2; \Theta_2) \right] \right\} \quad \Phi(h, H; \Theta_1, \Theta_2) 
\end{aligned} \tag{2.134} \]
A saddle point \((H_*, h_*; \Theta_1^*, \Theta_2^*)\) in this problem does exist, and the induced quadratic detector
\[ \phi_*(\omega) = \frac{1}{2} \omega^T H_* \omega + h_*^T \omega + \frac{1}{2} \left[ \Phi_{A_1,Z_1}(-h_1, H_*; \Theta_1^*) - \Phi_{A_2,Z_2}(h_2, H_*; \Theta_2^*) \right], \tag{2.135} \]
when applied to the families of Gaussian distributions $G_\chi$, $\chi = 1, 2$, has the risk
\[
\text{Risk}[\phi_\star | G_1, G_2] \leq \epsilon_\star := e^{SV},
\]
that is,
\[
\begin{align*}
(a) \quad & \int_{\mathbb{R}^d} e^{-\phi_\star(\omega)} P(d\omega) \leq \epsilon_\star \quad \forall P \in G_1, \\
(b) \quad & \int_{\mathbb{R}^d} e^{\phi_\star(\omega)} P(d\omega) \leq \epsilon_\star \quad \forall P \in G_2.
\end{align*}
\]
(2.136)

For proof, see Section 2.11.5.

Remark 2.44. Note that the computational effort to solve (2.134) reduces dramatically in the “easy case” of the situation described in item (ii) of Proposition 2.43 where

- the observations are direct, meaning that $A_\chi[u; 1] = u$, $u \in \mathbb{R}^d$, $\chi = 1, 2$;
- 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

$$
\mathcal{Z}_\chi = \{ Z \in S_+^{d+1} : Z \succeq 0, \quad \text{Tr}(ZQ^\chi_j) \leq q^\chi_j, \quad 1 \leq j \leq J_\chi \}
$$

with diagonal matrices $Q_j^\chi$; \(^{23}\) and these sets intersect the interior of the positive semidefinite cone $S_+^{d+1}$.

In this case, the convex-concave saddle point problem (2.134) admits a saddle point $(h_\star, H_\star; \Theta_1^\star, \Theta_2^\star)$ where $h_\star = 0$ and $H_\star$ is diagonal.

Justifying the remark. In the easy case, we have $B_\chi = I_{d+1}$ and therefore

$$
M_\chi(h, H) := B^T_\chi \left[ \begin{bmatrix} H & h \\ h^T & h^T \end{bmatrix} + [H, h]^T [\Theta_\chi]^{-1} - H \right]^{-1} [H, h] B_\chi
$$

and

$$
\phi_{Z_\chi}(Z) = \max_W \{ \text{Tr}(ZW) : W \succeq 0, \quad \text{Tr}(WQ_j^\chi) \leq q_j^\chi, \quad 1 \leq j \leq J_\chi \}
$$

$$
= \min_\lambda \left\{ \sum_j q_j^\chi \lambda_j : \lambda \geq 0, \quad Z \succeq \sum_j \lambda_j Q_j^\chi \right\},
$$

where the last equality is due to semidefinite duality. \(^{24}\) From the second representation of $\phi_{Z_\chi}(\cdot)$ and the fact that all $Q_j^\chi$ are diagonal it follows that $\phi_{Z_\chi}(M_\chi(-h, H)) = \phi_{Z_\chi}(M_\chi(h, H))$ (indeed, with diagonal $Q_j^\chi$, if $\lambda$ 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 $M_\chi(-h, H)$). This, in turn, combines straightforwardly with (2.130) to imply that when replacing $h_\star$ with 0 in a saddle point $(h_\star, H_\star; \Theta_1^\star, \Theta_2^\star)$ of (2.134), we end up with another saddle point of (2.134). In other words, when solving (2.134), we can from the very beginning set $h$ to 0, thus

\(^{23}\) In terms of the sets $U_\chi$, this assumption means that the latter sets are given by linear inequalities on the squares of entries in $u$.

\(^{24}\) See Section 4.1 (or [184, Section 7.1] for more details).
HYPOTHESIS TESTING

converting (2.134) into the convex-concave saddle point problem

\[
SV = \min_{H: (0,H) \in H_1 \cap H_2} \max_{\Theta_1 \in V_1, \Theta_2 \in V_2} \Phi(0, H; \Theta_1, \Theta_2).
\]  

(2.137)

Taking into account that we are in the case where all matrices from the sets \(V_\chi\), same as the matrices \(\Theta_\chi\) and all the matrices \(Q_\chi\), \(\chi = 1, 2\), are diagonal, it is immediate to verify that \(\Phi(0, H; \Theta_1, \Theta_2) = \Phi(0, EHE; \Theta_1, \Theta_2)\) for any \(d \times d\) diagonal matrix \(E\) with diagonal entries \(\pm 1\). Due to convexity-concavity of \(\Phi\) this implies that (2.137) admits a saddle point \((0, H_*, \Theta_1^*, \Theta_2^*)\) with \(H_ *\) invariant w.r.t. transformations \(H_ * \mapsto EHE\) with the above \(E\), that is, with diagonal \(H_ *\), as claimed.

\(\square\)

2.9.3 Quadratic lifting – does it help?

Assume that for \(\chi = 1, 2\), we are given

- affine mappings \(u \mapsto A_\chi(u) = A_\chi[u; 1]: \mathbb{R}^n \rightarrow \mathbb{R}^d\),
- nonempty convex compact sets \(U_\chi \subset \mathbb{R}^n\),
- nonempty convex compact sets \(\chi \subset \mathbb{R}^d\).

These data define families \(\mathcal{G}_\chi\) of Gaussian distributions on \(\mathbb{R}^d\): \(\mathcal{G}_\chi\) is comprised of all distributions \(\mathcal{N}(A_\chi(u), \Theta)\) with \(u \in U_\chi\) and \(\Theta \in \chi\). The data define also families \(\mathcal{SG}_\chi\) of sub-Gaussian distributions on \(\mathbb{R}^d\): \(\mathcal{SG}_\chi\) is comprised of all sub-Gaussian distributions with parameters \((A_\chi(u), \Theta)\) with \((u, \Theta) \in U_\chi \times \chi\).

Assume we observe random variable \(\zeta \in \mathbb{R}^d\) drawn from a distribution \(P\) known to belong to \(\mathcal{G}_1 \cup \mathcal{G}_2\), and our goal is to decide from a 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, and we assume from now on that we are in this case\(^{25}\).

At present, we have developed two approaches to building detector-based tests for \(H_1, H_2\):

**A.** Utilizing the affine in \(\zeta\) detector \(\phi_{\text{aff}}\) given by solution to the saddle point problem (see (2.122), (2.123)) and set \(\theta_\chi = A_\chi(u_\chi)\) with \(u_\chi\) running through \(U_\chi\)

\[
\text{SadVal}_{\text{aff}} = \min_{h \in \mathbb{R}^d} \max_{\theta_1 \in V_1, \theta_2 \in V_2} \frac{1}{2} \left[ h^T [A_2(u_2) - A_1(u_1)] + \frac{1}{2} h^T [\Theta_1 + \Theta_2] h \right];
\]

this detector satisfies the risk bound

\[
\text{Risk}[\phi_{\text{aff}}|\mathcal{G}_1, \mathcal{G}_2] \leq \exp(\text{SadVal}_{\text{aff}}).
\]

**Q.** Utilizing the quadratic in \(\zeta\) detector \(\phi_{\text{lift}}\) given by Proposition 2.43.ii, with the risk bound

\[
\text{Risk}[\phi_{\text{lift}}|\mathcal{G}_1, \mathcal{G}_2] \leq \exp(\text{SadVal}_{\text{lift}}),
\]

with \(\text{SadVal}_{\text{lift}}\) given by (2.134).

A natural question is, which of these options results in a better risk bound. Note

\(^{25}\)It is easily seen that what follows can be straightforwardly extended to the sub-Gaussian case, where the hypotheses we would decide upon state that \(P \in \mathcal{SG}_\chi\).
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 2.43, relation (2.133) is inequality rather than equality.\footnote{One cannot make (2.133) an equality by redefining the right hand side function – it will lose the required in our context convexity-concavity properties.} We are about to show that under reasonable assumptions, the second option indeed is better:

**Proposition 2.45.** In the situation in question, assume that the sets $\mathcal{V}_\chi$, $\chi = 1, 2$, contain the $\geq$-largest elements, and that these elements are taken as the matrices $\Theta^{(x)}_\chi$ participating in Proposition 2.43.ii. Let, further, the convex compact sets $\mathcal{Z}_\chi$, participating in Proposition 2.43.ii satisfy

$$
\mathcal{Z}_\chi \subset \bar{\mathcal{Z}}_\chi := \left\{ Z = \begin{bmatrix} W & u \\ u & 1 \end{bmatrix} \succeq 0, u \in U_\chi \right\}
$$

(\textit{this assumption does not restrict generality, since} $\bar{\mathcal{Z}}_\chi$, is, along with $U_\chi$, a closed convex set which clearly contains all matrices $[u; 1][u; 1]^T$ with $u \in U_\chi$. Then)

$$
\text{SadVal}_{\text{lift}} \leq \text{SadVal}_{\text{aff}},
$$

that is, option $\text{Q}$ is at least as efficient as option $\text{A}$.

**Proof.** Let $A_\chi = [\bar{A}_\chi, a_\chi]$. Looking at (2.122) (where one should substitute $\theta_\chi = A_\chi(u_\chi)$ with $u_\chi$ running through $U_\chi$) and taking into account that $\Theta_\chi \leq \Theta^{(x)}_\chi \in \mathcal{V}_\chi$, when $\Theta_\chi \in \mathcal{V}_\chi$, we conclude that

$$
\text{SadVal}_{\text{aff}} = \min_{h} \max_{u_1 \in U_1, u_2 \in U_2} \frac{1}{2} \left[ h^T [\bar{A}_2 u_2 - \bar{A}_1 u_1 + a_2 - a_1] + \frac{1}{2} h^T \left[ \Theta^{(1)} + \Theta^{(2)} \right] h \right].
$$

At the same time, we have by Proposition 2.43.ii:

$$
\text{SadVal}_{\text{lin}} = \min_{(h, H) \in \mathcal{H}_1 \cap \mathcal{H}_2} \max_{\Theta_1 \in V_1, \Theta_2 \in V_2} \frac{1}{2} \left[ \Phi_{A_1, z_1} (-h, -H; \Theta_1) + \Phi_{A_2, z_2} (h, H; \Theta_2) \right]
$$

\begin{align*}
&\leq \min_{h \in \mathbb{R}^d} \max_{\Theta_1 \in V_1, \Theta_2 \in V_2} \frac{1}{2} \left[ \frac{1}{2} \max_{z_1 \in z_1} \text{Tr} \left( Z_1 \begin{bmatrix} -\bar{A}_1^2 h & -A_1 h \\ -h^T A_1 & -2h^T a_1 + h^T \Theta^{(1)} h \end{bmatrix} \right) \\
&\quad + \frac{1}{2} \max_{z_2 \in z_2} \text{Tr} \left( Z_2 h^T A_2 \begin{bmatrix} 2h^T a_2 + h^T \Theta^{(2)} h \\ \bar{A}_2^2 h \end{bmatrix} \right) \right] \\
&\quad [\text{by direct computation utilizing (2.130)}]
\end{align*}

\begin{align*}
&\leq \min_{h \in \mathbb{R}^d} \frac{1}{2} \left[ \frac{1}{2} \max_{u_1 \in U_1} \left( -2u_1^T \bar{A}_1^2 h - 2a_1^T h + h^T \Theta^{(1)} h \right) + \\
&\quad \frac{1}{2} \max_{u_2 \in U_2} \left( 2u_2^T \bar{A}_2^2 h + 2a_2^T h + h^T \Theta^{(2)} h \right) \right] \\
&\quad [\text{due to (2.138)}]
\end{align*}

$$
= \text{SadVal}_{\text{aff}},
$$

where the concluding equality is due to (2.140).

\hfill \Box

**Numerical illustration.** To get an impression of the performance of quadratic detectors as compared to affine ones under the premise of Proposition 2.45, we present here the results of an experiment where $U_1 = U_1^0 = \{ u \in \mathbb{R}^{12} : u_i \geq \}$
\[ \rho, 1 \leq i \leq 12 \}, U_2 = U_2^\rho = -U_1^\rho, A_1 = A_2 \in \mathbb{R}^{8 \times 13}, \text{ and } \mathcal{V}_\chi = \{ \Theta(\chi) = \sigma^2 \chi I_8 \} \] 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 2.41 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 \( \Theta^{(1)}(1) \) differs from \( \Theta^{(2)}(2) \). 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, making affine detectors useless. In contrast, quadratic detectors are able to utilize information (valuable when \( \Theta^{(1)}(1) \neq \Theta^{(2)}(2) \)) “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 may increase significantly the attainable risk.

### 2.9.4 Quadratic lifting: sub-Gaussian case

Sub-Gaussian version of Proposition 2.43 is as follows:

**Proposition 2.46.**

(i) Assume we are given

- a nonempty and bounded subset \( U \) of \( \mathbb{R}^n \);
- a convex compact set \( \mathcal{V} \) contained in the interior of the cone \( \mathbb{S}^d_+ \) 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 liftings \( (\zeta, \zeta \zeta^T) \) of sub-Gaussian random vectors \( \zeta \) with sub-Gaussianity parameters \( A[u; 1] \), \( \Theta \) stemming from \( u \in U \) and \( \Theta \in \mathcal{V} \).

Let us select somehow

1. reals \( \gamma, \gamma^+ \) such that \( 0 < \gamma < \gamma^+ < 1 \),

<table>
<thead>
<tr>
<th>( \rho )</th>
<th>( \sigma_1 )</th>
<th>( \sigma_2 )</th>
<th>unrestricted ( H ) and ( h )</th>
<th>( H = 0 )</th>
<th>( h = 0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>2</td>
<td>2</td>
<td>0.31</td>
<td>0.31</td>
<td>1.00</td>
</tr>
<tr>
<td>0.5</td>
<td>1</td>
<td>4</td>
<td>0.24</td>
<td>0.39</td>
<td>0.62</td>
</tr>
<tr>
<td>0.01</td>
<td>1</td>
<td>4</td>
<td>0.41</td>
<td>1.00</td>
<td>0.41</td>
</tr>
</tbody>
</table>

Table 2.2: Risk of quadratic detector \( \phi(\zeta) = h^T \zeta + \frac{1}{2} \zeta^T H \zeta + \kappa \)
2. convex compact subset $Z$ of the set $\mathbb{Z}^n = \{Z \in \mathbb{S}^{n+1} : Z \succeq 0, Z_{n+1,n+1} = 1\}$ such that relation (2.128) takes place.

3. positive definite $d \times d$ matrix $\Theta_\star \in \mathbb{S}_+^d$ and $\delta \in [0, 2]$ such that (2.129) takes place.

These data specify the closed convex sets

$$\mathcal{H} = \mathcal{H}^\gamma := \{(h, H) \in \mathbb{R}^d \times \mathbb{S}^d : -\gamma \Theta_\star^{-1} \preceq H \preceq \gamma \Theta_\star^{-1}\},$$

$$\mathcal{H} = \mathcal{H}^\gamma \cap \mathcal{H}^\delta := \{(h, H, G) \in \mathbb{R}^d \times \mathbb{S}^d \times \mathbb{S}^d : -\gamma \Theta_\star^{-1} \preceq H \preceq \gamma \Theta_\star^{-1}, 0 \preceq G \preceq \gamma^\top \Theta_\star^{-1}, H \preceq G\}$$

and the functions

$$\Psi_{A, Z}(h, H, G) = -\frac{1}{2} \ln \det(I - \Theta_\star^{1/2}G\Theta_\star^{1/2}) + \frac{1}{2} \phi_Z(B^T \begin{bmatrix} H & h \end{bmatrix} + [H, h]^T[\Theta_\star^{-1} - G]^{-1}[H, h]B) : \mathcal{H} \times \mathbb{Z} \to \mathbb{R},$$

$$\Phi_{A, Z}(h, H, \Theta) = -\frac{1}{2} \ln \det(I - \Theta_\star^{1/2}G\Theta_\star^{1/2}) + \frac{1}{2} \text{Tr}[(\Theta - \Theta_\star)G] + \frac{\delta(2-d)}{2(1-\|\Theta_\star^{1/2}G\Theta_\star^{1/2}\|_F^2)}\|\Theta_\star^{1/2}G\Theta_\star^{1/2}\|_F^4 + \frac{1}{2} \phi_Z(B^T \begin{bmatrix} H & h \end{bmatrix} + [H, h]^T[\Theta_\star^{-1} - G]^{-1}[H, h]B) : \mathcal{H} \times \{0 \preceq \Theta \preceq \Theta_\star\} \to \mathbb{R}$$

where $B$ is given by (2.131) and $\phi_Z(\cdot)$ is the support function of $Z$ given by (2.132), along with

$$\Phi_{A, Z}(h, H) = \min_{G \in \mathbb{S}} \{\Psi_{A, Z}(h, H, G) : (h, H, G) \in \mathcal{H}\} : \mathcal{H} \to \mathbb{R},$$

$$\Phi_{A, Z}^\delta(h, H, \Theta) = \min_{G \in \mathbb{S}} \{\Psi_{A, Z}^\delta(h, H, G) : (h, H, G) \in \mathcal{H}\} : \mathcal{H} \times \{0 \preceq \Theta \preceq \Theta_\star\} \to \mathbb{R},$$

$\Phi_{A, Z}(h, H)$ is convex and continuous on its domain, and $\Phi_{A, Z}^\delta(h, H; \Theta)$ is continuous on its domain, convex in $(h, H) \in \mathcal{H}$ and concave in $\Theta \in \{0 \preceq \Theta \preceq \Theta_\star\}$.

Besides this,

(\#) Whenever $u \in \mathbb{R}^n$ 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 $\zeta$ satisfies the relation

$$\forall(h, H) \in \mathcal{H} :$$

(a) $\ln \left\{ e^{\frac{1}{2} h^T H\zeta + h^T \zeta} \right\} \leq \Phi_{A, Z}(h, H),$

(b) $\ln \left\{ e^{\frac{1}{2} h^T H\zeta + h^T \zeta} \right\} \leq \Phi_{A, Z}^\delta(h, H; \Theta),$

which combines with (2.128) to imply that

$$SG_A[U, V] \subset S[H, V, \Phi_{A, Z}] & SG_A[U, V] \subset S[H, V, \Phi_{A, Z}^\delta].$$

In addition, $\Phi_{A, Z}$ and $\Phi_{A, Z}^\delta$ are coercive in $(h, H)$: $\Phi_{A, Z}(h_i, H_i) \to +\infty$ and $\Phi_{A, Z}^\delta(h_i, H_i; \Theta) \to +\infty$ 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): $(\mathcal{V}_\chi, \Theta_\star^{(\chi)}, \delta_\chi, \gamma_\chi, \gamma_\chi^\top, A_\chi, Z_\chi)$, $\chi = 1, 2$, with common $d$ be given, giving rise to the sets $\mathcal{H}_\chi$, matrices $B_\chi$, and functions $\Phi_{A_\chi, Z_\chi}(h, H), \Phi_{A_\chi, Z_\chi}^\delta(h, H; \Theta)$, $\chi = 1, 2$. These collections specify the families $SG_\chi = SG_{A_\chi}[U, V_\chi]$ of sub-Gaussian distributions.
Consider the convex-concave saddle point problem

\[
SV = \min_{(h, H) \in H_1 \cap H_2} \max_{\Theta_1 \in \Theta_1, \Theta_2 \in \Theta_2} \frac{1}{2} \left[ \Phi_{A_1, Z_1}(h, -H; \Theta_1) + \Phi_{A_2, Z_2}(h, H; \Theta_2) \right].
\]  

(2.144)

A saddle point \((h_*, H_*; \Theta_1^*, \Theta_2^*)\) in this problem does exist, and the induced quadratic detector

\[
\phi_*(\omega) = \frac{1}{2} \omega^T H_* \omega + h_*^T \omega + \frac{1}{2} \left[ \Phi_{A_1, Z_1}(h_*, -H_*; \Theta_1^*) - \Phi_{A_2, Z_2}(h_*, H_*; \Theta_2^*) \right],
\]

when applied to the families of sub-Gaussian distributions \(SG_\chi, \chi = 1, 2\), has the risk

\[
\text{Risk}[\phi_*|SG_1, SG_2] \leq \epsilon_* := e^{SV}.
\]

As a result,

\[
\begin{align*}
(a) \quad & \int_{\mathbb{R}^d} e^{-\phi_*(\omega)} P(d\omega) \leq \epsilon_* & \forall P \in SG_1, \\
(b) \quad & \int_{\mathbb{R}^d} e^{\phi_*(\omega)} P(d\omega) \leq \epsilon_* & \forall P \in SG_2.
\end{align*}
\]

Similarly, the convex minimization problem

\[
\text{Opt} = \min_{(h, H) \in H_1 \cap H_2} \frac{1}{2} \left[ \Phi_{A_1, Z_1}(-h, -H) + \Phi_{A_2, Z_2}(h, H) \right].
\]  

(2.145)

is solvable, and the induced by its optimal solution \((h_*, H_*)\) quadratic detector

\[
\phi_*(\omega) = \frac{1}{2} \omega^T H_* \omega + h_*^T h_* + \frac{1}{2} \left[ \Phi_{A_1, Z_1}(-h_*, -H_*; \Theta_1^*) - \Phi_{A_2, Z_2}(h_*, H_*; \Theta_2^*) \right],
\]

when applied to the families of sub-Gaussian distributions \(SG_\chi, \chi = 1, 2\), has the risk

\[
\text{Risk}[\phi_*|SG_1, SG_2] \leq \epsilon_* := e^{\text{Opt}},
\]

so that for just defined \(\phi_*\) and \(\epsilon_*\) relation (2.145) takes place.

For proof, see Section 2.11.6.

Remark 2.47. Proposition 2.46 offers two options for building quadratic detectors for the families \(SG_1, SG_2\), those based on saddle point of (2.144) and on optimal solution to (2.145). Inspecting the proof, the number of options can be increased to 4: we can replace any of the functions \(\Phi_{A_\chi, Z_\chi}, \chi = 1, 2\) (or both these functions simultaneously) with \(\Phi_{A_\chi, Z_\chi}\). The second of the original two options is exactly what we get when replacing both \(\Phi_{A_\chi, Z_\chi}, \chi = 1, 2\), with \(\Phi_{A_\chi, Z_\chi}\) It is easily seen that depending on the data, each 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.
2.9.5 Generic application: quadratically constrained hypotheses

Propositions 2.43, 2.46 operate with Gaussian/sub-Gaussian observations $\zeta$ with matrix parameters $\Theta$ running through convex compact subsets $\mathcal{V}$ of $\text{int} \mathbf{S}_{++}^d$, and means of the form $A[u; 1]$, with “signals” $u$ running through given sets $U \subset \mathbb{R}^n$. The constructions, however, involved additional entities – convex compact sets $Z \subset \mathbb{R}^n := \{ Z \in \mathbf{S}_{++}^{n+1} : Z_{n+1,n+1} = 1 \}$ containing quadratic liftings $[u; 1][u; 1]^T$ of all signals $u \in U$. Other things being equal, the smaller $Z$, the smaller the associated function $\Phi_{A,Z}$ (or $\Phi_{A,Z}^d$), and consequently, the smaller the (upper bounds on the) risks of quadratic in $\zeta$ detectors we end up with. In order to implement these constructions, we need to understand how to build the required sets $Z$ in an “economical” way. There is a relatively simple case when it is easy to get reasonable candidates for the role of $Z$ – the case of quadratically constrained signal set $U$:

$$U = \{ u \in \mathbb{R}^n : f_k(u) := u^T Q_k u + 2q_k^T u \leq b_k, 1 \leq k \leq K \}.$$  \tag{2.147}

Indeed, the constraints $f_k(u) \leq b_k$ are just linear constraints on the quadratic lifting $[u; 1][u; 1]^T$ of $u$:

$$u^T Q_k u + 2q_k^T u \leq b_k \iff \text{Tr}(F_k[u; 1][u; 1]^T) \leq b_k, \quad F_k = \begin{bmatrix} Q_k & q_k \\ q_k^T & b_k \end{bmatrix} \in \mathbf{S}_{++}^{n+1}.$$  

Consequently, in the case of (2.147), the simplest candidate on the role of $Z$ is the set

$$Z = \{ Z \in \mathbf{S}^n : Z \succeq 0, Z_{n+1,n+1} = 1, \text{Tr}(F_k Z) \leq b_k, 1 \leq k \leq K \}.$$  \tag{2.148}

This set clearly is closed and convex (the latter – even when $U$ itself is not convex), and indeed contains the quadratic liftings $[u; 1][u; 1]^T$ of all points $u \in U$. We need also the compactness of $Z$; the latter definitely takes place when the quadratic constraints describing $U$ contain constraint of the form $u^T u \leq R^2$, which, in turn, can be ensured, basically “for free,” when $U$ is bounded. It should be stressed that the “ideal” choice of $Z$ would be the convex hull $Z[U]$ 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 liftings of all points from $U$. Moreover, $Z[U]$ is closed and bounded, provided $U$ is so. The difficulty is that $Z[U]$ can be computationally intractable (and thus useless in our context) already for pretty simple sets $U$ of the form (2.147). The set (2.148) is a simple outer approximation of $Z[U]$, and this approximation can be very loose: for instance, when $U = \{ u : -1 \leq u_k \leq 1, 1 \leq k \leq n \}$ is just the unit box in $\mathbb{R}^n$, the set (2.148) 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 $Z[U]$ clearly is bounded. There is, essentially, just one generic case when the set (2.148) is exactly equal to $Z[U]$ – the case where

$$U = \{ u : u^T Q u \leq c \}, Q \succ 0$$

is an ellipsoid centered at the origin; the fact that in this case the set given by (2.148) is exactly $Z[U]$ is a consequence of what is called S-Lemma.

Though, in general, the set $Z$ can be a very loose outer approximation of $Z[U]$, this does not mean that this construction cannot be improved. As an instructive
example, let \( U = \{ u \in \mathbb{R}^n : \|u\|_\infty \leq 1 \} \). We get a much better than above approximation of \( Z[U] \) when applying (2.148) to an equivalent description of the box by quadratic constraints:

\[
U := \{ u \in \mathbb{R}^n : \|u\|_\infty \leq 1 \} = \{ u \in \mathbb{R}^n : u_k^2 \leq 1, 1 \leq k \leq n \}.
\]

Applying the recipe of (2.148) to the latter description of \( U \), we arrive at a significantly less conservative outer approximation of \( Z[U] \), specifically,

\[
Z = \{ Z \in \mathbb{S}^{n+1} : Z \succeq 0, Z_{n+1,n+1} = 1, Z_{kk} \leq 1, 1 \leq k \leq n \}.
\]

Not only the resulting set \( Z \) is bounded; we can get a reasonable “upper bound” on the discrepancy between \( Z \) and \( Z[U] \). Namely, denoting by \( Z^o \) the matrix obtained from a symmetric \( n \times n \) matrix \( Z \) by zeroing out the entry \( Z_{n+1,n+1} \) and keeping the remaining entries intact, we have

\[
Z^o[U] := \{ Z^o : Z \in Z[U] \} \subset Z^o := \{ Z^o : Z \in Z \} \subset O(1) \ln(n + 1) Z^o.
\]

This is a particular case of a general result (which goes back to [187]; we shall get this result as a byproduct of our forthcoming considerations, specifically, Proposition 4.6) as follows:

Let \( U \) be a bounded set given by a system of convex quadratic constraints without linear terms:

\[
U = \{ u \in \mathbb{R}^n : u^T Q_k u \leq c_k, 1 \leq k \leq K \}, Q_k \succeq 0, 1 \leq k \leq K,
\]

and let \( Z \) be the associated set (2.148):

\[
Z = \{ Z \in \mathbb{S}^{n+1} : Z \succeq 0, Z_{n+1,n+1} = 1, \text{Tr}(Z \text{Diag}(Q_k, 1)) \leq c_k, 1 \leq k \leq K \}
\]

Then

\[
Z^o[U] := \{ Z^o : Z \in Z[U] \} \subset Z^o := \{ Z^o : Z \in Z \} \subset 3 \ln(\sqrt{3}(K + 1)) Z^o[U].
\]

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 \( 3 \ln(\sqrt{3}(K + 1)) \) 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 the outer approximation (2.148) of \( Z[U] \) 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 \( a_i^T u \leq c_i, a_j^T u \leq c_j \), 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 (2.148). Informally speaking, transition from (2.147) to (2.148) 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, while “computationally
tractable” assistance of this type allows to reduce the conservatism of (2.148), it usually does not allow to eliminate it completely: a grave “fact of life” is that even in the case of the unit box $U$, the set $Z[U]$ is computationally intractable. Scientifically speaking: maximizing quadratic forms over the unit box $U$ is provably an NP-hard problem; were we able to get a computationally tractable description of $Z[U]$, 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, “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.

### 2.9.5.1 Simple change detection

In Figure 2.8, you see a sample of frames from a “movie” in which a noisy picture of a dog gradually transforms into a noisy picture of a lady; several initial frames differ just by realizations of noise, and starting from some instant, the “signal” (the deterministic component of the image) starts to drift from the dog 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)

$$\omega_t = x_t + \xi_t, \quad t = 1, 2, ..., K,$$

(2.149)

where $x_t$ are deterministic components of the observations and $\xi_t$ 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 $x_\tau$ differs from $x_{\tau-1}$ (“$\tau$ is a change point”); if it is the case, $\tau$ is uniquely defined by $x^K = (x_1, ..., x_K)$. An alternative is that $x_t$ is independent of $t$, for all $1 \leq t \leq K$ (“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 $\tau$” is understood in the most straightforward way - as the fact that $x_\tau$ 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

$$\omega_t = \sum_{s=1}^{t} A_{ts} x_s + \xi_t, \quad t = 1, ..., K,$$

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
Figure 2.8: Frames from a “movie”
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 (2.149), 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 the assumptions on the noise,
- what happens with \( x_i \)'s after the change – do they jump from their common value prior to time \( \tau \) 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 [52]. 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 (2.149) with independent across \( t = 1, \ldots, K \) noises \( \xi_t \sim N(0, \sigma^2 I_d) \). We do not known \( \sigma \) a priori, what we know is that \( \sigma \) is independent of \( t \) and belongs to a given segment \( [\sigma, \sigma'] \), with \( 0 < \sigma \leq \sigma' \);
2. Observations (2.149) arrive one by one, so that at time \( t, 2 \leq t \leq K \) we have at our disposal observation \( \omega^t = (\omega_1, \ldots, \omega_t) \). Our goal is to build a system of inferences \( T_t, 2 \leq t \leq K \), such that \( T_t \) as applied to \( \omega^t \) 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 our collection of inferences to satisfy the bound \( \epsilon \) on the probability of false alarm (i.e., on the probability of terminating somewhere on time horizon \( 2, 3, \ldots, K \) with “there was a change” conclusion in the situation where there was no change: \( x_1 = \ldots = x_K \)). 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 \( \{T_t : 2 \leq t \leq K\} \) such that

- the probability of false alarm is at most \( \epsilon \), and
- the probability of “\( \rho \)-miss” – the probability to detect no change when there was a change of energy \( \geq \rho^2 \) (i.e., when there was a change a time \( \tau \), and, moreover, it holds \( \|x_\tau - x_1\|^2 \geq \rho^2 \)) is at most \( \varepsilon \).

What we are interested in, is to achieve the just formulated goal with as small \( \rho \) 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 $\|x_t - x_{t-1}\|^2 \leq R^2$.\footnote{$R$ is needed for 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.} Let us associate with $\rho > 0$ “signal hypotheses” $H^\rho_t$, $t = 2, 3, \ldots, K$, on the distribution of observation $\omega^K_t$ given by (2.149), with $H^\rho_t$ stating that at time $t$ there is a change, of energy at least $\rho^2$, in the time series $\{x_t\}_{t=1}^K$ underlying the observation $\omega^K_t$:

$$x_1 = x_2 = \ldots = x_{t-1} \quad \& \quad \|x_t - x_{t-1}\|^2 = \|x_t - x_1\|^2 \geq \rho^2$$

(and on the top of it, $\|x_t - x_{\tau}\|^2 \leq R^2$ for all $t, \tau$). Let us augment these hypotheses by the null hypothesis $H_0$ stating that there is no change at all – the observation $\omega^K_t$ stems from a stationary time series $x_1 = x_2 = \ldots = x_K$. We are about to use our machinery of detector-based tests in order to build a system of tests deciding, with partial risks $\epsilon, \varepsilon$, on the null hypothesis vs. the “signal alternative” $\bigcup_t H^\rho_t$ for as small $\rho$ as possible.

The implementation is as follows. Given $\rho > 0$ such that $\rho^2 < R^2$, consider two hypotheses, $G_1$ and $G_2^\rho$, on the distribution of observation

$$\zeta = x + \xi \in \mathbb{R}^d.$$ 

Both hypotheses state that $\xi \sim N(0, \sigma^2 I_d)$ with unknown $\sigma$ known to belong to a given segment $\Delta := [\sqrt{2} \zeta, \sqrt{2} \bar{\sigma}]$. In addition, $G_1$ states that $x = 0$, and $G_2^\rho$ - that $\rho^2 \leq \|x\|^2 \leq R^2$. We can use the result of Proposition 2.43.ii to build a quadratic in $\zeta$ detector for the families of distributions $P_1, P_2$ obeying the hypotheses $G_1, G_2^\rho$, respectively. To this end it suffices to apply the proposition to the collections

$$\mathcal{V}_\chi = \{ \sigma^2 I_d : \sigma \in \Delta \}, \Theta_{\chi}^{(x)} = 2\pi^2 I_d, \delta_{\chi} = 1 - \sigma / \bar{\sigma}, \gamma_{\chi} = 0.999, A_\chi = I_d, \mathcal{Z}_\chi, \quad [\chi = 1, 2]$$

where

$$\mathcal{Z}_1 = \{ [0; \ldots; 0; 1][0; \ldots; 0; 1]^T \} \subset S^{d+1}_+,$$

$$\mathcal{Z}_2 = \mathcal{Z}_2^\rho = \{ Z \in S^{d+1}_+ : Z_{d+1, d+1} = 1, 1 + R^2 \geq \text{Tr}(Z) \geq 1 + \rho^2 \}.$$ 

The (upper bound on the) risk of the quadratic in $\zeta$ detector yielded by a saddle point of function (2.134), as given by Proposition 2.43.ii, is immediate: by the same argument as used when justifying Remark 2.44, in the situation in question one can look for saddle point with $h = 0$, $H = \eta I_d$, and identifying the required $\eta$ reduces to solving univariate convex problem

$$\text{Opt}(\rho) = \min_{\eta} \frac{1}{2} \left\{ - \frac{\rho}{4} \ln(1 - \bar{\sigma}^4 \eta^2) - \frac{\rho^2}{8} \bar{\sigma}^2 (1 - \bar{\sigma}^2 / \bar{\sigma}^2 \eta^2) \eta + \frac{d(2 + \delta) \bar{\sigma}^4 \eta^2}{1 + \bar{\sigma}^2 \eta} \right\} + \frac{\rho^2 \eta}{\pi(1 - \bar{\sigma}^2 \eta)} : - \gamma \leq \bar{\sigma}^2 \eta \leq 0 \right\} \left[ \bar{\sigma} = \sqrt{2} \bar{\sigma}, \delta = 1 - \sigma / \bar{\sigma} \right]$$

which can be done in no time by Bisection. The resulting detector and the upper bound on its risk are given by the optimal solution $\eta(\rho)$ to the latter problem.
that there was no change, and proceed to time change has already taken place and terminate, otherwise we claim that so far, there 
and compute the quantity \( \phi \)
ming from (2.149), let us form the vector 

\[
\phi^*\left(\zeta\right) = \frac{1}{2} \eta(\rho) z^T \zeta + \frac{d}{4} \left[ \ln \left( \frac{1 - \hat{\sigma}^2 \eta(\rho)}{1 + \hat{\sigma}^2 \eta(\rho)} \right) - \hat{\sigma}^2 (1 - \sigma^2 / \hat{\sigma}^2) \eta(\rho) - \frac{\rho^2 \eta(\rho)}{d(1 - \hat{\sigma}^2 \eta(\rho))} \right]_{a(\rho)}
\]

with 

\[
\text{Risk}\left[\phi^*_{\rho}, P_1, P_2\right] \leq \text{Risk}(\rho) := c^{\text{opt}(\rho)}
\]

(observe that \( R \) does not appear nor in the definition of the optimal detector nor in the risk bound). It is immediately seen that \( \text{Opt}(\rho) \to 0 \) as \( \rho \to +\infty \) and \( \text{Opt}(\rho) \to -\infty \) as \( \rho \to +\infty \), implying that given \( \kappa \in (0,1) \), we can easily find by bisection \( \rho = \rho(\kappa) \) such that \( \text{Risk}(\rho) = \kappa \); in what follows, we assume w.l.o.g. that \( R > \rho(\kappa) \) for the value of \( \kappa \) 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 \( T^\kappa \) which, given observation \( \zeta \), 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

\[
\text{Risk}_1(T^\kappa | G_1, G_2^{\rho(\kappa)}) \leq \frac{\kappa^2}{\varepsilon}, \quad \text{Risk}_2(T^\kappa | G_1, G_2^{\rho(\kappa)}) \leq \varepsilon,
\]

see Proposition 2.14 and (2.48).

We are nearly done. Given \( \kappa \in (0,1) \), consider the system of tests \( T^\kappa_t \), \( t = 2, 3, ..., K \), as follows. At time \( t \in \{2, 3, ..., K\} \), given observations \( \omega_1, ..., \omega_t \) stemming from (2.149), 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 \( t = 2, 3, ..., K \):

- if there is no change on time horizon \( 1, ..., t \): \( x_1 = x_2 = ... = x_t \) (case A) the probability for \( T^\kappa_t \) to conclude that there was a change is at most \( \kappa^2 / \varepsilon \);
- if, on the other hand, \( \|x_t - x_1\|^2 \geq \rho^2(\kappa) \) (case B), then the probability for \( T^\kappa_t \) to conclude that so far there was no change is at most \( \varepsilon \).

Indeed, we clearly have

\[
\zeta_t = [x_t - x_1] + \xi_t,
\]

where \( \xi_t = \xi_t - \xi_1 \sim N(0, \sigma^2 I_d) \) with \( \sigma \in [\sqrt{2}\sigma, \sqrt{2}\sigma] \). Our action at time \( t \) is nothing but application of the test \( T^\kappa \) to the observation \( \zeta_t \). In case A the distribution of this observation obeys the hypothesis \( G_1 \), and the probability for \( T^\kappa_t \) to claim that there was a change is at most \( \kappa^2 / \varepsilon \) by the first inequality
HYPOTHESIS TESTING

In (2.151). In case B, the distribution of $\zeta$ obeys the hypothesis $G_2^{\rho(\kappa)}$, and thus the probability for $T^*_{\kappa}$ to claim that there was no change on time horizon $1,\ldots,t$ is $\leq \varepsilon$ by the second inequality in (2.151).

In view of (!), the probability of false alarm for the system of inferences $\{T^*_{\kappa}\}_{i=2}^K$ is at most $(K-1)\kappa^2/\varepsilon$, and specifying $\kappa$ as

$$\kappa = \sqrt{\varepsilon/(K-1)},$$

we make this probability $\leq \varepsilon$. The resulting procedure, by the same (!), detects a change at time $t \in \{2,3,\ldots,K\}$ with probability at least $1-\varepsilon$, provided that the energy of this change is at least $\rho^2_*$, with

$$\rho_* = \rho \left(\sqrt{\varepsilon/(K-1)}\right),$$

(2.152)

In fact we can say a bit more:

**Proposition 2.48.** Let the deterministic sequence $x_1,\ldots,x_K$ underlying observations (2.149) be such that for some $t$ it holds $||x_t-x_1||^2_2 \geq \rho^2_*$, with $\rho_*$ given by (2.152). 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 $T^*_{\kappa}$ 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 $2,3,\ldots,t$.

**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 $256 \times 256$, 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 $256 \times 256$ 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 $\sigma$ to be 10, and used $\sigma = \sigma/\sqrt{2}$. The reliability tolerances $\epsilon, \varepsilon$ 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 $\sigma^2 = 100$). The resulting detector is

$$\phi_*(\zeta) = -2.7138 \frac{\zeta^T \zeta}{10^6} + 366.9548.$$

In other words, test $T^*_{\kappa}$ 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 is obtained from the “basic sequence”

$$\bar{x}_t = D + 0.0357(t-1)(L-D), t = 1,2,\ldots$$

(2.153)
where \( D \) is the image of the dog and \( L \) is the image of the lady (up to noise, these are the first and the last frames on Figure 2.8). To get the observations in a particular simulation, we augment this sequence from the left by a random number of images \( D \) in such a way that with probability \( 1/2 \) there was no change of image on the time horizon \( 1,2,...,9 \), and with probability \( 1/2 \) there was a change at time instant \( \tau \) chosen at random from the uniform distribution on \( \{2,3,...,9\} \).

The observation is obtained by taking the first 9 images in the resulting sequence, and adding to them independent across the images observation noises drawn at random from \( \mathcal{N}(0,100I_{65536}) \).

In the series of 3000 simulations of this type we have not observed a single false alarm, while the empirical probability of a miss was 0.0553. Besides, the change at time \( t \), if detected, was never detected with a delay more than 1.

Finally, in the particular “movie” in Figure 2.8 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 when you managed to detect the change?

“Numerical near-optimality.” Recall that 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, that of deciding upon hypotheses \( G_1 \) and \( G_2^o \) via observation \( (2.150) \), and suppose that we want 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 \( \rho \); let us denote by \( \rho^+ \) the smallest \( \rho \) for which we can meet the risk specification with our detector-based approach (\( \rho^+ \) is nothing but what was above called \( \rho(\epsilon) \)), and by \( \bar{\rho} \) the smallest \( \rho \) for which “in the nature” there exists a simple test deciding on \( G_1 \) vs. \( G_2^o \) with risk \( \leq \epsilon \). We can consider the ratio \( \rho^+ / \bar{\rho} \) as the “index of conservatism” of our approach. Now, \( \rho^+ \) is given by an efficient computation; what about \( \rho^+ / \bar{\rho} \)? Well, there is a simple way to get a lower bound on \( \rho^+ / \bar{\rho} \), namely, as follows. Observe that if the composite hypotheses \( G_1, G_2^o \) can be decided upon with risk \( \leq \epsilon \), the same holds true for two simple hypotheses stating that the distribution of observation \( (2.150) \) is \( P_1 \), respectively, \( P_2 \), where \( P_1, P_2 \) correspond to the cases where

- \( (P_1) \): \( \zeta \) is drawn from \( \mathcal{N}(0,2\pi^2I_d) \)
- \( (P_2) \): \( \zeta \) is obtained by adding \( \mathcal{N}(0,2\pi^2I_d) \)-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 \( G_2^o \); as a result, a simple test \( T \) deciding \((1-\epsilon)\)-reliably on \( G_1 \) vs. \( G_2^o \) would induce a test deciding equally reliably on \( P_1 \) vs. \( P_2 \), specifically, the test which, given observation \( \zeta \), accepts \( P_1 \) if \( T \) on the same observation accepts \( G_1 \), and accepts \( P_2 \) otherwise.

We can now use two-point lower bound (Proposition 2.2) to lower-bound the risk of deciding on \( P_1 \) vs. \( P_2 \). Because 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

\[ \rho_{\text{smallest}} \]

The coefficient 0.0357 corresponds to 28-frame linear transition from \( D \) to \( L \).
easy to compute. The resulting lower risk bound depends on \( \rho \), and we can find the smallest \( \rho \) for which the bound is \( \geq 0.01 \), and use this \( \rho \) in the role of \( \rho \); 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, \( \epsilon = 0.01 \), \( d = 256^2 = 65536 \), \( \sigma = 10 \), \( \bar{\sigma} = \sigma / \sqrt{2} \). Computation shows that in this case we have

\[
\rho^+ = 2702.4, \quad \rho^+ / \bar{\rho} \leq 1.04
\]

– nearly no conservatism at all! When eliminating the uncertainty in the intensity of noise by increasing \( \bar{\sigma} \) from \( \sigma / \sqrt{2} \) to \( \sigma \), we get

\[
\rho^+ = 668.46, \quad \rho^+ / \bar{\rho} \leq 1.15
\]

– still not that much of conservatism!

2.10 EXERCISES FOR CHAPTER 2

2.10.1 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 \( \Omega = \{1, \ldots, d\} \), 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?

2.10.2 Around Euclidean Separation

**Exercise 2.2.**

Justify the “immediate observation” in Section 2.2.2.3.B.

**Exercise 2.3.**

1) Prove Proposition 2.9.

**Hint:** You can find useful the following simple observation (prove it, provided you indeed use it):

Let \( f(\omega), g(\omega) \) be probability densities taken w.r.t. a reference measure \( P \) on an observation space \( \Omega \), and let \( \epsilon \in (0, 1/2] \) be such that

\[
2\bar{\epsilon} := \int_{\Omega} \min[f(\omega), g(\omega)]P(d\omega) \leq 2\epsilon.
\]

Then

\[
\int_{\Omega} \sqrt{f(\omega)g(\omega)}P(d\omega) \leq 2\sqrt{\epsilon(1-\epsilon)}.
\]
2) Justify the illustration in Section 2.2.3.2.C.

2.10.3 Hypothesis testing via $\ell_1$-separation

Let $d$ be a positive integer, and the observation space $\Omega$ be the finite set $\{1, \ldots, d\}$ equipped with the counting reference measure. Probability distributions on $\Omega$ can be identified with points $p$ of $d$-dimensional probabilistic simplex

$$\Delta_d = \{ p \in \mathbb{R}^d : p \geq 0, \sum_i p_i = 1 \};$$

$i$-th entry $p_i$ in $p \in \Delta_d$ is the probability for the distributed according to $p$ random variable to take value $i \in \{1, \ldots, d\}$. With this interpretation, $p$ is the probability density taken w.r.t. the counting measure on $\Omega$.

Assume $B$ and $W$ are two nonintersecting nonempty closed convex subsets of $\Delta_d$; we interpret $B$ and $W$ as black and white probability distributions on $\Omega$, and our goal is to find optimal, in terms of its total risk, test deciding on the hypotheses

$$H_1 : p \in B, \quad H_2 : p \in W$$

via a single observation $\omega \sim p$.

Warning: Everywhere in this section, “test” means “simple test.”

Exercise 2.4.

Our first goal is to find optimal, in terms of its total risk, test deciding on the hypotheses $H_1, H_2$ via a single observation $\omega \sim p \in B \cup W$. To this end we consider the convex optimization problem

$$\text{Opt} = \min_{p \in B, q \in W} \left[ f(p, q) := \sum_{i=1}^d |p_i - q_i| \right] \quad (2.154)$$

and let $(p^*, q^*)$ be an optimal solution to this problem (it clearly exists).

1. Extract from optimality conditions that there exist reals $\rho_i \in [-1, 1], 1 \leq i \leq n$, such that

$$\rho_i = \begin{cases} 1, & p^*_i > q^*_i \\ -1, & p^*_i < q^*_i \end{cases} \quad (2.155)$$

and

$$\rho^T (p - p^*) \geq 0 \forall p \in B \& \rho^T (q - q^*) \leq 0 \forall q \in W. \quad (2.156)$$

2. Extract from the previous item that the test $T$ which, given an observation $\omega \in \{1, \ldots, d\}$, accepts $H_1$ with probability $\pi_\omega = (1 + \rho_\omega)/2$ and accepts $H_2$ with complementary probability, has its total risk equal to

$$\sum_{\omega \in \Omega} \min[p^*_\omega, q^*_\omega], \quad (2.157)$$

and thus is minimax optimal in terms of the total risk.

---

Counting measure is the measure on a discrete (finite or countable) set $\Omega$ which assigns every point of $\Omega$ with mass 1, so that the measure of a subset of $\Omega$ is the cardinality of the subset when it is finite and is $+\infty$ otherwise.
Comments. Exercise 2.4 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 typically 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 \( \omega_1, \ldots, \omega_K \) of the random variable in question. And construction presented in Exercise 2.4 says nothing on how to adjust the test to the case of repeated observation. Of course, when \( \omega^K = (\omega_1, \ldots, \omega_K) \) is \( K \)-element i.i.d. sample drawn from a probability distribution \( p \) on \( \Omega = \{1, \ldots, d\} \), \( \omega^K \) can be thought of as a single observation of discrete random variable taking value in the set \( \Omega^K = \underbrace{\Omega \times \ldots \times \Omega}_K \), the probability distribution \( p^K \) of \( \omega^K \) being readily given by \( p \). So, why not to apply the construction from Exercise 2.4 to \( \omega^K \) in the role of \( \omega \)? On a close inspection, this idea fails. One of the reasons for this failure is that the cardinality of \( \Omega^K \) (which, among other factors, is responsible for the computational complexity of implementing the test in Exercise 2.4) 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 \( \omega^K \) induced by a convex family of distributions \( p \) of \( \omega \), 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 2.4 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 2.4.30

In spite of the above pessimistic conclusions, the single-observation test from Exercise 2.4 admits a meaningful multi-observation modification, which is the subject of our next exercise.

Exercise 2.5.

There is a straightforward way to use the optimal, in terms of its total risk, single-observation test built in Exercise 2.4 in the “multi-observation” environment. Specifically, following the notation from Exercise 2.4, let \( \rho \in \mathbb{R}^d, p^*, q^* \) be the entities built in this Exercise, so that \( p^* \in B, q^* \in W \), all entries in \( \rho \) 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 \( \omega^K = (\omega_1, \ldots, \omega_K) \) with \( \omega_i \sim p \), 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 \( \omega^K \), 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

30 Though directly extending the optimal single-observation test to the case of repeated observations encounters significant technical difficulties, it was carried on in some specific situations. For instance, in [119, 120] such extension has been proposed for the case of sets \( B \) and \( W \) of distributions which are dominated by bi-alternating capacities (see, e.g., [8, 12, 35], and references therein); explicit constructions of the test were proposed for some special sets of distributions [118, 193, 206].
\( p \in W. \) Consequently, selecting a “comparison level” \( \ell \in (\beta, \alpha), \) we can decide on the hypotheses \( p \in B \) vs. \( p \in W \) by computing \( \zeta^K, \) comparing the result to \( \ell, \) accepting the hypothesis \( p \in B \) when \( \zeta^K \geq \ell, \) and accepting the alternative \( p \in W \) otherwise. 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 \( \mathbb{P}_{pK} \{\zeta^K \leq \ell\}? \)

B. For \( p \in W, \) how to upper-bound the probability \( \mathbb{P}_{pK} \{\zeta^K \geq \ell\}? \)

Here \( p_K \) is the probability distribution of the i.i.d. sample \( \omega^K = (\omega_1, \ldots, \omega_K) \) with \( \omega_i \sim p. \)

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 \( \{1, 2, \ldots, d\} \) it holds

\[
\mathbb{P}_{pK} \{\zeta^K \leq \ell\} \exp\{-\gamma \ell\} \leq \mathbb{E}_{pK} \{\exp\{-\gamma \zeta^K\}\} = \left[\sum_{i=1}^d p_i \exp\left\{-\frac{1}{K} \gamma \rho_i\right\}\right]^K,
\]

whence

\[
\ln(\pi_{K,-[p]} \leq K \ln \left(\sum_{i=1}^d p_i \exp\left\{-\frac{1}{K} \gamma \rho_i\right\}\right) + \gamma \ell,
\]

implying, via substitution \( \gamma = \mu K, \)

\[
\forall \mu \geq 0 : \ln(\pi_{K,-[p]} \leq K \psi_-(\mu, p), \psi_-(\mu, p) = \ln \left(\sum_{i=1}^d p_i \exp\left\{-\mu \rho_i\right\}\right) + \mu \ell.
\]

Similarly, setting \( \pi_{K,+[p]} = \mathbb{P}_{pK} \{\zeta^K \geq \ell\}, \) we get

\[
\forall \nu \geq 0 : \ln(\pi_{K,+[p]} \leq K \psi_+(\nu, p), \psi_+(\nu, p) = \ln \left(\sum_{i=1}^d p_i \exp\left\{\nu \rho_i\right\}\right) - \nu \ell.
\]

Now goes the exercise:

1. Extract from the above observations that

\[
\text{Risk}(T^{K,\ell}|H_1, H_2) \leq \exp\{K \kappa\}, \quad \kappa = \max \left[\max_{p \in B} \inf_{\mu \geq 0} \psi_-(\mu, p), \max_{q \in W} \inf_{\nu \geq 0} \psi_+(\nu, q)\right],
\]

where \( T^{K,\ell} \) 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 \( \mu \) and concave in \( p, \) and similarly for \( \psi_+(\nu, q), \) so that

\[
\max \inf_{p \in B} \psi_-(\mu, p) = \inf_{\mu \geq 0} \max_{p \in B} \psi_-(\mu, p), \quad \max \inf_{q \in W} \psi_+(\nu, q) = \inf_{\nu \geq 0} \max_{q \in W} \psi_+(\nu, q).
\]

Thus, computing \( \kappa \) 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

$$\kappa \leq - \frac{1}{12} \Delta^2, \quad \Delta = \alpha - \beta = \|p^* - q^*\|_1.$$ 

Note that the above test and the quantity $\kappa$ 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 $\ell$ is to minimize $\kappa$ over an equidistant grid $\Gamma \subset (\beta, \alpha)$, of small cardinality, of values of $\ell$.

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 risk randomized single-observation, test\footnote{This test can differ from the test built in Exercise 2.4 – the latter test is optimal in terms of the sum, rather than the maximum, of its partial risks.} 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 2.6.**

Let, as above, $B$ and $W$ be nonempty nonintersecting closed convex subsets of probabilistic simplex $\Delta_d$. Show that the problem of finding the best, in terms of its risk, randomized single-observation test deciding on $H_1 : p \in B$ versus $H_2 : p \in W$ 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:

$$B = \{ p : \exists u : P_B p + Q_B u \leq a_B \},$$

$$W = \{ p : \exists u : P_W p + Q_W u \leq a_W \}.$$ 

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, is the risk of this test $< \frac{1}{2}$ – this is what we need for the majority version of the minimum-risk single-observation test to be consistent.

**Exercise 2.7.**

Extract from Exercise 2.4 that in the situation of this section, denoting by $\Delta$ the optimal value in the optimization problem (2.154), one has

1. The risk of any single-observation test, deterministic or randomized, 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 2.6 does not exceed $\frac{1}{2} - \frac{\Delta}{8} < 1/2$ as well.

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 $\epsilon$ 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 2.4, and then to pass to its multi-observation version built in Exercise 2.5;
2. To use the majority version of the minimum-risk randomized single-observation test built in Exercise 2.6;
3. To use the test based on the minimum risk detector for $B,W$, as explained in the main body of Chapter 2.

In all cases, we have to specify the number $K$ of observations which guarantees that the risk of the resulting multi-observation test is at most a given target $\varepsilon$. A bound on $K$ can be easily obtained by utilizing the results on the risk of a detector-based test in a Discrete o.s. from the main body of Chapter 2 along with risk-related results of Exercises 2.5, 2.6 and 2.7.

**Exercise 2.8.**

Run numerical experimentation to get see if one of the three options above always dominates the others (that is, requires smaller sample of observations to ensure the same risk).

Let us now 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 2.10.3. Given $\varepsilon \in (0,1)$, the corresponding sample sizes $K_d$ and $K_m$ are completely determined by the relevant “measure of closeness” between $B$ and $W$. Specifically,

- For $K_d$, the closeness measure is
  \[
  \rho_d(B,W) = 1 - \max_{p \in B, q \in W} \sum_{\omega} \sqrt{p_\omega q_\omega};
  \]
  $1 - \rho_d(B,W)$ is the minimal risk of a detector for $B,W$, and for $\rho_d(B,W)$ and $\varepsilon$ small, we have $K_d \approx \ln(1/\varepsilon)/\rho_d(B,W)$ (why?).

- Given $\varepsilon$, $K_m$ is fully specified by the minimal risk $\rho$ of simple randomized single-observation test $T$ deciding on the associated with $B,W$ hypotheses. By Exercise 2.7, we have $\rho = \frac{1}{2} - \delta$, where $\delta$ is within absolute constant factor of the optimal value $\Delta = \min_{p \in B, q \in W} \|p - q\|_1$ of (2.154). The risk bound for the $K$-observation majority version of $T$ 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 $\rho$ is not close to $0$ and $\varepsilon$ is small, the $(1-\varepsilon)$-quantile of the number of heads in our $K$ coin tosses is $K\rho + O(1)\sqrt{K \ln(1/\varepsilon)} = K/2 - \delta K + O(1)\sqrt{K \ln(1/\varepsilon)}$ (why?). $K_m$ is the smallest $K$ for which this quantile is $< K/2$, so that $K_m$ is of the order of $\ln(1/\varepsilon)/\delta^2$, or, which is the same, of the order of $\ln(1/\varepsilon)/\Delta^2$. We see that the “responsible for $K_m$” closeness between $B$ and $W$ is
  \[
  \rho_m(B,W) = \Delta^2 = \left[ \min_{p \in B, q \in W} \|p - q\|_1 \right]^2,
  \]
  and $K_m$ is of the order of $\ln(1/\varepsilon)/\rho_m(B,W)$.

The goal of the next exercise is to compare $\rho_b$ and $\rho_m$.

**Exercise 2.9.**
Prove that in the situation of this Section one has
\[ \frac{1}{2} \rho_m(B, W) \leq \rho_d(B, W) \leq \frac{1}{2} \sqrt{\rho_m(B, W)}. \]  
(2.159)

Relation (2.159) suggests that while \( K_d \) never is “much larger” than \( K_m \) (this we know in advance: in repeated version of Discrete o.s., properly built detector-based test provably is nearly optimal), \( K_m \) might be much larger than \( K_d \). This indeed is the case:

Exercise 2.10.

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 \( K_d \) and \( K_m \) resulting in a given risk \( \epsilon \ll 1 \) of multi-observation tests, as functions of \( \delta \) are proportional to \( 1/\delta \) and \( 1/\delta^2 \), respectively. Compare the numbers when \( \epsilon = 0.01 \) and \( \delta \in \{0.01; 0.05; 0.1\} \).

2.10.4 Miscellaneous exercises

Exercise 2.11.

Prove that the conclusion in Proposition 2.18 remains true when the test \( T \) in the premise of the proposition is randomized.

Exercise 2.12.

Let \( p_1(\omega), p_2(\omega) \) be two positive probability densities, taken w.r.t. a reference measure \( \Pi \) on an observation space \( \Omega \), and let \( P_\chi = \{p_\chi\}, \chi = 1, 2 \). Find the optimal, in terms of its risk, balanced detector for \( P_\chi, \chi = 1, 2 \).

Exercise 2.13.

Recall that the exponential, with parameter \( \mu > 0 \), distribution on \( \Omega = \mathbb{R}_+ \) 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, \( P_1 = \{p_\mu : 0 < \mu \leq \alpha\} \), and \( P_2 = \{p_\mu : \mu \geq \beta\} \). Build the optimal, in terms of its risk, balanced detector for \( P_1, P_2 \). What happens with the risk of the detector you have built when the families \( P_\chi, \chi = 1, 2 \), are replaced with their convex hulls?

Exercise 2.14.

[Follow-up to Exercise 2.13] Assume that the “lifetime” \( \zeta \) 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 lifespan of a lightbulb in this model is \( 1/\mu \)).\(^{32}\) 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

\(^{32}\)In Reliability, probability distribution of the lifespan \( \zeta \) of an organism or a technical device is characterized by the failure rate \( \lambda(t) = \lim_{\delta \to 0} \frac{\text{Prob} \{t \leq \zeta \leq t + \delta\}}{\text{Prob} \{\zeta \geq t\} \cdot \delta} \) (so that for small \( \delta \), \( \lambda(t) \delta \) is the conditional probability to “die” in the time interval \( [t, t + \delta] \) provided the organism or device is still “alive” at time \( t \)). The exponential distribution corresponds to the case of failure rate independent of \( t \); in applications, this indeed is often the case except for “very small” and “very large” values of \( t \).
conclusion? Answer this question in the situations when the observation \( \omega \) in a test is

1. the lifespan of a lightbulb (i.e., \( \omega \sim p_\mu(\cdot) \))
2. the minimum \( \omega = \min[\zeta, \delta] \) of the lifespan \( \zeta \sim p_\mu(\cdot) \) and the allowed duration \( \delta > 0 \) of your test (i.e., if the lightbulb you are testing does not “die” on time horizon \( \delta \), 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 lifespan of a lightbulb, and \( \delta > 0 \) is the allowed test duration (i.e., you observe whether or not a lightbulb “dies” on time horizon \( \delta \), but do not register the lifespan when it is \( < \delta \)).

Consider the values 0.25, 0.5, 1, 2, 4 of \( \delta \).

Exercise 2.15.

[Follow-up to Exercise 2.14] In the situation of Exercise 2.14, build a sequential test for deciding on null hypothesis “the lifespan 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 lifespan 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 \( \mu \). 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 the exercise.

Exercise 2.16.

In Section 2.6, we consider 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 2.17.

Work out the following extension of the Opinion Poll Design problem. You are given two finite sets, \( \Omega_1 = \{1, ..., I\} \) and \( \Omega_2 = \{1, ..., M\} \), along with \( L \) nonempty closed convex subsets \( Y_\ell \) of the set

\[
\Delta_{IM} = \left\{ y_{im} > 0 | y_{im}: \sum_{i=1}^{I} \sum_{m=1}^{M} y_{im} = 1 \right\}
\]

of all non-vanishing probability distributions on \( \Omega = \Omega_1 \times \Omega_2 = \{(i, m) : 1 \leq i \leq I, 1 \leq m \leq M\} \). Sets \( Y_\ell \) are such that all distributions from \( Y_\ell \) 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 $\omega_1, \omega_2, \ldots$ 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 $Q$ of the positive part of $I$-dimensional probabilistic simplex. Let $y_{i\ell}$ be the conditional, $i$ being given, distribution of $m \in \Omega_2$ induced by $y$, so that $y_{i\ell} = y_{im}$.

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 $y_{i\ell}$.

Given closeness relation $C$, your goal is to decide, up to closeness $C$, on the hypotheses $H_1, \ldots, H_L$, with $H_\ell$ stating that the distribution $y$ selected by the nature belongs to $Y_\ell$. Given "observation budget” (a number $K$ of observations $\omega_k$ you can use), you want to find a probabilistic vector $q$ which results in the test with as small $C$-risk as possible. Pose this Measurement Design problem as an efficiently solvable convex optimization problem.

Exercise 2.18.

[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 $\mathcal{H} \subset \Omega = \mathbb{R}^d, M, \Phi$ be a regular data such that $0 \in \text{int} \mathcal{H}, M$ is compact, $\Phi(0; \mu) = 0 \forall \mu \in M$, and $\Phi(h; \mu)$ is differentiable at $h = 0$ for every $\mu \in M$. Let, further, $\bar{P} \in S[\mathcal{H}, M, \Phi]$ and let $\bar{\mu} \in M$ 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 $\Omega$ it holds
   
   \[ \pi := \bar{P}\{ \omega : e^T(\omega - e[\bar{P}]) \geq 1 \} \leq \exp \left\{ \inf \left\{ t \geq 0 : te \in \mathcal{H} \right\} \Phi(te; \bar{\mu}) - te^T \nabla_h \Phi(0; \bar{\mu}) - t \right\} \] (2.160)

What are the consequences of (2.160) for sub-Gaussian distributions?

Exercise 2.19.

[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 $U_\chi \subset \mathbb{R}^{n_\chi}$
- regular data $\mathcal{H}_k \subset \mathbb{R}^{d_k}, M_k, \Phi_k$, and affine mappings
  
  $u_\chi \mapsto A_k[u_\chi; 1] : \mathbb{R}^{n_\chi} \to \mathbb{R}^{d_k}$

such that

$u_\chi \in U_\chi \Rightarrow A_k[u_\chi; 1] \in M_k$

$1 \leq k \leq K, 1 \leq \ell \leq L$. 


3. \( \bar{\Theta} = \Theta + 1 \)

We can associate with the outlined data families of probability distributions \( \mathcal{P}_x \) on the observation space \( \Omega = \mathbb{R}^d \times \ldots \times \mathbb{R}^{d_K} \) as follows. For \( x = 1, 2 \), \( \mathcal{P}_x \) is comprised of all probability distributions \( P \) of random vectors \( \omega^K = [\omega_1; \ldots; \omega_K] \in \Omega \) generated as follows:

We select

- a signal \( u_x \in U_x \),
- a collection of probability distributions \( P_{k\ell} \in \mathcal{S}[\mathcal{H}^x_{k\ell}, \mathcal{M}^x_{k\ell}, \Phi^x_{k\ell}], 1 \leq k \leq K, 1 \leq \ell \leq L \), in such a way that \( A^x_{k\ell}[u_x; 1] \) is a parameter of \( P_{k\ell} \):

\[
\forall h \in \mathcal{H}_{k\ell}^x : \ln \left( \mathbb{E}_{\omega_k \sim P_{k\ell}} \{ e^{h^T \omega_k} \} \right) \leq \Phi^x_{k\ell}(h_k; A^x_{k\ell}[u_x; 1]);
\]

- we generate the components \( \omega_k, k = 1, \ldots, K \), independently across \( k \), from \( \mu^k \)-mixture \( \Pi \{ P_{k\ell} \} L \), of distributions \( P_{k\ell}, \ell = 1, \ldots, L \), that is, draw at random, from distribution \( \mu^k \) on \( \{ 1, \ldots, L \} \), index \( \ell \), and then draw \( \omega_k \) from the distribution \( P_{k\ell} \).

Prove that when setting

\[
\begin{align*}
\mathcal{H}_x &= \{ h = [h_1; \ldots; h_K] \in \mathbb{R}^{d_1 + \ldots + d_K} : h_k \in \bigcap_{\ell=1}^L \mathcal{H}^x_{k\ell}, 1 \leq k \leq K \}, \\
\mathcal{M}_x &= \{ 0 \} \subset \mathbb{R}, \\
\Phi_x(h; \mu) &= \sum_{k=1}^K \ln \left( \sum_{\ell=1}^L \mu^k \exp \left( \max_{u_x \in U_x} \Phi^x_{k\ell}(h_k; A^x_{k\ell}[u_x; 1]) \right) \right) : \mathcal{H}_x \times \mathcal{M}_x \to \mathbb{R},
\end{align*}
\]

we obtain the regular data such that

\[
\mathcal{P}_x \subset \mathcal{S}[\mathcal{H}_x, \mathcal{M}_x, \Phi_x].
\]

Exercise 2.20.

[mixture of sub-Gaussian distributions] Let \( P_\ell \) be sub-Gaussian distributions on \( \mathbb{R}^d \) with sub-Gaussianity parameters \( \theta_\ell, \Theta \), \( 1 \leq \ell \leq L \), with a common \( \Theta \)-parameter, and let \( \nu = [\nu_1; \ldots; \nu_L] \) be a probabilistic vector. Consider the \( \nu \)-mixture \( P = \Pi [ P^L, \nu ] \) of distributions \( P_\ell \), so that \( \omega \sim P \) is generated as follows: we draw at random from distribution \( \nu \) index \( \ell \) and then draw \( \omega \) at random from distribution \( P_\ell \). Prove that \( P \) is sub-Gaussian with sub-Gaussianity parameters \( \bar{\Theta} = \sum_\ell \nu_\ell \theta_\ell \) and \( \Theta \), with (any) \( \Theta \) chosen to satisfy

\[
\bar{\Theta} \geq \Theta + 6 \frac{\theta_\ell - \bar{\theta}}{\theta_\ell - \bar{\theta}} T \forall \ell,
\]

in particular, according to any one of the following rules:

1. \( \bar{\Theta} = \Theta + \left( 6 \max_\ell \| \theta_\ell - \bar{\theta} \|_2 \right) I_d \),
2. \( \bar{\Theta} = \Theta + \left( 6 \sum_\ell (\theta_\ell - \bar{\theta}) (\theta_\ell - \bar{\theta})^T \right) \),
3. \( \bar{\Theta} = \Theta + \frac{6}{5} \sum_\ell \theta_\ell \theta_\ell^T \), provided that \( \nu_1 = \ldots = \nu_L = 1/L \).

Exercise 2.21.
HYPOTHESIS TESTING

The goal of this exercise is to give a simple sufficient condition for quadratic lifting “to work” in the Gaussian case. Namely, let \( A_\chi, \ U_\chi, \ V_\chi, \ G_\chi, \ \chi = 1, 2, \) be as in Section 2.9.3, with the only difference that now we do not assume the compact sets \( U_\chi \) to be convex, and let \( Z_\chi \) be convex compact subsets of the sets \( Z^{n_\chi}, \) see item i.2. in Proposition 2.43, such that

\[
[u_\chi; 1][u_\chi; 1]^T \in Z_\chi \ \forall u_\chi \in U_\chi, \ \chi = 1, 2.
\]

Augmenting the above data with \( \Theta_\chi^{(*)}, \ \delta_\chi \) such that \( V = V_\chi, \ \Theta_* = \Theta_\chi^{(*)}, \ \delta = \delta_\chi \) satisfy (2.129), \( \chi = 1, 2, \) and invoking Proposition 2.43.ii, we get at our disposal a quadratic detector \( \phi_{\text{lift}} \) such that

\[
\text{Risk}[\phi_{\text{lift}}|G_1, G_2] \leq \exp\{\text{SadVal}_{\text{lift}}\},
\]

with \( \text{SadVal}_{\text{lift}} \) given by (2.134). A natural question is, when \( \text{SadVal}_{\text{lift}} \) 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 are consistent – able to decide on the hypotheses \( H_\chi : P \in G_\chi, \ \chi = 1, 2, \) 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 \( V_1 = V_2 = \{\Theta_*\}, \) which allows to set \( \Theta_\chi^{(*)} = \Theta_* , \ \delta_\chi = 0, \ \chi = 1, 2. \) Prove that in this case a necessary and sufficient condition for \( \text{SadVal}_{\text{lift}} \) to be negative is that the convex compact sets

\[
U_\chi = \{B_\chi Z B_\chi^T : Z \in Z_\chi\} \subset S^{d+1}_+, \ \chi = 1, 2
\]

do not intersect with each other.

Exercise 2.22.

Prove that if \( X \) is a nonempty convex compact set in \( \mathbb{R}^d, \) then the function \( \hat{\Phi}(h; \mu) \) given by (2.114) is real-valued and continuous on \( \mathbb{R}^d \times X \) and is convex in \( h \) and concave in \( \mu. \)

Exercise 2.23.

The goal of what follows is to refine the change detection procedure (let us refer to it as to “basic” one) developed in Section 2.9.5.1. The idea is pretty simple. With the notation from Section 2.9.5.1, in the basic procedure, when testing the null hypothesis \( H_0 \) vs. signal hypothesis \( H_1^0, \) we look at the difference \( \zeta_t = \omega_t - \omega_1 \) and try to decide whether the energy of the deterministic component \( x_t - x_1 \) of \( \zeta_t \) is 0, as is the case under \( H_0, \) or is \( \geq \rho^2, \) as is the case under \( H_1^0. \) Note that if \( \sigma \in [\underline{\sigma}, \overline{\sigma}] \) is the actual intensity of the observation noise, then the noise component of \( \zeta_t \) is \( \mathcal{N}(0, 2\sigma^2 I_d); \) other things being equal, the larger is the noise in \( \zeta_t, \) the larger should be \( \rho \) to allow for a reliable, with a given reliability level, decision. Now note that under the hypothesis \( H_1^0, \) we have \( x_1 = \ldots = x_{t-1}, \) so that the deterministic component of the difference \( \zeta_t = \omega_t - \omega_1 \) is exactly the same as for the difference \( \zeta_t = \omega_t - \omega_1 - \frac{1}{t-1} \sum_{s=1}^{t-1} \omega_s \), while the noise component in \( \zeta_t \) is \( \mathcal{N}(0, \sigma_t^2 I_d) \) with \( \sigma_t^2 = \sigma^2 + \frac{1}{t-1} \sigma^2 = \frac{1}{t-1} \sigma^2. \) Thus, the intensity of noise in \( \zeta_t \) is at most the
same as in $\zeta_t$, and this intensity, in contrast to that for $\zeta_t$, decreases as $t$ grows.

Here goes the exercise:

Let reliability tolerances $\epsilon, \varepsilon \in (0, 1)$ be given, and let our goal be to design a system of inferences $T_t, t = 2, 3, \ldots, K$, which, when used in the same fashion as tests $T^K_\ell$ were used in the basic procedure, results in false alarm probability at most $\epsilon$ and in probability to miss a change of energy $\geq \rho^2$ at most $\varepsilon$. Needless to say, we want to achieve this goal with as small $\rho$ as possible. Think how to utilize the above observation to refine the basic procedure eventually reducing (and provably not increasing) the required value of $\rho$. Implement the basic and the refined change detection procedures and compare their quality (the resulting values of $\rho$) e.g., on the data used in the experiment reported in Section 2.9.5.1.

2.11 PROOFS

2.11.1 Proof of the observation in Remark 2.8

We have to prove that if $p = [p_1; \ldots; p_K] \in B = [0, 1]^K$ 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

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

(2.161)

(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 (2.161) is exactly the probability, taken w.r.t. the uniform distribution on $B$, of the event “at least $M$ of the events $E_1, \ldots, E_K$ take place”). But the right hand side in (2.31) clearly is nondecreasing in $p \in B$, since $\chi_p$, 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 $p$ increases entrywise. \qed

2.11.2 Proof of Proposition 2.6 in the case of quasi-stationary $K$-repeated observations

2.11.2.A Situation and goal. We are in the case QS, see Section 2.2.3.1, of the setting described in the beginning of Section 2.2.3. It suffices to verify that if $H_\ell$, $\ell \in \{1, 2\}$, is true then the probability for $T^K_{\max}$ to reject $H_\ell$ is at most the quantity $\epsilon_K$ defined in (2.23). 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 \( \theta_k(\zeta^k = (\zeta_1, ..., \zeta_k)) \) taking values in \( \Omega = \mathbb{R}^d \) such that our \( k \)-th observation is \( \omega_k = \theta_k(\zeta^k) \). Additionally, the conditional, \( \zeta^{k-1} \) given, distribution \( P_{\omega_k|\zeta^{k-1}} \) of \( \omega_k \) 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 \( \xi \) belongs to \( \mathcal{P}_1^d \).

- There exist deterministic functions \( \chi_k : \Omega \to \{0,1\} \) and integer \( M, 1 \leq M \leq K \), such that the test \( T_K^{\text{maj}} \), 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 2.6, \( M = \lfloor K/2 \rfloor \) and \( \chi_k(\cdot) \) are in fact independent of \( k \) : \( \chi_k(\omega) = 1 \) if and only if \( \phi(\omega) \leq 0 \).

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

\[
P_{\omega_k|\zeta^{k-1}}\{\omega_k : \chi_k(\omega_k) = 1\} \leq \epsilon_* \forall \zeta^{k-1}.
\]

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

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

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 \( \epsilon_* \) which upper-bounds the risk of the single-observation test underlying \( T_K^{\text{maj}} \).

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

\[
P_{\omega^K}\{\omega^K = (\omega_1, ..., \omega_K) : \text{Card}\{k \leq K : \chi_k(\omega_k) = 1\} \geq M\} \leq \epsilon_M = \sum_{M \leq k \leq K} \binom{K}{k} \epsilon_*^{k}(1 - \epsilon_*)^{K-k}, \tag{2.162}
\]

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

2.11.2.B Achieving the goal, step 1. Our reasoning, inspired by that used to justify Remark 2.8, is as follows. Consider a sequence of random variables \( \eta_k, 1 \leq k \leq K \), uniformly distributed on \([0,1]\) and independent of each other and of

\[33\] as always, given a \( K \)-element sequence, say, \( \zeta_1, ..., \zeta_K \), we write \( \zeta^t, t \leq K \), as a shorthand for the fragment \( \zeta_1, ..., \zeta_t \) of this sequence.

\[34\] 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 \( \chi_k \) 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.
\(\zeta^K\), and consider new driving factors \(\lambda_k = [\zeta_k; \eta_k]\) and new observations\(^{35}\)

\[
\mu_k = [\omega_k = \theta_k(\zeta^k); \eta_k] = \Theta_k(\lambda^k = (\lambda_1, \ldots, \lambda_k))
\] (2.163)
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 \(\lambda^k\), and the conditional, \(\lambda^{k-1} = [\zeta^{k-1}; \eta^{k-1}]\) given, distribution \(P_{\mu_k|\lambda^{k-1}}\) of \(\mu_k\) is the product distribution \(P_{\omega_k|\zeta^{k-1}} \times \mathcal{U}\) on \(\Omega \times [0, 1]\), where \(\mathcal{U}\) is the uniform distribution on \([0, 1]\). In particular,

\[
\pi_k(\lambda^{k-1}) := P_{\mu_k|\lambda^{k-1}}\{\mu_k = [\omega_k; \eta_k] : \chi_k(\omega_k) = 1\} = P_{\omega_k|\zeta^{k-1}}\{\omega_k : \chi_k(\omega_k) = 1\} \leq \epsilon_*.
\] (2.164)

- We have

\[
P_{\lambda^K} \{\lambda^K : \text{Card}\{k \leq K : \psi_k(\mu_k = \Theta_k(\lambda^k)) = 1\} \geq M\} = P_{\omega^K} \{\omega^K = (\omega_1, \ldots, \omega_K) : \text{Card}\{k \leq K : \chi_k(\omega_k) = 1\} \geq M\}
\] (2.165)

where \(P_{\omega^K}\) is as in (2.162), and \(\Theta_k(\cdot)\) is defined in (2.163).

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

\[
\eta_k \leq \gamma_k(\lambda^{k-1}) := \frac{\epsilon_* - \pi_k(\lambda^{k-1})}{1 - \pi_k(\lambda^{k-1})}
\]

and \(\psi^+_k(\lambda^k) = 0\) otherwise.

Let us make the following immediate observations:

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

(B) The conditional, given \(\lambda^{k-1} = [\zeta^{k-1}; \eta^{k-1}]\), probability of the event

\[
\psi^+_k(\lambda^k) = 1
\]

is exactly \(\epsilon_*\).

Indeed, let \(P_{\lambda_k|\lambda^{k-1}}\) be the conditional, given \(\lambda^{k-1}\), distribution of \(\lambda_k\). Let us fix \(\lambda^{k-1}\). The event \(E = \{\lambda_k : \psi^+_k(\lambda^k) = 1\}\), by construction, is the union of two

\(^{35}\)In this display, 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 \(\zeta\) and \(\mu\) are components of \(\lambda\), \(\omega = \Theta(\lambda)\), and \(\omega = \theta(\zeta)\). To remind about these "hidden relations," we sometimes write something like \(\phi(\omega_k = \theta_k(\zeta^k))\) to stress that we are speaking about the value of function \(\phi\) at the point \(\omega_k = \theta_k(\zeta^k)\).
nonoverlapping events:

\[ E_1 = \{ \lambda_k = [\zeta_k; \eta_k] : \chi_k(\theta_k(\zeta_k^k)) = 1 \}, \]

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

Taking into account that the conditional, \( \lambda^{k-1} \) being fixed, distribution of \( \mu_k = [\omega_k = \theta_k(\zeta_k^k); \eta_k] \) is the product distribution \( P_{\omega_k|\zeta^{k-1} \times U} \), we conclude in view of \( (2.164) \) that

\[ P_{\lambda_k|\zeta^{k-1}}(E_1) = P_{\omega_k|\zeta^{k-1}}(\omega_k : \chi_k(\omega_k) = 1) = \pi_k(\lambda^{k-1}), \]

\[ P_{\lambda_k|\zeta^{k-1}}(E_2) = P_{\omega_k|\zeta^{k-1}}(\omega_k : \chi_k(\omega_k) = 0) U(\eta \leq \gamma_k(\lambda^{k-1})) \]

\[ = (1 - \pi_k(\lambda^{k-1})) \gamma_k(\lambda^{k-1}), \]

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

**2.11.2.C Achieving the goal, step 2.** By (A) combined with \( (2.165) \) we have

\[ P_{\omega_k \{\lambda^K : \text{Card}\{k \leq K : \chi_k(\omega_k) = 1\} \geq M\}} \]

\[ = P_{\lambda^K \{\lambda^K : \text{Card}\{k \leq K : \psi_k(\mu_k = \Theta_k(\lambda^k)) = 1\} \geq M\}} \]

\[ \leq P_{\lambda^K \{\lambda^K : \text{Card}\{k \leq K : \psi_k^+(\lambda^k) = 1\} \geq M\}}, \]

and all we need to verify is that the first quantity in this chain is upper-bounded by the quantity \( \epsilon_M \) given by \( (2.162) \). Invoking \( (B) \), it is enough to prove 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, \( \lambda^{k-1} \) being fixed, probability for \( \psi_k^+(\lambda^k) \) to take value 1 is, for all \( \lambda^{k-1} \), equal to \( \epsilon_+ \). Then the \( P \)-probability of the event

\[ \{\lambda^K : \text{Card}\{k \leq K : \psi_k^+(\lambda_k) = 1\} \geq M\} \]

is equal to \( \epsilon_M \) given by \( (2.162) \).

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

\[ \{\lambda^k : \text{Card}\{t \leq k : \psi_t^+(\lambda^t) = 1\} = m\}, \]

and let

\[ \pi_m^k = P\{\lambda^K : \chi_m^k(\lambda^k) = 1\}. \]

We have the following evident recurrence:

\[ \chi_m^k(\lambda^k) = \chi_m^{k-1}(\lambda^{k-1})(1 - \psi_k^+(\lambda^k)) + \chi_{m-1}^{k-1}(\lambda^{k-1})\psi_k^+(\lambda^k), k = 1, 2, ... \]

augmented by the “boundary conditions” \( \chi_m^0 = 0, m > 0, \chi_0^0 = 1, \chi_{m-1}^{k-1} = 0 \) for all \( k \geq 1 \). Taking expectation w.r.t. \( P \) and utilizing the fact that conditional, given
\( \lambda^{k-1} \), expectation of \( \psi_k^+(\lambda^k) \) is, identically in \( \lambda^{k-1} \), equal to \( \epsilon_* \), we get

\[
\pi^k_m = \pi^k_{m-1}(1 - \epsilon_*) + \pi^k_{m-1}\epsilon_*, \quad k = 1, \ldots, K,
\]
\[
\pi^0_m = \begin{cases} 1, & m = 0, \\ 0, & m > 0, \end{cases}
\]
whence

\[
\pi^k_m = \begin{cases} \frac{1}{m^k}\epsilon^m (1 - \epsilon_*)^{k-m}, & m \leq k, \\ 0, & m > k. \end{cases}
\]

Therefore,

\[
P\{\lambda^K : \text{Card}\{k \leq K : \psi_k^+(\lambda^k) = 1\} \geq M\} = \sum_{M \leq k \leq K} \pi^K_k = \epsilon_M,
\]
as required. \( \square \)

### 2.11.3 Proof of Theorem 2.23

1°. Since \( \mathcal{O} \) is a simple o.s., the function \( \Phi(\phi, [\mu; \nu]) \) given by (2.56) 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 \( \Phi \) on \( \mathcal{F} \times (\mathcal{M} \times \mathcal{M}) \) imply the continuity of \( \Phi \) on the indicated domain. As a consequence, \( \Phi \) is convex-concave continuous real-valued function on \( \mathcal{F} \times (\mathcal{M} \times \mathcal{M}) \).

Now let

\[
\Phi(\mu, \nu) = \inf_{\phi \in \mathcal{F}} \Phi(\phi, [\mu; \nu]).
\]

Note that \( \Phi \), 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 (2.166), so that

\[
\forall (\mu \in M_1, \nu \in M_2) : \Phi([\mu; \nu]) := \inf_{\phi \in \mathcal{F}} \Phi(\phi, [\mu; \nu]) = \Phi(\phi_{\mu, \nu}, [\mu; \nu]) = \ln \left( \int_{\Omega} \sqrt{p_\mu(\omega)p_\nu(\omega)} d\omega \right).
\]

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 \( \Phi(\phi_{\mu, \nu}, [\mu; \nu]) = \ln \left( \int_{\Omega} g(\omega) d\omega \right) \). On the other hand, for \( \phi(\cdot) = \phi_{\mu, \nu}(\cdot) + \delta(\cdot) \in \mathcal{F} \) we have

\[
\int_{\Omega} g(\omega) d\omega = \int_{\Omega} \left[ \sqrt{g(\omega)\exp(-\delta(\omega)/2)} \right] \left[ \sqrt{g(\omega)\exp(\delta(\omega)/2)} \right] d\omega
\]

\[
\leq \left( \int_{\Omega} g(\omega) \exp(-\delta(\omega)/2) d\omega \right)^{1/2} \left( \int_{\Omega} g(\omega) \exp(\delta(\omega)/2) d\omega \right)^{1/2}
\]

\[
= \left( \int_{\Omega} \exp(-\phi(\omega))p_\mu(\omega) d\omega \right)^{1/2} \left( \int_{\Omega} \exp(\phi(\omega))p_\nu(\omega) d\omega \right)^{1/2} \left[ \text{by (2.168)} \right]
\]

\[
\Rightarrow \ln \left( \int_{\Omega} g(\omega) d\omega \right) \leq \Phi(\phi, [\mu; \nu]),
\]
and thus \( \Phi(\phi_{\mu,\nu}, [\mu; \nu]) \leq \Phi(\phi, [\mu; \nu]) \) for every \( \phi \in \mathcal{F} \).

**Remark 2.49.** Note that the above reasoning did not use the fact that the minimization in the right hand side of \((2.166)\) is over \( \phi \in \mathcal{F} \); in fact, this reasoning shows that \( \phi_{\mu,\nu}(\cdot) \) minimizes \( \Phi(\phi, [\mu; \nu]) \) over all functions \( \phi \) 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 2.50.** Note that the inequality in \((b)\) can be equality only when the inequality in \((a)\) is so. In other words, if \( \tilde{\phi} \) is a minimizer of \( \Phi(\phi, [\mu; \nu]) \) over \( \phi \in \mathcal{F} \), setting \( \delta(\cdot) = \tilde{\phi}(\cdot) - \phi_{\mu,\nu}(\cdot) \), the functions \( \sqrt{g(\omega)} \exp\{-\delta(\omega)/2\} \) and \( g(\omega) \exp\{\delta(\omega)/2\} \), considered as elements of \( L_2[\Omega, \Pi] \), are proportional to each other. Since \( g \) is positive and \( g, \delta \) are continuous, while the support of \( \Pi \) is the entire \( \Omega \), 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 \( \phi(\omega) = \phi_{\mu,\nu}(\omega) + \text{const} \).

Let us verify that \( \Phi(\phi, [\mu; \nu]) \) has a saddle point (min in \( \phi \in \mathcal{F} \), max in \([\mu; \nu] \in M_1 \times M_2 \)). First, observe that on the domain of \( \Phi \) it holds

\[
\Phi(\phi(\cdot) + a, [\mu; \nu]) = \Phi(\phi(\cdot), [\mu; \nu]) \quad \forall (a \in \mathbb{R}, \phi \in \mathcal{F}). \tag{2.169}
\]

Let us select somehow \( \bar{\mu} \in \mathcal{M} \), and let \( \mathcal{P} \) be the measure on \( \Omega \) with density \( p_{\bar{\mu}} \) w.r.t. \( \Pi \). For \( \phi \in \mathcal{F} \), the integrals \( \int_\Omega e^{\pm \phi(\omega)} \mathcal{P}(d\omega) \) are finite (since \( \mathcal{O} \) is simple), implying that \( \phi \in L_1[\Omega, \mathcal{P}] \); note also that \( \mathcal{P} \) is a probabilistic measure. Let now \( \mathcal{F}_0 = \{ \phi \in \mathcal{F} : \int_\Omega \phi(\omega) \mathcal{P}(d\omega) = 0 \} \), 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. Now, by \((2.169)\), to prove the existence of a saddle point of \( \Phi \) on \( \mathcal{F} \times (M_1 \times M_2) \) is exactly the same as to prove the existence of a saddle point of \( \Phi \) on \( \mathcal{F}_0 \times (M_1 \times M_2) \). Let us verify that \( \Phi(\phi, [\mu; \nu]) \) indeed has a saddle point on \( \mathcal{F}_0 \times (M_1 \times M_2) \). Because \( M_1 \times M_2 \) is a convex compact set, and \( \Phi \) 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 prove the existence of a saddle point is to verify that \( \Phi \) is coercive in the first argument. In other words, we have to show that 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

\[
\Theta(\phi) = \Phi(\phi, [\mu; \nu]) = \frac{1}{2} \left[ \ln \left( \int_{\omega} e^{-\phi(\omega)} p_\mu(\omega) \Pi(d\omega) \right) + \ln \left( \int_{\omega} e^{\phi(\omega)} p_\nu(\omega) \Pi(d\omega) \right) \right]
\]

and taking into account that \( \Theta \) is convex and finite on \( \mathcal{F}_0 \), in order to prove that \( \Theta \) 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 \( \int_\Omega \phi(\omega) \mathcal{P}(d\omega) = 0 \) and \( \phi \) is nonzero, we have \( \int_\Omega \max[\phi(\omega), 0] \mathcal{P}(d\omega) = \int_\Omega \max[\phi(\omega), 0] \mathcal{P}(d\omega) > 0 \), whence \( \phi > 0 \) and \( \phi < 0 \) on sets of \( \Pi \)-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 continuous and everywhere positive.

3'. Now let \( (\phi_\ast(\cdot); [\mu_\ast; \nu_\ast]) \) be a saddle point of \( \Phi \) on \( \mathcal{F} \times (M_1 \times M_2) \). Shifting, if necessary, \( \phi_\ast(\cdot) \) by a constant (by \((2.169)\), this does not affect the fact that \( (\phi_\ast, [\mu_\ast; \nu_\ast]) \) is a saddle point of \( \Phi \)), we can assume that

\[
\varepsilon_\ast := \int_\Omega \exp(-\phi_\ast(\omega)) p_{\mu_\ast}(\omega) \Pi(d\omega) = \int_\Omega \exp(\phi_\ast(\omega)) p_{\nu_\ast}(\omega) \Pi(d\omega),
\]
so that the saddle point value of $\Phi$ is

$$\Phi_* := \max_{\mu, \nu} \min_{\phi, [\mu; \nu]} \Phi(\phi, [\mu; \nu]) = \Phi(\phi_*, [\mu_*; \nu_*]) = \ln(\varepsilon_*). \quad (2.170)$$

as claimed in item (i) of Theorem.

Now let us prove $(2.58)$. For $\mu \in M_1$, we have

$$\ln(\varepsilon_*) = \Phi_* \geq \Phi(\phi_*, [\mu_*; \nu_*])$$

$$= \frac{1}{2} \ln \left( \int_{\Omega} \exp(-\phi_*(\omega)) p_{\mu}(\omega) \Pi(d\omega) \right) + \frac{1}{2} \ln \left( \int_{\Omega} \exp(\phi_*(\omega)) p_{\nu}(\omega) \Pi(d\omega) \right)$$

$$= \frac{1}{2} \ln \ln(\varepsilon_*) + \frac{1}{2} \ln(\varepsilon_*)$$

Hence,

$$\ln \left( \int_{\Omega} \exp\{-\phi_*(\omega)\} p_{\mu}(\omega) \Pi(d\omega) \right) = \ln \left( \int_{\Omega} \exp\{-\phi_*(\omega)\} p_{\mu}(\omega) P(d\omega) \right) + a$$

$$\leq \ln(\varepsilon_*) + a,$$

and $(2.58.a)$ follows. Similarly, when $\nu \in M_2$, we have

$$\ln(\varepsilon_*) = \Phi_* \geq \Phi(\phi_*, [\mu_*; \nu_*])$$

$$= \frac{1}{2} \ln \left( \int_{\Omega} \exp(-\phi_*(\omega)) p_{\nu}(\omega) \Pi(d\omega) \right) + \frac{1}{2} \ln \left( \int_{\Omega} \exp(\phi_*(\omega)) p_{\nu}(\omega) \Pi(d\omega) \right)$$

$$= \frac{1}{2} \ln(\varepsilon_*) + \frac{1}{2} \ln(\varepsilon_*)$$

so that

$$\ln \left( \int_{\Omega} \exp\{\phi_*(\omega)\} p_{\nu}(\omega) \Pi(d\omega) \right) = \ln \left( \int_{\Omega} \exp\{\phi_*(\omega)\} p_{\nu}(\omega) P(d\omega) \right) - a$$

$$\leq \ln(\varepsilon_*) - a,$$

and $(2.58.b)$ follows.

We have proved all statements of item (i), except for the claim that the just defined $\phi_*, \varepsilon_*$ form an optimal solution to $(2.59)$. Note that by $(2.58)$ as applied with $a = 0$, the pair in question is feasible for $(2.59)$. Assuming that the problem admits a feasible solution $(\phi, \epsilon)$ with $\epsilon < \varepsilon_*$, let us lead this assumption to a contradiction. Note that $\tilde{\phi}$ should be such that

$$\int_{\Omega} \exp(-\phi_*(\omega)) p_{\mu}(\omega) \Pi(d\omega) < \varepsilon_* \& \int_{\Omega} \exp(\phi_*(\omega)) p_{\nu}(\omega) \Pi(d\omega) < \varepsilon_*,$$

and consequently $\Phi(\tilde{\phi}, [\mu_*; \nu_*]) < \ln(\varepsilon_*)$. On the other hand, Remark 2.49 says that $\Phi(\phi_*, [\mu_*; \nu_*])$ cannot be less than $\min_{\phi \in F} \Phi(\phi_* [\mu_*; \nu_*])$, and the latter quantity is $\Phi(\phi_*, [\mu_*; \nu_*])$ because $(\phi_*, [\mu_*; \nu_*])$ is a saddle point of $\Phi$ on $F \times (M_1 \times M_2)$. Thus, assuming that the optimal value in $(2.59)$ is $< \varepsilon_*$, we conclude that $\Phi(\phi_*, [\mu_*; \nu_*]) < \Phi(\phi_*, [\mu_*; \nu_*]) < \ln(\varepsilon_*)$, contradicting $(2.170)$. Item (i) of Theorem 2.23 is proved.

4°. Let us prove item (ii) of Theorem 2.23. Relation $(2.60)$ and concavity of the right hand side of this relation in $[\mu; \nu]$ were already proved; moreover, these relations were proved in the range $M \times M$ of $[\mu; \nu]$. Since this range coincides with its relative interior, the real-valued concave function $\Phi$ is continuous on $M \times M$ and thus is continuous on $M_1 \times M_2$. Next, let $\phi_*$ be the $\phi$-component of a saddle point of $\Phi$ on $F \times (M_1 \times M_2)$ (we already know that such saddle point exists). By Proposition 2.21, the $[\mu; \nu]$-components of saddle points of $\Phi$ on $F \times (M_1 \times M_2)$ are exactly the maximizers of $\Phi$ on $M_1 \times M_2$: let $[\mu_*; \nu_*]$ be such a maximizer. By the same proposition, $(\phi_*, [\mu_*; \nu_*])$ is a saddle point of $\Phi$, whence $\Phi(\phi_*, [\mu_*; \nu_*])$ attains its minimum over $\phi \in F$ 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 2.50 to imply that \( \phi_* \) and \( \phi_{\mu_*, \nu_*} \) differ by a constant, which, in view of (2.169), means that \( (\phi_{\mu_*, \nu_*}, [\mu_*; \nu_*]) \) is a saddle point of \( \Phi \) along with \( (\phi_*, [\mu_*; \nu_*]) \). (ii) is proved.

5°. It remains to prove item (iii) of Theorem 2.23. In the notation from (iii), simple hypotheses \((A)\) and \((B)\) can be decided with the total risk \( \leq 2\varepsilon \), and therefore, by Proposition 2.2,

\[
2\varepsilon := \int_{\Omega} \min[p(\omega)q(\omega)]\Pi(d\omega) \leq 2\varepsilon.
\]

On the other hand, we have seen that the saddle point value of \( \Phi \) is \( \ln(\varepsilon_*) \); since \([\mu_*; \nu_*] \) is a component of a saddle point of \( \Phi \), it follows that \( \min_{\phi \in \mathcal{F}} \Phi(\phi, [\mu_*; \nu_*]) = \ln(\varepsilon_*) \).

The left hand side in this equality, by item 1°, is \( \Phi(\phi_{\mu_*, \nu_*}, [\mu_*; \nu_*]) \), and we arrive at

\[
\ln(\varepsilon_*) = \Phi(\frac{1}{2} \ln(p_{\mu_*}(\cdot)/p_{\nu_*}(\cdot)), [\mu_*; \nu_*]) = \ln \left( \int_{\Omega} \sqrt{p_{\mu_*}(\omega)p_{\nu_*}(\omega)\Pi(d\omega)} \right),
\]

so that

\[
\varepsilon_* = \int_{\Omega} \sqrt{p_{\mu_*}(\omega)p_{\nu_*}(\omega)\Pi(d\omega)} = \int_{\Omega} \sqrt{p(\omega)q(\omega)\Pi(d\omega)}.
\]

We now have

\[
\varepsilon_* = \int_{\Omega} \sqrt{p(\omega)q(\omega)\Pi(d\omega)} = \int_{\Omega} \sqrt{\min[p(\omega), q(\omega)]\max[p(\omega), q(\omega)]\Pi(d\omega)} \leq \left( \int_{\Omega} \min[p(\omega), q(\omega)]\Pi(d\omega) \right)^{1/2} \left( \int_{\Omega} \max[p(\omega), q(\omega)]\Pi(d\omega) \right)^{1/2} = \sqrt{2} \sqrt{2 - 2\varepsilon} \leq 2 \sqrt{1 - \varepsilon} \varepsilon,
\]

where the concluding inequality is due to \( \varepsilon \leq \varepsilon < 1/2 \). (iii) is proved, and the proof of Theorem 2.23 is complete. \( \square \)

2.11.4 Proof of Proposition 2.37

All we need is to verify (2.107) 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 (2.107), let us fix \( P \in \mathcal{P}[X] \) and \( h \in \mathbb{R}^d \) and set

\[
\nu = h^T \epsilon[P],
\]

so that \( \nu \) is the expectation of \( h^T \omega \) with \( \omega \sim P \). Note that for \( \omega \sim P \) we have

\[
h^T \omega \in [-\phi_X(-h), \phi_X(h)] \text{ with } P\text{-probability } 1, \text{ whence } -\phi_X(-h) \leq \nu \leq \phi_X(h).
\]

In particular, when \( \phi_X(h) + \phi_X(-h) = 0 \), \( h^T \omega = \nu \) with \( P\)-probability 1, so that (2.107) definitely holds true. Now let

\[
\eta := \frac{1}{2} [\phi_X(h) + \phi_X(-h)] > 0,
\]

and let

\[
a = \frac{1}{2} [\phi_X(h) - \phi_X(-h)], \quad \beta = (\nu - a)/\eta.
\]

Denoting by \( P_h \) the distribution of \( h^T \omega \) induced by the distribution \( P \) of \( \omega \) and noting that this distribution is supported on \([-\phi_X(-h), \phi_X(h)] = [a - \eta, a + \eta] \) and
has expectation $\nu$, we get $\beta \in [-1, 1]$ and
\[
\gamma := \int \exp \{ h^T \omega \} P(d\omega) = \int_{a-\eta}^{a+\eta} [e^s - \lambda(s - \nu)] P_h(ds)
\]
for all $\lambda \in \mathbb{R}$. Hence,
\[
\ln(\gamma) \leq \inf_{\lambda} \ln \left( \max_{a-\eta \leq s \leq a+\eta} [e^s - \lambda(s - \nu)] \right)
\]
\[
= a + \inf_{\rho} \ln \left( \max_{-\eta \leq t \leq \eta} [e^t - \rho(t - [\nu - a])] \right) \quad [\text{substituting } \lambda = e^s \rho, s = a + t]
\]
\[
= a + \inf_{\rho} \ln \left( \max_{-\eta \leq t \leq \eta} [e^t - \rho(t - \eta \beta)] \right) \leq a + \ln \left( \max_{-\eta \leq t \leq \eta} [e^t - \bar{\rho}(t - \eta \beta)] \right)
\]
with $\bar{\rho} = (2\eta)^{-1}(e^\eta - e^{-\eta})$. The function $g(t) = e^t - \bar{\rho}(t - \eta \beta)$ 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
\[
\ln(\gamma) \leq a + \ln(\cosh(\eta) + \beta \sinh(\eta)). \quad (2.171)
\]
Thus, all we need to verify is that
\[
\forall (\eta > 0, \beta \in [-1, 1]) : \beta \eta + \frac{1}{2} \eta^2 - \ln(\cosh(\eta) + \beta \sinh(\eta)) \geq 0. \quad (2.172)
\]
Indeed, if (2.172) holds true (2.171) implies that
\[
\ln(\gamma) \leq a + \beta \eta + \frac{1}{2} \eta^2 = \nu + \frac{1}{2} \eta^2,
\]
which, recalling what $\gamma$, $\nu$ and $\eta$ are, is exactly what we want to prove.

Verification of (2.172) is as follows. The left hand side in (2.172) is convex in $\beta$ for $\beta > -\coth(\eta)/\sinh(\eta)$ containing, due to $\eta > 0$, the range of $\beta$ in (2.172). Furthermore, the minimum of the left hand side of (2.172) over $\beta > -\coth(\eta)$ is attained at $\beta = \frac{\sinh(\eta) - \eta \cosh(\eta)}{\eta \sinh(\eta)}$ 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 $\eta > 0$. \hfill \Box

2.11.5 Proof of Proposition 2.43
2.11.5.A Proof of Proposition 2.43.i
1°. Let \( b = [0; ..., 0; 1] \in \mathbb{R}^{n+1} \), so that \( B = \begin{bmatrix} A \\ b^T \end{bmatrix} \), and let \( A(u) = A[u; 1] \). For any \( u \in \mathbb{R}^n \), \( h \in \mathbb{R}^d \), \( \Theta \in \mathcal{S}^d_+ \) and \( H \in \mathcal{S}^d \) such that \( -I < \Theta^{1/2} H \Theta^{1/2} < I \) we have

\[
\Psi(h; H; u, \Theta) := \ln(\mathbb{E}_{z \sim \mathcal{N}(H(u, \Theta))}\{\exp\{h^T \zeta + \frac{1}{2} \zeta^T H \zeta\}\})
\]

\[
= \ln(\mathbb{E}_{z \sim \mathcal{N}(0,I)\} \{\exp\{h^T[A(u) + \Theta^{1/2}\zeta] + \frac{1}{2}A(u) + \Theta^{1/2}\zeta]\} \}
\]

\[
= -\frac{1}{2} \ln \det(I - \Theta^{1/2} H \Theta^{1/2}) + h^T A(u) + \frac{1}{2} A(u)^T H A(u)
\]

\[
+ \frac{1}{2} [H A(u) + h]^T \Theta^{1/2} [I - \Theta^{1/2} H \Theta^{1/2}]^{-1} \Theta^{1/2} [H A(u) + h]
\]

\[
= -\frac{1}{2} \ln \det(I - \Theta^{1/2} H \Theta^{1/2}) + \frac{1}{2} [u; 1]^T \begin{bmatrix} h & A \end{bmatrix} + A^T H A + B^T [H, h]^T \Theta^{1/2} [I - \Theta^{1/2} H \Theta^{1/2}]^{-1} \Theta^{1/2} [H, h] B [u; 1]
\]

\[
\text{(2.173)}
\]

due to

\[ h^T A(u) = [u; 1]^T h b^T A[u; 1] = [u; 1]^T A^T h b^T [u; 1] \]

and \( H A(u) + h = [H, h] B [u; 1] \).

Observe that when \((h, H) \in \mathcal{H}^0\), we have

\[ \Theta^{1/2} [I - \Theta^{1/2} H \Theta^{1/2}]^{-1} \Theta^{1/2} = [\Theta^{-1} - H]^{-1} \leq [\Theta^{-1} - H]^{-1}, \]

so that (2.173) implies that for all \( u \in \mathbb{R}^n \), \( \Theta \in \mathcal{V} \), and \( (h, H) \in \mathcal{H}^0 \),

\[
\Psi(h; H; u, \Theta) \leq -\frac{1}{2} \ln \det(I - \Theta^{1/2} H \Theta^{1/2}) + \frac{1}{2} [u; 1]^T \begin{bmatrix} h & A \end{bmatrix} + A^T H A + B^T [H, h]^T [\Theta^{-1} - H]^{-1} [H, h] B [u; 1]
\]

\[
\leq -\frac{1}{2} \ln \det(I - \Theta^{1/2} H \Theta^{1/2}) + \frac{1}{2} \text{Tr}(Q[H, h] Z(u)) + \frac{1}{2} \text{Tr}(Q[H, h]) + \Gamma_Z(h, \mathcal{H})
\]

\[
\Gamma_Z(h, \mathcal{H}) = \frac{1}{2} \phi_Z(Q[H, h])
\]

\[
\text{(2.174)}
\]

(we have taken into account that \( Z(u) \in \mathcal{Z} \) when \( u \in \mathcal{U} \) (premise of the proposition) and therefore \( \text{Tr}(Q[H, h] Z(u)) \leq \phi_Z(Q[H, h]) \)). Note that the above function \( Q[H, h] \) is nothing but

\[
Q[H, h] = B^T \left( \begin{bmatrix} H & h \\ h^T & h^T \end{bmatrix} + [H, h]^T [\Theta^{-1} - H]^{-1} [H, h] \right) B.
\]

\[
\text{(2.175)}
\]

2°. We need the following

**Lemma 2.51.** Let \( \Theta_* \) be a \( d \times d \) symmetric positive definite matrix, let \( \delta \in [0, 2] \), and let \( \mathcal{V} \) be a closed convex subset of \( \mathcal{S}_d^+ \) such that

\[ \Theta \in \mathcal{V} \Rightarrow \{\Theta \leq \Theta_*\} & \{||\Theta^{1/2} \Theta_*^{-1/2} - I_d|| \leq \delta\} \]

\[
(\text{cf. (2.129)}) \]

Let also \( \mathcal{H}^0 := \{H \in \mathcal{S}^d : -\Theta_*^{-1} < H < \Theta_*^{-1}\} \). Then

\[
\forall (H, \Theta) \in \mathcal{H}^0 \times \mathcal{V}:
\]

\[
G(H; \Theta) := -\frac{1}{2} \ln \det(I - \Theta^{1/2} H \Theta^{1/2})
\]

\[
\leq G^+(H; \Theta) := -\frac{1}{2} \ln \det(I - \Theta_*^{1/2} H \Theta_*^{1/2}) + \frac{1}{2} \text{Tr}((\Theta - \Theta_*) H)
\]

\[
+ \frac{\delta(2+\delta)}{2(1-\||\Theta_*^{1/2} H \Theta_*^{1/2}\|)^2} \|\Theta^{1/2} H \Theta^{1/2}\|_F^2,
\]

\[
\text{(2.177)}
\]
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
\[
\|\Theta^{1/2} H \Theta^{1/2}\| = \|\Theta^{1/2} \Theta^{-1/2} H \Theta^{-1/2} \Theta^{1/2}\| \leq \|\Theta^{1/2} \Theta^{-1/2}\|^2 \|\Theta^{1/2} H \Theta^{1/2}\| \leq \|\Theta^{1/2} H \Theta^{1/2}\| = d(H)
\]
(we have used the fact that \( 0 \leq \Theta \leq \Theta_* \) implies \( \|\Theta^{1/2} \Theta^{-1/2}\| \leq 1 \)). Noting that \( \|AB\|_F \leq \|A\|_F \|B\|_F \), computation completely similar to the one in (2.178) yields
\[
\|\Theta^{1/2} H \Theta^{1/2}\|_F \leq \|\Theta^{1/2} H \Theta^{1/2}\|_F =: D(H)
\]
(2.179)

Besides this, setting \( F(X) = -\ln \det(X) : \text{int} \mathbb{S}^d_+ \rightarrow \mathbb{R} \) and equipping \( \mathbb{S}^d \) 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 \( R_\lambda = \lambda R_0 + (1 - \lambda) R_1 \):
\[
F(I - R_1) = F(I - R_0 - \Delta) = F(I - R_0) + \langle \nabla F(I - R_\lambda), -\Delta \rangle = F(I - R_0) + \langle (I - R_\lambda)^{-1}, \Delta \rangle = F(I - R_0) + \langle I, \Delta \rangle + \langle (I - R_\lambda)^{-1} - I, \Delta \rangle.
\]
We conclude that
\[
F(I - R_1) \leq F(I - R_0) + \text{Tr}(\Delta) + \|I - (I - R_\lambda)^{-1}\|_F \|\Delta\|_F. \tag{2.180}
\]

Denoting by \( \mu_i \) the eigenvalues of \( R_\lambda \) and noting that \( \|R_\lambda\| \leq \max\|R_0\|, \|R_1\| = d(H) \) (see (2.178)), 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 (2.179), we conclude that \( \|I - (I - R_\lambda)^{-1}\|_F \leq D(H)/(1 - d(H)) \), so that (2.180) yields
\[
F(I - R_1) \leq F(I - R_0) + \text{Tr}(\Delta) + D(H) \|\Delta\|_F/(1 - d(H)). \tag{2.181}
\]

Further, by (2.129) the matrix \( D = \Theta^{1/2} \Theta_*^{-1/2} - I \) satisfies \( \|D\| \leq \delta \), whence
\[
\Delta = \Theta^{1/2} H \Theta^{1/2} \cdot \Theta_*^{-1/2} H \Theta_*^{1/2} = (I + D)R_0(I + D^T)-R_0 = DR_0+R_0D^T+DR_0D^T.
\]

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 (2.181) and the relation
\[
\text{Tr}(\Delta) = \text{Tr}(\Theta^{1/2}H\Theta^{1/2} - \Theta_s^{1/2}H\Theta_s^{1/2}) = \text{Tr}((\Theta - \Theta_s)H)
\]
to yield
\[
F(I - R_1) \leq F(I - R_0) + \text{Tr}((\Theta - \Theta_s)H) + \frac{\delta(2+\delta)}{1-\parallel H \parallel^{1+\delta}} D(H)
\]
\[
= F(I - R_0) + \text{Tr}((\Theta - \Theta_s)H) + \frac{\delta(2+\delta)}{1-\parallel H \parallel^{1+\delta}} \parallel \Theta_s^{1/2}H\Theta_s^{1/2} \parallel^2_F,
\]
and we arrive at (2.177). It remains to prove that \( G^+(H; \Theta) \) is convex-concave and continuous on \( H^o \times V \). The only component of this claim which is not completely evident is convexity of the function in \( H \in H^o \). To see that it is the case, note that \( \ln \det(S) \) is concave on the interior of the semidefinite cone, the function \( f(u, v) = \frac{u^2}{v^2} \) is convex and nondecreasing in \( u, v \) in the convex domain \( \Pi = \{(u, v) : u \geq 0, v < 1\} \), and the function \( \frac{1}{1-\parallel H \parallel^{1+\delta}} \) is obtained from \( f \) by convex substitution of variables \( H \mapsto (\parallel \Theta_s^{1/2}H\Theta_s^{1/2} \parallel_F, \parallel \Theta_s^{1/2}H\Theta_s^{1/2} \parallel) \) mapping \( H^o \) into \( \Pi \).

3°. Combining (2.177), (2.174), (2.130) and the origin of \( \Psi \), see (2.173), we arrive at
\[
\forall ((u, \Theta) \in U \times V, (h, H) \in H^\gamma = H) : \\
\ln \left( \mathbb{E}_{\zeta \sim \mathcal{N}(A[u; 1], i)} \{ \exp \{ h^T \zeta + \frac{1}{2} \zeta^T H \zeta \} \} \right) \leq \Phi_{A, Z}(h, H; \Theta),
\]
as claimed in (2.133).

4°. Now let us check that \( \Phi_{A, Z}(h, H; \Theta) : H \times V \rightarrow R \) is continuous and convex-concave. Recalling that the function \( G^+(H; \Theta) \) from (2.177) is convex-concave and continuous on \( H^o \times V \), all we need to verify is that \( \Gamma_Z(h, H) \) is convex and continuous on \( H \). Recalling that \( Z \) is nonempty compact set, the function \( \phi_Z(\cdot) : S^{d+1} \rightarrow R \) is continuous, implying the continuity of \( \Gamma_Z(h, H) = \frac{1}{2} \phi_Z(Q[H, h]) \) on \( H = H^\gamma \) \((Q[H, h] \text{ is defined in (2.175))}. To prove convexity of \( \Gamma_Z \), note that \( Z \) is contained in \( S_n^{d+1} \), implying that \( \phi_Z(\cdot) \) is convex and \( \geq \)-monotone. On the other hand, by Schur Complement Lemma, we have
\[
S := \{(h, H, G) : G \succeq Q[H, h], (h, H) \in H^\gamma \}
= \left\{(h, H, G) : \begin{bmatrix} G - \left[ hh^T A + A^T hb^T + A^T HA \right] \left[ H, h \right] B \left[ H, h \right]^T \Theta_s^{-1} - H \end{bmatrix} \geq 0, \right. \\
\left. (h, H) \in H^\gamma \right\},
\]
implying that \( S \) is convex. Since \( \phi_Z(\cdot) \) is \( \geq \)-monotone, we have
\[
\{(h, H, \tau) : (h, H) \in H^\gamma, \tau \geq \Gamma_Z(h, H) \}
= \{(h, H, \tau) : \exists G : G \succeq Q[H, h], 2\tau \geq \phi_Z(G), (h, H) \in H^\gamma \},
\]
and we see that the epigraph of \( \Gamma_Z \) is convex (since the set \( S \) and the epigraph of \( \phi_Z \) are so), as claimed.

5°. It remains to prove that \( \Phi_{A, Z} \) is coercive in \( H, h \). Let \( \Theta \in V \) and \( (h_i, H_i) \in H^\gamma \) with \( \parallel (h_i, H_i) \parallel \rightarrow \infty \) as \( i \rightarrow \infty \), and let us prove that \( \Phi_{A, Z}(h_i, H_i; \Theta) \rightarrow \infty \).
Looking at the expression for $\Phi_{A_i}(h_i, H_i; \Theta)$, it is immediately seen that all terms in this expression, except for the terms coming from $\phi_Z(\cdot)$, remain bounded as $i$ grows, so that all we need to verify is that the $\phi_Z(\cdot)$-term goes to $\infty$ as $i \to \infty$. Observe that $H_i$ are uniformly bounded due to $(h_i, H_i) \in \mathcal{H}^7$, implying that $\|h_i\|_2 \to \infty$ as $i \to \infty$. Denoting $e = [0; \ldots; 0; 1] \in \mathbb{R}^{d+1}$ and, as before, $b = [0; \ldots; 0; 1] \in \mathbb{R}^{n+1}$, note that, by construction, $B^T e = b$. Now let $W \in \mathcal{Z}$, so that $W_{a+1,n+1} = 1$. Taking into account that the matrices $[\Theta^{-1}_s - H_i]^{-1}$ satisfy $\alpha I_d \leq [\Theta^{-1}_s - H_i]^{-1} \leq \beta I_d$ for some positive $\alpha, \beta$ due to $H_i \in \mathcal{H}^7$, observe that

$$
\left[ \begin{array}{c} \frac{H_i}{h_i} \\ h_i \end{array} \right]^{[\Theta^{-1}_s - H_i]^{-1}} = \left[ \begin{array}{c} h_i^T [\Theta^{-1}_s - H_i]^{-1} h_i \\ \alpha \|h_i\|_2^2 \end{array} \right] ee^T + R_i,
$$

where $\alpha_i \geq \alpha > 0$ and $\|R_i\|_F \leq C(1 + \|h_i\|_2)$. As a result,

$$
\phi_Z(B^T Q_i B) \geq \text{Tr}(WB^T Q_i B) = \text{Tr}(WB^T[\alpha_i \|h_i\|_2^2 ee^T + R_i]B) = \alpha_i \|h_i\|_2^2 \text{Tr}(Wbb^T) - \|BW B^T\|_F \|R_i\|_F = W_{a+1,n+1} = 1,
$$

$$
\geq \alpha \|h_i\|_2^2 - C(1 + \|h_i\|_2) \|BW B^T\|_F,
$$

and the concluding quantity tends to $\infty$ as $i \to \infty$ due to $\|h_i\|_2 \to \infty$, $i \to \infty$. Part (i) is proved.

### 2.11.5.B Proof of Proposition 2.43.ii

By (i) the function $\Phi(h, H; \Theta_1, \Theta_2)$, as defined in (2.134), is continuous and convex-concave on the domain $(\mathcal{H}_1 \cap \mathcal{H}_2) \times (\mathcal{V}_1 \times \mathcal{V}_2)$ and is coercive in $(h, H)$, $\mathcal{H}$ and $\mathcal{V}$ are closed and convex, and $\mathcal{V}$ in addition is compact, so that saddle point problem (2.134) is solvable (Sion-Kakutani Theorem, a.k.a. Theorem 2.22). Now let $(h_s, H_s; \Theta_1^*, \Theta_2^*)$ be a saddle point. To prove (2.136), let $P \in \mathcal{G}_1$, that is, $P = N(A_1[u; 1], \Theta_1)$ for some $\Theta_1 \in \mathcal{V}_1$ and some $u$ with $[u; 1][u; 1]^T \in \mathbb{Z}_1$. Applying (2.133) to the first collection of data, with $a$ given by (2.135), we get the first $\leq$ in the following chain:

$$
\ln \left( \int e^{-\frac{1}{2} \omega^T H_s \omega - \alpha a} P(d\omega) \right) \leq \Phi_{A_s, z_1}(-h_s, -H_s; \Theta_1) - a
$$

(a)

$$
\leq \Phi_{A_s, z_1}(-h_s, -H_s; \Theta_1^*) - a = SV,
$$

(b)

where (a) is due to the fact that $\Phi_{A_1, z_1}(-h_s, -H_s; \Theta_1) + \Phi_{A_2, z_2}(h_s, H_s; \Theta_2)$ attains its maximum over $(\Theta_1, \Theta_2) \in \mathcal{V}_1 \times \mathcal{V}_2$ at the point $(\Theta_1^*, \Theta_2^*)$, and (b) is due to the origin of $a$ and the relation $SV = \frac{1}{2} [\Phi_{A_1, z_1}(-h_s, -H_s; \Theta_1^*) + \Phi_{A_2, z_2}(h_s, H_s; \Theta_2^*)]$. The bound in (2.136a) is proved. Similarly, let $P \in \mathcal{G}_2$, that is, $P = N(A_2[u; 1], \Theta_2)$ for some $\Theta_2 \in \mathcal{V}_2$ and some $u$ with $[u; 1][u; 1]^T \in \mathbb{Z}_2$. Applying (2.133) to the second collection of data, with the same $a$ as above, we get the first $\leq$ in the following
chain:
\[
\ln \left( \int e^{\frac{1}{2}\omega^T H \omega + \omega^T h + \alpha P(d\omega)} \right) \leq \Phi_{A_2, z_2}(h, H; \Theta) + a
\]
\[
\leq \Phi_{A_2, z_2}(h, H; \Theta^*) + a = SV,
\]
with exactly the same as above justification of (a) and (b). The bound in (2.136.b) is proved.

2.11.6 Proof of Proposition 2.46

2.11.6.A Preliminaries

We start with the following result:

Lemma 2.52. Let \( \Theta \) be a positive definite \( d \times d \) matrix, \( B = \begin{bmatrix} A & \cdots & 0 \end{bmatrix} \), and let
\[
u \mapsto C(\nu) = A[\nu; 1]
\]
be an affine mapping from \( \mathbb{R}^d \) into \( \mathbb{R}^d \). Finally, let \( h \in \mathbb{R}^d \), \( H \in \mathbb{S}^d \) and \( P \in \mathbb{S}^d \) satisfy the relations
\[
0 \preceq P \prec I_d \& P \succeq \Theta^{1/2} H \Theta^{1/2}.
\]
Then, \( \zeta \sim SG(C(\nu), \Theta) \) and for every \( \nu \in \mathbb{R}^n \) it holds
\[
\ln \left( E_\zeta \left\{ e^{h^T \zeta + \frac{1}{2} \zeta^T H \zeta} \right\} \right) \leq -\frac{1}{2} \ln \text{Det}(I - P)
\]
\[
+ \frac{1}{2} \nu^T B^T \left[ \begin{bmatrix} H & h^T \end{bmatrix} + [H, h]^T \Theta^{1/2} [I - P]^{-1} \Theta^{1/2} [H, h] \right] B \nu + 1
\]
Equivalently (set \( G = \Theta^{-1/2} P \Theta^{-1/2} \)): whenever \( h \in \mathbb{R}^d \), \( H \in \mathbb{S}^d \) and \( G \in \mathbb{S}^d \) satisfy the relations
\[
0 \preceq G \prec \Theta^{-1} \& G \succeq H,
\]
one has for \( \zeta \sim SG(C(\nu), \Theta) \) and every \( \nu \in \mathbb{R}^n \):
\[
\ln \left( E_\zeta \left\{ e^{h^T \zeta + \frac{1}{2} \zeta^T H \zeta} \right\} \right) \leq -\frac{1}{2} \ln \text{Det}(I - \Theta^{1/2} G \Theta^{1/2})
\]
\[
+ \frac{1}{2} \nu^T B^T \left[ \begin{bmatrix} H & h^T \end{bmatrix} + [H, h]^T \Theta^{-1} - G^{-1} [H, h] \right] B \nu + 1
\]

Proof. 1°. Let us start with the following observation:

Lemma 2.53. Let \( \Theta \in \mathbb{S}_+^d \) and \( S \in \mathbb{R}^{d \times d} \) be such that \( S \Theta S^T \prec I_d \). Then for every \( \nu \in \mathbb{R}^d \) one has
\[
\ln \left( E_{\zeta \sim SG(0, \Theta)} \left\{ e^{\nu^T S \zeta + \frac{1}{2} \zeta^T S^T S \zeta} \right\} \right) \leq \ln \left( E_{\eta \sim N(\nu, I_d)} \left\{ e^{\frac{1}{2} \nu^T S \Theta S^T \nu} \right\} \right)
\]
\[
= -\frac{1}{2} \ln \text{Det}(I_d - S \Theta S^T) + \frac{1}{2} \nu^T \left[ S \Theta S^T (I_d - S \Theta S^T)^{-1} \right] \nu.
\]

Indeed, let \( \zeta \sim SG(0, \Theta) \) and \( \eta \sim N(\nu, I_d) \) be independent. We have
\[
E_\zeta \left\{ e^{\nu^T S \zeta + \frac{1}{2} \zeta^T S^T S \zeta} \right\} = E_\zeta \left\{ E_\eta \left\{ e^{\nu^T \zeta} \right\} \right\} = E_\eta \left\{ E_\zeta \left\{ e^{\nu^T \eta S^T \zeta} \right\} \right\}
\]
\[
\leq E_\eta \left\{ e^{\frac{1}{2} \nu^T S \Theta S^T \nu} \right\}.
\]
where \(a\) is due to \(\eta \sim \mathcal{N}(\nu, I_d)\) and \(b\) is due to \(\xi \sim \mathcal{SG}(0, \Theta)\). We have verified the inequality in (2.186); the equality in (2.186) is given by direct computation. \(\square\)

2*. Now, in the situation described in Lemma 2.52, by continuity it suffices to prove (2.183) in the case when \(P \geq 0\) in (2.182) is replaced with \(P > 0\). Under the premise of the lemma, given \(u \in \mathbb{R}^n\) and assuming \(P > 0\), let us set \(\mu = \mathcal{C}(u) = A[u; 1]\), \(\nu = P^{1/2} \tilde{\Theta}^{1/2}[H \mu + h]\), \(S = P^{1/2} \tilde{\Theta}^{-1/2}\), so that \(S \Theta S^T = P < I_d\), and let \(G = \Theta^{-1/2} P \Theta^{-1/2}\), so that \(G \succeq H\). Let \(\zeta \sim \mathcal{SG}(\mu, \Theta)\). Representing \(\zeta = \mu + \xi\) with \(\xi \sim \mathcal{SG}(0, \Theta)\), we have

\[
\ln \left( E_\xi \left\{ e^{h^T \zeta + \frac{1}{2} \xi^T H \xi} \right\} \right) = h^T \mu + \frac{1}{2} h^T H \mu + \ln \left( E_\xi \left\{ e^{(h^T + h \mu)^T \xi + \frac{1}{2} \xi^T G \xi} \right\} \right)
\]

\[
\leq h^T \mu + \frac{1}{2} h^T H \mu + \ln \left( E_\xi \left\{ e^{(h^T + h \mu)^T \xi + \frac{1}{2} \xi^T G \xi} \right\} \right)
\]

\[
\leq h^T \mu + \frac{1}{2} h^T H \mu + \ln \left( E_\xi \left\{ e^{v^T \zeta + \frac{1}{2} \xi^T S \xi} \right\} \right)
\]

\[
\leq h^T \mu + \frac{1}{2} h^T H \mu - \frac{1}{2} \ln \det(\Theta - \Theta_{\xi, 1/2}) + \frac{1}{2} \nu^T \left[ \Theta^{1/2} (I_d - \Theta S T^{-1}) \right] \nu
\]

\[
= h^T \mu + \frac{1}{2} h^T H \mu - \frac{1}{2} \ln \det(I_d - P) + \frac{1}{2} (H \mu + h)^T \Theta^{1/2} (I_d - P)^{-1} \Theta^{1/2} (H \mu + h)
\]

\[
\text{[plugging in} \ S \text{and} \ \nu]\]

It is immediately seen that the concluding quantity in this chain is nothing but the right hand side quantity in (2.183). \(\square\)

2.11.6.B Completing the proof of Proposition 2.46.

1°. Let us prove (2.142.a). By Lemma 2.52 (see (2.185)) applied with \(\Theta = \Theta_{\ast}\), setting \(\mathcal{C}(u) = A[u; 1]\), we have

\[
\forall ((h, H) \in \mathcal{H}, G : 0 \preceq G \preceq \gamma^+ \Theta_{\ast, 1}^{-1}, G \succeq H, u \in \mathbb{R}^n : [u; 1][u; 1]^T \in \mathcal{Z}) :
\]

\[
\ln \left( E_{\zeta \sim \mathcal{SG}(\mathcal{C}(u), \Theta_{\ast})} \left\{ e^{h^T \zeta + \frac{1}{2} \xi^T H \xi} \right\} \right) \leq -\frac{1}{2} \ln \det(I - \Theta_{\ast, 1/2} G \Theta_{\ast, 1/2})
\]

\[
+ \frac{1}{2} [u; 1]^T B^T \left[ \begin{bmatrix} H & h \end{bmatrix} \right] + [H, h]^T [\Theta_{\ast, 1}^{-1} - G^{-1} [H, h]] B [u; 1]
\]

\[
\leq -\frac{1}{2} \ln \det(I - \Theta_{\ast, 1/2} G \Theta_{\ast, 1/2})
\]

\[
+ \frac{1}{2} \Phi_{\mathcal{A}, \mathcal{Z}} \left[ \begin{bmatrix} H & h \end{bmatrix} \right] + [H, h]^T [\Theta_{\ast, 1}^{-1} - G^{-1} [H, h]] B \right) = \Psi_{\mathcal{A}, \mathcal{Z}}(h, H, G).
\]

(2.187)

implying, due to the origin of \(\Phi_{\mathcal{A}, \mathcal{Z}}\), that under the premise of (2.187) we have

\[
\ln \left( E_{\zeta \sim \mathcal{SG}(\mathcal{C}(u), \Theta_{\ast})} \left\{ e^{h^T \zeta + \frac{1}{2} \xi^T H \xi} \right\} \right) \leq \Phi_{\mathcal{A}, \mathcal{Z}}(h, H), \forall (h, H) \in \mathcal{H}.
\]

Taking into account that when \(\zeta \sim \mathcal{SG}(\mathcal{C}(u), \Theta)\) with \(\Theta \in \mathcal{V}\), we have also \(\zeta \sim \mathcal{SG}(\mathcal{C}(u), \Theta_{\ast})\), (2.142.a) follows.

2°. Now let us prove (2.142.b). All we need is to verify the relation

\[
\forall ((h, H) \in \mathcal{H}, G : 0 \preceq G \preceq \gamma^+ \Theta_{\ast, 1}^{-1}, G \succeq H, u \in \mathbb{R}^n : [u; 1][u; 1]^T \in \mathcal{Z}, \Theta \in \mathcal{V}) :
\]

\[
\ln \left( E_{\zeta \sim \mathcal{SG}(\mathcal{C}(u), \Theta)} \left\{ e^{h^T \zeta + \frac{1}{2} \xi^T H \xi} \right\} \right) \leq \Phi_{\mathcal{A}, \mathcal{Z}}(h, H, G; \Theta);
\]

(2.188)

with this relation at our disposal (2.142.b) can be obtained by the same argument as the one we used in item 1° to derive (2.142.a).
To establish (2.188), let us fix $h, H, G, u, \Theta$ satisfying the premise of (2.188); recall that under the premise of Proposition 2.46.i, we have $0 \leq \Theta \leq \Theta_{\star}$. Now let $\lambda \in (0, 1)$, and let $\Theta_{\lambda} = \Theta + \lambda(\Theta_{\star} - \Theta)$, so that $0 < \Theta_{\lambda} \leq \Theta_{\star}$, and let $\delta_{\lambda} = \|\Theta_{\lambda}^{1/2}G\Theta_{\lambda}^{-1/2} - I\|$, implying that $\delta_{\lambda} \in [0, 2]$. We have $0 \leq G \leq \lambda + \Theta_{\star}^{-1} \leq \lambda + \Theta_{\lambda}^{-1}$, that is, $H, G$ satisfy (2.184) w.r.t. $\Theta = \Theta_{\lambda}$. As a result, for our $h, G, H, u$, the just defined $\Theta$ and $\zeta \in S_{\ast}G(C(u), \Theta_{\lambda})$ relation (2.185) holds true:

$$
\ln \left( E_{\zeta} \left\{ e^{h^T \zeta + \frac{1}{2} \zeta^T H \zeta} \right\} \right) \leq - \frac{1}{2} \ln \det(I - \Theta_{\lambda}^{1/2}G\Theta_{\lambda}^{-1/2}) + \frac{1}{2} [u; 1]^T B^T \left[ \frac{H}{h^T} - \frac{h}{h^T} \right] [H, h]B[u; 1] + \frac{1}{2} \phi_{\zeta} \left( B^T \left[ \frac{H}{h} \right] + [H, h]^T \Theta_{\lambda}^{-1} - [H, h]B \right)
$$

(2.189)

(recall that $[u; 1]^T \in \mathcal{Z}$). As a result,

$$
\ln \left( E_{\zeta} \sim S_{G}(C(u), \Theta) \left\{ e^{h^T \zeta + \frac{1}{2} \zeta^T H \zeta} \right\} \right) \leq - \frac{1}{2} \ln \det(I - \Theta_{\lambda}^{1/2}G\Theta_{\lambda}^{-1/2}) + \frac{1}{2} \phi_{\zeta} \left( B^T \left[ \frac{H}{h^T} - \frac{h}{h^T} \right] [H, h]B \right)
$$

(2.190)

When deriving (2.190) from (2.189), we have used that

- $\Theta \leq \Theta_{\lambda}$, so that when $\zeta \sim S_{G}(C(u), \Theta)$, we have also $\zeta \sim S_{G}(C(u), \Theta_{\lambda})$,
- $0 \leq \Theta_{\lambda} \leq \Theta$, and $G < \Theta^{-1}$, whence $[\Theta_{\lambda}^{-1} - G]^{-1} \preceq [\Theta_{\star}^{-1} - G]^{-1}$,
- $\mathcal{Z} \subset S_{\ast}^{+}$, whence $\phi_{\zeta}$ is $\succ$-monotone: $\phi_{\zeta}(M) \leq \phi_{\zeta}(N)$ whenever $M \preceq N$.

By Lemma 2.51 applied with $\Theta_{\lambda}$ in the role of $\Theta$ and $\delta_{\lambda}$ in the role of $\delta$, we have

$$
- \frac{1}{2} \ln \det(I - \Theta_{\lambda}^{1/2}G\Theta_{\lambda}^{-1/2}) \leq - \frac{1}{2} \ln \det(I - \Theta_{\lambda}^{1/2}G\Theta_{\lambda}^{1/2}) + \frac{1}{2} \text{Tr}([\Theta_{\lambda} - \Theta_{\star}]G) + \frac{\delta_{\lambda}}{2(1 - \|G\Theta_{\lambda}^{1/2}\|^2)} \|G\Theta_{\lambda}^{1/2}\|^2.
$$

Consequently, (2.190) implies that

$$
\ln \left( E_{\zeta} \sim S_{G}(C(u), \Theta) \left\{ e^{h^T \zeta + \frac{1}{2} \zeta^T H \zeta} \right\} \right) \leq - \frac{1}{2} \ln \det(I - \Theta_{\lambda}^{1/2}G\Theta_{\lambda}^{1/2}) + \frac{1}{2} \text{Tr}([\Theta_{\lambda} - \Theta_{\star}]G) + \frac{\delta_{\lambda}}{2(1 - \|G\Theta_{\lambda}^{1/2}\|^2)} \|G\Theta_{\lambda}^{1/2}\|^2 + \frac{1}{2} \phi_{\zeta} \left( B^T \left[ \frac{H}{h^T} - \frac{h}{h^T} \right] [H, h]B \right).
$$

The resulting inequality holds true for all small positive $\lambda$; taking lim inf of the right hand side as $\lambda \to +0$, and recalling that $\Theta_{0} = \Theta$, we get

$$
\ln \left( E_{\zeta} \sim S_{G}(C(u), \Theta) \left\{ e^{h^T \zeta + \frac{1}{2} \zeta^T H \zeta} \right\} \right) \leq - \frac{1}{2} \ln \det(I - \Theta_{\star}^{1/2}G\Theta_{\star}^{-1/2}) + \frac{1}{2} \text{Tr}([\Theta_{\star} - \Theta_{\star}]G) + \frac{\delta_{\lambda}}{2(1 - \|G\Theta_{\star}^{1/2}\|^2)} \|G\Theta_{\star}^{1/2}\|^2 + \frac{1}{2} \phi_{\zeta} \left( B^T \left[ \frac{H}{h^T} - \frac{h}{h^T} \right] [H, h]B \right)
$$

(note that under the premise of Proposition 2.46.i we clearly have lim inf$_{\lambda \to +0} \delta_{\lambda} \leq \delta$). The right hand side of the resulting inequality is nothing but $\Psi_{A, \mathcal{Z}}^{\delta}(h, H, G; \Theta)$, see (2.141), and we arrive at the inequality required in the conclusion of (2.188).

3°. To complete the proof of Proposition 2.46.i, it remains to show that functions $\Phi_{A, \mathcal{Z}}$, $\Phi_{A, \mathcal{Z}}^{\delta}$, as announced in the proposition, possess continuity, convexity-concavity, and coerciveness properties. Let us verify that this indeed is so for $\Phi_{A, \mathcal{Z}}^{\delta}$;
the reasoning which follows, with obvious simplifications, is applicable to $\Phi_{A,Z}$ as well.

Observe, first, that for exactly the same reasons as in item 4o of the proof of Proposition 2.43, the function $\Psi_{A,Z}(h,H,G;\Theta)$ is real valued, continuous and convex-concave on the domain

$$\tilde{H} \times \mathcal{V} = \{(h,H,G) : -\gamma^+\Theta^{-1}_z \preceq H \preceq \gamma^+\Theta^{-1}_z, 0 \preceq G \preceq \gamma^+\Theta^{-1}_z, H \preceq G\} \times \mathcal{V}.$$

The function $\Phi_{A,Z}(h,H;\Theta) : \mathcal{H} \times \mathcal{V} \rightarrow \mathbb{R}$ is obtained from $\Psi_{A,Z}(h,H,G;\Theta)$ by the following two operations: we first minimize $\Psi_{A,Z}(h,H,G;\Theta)$ over $G$ linked to $(h,H)$ by the convex constraints $0 \preceq G \preceq \gamma^+\Theta^{-1}_z$ and $G \preceq H$, thus obtaining a function

$$\bar{\Phi}(h,H;\Theta) : \{(h,H) : -\gamma^+\Theta^{-1}_z \preceq H \preceq \gamma^+\Theta^{-1}_z\} \times \mathcal{V} \rightarrow \mathbb{R} \cup \{+\infty\} \cup \{-\infty\}.$$

Second, we restrict the function $\bar{\Phi}(h,H;\Theta)$ from $\tilde{H} \times \mathcal{V}$ onto $\mathcal{H} \times \mathcal{V}$. For $(h,H) \in \tilde{H}$, 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 $\tilde{H} \times \mathcal{V}$. From continuity of $\Psi_{A,Z}$ on its domain it immediately follows that $\Psi_{A,Z}$ is bounded and uniformly continuous on every bounded subset of this domain. This implies 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 $\tilde{H}$, and is continuous on $\bar{B} \times \mathcal{V}$ in $\Theta \in \mathcal{V}$ with properly selected modulus of continuity independent of $(h,H) \in \bar{B}$. Furthermore, by construction, $\mathcal{H} \subset \text{int} \tilde{H}$, 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 $\tilde{H}$. 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 $(h,H) \in \bar{B}$ 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 $\Phi_{A,Z}$ is not just real-valued convex-concave function on the domain $\mathcal{H} \times \mathcal{V}$, but is also continuous on this domain.

Coerciveness of $\Phi_{A,Z}(h,H;\Theta)$ in $(h,H)$ is proved in exactly the same way as the similar property of function (2.130), see item 5o in the proof of Proposition 2.43. The proof of item (i) of Proposition 2.46 is complete.

4o. Item (ii) of Proposition 2.46 can be derived from item (i) of the proposition following the steps of the proof of (ii) of Proposition 2.43. □
Chapter Three

From Hypothesis Testing to Estimating Functionals

In this chapter we extend the techniques developed in Chapter 2 beyond the hypothesis testing problem and apply them to estimating properly structured scalar functionals of the unknown signal, specifically:

- In simple observation schemes — linear (and more generally, $N$-convex, see Section 3.2) functionals on unions of convex sets (Sections 3.1 and 3.2);
- Beyond simple observation schemes — linear and quadratic functionals on convex sets (Sections 3.3 and 3.4).

3.1 ESTIMATING LINEAR FORMS ON UNIONS OF CONVEX SETS

The key to the subsequent developments in this section and in Sections 3.3, 3.4 is the following simple observation. Let $P = \{P_x : x \in \mathcal{X}\}$ be a parametric family of distributions on $\mathbb{R}^d$, $\mathcal{X}$ being a convex subset of some $\mathbb{R}^m$. Suppose that given a linear form $g^T x$ on $\mathbb{R}^m$ and an observation $\omega \sim P_x$ stemming from unknown signal $x \in \mathcal{X}$, we want to recover $g^T x$, and intend to use for this purpose an affine function $h^T \omega + \kappa$ of the observation. How to ensure that the recovery, with a given probability $1 - \epsilon$, deviates from $g^T x$ by at most a given margin $\rho$, for all $x \in \mathcal{X}$?

Let us focus on a “half” of the answer: how to ensure that the probability of the event $h^T \omega + \kappa > g^T x + \rho$ does not exceed $\epsilon/2$, for every $x \in \mathcal{X}$. The answer becomes easy when assuming that we have at our disposal an upper bound on exponential moments of the distributions from the family — a function $\Phi(h; x)$ such that

$$
\ln \left( \int e^{h^T \omega} P_x(d\omega) \right) \leq \Phi(h; x) \forall (h \in \mathbb{R}^n, x \in \mathcal{X}).
$$

Indeed, for obvious reasons, in this case the $P_x$-probability of the event $h^T \omega + \kappa > g^T x + \rho$ is at most

$$
\exp \{ \Phi(h; x) - (g^T x + \rho - \kappa) \}.
$$

To add some flexibility, note that when $\alpha > 0$, the event in question is the same as the event $(h/\alpha)^T \omega + \kappa/\alpha > [g^T x + \rho]/\alpha$, thus we arrive at a parametric family of upper bounds

$$
\exp \{ \Phi(h/\alpha; x) - [g^T x + \rho - \kappa]/\alpha \}, \alpha > 0,
$$

on the $P_x$-probability of our “bad” event. It follows that a sufficient condition for this probability to be $\leq \epsilon/2$, for a given $x \in \mathcal{X}$, is the existence of $\alpha > 0$ such that

$$
\exp \{ \Phi(h/\alpha; x) - [g^T x + \rho - \kappa]/\alpha \} \leq \epsilon/2;
$$

same as

$$
\Phi(h/\alpha; x) - [g^T x + \rho - \kappa]/\alpha \leq \ln(\epsilon/2),
$$
or, which again is the same, the existence of \( \alpha > 0 \) such that
\[
\alpha \Phi \left( \frac{h}{\alpha}; x \right) + \alpha \ln \left( \frac{2}{\epsilon} \right) - g^T x \leq \rho - \kappa.
\]
In other words, a sufficient condition for the relation
\[
\text{Prob}_{\omega \sim P_x} \{ h^T \omega + \kappa > g^T x + \rho \} \leq \epsilon/2
\]
is
\[
\inf_{\alpha > 0} \left[ \alpha \Phi \left( \frac{h}{\alpha}; x \right) + \alpha \ln \left( \frac{2}{\epsilon} \right) - g^T x \right] \leq \rho - \kappa.
\]
If we want the bad event in question to take place with \( P_x \)-probability \( \leq \epsilon/2 \) whatever be \( x \in X \), the sufficient condition for this is
\[
\sup_{x \in X} \inf_{\alpha > 0} \left[ \alpha \Phi \left( \frac{h}{\alpha}; x \right) + \alpha \ln \left( \frac{2}{\epsilon} \right) - g^T x \right] \leq \rho - \kappa.
\] (3.1)

Now assume that \( X \) is convex and compact, and \( \Phi(h; x) \) is continuous, convex in \( h \) and concave in \( x \). In this case the function \( \alpha \Phi(h/\alpha; x) \) is convex in \( (h, \alpha) \) in the domain \( \alpha > 0 \) and is concave in \( x \), so that we can switch sup and inf, thus arriving at the sufficient condition
\[
\exists \alpha > 0 : \max_{x \in X} \left[ \alpha \Phi(h/\alpha; x) + \alpha \ln \left( \frac{2}{\epsilon} \right) - g^T x \right] \leq \rho - \kappa,
\] (3.2)
for the validity of the relation
\[
\forall x \in X : \text{Prob}_{\omega \sim P_x} \{ h^T \omega + \kappa - g^T x \leq \rho \} \geq 1 - \epsilon/2.
\]

Note that our sufficient condition is expressed in terms of a convex constraint on \( h, \kappa, \rho, \alpha \). Consider also the dramatic simplification allowed by convexity-concavity of \( \Phi \): in (3.1), every \( x \in X \) should be “served” by its own \( \alpha \), so that (3.1) is an infinite system of constraints on \( h, \rho, \kappa \). In contrast, in (3.2) all \( x \in X \) are “served” by a single \( \alpha \).

The developments in this section and Sections 3.3, 3.4 are no more that implementations, under various circumstances, of the simple idea we have just outlined.

### 3.1.1 The problem

Let \( \mathcal{O} = (\Omega, \Pi, \{p_\mu(\cdot) : \mu \in \mathcal{M}\}, \mathcal{F}) \) be a simple observation scheme (see Section 2.4.2). The problem we consider in this section is as follows:

We are given a positive integer \( K \) and \( I \) nonempty convex compact sets \( X_j \subset \mathbb{R}^n \), along with affine mappings \( A_j(\cdot) : \mathbb{R}^n \to \mathbb{R}^M \) 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 \( \mathbb{R}^n \). Given random observation
\[
\omega^K = (\omega_1, \ldots, \omega_K)
\]
with \( \omega_k \) drawn, independently across \( k \), from \( p_{A_j(x)} \) with \( j \leq I \) and \( x \in X_j \),

---

1This is due to the following standard fact: if \( f(h) \) is a convex function, then the projective transformation \( \alpha f(h/\alpha) \) of \( f \) is convex in \( (h, \alpha) \) in the domain \( \alpha > 0 \).
we want to recover \( g^T x \).

It should be stressed that we do not know \( j \) and \( x \) underlying our observation.

Given reliability tolerance \( \epsilon \in (0, 1) \), we quantify the performance of a candidate estimate – a Borel function \( \hat{g}(\cdot) : \Omega \to \mathbb{R} \) – by the worst case, over \( j \) and \( x \), width of \( (1 - \epsilon) \)-confidence interval. Precisely, we say that \( \hat{g}(\cdot) \) is \((\rho, \epsilon)\)-reliable, if

\[
\forall (j \leq I, x \in X_j) : \text{Prob}_{\omega \sim \mathcal{P}_{A_j(\cdot)}} \{ |\hat{g}(\omega) - g^T x| > \rho \} \leq \epsilon. \tag{3.3}
\]

We define the \( \epsilon \)-risk of the estimate as

\[
\text{Risk}_\epsilon[\hat{g}] = \inf \{ \rho : \hat{g} \text{ is } (\rho, \epsilon)\text{-reliable} \},
\]

i.e., \( \text{Risk}_\epsilon[\hat{g}] \) is the smallest \( \rho \) such that \( \hat{g} \) is \((\rho, \epsilon)\)-reliable.

The technique we are about to develop originates from [130] where estimating a linear form on a convex compact set in a simple o.s. (i.e., the case \( I = 1 \) of the problem at hand) was considered, and where it was proved that in this situation the estimate

\[
\hat{g}(\omega^K) = \sum_k \phi(\omega_k) + \kappa
\]

with properly selected \( \phi \in \mathcal{F} \) and \( \kappa \in \mathbb{R} \) is near-optimal. The problem of estimating linear functionals of a signal in Gaussian o.s. has a long history, see, e.g. [38, 40, 122, 125, 126, 123, 166, 176] and references therein. In particular, in the case of \( I = 1 \), using different techniques, a similar fact was proved by D. Donoho [65] in 1991, related results in the case of \( I > 1 \) are available in [41, 42].

### 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

\[
\Phi_{\mathcal{O}}(\phi; \mu) = \ln \left( \int e^{\phi(\omega)} p_\mu(\omega) \Pi(d\omega) \right), \quad (\phi, \mu) \in \mathcal{F} \times \mathcal{M}.
\]

Recall that by definition of a simple o.s., this function is real-valued on \( \mathcal{F} \times \mathcal{M} \), 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 \((i, j)\), \( 1 \leq i, j \leq I \), the functions

\[
\Phi_{ij}(\alpha, \phi; x, y) = \tfrac{1}{2} [K\alpha \Phi_{\mathcal{O}}(\phi/\alpha; A_i(x)) + K\alpha \Phi_{\mathcal{O}}(-\phi/\alpha; A_j(y)) + g^T(y - x) + 2\alpha \ln(2I/\epsilon)] : \{ \alpha > 0, \phi \in \mathcal{F} \} \times [X_i \times X_j] \to \mathbb{R},
\]

\[
\Psi_{ij}(\alpha, \phi) = \max_{x \in X_i, y \in X_j} \Phi_{ij}(\alpha, \phi; x, y) = \tfrac{1}{2} [\Psi_{i,+}(\alpha, \phi) + \Psi_{j,-}(\alpha, \phi)] : \{ \alpha > 0 \} \times \mathcal{F} \to \mathbb{R}
\]

where

\[
\Psi_{i,+}(\beta, \psi) = \max_{x \in X_i} [K\beta \Phi_{\mathcal{O}}(\psi/\beta; A_i(x)) - g^T x + \beta \ln(2I/\epsilon)] : \{ \beta > 0, \psi \in \mathcal{F} \} \to \mathbb{R},
\]

\[
\Psi_{i,-}(\beta, \psi) = \max_{x \in X_i} [K\beta \Phi_{\mathcal{O}}(-\psi/\beta; A_i(x)) + g^T x + \beta \ln(2I/\epsilon)] : \{ \beta > 0, \psi \in \mathcal{F} \} \to \mathbb{R}.
\]
Note that the function $\alpha \Phi_O(\phi/\alpha; A_i(x))$ is obtained from the continuous convex-concave function $\Phi_O(\cdot, \cdot)$ 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 $\{\alpha > 0, \phi \in X\} \times X_i$. By similar argument, the function $\alpha \Phi_O(-\phi/\alpha; A_j(y))$ is convex-concave and continuous on the domain $\{\alpha > 0, \phi \in F\} \times X_j$. These observations combine with compactness of $X_i$ and $X_j$ to imply that $\Psi_{ij}(\alpha, \phi)$ is a real-valued continuous convex function on the domain $F^+ = \{\alpha > 0\} \times F$.

Observe that functions $\Psi_{ii}(\alpha, \phi)$ are nonnegative on $F^+$. Indeed, selecting somehow $\bar{x} \in X_i$, and setting $\mu = A_i(\bar{x})$, we have

$$\Psi_{ii}(\alpha, \phi) \geq \Phi_{ii}(\alpha, \phi; \bar{x}, \bar{x}) = \alpha \left[ \Phi_O(0; \mu) \right] + \ln(2I/\epsilon) \geq 0$$

(we have used convexity of $\Phi_O$ in the first argument).

Functions $\Psi_{ij}$ give rise to convex and feasible optimization problems

$$\text{Opt}_{ij} = \text{Opt}_{ij}(K) = \min_{(\alpha, \phi) \in F^+} \Psi_{ij}(\alpha, \phi). \tag{3.4}$$

By its origin, $\text{Opt}_{ij}$ is either a real, or $-\infty$; by the observation above, $\text{Opt}_{ii}$ are nonnegative. Our estimate is as follows.

1. For $1 \leq i, j \leq I$, we select somehow feasible solutions $\alpha_{ij}, \phi_{ij}$ to problems (3.4) (the less the values of the corresponding objectives, the better) and set

$$\rho_{ij} = \Psi_{ij}(\alpha_{ij}, \phi_{ij}) = \frac{1}{2} \left[ \Psi_{i,+}(\alpha_{ij}, \phi_{ij}) + \Psi_{j,-}(\alpha_{ij}, \phi_{ij}) \right]$$

$$\kappa_{ij} = \frac{1}{2} \left[ \Psi_{j,-}(\alpha_{ij}, \phi_{ij}) - \Psi_{i,+}(\alpha_{ij}, \phi_{ij}) \right]$$

$$g_{ij}(\omega^K) = \sum_{k=1}^{K} \phi_{ij}(\omega_k) + \kappa_{ij}$$

$$\rho = \max_{1 \leq i, j \leq I} \rho_{ij} \tag{3.5}$$

2. Given observation $\omega^K$, we specify the estimate $\hat{g}(\omega^K)$ as follows:

$$r_i = \max_{1 \leq j \leq I} g_{ij}(\omega^K)$$

$$c_j = \min_{1 \leq i \leq I} g_{ij}(\omega^K)$$

$$\hat{g}(\omega^K) = \frac{1}{2} \left[ \min_{1 \leq i \leq I} r_i + \max_{1 \leq j \leq I} c_j \right]. \tag{3.6}$$

3.1.3 Main result

**Proposition 3.1.** The $\epsilon$-risk of the estimate $\hat{g}(\omega^K)$ can be upper-bounded as follows:

$$\text{Risk}_\epsilon[\hat{g}] \leq \rho. \tag{3.7}$$

**Proof.** Let the common distribution $p$ of independent across $k$ components $\omega_k$ in observation $\omega^K$ be $p_{A_i(u)}$ for some $\ell \leq I$ and $u \in X_{\ell}$. Let us fix these $\ell$ and $u$; we denote $\mu = A_\ell(u)$ and let $p^K$ stand for the distribution of $\omega^K$. 
FROM HYPOTHESIS TESTING TO ESTIMATING FUNCTIONALS

1°. We have

$$\Psi_{\epsilon, +}(\alpha_{ij}, \phi_{ij}) = \max_{i \in X_i} \left[ K\alpha_{ij} \Phi_{\omega}(\phi_{ij}/\alpha_{ij}, A_t(x)) - g^T x \right] + \alpha_{ij} \ln(21/e)$$

$$\geq K\alpha_{ij} \Phi_{\omega}(\phi_{ij}/\alpha_{ij}, \mu) - g^T u + \alpha_{ij} \ln(21/e) \quad \text{[since } u \in X_i \text{ and } \mu = A_t(u)]$$

$$= K\alpha_{ij} \ln \left( \int \exp\left( \phi_{ij}(\omega)/\alpha_{ij} \right) p_{i}(\omega) d(\omega) \right) - g^T u + \alpha_{ij} \ln(21/e)$$

$$\geq \alpha_{ij} \ln \left( \int \exp\left( \phi_{ij}(\omega)/\alpha_{ij} \right) p_{i}(\omega) d(\omega) \right) - g^T u + \alpha_{ij} \ln(21/e)$$

$$\geq \alpha_{ij} \ln \left( \exp\left( \alpha_{ij}^{-1}[g(t_{ij} - x_{ij})] \right) \right) - g^T u + \alpha_{ij} \ln(21/e)$$

$$\geq \alpha_{ij} \ln \left( \exp\left( \alpha_{ij}^{-1}[g(t_{ij} - x_{ij})] \right) \right) + \rho_{ij} - x_{ij} + \alpha_{ij} \ln(21/e)$$

$$= \alpha_{ij} \ln \left( \exp\left( \alpha_{ij}^{-1}[g(t_{ij} - x_{ij})] \right) \right) + \rho_{ij} - x_{ij} + \alpha_{ij} \ln(21/e)$$

$$\geq \alpha_{ij} \ln \left( \exp\left( \alpha_{ij}^{-1}[g(t_{ij} - x_{ij})] \right) \right) + \rho_{ij} - x_{ij} + \alpha_{ij} \ln(21/e)$$

and we arrive at

$$\text{Prob}_{\alpha_{ij} \sim \kappa_{ij}} \{ g(t_{ij}(\omega^K)) > g^T u + \rho_{ij} \} \leq \frac{\epsilon}{21}. \quad (3.8)$$

Similarly,

$$\Psi_{\epsilon, -}(\alpha_{ij}, \phi_{ij}) = \max_{i \in X_i} \left[ K\alpha_{ij} \Phi_{\omega}(-\phi_{ij}/\alpha_{ij}, A_t(y)) + g^T y \right] + \alpha_{ij} \ln(21/e)$$

$$\geq K\alpha_{ij} \Phi_{\omega}(-\phi_{ij}/\alpha_{ij}, \mu) + g^T u + \alpha_{ij} \ln(21/e) \quad \text{[since } u \in X_i \text{ and } \mu = A_t(u)]$$

$$= K\alpha_{ij} \ln \left( \int \exp\left( -\phi_{ij}(\omega)/\alpha_{ij} \right) p_{i}(\omega) d(\omega) \right) + g^T u + \alpha_{ij} \ln(21/e)$$

$$\geq \alpha_{ij} \ln \left( \int \exp\left( -\phi_{ij}(\omega)/\alpha_{ij} \right) p_{i}(\omega) d(\omega) \right) + g^T u + \alpha_{ij} \ln(21/e)$$

$$\geq \alpha_{ij} \ln \left( \exp\left( \alpha_{ij}^{-1}[g(t_{ij} + x_{ij})] \right) \right) + \rho_{ij} + x_{ij} - \alpha_{ij} \ln(21/e)$$

$$= \alpha_{ij} \ln \left( \exp\left( \alpha_{ij}^{-1}[g(t_{ij} + x_{ij})] \right) \right) + \rho_{ij} + x_{ij} - \alpha_{ij} \ln(21/e)$$

and we arrive at

$$\text{Prob}_{\alpha_{ij} \sim \kappa_{ij}} \left\{ g(t_{ij}(\omega^K)) < g^T u - \rho_{ij} \right\} \leq \frac{\epsilon}{21}. \quad (3.9)$$

2°. Let

$$E = \{ \omega^K : g(t_{ij}(\omega^K)) \leq g^T u + \rho_{ij}, \quad g(t_{ij}(\omega^K)) \geq g^T u - \rho_{ij}, 1 \leq i, j \leq I \}.$$

From (3.8), (3.9) and the union bound it follows that $p^K$-probability of the event $E$ is $\geq 1 - \epsilon$. As a result, all we need to complete the proof of the proposition is to verify that

$$\omega^K \in E \Rightarrow |g(\omega^K) - g^T u| \leq \rho. \quad (3.10)$$

To this end, let us fix $\omega^K \in E$, 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 (3.6), is the maximum of the entries in the $i$-th row of $E$, while the quantity $c_j$ is the minimum of the entries in the $j$-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, let

$$\Delta = [g^T u - \rho, g^T u + \rho].$$
Since $\omega^K \in \mathcal{E}$, we have $E_{i\ell} = g_{i\ell}(\omega^K) \geq g^T u - \rho_{i\ell} \geq g^T u - \rho$ and $E_{i\ell} = g_{i\ell}(\omega^K) \leq g^T u + \rho_{i\ell} \leq g^T u + \rho$ for all $j$, implying that $r_\ell = \max_j E_{i\ell} \in \Delta$. Similarly, $\omega^K \in \mathcal{E}$ implies that $E_{i\ell} = g_{i\ell}(\omega^K) \leq g^T u + \rho$ and $E_{i\ell} = g_{i\ell}(\omega^K) \geq g^T u - \rho$ for all $i$, implying that $c_\ell = \min_i E_{i\ell} \in \Delta$. We see that both $r_\ell$ and $c_\ell$ belong to $\Delta$; 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, \( \hat{g}(\omega^K) = \frac{1}{2}[r^* + c^*] \), that is, \( \hat{g}(\omega^K) \in \Delta \), and the conclusion in (3.10) indeed takes place.

Remark 3.2. Let us consider a special case of $I = 1$. In this case, given a $K$-repeated observation of the signal in a simple o.s., our construction yields an estimate of a linear form $g^T x$ of unknown signal $x$, known to belong to a given convex compact set $X_1$. This estimate is

\[
\hat{g}(\omega^K) = \sum_{k=1}^{K} \phi(\omega_k) + \kappa, \quad (3.11)
\]

and is associated with the optimization problem

\[
\min_{\alpha > 0, \phi \in \mathcal{F}} \{ \Psi(\alpha, \phi) := \frac{1}{2}[\Psi_+(\alpha, \phi) + \Psi_-(\alpha, \phi)] \},
\]

\[
\Psi_+(\alpha, \phi) = \max_{x \in X_1} \left[ K \alpha \Phi_O(\phi/\alpha, A_1(x)) - g^T x + \alpha \ln(2/\epsilon) \right],
\]

\[
\Psi_-(\alpha, \phi) = \max_{x \in X_1} \left[ K \alpha \Phi_O(-\phi/\alpha, A_1(x)) + g^T x + \alpha \ln(2/\epsilon) \right].
\]

By Proposition 3.1, when $\alpha, \phi$ is a feasible solution to the problem and

\[
\kappa = \frac{1}{2} [\Psi_-(\alpha, \phi) - \Psi_+(\alpha, \phi)],
\]

the $\epsilon$-risk of estimate (3.11) does not exceed $\Psi(\alpha, \phi)$.

3.1.4 Near-optimality

Observe that by properly selecting $\phi_{ij}$ and $\alpha_{ij}$ we can make, in a computationally efficient manner, the upper bound $\rho$ on the $\epsilon$-risk of the above estimate arbitrarily close to

\[
\text{Opt}(K) = \max_{1 \leq i, j \leq I} \text{Opt}_{ij}(K).
\]

We are about to demonstrate that the quantity $\text{Opt}(K)$ “nearly lower-bounds” the minimax optimal $\epsilon$-risk

\[
\text{Risk}^*_\epsilon(K) = \inf_{\hat{g}} \text{Risk}_\epsilon[\hat{g}],
\]

the infimum being taken over all estimates (all Borel functions of $\omega^K$). The precise statement is as follows:

**Proposition 3.3.** In the situation of this section, let $\epsilon \in (0, 1/2)$ and $K$ be a positive integer. Then for every integer $K$ satisfying

\[
\frac{K}{K} > \frac{2 \ln(2I/\epsilon)}{\ln \left( \frac{1}{\epsilon(1-\epsilon)} \right)} \quad (3.12)
\]
one has
\[
\text{Opt}(K) \leq \text{Risk}_c^*(K). \quad (3.13)
\]
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
\[
K \geq K \Rightarrow \text{Opt}(K) \leq \frac{2 \ln(2I/c)}{\ln \left(\frac{1}{4\epsilon(1-\epsilon)}\right)} \text{Risk}_c^*(K). \quad (3.14)
\]
For proof, see Section 3.6.1.

3.1.5 Illustration

We illustrate our construction with the simplest possible example in which \(X_i = \{x_i\}\) are singletons in \(\mathbb{R}^n\), \(i = 1, \ldots, I\), and the observation scheme is Gaussian. Thus, setting \(y_i = A_i(x_i) \in \mathbb{R}^m\), the observation’s components \(\omega_k\), \(1 \leq k \leq K\), stemming from the 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 \(\phi\) 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 \(\mathbb{R}^m\)); we identify \(\phi \in \mathcal{F}\) with the pair \((\phi_0, \varphi)\). The function \(\Psi_O\) associated with the Gaussian observation scheme with \(m\)-dimensional observations is
\[
\Phi_O(\phi; \mu) = \phi_0 + \varphi^T \mu + \frac{1}{2} \varphi^T \varphi : (\mathbb{R} \times \mathbb{R}^m) \times \mathbb{R}^m \to \mathbb{R},
\]
a straightforward computation shows that in the case in question, setting \(\theta = \ln(2I/c)\), we have
\[
\Psi_{i,+}(\alpha, \phi) = \frac{1}{2} y_i^T x_i - \frac{1}{2} \varphi^T \varphi + \frac{1}{2} \varphi^T y_i + \frac{1}{2} \varphi^T x_i + g^T x_i + \frac{1}{2} \varphi^T \varphi + \alpha \theta,
\]
\[
\Psi_{j,-}(\alpha, \phi) = \frac{1}{2} y_j^T x_j - \frac{1}{2} \varphi^T \varphi + \frac{1}{2} \varphi^T y_j + \frac{1}{2} \varphi^T x_j + g^T x_j + \frac{1}{2} \varphi^T \varphi + \alpha \theta,
\]
\[
\text{Opt}_{ij} = \inf_{\alpha > 0, \varphi} \left[ \frac{1}{2} \Psi_{i,+}(\alpha, \phi) + \Psi_{j,-}(\alpha, \phi) \right]
\]
\[
= \frac{1}{2} g^T [x_j - x_i] + \inf_{\varphi} \left[ \frac{1}{2} \varphi^T [y_i - y_j] + \inf_{\alpha > 0} \left[ \frac{1}{2} \varphi^T \varphi + \alpha \theta \right] \right]
\]
\[
= \frac{1}{2} g^T [x_j - x_i] + \inf_{\varphi} \left[ \frac{1}{2} \varphi^T [y_i - y_j] + \sqrt{2\theta} \|\varphi\|_2 \right]
\]
\[
= \begin{cases} 
\frac{1}{2} g^T [x_j - x_i], & \|y_i - y_j\|_2 \leq 2\sqrt{2\theta / K} \\
-\infty, & \|y_i - y_j\|_2 > 2\sqrt{2\theta / K}.
\end{cases}
\]
We see that we can put \(\phi_0 = 0\), and that setting
\[
\mathcal{I} = \{(i, j) : \|y_i - y_j\|_2 \leq 2\sqrt{2\theta / K}\},
\]
Opt_{ij}(K) is finite if and only if \((i, j) \in \mathcal{I}\) and is \(-\infty\) otherwise. In both cases, the optimization problem specifying \(\text{Opt}_{ij}\) has no optimal solution.\(^2\) Indeed, this clearly is the case when \((i, j) \notin \mathcal{I}\); 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 limit is not in the minimization domain (on this domain, \(\alpha\) should be positive). In this particular case, the simplest way to overcome the difficulty is to restrict the optimization domain \(\mathcal{F}^+\) in (3.4) with its compact subset \(\{\alpha \geq 1/R, \phi_0 = 0, \|\varphi\|_2 \leq R\}\) with large \(R\), like \(R = 10^{10}\) or \(10^{20}\). Then we specify the entities participating in (3.5) as

\[
\begin{align*}
\phi_{ij}(\omega) &= \varphi^T_{ij}\omega, \\
\varphi_{ij} &= \left\{ \begin{array}{ll}
0, & (i, j) \in \mathcal{I} \\
1/R, & (i, j) \notin \mathcal{I}
\end{array} \right., \\
\alpha_{ij} &= \left\{ \begin{array}{ll}
1/R, & (i, j) \in \mathcal{I} \\
\sqrt{K/R}, & (i, j) \notin \mathcal{I}
\end{array} \right.
\end{align*}
\]

resulting in

\[
\begin{align*}
\zeta_{ij} &= \frac{1}{2} \left[ \Psi_{i,-}(\alpha_{ij}, \phi_{ij}) - \Psi_{i,+}(\alpha_{ij}, \phi_{ij}) \right] \\
&= \frac{1}{2} \left[ -K \varphi^T_{ij} y_j + g^T \varphi_{ij} x_j + \frac{K}{2\alpha_{ij}} \varphi^T_{ij} \varphi_{ij} + \alpha_{ij} \theta - K \varphi^T_{ij} y_i + g^T x_i - \frac{K}{2\alpha_{ij}} \varphi^T_{ij} \varphi_{ij} - \alpha_{ij} \theta \right] \\
&= \frac{1}{2} g^T [x_i + x_j] - \frac{K}{2 \alpha_{ij}} \varphi^T_{ij} y_i + y_j
\end{align*}
\]

and

\[
\begin{align*}
\rho_{ij} &= \frac{1}{2} \left[ \Psi_{i,+}(\alpha_{ij}, \phi_{ij}) + \Psi_{i,-}(\alpha_{ij}, \phi_{ij}) \right] \\
&= \frac{1}{2} \left[ K \varphi^T_{ij} y_i - g^T x_i + \frac{K}{2\alpha_{ij}} \varphi^T_{ij} \varphi_{ij} + \alpha_{ij} \theta - K \varphi^T_{ij} y_j + g^T x_j + \frac{K}{2\alpha_{ij}} \varphi^T_{ij} \varphi_{ij} - \alpha_{ij} \theta \right] \\
&= \frac{K}{2\alpha_{ij}} \varphi^T_{ij} \phi_{ij} + \alpha_{ij} \theta + \frac{1}{2} g^T [x_j - x_i] + \frac{K}{2} \varphi^T_{ij} y_i - y_j
\end{align*}
\]

\[
\begin{align*}
= \left\{ \begin{array}{ll}
\frac{1}{2} g^T [x_j - x_i] + R^{-1} \theta, & (i, j) \in \mathcal{I}, \\
\frac{1}{2} g^T [x_j - x_i] + [\sqrt{2K} \theta - \frac{K}{2} \|y_i - y_j\|_2]R, & (i, j) \notin \mathcal{I}
\end{array} \right.
\]

(3.15)

In the numerical experiment we report on we use \(n = 20, m = 10, \) 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). The linear form to be recovered is the first coordinate of \(x\), the confidence parameter is set to \(\epsilon = 0.01\), and \(R = 10^{20}\). Results of a typical experiment are presented in Figure 3.1.

### 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.

Given are:

- a simple o.s. \(\mathcal{O} = (\Omega, \Pi; \{p_{\mu}: \mu \in \mathcal{M}\}; \mathcal{F})\),
- a signal space \(X \subset \mathbb{R}^n\) along with the affine mapping \(x \mapsto Ax: X \to \mathcal{M}\),
- a real-valued function \(f\) on \(X\).

\(^2\)Handling this case was exactly the reason why in our construction we required \(\phi_{ij}, \alpha_{ij}\) to be feasible, and not necessary optimal, solutions to the optimization problems (3.4).
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

Though the estimator we develop is, in a nutshell, quite simple, its formal description turns out to be rather involved.\footnote{It should be mentioned that the proposed estimation procedure is a “close relative” of the binary search algorithm of [77].} For this reason we start its presentation with an informal outline, which exposes some simple ideas underlying its construction.

Consider the situation where the signal space $X$ is the 2D rectangle as presented on the top of Figure 3.2.(a), and let the function to be recovered be $f(x) = x_1$. Thus, “the nature” has somehow selected $x = [x_1, x_2]$ in the rectangle, and we observe a Gaussian random vector 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 our 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 that neither of the hypotheses 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 the Bisection might look like.

1. Were we able to decide reliably on the left and the right hypotheses in Figure 3.2.(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 3.2.(a), and then iterate this process. The difficulty, of course, is that our left and right hypotheses intersect, so that is impossible to decide on them reliably.

2. In order to make left and right hypotheses distinguishable from each other, we could act as shown in Figure 3.2.(b), by shrinking the left and the right rectangles and inserting in between the middle ("no man’s land") rectangle. Assuming that the width of the middle rectangle allows to decide reliably on our new left and right hypotheses and utilizing available observation, we can localize $x$ either in the left, or in the right rectangles as shown in Figure 3.2.(b). Specifically, assume that our "left vs. right" test rejected correctly the right hypothesis. Then $x$ can be located either in the left, or in the middle rectangle shown on the top, and thus $x$ is in the new left localizer which is the union of the left and the middle original rectangles. Similarly, if our test rejects correctly the left hypothesis, then we can take, as the new localizer of $x$, the union of the original right and middle rectangles. Note that our localization is as reliable as our test is, and that it reduces the width of localizer by a factor close to 2, provided the width of the middle rectangle is small if compared to the width of the original localizer of $x$. We can iterate this process, until we arrive at a localizer so narrow that the corresponding separator – "no man’s land" (this part cannot be too narrow, since it should allow for reliable decision on the current left and right hypotheses) becomes too large to allow reducing significantly the localizer’s width.

Note that in this implementation of the binary search (same as in the implementation proposed in [77]), starting from the second step of the Bisection, the hypotheses to decide upon depend on the observations (e.g., when $x$ belongs to the middle part of the three-rectangle localizer in Figure 3.2, deciding on “left vs. right” can, depending on observation, result in accepting either left or right hypothesis, thus leading to different updated localizers). Analysing this situation usually
brings about complications we would like to avoid.

3. A simple modification of the Bisection allows to circumvent the difficulties related to testing random hypotheses. Indeed, let us consider the following construction: given current localizer for $x$ (at the first step – the initial rectangle), we consider two “three-rectangle” partitions of it as presented in Figure 3.2.(c). In the first partition, the left rectangle is the left half of the original rectangle, in the second partition the right rectangle is the right half of the original rectangle. We then run two “left vs. right” tests, the first on the pair of left and right hypotheses stemming from the first partition, and the second on the pair of left and right 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 the following conclusions:

- when both tests reject the right 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 the left hypotheses from the corresponding pairs, $x$ is located in the right half of the original rectangle (for the exactly same reasons as in the previous case);
- when the tests “disagree,” rejecting hypotheses of different types (like left in the firsts, and right in the second test), $x$ is located in the union of the two middle rectangles we deal with. Indeed, otherwise $x$ should be either in the left rectangles of both our three-rectangle partitions, or in the right 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 right hypotheses, and both should reject left hypotheses in the second, while none of these events took place.

Now, in the first two cases we can safely say to which of the “halves” – left or right – of the initial rectangle $x$ belongs, and take this half as the new localizer. In the third case, we take as a new localizer for $x$ the middle rectangle shown at the bottom of Figure 3.2 and terminate our estimation process – the new localizer already is narrow! In the proposed algorithm, unless we terminate at the very first step, we carry out the second step exactly in the same way 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 way, 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 the present situation, in contrast to that 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 deterministic Bisection, that is, were updating subsequent rectangles $\Delta_t$ containing $x$ according to the rules

- $\Delta_1$ is a given in advance rectangle containing $x$,
- $\Delta_{t+1}$ is precisely the half of $\Delta_t$ containing $x$ (say, the left half in the case of a tie).
Thus, given \( x \) and assuming that there are no wrong rejections, the situation is as if a single observation were used in \( L \) tests run in “parallel” rather than sequentially. The only elaboration caused by the sequential nature of our process is the “risk accumulation” – we want the probability of error in one or more of our \( L \) tests to be less than the desired risk \( \epsilon \) of wrong “bracketing” of \( f(x) \), implying, in the absence of something better, that the risks of the individual tests should be at most \( \epsilon/L \). These risks, in turn, define the allowed width of separators 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’s land” which at the concluding Bisection steps is of order of \( 2^{-L} \) would be too small to allow for deciding on the concluding pairs of our hypotheses with risk \( \epsilon/L \), at least when our observations possess non-negligible volatility). As a result, “the price” of Bisection turns out to be significantly lower than in the case where every test uses its own observation.

From the above sketch of our construction 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 2.23. 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. We are given

1. simple o.s. \( \mathcal{O} = (\Omega, P, \{ p_\mu(\cdot) : \mu \in \mathcal{M} \}, \mathcal{F}) \),
2. convex compact set \( X \subset \mathbb{R}^n \) along with a collection of \( I \) convex compact sets \( X_i \subset X \),
3. affine mapping \( x \mapsto A(x) : X \rightarrow \mathcal{M} \),
4. a continuous function \( f(x) : X \rightarrow \mathbb{R} \) which is \( N \)-convex, meaning that for every \( a \in \mathbb{R} \) the sets \( X^a_{\geq} = \{ x \in X : f(x) \geq a \} \) and \( X^a_{\leq} = \{ x \in X : f(x) \leq a \} \) can be represented as the unions of at most \( N \) closed convex sets \( X^a_{\geq, \nu}, X^a_{\leq, \nu} \):

\[
X^a_{\geq} = \bigcup_{\nu=1}^{N} X^a_{\geq, \nu}, \quad X^a_{\leq} = \bigcup_{\nu=1}^{N} X^a_{\leq, \nu}.
\]

For some unknown \( x \) known to belong to \( X = \bigcup_{i=1}^{I} X_i \), we have at our disposal observation \( \omega^K = (\omega_1, ..., \omega_K) \) with i.i.d. \( \omega_i \sim p_{A(x)}(\cdot) \), 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 \( \hat{f}(\omega^K) \) \( (\rho, \epsilon) \)-reliable (cf. (3.3)), if for every \( x \in X \), with the \( p_{A(x)} \)-probability at least \( 1 - \epsilon \) it holds \( |\hat{f}(\omega^K) - f(x)| \leq \rho \) or, which is the same, if

\[
\forall (x \in X) : \text{Prob}_{\omega^K \sim p_{A(x)} \times ... \times p_{A(x)}} \left\{ |\hat{f}(\omega^K) - f(x)| > \rho \right\} \leq \epsilon.
\]
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 $g^\nu(x)$, $h^\nu(x)$ are affine functions on $\mathcal{X}$ and $h^\nu$ 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 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, $i = 1, 2, \ldots, I$, then $f(x) = \max_i f_i(x)$ is $N$-convex with

$$N = \max \left[ \prod_i N_i, \sum_i N_i \right],$$

due to

$$\{x \in \mathcal{X} : f(x) \leq a\} = \bigcap_{i=1}^I \{x : f_i(x) \leq a\},$$
$$\{x \in \mathcal{X} : f(x) \geq a\} = \bigcup_{i=1}^I \{x : f_i(x) \geq a\}.$$

Note that the first 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 set is the union of $I$ unions of convex sets with $N_i$ elements in the $i$-th union, and thus is the union of $\sum_i N_i$ convex sets.

Example 3.2. [Conditional quantile] Let $S = \{s_1 < s_2 < \ldots < s_M\} \subset \mathbb{R}$. For a nonvanishing probability distribution $q$ on $S$ and $\alpha \in [0, 1]$, let $\chi_\alpha[q]$ be the regularized $\alpha$-quantile of $q$ defined as follows: we pass from $q$ to the distribution on $[s_1, s_M]$ by spreading uniformly the mass $q_{\nu}$, $1 < \nu \leq M$, over $[s_{\nu-1}, s_\nu]$, and assigning mass $q_1$ to the point $s_1$; $\chi_\alpha[q]$ is the usual $\alpha$-quantile of the resulting distribution $\bar{q}$:

$$\chi_\alpha[q] = \min\{s \in [s_1, s_M] : \bar{q}\{[s_1, s]\} \geq \alpha\}.$$

Regularized quantile as function of $\alpha$, $M = 4$

Given, along with $S$, a finite set $T$, let $\mathcal{X}$ be a convex compact set in the space of nonvanishing probability distributions on $S \times T$. For $\tau \in T$, consider the conditional
to \( t = \tau \), distribution \( p_r(\cdot) \) of \( s \in S \) induced by a distribution \( p(\cdot,\cdot) \in \mathcal{X} \):

\[
p_r(\mu) = \frac{p(\mu, \tau)}{\sum_{\nu=1}^{M} p(\nu, \tau)}, 1 \leq \mu \leq M,
\]

where \( p(\mu, \tau) \) is the \( p \)-probability for \((s,t)\) to take value \((s_\mu, \tau)\), and \( p_r(\mu) \) is the \( p \)-probability for \( s \) to take value \( s_\mu \), \( 1 \leq \mu \leq M \).

The function \( \chi_\alpha[p_r] : \mathcal{X} \to \mathbb{R} \) turns out to be 1-convex, for verification see Section 3.6.2.

### 3.2.3 Bisection estimate: construction

While the construction to be presented admits numerous refinements, we focus here on its simplest version.

#### 3.2.3.1 Preliminaries

**Upper and lower feasibility/infeasibility, sets** \( Z_{i,\geq}^a \) and \( Z_{i,\leq}^a \). Let \( a \) be a real. We associate with \( a \) a collection of upper \( a \)-sets defined as follows: we look at the sets \( X_i \cap X_{a,\geq}^\nu \), \( 1 \leq i \leq I \), \( 1 \leq \nu \leq N \), and arrange the nonempty sets from this family into a sequence \( Z_{i,\geq}^a \), \( 1 \leq i \leq I_{a,\geq} \). Here \( I_{a,\geq} = 0 \) if all sets in the family are empty; in the latter case, we call \( a \) upper-infeasible, and call it upper-feasible otherwise. Similarly, we associate with \( a \) the collection of lower \( a \)-sets \( Z_{i,\leq}^a \), \( 1 \leq i \leq I_{a,\leq} \) by arranging into a sequence all nonempty sets from the family \( X_i \cap X_{a,\leq}^\nu \), 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

\[
X_{a,\geq} := \{ x \in X : f(x) \geq a \} = \bigcup_{1 \leq i \leq I_{a,\geq}} Z_{i,\geq}^a,
\]

\[
X_{a,\leq} := \{ x \in X : f(x) \leq a \} = \bigcup_{1 \leq i \leq I_{a,\leq}} Z_{i,\leq}^a.
\]

**Right tests.** Given a segment \( \Delta = [a,b] \) of positive length with lower-feasible \( a \), we associate with this segment a right test – a function \( T^K_{\Delta,x}(\omega^K) \) taking values right and left, and risk \( \sigma_{\Delta,x} \geq 0 \) – as follows:

1. if \( b \) is upper-infeasible, \( T^K_{\Delta,x}(\cdot) \equiv \text{left} \) and \( \sigma_{\Delta,x} = 0 \).
2. if \( b \) is upper-feasible, the collections of “right sets” \( \{ A(Z_{i,\geq}^a) \}_{1 \leq i \leq I_{a,\geq}} \) and of “left sets” \( \{ A(Z_{j,\leq}^a) \}_{j \leq I_{a,\leq}} \) 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 \( O \), specifically,

   - for \( 1 \leq i \leq I_{a,\geq}, 1 \leq j \leq I_{a,\leq}, \) we build the detectors

   \[
   \phi^K_{ij\Delta}(\omega^K) = \sum_{i=1}^{K} \phi_{ij\Delta}(\omega^K),
   \]
with \( \phi_{ij\Delta}(\omega) \) given by

\[
\begin{align*}
(r_{ij\Delta}, s_{ij\Delta}) & \in \arg\min_{r \in \mathbb{Z}_+^n, s \in \mathbb{Z}_+^n} \ln \left( \int_{\Omega} \sqrt{p_{A(r)}(\omega)p_{A(s)}(\omega)\Pi(d\omega)} \right), \\
\phi_{ij\Delta}(\omega) &= \frac{1}{2} \ln \left( \frac{p_{A(r_{ij\Delta})}(\omega)}{p_{A(s_{ij\Delta})}(\omega)} \right).
\end{align*}
\]

We set

\[
\epsilon_{ij\Delta} = \int_{\Omega} \sqrt{p_{A(r_{ij\Delta})}(\omega)p_{A(s_{ij\Delta})}(\omega)\Pi(d\omega)}
\]

and build the \( I_{a, \geq} \times I_{a, \leq} \) matrix \( E_{\Delta, r} = \left[ \epsilon_{ij\Delta} \right]_{1 \leq i \leq I_{a, \geq}, 1 \leq j \leq I_{a, \leq}} \).

- We define \( \sigma_{\Delta, r} \) as the spectral norm of \( E_{\Delta, r} \). We compute the Perron-Frobenius eigenvector \( [g^{\Delta, r}; h^{\Delta, r}] \) of the matrix \( \left[ \frac{E_{\Delta, r}^T}{E_{\Delta, r}} \right] \), so that (see Section 2.5.1.2)

\[
g^{\Delta, r} > 0, \; h^{\Delta, r} > 0, \; \sigma_{\Delta, r} g^{\Delta, r} = E_{\Delta, r} h^{\Delta, r}, \; \sigma_{\Delta, r} h^{\Delta, r} = E_{\Delta, r}^T g^{\Delta, r}.
\]

Finally, we define the matrix-valued function

\[
D_{\Delta, r}(\omega^K) = [\phi_{ij\Delta}(\omega^K) - \ln(h^{\Delta, r}_j) + \ln(g^{\Delta, r}_i)]_{1 \leq i \leq I_{a, \geq}, 1 \leq j \leq I_{a, \leq}}.
\]

Test \( T_{\Delta, r}^K(\omega^K) \) takes value \textit{right} if the matrix \( D_{\Delta, r}(\omega^K) \) has a nonnegative row, and takes value \textit{left} otherwise.

Given \( \delta > 0, \kappa > 0 \), we call segment \( \Delta = [a, b] \) \( \delta \)-\textit{good (right)}, if \( a \) is lower-feasible, \( b > a \), and \( \sigma_{\Delta, r} \leq \delta \). We call a \( \delta \)-\textit{good (right) segment} \( \Delta = [a, b] \) \( \kappa \)-\textit{maximal}, if the segment \( [a, b - \kappa] \) is not \( \delta \)-\textit{good (right)}.

**Left 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 a \textit{left test} – a function \( T_{\Delta, l}^K(\omega^K) \) taking values \textit{right} and \textit{left}, and risk \( \sigma_{\Delta, l} \geq 0 \) – as follows:

1. if \( a \) is lower-infeasible, \( T_{\Delta, l}^K(\cdot) \equiv \text{right} \) and \( \sigma_{\Delta, l} = 0 \).
2. if \( a \) is lower-feasible, we set \( T_{\Delta, l}^K = T_{\Delta, r}^K, \sigma_{\Delta, l} = \sigma_{\Delta, r} \).

Given \( \delta > 0, \kappa > 0 \), we call segment \( \Delta = [a, b] \) \( \delta \)-\textit{good (left)}, if \( b \) is upper-feasible, \( b > a \), and \( \sigma_{\Delta, l} \leq \delta \). We call a \( \delta \)-\textit{good (left) segment} \( \Delta = [a, b] \) \( \kappa \)-\textit{maximal}, if the segment \( [a + \kappa, b] \) is not \( \delta \)-\textit{good (left)}.

**Explanation:** When \( a < b \) and \( a \) is lower-feasible, \( b \) is upper-feasible, so that the sets

\[
X^{a, \leq} = \{ x \in X : f(x) \leq a \}, \; X^{b, \geq} = \{ x \in X : f(x) \geq b \}
\]

are nonempty, the right and the left tests \( T_{\Delta, l}^K, T_{\Delta, r}^K \) 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 “location” of the distribution \( p_{A(x)} \) underlying observations – whether this location is \textit{left} (left hypothesis stating that \( x \in X \) and \( f(x) \leq a \), whence \( A(x) \in \bigcup_{1 \leq i \leq I_{b, \leq}} A(Z^{a, \leq}_i) \)), or \textit{right} (right hypothesis
stating that \( x \in X \) and \( f(x) \geq b \), whence \( A(x) \in \bigcup_{1 \leq i \leq I_{\geq}} A(Z_i^{b_{\geq}}) \). When \( a \) is lower-feasible and \( b \) is not upper-feasible, the right hypothesis is empty, and the left test associated with \([a, b]\), naturally, always accepts the left hypothesis; similarly, when \( a \) is lower-infeasible and \( b \) is upper-feasible, the right test associated with \([a, b]\) always accepts the right hypothesis.

A segment \([a, b]\) with \( a < b \) is \( \delta \)-good (left), if the corresponding to the segment right hypothesis is nonempty, and the left test \( T_{\Delta_0} \) associated with \([a, b]\) decides on the right and the left hypotheses with risk \( \leq \delta \), and similarly for \( \delta \)-good (right) segment \([a, b]\).

### 3.2.4 Building Bisection estimate

#### 3.2.4.1 Control parameters

The control parameters of the Bisection estimate are:

1. positive integer \( L \) – the maximum allowed number of bisection steps,
2. tolerances \( \delta \in (0, 1) \) and \( \kappa > 0 \).

#### 3.2.4.2 Bisection estimate: construction

The estimate of \( f(x) \) (\( x \) is the signal underlying our observations: \( \omega_i \sim p_{A(x)} \)) is given by the following recurrence run on the observation \( \omega^K = (\omega_1, ..., \omega_K) \) at our disposal:

1. **Initialization.** We find a valid upper bound \( b_0 \) on \( \max_{a \in X} f(u) \) and valid lower bound \( a_0 \) on \( \min_{a \in X} f(u) \) 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: \( f(x) \in \Delta_0 \).

2. **Bisection Step** \( \ell, 1 \leq \ell \leq L \). Given the 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 \( c_\ell \) is not upper-feasible, we set \( \Delta_\ell = [a_{\ell-1}, c_\ell] \) and pass to 2e, and if \( c_\ell \) is not lower-feasible, we set \( \Delta_\ell = [c_\ell, b_{\ell-1}] \) and pass to 2e.
   
   Note: When the rule requires to pass to 2e, the set \( \Delta_{\ell-1} \setminus \Delta_\ell \) does not intersect with \( f(X) \); in particular, in such case \( f(x) \in \Delta_\ell \) provided that \( f(x) \in \Delta_{\ell-1} \).

   b) When \( c_\ell \) is both upper- and lower-feasible, we check whether the segment \([c_\ell, b_{\ell-1}] \) is \( \delta \)-good (right). If it is not the case, we terminate and claim that \( f(x) \in \Delta := \Delta_{\ell-1} \), otherwise find \( v_\ell, c_\ell < v_\ell \leq b_{\ell-1} \), such that the segment \( \Delta_{\ell, \text{rg}} = [c_\ell, v_\ell] \) is \( \delta \)-good (right) \( \kappa \)-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 \( \delta \)-good (right) corresponds to the case when the current localizer is too small to allow for the “no-man’s land” wide enough to ensure low-risk decision on the left and the right hypotheses.

   To find \( v_\ell \), we check the candidates with \( v^k_\ell = b_{\ell-1} - k\kappa, k = 0, 1, ... \) until arriving for the first time at segment \([c_\ell, v^k_\ell] \) which is not \( \delta \)-good (right), and take as \( v_\ell \) the quantity \( v^{k-1}_\ell \) (because \( k \geq 1 \) the resulting value of \( v_\ell \) is well-defined and clearly meets the above requirements).

   c) Similarly, we check whether the segment \([a_{\ell-1}, c_\ell] \) is \( \delta \)-good (left). If it is not the case, we terminate and claim that \( f(x) \in \Delta := \Delta_{\ell-1} \), otherwise find
\[ u_\ell, \ a_{\ell-1} \leq u_\ell < c_\ell, \ \text{such that the segment } \Delta_{\ell, l} = [u_\ell, c_\ell] \text{ is } \delta\text{-good (left) } \kappa\text{-maximal.} \]

Note: The rules for building \( u_\ell \) are completely similar to those for \( v_\ell \).

d) We compute \( T^K_{\Delta_{\ell, r_g r_r}(\bar{\omega}^K)} \) and \( T^K_{\Delta_{\ell, l_f l_r}(\bar{\omega}^K)} \). If \( T^K_{\Delta_{\ell, r_g r_r}(\bar{\omega}^K)} = T^K_{\Delta_{\ell, l_f l_r}(\bar{\omega}^K)} \) ("consensus"), we set \( \Delta_\ell = [a_\ell, b_\ell] = \left\{ \begin{array}{ll} [c_\ell, b_\ell - 1] & \text{if } T^K_{\Delta_{\ell, r_g r_r}(\bar{\omega}^K)} = \text{right,} \\
[a_\ell - 1, c_\ell] & \text{if } T^K_{\Delta_{\ell, l_f l_r}(\bar{\omega}^K)} = \text{left} \end{array} \right. \) (3.16)

and pass to 2e. Otherwise ("disagreement") we terminate and claim that \( f(x) \in \bar{\Delta} = [u_\ell, v_\ell] \).

e) We pass to step \( \ell + 1 \) when \( \ell < L \), otherwise we terminate with the claim that \( f(x) \in \Delta := \Delta_L \).

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 3.4.** 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

\[ L, \ \delta = \frac{\epsilon}{2L}, \ \kappa \]  

is \((1 - \epsilon)\)-reliable: for every \( x \in X \), the \( p_{\mathcal{A}(x)} \)-probability of the event

\( f(x) \in \bar{\Delta} \)

(\( \bar{\Delta} \) is the Bisection output as defined above) is at least \( 1 - \epsilon \).

(ii) [near-optimality] Let \( \rho > 0 \) and positive integer \( K \) be such that in the nature there exists a \((\rho, \epsilon)\)-reliable estimate \( \hat{f}(\cdot) \) of \( f(x), \ x \in X := \bigcup_{i \leq I} X_i \), via stationary \( K\)-repeated observation \( \omega^K \) with \( \omega_k \sim p_{\mathcal{A}(x)}, 1 \leq k \leq K \). Given \( \hat{\rho} > 2\rho \), the Bisection estimate utilizing stationary \( K\)-repeated observations, with

\[ K \geq \frac{2 \ln(2LNÎ/\epsilon)}{\ln \left( \frac{1}{4\epsilon(1-\epsilon)} \right) K}, \]  

the control parameters of the estimate being

\[ L = \left\lfloor \log_2 \left( \frac{b_0 - a_0}{2\hat{\rho}} \right) \right\rfloor, \ \delta = \frac{\epsilon}{2L}, \ \kappa = \hat{\rho} - 2\rho, \]  

is \((\hat{\rho}, \epsilon)\)-reliable. Not that \( K \) is only "slightly larger" than \( K \).

For proof, see Section 3.6.3.

Note that the running time \( K \) of Bisection estimate as given by (3.18) is just
by (at most) logarithmic in \( N, I, L \) and \( 1/\epsilon \) factor larger than \( \overline{K} \); note also that \( L \) is just logarithmic in \( 1/\hat{\rho} \). Assume, e.g., that for some \( \gamma > 0 \) “in the nature” there exist \((\epsilon', \epsilon)\)-reliable estimates, parameterized by \( \epsilon \in (0, 1/2) \), utilizing \( K = \overline{K}(\epsilon) \) observations. Then Bisection with the volume of observation and control parameters given by (3.18) and (3.19), where \( \hat{\rho} = 3\rho = 3\epsilon' \) and \( \overline{K} = \overline{K}(\epsilon) \), is \((3\epsilon', \epsilon)\)-reliable and requires \( K = K(\epsilon) \)-repeated observations with \( \lim_{\epsilon \rightarrow +0} K(\epsilon)/\overline{K}(\epsilon) \leq 2 \).

### 3.2.6 Illustration

To illustrate bisection-based estimation of \( N \)-convex function, consider the following situation. There are \( M \) devices (“receivers”) recording a signal \( u \) known to belong to a given convex compact and nonempty set \( U \subset \mathbb{R}^n \); the output of the \( i \)-th receiver is the vector

\[
y_i = A_i u + \sigma \xi \in \mathbb{R}^m \quad [\xi \sim \mathcal{N}(0, I_m)]
\]

where \( A_i \) are given \( m \times n \) matrices (you may think of \( M \) allowed positions for a single receiver, and of \( y_i \) as on the output of the receiver when the latter is in position \( i \)). Our observation \( \omega \) is one of the vectors \( y_i \), \( 1 \leq i \leq M \) with unknown to us index \( i \) (“we observe a noisy record of a signal, but do not know the position in which this record was taken”). Given \( \omega \), we want to recover a given linear function

\[
g(x) = e^T u \quad \text{of the signal.}
\]

The problem can be modeled as follows. Consider \( M \) sets

\[
X_i = \{ x = [x^1; \ldots; x^M] \in \mathbb{R}^{Mn} = \bigcup_{i=1}^{M} \mathbb{R}^n \times \cdots \times \mathbb{R}^n : x^j = 0, j \neq i; x^i \in U \}
\]

along with the linear mapping

\[
A[x^1; \ldots; x^M] = \sum_{i=1}^{M} A_i x^i : \mathbb{R}^{Mn} \rightarrow \mathbb{R}^m
\]

and linear function

\[
f([x^1; \ldots; x^M]) = e^T \sum_{i} x^i : \mathbb{R}^{Mn} \rightarrow \mathbb{R}.
\]

Let \( \mathcal{X} \) be a convex compact set in \( \mathbb{R}^{Mn} \) containing all the sets \( X_i \), \( 1 \leq i \leq m \). Observe that the problem we are interested in is nothing but the problem of recovering \( f(x) \) via observation

\[
\omega = Ax + \sigma \xi, \quad \xi \sim \mathcal{N}(0, I_m), \quad (3.20)
\]

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 developed in this section.

**Numerical illustration.** In the numerical experiments to be reported, we use \( n = 128, m = 64 \) and \( M = 2 \). The data is generated as follows:

- The set \( U \subset \mathbb{R}^{128} \) of candidate signals is comprised of restrictions onto the
Table 3.1: Data of 10 Bisection experiments, \( \sigma = 0.01 \). In the table, “error bound” is the half-length of the final localizer, which is an 0.99-reliable upper bound on the estimation error, “actual error” is the actual estimation error.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>min</th>
<th>median</th>
<th>mean</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>error bound</td>
<td>0.008</td>
<td>0.015</td>
<td>0.014</td>
<td>0.015</td>
</tr>
<tr>
<td>actual error</td>
<td>0.001</td>
<td>0.002</td>
<td>0.002</td>
<td>0.005</td>
</tr>
<tr>
<td># of Bisection steps</td>
<td>5</td>
<td>7.00</td>
<td>6.60</td>
<td>8</td>
</tr>
</tbody>
</table>

Equidistant \((n = 128)\)-point grid in \([0, 1]\) of twice differentiable functions \(h(t)\) of continuous argument \(t \in [0, 1]\) satisfying the relations \(|h(0)| \leq 1\), \(|h'(0)| \leq 1\), \(|h''(t)| \leq 1\), \(0 \leq t \leq 1\). For the discretized signal \(u = [h(0); h(1/n); \ldots; h(1-1/n)]\) this translates into the system of convex constraints

\[ |u_1| \leq 1, \ n|u_2 - u_1| \leq 1, \ n^2|u_{i+1} - 2u_i + u_{i-1}| \leq 1, \ 2 \leq i \leq n - 1. \]

- We look to estimate the discretized counterpart of the integral \(\int_0^1 h(t)dt\), specifically, the quantity \(e^Tu = \alpha \sum_i u_i\). The normalizing constant \(\alpha\) is selected to ensure \(\max_{u \in U} e^Tu = 1\), \(\min_{u \in U} e^Tu = -1\), allowing to run Bisection over \(\Delta_0 = [-1; 1]\).

- We generate \(A_1\) as \((m = 64) \times (n = 128)\) matrix with singular values \(\sigma_i = \theta^{-i-1}, 1 \leq i \leq m, \) with \(\theta\) selected from the requirement \(\sigma_m = 0.1\). The system of left singular vectors of \(A_1\) is obtained from the system of basic orths in \(\mathbb{R}^n\) by random rotation.
Matrix \(A_2\) was selected as \(A_2 = A_1S\), where \(S\) is a symmetry w.r.t. the axis \(e\), that is,

\[ Se = e \& Sh = -h \text{ whenever } h \text{ is orthogonal to } e. \]  

(3.21)

Signals \(u\) underlying the observations are selected at random in \(U\).

- The reliability \(1 - \epsilon\) of the estimate is set to 0.99, while the maximal allowed number \(L\) of Bisection steps set to 8. We use single observation \((3.20)\) (i.e., used \(K = 1\) in our general scheme) with \(\sigma = 0.01\).

The results of our experiments are presented in Table 3.1. Observe that in the considered problem 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 \(A_1u = A_2u'\) and \(e^Tu \neq e^Tu'\). Since we do not know which of the matrices, \(A_1\) or \(A_2\), underlies the observation and \(A_1u = A_2u'\), there is no way to distinguish between the two cases we have described, implying that the quantity

\[ \rho = \max_{u, u' \in U} \left\{ \frac{1}{2}[e^T(u - u')] : A_1u = A_2u' \right\} \]

(3.22)
is a lower bound on the worst-case, over signals from \(U\), error of a reliable recovery of \(e^Tu\), independently of how small is the noise. In the reported experiments, we used \(A_2 = A_1S\) with \(S\) linked to \(e\), see (3.21); with this selection of \(S, e\) and \(A_2, \) and invertible \(A_1\), the lower bound \(\rho\) would be trivial – just zero. Note that the selected \(A_1\) is not invertible, resulting in a positive \(\rho\). However, computation shows that with our data, this positive \(\rho\) 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 $A_2 = A_1S$ exactly as they are, but replace the linear form to be estimated with $e^T u$, $e$ being randomly selected.\footnote{In the experiments to be reported, $e$ is selected as follows: we start with a random unit vector drawn from the uniform distribution on the unit sphere in $\mathbb{R}^n$ and then normalize it to have $\max_{u \in U} e^T u - \min_{u \in U} e^T u = 2$.} The corresponding results are presented in Table 3.2. The data in the top part of the table match “difficult” signals $u$ – those participating in forming the lower bound (3.22) on the recovery error, while the data in the bottom part of the table correspond to randomly selected signals.\footnote{Precisely, to generate a signal $u$, we draw 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 $U$.} Observe that when estimating 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 a reasonably good agreement with the lower bound $\rho$, illustrating the basic property of nearly optimal estimates: the guaranteed performance of an estimate can be bad or good, but it is always nearly as good as is possible under the circumstances. As about actual estimation errors, they in some experiments are significantly less than the error bounds, especially when random signals were used.

### Table 3.2: Results of experiments with randomly selected linear form, $\sigma = 0.01$.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>min</th>
<th>median</th>
<th>mean</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>error bound</td>
<td>0.057</td>
<td>0.457</td>
<td>0.441</td>
<td>1.000</td>
</tr>
<tr>
<td>actual error</td>
<td>0.001</td>
<td>0.297</td>
<td>0.350</td>
<td>1.000</td>
</tr>
<tr>
<td># of Bisection steps</td>
<td>1</td>
<td>1.00</td>
<td>2.20</td>
<td>5</td>
</tr>
</tbody>
</table>

“Difficult” signals, data over 10 experiments

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>0.022</th>
<th>0.028</th>
<th>0.154</th>
<th>0.176</th>
<th>0.213</th>
<th>0.248</th>
<th>0.250</th>
<th>0.500</th>
<th>0.605</th>
<th>0.924</th>
</tr>
</thead>
<tbody>
<tr>
<td>error</td>
<td>0.057</td>
<td>0.063</td>
<td>0.219</td>
<td>0.239</td>
<td>0.406</td>
<td>0.516</td>
<td>0.625</td>
<td>0.773</td>
<td>1.000</td>
<td></td>
</tr>
</tbody>
</table>

Error bound vs. $\rho$, experiments sorted according to the values of $\rho$

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>min</th>
<th>median</th>
<th>mean</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>error bound</td>
<td>0.016</td>
<td>0.274</td>
<td>0.348</td>
<td>1.000</td>
</tr>
<tr>
<td>actual error</td>
<td>0.005</td>
<td>0.066</td>
<td>0.127</td>
<td>0.556</td>
</tr>
<tr>
<td># of Bisection steps</td>
<td>1</td>
<td>2.00</td>
<td>2.80</td>
<td>7</td>
</tr>
</tbody>
</table>

Random signals, data over 10 experiments

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>0.010</th>
<th>0.085</th>
<th>0.177</th>
<th>0.243</th>
<th>0.294</th>
<th>0.334</th>
<th>0.337</th>
<th>0.554</th>
<th>0.630</th>
<th>0.762</th>
</tr>
</thead>
<tbody>
<tr>
<td>error</td>
<td>0.016</td>
<td>0.182</td>
<td>0.376</td>
<td>0.438</td>
<td>0.602</td>
<td>0.029</td>
<td>0.031</td>
<td>0.688</td>
<td>0.125</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Error bound vs. $\rho$, experiments sorted according to the values of $\rho$

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 of Section 3.2.2, we start with computing
the range $\Delta$ of function $f$ on the set $X = \bigcup_{i \leq I} X_i$, that is, we compute the quantities

$$f = \min_{x \in X} f(x), \quad \overline{f} = \max_{x \in X} f(x)$$

and set $\Delta = [f, \overline{f}]$. We assume that this segment is not a singleton, otherwise estimating $f$ is trivial. Let $L \in \mathbb{Z}_+$ and let $\delta_L = (\overline{f} - f)/L$ be the desired estimation accuracy. We split $\Delta$ into $L$ segments $\Delta_\ell$ of equal length $\delta_L$ and consider the sets

$$X_{i\ell} = \{x \in X_i : f(x) \in \Delta_\ell\}, 1 \leq i \leq I, 1 \leq \ell \leq L.$$ 

Since $f$ is $N$-convex, each set $X_{i\ell}$ is a union of $M_{i\ell} \leq N^2$ convex compact sets $X_{i\ell j}, 1 \leq j \leq M_{i\ell}$. Thus, we have 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 set in Section 3.2.2 can be reformulated as follows:

For some unknown $x$ known to belong to $X = \bigcup_{s=1}^M Y_s$, we have at our disposal observation $\omega^K = (\omega_1, ..., \omega_K)$ with i.i.d. $\omega_1 \sim p_{A(x)}(\cdot)$; we aim at estimating the quantity $f(x)$ from this observation.

The sets $Y_s$ give rise to $M$ hypotheses $H_1, ..., H_M$ on the distribution of the observations $\omega_1, 1 \leq t \leq K$; according to $H_s$, $\omega_1 \sim p_{A(x)}(\cdot)$ with some $x \in Y_s$.

Let us define a closeness $C$ 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 $C$-close, if the segments $\Delta_{\ell(s)}$ and $\Delta_{\ell(s')}^*$ intersect. Observe that when $H_s$ and $H_{s'}$ are not $C$-close, the convex compact sets $Y_s$ and $Y_{s'}$ 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 $C$, see Section 2.5.2. Assuming that whenever $H_s$ and $H_{s'}$ are not $C$-close, the risks $\epsilon_{s,s'}$ defined in Section 2.5.2.2 are $< 1$, the way we have observed that $H_s$ is $K$-repeated observation $\omega^K$ allows to decide $(1-\epsilon)$-reliably on $H_1, ..., H_M$ up to closeness $C$. As applied to $\omega^K$, the corresponding test $T^K$ 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 $\hat{f}(\omega^K)$ of $f(x), x \in X = \bigcup_{s \leq M} Y_s$ being the signal underlying our observation, as follows:

- when $S = \emptyset$ the estimate is, say $(\overline{f} + f)/2$;
- 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 that if the signal $x$ underlying our stationary $K$-repeated observation $\omega^K$ belongs to some $Y_{s_s}$, so that the hypothesis $H_{s_s}$ is true, and the outcome $S$ of $T^K$ contains $s_s$ and is such that for all $s \in S$ $H_s$ and $H_{s_s}$ are $C$-close to each other, we have $|f(x) - \hat{f}(\omega^K)| \leq \delta_L$. Note that since the $C$-risk of

\footnote{In 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. Because for $s, s’$, $Y_s$ and $Y_{s’}$ do not intersect this definitely is the case when $A(\cdot)$ is an embedding.}
$\mathcal{T}^K$ is $\leq \epsilon$, the $p_{A(x)}$-probability to get such a “good” outcome, and thus to get $|f(x) - \hat{f}(x^K)| \leq \delta_L$, 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 $(i_*, j_*)$:

$$r_{i_*, j_*} = \frac{V_{j_*} - V_{i_*}}{I_{i_*, j_*}}.$$  

The observation noises are assumed to be $\mathcal{N}(0, \sigma^2)$ and independent across the measurements.

In our experiment, we work with the data as follows:

- Currents are measured in all arcs except for $a$, $b$
- Voltages are measured at all nodes except for $c$
- We want to recover resistance of arc $b$
- $X$: \begin{align*}
    & \text{conservation of current, except for the input/output nodes} \\
    & \text{zero voltage at input node, nonnegative currents} \\
    & \text{current in arc } b \text{ at least 1, total of currents at most 33} \\
    & \text{Ohm law, resistances of arcs between 1 and 10}
\end{align*}

We are in the situation of $N = 1$ and $I = 1$, 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
\((V_{j*} - V_{i*}, I_{i,j*})\) are the “stripes” shown below:

The range of the unknown resistance turns out to be \(\Delta = [1, 10]\).

We set \(\epsilon = 0.01\), and 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 look for the largest observation noise \(\sigma\) allowing to achieve the desired recovery with a single observation. The results for \(L = 8, 16, 32\) are as follows

<table>
<thead>
<tr>
<th>(L)</th>
<th>8</th>
<th>16</th>
<th>32</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\delta_{L})</td>
<td>(9/8 \approx 1.13)</td>
<td>(9/16 \approx 0.56)</td>
<td>(9/32 \approx 0.28)</td>
</tr>
<tr>
<td>(\sigma)</td>
<td>0.024</td>
<td>0.010</td>
<td>0.005</td>
</tr>
<tr>
<td>(\sigma_{\text{opt}}/\sigma \leq)</td>
<td>1.31</td>
<td>1.31</td>
<td>1.33</td>
</tr>
<tr>
<td>(\sigma)</td>
<td>0.031</td>
<td>0.013</td>
<td>0.006</td>
</tr>
<tr>
<td>(\sigma_{\text{opt}}/\sigma \leq)</td>
<td>1.01</td>
<td>1.06</td>
<td>1.08</td>
</tr>
</tbody>
</table>

In the above table:

- \(\sigma_{\text{opt}}\) is the largest \(\sigma\) for which “in the nature” there exists a test deciding on \(H_1, ..., H_L\) with \(C\)-risk \(\leq 0.01\);
- Underlined data: Risks \(\epsilon_{ss'}\) of pairwise tests are bounded via risks of optimal detectors, \(C\)-risk of \(T\) is bounded by

\[
\left\| \epsilon_{ss'} \chi_{ss'} \right\|_{L,s,s'=1} \leq \chi_{ss'} = \begin{cases} 
1, & (s, s') \notin C \\
0, & (s, s') \in C 
\end{cases}
\]

see Proposition 2.29;
- “Slanted” data: Risks \(\epsilon_{ss'}\) of pairwise tests are bounded via the error function, \(C\)-risk of \(T\) is bounded by

\[
\max_s \sum_{s' : (s, s') \notin C} \epsilon_{ss'}
\]

(it is immediately seen that in the case of Gaussian o.s., this indeed is a legitimate risk bound).
Figure 3.3: A circuit with 9 nodes and 16 arcs.  $a$ is the arc of interest; the current is measured in arcs $b$; external current and voltage are measured at the input node $c$.

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 described Section 2.9 to yield, under favorable circumstances, estimates of quadratic and quadratic fractional functions. We are about to consider an instructive example of this type. Figure 3.3 represent a DC circuit. We have access to repeated noisy measurements of currents in some arcs and voltages at some 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.” The a priori information is as follows:

- the (unknown) arc resistances 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 $c$.
  - the voltages and the currents are linked by Ohm’s law: for every (inner) arc $\gamma$, we have
    \[ I_\gamma r_\gamma = V_{j(\gamma)} - V_{i(\gamma)} \]
    where $I_\gamma$ is the current in the arc, $r_\gamma$ is the arc’s resistance, $V_s$ is the voltage at node $s$, and $i(\gamma)$, $j(\gamma)$ are the initial and the terminal nodes linked by arc $\gamma$;
- 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 $\theta$ from a given range $[\sigma, \sigma]$. 
Processing the problem. We specify the “signal” underlying our observation as collection $u$ of the voltages at 9 nodes and currents $I_s$ in 16 (inner) arcs $\gamma$ of the circuit, augmented by the external current $I_o$ at the input node (so that $-I_o$ is the external current at the ground node). Thus, our single-time observation is

$$\zeta = Au + \theta \xi,$$

where $A$ extracts from $u$ four entries (currents in two arcs $b$ and external current and voltage at the input node $c$), $\xi \sim \mathcal{N}(0, I_s)$, and $\theta \in [\mathcal{Z}, \mathcal{S}]$. Our a priori information on $u$ states that $u$ belongs to the compact set $U$ given by the quadratic constraints, namely, as follows:

$$U = \left\{ u = \{I_s, I_o, V_i\} : \begin{array}{l}
I_s^2 \leq 1, V_i^2 \leq 1 \forall \gamma, i;\ u^T J^TJu \leq 0 \\
I_s[V_{j(\gamma)} - V_{i(\gamma)}]^2/R - I_o[V_{j(\gamma)} - V_{i(\gamma)}] \leq 0 \\
I_s[V_{j(\gamma)} - V_{i(\gamma)}] - [V_{j(\gamma)} - V_{i(\gamma)}]^2/r \leq 0 \\
\gamma I_s^2 - I_o[V_{j(\gamma)} - V_{i(\gamma)}] \leq 0 \\
I_o[V_{j(\gamma)} - V_{i(\gamma)}] - RI_o^2 \leq 0
\end{array} \forall \gamma \right\} \quad (a)$$

$$U = \left\{ u = \{I_s, I_o, V_i\} : \begin{array}{l}
I_s^2 \leq 1, V_i^2 \leq 1 \forall \gamma, i;\ u^T J^TJu \leq 0 \\
I_s[V_{j(\gamma)} - V_{i(\gamma)}]^2/R - I_o[V_{j(\gamma)} - V_{i(\gamma)}] \leq 0 \\
I_s[V_{j(\gamma)} - V_{i(\gamma)}] - [V_{j(\gamma)} - V_{i(\gamma)}]^2/r \leq 0 \\
\gamma I_s^2 - I_o[V_{j(\gamma)} - V_{i(\gamma)}] \leq 0 \\
I_o[V_{j(\gamma)} - V_{i(\gamma)}] - RI_o^2 \leq 0
\end{array} \forall \gamma \right\} \quad (b)$$

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 exact resistances but only their range $[r, R]$. Note that groups $(a)$, $(b)$ of constraints in (3.24) are “logical consequences” of each other, and thus one of groups seems to be redundant. However, on closer inspection, valid on $U$ quadratic inequalities do not tighten the outer approximation $Z$ of $Z[U]$ and thus are redundant in our context only when these inequalities can be obtained from the inequalities we do include into the description of $Z$ “in a linear fashion” – by taking weighted sums with nonnegative coefficients. This is not how $(b)$ is obtained from $(a)$. As a result, to get a smaller $Z$, it makes sense to keep both $(a)$ and $(b)$.

The dissipated power we are interested to estimate is the quadratic function

$$f(u) = I_{s*}[V_{j*} - V_{i*}] = [u; 1]^T G[u; 1]$$

where $\gamma_s = (i_s, j_s)$ is the arc of interest, and $G \in \mathcal{S}^{n+1}$, $n = \text{dim } u$, is a properly built matrix.

In order to build an estimate, we “lift quadratically” the observations:

$$\zeta \mapsto \omega = (\zeta, \zeta^T)$$

and pass from the domain $U$ of actual signals to the outer approximation $Z$ of the quadratic lifting of $U$:

$$Z := \left\{ Z \in \mathcal{S}^{n+1} : Z \geq 0, Z_{n+1,n+1} = 1, \text{Tr}(Q_s Z) \leq c_s, 1 \leq s \leq S \right\} \supset \{ [u; 1][u; 1]^T : u \in \mathcal{V} \}.$$

Here the matrix $Q_s \in \mathcal{S}^{n+1}$ represents the left hand side $F_s(u)$ of the $s$-th quadratic constraint in the description (3.24) of $U$: $F_s(u) \equiv [u; 1]^T Q_s [u; 1]$, and $c_s$ is the right hand side of the $s$-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 $f$ 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...
\[ f = \max_{Z \in \mathbb{Z}} \text{Tr}(GZ) \]

2. Given a positive integer \( L \), we split the range \([f, \bar{f}]\) into \( L \) segments \( \Delta_\ell = [a_{\ell-1}, a_\ell] \) of equal length \( \delta_L = (\bar{f} - f) / L \) and define convex compact sets \( Z_\ell = \{ Z \in \mathbb{Z} : a_{\ell-1} \leq \text{Tr}(GZ) \leq a_\ell \} \), 1 \( \leq \ell \leq L \), so that \( u \in U, f(u) \in \Delta_\ell \Rightarrow [u; 1]|[u; 1]^T \in Z_\ell, 1 \leq \ell \leq L \);

3. We specify \( L \) quadratically constrained hypotheses \( H_1, ..., H_L \) on the distribution of observation (3.23), with \( H_\ell \) stating that \( \zeta \sim \mathcal{N}(Au, \theta^2 I_4) \) with some \( u \in U \) satisfying \( f(u) \in \Delta_\ell \) (so that \([u; 1]|[u; 1]^T \in Z_\ell\)), and \( \theta \) belongs to the above segment \([\sigma, \bar{\sigma}]\).

We equip our hypotheses with a closeness relation \( C \), specifically, we consider \( H_\ell \) and \( H_{\ell'} \) \( C \)-close if and only if the segments \( \Delta_\ell \) and \( \Delta_{\ell'} \) intersect.

4. We use Propositions 2.43.ii, 2.40 to build quadratic in \( \zeta \) detectors \( \phi_{\ell\ell'} \) for the families of distributions obeying \( H_\ell \) and \( H_{\ell'} \), respectively, along with upper bounds \( \epsilon_{\ell\ell'} \) on the risks of these detectors. Finally, we use the machinery from Section 2.5.2 to find the smallest \( K \) and a test \( T_c^K \), based on stationary \( K \)-repeated version of observation (3.23), able to decide on \( H_1, ..., H_L \) with \( C \)-risk \( \leq \epsilon \), where \( \epsilon \in (0, 1) \) is a given tolerance.

Finally, given stationary \( K \)-repeated observation (3.23), we apply to it test \( T_c^K \), look at the hypotheses, if any, accepted by the test, and build the union \( \Delta \) of the corresponding segments \( \Delta_\ell \). If \( \Delta = \emptyset \), we estimate \( f(u) \) by the midpoint of the power range \([f, \bar{f}]\), otherwise the estimate is the mean of the largest and the smallest points in \( \Delta \). It is easily seen that for this estimate, the probability for the estimation error to be \( \geq \delta_\ell \) is \( \leq \epsilon \).

**Numerical results** We present here correspond to the circuit presented in Figure 3.3. We set \( \bar{\sigma} = 0.01, \sigma = \bar{\sigma}/\sqrt{2}, [r, R] = [1, 2], \epsilon = 0.01, \) and \( L = 8 \). The simulation setting is as follows: the computed range \([f, \bar{f}]\) of the dissipated power is \([0, 0.821]\), so that the estimate built recovers the dissipated power within accuracy 0.103 and reliability 0.99. The resulting value of \( K \) is \( K = 95 \).

In all 500 simulation runs, the actual recovery error was less than the bound 0.103, and the average error was as small as 0.041.

### 3.3 Estimating Linear Forms Beyond Simple Observation Schemes

We are about to show 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 do not estimate general \( N \)-convex functions nor 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.\footnote{\textsuperscript{8}} As a compensation, we are not limited to simple observation schemes anymore – our new approach is in fact a natural extension of the approach developed in Section 3.1 beyond simple o.s.’s.

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

- a nonempty convex compact set $\mathcal{X} \subset \mathcal{E}_X$,
- an affine mapping $x \mapsto A(x) : \mathcal{X} \rightarrow \mathcal{E}_M$ such that $A(\mathcal{X}) \subset \mathcal{M}$,
- a vector $g \in \mathcal{E}_X$ and a constant $c$ specifying the linear form $G(x) = \langle g, x \rangle + c$:
- $\mathcal{E}_X \rightarrow \mathbb{R}$,
- a tolerance $\epsilon \in (0, 1)$.

These data specify, in particular, the family

$$\mathcal{P} = S[\mathcal{H}, \mathcal{M}, \Phi]$$

of probability distributions on $\Omega = \mathcal{E}_H$, see Section 2.8.1.1. Given random observation

$$\omega \sim P(\cdot)$$

(3.25)

where $P \in \mathcal{P}$ is such that

$$\forall h \in \mathcal{H} : \ln \left( \int_{\mathcal{E}_H} e^{\langle h, \omega \rangle} P(d\omega) \right) \leq \Phi(h; A(x))$$

(3.26)

for some $x \in \mathcal{X}$ (that is, $A(x)$ is a parameter, as defined in Section 2.8.1.1, of distribution $P$), we want to recover the quantity $G(x)$.

**$\epsilon$-risk.** Given $\rho > 0$, we call an estimate $\hat{g}(\cdot) : \mathcal{E}_H \rightarrow \mathbb{R}$ $(\rho, \epsilon, v(\cdot))$-accurate, if for all pairs $x \in \mathcal{X}$, $P \in \mathcal{P}$ satisfying (3.26) it holds

$$\text{Prob} \{ |\hat{g}(\omega) - G(x)| > \rho + v(x) \} \leq \epsilon.$$ 

If $\rho_*$ is the infimum of those $\rho$ for which estimate $\hat{g}$ is $(\rho, \epsilon, v(\cdot))$-accurate, then clearly $\hat{g}$ is $(\rho_*, \epsilon, v(\cdot))$-accurate; we shall call $\rho_*$ the $\epsilon$-risk of the estimate $\hat{g}$ taken

\footnote{\textsuperscript{8}}The latter is just for the sake of simplicity, to not overload the presentation to follow. An interested reader will certainly be able to reproduce the corresponding construction of Section 3.1 in the situation of this section.
\footnote{\textsuperscript{9}}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.}
w.r.t. the data $G(\cdot), \mathcal{X}, v(\cdot)$ and $(\mathcal{A}, \mathcal{H}, \mathcal{M}, \Phi)$:

$$\text{Risk}_r(\tilde{g}(\cdot)|G, \mathcal{X}, v, \mathcal{A}, \mathcal{H}, \mathcal{M}, \Phi) = \min \left\{ \rho : \text{Prob}_{\omega \sim P}\{\omega : |\tilde{g}(\omega) - G(x)| > \rho + v(x)\} \leq \epsilon \right\}$$

(3.27)

When $G, \mathcal{X}, v, \mathcal{A}, \mathcal{H}, \mathcal{M}$ and $\Phi$ are clear from the context, we shorten $\text{Risk}_r(\tilde{g}(\cdot)|G, \mathcal{X}, v, \mathcal{A}, \mathcal{H}, \mathcal{M}, \Phi)$ to $\text{Risk}_r(\tilde{g}(\cdot))$.

Given the data listed in the beginning of this section, we are about to build, in a computationally efficient fashion, an affine estimate $\tilde{g}(\omega) = \langle h_*, \omega \rangle + \kappa$ along with $\rho_*$ such that the estimate is $(\rho_*, \epsilon, v(\cdot))$-accurate.

### 3.3.2 Construction and 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 $\mathcal{E}_H \times \mathbb{R}_+$, and let

(a) $\Psi_+(h, \alpha) = \sup_{x \in \mathcal{X}} [\alpha \Phi(h/\alpha, A(x)) - G(x) - v(x)] : \mathcal{H}^+ \to \mathbb{R},$

(b) $\Psi_-(h, \beta) = \sup_{x \in \mathcal{X}} [\beta \Phi(-h/\beta, A(x)) + G(x) - v(x)] : \mathcal{H}^+ \to \mathbb{R},$ (3.28)

so that $\Psi_{\pm}$ are convex real-valued functions on $\mathcal{H}^+$ (recall that $\Phi$ is convex-concave and continuous on $\mathcal{H} \times \mathcal{M}$, while $A(\mathcal{X})$ is a compact subset of $\mathcal{M}$).

Our starting point is quite simple:

**Proposition 3.5.** Given $\epsilon \in (0, 1)$, let $\bar{h}, \bar{\alpha}, \bar{\beta}, \bar{\kappa}, \bar{\rho}$ be a feasible solution to the system of convex constraints

$$\begin{align*}
(a_1) & (h, \alpha) \in \mathcal{H}^+ \\
(a_2) & (h, \beta) \in \mathcal{H}^+ \\
(b_1) & \alpha \ln(\epsilon/2) \geq \Psi_+(h, \alpha) - \rho + \kappa \\
(b_2) & \beta \ln(\epsilon/2) \geq \Psi_-(h, \beta) - \rho - \kappa
\end{align*}$$

(3.29)

in variables $h, \alpha, \beta, \rho, \kappa$. Setting

$$\tilde{g}(\omega) = \langle \bar{h}, \omega \rangle + \bar{\kappa},$$

we obtain an estimate with $\epsilon$-risk at most $\bar{\rho}$.

**Proof.** Let $\epsilon \in (0, 1), \bar{h}, \bar{\alpha}, \bar{\beta}, \bar{\kappa}, \bar{\rho}$ satisfy the premise of the proposition, and let $x \in \mathcal{X}, P$ satisfy (3.26). We have

$$\text{Prob}_{\omega \sim P}\{\tilde{g}(\omega) > G(x) + \bar{\rho} + v(x)\} = \text{Prob}_{\omega \sim P}\left\{\frac{\langle \bar{h}, \omega \rangle}{\alpha} > \frac{G(x) + \bar{\rho} - \kappa + v(x)}{\alpha}\right\}$$

$$\Rightarrow \text{Prob}_{\omega \sim P}\{\tilde{g}(\omega) > G(x) + \bar{\rho} + v(x)\} \leq \left[\int e^{\langle \bar{h}, \omega \rangle/\bar{\alpha}} P(d\omega) e^{G(x) + \bar{\rho} - \kappa + v(x)} \right]$$

$$\leq e^{\Phi(\bar{h}/\alpha, A(x))} e^{-\frac{G(x) + \bar{\rho} - \kappa + v(x)}{\alpha}}.$$


As a result,
\[
\bar{\alpha} \ln \left( \Pr_{\omega \sim P} \{ \hat{g}(\omega) > G(x) + \bar{\rho} + v(x) \} \right)
\leq \bar{\alpha} \Phi \left( \frac{\langle h, \omega \rangle}{\beta}, A(x) \right) - G(x) - \bar{\rho} + \bar{z} - v(x)
\leq \Psi_+(\bar{h}, \bar{\alpha}) - \bar{\rho} + \bar{z} \quad \text{[by definition of } \Psi_+ \text{ and due to } x \in X]\]
\[
\leq \bar{\alpha} \ln(\epsilon/2) \quad \text{[by (3.29.b1)]}
\]
so that
\[
\Pr_{\omega \sim P} \{ \hat{g}(\omega) > G(x) + \bar{\rho} + v(x) \} \leq \epsilon/2,
\]

Similarly
\[
\Pr_{\omega \sim P} \{ \hat{g}(\omega) < G(x) - \bar{\rho} - v(x) \} = \Pr_{\omega \sim P} \left\{ \frac{-\langle h, \omega \rangle}{\beta} > \frac{-G(x) + \bar{\rho} + \bar{z} + v(x)}{\beta} \right\}
\Rightarrow \Pr_{\omega \sim P} \{ \hat{g}(\omega) < G(x) - \bar{\rho} - v(x) \} \leq \int e^{-\langle h, \omega \rangle/\beta} P(\omega) e^{-G(x) + \bar{\rho} + \bar{z} + v(x)}
\leq e^{\Phi_-(\bar{h}, \bar{\beta}, A(x))} e^{G(x) - \bar{\rho} - \bar{z} - v(x)}.
\]
Thus
\[
\bar{\beta} \ln \left( \Pr_{\omega \sim P} \{ \hat{g}(\omega) < G(x) - \bar{\rho} - v(x) \} \right)
\leq \bar{\beta} \Phi \left( \frac{-\langle h, \omega \rangle}{\beta}, A(x) \right) + G(x) - \bar{\rho} - \bar{z} - v(x)
\leq \Psi_-(\bar{h}, \bar{\beta}) - \bar{\rho} - \bar{z} \quad \text{[by definition of } \Psi_- \text{ and due to } x \in X]\]
\[
\leq \bar{\beta} \ln(\epsilon/2) \quad \text{[by (3.29.b2)]}
\]
and
\[
\Pr_{\omega \sim P} \{ \hat{g}(\omega) < G(x) - \bar{\rho} - v(x) \} \leq \epsilon/2. \quad \Box
\]

**Corollary 3.6.** In the situation described in Section 3.3.1, let \( \Phi \) satisfy the relation
\[
\Phi(0; \mu) \geq 0 \quad \forall \mu \in M. \tag{3.30}
\]
Then
\[
(a) \quad \hat{\Psi}_+(h) := \inf_{\alpha \in \mathcal{X}} \{ \Psi_+(h, \alpha) + \alpha \ln(2/\epsilon) \} : \alpha > 0, (h, \alpha) \in \mathcal{H}^+ \}
\quad = \sup_{x \in X} \inf_{\alpha > 0, (h, \alpha) \in H^+} \left\{ \alpha \Phi(h/\alpha, A(x)) - G(x) - v(x) + \alpha \ln(2/\epsilon) \right\},
\]
\[
(b) \quad \hat{\Psi}_-(h) := \inf_{\alpha \in \mathcal{X}} \{ \Psi_-(h, \alpha) + \alpha \ln(2/\epsilon) \} : \alpha > 0, (h, \alpha) \in \mathcal{H}^+ \}
\quad = \sup_{x \in X} \inf_{\alpha > 0, (h, \alpha) \in H^+} \left\{ \alpha \Phi(-h/\alpha, A(x)) + G(x) - v(x) + \alpha \ln(2/\epsilon) \right\}, \tag{3.31}
\]
and functions \( \hat{\Psi}_\pm : \mathcal{E}_H \rightarrow \mathbb{R} \) are convex. Furthermore, let \( h, z, \rho \) be a feasible solution to the system of convex constraints
\[
\hat{\Psi}_+(h) \leq \rho - z, \quad \hat{\Psi}_-(h) \leq \rho + z \tag{3.32}
\]
in variables \( h, \rho, z \). Then the estimate
\[
\hat{g}(\omega) = \langle \bar{h}, \omega \rangle + \bar{z},
\]
of \( G(x), x \in X, \) has the \( \epsilon \)-risk at most \( \bar{\rho} \):
\[
\text{Risk}_\epsilon(\hat{g}(\cdot) | G, X, v, A, \mathcal{H}, M, \Phi) \leq \bar{\rho}. \tag{3.33}
\]
Relation (3.32) (and thus the risk bound (3.33)) clearly holds true when \( \bar{h} \) is a candidate solution to the convex optimization problem

\[
\text{Opt} = \min_h \left\{ \hat{\Psi}(h) := \frac{1}{2} \left[ \hat{\Psi}_+(h) + \hat{\Psi}_-(h) \right] \right\},
\]

(3.34)

\( \bar{\rho} = \hat{\Psi}(\bar{h}), \) and

\( \bar{\kappa} = \frac{1}{2} \left[ \hat{\Psi}_-(\bar{h}) - \hat{\Psi}_+(\bar{h}) \right]. \)

As a result, by properly selecting \( \bar{h} \), we can make (an upper bound on) the \( \epsilon \)-risk of estimate \( \hat{g}(\cdot) \) arbitrarily close to \( \text{Opt} \), and equal to \( \text{Opt} \) when optimization problem (3.34) is solvable.

**Proof.** Let us first verify the identities in (3.31). The function

\[
\Theta_+(h, \alpha; x) = \alpha \Phi(h/\alpha; \mathcal{A}(x)) - G(x) - v(x) + \alpha \ln(2/\epsilon) : \mathcal{H}^+ \times \mathcal{X} \rightarrow \mathbb{R}
\]

is convex-concave and continuous, and \( \mathcal{X} \) is compact, whence by Sion-Kakutani Theorem

\[
\hat{\Psi}_+(h) := \inf_{\alpha > 0, (h, \alpha) \in \mathcal{H}^+} \{ \Theta_+(h, \alpha; x) \}
\]

\( = \inf_{\alpha > 0, (h, \alpha) \in \mathcal{H}^+} \max_{x \in \mathcal{X}} \Theta_+(h, \alpha; x) \)

\( = \sup_{x \in \mathcal{X}} \inf_{\alpha > 0, (h, \alpha) \in \mathcal{H}^+} \Theta_+(h, \alpha; x) \)

\( = \sup_{x \in \mathcal{X}} \inf_{\alpha > 0, (h, \alpha) \in \mathcal{H}^+} [\alpha \Phi(h/\alpha; \mathcal{A}(x)) - G(x) - v(x) + \alpha \ln(2/\epsilon)] \)

as required in (3.31.a). As we know, \( \hat{\Psi}_+(h, \alpha) \) is real-valued continuous function on \( \mathcal{H}^+ \), so that \( \hat{\Psi}_+ \) 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 \( \Phi(h) = \Phi(h; \mathcal{A}(\bar{x})) \) taken at \( h = 0 \). For \( h \in \mathcal{E}_H \) and all \( \alpha > 0 \) such that \( (h, \alpha) \in \mathcal{H}^+ \) we have

\[
\Psi_+(h, \alpha) \geq \alpha \Phi(h/\alpha; \mathcal{A}(\bar{x})) - G(\bar{x}) - v(\bar{x}) + \alpha \ln(2/\epsilon)
\]

\( \geq \alpha [\Phi(0; \mathcal{A}(\bar{x})) + (e, h/\alpha)] - G(\bar{x}) - v(\bar{x}) + \alpha \ln(2/\epsilon) \)

\( \geq \langle e, h \rangle - G(\bar{x}) - v(\bar{x}) \)

(we have used (3.30)), and therefore \( \Psi_+(h, \alpha) \) as a function of \( \alpha \) is below bounded 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, \( \hat{\Psi}_+ \) is real-valued and convex on \( \mathcal{E}_H \). Verification of (3.31.b) and of the fact that \( \hat{\Psi}_-(h) \) is real-valued convex function on \( \mathcal{E}_H \) is completely similar.

Now, given a feasible solution \((\bar{h}, \bar{\kappa}, \bar{\rho})\) to (3.32), let us select somehow \( \bar{\rho} > \bar{\rho} \). Taking into account the definition of \( \hat{\Psi}_\pm \), we can find \( \bar{\alpha} \) and \( \bar{\beta} \) such that

\[
(\bar{h}, \bar{\alpha}) \in \mathcal{H}^+ \ \& \ \Psi_+(\bar{h}, \bar{\alpha}) + \bar{\alpha} \ln(2/\epsilon) \leq \bar{\rho} - \bar{\kappa},
\]

\[
(\bar{h}, \bar{\beta}) \in \mathcal{H}^+ \ \& \ \Psi_-(\bar{h}, \bar{\beta}) + \bar{\beta} \ln(2/\epsilon) \leq \bar{\rho} + \bar{\kappa},
\]

implying that the collection \((\bar{h}, \bar{\alpha}, \bar{\beta}, \bar{\kappa}, \bar{\rho})\) is a feasible solution to (3.29). Invoking Proposition 3.5, we get

\[
\text{Prob}_{\omega \sim P} \{ \omega : |\hat{g}(\omega) - G(x)| > \bar{\rho} + v(x) \} \leq \epsilon
\]

for all \((x \in \mathcal{X}, P \in \mathcal{P})\) satisfying (3.26). Since \( \bar{\rho} \) can be selected arbitrarily close to \( \bar{\rho} \), \( \hat{g}(\cdot) \) indeed is a \((\bar{\rho}, \epsilon, v(\cdot))\)-accurate estimate. \( \square \)
3.3.3 Estimation from repeated observations

Assume that in the situation described in Section 3.3.1 we have access to \( K \) observations \( \omega_1, \ldots, \omega_K \) sampled, independently of each other, from a probability distribution \( P \), and aim to build the 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

\( \mathcal{H} \subset \mathcal{E}_H, \mathcal{M} \subset \mathcal{E}_M, \Phi(\cdot; \cdot): \mathcal{H} \times \mathcal{M} \rightarrow \mathbb{R}, \mathcal{X} \subset \mathcal{E}_X, v(\cdot), A(\cdot), G(x) = \langle g, x \rangle + c, \)

see Section 3.3.1, and a positive integer \( K \), let us update part of the data, namely, replace \( \mathcal{H} \subset \mathcal{E}_H \) with

\[
\mathcal{H}^K := \mathcal{H} \times \ldots \times \mathcal{H} := \mathcal{E}_H \times \ldots \times \mathcal{E}_H.
\]

and replace \( \Phi(\cdot, \cdot): \mathcal{H} \times \mathcal{M} \rightarrow \mathbb{R} \) with

\[
\Phi^K(h^K = (h_1, \ldots, h_K); \mu) = \sum_{i=1}^K \Phi(h_i; \mu) : \mathcal{H}^K \times \mathcal{M} \rightarrow \mathbb{R}.
\]

It is immediately seen that the updated data satisfy all requirements imposed on the data in Section 3.3.1, and that whenever \( x \in \mathcal{X} \) and a Borel probability distribution \( P \) on \( \mathcal{E}_H \) are linked by (3.26), \( x \) and the distribution \( P^K \) of \( K \)-element i.i.d. sample \( \omega^K = (\omega_1, \ldots, \omega_K) \) drawn from \( P \) are linked by the relation

\[
\forall h^K = (h_1, \ldots, h_K) \in \mathcal{H}^K : \ln \left( \int_{\mathcal{E}_H} e^{\langle h^K, \omega^K \rangle} P^K(d\omega^K) \right) = \sum_{i=1}^K \ln \left( \int_{\mathcal{E}_H} e^{\langle h_i, \omega_i \rangle} P(d\omega_i) \right) \leq \Phi^K(h^K; A(x)).
\]

Applying to our new data the construction from Section 3.3.2, we arrive at “repeated observations” versions of Proposition 3.5 and Corollary 3.6. Note that the resulting convex constraints/objectives are symmetric w.r.t. permutations functions of the components \( h_1, \ldots, h_K \) 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 \( \bar{h}/K \). With this observation in mind, Proposition 3.5 and Corollary 3.6 translate into the following statements (we use the assumptions and the notation from the previous sections):

**Proposition 3.7.** Given \( \epsilon \in (0, 1) \) and positive integer \( K \), let

\[
(a) \quad \Psi_+(h, \alpha) = \sup_{x \in \mathcal{X}} \left[ \alpha \Phi(h/\alpha, A(x)) - G(x) - v(x) \right] : \mathcal{H}^+ \rightarrow \mathbb{R},
\]

\[
(b) \quad \Psi_-(h, \beta) = \sup_{x \in \mathcal{X}} \left[ \beta \Phi(-h/\beta, A(x)) + G(x) - v(x) \right] : \mathcal{H}^+ \rightarrow \mathbb{R},
\]

and let \( \bar{h}, \bar{\alpha}, \bar{\beta}, \bar{\kappa}, \bar{\rho} \) be a feasible solution to the system of convex constraints

\[
\begin{align*}
(a_1) \quad & (h, \alpha) \quad \in \quad \mathcal{H}^+ \\
(a_2) \quad & (h, \beta) \quad \in \quad \mathcal{H}^+ \\
(b_1) \quad & \alpha K^{-1} \ln(\epsilon/2) \quad \geq \quad \Psi_+(h, \alpha) - \rho + \kappa \\
(b_2) \quad & \beta K^{-1} \ln(\epsilon/2) \quad \geq \quad \Psi_-(h, \beta) - \rho - \kappa
\end{align*}
\]
in variables \( h, \alpha, \beta, \rho, \kappa \). Setting
\[
\hat{g}(\omega^K) = \left( \frac{1}{K} \sum_{i=1}^{K} \omega_i \right) + \bar{\kappa},
\]
we obtain an estimate of \( G(x) \) via independent \( K \)-repeated observations
\[
\omega_i \sim P, \ i = 1, ..., K,
\]
with the \( \epsilon \)-risk on \( X \) not exceeding \( \bar{\rho} \). In other words, whenever \( x \in X \) and a Borel probability distribution \( P \) on \( \mathcal{E}_H \) are linked by (3.26), one has
\[
\text{Prob}_{x \sim P} \left\{ \omega^K : |\hat{g}(\omega^K) - G(x)| > \bar{\rho} + \nu(x) \right\} \leq \epsilon. \tag{3.36}
\]

**Corollary 3.8.** In the situation described in the beginning of Section 3.3.1, let \( \Phi \) satisfy relation (3.30), and let a positive integer \( K \) be given. Then
\[
(a) \ \hat{\Psi}_{+,K}(h) := \inf_{h} \left\{ \Psi_{+,K}(h) + K^{-1} \alpha \ln(2/\epsilon) : \alpha > 0, (h, \alpha) \in H^+ \right\} = \sup_{x \in X} \inf_{\alpha > 0, (h, \alpha) \in H^+} \left[ \alpha \Phi(h/\alpha, A(x)) - G(x) - \nu(x) + K^{-1} \alpha \ln(2/\epsilon) \right],
\]
\[
(b) \ \hat{\Psi}_{-,K}(h) := \inf_{h} \left\{ \Psi_{-,K}(h) + K^{-1} \alpha \ln(2/\epsilon) : \alpha > 0, (h, \alpha) \in H^+ \right\} = \sup_{x \in X} \inf_{\alpha > 0, (h, \alpha) \in H^+} \left[ \alpha \Phi(-h/\alpha, A(x)) + G(x) - \nu(x) + K^{-1} \alpha \ln(2/\epsilon) \right].
\]
and functions \( \hat{\Psi}_{\pm,K} : \mathcal{E}_H \to \mathbb{R} \) are convex. Furthermore, let \( \bar{h}, \bar{\kappa}, \bar{\rho} \) be a feasible solution to the system of convex constraints
\[
\hat{\Psi}_{+,K}(h) \leq \rho - \kappa, \ \hat{\Psi}_{-,K}(h) \leq \rho + \kappa \tag{3.37}
\]
in variables \( h, \rho, \kappa \). Then the \( \epsilon \)-risk of the estimate
\[
\hat{g}(\omega^K) = \left( \frac{1}{K} \sum_{i=1}^{K} \omega_i \right) + \bar{\kappa},
\]
of \( G(x), \ x \in X \), is at most \( \hat{\Psi}(\bar{h}) \), implying that whenever \( x \in X \) and a Borel probability distribution \( P \) on \( \mathcal{E}_H \) are linked by (3.26), relation (3.36) holds true.

Relation (3.37) clearly holds true when \( \bar{h} \) is a candidate solution to the convex optimization problem
\[
\text{Opt}_K = \min_{\bar{h}} \left\{ \bar{\Psi}_K(h) := \frac{1}{2} \left[ \hat{\Psi}_{+,K}(h) + \hat{\Psi}_{-,K}(h) \right] \right\}, \tag{3.38}
\]
\( \bar{\rho} = \hat{\Psi}_K(\bar{h}) \) and
\[
\bar{\kappa} = \frac{1}{2} \left[ \hat{\Psi}_{-,K}(\bar{h}) - \hat{\Psi}_{+,K}(\bar{h}) \right].
\]
As a result, by properly selecting \( \bar{h} \) we can make (an upper bound on) the \( \epsilon \)-risk of the estimate \( \hat{g}(\cdot) \) arbitrarily close to \( \text{Opt} \), and equal to \( \text{Opt} \) when optimization problem (3.38) is solvable.

From now on, if not explicitly stated otherwise, 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 forms of sub-Gaussianity parameters

Consider the simplest case of the situation from Sections 3.3.1, 3.3.3, where
\[ \mathcal{H} = \mathcal{E}_H = \mathbb{R}^d, \mathcal{M} = \mathcal{E}_M = \mathbb{R}^d \times \mathbb{S}_+^d, \]
\[ \Phi(h; \mu, M) = h^T \mu + \frac{1}{2} h^T M h : \mathbb{R}^d \times (\mathbb{R}^d \times \mathbb{S}_+^d) \to \mathbb{R}, \]
so that \( S[\mathcal{H}, \mathcal{M}, \Phi] \) is the family of all sub-Gaussian distributions on \( \mathbb{R}^d \);
\bullet \quad \mathcal{X} \subset \mathcal{E}_X = \mathbb{R}^{n_x} \) is a nonempty convex compact set, and
\bullet \quad A(x) = (Ax + a, M(x)), where \( A \) is \( d \times n_x \) matrix, and \( M(x) \) is an affinely depending on \( x \) symmetric \( d \times d \) matrix such that \( M(x) \) is \( \succeq 0 \) when \( x \in \mathcal{X} \),
\bullet \quad v(x) \) is a convex continuous function on \( \mathcal{X} \),
\bullet \quad G(x) \) is an affine function on \( \mathcal{E}_X \).

In the case in question, (3.30) clearly takes place, and the left hand sides in constraints (3.37) become
\[ \hat{\Psi}_{+,K}(h) = \sup_{x \in \mathcal{X}} \inf_{\alpha > 0} \left\{ h^T [Ax + a] + \frac{1}{2} h^T M(x) h + K^{-1} \alpha \ln(2/\epsilon) - G(x) - v(x) \right\} = \max_{x \in \mathcal{X}} \left\{ \sqrt{2K^{-1} \ln(2/\epsilon)} h^T M(x) h + h^T [Ax + a] - G(x) - v(x) \right\}, \]
\[ \hat{\Psi}_{-,K}(h) = \sup_{x \in \mathcal{X}} \inf_{\alpha > 0} \left\{ -h^T [Ax + a] + \frac{1}{2} h^T M(x) h + K^{-1} \alpha \ln(2/\epsilon) + G(x) - v(x) \right\} = \max_{x \in \mathcal{X}} \left\{ \sqrt{2K^{-1} \ln(2/\epsilon)} h^T M(x) h - h^T [Ax + a] + G(x) - v(x) \right\}. \]

Thus, system (3.37) reads
\[ a^T h + \max_{x \in \mathcal{X}} \left[ \sqrt{2K^{-1} \ln(2/\epsilon)} h^T M(x) h + h^T Ax - G(x) - v(x) \right] \leq \rho - \kappa, \]
\[ -a^T h + \max_{x \in \mathcal{X}} \left[ \sqrt{2K^{-1} \ln(2/\epsilon)} h^T M(x) h - h^T Ax + G(x) - v(x) \right] \leq \rho + \kappa. \]

We arrive at the following version of Corollary 3.8:

**Proposition 3.9.** 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
\[ \text{Opt}_K = \min_{h \in \mathbb{R}^d} \hat{\Psi}_K(h) \] (3.39)

where
\[ \hat{\Psi}_K(h) := \frac{1}{2} \left[ \max_{x \in \mathcal{X}} \left\{ \sqrt{2K^{-1} \ln(2/\epsilon)} h^T M(x) h + h^T Ax - G(x) - v(x) \right\} + a^T h \right] + \max_{y \in \mathcal{Y}} \left\{ \sqrt{2K^{-1} \ln(2/\epsilon)} h^T M(y) h - h^T Ay + G(y) - v(y) \right\} - a^T h. \]

Then, setting
\[ \bar{\kappa} = \frac{1}{2} \left[ \hat{\Psi}_{-,K}(\bar{h}) - \hat{\Psi}_{+,K}(\bar{h}) \right], \quad \bar{\rho} = \hat{\Psi}_K(\bar{h}), \]
the affine estimate
\[ \hat{g}(\omega^K) = \frac{1}{K} \sum_{i=1}^K h^T \omega_i + \bar{\kappa} \]
has $\epsilon$-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 (3.39) is solvable, provided that

$$\bigcap_{x \in \mathcal{X}} \text{Ker}(M(x)) = \{0\},$$

and an optimal solution $h_*$ to the problem, taken along with

$$\mathcal{K}_* = \frac{1}{2} [\hat{\Psi}_{-,K}(h_*) - \hat{\Psi}_{+,K}(h_*)],$$

(3.40)
yields the affine estimate

$$\hat{g}_*(\omega) = \frac{1}{K} \sum_{i=1}^{K} h_i^T \omega_i + \mathcal{K}_*$$

with $\epsilon$-risk, taken w.r.t. the data listed in the beginning of this section, at most $\text{Opt}_K$.

3.3.4.1 Consistency

Assuming $\nu(x) \equiv 0$, 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 $\epsilon$-risk, provided that $K$ is large enough. Specifically, denoting by $g^T_x$ the linear part of $G(x)$: $G(x) = g^T x + c$, from Proposition 3.9 it is immediately seen that a necessary and sufficient condition for consistency is the existence of $\bar{h} \in \mathbb{R}^d$ such that $\bar{h}^T A x = g^T x$ 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 $\lim_{K \to \infty} \hat{\Psi}_K(\bar{h}) = 0$, implying that $\lim_{K \to \infty} \text{Opt}_K = 0$, with $\hat{\Psi}_K$ and $\text{Opt}_K$ given by (3.39). On the other hand, if the condition is violated, then there exist $x', x'' \in \mathcal{X}$ such that $A x' = A x''$ and $G(x') \neq G(x'')$; we lose noting when assuming that $G(x'') > G(x')$. Looking at (3.39), we see that

$$\hat{\Psi}_K(h) \geq \frac{1}{2} \left( \sqrt{2K^{-1} \ln(2/\epsilon)} |h^T M(x') h| + h^T A x' - G(x') \right) + a^T h$$

$$\geq \left( \sqrt{2K^{-1} \ln(2/\epsilon)} |h^T M(x'') h| - h^T A x'' + G(x'') \right) - a^T h$$

whence $\text{Opt}_K$, for all $K$, is lower-bounded by $G(x'') - G(x') > 0$.

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 $\mathcal{X} = U \times V$, with convex compact sets $U \subset \mathcal{E}_U = \mathbb{R}^{n_u}$ and $V \subset \mathcal{E}_V = \mathbb{R}^{n_v}$,
- $A(x = (u,v)) = [Au + a, M(v)] : U \times V \to \mathbb{R}^d \times \mathbb{S}^d$, with $M(v) \succeq 0$ for $v \in V$,
- $G(x = (u,v)) = g^T u + c$ depends solely on $u$, and
• \( v(x = (u, v)) = g(u) \) depends solely on \( u \).

It is immediately seen that in the direct product case problem (3.39) reads

\[
\text{Opt}_K = \min_{h \in \mathbb{R}^d} \left\{ \frac{\phi_U(A^T h - g) + \phi_U(-A^T h + g)}{2} + \max_{v \in V} \sqrt{2K^{-1} \ln(2/\epsilon)h^T M(v)h} \right\},
\]

where

\[
\phi_U(f) = \max_{u \in U} [u^T f - g(u)].
\]

Assuming \( \bigcap_{v \in V} \text{Ker}(M(v)) = \{0\} \), the problem is solvable, and its optimal solution \( h^*_v \) gives rise to the affine estimate

\[
\hat{g}_v(\omega^K) = \frac{1}{K} \sum_i h^*_i \omega_i + \kappa_v, \quad \kappa_v = \frac{1}{2} [\phi_U(-A^T h + g) - \phi_U(A^T h - g)] - a^T h_v + c
\]

with \( \epsilon \)-risk \( \leq \text{Opt}_K \).

**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 (3.39) reads

\[
\text{Opt}_K = \min_{h} \max_{v \in V} \left\{ \Theta(h, v) := \frac{1}{2} [\phi_U(A^T h - g) + \phi_U(-A^T h + g)] + \sqrt{2K^{-1} \ln(2/\epsilon)h^T M(v)h} \right\}.
\]

Hence, 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

\[
\text{Opt}_K = \max_{v \in V} \left[ \text{Opt}(v) = \min_h \frac{1}{2} [\phi_U(A^T h - g) + \phi_U(-A^T h + g)] + \sqrt{2K^{-1} \ln(2/\epsilon)h^T M(v)h} \right].
\]

Now consider the problem of estimating \( g^T u \) from independent observations \( \omega_i, \ i \leq K \), 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 \( \epsilon \)-risk of recovery:

\[
\rho_\epsilon(v) = \inf_{\hat{g}(\cdot)} \left\{ \rho : \text{Prob}_{\omega_i \sim \mathcal{N}(Au + a, M(v))} \left[ \omega^K : |\hat{g}^{\epsilon}(\omega^K) - g^T u| > \rho \right] \leq \epsilon \forall u \in U \right\},
\]

where inf is taken over all Borel functions \( \hat{g}(\cdot) : \mathbb{R}^{Kd} \rightarrow \mathbb{R} \). Invoking [130, Theorem 3.1], it is immediately seen that whenever \( \epsilon < 1/4 \), one has

\[
\rho_\epsilon(v) \geq \left[ \frac{2 \ln(2/\epsilon)}{\ln \left( \frac{1}{4\epsilon} \right)} \right]^{-1} \text{Opt}(v).
\]

Since the family \( \mathcal{SG}(U, V) \) of all sub-Gaussian, with parameters \( (Au + a, M(v)) \), \( u \in U \), \( v \in V \), distributions on \( \mathbb{R}^d \) 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 3.10. In the just described situation, the minimax optimal $\epsilon$-risk

$$\Risk_{\text{opt}}^\epsilon(K) = \inf_{\hat{g}()} \Risk_{\epsilon}(\hat{g}(\cdot)),$$

of recovering $g^T u$ from a $K$-repeated i.i.d. sub-Gaussian, with parameters $(Au + a, M(v))$, $(u, v) \in U \times V$, observation is within a moderate factor of the upper bound $\Opt_K$ on the $\epsilon$-risk, taken w.r.t. the same data, of the affine estimate $\hat{g}_*(\cdot)$ yielded by an optimal solution to (3.41), namely,

$$\Opt_K \leq \frac{2 \ln(2/\epsilon)}{\ln \left(\frac{1}{4}\epsilon\right)} \Risk_{\text{opt}}^\epsilon(K).$$

3.3.4.3 Numerical illustration

The numerical illustration we are about to discuss models the situation in which 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

$$\omega \sim SG(Ax, M(x)),$$

where

$$x \in X = \{x \in \mathbb{R}^n : 0 \leq x_j \leq j^{-\alpha}, 1 \leq j \leq n\}, \quad M(x) = \sigma^2 \sum_{j=1}^{n} x_j \Theta_j \quad (3.44)$$

where $A \in \mathbb{R}^{d \times n}$ and $\Theta_j \in \mathbb{S}_d^+$, $j = 1, ..., n$, are given matrices; the linear form to be estimated 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 $g, A, \Theta_j$ are as follows:

- $g \geq 0$ is selected at random and then normalized to have
  \[ \max_{x \in X} g^T x = \max_{x,y \in X} g^T |x - y| = 2; \]

- we deal with the case of $n > d$ (“deficient observations”); the $d$ nonzero singular values of $A$ were set to $\theta^{-\frac{d-n}{2}}$, where “condition number” $\theta \geq 1$ is a parameter; the orthonormal systems $U$ and $V$ of the first $d$ left and, respectively, right singular vectors of $A$ were drawn at random from rotationally invariant distributions;

- the positive semidefinite $d \times d$ matrices $\Theta_j$ are orthogonal projectors on randomly selected subspaces in $\mathbb{R}^d$ of dimension $\lfloor d/2 \rfloor$;

- in all our experiments, we consider single-observation case $K = 1$ and use $v(\cdot) \equiv 0$.

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 the considered experiment is to understand how much we might loose were we replacing $M(\cdot)$
with \( \hat{M}(x) \equiv M(\bar{x}) \), that is, were we ignoring the fact that small signals result in low-noise observations.

In our experiment we use \( d = 32, m = 48, \alpha = 2, \theta = 2, \) and \( \sigma = 0.01 \). With 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 each problem, we apply the outlined machinery to build affine in \( \omega \) estimate of \( g^T x \) as yielded by optimal solution to (3.39), and compute upper bound \( \text{Opt} \) on \((\epsilon = 0.01)\)-risk of this estimate. In fact, for each problem, 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 conservative \( x \mapsto \hat{M}(x) := M(\bar{x}) \). The results are as follows:

<table>
<thead>
<tr>
<th></th>
<th>min</th>
<th>median</th>
<th>mean</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.138</td>
<td>0.190</td>
<td>0.212</td>
<td>0.299</td>
</tr>
<tr>
<td></td>
<td>0.150</td>
<td>0.210</td>
<td>0.227</td>
<td>0.320</td>
</tr>
</tbody>
</table>

Upper bounds on 0.01-risk, data over 10 estimation problems

\([d = 32, m = 48, \alpha = 2, \theta = 2, \sigma = 0.01]\)

First row: \( \omega \sim SG(Ax, M(x)) \); second row: \( \omega \sim SG(Ax, M(\bar{x})) \)

Note significant “noise amplification” in the estimate (about 20 times the observation noise level \( \sigma \)) and high risk variability across the experiments. Seemingly, both these phenomena stem from the fact that we have highly deficient observations \((n/d = 1.5)\) combined with random 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 (3.25) to their quadratic lifting, we can use the just developed machinery to estimate quadratic, rather than linear, forms of the underlying parameters. We investigate the related possibilities in the cases of Gaussian and sub-Gaussian observations. The results of this section form an essential extension of the results of [39, 81] where a similar approach to estimating quadratic functionals of the mean of a Gaussian vector was used.

#### 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 \( \mathbb{R}^m \);
- a nonempty convex compact subset \( \mathcal{V} \) of the positive semidefinite cone \( S^d_+ \);
- a matrix \( \Theta_\ast \succ 0 \) such that \( \Theta_\ast \succeq \Theta \) for all \( \Theta \in \mathcal{V} \);
- an affine mapping \( u \mapsto A[u; 1] : \mathbb{R}^m \to \Omega = \mathbb{R}^d \), where \( A \) is a given \( d \times (m + 1) \) matrix,
- a convex continuous function \( \varrho(\cdot) \) on \( S^m_+ \).
A pair \((u \in U, \Theta \in V)\) specifies Gaussian random vector \(\zeta \sim \mathcal{N}(A[u; 1], \Theta)\) and thus specifies probability distribution \(P[u, \Theta]\) of \((\zeta, \zeta^T)\). Let \(Q(U, V)\) be the family of probability distributions on \(\Omega = \mathbb{R}^d \times S^d\) stemming this way from Gaussian distributions with parameters from \(U \times V\). Our goal is to cover the family \(Q(U, V)\) by a family of the type \(S[N, M, \Phi]\).

It is convenient to represent a linear form on \(\Omega = \mathbb{R}^d \times S^d\) as
\[
h^T z + \frac{1}{2} \text{Tr}(H Z),
\]
where \((h, H) \in \mathbb{R}^d \times S^d\) is the “vector of coefficients” of the form, and \((z, Z) \in \mathbb{R}^d \times S^d\) is the argument of the form.

We assume that for some \(\delta \in [0, 2]\) it holds
\[
\|\Theta^{1/2} \Theta^{-1/2} - I_d\| \leq \delta \ \forall \Theta \in V,
\]
where \(\|\cdot\|\) is the spectral norm (cf. (2.129)). Finally, we set
\[
b = [0; \ldots; 0; 1] \in \mathbb{R}^m + 1, \quad B = \begin{bmatrix} A \\ b^T \end{bmatrix}
\]
and
\[
Z^+ = \{W \in S^{m+1}_+ : W_{m+1,m+1} = 1\}.
\]
The statement below is nothing but a straightforward reformulation of Proposition 2.43.i:

**Proposition 3.11.** In the just described situation, let us select \(\gamma \in (0, 1)\) and set
\[
\mathcal{H} = \mathcal{H}_\gamma := \{(h, H) \in \mathbb{R}^d \times S^d : \gamma \Theta^{-1}_- \preceq H \preceq \gamma \Theta^{-1}_+\},
\]
\[
\mathcal{M}^+ = V \times Z^+,
\]
\[
\Phi(h, H; \Theta, Z) = -\frac{1}{2} \ln \det(I - \Theta^{-1}_+ H \Theta^{-1}_+^T) + \frac{1}{2} \text{Tr}(\Theta - \Theta_1 H) + \frac{1}{2(2+\delta)}\|\Theta^{1/2} H \Theta^{1/2}_+\|_F^2 + \Gamma(h, H; Z) : \mathcal{H} \times \mathcal{M}^+ \to \mathbb{R},
\]
where \(\|\cdot\|\) is the spectral, \(\|\cdot\|_F\) is the Frobenius norm, and
\[
\Gamma(h, H; Z) = \frac{1}{2} \text{Tr}(Z[h h^T + A^T h b^T + A^T H A + B^T [H, h]^T [\Theta^{-1}_- - H]^{-1} [H, h] B])
\]
\[
= \frac{1}{2} \text{Tr}(ZB^T \begin{bmatrix} \frac{H}{h^T h} & [H, h]^T [\Theta^{-1}_- - H]^{-1} [H, h] B \\ 1 & 0 \end{bmatrix} Z).
\]
Then \(\mathcal{H}, \mathcal{M}^+, \Phi\) is a regular data, and for every \((u, \Theta) \in \mathbb{R}^m \times V\) it holds
\[
\forall (h, H) \in \mathcal{H} : \ln \left(\mathbf{E}_{\zeta \sim \mathcal{N}(A[u; 1], \Theta)} \left\{ \exp \left(\frac{1}{2} \zeta^T H \zeta \right) \right\} \right) \leq \Phi(h, H; \Theta, [u; 1][u; 1]^T).
\]

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 and goal

Let us assume that we are given a sample \(\zeta^K = (\zeta_1, \ldots, \zeta_K)\) of independent across \(i\) and identically distributed observations
\[
\zeta_i \sim \mathcal{N}(A[u; 1], M(v)), \ 1 \leq i \leq K,
\]
(3.46)
where

- \((u, v)\) is an unknown “signal” known to belong to a given set \(U \times V\), where
  - \(U \subset \mathbb{R}^m\) is a compact set, and
  - \(V \subset \mathbb{R}^k\) is a compact convex set;
- \(A\) is a given \(d \times (m + 1)\) matrix, and \(v \mapsto M(v) : \mathbb{R}^k \to \mathbb{S}_d\) is affine mapping such that \(M(v) \succeq 0\) whenever \(v \in V\).

We are also given a convex calibrating function \(\varrho(Z) : \mathbb{S}_+^{m+1} \to \mathbb{R}\) and “functional of interest”

\[
F(u, v) = [u; 1]^T Q[u; 1] + q^T v, \tag{3.47}
\]

where \(Q\) and \(q\) are known \((m + 1) \times (m + 1)\) symmetric matrix and \(k\)-dimensional vector, respectively. Our goal is to estimate the value \(F(u, v)\), for unknown \((u, v)\) known to belong to \(U \times V\). Given a tolerance \(\epsilon \in (0, 1)\), we quantify the quality of a candidate estimate \(\hat{g}(\zeta^K)\) of \(F(u, v)\) by the smallest \(\rho\) such that for all \((u, v) \in U \times V\) it holds

\[
\Pr_{\zeta^K \sim \mathcal{N}(\mathcal{A}[u; 1], M(v))} \left\{ |\hat{g}(\zeta^K) - F(u, v)| > \rho + \varrho([u; 1][u; 1]^T) \right\} \leq \epsilon.
\]

### 3.4.1.3 Construction and result

Let

\[
\mathcal{V} = \{ M(v) : v \in V \},
\]

so that \(\mathcal{V}\) is a convex compact subset of the positive semidefinite cone \(\mathbb{S}_+^d\). Let us select somehow

1. a matrix \(\Theta_\ast > 0\) such that \(\Theta_\ast \succeq \Theta\), for all \(\Theta \in \mathcal{V}\);
2. a convex compact subset \(Z\) of the set \(Z^+ = \{ Z \in \mathbb{S}_+^{m+1} : Z_{m+1,m+1} = 1 \}\) such that \([u; 1][u; 1]^T \in Z\) for all \(u \in U\);
3. a real \(\gamma \in (0, 1)\) and a nonnegative real \(\delta\) such that (3.45) takes place.

We further set (cf. Proposition 3.11)

\[
B = \begin{bmatrix} A & \mathbb{0} \\ \mathbb{0}, \ldots, \mathbb{0} \end{bmatrix} \in \mathbb{R}^{(d+1) \times (m+1)},
\]

\[
\mathcal{H} = \mathcal{H}_\gamma := \{(h, H) \in \mathbb{R}^d \times \mathbb{S}_d : -\gamma \Theta_\ast^{-1} \preceq H \preceq \gamma \Theta_\ast^{-1}\},
\]

\[
\mathcal{M} = \mathcal{V} \times Z,
\]

\[
\Phi(h, H; \Theta, Z) = -\frac{1}{2} \ln \det(I - \Theta_\ast^{1/2} H \Theta_\ast^{1/2}) + \frac{1}{2} \text{Tr}(\Theta - \Theta_\ast H) + \frac{\delta(2+\delta)}{2(1-\|\Theta_\ast^{1/2} H \Theta_\ast^{1/2}\|^2)} \|\Theta_\ast^{1/2} H \Theta_\ast^{1/2}\|^2 \Gamma(h, H; Z) : \mathcal{H} \times \mathcal{M} \to \mathbb{R},
\]

where

\[
\Gamma(h, H; Z) = \frac{1}{2} \text{Tr} \left( Z [hb^T + A^T h b^T + A^T H A + B^T [H, h]^T [\Theta_\ast^{-1} - H]^{-1} [H, h] B] \right) + \frac{1}{2} \text{Tr} \left( Z B^T \left[ \begin{bmatrix} h & [H, h] \end{bmatrix} \right] [\Theta_\ast^{-1} - H]^{-1} [H, h] B \right)
\]

and treat, as observation, the quadratic lifting of observation (3.46), that is, our observation is

\[
\omega^K = \{ \omega_i = (\zeta_i, \zeta_i^T) \}_{i=1}^K, \quad \text{with independent } \zeta_i \sim \mathcal{N}(\mathcal{A}[u; 1], M(v)). \tag{3.49}
\]
Note that by Proposition 3.11 function $\Phi(h, H; \Theta, Z) : \mathcal{H} \times \mathcal{M} \to \mathbb{R}$ is a continuous convex-concave function which is coercive in convex argument and is such that

$$\forall (u \in U, v \in V, (h, H) \in \mathcal{H}) : \ln \left( \mathbf{E}_{\zeta \sim \mathcal{N}(A[u;1], M(v))} \left\{ e^{\frac{1}{2} \zeta^T H \zeta + h^T \zeta} \right\} \right) \leq \Phi(h, H; M(v), [u;1][u;1]^T). \quad (3.50)$$

We are about to demonstrate that when estimating the functional of interest (3.47) at a point $(u,v) \in U \times V$ via observation (3.49), we are in the situation considered in Section 3.3 and can utilize the corresponding machinery. Indeed, let us specify the following data introduced in Section 3.3.1:

- $\mathcal{H} = \{ f = (h, H) \in \mathcal{H} \} \subset \mathcal{E}_H = \mathbb{R}^d \times S^d$, with $\mathcal{H}$ defined in (3.48), and the inner product on $\mathcal{E}_H$ defined as
  $$\langle (h, H), (h', H') \rangle = h^T h' + \frac{1}{2} \text{Tr}(HH'),$$

- $\mathcal{E}_M = S^d \times S^{m+1}$, and $\mathcal{M}, \Phi$ are as defined in (3.48);
- $\mathcal{E}_X = \mathbb{R}^k \times S^{m+1}, \mathcal{X} = V \times Z$;
- $\mathcal{A}(x = (v, Z)) = (M(v), Z)$; note that $\mathcal{A}$ is an affine mapping from $\mathcal{E}_X$ into $\mathcal{E}_M$ which maps $\mathcal{X}$ into $\mathcal{M}$, as required in Section 3.3.1. Observe that when $u \in U$ and $v \in V$, the common distribution $P = P_{u,v}$ of i.i.d. observations $\omega_i$ defined by (3.49) satisfies the relation
  $$\forall (f = (h, H) \in \mathcal{H}) : \ln \left( \mathbf{E}_{\omega \sim P} \left\{ e^{\langle f, \omega \rangle} \right\} \right) = \ln \left( \mathbf{E}_{\zeta \sim \mathcal{N}(A[u;1], M(v))} \left\{ e^{h^T \zeta + \frac{1}{2} \zeta^T H \zeta} \right\} \right) \leq \Phi(h, H; M(v), [u;1][u;1]^T), \quad (3.51)$$
  see (3.50); 
- $\psi(x = (v, Z)) = \varphi(Z) : \mathcal{X} \to \mathbb{R}$, 
- we define affine functional $G(x)$ on $\mathcal{E}_X$ by the relation
  $$\langle g, x := (v, Z) \rangle = q^T v + \text{Tr}(QZ),$$
  see (3.47). As a result, for $x = (v, [u;1][u;1]^T)$ with $v \in V$ and $u \in U$ we have
  $$F(u, v) = G(x).$$

Applying Corollary 3.8 to the just specified data (which is legitimate, because our $\Phi$ clearly satisfies (3.30)), we arrive at the result as follows:
Proposition 3.12. In the just described situation, let us set
\[
\hat{\Psi}_{+,K}(h, H) := \inf \alpha \left\{ \max_{(v,Z) \in V \times Z} \left[ \alpha \Phi(h/\alpha, H/\alpha; M(v), Z) - G(v, Z) - g(Z) + K^{-1} \alpha \ln(2/\epsilon) \right] : \right. \\
\left. \alpha > 0, -\gamma_\alpha \Theta^{-1} \leq H \leq \gamma_\alpha \Theta^*-1 \right\}
\]
\[
= \max_{(v,Z) \in V \times Z} \inf_{\alpha > 0} \left[ \alpha \Phi(h/\alpha, H/\alpha; M(v), Z) - G(v, Z) - g(Z) + K^{-1} \alpha \ln(2/\epsilon) \right],
\]
\[
\hat{\Psi}_{-,K}(h, H) := \inf \alpha \left\{ \max_{(v,Z) \in V \times Z} \left[ \alpha \Phi(-h/\alpha, -H/\alpha; M(v), Z) + G(v, Z) - g(Z) + K^{-1} \alpha \ln(2/\epsilon) \right] : \right. \\
\left. \alpha > 0, -\gamma_\alpha \Theta^{-1} \leq H \leq \gamma_\alpha \Theta^*-1 \right\}
\]
\[
= \max_{(v,Z) \in V \times Z} \inf_{\alpha > 0} \left[ \alpha \Phi(-h/\alpha, -H/\alpha; M(v), Z) + G(v, Z) - g(Z) + K^{-1} \alpha \ln(2/\epsilon) \right].
\]
so that functions \( \hat{\Psi}_{\pm,K}(h, H) : \mathbb{R}^d \times S^d \to \mathbb{R} \) are convex. Furthermore, whenever \( h, H, \hat{\rho}, \hat{\nu} \) form a feasible solution to the system of convex constraints
\[
\hat{\Psi}_{+,K}(h, H) \leq \rho - \epsilon, \quad \hat{\Psi}_{-,K}(h, H) \leq \rho + \epsilon
\]
in variables \((h, H) \in \mathbb{R}^d \times S^d, \rho \in \mathbb{R}, \epsilon \in \mathbb{R}, \) setting
\[
\hat{g}(\zeta^K := (\zeta_1, ..., \zeta_K)) = \frac{1}{K} \sum_{i=1}^K \left[ h^T \zeta_i + \frac{1}{2} \zeta_i^T H \zeta_i \right] + \hat{\nu},
\]
we get an estimate of the functional of interest \( F(u, v) = [u; 1]^T Q[u; 1] + q^T v \) via \( K \) independent observations
\[
\zeta_i \sim \mathcal{N}(A[u; 1], M(v)), \ i = 1, ..., K,
\]
with the following property:
\[
\forall (u, v) \in U \times V : \quad \Pr_{\zeta^K \sim [\mathcal{N}(A[u; 1], M(v))]} \left\{ |F(u, v) - \hat{g}(\zeta^K)| > \hat{\rho} + g([u; 1][u; 1]^T) \right\} \leq \epsilon.
\]
It remains to note that by construction for \( x = (v, Z) \) in question it holds
\[
G(x) = q^Tv + \text{Tr}(QZ) = q^Tv + \text{Tr}(Q[u;1][u;1]^T) = q^Tv + [u;1]^TQ[u;1] = F(u,v).
\]
An immediate consequence of Proposition 3.12 is as follows:

**Corollary 3.13.** Under the premise and in the notation of Proposition 3.12, let \((h, H) \in \mathbb{R}^d \times \mathbb{S}^d\). Setting
\[
\begin{align*}
\rho &= \frac{1}{2} \left[ \hat{\Psi}_{+,K}(h, H) + \hat{\Psi}_{-,K}(h, H) \right], \\
\kappa &= \frac{1}{2} \left[ \hat{\Psi}_{-,K}(h, H) - \hat{\Psi}_{+,K}(h, H) \right],
\end{align*}
\]
the \( \epsilon \)-risk of estimate (3.54) does not exceed \( \rho \).

Indeed, with \( \rho \) and \( \kappa \) given by (3.56), \( h, H, \rho, \kappa \) satisfy (3.53).

### 3.4.1.4 Consistency

We are about to present a simple sufficient condition for the estimator defined in Proposition 3.12 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 and \( M(v) \succ 0 \), which allows to set \( \Theta_\ast = M(\bar{v}) \), to satisfy (3.45) with \( \delta = 0 \), and to assume w.l.o.g. that
\[
F(u,v) = [u;1]^TQ[u;1], \ G(Z) = \text{Tr}(QZ);
\]

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 (3.48), are linearly independent, so that we can find \( (m+1) \times (d+1) \) matrix \( C \) such that \( CB = I_{m+1} \). Let us define \((\bar{h}, \bar{H}) \in \mathbb{R}^d \times \mathbb{S}^d\) from the relation
\[
\begin{bmatrix}
\bar{H} \\
\bar{h}^T
\end{bmatrix} = 2(C^TQC)^o,
\]
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 3.14.** In the just described situation and under assumptions A.1-3, given \( \epsilon \in (0,1) \), consider the estimate
\[
\hat{g}_K(\zeta^K) = \frac{1}{K} \sum_{k=1}^{K} [\hat{h}^T\zeta_k + \frac{1}{2} \zeta^T\bar{H}\zeta_k] + \kappa_K,
\]
where
\[
\kappa_K = \frac{1}{2} \left[ \hat{\Psi}_{-,K}(\bar{h}, \bar{H}) - \hat{\Psi}_{+,K}(\bar{h}, \bar{H}) \right]
\]
and \( \hat{\Psi}_{\pm,K} \) are given by (3.52). Then the \( \epsilon \)-risk of \( \hat{g}_K(\cdot) \) goes to 0 as \( K \to \infty \).
3.4.1.5 A modification

In the situation described in the beginning of Section 3.4.1.2, let a set $W \subset U \times V$ be given, and assume we are interested in estimating the value of $F(u, v)$, as defined in (3.47), at points $(u, v) \in W$ only. When reducing the “domain of interest” $U \times V$ to $W$, we hopefully can reduce the attainable $\epsilon$-risk of recovery. Let us assume that we can point out a convex compact set $W \subset V \times Z$ such that

$$(u, v) \in W \Rightarrow (v, [u; 1][u; 1]^T) \in W$$

A straightforward inspection justifies the following

Remark 3.15. In the just described situation, the conclusion of Proposition 3.12 remains valid when the set $U \times V$ participating in (3.55) is reduced to $W$, and the set $V \times Z$ participating in relations (3.52) is reduced to $W$. This modification enlarges the feasible set of (3.53) and thus reduces the 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 as follows: we are given $K$ i.i.d. observations

$$\zeta_i \sim SG(A[u; 1], M(v)), \ i = 1, ..., K$$

(i.e., $\zeta_i$ are sub-Gaussian random vectors with parameters $A[u; 1] \in \mathbb{R}^d$ and $M(v) \in S_{+}^d$), where

- $(u, v)$ is unknown “signal” known to belong to a given set $U \times V$, where
  - $U \subset \mathbb{R}^m$ is a compact set, and
  - $V \subset \mathbb{R}^k$ is a compact convex set;
- $A$ is a given $d \times (m + 1)$ matrix, and $v \mapsto M(v) : \mathbb{R}^k \to \mathbb{S}^d$ is affine mapping such that $M(v) \succeq 0$ whenever $v \in V$.

We are also given a convex calibrating function $\varphi(Z) : S_{+}^{m+1} \to \mathbb{R}$ and “functional of interest”

$$F(u, v) = [u; 1]^T Q[u; 1] + q^T v,$$  \hspace{1cm} (3.59)

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 (3.58).

Note that the only difference of our present setting and that 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 and result

Let

$$V = \{M(v) : v \in V\},$$
so that $\mathcal{V}$ is a convex compact subset of the positive semidefinite cone $S^d_{+}$. Let us select somehow 
1. a matrix $\Theta_{*} > 0$ such that $\Theta_{*} \succeq \Theta_{*}$ for all $\Theta \in \mathcal{V}$; 
2. a convex compact subset $\mathcal{Z}$ of the set $\mathcal{Z}^+ = \{ Z \in S^{m+1}_+ : Z_{m+1,m+1} = 1 \}$ 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 3.11)

$$
\mathcal{H} = \mathcal{H}_\gamma := \{(h, H) \in \mathbb{R}^d \times S^d : -\gamma \Theta_{*}^{-1} \preceq H \preceq \gamma \Theta_{*}^{-1}\},
$$

$$
B = \begin{bmatrix} A \\ [0, ..., 0, 1] \end{bmatrix} \in \mathbb{R}^{(d+1) \times (m+1)},
$$

$$
\mathcal{M} = \mathcal{V} \times \mathcal{Z},
$$

$$
\Psi(h, H, G; Z) = -\frac{1}{2} \ln \det(I - \Theta_{1/2} G \Theta_{1/2}^T)
+ \frac{1}{2} \text{Tr} \left( \begin{bmatrix} H \\ h \end{bmatrix}^T \begin{bmatrix} H \\ h \end{bmatrix} + [H, h]^T (\Theta_{*}^{-1} - G)^{-1} [H, h] \right) B : (\mathcal{H} \times \{ G : 0 \preceq G \preceq \gamma^+ \Theta_{*}^{-1} \}) \times \mathcal{Z} \to \mathbb{R},
$$

where

$$
\Psi_\delta(h, H, G; Z) = -\frac{1}{2} \ln \det(I - \Theta_{1/2} G \Theta_{1/2}^T) + \frac{1}{2} \text{Tr} \left( \begin{bmatrix} H \\ h \end{bmatrix}^T \begin{bmatrix} H \\ h \end{bmatrix} + [H, h]^T (\Theta_{*}^{-1} - G)^{-1} [H, h] \right) B : (\mathcal{H} \times \{ G : 0 \preceq G \preceq \gamma^+ \Theta_{*}^{-1} \}) \times \mathcal{Z} \to \mathbb{R},
$$

$$
\Phi(h, H; Z) = \min_{G} \left\{ \Psi(h, H, G; Z) : 0 \preceq G \preceq \gamma^+ \Theta_{*}^{-1}, G \succeq H \right\} : \mathcal{H} \times \mathcal{Z} \to \mathbb{R},
$$

$$
\Phi_\delta(h, H; Z) = \min_{G} \left\{ \Psi_\delta(h, H, G; Z) : 0 \preceq G \preceq \gamma^+ \Theta_{*}^{-1}, G \succeq H \right\} : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}.
$$

The following statement is a straightforward reformulation of Proposition 2.46.i:

**Proposition 3.16.** In the situation described in Sections 3.4.2.1, 3.4.2.2 we have 
(i) $\Phi$ 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 $(h, H) \in \mathcal{H}, u \in U, v \in \mathcal{V}$, and let $\zeta \sim SG(A[u; 1], M(v))$. Then

$$
\ln \left( \mathbb{E}_\zeta \left\{ \exp \{ h^T \zeta + \frac{1}{2} \zeta^T H \zeta \} \right\} \right) \leq \Phi(h, H; [u; 1][u; 1]^T). \tag{3.61}
$$

(ii) Assume that

$$
\forall \Theta \in \mathcal{V} : \| \Theta_{1/2} \Theta_{*}^{-1/2} - I_d \| \leq \delta. \tag{3.62}
$$

Then $\Phi_\delta(h, H; \Theta, Z)$ is a well-defined real-valued continuous function on the domain $\mathcal{H} \times (\mathcal{V} \times \mathcal{Z})$; it 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 $(h, H) \in \mathcal{H}, u \in U, v \in \mathcal{V}$, and let $\zeta \sim SG(A[u; 1], M(v))$. Then

$$
\ln \left( \mathbb{E}_\zeta \left\{ \exp \{ h^T \zeta + \frac{1}{2} \zeta^T H \zeta \} \right\} \right) \leq \Phi_\delta(h, H; M(v), [u; 1][u; 1]^T). \tag{3.63}
$$

**The estimate.** We construction of the estimate is completely similar to the case of Gaussian observations. Specifically, let us pass from observations (3.58) to their
Proposition 3.17. In the situation described in Sections 3.4.2.1, 3.4.2.2 let us set because our Φ clearly satisfies (3.30)), we arrive at the result as follows:

The result.

Applying to the just specified data Corollary 3.8 (which is legitimate, introduced in Section 3.3.1 and participating in the constructions of Section 3.3 as follows:

- H = {f = (h, H) ∈ H} ⊂ E_H = R^d × S^d, with H defined in (3.60), and the inner product on E_H defined as

\[ \langle (h, H), (h', H') \rangle = h^T h' + \frac{1}{2} \text{Tr}(H H'), \]

\[ E_M = S^d × S^{m+1}, \text{and } M, Φ \text{ are as defined in (3.60)}; \]
- \[ E_X = R^k × S^{m+1}, X = V × Z; \]

\[ Σ(x = (v, Z)) = (M(v), Z); \] note that Σ is an affine mapping from E_X into E_M, mapping X into M, as required in Section 3.3. Observe that when u ∈ U and v ∈ V, the common distribution P = P_{u,v} of i.i.d. observations ω_1 defined by (3.64) satisfies the relation

\[ \ln \left( E_{x_0} \{ e^{f(ω)} \} \right) = \ln \left( E_{h,h/H} \{ e^{h^T ζ + 1/2 ζ^T H ζ} \} \right) \]

see (3.61). Moreover, in the case of (3.62), we have also

\[ \ln \left( E_{x_0} \{ e^{f(ω)} \} \right) = \ln \left( E_{h,h/H} \{ e^{h^T ζ + 1/2 ζ^T H ζ} \} \right) \]

see (3.63); 

- we set v(x = (v, Z)) = q(Z),
- we define affine functional G(x) on E_X by the relation

\[ G(x := (v, Z)) = q^T v + \text{Tr}(QZ), \]

see (3.59). As a result, for x = (v, [u; 1][u; 1]^T) with v ∈ V and u ∈ U we have

\[ F(u, v) = G(x). \]

The result. Applying to the just specified data Corollary 3.8 (which is legitimate, because our Φ clearly satisfies (3.30)), we arrive at the result as follows:

Proposition 3.17. In the situation described in Sections 3.4.2.1, 3.4.2.2 let us set

\[ \hat{Ψ}_{+, K}(h, H) := \inf_{α} \left\{ \max_{(v, Z) ∈ V \times Z} \left[ α Φ(h/α, H/α; Z) - G(v, Z) - q(Z) + α K^{-1} \ln(2/α) \right] \right\} \]

\[ = \max_{(v, Z) ∈ V \times Z} \inf_{α > 0, -α Θ^{-1} ≤ H ≤ α Θ^{-1}} \left[ α Φ(h/α, H/α; Z) - G(v, Z) - q(Z) + α K^{-1} \ln(2/α) \right] , \]
Under the premise of the proposition, let us fix $F$ of $s$ satisfies for all $(v, Z) \in V \times Z$:

$$\max_{(v, Z) \in V \times Z} \left[ \alpha \Phi(-h/\alpha, -H/\alpha; Z) + G(v, Z) - \varrho(Z) + \alpha K^{-1} \ln(2/\epsilon) \right] : \alpha > 0, -\gamma \alpha \Theta^{-1}_s \leq H \leq \gamma \alpha \Theta^{-1}_s$$

Thus, functions $\hat{\Psi}_{+, K}(h, H) : \mathbb{R}^d \times \mathbb{S}^d \to \mathbb{R}$ are convex. Furthermore, whenever $\hat{h}, \hat{H}, \hat{\rho}, \hat{\varpi}$ form a feasible solution to the system of convex constraints

$$\hat{\Psi}_{+, K}(h, H) \leq \rho - \varpi, \hat{\Psi}_{-, K}(h, H) \leq \rho + \varpi$$

in variables $(h, H) \in \mathbb{R}^d \times \mathbb{S}^d$, $\rho \in \mathbb{R}$, $\varpi \in \mathbb{R}$, the estimate

$$\hat{g}(\zeta^K) = \frac{1}{K} \sum_{i=1}^{K} \left[ h^T \zeta_i + \frac{1}{2} \zeta_i^T H \zeta_i \right] + \bar{\varpi},$$

of $F(u, v) = [u; 1]^T Q[u; 1] + q^T v$ via i.i.d. observations

$$\zeta_i \sim SG(A[u; 1], M(v)), 1 \leq i \leq K,$$

satisfies for all $(u, v) \in U \times V$:

$$\text{Prob}_{\zeta^K \sim [SG(A[u; 1], M(v))]^K} \left\{ |F(u, v) - \hat{g}(\zeta^K)| > \hat{\rho} + \varrho([u; 1][u; 1]^T) \right\} \leq \epsilon,$$

Proof. Under the premise of the proposition, let us fix $u \in U, v \in V$, and let $x = (v, Z) := [u; 1][u; 1]^T$. Denoting by $P$ the distribution of $\omega := (\zeta, \zeta^T)$ with $\zeta \sim SG(A[u; 1], M(v))$, and invoking (3.65), we see that for the just defined $(x, P)$ relation (3.26) takes place. Applying Corollary 3.8, we conclude that

$$\text{Prob}_{\zeta^K \sim [N(A[u; 1], M(v))]^K} \left\{ |\hat{g}(\zeta^K) - G(x)| > \hat{\rho} + \varrho([u; 1][u; 1]^T) \right\} \leq \epsilon.$$

It remains to note that by construction for $x = (v, Z)$ in question it holds

$$G(x) = q^T v + \text{Tr}(QZ) = q^T v + [u; 1]^T Q[u; 1] = F(u, v).$$

Remark 3.18. In the situation described in Sections 3.4.2.1, 3.4.2.2 let $\delta \in [0, 2]$ be such that

$$\|\Theta^{1/2} \Theta^{-1/2}_s - I_d\| \leq \delta \quad \forall \Theta \in \mathcal{V}.$$ 

Then the conclusion of Proposition 3.17 remains valid when the function $\Phi$ in (3.67) is replaced with the function $\Phi_8$, that is, when $\hat{\Psi}_{+, K}$ are defined as

$$\hat{\Psi}_{+, K}(h, H) := \inf_{\alpha} \left\{ \max_{(v, Z) \in V \times Z} \left[ \alpha \Phi_8(h/\alpha, H/\alpha; M(v), Z) - G(v, Z) - \varrho(Z) + \alpha K^{-1} \ln(2/\epsilon) \right] : \alpha > 0, -\gamma \alpha \Theta^{-1}_s \leq H \leq \gamma \alpha \Theta^{-1}_s \right\}$$

$$= \max_{(v, Z) \in V \times Z} \left\{ \inf_{\alpha > 0} \left[ \alpha \Phi_8(h/\alpha, H/\alpha; M(v), Z) - G(v, Z) - \varrho(Z) + \alpha K^{-1} \ln(2/\epsilon) \right] : -\gamma \alpha \Theta^{-1}_s \leq H \leq \gamma \alpha \Theta^{-1}_s \right\}.$$
In Gaussian case, \( \eta \). We lose nothing when restricting ourselves with estimates of the form of unknown signal \( u \in \mathbb{R}^m \) known to belong to a given set \( U \), we want to recover the “energy” \( u^T u \) of \( u \). We are interested in quadratic in \( \zeta = u + \xi \) of \( \zeta \) estimate of \( u^T U \) with as small as possible \( \epsilon \)-risk on \( U \); here \( \epsilon \in (0, 1) \) is a given design parameter. The details of our setup are as follows:

- \( U \) is the “spherical layer” \( U = \{ u \in \mathbb{R}^m : r^2 \leq u^T u \leq R^2 \} \), where \( r \) and \( R, 0 \leq r < R < \infty \), are given. As a result, the “main ingredient” of constructions from Sections 3.4.1.3, 3.4.2.2 – the convex compact subset \( Z \) of the set \( Z \in \mathbb{S}^{m+1}_r : Z_{m+1,m+1} = 1 \) containing all matrices \( [u; 1][u; 1]^T, u \in U \), can be specified as

\[
Z = \{ Z \in \mathbb{S}^{m+1}_r : Z_{m+1,m+1} = 1, 1 + r^2 \leq \text{Tr}(Z) \leq 1 + R^2 \};
\]

- \( \xi \) is either \( \mathcal{N}(0, \Theta) \) (Gaussian case), or \( \mathcal{SG}(0, \Theta) \) (sub-Gaussian case), with matrix \( \Theta \) known to be diagonal with equal to each other 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 \( g(Z) \) is \( g(Z) = \zeta(\sum_{i=1}^m Z_{ii}) \), where \( \zeta \) is a convex continuous real-valued function on \( \mathbb{R}_+ \). Note that with this selection, the claim that \( \epsilon \)-risk of an estimate \( \hat{g}(\cdot) \) is \( \leq \rho \) means that whenever \( u \in U \), one has

\[
\text{Prob}\{ |\hat{g}(u + \xi) - u^T u| > \rho + \zeta(u^T u) \} \leq \epsilon. \quad (3.69)
\]

**Processing the problem.** It is easily seen that in the situation in question the apparatus in Sections 3.4.1 and 3.4.2 translates into the following:

1. We lose nothing when restricting ourselves with estimates of the form

\[
\hat{g}(\zeta) = \frac{1}{\eta} \eta^T \zeta + \xi, \quad (3.70)
\]

with properly selected scalars \( \eta \) and \( \xi \);

2. In Gaussian case, \( \eta \) and \( \xi \) are yielded by the convex optimization problem with
only 3 variables $\alpha_+, \alpha_-$ and $\eta$, namely the problem
\[
\min_{\alpha_\pm, g_\pm, \eta} \left\{ \Psi_+(\alpha_+, g_+, \eta) + \Psi_-(\alpha_-, g_-, \eta) \right\} : \sigma^2 g_\pm < \alpha_\pm
\]
(3.72)
where
\[
\Psi_+(\alpha_+, \eta) = -\frac{d\alpha_+}{2} \ln(1 - \sigma^2 g_+/\alpha_+) \quad \text{and} \quad \Psi_-(\alpha_-, \eta) = -\frac{d\alpha_-}{2} \ln(1 - \sigma^2 g_-/\alpha_-)
\]
with $\delta = 1 - \sqrt{\theta}$. Now, the $\eta$-component of a feasible solution to (3.71) augmented by the quantity
\[
\kappa = \frac{1}{2} \left[ \Psi_-(\alpha_-, \eta) - \Psi_+(\alpha_+, \eta) \right]
\]
yields estimate (3.70) with $\epsilon$-risk on $U$ not exceeding $\Psi(\alpha_+, \alpha_-, \eta)$;
3. In sub-Gaussian case, $\eta$ and $\kappa$ are yielded by convex optimization problem with 5 variables, $\alpha_\pm, g_\pm$ and $\eta$, namely, the problem
\[
\min_{\alpha_\pm, g_\pm, \eta} \left\{ \Psi_+(\alpha_+, g_+, \eta) + \Psi_-(\alpha_-, g_-, \eta) \right\} : \\
0 \leq \sigma^2 g_\pm < \alpha_\pm, -\alpha_- < \sigma^2 \eta < \alpha_-, \eta \leq g_+, -\eta \leq g_-
\]
(3.72)
where
\[
\Psi_+(\alpha_+, \eta) = -\frac{d\alpha_-}{2} \ln(1 - \sigma^2 g_+/\alpha_+) \quad \text{and} \quad \Psi_-(\alpha_-, \eta) = -\frac{d\alpha_+}{2} \ln(1 - \sigma^2 g_-/\alpha_-)
\]
The $\eta$-component of a feasible solution to (3.72) augmented by the quantity
\[
\kappa = \frac{1}{2} \left[ \Psi_-(\alpha_-, g_-, \eta) - \Psi_+(\alpha_+, g_+, \eta) \right]
\]
yields estimate (3.70) with $\epsilon$-risk on $U$ not exceeding $\Psi(\alpha_+, g_\pm, \eta)$.

Note that the Gaussian case of our “energy estimation” problem is well studied in the literature (see, among others, [19, 43, 81, 86, 89, 96, 117, 122, 144, 156]), mainly in the case $\xi \sim \mathcal{N}(0, \sigma^2 I_m)$ of white Gaussian noise with exactly known variance $\sigma^2$. Available results investigate analytically the interplay between the dimension $m$ of signal, noise intensity $\sigma^2$ and the parameters $R, r$ and offer provably optimal, up to absolute constant factors, estimates. A nice property of the proposed approach is that (3.71) automatically takes care of the parameters and results in estimates with seemingly near-optimal performance, as witnessed by the numerical experiments we are about to present.

**Numerical results.** In the first series of experiments we use the trivial calibrating function: $\zeta(\cdot) \equiv 0$. 

"
Table 3.3: Estimating the signal energy from direct observations

<table>
<thead>
<tr>
<th>$d$</th>
<th>$r$</th>
<th>$R$</th>
<th>$\theta$</th>
<th>Relative 0.01-risk, Gaussian case</th>
<th>Relative 0.01-risk, sub-Gaussian case</th>
<th>Optimality ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>64</td>
<td>0</td>
<td>16</td>
<td>1</td>
<td>0.34808</td>
<td>0.44469</td>
<td>1.22</td>
</tr>
<tr>
<td>64</td>
<td>0</td>
<td>16</td>
<td>0.5</td>
<td>0.43113</td>
<td>0.44469</td>
<td>1.28</td>
</tr>
<tr>
<td>64</td>
<td>0</td>
<td>128</td>
<td>1</td>
<td>0.04962</td>
<td>0.05181</td>
<td>1.28</td>
</tr>
<tr>
<td>64</td>
<td>0</td>
<td>128</td>
<td>0.5</td>
<td>0.05064</td>
<td>0.05181</td>
<td>1.28</td>
</tr>
<tr>
<td>64</td>
<td>8</td>
<td>80</td>
<td>1</td>
<td>0.08287</td>
<td>0.08376</td>
<td>1.28</td>
</tr>
<tr>
<td>64</td>
<td>8</td>
<td>80</td>
<td>0.5</td>
<td>0.08095</td>
<td>0.08376</td>
<td>1.28</td>
</tr>
<tr>
<td>256</td>
<td>0</td>
<td>32</td>
<td>1</td>
<td>0.15003</td>
<td>0.30457</td>
<td>2.13</td>
</tr>
<tr>
<td>256</td>
<td>0</td>
<td>32</td>
<td>0.5</td>
<td>0.28511</td>
<td>0.30457</td>
<td>2.13</td>
</tr>
<tr>
<td>256</td>
<td>0</td>
<td>512</td>
<td>1</td>
<td>0.01264</td>
<td>0.01314</td>
<td>1.28</td>
</tr>
<tr>
<td>256</td>
<td>0</td>
<td>512</td>
<td>0.5</td>
<td>0.01289</td>
<td>0.01314</td>
<td>1.28</td>
</tr>
<tr>
<td>256</td>
<td>16</td>
<td>160</td>
<td>1</td>
<td>0.03996</td>
<td>0.04501</td>
<td>1.28</td>
</tr>
<tr>
<td>256</td>
<td>16</td>
<td>160</td>
<td>0.5</td>
<td>0.04257</td>
<td>0.04501</td>
<td>1.28</td>
</tr>
<tr>
<td>1024</td>
<td>0</td>
<td>64</td>
<td>1</td>
<td>0.10242</td>
<td>0.21923</td>
<td>2.13</td>
</tr>
<tr>
<td>1024</td>
<td>0</td>
<td>64</td>
<td>0.5</td>
<td>0.10742</td>
<td>0.21923</td>
<td>2.13</td>
</tr>
<tr>
<td>1024</td>
<td>0</td>
<td>2048</td>
<td>1</td>
<td>0.00317</td>
<td>0.00330</td>
<td>1.28</td>
</tr>
<tr>
<td>1024</td>
<td>0</td>
<td>2048</td>
<td>0.5</td>
<td>0.00324</td>
<td>0.00330</td>
<td>1.28</td>
</tr>
<tr>
<td>1024</td>
<td>32</td>
<td>320</td>
<td>1</td>
<td>0.02273</td>
<td>0.02516</td>
<td>1.28</td>
</tr>
<tr>
<td>1024</td>
<td>32</td>
<td>320</td>
<td>0.5</td>
<td>0.02273</td>
<td>0.02516</td>
<td>1.28</td>
</tr>
</tbody>
</table>

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 of the estimates, that is, the plain risk as given by (3.71) divided by $R^2$; keeping this in mind, one will not be surprised that when extending the range $[r, R]$ of allowed 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 $\sigma$ is 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 upper risk bounds given by (3.71) in the Gaussian case, to (lower bounds on) the best possible under circumstances 0.01-risks $\text{Risk}_{0.01}^\epsilon$, defined as the infimum of the 0.01-risk over all estimates recovering $\|u\|_2$ via single observation $\omega = u + \zeta$. These lower bounds are obtained as follows. Let us select somehow values $r_1 < r_2$ in the allowed range $[r, R]$ of $\|u\|_2$, along with two values, $\sigma_1, \sigma_2$, in the allowed range $[\theta\sigma, \sigma] = [\theta, 1]$ of values of diagonal entries in diagonal matrices $\Theta$, and consider two distributions of observations $P_1$ and $P_2$ as follows: $P_\chi$ is the distribution of the random vector $x + \zeta$, where $x$ and $\xi$ 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 (the sum, over the two hypotheses in question, of probabilities for the test to reject the hypothesis when it is true) $\leq 2\epsilon$, the quantity $\delta = \frac{1}{2}(r_2^2 - r_1^2)$ is a lower bound on the optimal $\epsilon$-risk, $\text{Risk}_\epsilon^\star$. In other words, denoting by $p_\chi(\cdot)$ the density of $P_\chi$, we have

$$0.02 < \int_{\mathbb{R}^d} \min[p_1(\omega), p_2(\omega)]d\omega \Rightarrow \text{Risk}_{0.01}^\epsilon \geq \frac{1}{2}(r_2^2 - r_1^2).$$

Now, the densities $p_\chi$ are spherically symmetric, whence, denoting by $q_\chi(\cdot)$ the
univariate density of the energy \( \omega^T \omega \) of observation \( \omega \sim P_x \), we have

\[
\int_{\mathbb{R}^d} \min[p_1(\omega), p_2(\omega)]d\omega = \int_0^\infty \min[q_1(s), q_2(s)]ds,
\]

so that

\[
0.02 < \int_0^\infty \min[q_1(s), q_2(s)]ds \Rightarrow \text{Risk}^*_{0.01} \geq \frac{1}{2}(r^2_2 - r^2_1). \quad (3.73)
\]

On closer inspection, \( q_2 \) is the convolution of two univariate densities representable by explicit computation-friendly formulas, implying that given \( r_1, r_2, \sigma_1 \) and \( \sigma_2 \), we can check numerically whether the premise in (3.73) indeed takes place, and whenever the latter is the case, the quantity \( (r^2_2 - r^2_1)/2 \) is a lower bound on \( \text{Risk}^*_{0.01} \).

In our experiments, we implement a simple search strategy (not described here) aimed at crude maximizing this bound in \( r_1, r_2, \sigma_1 \) and \( \sigma_2 \) and use the resulting lower bounds on \( \text{Risk}^*_{0.01} \) to compute the optimality ratios presented in the table.\(^{10}\)

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 0.01-risk, the corresponding to the trivial calibrating function, in recovery \( u^T u \) 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 rather high “actual” risk. For example, when \( d := \dim u = 1024, \theta = 1 \), and \( U = \{ u \in \mathbb{R}^d : \|u\|_2 \leq 1.0e6 \} \), the 0.01-risk becomes as large as \( \rho \approx 6.5e6 \). For “relatively small” signals, like \( u^T u \approx 10^4 \), recovering \( u^T u \) within accuracy \( \rho \) if of no interest. In order to allow for “large” domains \( U \) it makes sense to pass from the trivial calibrating function to a nontrivial one, e.g., \( \zeta(t) = \alpha t \), with small positive \( \alpha \). With this calibrating function, (3.69) reads

\[
\text{Prob} \{ |\hat{g}(u + \xi) - u^T u| > \rho + \alpha u^T u \} \leq \epsilon.
\]

It turns out that (quite reasonable when \( U \) is large) “relative” risk quantification results in risk values essentially smaller than those of “absolute” risk. Here is some instructive numerical data:

<table>
<thead>
<tr>
<th>( r )</th>
<th>( R )</th>
<th>0.01-Risk, ( \alpha = 0 )</th>
<th>0.01-Risk, ( \alpha = 0.01 )</th>
<th>0.01-Risk, ( \alpha = 0.1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1.0e7</td>
<td>6.51e7/6.51e7</td>
<td>1.33e3/1.58e3</td>
<td>474/642</td>
</tr>
<tr>
<td>1.0e2</td>
<td>1.0e7</td>
<td>6.51e7/6.51e7</td>
<td>1.33e3/1.58e3</td>
<td>-123/92.3</td>
</tr>
<tr>
<td>1.1e3</td>
<td>1.0e7</td>
<td>6.51e7/6.51e7</td>
<td>-4.73e3/-4.48e3</td>
<td>-1.14e5/-1.14e5</td>
</tr>
</tbody>
</table>

\( U = \{ u \in \mathbb{R}^{1024} : r \leq \|u\|_2 \leq R \}, \theta = 1/2 \)

Left/Right: risks in Gaussian/sub-Gaussian cases

### 3.4.2.4 Numerical illustration, indirect observations

**The problem.** Let us consider the estimation problem as follows. Our observations are

\[
\zeta = Bu + \xi,
\]

\(^{10}\)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 on the grid 1.05\(^i\), \( i = 1, 2, \ldots \).
where

- \( B \) is a given \( d \times m \) matrix, with \( m > d \) ("deficient observations"),
- \( u \in \mathbb{R}^m \) is a signal known to belong to a compact set \( U \),
- \( \xi \sim \mathcal{N}(0, \Theta) \) (Gaussian case) or \( \xi \sim \mathcal{SG}(0, \Theta) \) (sub-Gaussian case) is the observation noise; \( \Theta \) is positive semidefinite \( d \times d \) matrix known to belong to a given convex compact set \( \mathcal{V} \subset \mathbb{S}_d^+ \).

Our goal is to estimate the energy

\[
F(u) = \frac{1}{n} \| u \|_2^2
\]

of the signal given observation (3.74).

In our experiment, the data is specified as follows:

1. We think of \( u \in \mathbb{R}^m \) as of discretization of a smooth function \( x(t) \) of continuous argument \( t \in [0; 1] \): \( u_i = x(t_i), 1 \leq i \leq m \). 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 analog 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 \( B \) 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 \( \theta^{-\frac{1}{d-1}}, 1 \leq i \leq d \).
3. The set \( V \) of admissible matrices \( \Theta \) is the set of all diagonal \( d \times d \) matrices with diagonal entries varying in \([0, \sigma^2]\).

Both \( \sigma \) and \( \theta \) are components of the experiment setup.

**Processing the problem.** The described estimation 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 \( \Theta_\sigma \) as \( \sigma^2 I_d \), \( M(\nu) \) as the identity mapping of \( V \) onto itself, the mapping \( u \mapsto A[u; 1] \) as the mapping \( u \mapsto Bu \), and the set \( Z \) (which should be a convex compact subset of the set \( \{ Z \in \mathbb{S}_d^{d+1} : Z_{d+1,d+1} = 0 \} \)) containing all matrices of the form \( [u; 1][u; 1]^T, u \in U \) as the set

\[
Z = \{ Z \in \mathbb{S}_d^{d+1} : Z_{d+1,d+1} = 1, \text{Tr} (Z \text{Diag}\{S^T S, 0\}) \leq 1 \}.
\]

As suggested by Propositions 3.12 (Gaussian case) and 3.17 (sub-Gaussian case), the linear in "lifted observation" \( \omega = (\zeta, \zeta^T) \) estimates of \( F(u) = \frac{1}{n} \| u \|_2^2 \) stem from the optimal solution \( (h_*, H_*) \) to the convex optimization problem

\[
\text{Opt} = \min_{h,H} \frac{1}{2} \left[ \hat{\Psi}_+(h, H) + \hat{\Psi}_-(h, H) \right], \tag{3.75}
\]

with \( \hat{\Psi}_\pm(\cdot) \) given by (3.52) in the Gaussian, and by (3.67) in the sub-Gaussian cases, with the number \( K \) of observations in (3.52) and (3.67) set to 1. The resulting estimate is

\[
\zeta \mapsto h_*^T \zeta + \frac{1}{2} \zeta^T H_* \zeta + \varkappa, \quad \varkappa = \frac{1}{2} \left[ \hat{\Psi}_-(h_*, H_*) - \hat{\Psi}_+(h_*, H_*) \right] \tag{3.76}
\]

and the \( \varepsilon \)-risk of the estimate is (upper-bounded by) \( \text{Opt} \).

Problem (3.75) 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 (or maximization) problems. However, applying conic duality, one can easily eliminate in (3.52), (3.67) the inner maxima over $v, Z$ to end up with reformulation which can be solved numerically by CVX [106], and this is how we process (3.75) in our experiments.

**Numerical results.** In the experiments to be reported, we use the trivial calibrating function: $g(\cdot) \equiv 0$.

We present some typical numerical results in Table 3.4. To qualify the performance of our approach, we present, along with the upper risk bounds for the computed estimates, simple lower bounds on $\epsilon$-risk. The origin of the lower bounds is as follows. Assume we have at our disposal a signal $w \in U$, and let $t(w) = \|Bw\|_2$, $\rho = 2\sigma \text{ErfcInv}(\epsilon)$, where ErfcInv is the inverse error function as defined in (1.26). Setting $\theta(w) = \max[1 - \rho/t(w), 0]$, observe that $w' := \theta(w)w \in U$ and $\|Bw - Bw'\|_2 \leq \rho$, which, due to the origin of $\rho$, implies that there is no way to decide via observation $Bu + \xi$, $\xi \sim \mathcal{N}(0, \sigma^2)$, with risk < $\epsilon$ on the two simple hypotheses $u = w$ and $u = w'$. As an immediate consequence, the quantity $\phi(w) := \frac{1}{2}\|w\|_2^2 - \frac{1}{2}\|w'\|_2^2 = \|w\|_2^2(1 - \theta^2(w))/2$ is a lower bound on the $\epsilon$-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 risk bound

$$\text{LwBnd} = \max_{w \in U} \left\{ \frac{1}{2}\|w\|_2^2(1 - \theta^2(w)) \right\}.$$ 

On closer 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 are built (we omit the details). We see that the $\epsilon$-risks of our estimates are within a moderate factor of the optimal ones.

Figure 3.4 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 $\mathcal{N}(0, \sigma^2 I_d)$ noise and selected signals in $U$ by maximizing over $U$ randomly selected linear forms. Finally, we note that already with fixed design parameters $d, m, \theta$ and $\sigma$ we deal with a family of estimation problems rather than with a single problem, the reason being that our $U$ is an ellipsoid with essentially different from each other half-axes. In this situation, attainable 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. Note also that the “sub-Gaussian experiments” were conducted on exactly the same data as “Gaussian experiments” of the same sizes $d, m$.

<table>
<thead>
<tr>
<th>$d, m$</th>
<th>Opt, Gaussian case</th>
<th>Opt, sub-Gaussian case</th>
<th>LwBnd</th>
</tr>
</thead>
<tbody>
<tr>
<td>8, 12</td>
<td>0.1362(+65%)</td>
<td>0.1382(+67%)</td>
<td>0.0825</td>
</tr>
<tr>
<td>16, 24</td>
<td>0.1614(+53%)</td>
<td>0.1640(+55%)</td>
<td>0.1058</td>
</tr>
<tr>
<td>32, 48</td>
<td>0.0687(+46%)</td>
<td>0.0692(+48%)</td>
<td>0.0469</td>
</tr>
</tbody>
</table>

Table 3.4: Upper bound (Opt) on the 0.01-risk of estimate (3.76), (3.75) vs. lower bound (LwBnd) on 0.01-risk attainable under the circumstances. In the experiments, $\sigma = 0.025$ and $\theta = 10$. Data in parentheses: excess of Opt over LwBnd.
3.5 EXERCISES FOR CHAPTER 3

Exercise 3.1.

In the situation of Section 3.3.4, design of a “good” estimate is reduced to solving convex optimization problem (3.39). Note that the objective in this problem is, in a sense, “implicit” – the design variable is $h$, and the objective is obtained from an explicit convex-concave function of $h$ and $(x, y)$ by maximization over $(x, y)$. There exist solvers capable to process problems of this type efficiently. However, commonly used off-the-shelf solvers, like cvx, cannot handle problems of this type. The goal of the exercise to follow is to reformulate (3.39) 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 $X \subset \mathbb{R}^n$ along with affine function $M(x)$ taking values in $\mathbb{S}^d$ and such that $M(x) \succeq 0$ when $x \in X$, and
- affine function $F(h) : \mathbb{R}^d \rightarrow \mathbb{R}^n$.

Given $\gamma > 0$, this data gives rise to the convex function

$$\Psi(h) = \max_{x \in X} \left\{ F^T(h)x + \gamma \sqrt{h^T M(x)h} \right\},$$

and we want to find a “nice” representation of this function, specifically, want to represent the inequality $\tau \geq \Psi(h)$ by a bunch of LMI’s in variables $\tau, h$, 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 \( \xi \), namely,
\[
X^+ = \left\{ (x, M) : \begin{array}{ll}
(a) & s_i - a_i^T x - b_i^T \xi - \text{Tr}(C_i M) \geq 0, \quad i \leq I \\
(b) & S - A(x) - B(\xi) - C(M) \succeq 0 \\
(c) & M \succeq 0 
\end{array} \right\}.
\]
Here \( s_i \in \mathbb{R}, \ S \in \mathbb{S}^N \) are some constants, and \( A(\cdot), B(\cdot), C(\cdot) \) are (homogeneous) linear functions taking values in \( \mathbb{S}^N \). 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).

Here goes the exercise:

1) Check that \( \Psi(h) \) is the optimal value in the semidefinite program
\[
\Psi(h) = \max_{x, M, \xi, t} \left\{ \begin{array}{ll}
F^T(h)x + \gamma t : \\
\{ s_i - a_i^T x - b_i^T \xi - \text{Tr}(C_i M) \geq 0, \quad i \leq I \\
S - A(x) - B(\xi) - C(M) \succeq 0 \\
M \succeq 0 \\
[\begin{array}{c}
h^T M h \\ t \\
1 \end{array}] \succeq 0 \right\}.
\]

2) Passing from \((P)\) to the semidefinite dual of \((P)\), build explicit semidefinite representation of \( \Psi \), that is, an explicit system \( S \) of LMI’s in variables \( h, \tau \) and additional variables \( u \) such that
\[
\{ \tau \geq \Psi(h) \} \iff \exists u : (\tau, h, u) \text{ satisfies } S. \]

Exercise 3.2.

Let us consider the situation as follows. Given an \( m \times n \) “sensing matrix” \( A \) which is stochastic – with columns from the probabilistic simplex
\[
\Delta_m = \left\{ v \in \mathbb{R}^m : v \geq 0, \sum_i v_i = 1 \right\}
\]
and a nonempty closed subset \( U \) of \( \Delta_n \), we observe \( M \)-element, \( M > 1 \), i.i.d. sample \( \zeta^M = (\zeta_1, ..., \zeta_M) \) with \( \zeta_k \) drawn from the discrete distribution \( Au_\ast \), where \( u_\ast \) is an unknown probabilistic vector (“signal”) known to belong to \( U \). We handle the discrete distribution \( Au \), \( u \in \Delta_n \), as a distribution on the vertices \( e_1, ..., e_m \) of \( \Delta_m \), so that possible values of \( \zeta_k \) are basic orths \( e_1, ..., e_m \) in \( \mathbb{R}^m \). Our goal is to recover the value \( F(u_\ast) \) of a given quadratic form
\[
F(u) = u^T Qu + 2q^T u.
\]
Observe that for \( u \in \Delta_n \), we have \( u = [uu^T]1_n \), where \( 1_n \) is the all-ones vector in \( \mathbb{R}^n \). This observation allows to rewrite \( F(u) \) as a homogeneous quadratic form:
\[
F(u) = u^T Q u, \quad Q = Q + [q1_n^T + 1_n q^T]. \quad (3.77)
\]
The goal of the exercise is to follow the approach developed in Section 3.4.1 for the Gaussian case in order to build an estimate \( \hat{g}(\xi^M) \) of \( F(u) \). To this end, consider the following construction.

Let

\[ J_M = \{(i, j): 1 \leq i < j \leq M\}, J_M = \text{Card}(J_M). \]

For \( \xi^M = (\xi_1, \ldots, \xi_M) \) with \( \xi_k \in \{e_1, \ldots, e_m\}, 1 \leq k \leq M \), let

\[ \omega_{ij}[\xi^M] = \frac{1}{2}[\xi_i^T \xi_j^T + \xi_j^T \xi_i^T], (i, j) \in J_M. \]

The estimates we are interested in are of the form

\[ \hat{g}(\xi^M) = \text{Tr}\left( h \left[ \frac{1}{n} \sum_{(i,j) \in J_M} \omega_{ij}[\xi^M] \right] \right) + \kappa \]

where \( h \in S^m \) and \( \kappa \in \mathbb{R} \) are the parameters of the estimate.

Now goes the exercise:

1) Verify that when \( \xi_k \)'s stem from signal \( u \in U \), the expectation of \( \omega[\xi^M] \) is a linear image \( A z[u] A^T \) of the matrix \( z[u] = uu^T \in S^n \): denoting by \( P^M_u \) the distribution of \( \xi^M \), we have

\[ \mathbb{E}_{\xi^M \sim P^M_u} \{ \omega[\xi^M] \} = A z[u] A^T. \quad (3.78) \]

Check that when setting

\[ Z_k = \{ \omega \in S^k : \omega \geq 0, 0 \leq \omega \leq 0, 1_k^T \omega 1_k = 1 \}, \]

where \( x \geq 0 \) for a matrix \( x \) means that \( x \) is entrywise nonnegative, the image of \( Z_k \) under the mapping \( z \mapsto A z A^T \) is contained in \( Z_m \).

2) Let \( \Delta^k = \{ z \in S^k : z \geq 0, 1_n^T z 1_n = 1 \} \), so that \( Z_k \) is the set of all positive semidefinite matrices from \( \Delta^k \). For \( \mu \in \Delta^m \), let \( P_\mu \) be the distribution of the random matrix \( w \) taking values in \( S^m \) 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 \( \mu_{ii} \), and for every \( i, j \) with \( i < j \), \( w \) takes value \( e^{ij} \) with probability \( 2\mu_{ij} \). Let us set

\[ \Phi_1(h; \mu) = \ln \left( \sum_{i,j=1}^m \mu_{ij} \exp\{h_{ij}\} \right) : S^m \times \Delta^m \rightarrow \mathbb{R}, \]

so that \( \Phi_1 \) is a continuous convex-concave function on \( S^m \times \Delta^m \).

2.1. Prove that

\[ \forall (h \in S^m, \mu \in Z_m) : \ln \left( \mathbb{E}_{w \sim P_\mu} \{ \exp(\text{Tr}(hw)) \} \right) = \Phi_1(h; \mu). \]

2.2. Derive from 2.1 that setting

\[ K = K(M) = \lfloor M/2 \rfloor, \quad \Phi_M(h; \mu) = K \Phi_1(h/K; \mu) : S^m \times \Delta^m \rightarrow \mathbb{R}, \]
Consider an alternative way to estimate \( F \rho \) of the functional of interest. Let \( P_{u,M} \) be the distribution of \( \omega = \omega_1[\zeta^M], \zeta^M \sim P^M_u \). Then for all \( u \in U, h \in S^m \),

\[
\ln \left( \mathbb{E}_{\omega \sim P_{u,M}} \{ \exp\{ \text{Tr}(h\omega) \} \} \right) \leq F(h; A[z|u]A^T), \ z[u] = uu^T. \tag{3.79}
\]

3) Combine the above observations with Corollary 3.6 to arrive at the following result:

**Proposition 3.19.** In the situation in question, let \( Z \) be a convex compact subset of \( Z_n \) such that \( uu^T \in Z \) for all \( u \in U \). Given \( \epsilon \in (0, 1) \), let

\[
\Psi_+(h, \alpha) = \max_{z \in Z} \left[ \alpha \Phi_M(h/\alpha, AzA^T) - \text{Tr}(Qz) \right] : S^m \times \{ \alpha > 0 \} \to \mathbb{R},
\]

\[
\Psi_-(h, \alpha) = \max_{z \in Z} \left[ \alpha \Phi_M(-h/\alpha, AzA^T) + \text{Tr}(Qz) \right] : S^m \times \{ \alpha > 0 \} \to \mathbb{R},
\]

\[
\tilde{\Psi}_+(h) := \inf_{\alpha > 0} \left[ \Psi_+(h, \alpha) + \alpha \ln(2/\epsilon) \right] = \max_{z \in Z, \alpha > 0} \left[ \alpha \Phi_M(h/\alpha, AzA^T) - \text{Tr}(Qz) + \alpha \ln(2/\epsilon) \right] \quad [\beta = K\alpha],
\]

\[
\tilde{\Psi}_-(h) := \inf_{\alpha > 0} \left[ \Psi_-(h, \alpha) + \alpha \ln(2/\epsilon) \right] = \max_{z \in Z, \alpha > 0} \left[ \alpha \Phi_M(-h/\alpha, AzA^T) + \text{Tr}(Qz) + \alpha \ln(2/\epsilon) \right] \quad [\beta = K\alpha].
\]

The functions \( \tilde{\Psi}_\pm \) are real valued and convex on \( S^m \), and every candidate solution \( h \) to the convex optimization problem

\[
\text{Opt} = \min_h \left\{ \tilde{\Psi}(h) := \frac{1}{2} \left[ \tilde{\Psi}_+(h) + \tilde{\Psi}_-(h) \right] \right\}, \tag{3.80}
\]

induces the estimate

\[
\bar{g}_h(\zeta^M) = \text{Tr}(h\omega_1[\zeta^M]) + \kappa(h), \ \kappa(h) = \frac{1}{2} [\tilde{\Psi}_-(h) - \tilde{\Psi}_+(h)]
\]

of the functional of interest (3.77) via observation \( \zeta^M \) with \( \epsilon \)-risk on \( U \) not exceeding \( \rho = \tilde{\Psi}(h) \):

\[
\forall (u \in U) : \text{Prob}_{\zeta^M \sim P^M_u} \{|F(u) - \bar{g}_h(\zeta^M)| > \rho \} \leq \epsilon.
\]

4) Consider an alternative way to estimate \( F(u) \), namely, as follows. Let \( u \in U \). Given a pair of independent observations \( \zeta_1, \zeta_2 \) 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 \( \mu(z) = AzA^T : \Delta^u \to \Delta^m \). 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 (3.77) is a linear function \( \text{Tr}(Qz|u) \) of \( z[u] \). Assume that we are given a set \( Z \) as in the premise of Proposition 3.19. Observe that we are in the situation as follows:

Given \( K \) i.i.d. observations \( \omega^K = (\omega_1, ..., \omega_K) \) with \( \omega_k \sim P_{\mu(z)}, \) where \( z \)
is unknown signal known to belong to $Z$, we want to recover the value at $z$ of linear function $G(v) = \text{Tr}(\bar{Q}v)$ of $v \in S^n$. Besides this, we know that $P_{\mu}$, for every $\mu \in \Delta^m$, satisfies the relation

$$\forall (h \in S^m) : \ln \left( \mathbb{E}_{\omega \sim P_{\mu}} \{ \exp\{ \text{Tr}(h\omega) \} \} \right) \leq \Phi_1(h; \mu).$$

This situation fits the setting of Section 3.3.3, with the data specified as $H = \mathcal{E}_H = S^m, \mathcal{M} = \Delta^m \subset \mathcal{E}_M = S^m, \Phi = \Phi_1, \mathcal{X} := Z \subset \mathcal{E}_X = S^n, A(z) = AzA^T$.

Therefore, we can use the apparatus developed in that section to upper-bound $\epsilon$-risk of the affine estimate $\text{Tr} \left( h \frac{1}{K} \sum_{k=1}^{K} \omega_k \right) + \kappa$ of $F(u) := G(z[u]) = u^T \bar{Q} u$ and to build the best, in terms of the upper risk bound, estimate, see Corollary 3.8. On closer inspection (carry it out!), the associated with the above data functions $\hat{\Psi}_\pm$ arising in (3.38) are exactly the functions $\hat{\Psi}_\pm$ specified in Proposition 3.19 for $M = 2K$. Thus, the just outlined approach to estimating $F(u)$ via stemming from $u \in U$ observations $\zeta^{2K}$ results in a family of estimates

$$\tilde{g}_h(\zeta^{2K}) = \text{Tr} \left( h \frac{1}{K} \sum_{k=1}^{K} \omega_k \left[ \zeta_{2k-1}; \zeta_{2k} \right] \right) + \kappa(h), h \in S^m.$$

The resulting upper bound on the $\epsilon$-risk of estimate $\tilde{g}_h$ is $\hat{\Psi}(h)$, where $\hat{\Psi}(\cdot)$ is associated with $M = 2K$ according to Proposition 3.19. In other words, this is exactly the upper bound on the $\epsilon$-risk of the estimate $\hat{g}_h$ offered by the proposition. Note, however, that the estimates $\tilde{g}_h$ and $\hat{g}_h$ are not identical:

$$\tilde{g}_h(\zeta^{2K}) = \text{Tr} \left( h \frac{1}{K} \sum_{k=1}^{K} \omega_{2k-1,2k} \left[ \zeta^{2K} \right] \right) + \kappa(h),$$

$$\hat{g}_h(\zeta^{2K}) = \text{Tr} \left( h \frac{1}{K(2K-1)} \sum_{1 \leq i<j \leq 2K} \omega_{ij} \left[ \zeta^{2K} \right] \right) + \kappa(h).$$

Now goes the question:

- Which of the estimates $\tilde{g}_h$ and $\hat{g}_h$ would you prefer? That is, which one of these estimates, in your opinion, exhibits better practical performance?

To check your intuition, compare the estimate performance by simulation. Consider the following story underlying the recommended simulation model:

“Tomorrow, tomorrow not today, all the lazy people say.” Is it profitable to be lazy? Imagine you are supposed to carry out a job, and should decide whether to do it today or tomorrow. The reward for the job is drawn at random “by nature,” 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 would be better – to complete the job today or tomorrow? In other words, is the probability that tomorrow reward is at least that of today greater than 0.5? What is this probability? How to estimate it from historical data?
State the above problem as that 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 of the estimates \( \hat{g}_h, \tilde{g}_h \) works better. To avoid the necessity to solve optimization problem (3.80), you can use \( h = \bar{Q} \), resulting in unbiased estimate of \( u^T \bar{Q} u \).

**Exercise 3.3.**

What follows is a variation of Exercise 3.2. Consider the situation as follows. We observe \( K \) realizations \( \eta_k, k \leq K \), of discrete random variable with \( p \) possible values, and \( L \geq K \) realizations \( \zeta_\ell, \ell \leq L \), of discrete random variable with \( q \) possible values. All realizations are independent of each other; \( \eta_k \)'s are drawn from distribution \( P_u \), and \( \zeta_\ell \) – from distribution \( Q_v \), where \( P \in \mathbb{R}^{p \times r}, Q \in \mathbb{R}^{q \times s} \) are given stochastic “sensing matrices,” and \( u, v \) are unknown “signals” known to belong to given subsets \( U, \text{ resp.}, V \) of probabilistic simplexes \( \Delta_r, \text{ resp.}, \Delta_s \). Our goal is to recover from observations \( \{ \eta_k, \zeta_\ell \} \) the value at \( u, v \) of a given bilinear function

\[
F(u, v) = u^T F v = \text{Tr}(F[u v^T]^T). \tag{3.81}
\]

A “covering story” could be as follows. Imagine that there are two possible actions, say, administering to a patient drug A and drug B. Let \( u \) be 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 the outcome of the first action to be better than the outcome of the second. This amounts to computing the probability \( \pi \) of the event “\( \eta > \zeta \),” where \( \eta, \zeta \) are independent of each other discrete real-valued random variables with distributions \( u, \text{ resp.}, v \), and \( \pi \) is a linear function of the “joint distribution” \( uv^T \) of \( \eta, \zeta \). 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 – we arrive at problems where \( P \) and \( Q \) can be general type stochastic matrices.

As always, we encode the \( p \) possible values of \( \eta_k \) by the basic orths \( e_1, \ldots, e_p \) in \( \mathbb{R}^p \), and the \( q \) possible values of \( \zeta_\ell \) by the basic orths \( f_1, \ldots, f_q \) in \( \mathbb{R}^q \).

We focus on estimates of the form

\[
\tilde{g}_{h, \kappa} (\eta^K, \zeta^L) = \left[ \frac{1}{K} \sum_k \eta_k \right]^T h \left[ \frac{1}{L} \sum_\ell \zeta_\ell \right] + \kappa \quad [h \in \mathbb{R}^{p \times q}, \kappa \in \mathbb{R}].
\]

This is what you are supposed to do:

1) (cf. item 2 in Exercise 3.2) Denoting by \( \Delta_{mn} \) the set of nonnegative \( m \times n \) matrices with unit sum of all entries (i.e., the set of all probability distributions on \( \{1, \ldots, m\} \times \{1, \ldots, n\} \)) and assuming \( L \geq K \), let us set

\[
\mathcal{A}(z) = PzQ^T : \mathbb{R}^{r \times s} \rightarrow \mathbb{R}^{p \times q}
\]
2) Combine (!) with Corollary 3.6 to arrive at the following analog of Proposition 3.19: 

\[ \Phi(h; \mu) = \ln \left( \sum_{i=1}^{\nu} \sum_{j=1}^{\omega} \mu_{ij} \exp(h_{ij}) \right); \quad R^{p \times q} \times \Delta_{pq} \to R, \]

\[ \Phi_K(h; \mu) = K\Phi(h/K; \mu); \quad R^{p \times q} \times \Delta_{pq} \to R. \]

Verify that \( A \) maps \( \Delta_{rs} \) into \( \Delta_{pq} \). \( \Phi \) and \( \Phi_K \) 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 \( \eta^K = (\eta_1, ..., \eta_K), \eta^L = (\zeta_1, ..., \zeta_L) \) with mutually independent \( \eta_1, ..., \zeta_k \) such that \( \eta_k \sim Pu, \eta_\ell \sim Qv \) for all \( k, \ell \), we have

\[ \ln \left( E_{\eta, \zeta} \left\{ \exp \left[ \left( \frac{1}{K} \sum_k \eta_k \right)^T h \left( \frac{1}{L} \sum_\ell \zeta_\ell \right) \right] \right\} \right) \leq \Phi_K(h; A(uv^T)). \quad (3.82) \]

2) Combine (!) with Corollary 3.6 to arrive at the following analog of Proposition 3.19:

**Proposition 3.20.** In the situation in question, let \( \mathcal{Z} \) be a convex compact subset of \( \Delta_{rs} \) such that \( uv^T \in \mathcal{Z} \) for all \( u \in U, v \in V \). Given \( \epsilon \in (0, 1) \), let

\[
\Psi_+(h, \alpha) = \max_{z \in \mathcal{Z}} \left[ \alpha \Phi_K(h/\alpha, PzQ^T) - \text{Tr}(Fz^T) \right]; \quad R^{p \times q} \times \{ \alpha > 0 \} \to R, \\
\Psi_-(h, \alpha) = \max_{z \in \mathcal{Z}} \left[ \alpha \Phi_K(-h/\alpha, PzQ^T) + \text{Tr}(Fz^T) \right]; \quad R^{p \times q} \times \{ \alpha > 0 \} \to R, \\
\hat{\Psi}_+(h) := \inf_{\alpha > 0} \left[ \Psi_+(h, \alpha) + \alpha \ln(2/\epsilon) \right] \\
= \max_{z \in \mathcal{Z}} \left[ \alpha \Phi_K(h/\alpha, PzQ^T) - \text{Tr}(Fz^T) + \alpha \ln(2/\epsilon) \right] \\
= \max_{z \in \mathcal{Z}} \left[ \alpha \Phi_K(-h/\alpha, PzQ^T) + \text{Tr}(Fz^T) + \alpha \ln(2/\epsilon) \right] \quad \beta = K\alpha, \\
\hat{\Psi}_-(h) := \inf_{\alpha > 0} \left[ \Psi_-(h, \alpha) + \alpha \ln(2/\epsilon) \right] \\
= \max_{z \in \mathcal{Z}} \left[ \alpha \Phi_K(h/\alpha, PzQ^T) + \text{Tr}(Fz^T) + \alpha \ln(2/\epsilon) \right] \\
= \max_{z \in \mathcal{Z}} \left[ \alpha \Phi_K(-h/\alpha, PzQ^T) - \text{Tr}(Fz^T) + \alpha \ln(2/\epsilon) \right] \quad \beta = K\alpha. 
\]

The functions \( \hat{\Psi}_\pm \) are real valued and convex on \( R^{p \times q} \), and every candidate solution \( h \) to the convex optimization problem

\[
\text{Opt} = \min_h \left\{ \hat{\Psi}(h) := \frac{1}{2} \left[ \hat{\Psi}_+(h) + \hat{\Psi}_-(h) \right] \right\},
\]

induces the estimate

\[
\hat{\theta}_h(\eta^K, \zeta^L) = \text{Tr} \left( h \left[ \left( \frac{1}{K} \sum_k \eta_k \right) \left( \frac{1}{L} \sum_\ell \zeta_\ell \right) \right]^T \right) + \kappa(h), \quad \kappa(h) = \frac{1}{2} \left[ \hat{\Psi}_-(h) - \hat{\Psi}_+(h) \right] 
\]

of the functional of interest (3.81) via observation \( \eta^K, \zeta^L \) with \( \epsilon \)-risk on \( U \times V \) not exceeding \( \rho = \hat{\Psi}(h) \):

\[
\forall (u \in U, v \in V): \text{Prob}\{|F(u,v) - \hat{\theta}_h(\eta^K, \zeta^L)| > \rho\} \leq \epsilon,
\]

the probability being taken w.r.t. the distribution of observations \( \eta^K, \zeta^L \) stemming from signals \( u, v \).
Exercise 3.4.
[recovering mixture weights] The problem to be addressed in this exercise is as follows. We are given \( K \) probability distributions \( P_1, \ldots, P_K \) on observation space \( \Omega \), and let these distributions have densities \( p_k(\cdot) \) w.r.t. some reference measure \( \Pi \) on \( \Omega \); we assume that \( \sum_k p_k(\cdot) \) is positive on \( \Omega \). We are given also \( N \) independent observations \( \omega_t \sim P_{\mu}, t = 1, \ldots, N, \) drawn from distribution \( P_{\mu} = \sum_{k=1}^K \mu_k P_k, \) where \( \mu \) is unknown “signal” known to belong to the probabilistic simplex \( \Delta_K = \{ \mu \in \mathbb{R}^K : \mu \geq 0, \sum_k \mu_k = 1 \} \). Given \( \omega^N = (\omega_1, \ldots, \omega_N) \), we want to recover the linear image \( G\mu \) of \( \mu \), where \( G \in \mathbb{R}^{\nu \times K} \) is given.

We intend to measure the risk of a candidate estimate \( \hat{G}(\omega^N) : \Omega \times \ldots \times \Omega \to \mathbb{R}^{\nu} \) by the quantity

\[
\text{Risk}[\hat{G}(\cdot)] = \sup_{\mu \in \Delta} \left[ \mathbb{E}_{\omega^N \sim P_{\mu} \times \ldots \times P_{\mu}} \left\{ \| \hat{G}(\omega^N) - G\mu \|^2_2 \right\} \right]^{1/2}
\]

3.4.A. Recovering linear form. Let us start with the case when \( G = g^T \) is a 1 \( \times \) \( K \) matrix.

3.4.A.1. Preliminaries. To motivate the construction to follow, consider the case when \( \Omega \) 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.: given stationary \( N \)-repeated observation stemming from discrete probability distribution \( P_{\mu} \) affinely parameterized by signal \( \mu \in \Delta_K \), we want to recover a linear form of \( \mu \). It is shown in Section 3.1, see Remark 3.2, that in this case a nearly optimal in terms of its \( \epsilon \)-risk estimate is of the form

\[
\hat{\mu}(\omega^N) = \frac{1}{N} \sum_{t=1}^N \phi(\omega_t)
\]

(3.83)

with properly selected \( \phi \). The difficulty with this approach is that as far as computations are concerned, optimal design of \( \phi \) requires solving a convex optimization problem of design dimension of order of the cardinality of \( \Omega \), and this cardinality could be huge, as is the case when \( \Omega \) is a discretization of a domain in \( \mathbb{R}^d \) with \( d \) is in the range of tens. To circumvent this problem, we are to simplify the outlined approach: from the construction of Section 3.1 we inherit the simple structure (3.83) of the estimator; taking this structure for granted, we are to develop an alternative design of \( \phi \). With this new design, 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.

3.4.A.2. Generic estimate. Let us select somehow \( L \) functions \( F_\ell(\cdot) \) on \( \Omega \) such that

\[
\int F_\ell^2(\omega)p_k(\omega)\Pi(d\omega) < \infty, \quad 1 \leq \ell \leq L, 1 \leq k \leq K
\]

(3.84)
With $\lambda \in \mathbb{R}^K$, consider estimate of the form

$$
\hat{g}_\lambda (\omega^N) = \frac{1}{N} \sum_{t=1}^{N} \Phi_\lambda (\omega_t), \quad \Phi_\lambda (\omega) = \sum_{\ell} \lambda_\ell F_\ell (\omega).
$$

(3.85)

1) Prove that

$$
\text{Risk}[\hat{g}_\lambda] \leq \hat{\text{Risk}}(\lambda) := \max_{k \leq K} \left[ \frac{1}{N} \left( \int [\sum_{\ell} \lambda_\ell F_\ell (\omega)]^2 p_k (\omega) \Pi (d\omega) \right) \right]^{1/2} + \left[ \int \left( \sum_{\ell} \lambda_\ell F_\ell (\omega) \right) p_k (\omega) \Pi (d\omega) - g^T e_k \right]^2 \right]^{1/2},
$$

(3.86)

where

$$
M = \left[ M_{k\ell} := \int F_\ell (\omega) p_k (\omega) \Pi (d\omega) \right]_{k \leq K, \ell \leq L},
$$

$$
W_k = \left[ [W_k]_{\ell\ell'} := \int F_\ell (\omega) F_{\ell'} (\omega) p_k (\omega) \Pi (d\omega) \right]_{\ell \leq L, 1 \leq k \leq K},
$$

and $e_1, ..., e_K$ are the standard basic orths in $\mathbb{R}^K$. Note that $\hat{\text{Risk}}(\lambda)$ is a convex function of $\lambda$; 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

$$
\text{Opt} = \min_{\lambda \in \mathbb{R}^K} \hat{\text{Risk}}(\lambda)
$$

(3.87)

and use the estimate (3.85) associated with optimal solution to this problem; the risk of this estimate will be upper-bounded by Opt.

3.4.A.3. Implementation. When implementing the generic estimate we arrive at the “Measurement Design” question: how to select the value of $L$ and functions $F_\ell, 1 \leq \ell \leq L$ resulting in small (upper bound Opt on the) risk of the estimate (3.85) yielded by an optimal solution to (3.87)? 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 (3.84). For the sake of definiteness, consider the “Gaussian case,” where $\Omega = \mathbb{R}^d, \Pi$ is the Lebesgue measure, and $p_k$ is Gaussian distribution with parameters $\nu_k, \Sigma_k$:

$$
p_k (\omega) = \frac{1}{(2\pi)^{d/2} \sqrt{\text{Det}(\Sigma_k)}} \exp \left\{ -\frac{1}{2} (\omega - \nu_k)^T \Sigma_k^{-1} (\omega - \nu_k) \right\}.
$$

In this case, the Naive option leads to easily computable matrices $M$ and $W_k$ appearing in (3.86).

2) Check that in the Gaussian case, when setting

$$
\Sigma_{k\ell} = [\Sigma_k^{-1} + \Sigma_\ell^{-1}]^{-1}, \quad \Sigma_{k\ell m} = [\Sigma_k^{-1} + \Sigma_\ell^{-1} + \Sigma_m^{-1}]^{-1}, \quad \chi_k = \Sigma_k^{-1} \nu_k,
$$

$$
\alpha_{k\ell} = \sqrt{\frac{\text{Det}(\Sigma_{k\ell})}{(2\pi)^d \text{Det}(\Sigma_k) \text{Det}(\Sigma_\ell)}}, \quad \beta_{k\ell m} = (2\pi)^{-d/2} \sqrt{\frac{\text{Det}(\Sigma_{k\ell m})}{\text{Det}(\Sigma_k) \text{Det}(\Sigma_\ell) \text{Det}(\Sigma_m)}},
$$

...
we have
\[ M_{kt} := \int p_\ell(\omega)p_k(\omega)\Pi(d\omega) \]
\[ = \int \alpha_{kt} \exp \{ \frac{1}{2} \left[ \left[ \chi_k + \chi_\ell \right]^T \Sigma_{kt} [\chi_k + \chi_\ell] - \chi_k^T \Sigma_k \chi_k - \chi_\ell^T \Sigma_\ell \chi_\ell \right] \}, \]
\[ [W_{k}]_{\ell m} := \int p_\ell(\omega)p_m(\omega)p_k(\omega)\Pi(d\omega) \]
\[ = \int \beta_{k\ell m} \exp \{ \frac{1}{2} \left[ \left[ \chi_k + \chi_\ell + \chi_m \right]^T \Sigma_{k\ell m} [\chi_k + \chi_\ell + \chi_m] - \chi_k^T \Sigma_k \chi_k - \chi_\ell^T \Sigma_\ell \chi_\ell - \chi_m^T \Sigma_m \chi_m \right] \}.

**Basic option.** Though simple, Naive option does not make much sense: when replacing the reference measure \( \Pi \) with another measure \( \Pi' \) which has positive density \( \theta(\cdot) \) w.r.t. \( \Pi \), the densities \( p_k \) are updated according to \( p_k(\cdot) \mapsto p_k'(\cdot) = \theta^{-1}(\cdot)p(\cdot) \), so that selecting \( F_\ell' = p_\ell' \), the matrices \( M \) and \( W_k \) become \( M' \) and \( W'_k \) with
\[ M'_{k\ell} = \int p_k(\omega)p_\ell(\omega)\Pi'(d\omega) = \int \frac{p_k(\omega)p_\ell(\omega)}{\theta(\omega)}\Pi(d\omega), \]
\[ [W'_k]_{\ell m} = \int p_k(\omega)p_\ell(\omega)p_m(\omega)\Pi'(d\omega) = \int \frac{p_k(\omega)p_\ell(\omega)p_m(\omega)}{\theta(\omega)}\Pi(d\omega). \]

We see that in general \( M \neq M' \) and \( W_k \neq W'_k \), which makes the Naive option rather unnatural. In the alternative Basic option we set
\[ L = K, \quad F_\ell(\omega) = \pi(\omega) := \frac{p_\ell(\omega)}{\sum_k p_k(\omega)}. \]

The motivation is that the functions \( F_\ell \) are invariant when replacing \( \Pi \) with \( \Pi' \), so that here \( M = M' \) and \( W_k = W'_k \). Besides this, there are statistical arguments in favor of Basic option, namely, as follows. Let \( \Pi_* \) be the measure with the density \( \sum_k p_k(\cdot) \) w.r.t. \( \Pi \); taken w.r.t. \( \Pi_* \), the densities of \( p_k \) are exactly the above \( \pi_k(\cdot) \), and \( \sum_k \pi_k(\omega) = 1 \). Now, (3.86) says that the risk of estimate \( \hat{\theta}_\lambda \) can be upper-bounded by the function \( \text{Risk}(\lambda) \) defined in (3.86), and this function, in turn, can be upper-bounded by the function
\[ \text{Risk}^+(\lambda) := \left[ \frac{1}{N} \sum_k \int \left[ \sum_\ell \lambda_\ell F_\ell(\omega) \right]^2 p_k(\omega)\Pi(d\omega) \right]^{1/2} \]
\[ + \max_k \left[ \int \left[ \sum_\ell \lambda_\ell F_\ell(\omega) \right] p_k(\omega)\Pi(d\omega) - g^T e_k \right]^2 \]
\[ \leq K \text{Risk}(\lambda) \]
(we have said that the maximum of \( K \) nonnegative quantities is at most their sum, and the latter is at most \( K \) times the maximum of the quantities). Consequently, the risk of the estimate (3.85) stemming from an optimal solution to (3.87) can be upper-bounded by the quantity
\[ \text{Opt}^+ := \min_\lambda \text{Risk}^+(\lambda) \quad [\geq \text{Opt} := \max_\lambda \text{Risk}(\lambda)]. \]

And here goes the punchline:

3.1) Prove that both the quantities \( \text{Opt} \) defined in (3.87) and the above \( \text{Opt}^+ \) depend
only on the linear span of the functions $F_\ell$, $\ell = 1, \ldots, L$, not on how the functions $F_\ell$ are selected in this span.

3.2) Prove that the selection $F_\ell = \pi_\ell$, $1 \leq \ell \leq L = K$, minimizes $\text{Opt}^+$ among all possible selections $L$, \{ $F_\ell$ \}_1^{L}$ satisfying (3.84).

Conclude that the selection $F_\ell = \pi_\ell$, $1 \leq \ell \leq L = K$, while not necessary optimal in terms of $\text{Opt}$, definitely is meaningful: this selection optimizes the natural upper bound $\text{Opt}^+$ on $\text{Opt}$. Observe that $\text{Opt}^+ \leq K \text{Opt}$, so that optimizing instead of $\text{Opt}$ the upper bound $\text{Opt}^+$, although rough, 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 (3.86). For example, in the Gaussian case, Naive choice of $F_\ell$’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 $P_1, \ldots, P_K$. 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. $P_1, \ldots, P_K$, of bounded in magnitude by $1$, and thus well-suited for Monte-Carlo simulation, functions of $\omega$.

**Maximum Likelihood option.** This choice of \{ $F_\ell$ \}_1^{L}$ follows straightforwardly the idea of discretization we started with in this exercise. Specifically, we split $\Omega$ into $L$ cells $\Omega_1, \ldots, \Omega_L$ in such a way that the intersection of any two different cells is of $\Pi$-measure zero, and treat as our observations not the actual observations $\omega_t$, but the indexes of the cells $\omega_t$’s belong to. With our estimation scheme, this is the same as to select $F_\ell$ as the characteristic function of $\Omega_\ell$, $\ell \leq L = K$. Assuming that for distinct $k$, $k'$ the densities $p_k$, $p_{k'}$ differ from each other $\Pi$-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 $F_\ell$’s, the characteristic functions of the just defined sets $\Omega_\ell$, $1 \leq \ell \leq L = K$. 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 $F_\ell$’s. In applications, one can compute the upper risk bounds $\text{Opt}$, see (3.87), associated with each option, and to use the option with the best – the smallest – risk bound (“smart” choice of $F_\ell$’s). 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 $F_\ell$’s, the smaller is the associated $\text{Opt}$, so that the only price for combining different selections is in increasing the computational cost of solving (3.87).

**3.4.A.4. Illustration.** In the experimental part of this exercise your are expected to

4.1) Run numerical experiments to compare the estimates yielded by the above three options (Naive, Basic, ML). Recommended setup:

- $d = 8$, $K = 90$;
• Gaussian case with the covariance matrices $\Sigma_k$ of $P_k$ selected at random:

$$S_k = \text{rand}(d, d), \quad \Sigma_k = \frac{S_k S_k^T}{\|S_k\|^2}$$

and the expectations $\nu_k$ of $P_k$ selected at random from $\mathcal{N}(0, \sigma^2 I_d)$, with $\sigma = 0.1$;

• values of $N$: $\{10^3, s = 0, 1, \ldots, 5\}$;

• linear form to be recovered: $g^T \mu = \mu_1$.

4.21. Utilize Cramer-Rao lower risk bound (see Proposition 4.37, Exercise 4.22) to upper-bound the level of conservatism $\frac{\text{Opt}}{\text{Risk}}$ of the estimates built in item 4.1. Here Risk, is the minimax risk in our estimation problem:

$$\text{Risk}_\ast = \inf_{\hat{g}(\cdot)} \text{Risk}[\hat{g}(\omega^N)] = \inf_{\hat{g}(\cdot)} \sup_{\omega \in \Delta} \left\{\mathbb{E}_{\omega^N \sim P_\mu \times \ldots \times P_\mu} \left\{\|\hat{g}(\omega^N) - g^T \mu\|^2\right\}\right\}^{1/2},$$

where inf is taken over all estimates.

3.4.B. Recovering linear images. Now consider the case when $G$ is a general $\nu \times K$ matrix. The analog of the estimate $\hat{g}_\lambda(\cdot)$ is now as follows: with somehow chosen $F_1, \ldots, F_L$ satisfying (3.84), we select a $\nu \times L$ matrix $\Lambda = [\lambda_{i\ell}]$, set

$$\Phi_\Lambda(\omega) = \left[\sum_{\ell} \lambda_{1\ell} F_\ell(\omega); \sum_{\ell} \lambda_{2\ell} F_\ell(\omega); \ldots; \sum_{\ell} \lambda_{KL} F_\ell(\omega)\right]$$

and estimate $G\mu$ by

$$\hat{G}_\Lambda(\omega^N) = \frac{1}{N} \sum_{i=1}^N \Phi_\Lambda(\omega_i).$$

5) Prove the following counterpart of the results of item 3.4.A:

**Proposition 3.21.** The risk of the proposed estimator can be upper-bounded as follows:

$$\text{Risk}[\hat{G}_\Lambda] := \max_{\mu \in \Delta_K} \left\{\mathbb{E}_{\omega^N \sim P_\mu \times \ldots \times P_\mu} \left\{\|\hat{G}(\omega^N) - G\mu\|^2\right\}\right\}^{1/2},$$

$$\ast \leq \max_{\Lambda \leq K} \Psi(\Lambda, \mu),$$

$$\Psi(\Lambda, \mu) = \left\{\frac{1}{N} \sum_{k=1}^K \mu_k \mathbb{E}_{\omega \sim P_k} \left\{\|\Phi_\Lambda(\omega)\|^2\right\} + \left[\|\hat{G}_\Lambda(\omega^N) - G\mu\|^2\right]\right\}^{1/2},$$

where

$$\text{Col}_k[\psi_\Lambda] = \mathbb{E}_{\omega \sim P_k(\cdot)} \Phi_\Lambda(\omega) = \left[\int [\sum_{\ell} \lambda_{1\ell} F_\ell(\omega)] P_k(d\omega) \right], \ldots, 1 \leq k \leq K,$$

and $e_1, \ldots, e_K$ are the standard basic orths in $\mathbb{R}^K$.

Note that exactly the same reasoning as in the case of the scalar $G\mu \equiv g^T \mu$ demonstrates that a reasonable way to select $L$ and $F_\ell$, $\ell = 1, \ldots, L$, is to set $L = K$ and $F_\ell(\cdot) = \pi_\ell(\cdot), 1 \leq \ell \leq L$. 

3.6 PROOFS

3.6.1 Proof of Proposition 3.3

Observe that $\text{Opt}_{ij}(K)$ is the saddle point value in the convex-concave saddle point problem:

$$\text{Opt}_{ij}(K) = \inf_{\alpha > 0, \phi \in \mathcal{F}} \max_{x \in X_i, \nu \in \mathcal{X}_j} \left[ \frac{1}{2} K \alpha \left[ \Phi(\phi/\alpha; A_i(x)) + \Phi(-\phi/\alpha; A_j(y)) \right] + \frac{1}{2} g^T[y - x] + \alpha \ln(2I/\epsilon) \right].$$

The domain of the maximization variable is compact and the cost function is continuous on its domain, whence, by Sion-Kakutani Theorem, we also have

$$\text{Opt}_{ij}(K) = \max_{x \in X_i, \nu \in \mathcal{X}_j} \Theta_{ij}(x, y),$$

$$\Theta_{ij}(x, y) = \inf_{\alpha > 0, \phi \in \mathcal{F}} \left[ \frac{1}{2} K \alpha \left[ \Phi(\phi/\alpha; A_i(x)) + \Phi(-\phi/\alpha; A_j(y)) \right] + \frac{1}{2} g^T[y - x] \right] + \alpha \ln(2I/\epsilon).$$

Note that

$$\Theta_{ij}(x, y) = \inf_{\alpha > 0, \phi \in \mathcal{F}} \left[ \frac{1}{2} K \alpha \left[ \Phi(\phi/\alpha; A_i(x)) + \Phi(-\phi/\alpha; A_j(y)) \right] + \frac{1}{2} g^T[y - x] \right] = \frac{1}{2} \alpha K \inf_{\psi \in \mathcal{F}} \left[ \Phi(\psi/\alpha; A_i(x)) + \Phi(-\psi/\alpha; A_j(y)) \right] + \alpha \ln(2I/\epsilon).$$

Given $x \in X_i, y \in X_j$ and setting $\mu = A_i(x), \nu = A_j(y)$, we obtain

$$\inf_{\psi \in \mathcal{F}} \left[ \Phi(\psi/\alpha; A_i(x)) + \Phi(-\psi/\alpha; A_j(y)) \right] = \inf_{\psi \in \mathcal{F}} \left[ \ln \left( \int \exp(\psi(\omega))p_\mu(\omega)\Pi(\omega) \right) + \ln \left( \int \exp(-\psi(\omega))p_\nu(\omega)\Pi(\omega) \right) \right].$$

Since $\mathcal{O}$ is a good o.s., the function $\bar{\psi}(\omega) = \frac{1}{2} \ln(p_\mu(\omega)/p_\nu(\omega))$ belongs to $\mathcal{F}$, and

$$\inf_{\psi \in \mathcal{F}} \left[ \ln \left( \int \exp(\psi(\omega))p_\mu(\omega)\Pi(\omega) \right) + \ln \left( \int \exp(-\psi(\omega))p_\nu(\omega)\Pi(\omega) \right) \right] = \inf_{\delta \in \mathcal{F}} \left[ \ln \left( \int \exp(\delta(\omega))\sqrt{p_\mu(\omega)p_\nu(\omega)}\Pi(\omega) \right) + \ln \left( \int \exp(-\delta(\omega))\sqrt{p_\mu(\omega)p_\nu(\omega)}\Pi(\omega) \right) \right] = \inf_{\delta \in \mathcal{F}} \left[ \ln \left( \int \exp(-\delta(\omega))\sqrt{p_\mu(\omega)p_\nu(\omega)}\Pi(\omega) \right) + \ln \left( \int \exp(-\delta(\omega))\sqrt{p_\mu(\omega)p_\nu(\omega)}\Pi(\omega) \right) \right].$$

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

$$\inf_{\psi \in \mathcal{F}} \left[ \Phi(\psi/\alpha; A_i(x)) + \Phi(-\psi/\alpha; A_j(y)) \right] = 2 \ln \left( \int \sqrt{p_{A_i(x)}(\omega)p_{A_j(y)}(\omega)}\Pi(\omega) \right),$$

(3.89)
and
\[
\Theta_{ij}(x, y) = \inf_{\alpha > 0} \left\{ K \ln \left( \int \sqrt{p_{A_i(x)}(\omega)p_{A_j(y)}(\omega)} \Pi(d\omega) \right) + \ln(2I/\epsilon) \right\} + \frac{1}{2}g^T[y - x]
\]
\[
= \begin{cases} \frac{1}{2}g^T[y - x], & K \ln \left( \int \sqrt{p_{A_i(x)}(\omega)p_{A_j(y)}(\omega)} \Pi(d\omega) \right) + \ln(2I/\epsilon) \geq 0, \\ -\infty, & \text{otherwise.} \end{cases}
\]

This combines with (3.88) to imply that
\[
\text{Opt}_{ij}(K) = \max_{x,y} \left\{ \frac{1}{2}g^T[y - x] : x \in X_i, y \in X_j, \left[ \int \sqrt{p_{A_i(x)}(\omega)p_{A_j(y)}(\omega)} \Pi(d\omega) \right]^K \geq \frac{\epsilon}{2I} \right\},
\]
(3.90)

\(2^\circ\). We claim that under the premise of the proposition, for all \(i, j, 1 \leq i, j \leq I\), one has
\[
\text{Opt}_{ij}(K) \leq \text{Risk}^*_i(K),
\]
implying the validity of (3.13). Indeed, assume that for some pair \(i, j\) the opposite inequality holds true:
\[
\text{Opt}_{ij}(K) > \text{Risk}^*_i(K),
\]
and let us lead this assumption to a contradiction. Under our assumption optimization problem in (3.90) has a feasible solution \((\bar{x}, \bar{y})\) such that
\[
r := \frac{1}{2}g^T[\bar{y} - \bar{x}] > \text{Risk}^*_i(K),
\]
(3.91)
implying, due to the origin of \(\text{Risk}^*_i(K)\), that there exists an estimate \(\bar{g}(\omega^K)\) such that for \(\mu = A_i(\bar{x}), \nu = A_j(\bar{y})\) it holds
\[
\text{Prob}_{\omega \sim \pi \sim p^\omega} \left\{ \bar{g}(\omega^K) \leq \frac{1}{2}g^T[\bar{x} + \bar{y}] \right\} \leq \text{Prob}_{\omega \sim \pi \sim p^\omega} \left\{ \bar{g}(\omega^K) - g^T[\bar{y}] \geq r \right\} \leq \epsilon
\]
\[
\text{Prob}_{\omega \sim \pi \sim p^\omega} \left\{ \bar{g}(\omega^K) \geq \frac{1}{2}g^T[\bar{x} + \bar{y}] \right\} \leq \text{Prob}_{\omega \sim \pi \sim p^\omega} \left\{ |\bar{g}(\omega^K) - g^T[\bar{x}]| \geq r \right\} \leq \epsilon.
\]
In other words, we can decide on two simple hypotheses stating that observation \(\omega^K\) obeys distribution \(p^n(\omega^K)\), resp., \(p^n(\omega^K)\), with risk \(\leq \epsilon\). Consequently, setting \(\Pi^K = \Pi \times \ldots \times \Pi\) and \(p^K(\omega^K) = \prod_{k=1}^K p(\omega_k)\), we have
\[
\int \min \left[ p^K(\omega^K), p^K(\omega^K) \right] \Pi^K(d\omega^K) \leq 2\epsilon.
\]
Hence,
\[
\left[ \int \sqrt{p^n(\omega)p^n(\omega) \Pi(d\omega)} \right]^K = \left[ \int \sqrt{p^K(\omega^K)p^K(\omega^K) \Pi^K(d\omega^K)} \right] = \left[ \int \min \left[ p^K(\omega^K), p^K(\omega^K) \right] \Pi^K(d\omega^K) \right]^1/2 \left[ \max \left[ p^K(\omega^K), p^K(\omega^K) \right] \Pi^K(d\omega^K) \right]^{1/2}
\]
\[
\leq \left( \int \min \left[ p^K(\omega^K), p^K(\omega^K) \right] \Pi^K(d\omega^K) \right)^{1/2} \left( \int \max \left[ p^K(\omega^K), p^K(\omega^K) \right] \Pi^K(d\omega^K) \right)^{1/2}
\]
\[
\leq \left( \int \left[ p^K(\omega^K) + p^K(\omega^K) - \min \left[ p^K(\omega^K), p^K(\omega^K) \right] \right] \Pi^K(d\omega^K) \right)^{1/2}
\]
\[
\leq 2\sqrt{\varepsilon(1 - \varepsilon)}.\]
Therefore, for $K$ satisfying \eqref{eq:3.12} we have
\[
\left[ \int \sqrt{\rho_{\mu}(\omega)p_{\nu}(\omega)} \Pi(d\omega) \right]^K \leq [2\sqrt{\epsilon(1-\epsilon)}]^{K/R} < \frac{\epsilon}{2I},
\]
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 \eqref{eq:3.90}).

3º. Now let us prove that under the premise of the proposition, \eqref{eq:3.14} takes place. To this end, let us set
\[
w_{ij}(s) = \max_{x \in X_i, y \in X_j} \left\{ \frac{1}{2} g^T [y - x] : K \ln \left( \int \sqrt{\rho_{A_i(x)}(\omega)p_{A_j(y)}(\omega)} \Pi(d\omega) \right) + s \geq 0 \right\}.
\]
As we have seen in item 1º, see \eqref{eq:3.89}, one has
\[
H(x, y) = \inf_{\psi \in \mathcal{F}} \frac{1}{2} [\Phi_{\psi}(A_i(x)) + \Phi_{\psi}(A_j(y))],
\]
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 \eqref{eq:3.92} is feasible whenever $s \geq 0$, a feasible solution being $y = x = x_{ij}$. At this feasible solution we have $\psi^T [y - x] = 0$, 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
\[
w_{ij}(\bar{s}) \leq \text{Risk}^*_\epsilon(K), \quad \bar{s} = -\ln(2\sqrt{\epsilon(1-\epsilon)}).
\]
Indeed, $w_{ij}(s)$ is nonnegative, concave and bounded (since $X_i, X_j$ are compact) on $\mathbb{R}_+$, implying that $w_{ij}(s)$ is continuous on $\{s > 0\}$. Assuming, on the contrary to what we need to prove, that $w_{ij}(\bar{s}) > \text{Risk}^*_\epsilon(K)$, there exists $s' \in (0, \bar{s})$ such that $w_{ij}(s') > \text{Risk}^*_\epsilon(K)$ 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 \eqref{eq:3.91} takes place. We have seen in item 2º that the latter relation implies that for $\mu = A_i(\bar{x})$, $\nu = A_j(\bar{y})$ it holds
\[
\left[ \int \sqrt{\rho_{\mu}(\omega)p_{\nu}(\omega)} \Pi(d\omega) \right]^K \leq 2\sqrt{\epsilon(1-\epsilon)};
\]
that is,
\[
K \ln \left( \int \sqrt{\rho_{\mu}(\omega)p_{\nu}(\omega)} \Pi(d\omega) \right) + \bar{s} \leq 0.
\]
Hence,
\[
K \ln \left( \int \sqrt{\rho_{\mu}(\omega)p_{\nu}(\omega)} \Pi(d\omega) \right) + s' < 0,
\]
contradicting the feasibility of $(\bar{x}, \bar{y})$ to the optimization problem specifying $w_{ij}(s')$. It remains to note that \eqref{eq:3.93} combines with concavity of $w_{ij}(\cdot)$ and the relation $w_{ij}(0) \geq 0$ to imply that
\[
w_{ij}(\ln(2I/\epsilon)) \leq \partial w_{ij}(\bar{s}) \leq \partial \text{Risk}^*_\epsilon(K)
\]
where
\[ \vartheta = \ln(2l/e)/s = \frac{2\ln(2l/e)}{\ln([4e(1-e)]^{-1})}. \]

Invoking (3.90), we conclude that
\[ \text{Opt}_{ij}(K) = w_{ij}(\ln(2l/e)) \leq \vartheta \text{Risk}_n^*(K) \forall i,j. \]

Finally, from (3.90) it immediately follows that Opt_{ij}(K) is nonincreasing in K (as K grows, the feasible set of the optimization problem in (3.90) shrinks), so that for \( K \geq K \) we have
\[ \text{Opt}(K) \leq \text{Opt}(K) = \max_{i,j} \text{Opt}_{ij}(K) \leq \vartheta \text{Risk}_n^*(K), \]
and (3.14) follows. \( \square \)

### 3.6.2 Verifying 1-convexity of the conditional quantile

Let r be a nonvanishing probability distribution on S, and let
\[ F_m(r) = \sum_{i=1}^{m} r_i, 1 \leq m \leq M, \]
so that \( 0 < F_1(r) < F_2(r) < \ldots < F_M(r) = 1 \). Denoting by \( \mathcal{P} \) the set of all nonvanishing probability distributions on S, observe that for every \( p \in \mathcal{P} \) \( \chi_\alpha[r] \) is a piecewise-linear function of \( \alpha \in [0,1] \) with breakpoints \( 0, F_1(r), F_2(r), F_3(r), \ldots, F_M(r) \), the values of the function at these breakpoints being \( s_1, s_2, s_3, \ldots, s_M \). In particular, this function is equal to \( s_1 \) on \([0,F_1(r)]\) and is strictly increasing on \([F_1(r),1]\).

Now let \( s \in \mathbb{R} \), and let
\[ \mathcal{P}_\alpha^< [s] = \{ r \in \mathcal{P} : \chi_\alpha[r] \leq s \}, \quad \mathcal{P}_\alpha^> [s] = \{ r \in \mathcal{P} : \chi_\alpha[r] > s \}. \]
Observe that the just introduced sets are cut off \( \mathcal{P} \) by nonstrict linear inequalities, specifically,

- when \( s < s_1 \), we have \( \mathcal{P}_\alpha^< [s] = \emptyset \), \( \mathcal{P}_\alpha^> [s] = \mathcal{P} \);
- when \( s = s_1 \), we have \( \mathcal{P}_\alpha^< [s] = \{ r \in \mathcal{P} : F_1(r) \geq \alpha \}, \mathcal{P}_\alpha^> [s] = \mathcal{P} \);
- when \( s > s_M \), we have \( \mathcal{P}_\alpha^< [s] = \mathcal{P}, \mathcal{P}_\alpha^> [s] = \emptyset \);
- when \( s_1 < s \leq s_M \), for every \( r \in \mathcal{P} \) the equation \( \chi_\gamma[r] = s \) in variable \( \gamma \in [0,1] \) has exactly one solution \( \gamma(r) \) which can be found as follows: we specify \( k = k^s \in \{1,...,M-1\} \) such that \( s_k < s \leq s_{k+1} \) and set
\[ \gamma(r) = \frac{(s_{k+1} - s)F_k(r) + (s - s_k)F_{k+1}(r)}{s_{k+1} - s_k}. \]

Since \( \chi_\alpha[r] \) is strictly increasing in \( \alpha \) when \( \alpha \in [F_1(p),1] \), for \( s \in (s_1,s_M] \) we have
\[ \mathcal{P}_\alpha^< [s] = \{ r \in \mathcal{P} : \alpha \leq \gamma(r) \} = \left\{ r \in \mathcal{P} : \frac{(s_{k+1} - s)F_k(r) + (s - s_k)F_{k+1}(r)}{s_{k+1} - s_k} \geq \alpha \right\}, \]
\[ \mathcal{P}_\alpha^> [s] = \{ r \in \mathcal{P} : \alpha > \gamma(r) \} = \left\{ r \in \mathcal{P} : \frac{(s_{k+1} - s)F_k(r) + (s - s_k)F_{k+1}(r)}{s_{k+1} - s_k} \leq \alpha \right\}. \]
As an immediate consequence of this description, given \( \alpha \in [0, 1] \) and \( \tau \in T \) and setting

\[
G_{\tau, \mu}(p) = \sum_{i=1}^{\mu} p(i, \tau), \ 1 \leq \mu \leq M,
\]

and

\[
X_{s, \leq} = \{p(\cdot, \cdot) \in \mathcal{X} : \chi_\alpha[p(x) \leq s]\}, \quad X_{s, \geq} = \{p(\cdot, \cdot) \in \mathcal{X} : \chi_\alpha[p(x) \geq s]\},
\]

we get

\[
s < s_1 \implies X_{s, \leq} = \emptyset, \quad X_{s, \geq} = \mathcal{X},
\]

\[
s = s_1 \implies X_{s, \leq} = \{p \in \mathcal{X} : G_{\tau, 1}(p) \leq s_1 G_{\tau, M}(p)\}, \quad X_{s, \geq} = \mathcal{X},
\]

\[
s > s_M \implies X_{s, \leq} = \mathcal{X}, \quad X_{s, \geq} = \emptyset.
\]

\[
s_1 < s \leq s_M \implies
\begin{align*}
X_{s, \leq} &= \left\{p \in \mathcal{X} : \frac{(s_{k+1} - s_k) G_{\tau, k}(r) + (s_k - s_{k-1}) G_{\tau, k-1}(r)}{s_{k+1} - s_k} \geq \alpha G_{\tau, M}(p)\right\}, \\
X_{s, \geq} &= \left\{p \in \mathcal{X} : \frac{(s_{k+1} - s_k) G_{\tau, k}(r) + (s_k - s_{k-1}) G_{\tau, k-1}(r)}{s_{k+1} - s_k} \leq \alpha G_{\tau, M}(p)\right\},
\end{align*}
\]

\[
k = k_s : s_k < s \leq s_{k+1}.
\]

implying 1-convexity of the conditional quantile on \( \mathcal{X} \) (recall that \( G_{\tau, \mu}(p) \) are linear in \( p \)). \( \square \)

### 3.6.3 Proof of Proposition 3.4

#### 3.6.3.1 Proof of Proposition 3.4.i

We call step \( \ell \) essential, if at this step rule 2d is invoked.

1°. Let \( x \in \mathcal{X} \) be the true signal underlying the observation \( \tilde{\omega}^K \), so that \( \tilde{\omega}_1, \ldots, \tilde{\omega}_K \) are independently of each other drawn from the distribution \( p_{A(x)} \). Consider the "ideal" estimate given by exactly the same rules as the Bisection procedure in Section 3.2.4.2 (in the sequel, we refer to the latter as "true" one), with tests \( T^K_{\Delta, r, t}(\cdot) \), \( T^K_{\Delta, l, t}(\cdot) \) in rule 2d replaced with "ideal tests"

\[
\tilde{T}_{\Delta, r, t} = \tilde{T}_{\Delta, l, t} = \begin{cases} \text{right}, & f(x) > c_\ell, \\ \text{left}, & f(x) \leq c_\ell. \end{cases}
\]

Marking by * the entities produced by the resulting fully deterministic procedure, we arrive at the sequence of nested segments \( \Delta^*_\ell = [a^*_\ell, b^*_\ell], \ 0 \leq \ell \leq L^* \leq L \), along with subsegments \( \Delta^*_\ell, r = [c^*_\ell, u^*_\ell], \ \Delta^*_\ell, l = [u^*_\ell, c^*_\ell] \) of \( \Delta^*_{\ell-1} \), defined for all *-essential values of \( \ell \), and the output segment \( \Delta^* \) 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 \( \Delta^*_\ell \), \( 0 \leq \ell \leq L^* \), same as \( f(x) \in \Delta^* \).

Let \( \mathcal{L}^* \) be the set of all *-essential values of \( \ell \). For \( \ell \in \mathcal{L}^* \), let the event \( \mathcal{E}_\ell[x] \)
parameterized by $x$ be defined as follows:

$$E_x = \begin{cases} \omega^K_\Delta_{i,e,t}\left(\omega^K\right) = \text{right} \text{ or } T^K_{\Delta^{i,e}_{t,r,x}}\left(\omega^K\right) = \text{right}, & f(x) \leq u^*_t, \\ \omega^K_\Delta_{i,e,t}\left(\omega^K\right) = \text{right}, & u^*_t \leq f(x) \leq c^*_t, \\ \omega^K_\Delta_{i,e,t}\left(\omega^K\right) = \text{left}, & c^*_t \leq f(x) < v^*_t, \\ \omega^K_\Delta_{i,e,t}\left(\omega^K\right) = \text{left or } T^K_{\Delta^{i,e}_{t,r,x}}\left(\omega^K\right) = \text{left}, & f(x) \geq v^*_t. \end{cases}$$

(3.94)

2°. Observe that by construction and in view of Proposition 2.27 we have

$$\forall \ell \in \mathcal{L}^*: \text{Prob}_{\omega^K \sim \mathcal{P}(A_x) \times \ldots \times \mathcal{P}(A_z)}\{E_x[x]\} \leq 2\delta. \quad (3.95)$$

Indeed, let $\ell \in \mathcal{L}^*$.

- When $f(x) \leq u^*_t$, we have $x \in X$ and $f(x) \leq u^*_t \leq c^*_t$, implying that $E_x[x]$ takes place only when either the left test $T^K_{\Delta^{i,e}_{t,r,x}}$ or the right test $T^K_{\Delta^{i,e}_{t,r,x}}$, or both, accept wrong – right – hypotheses from the pairs of right and left hypotheses. Since the corresponding intervals ($[u^*_t, c^*_t]$, $[c^*_t, v^*_t]$) for the right side one) are $\delta$-good left/right, respectively, the risks of the tests do not exceed $\delta$, and the $p_{A(x)}$-probability of the event $E_x[x]$ is at most $2\delta$;
- when $u^*_t < f(x) \leq c^*_t$, the event $E_x[x]$ takes place only when the right side test $T^K_{\Delta^{i,e}_{t,r,x}}$ accepts wrong – right – hypothesis from the pair; similarly to the above, this can happen with $p_{A(x)}$-probability at most $\delta$;
- when $c^*_t < f(x) \leq v^*_t$, the event $E_x[x]$ takes place only if the left test $T^K_{\Delta^{i,e}_{t,r,x}}$ accepts wrong – left – hypothesis from the pair it was applied to, which again happens with $p_{A(x)}$-probability at most $\delta$;
- finally, when $f(x) > v^*_t$, the event $E_x[x]$ takes place only when either the left test $T^K_{\Delta^{i,e}_{t,r,x}}$ or the right side test $T^K_{\Delta^{i,e}_{t,r,x}}$, or both, accept wrong – left – hypotheses from the pairs the tests are applied to; same as above, this can happen with $p_{A(x)}$-probability at most $2\delta$.

3°. Let $\bar{L} = \bar{L}(\omega^K)$ be the last step of true estimating procedure as run on the observation $\omega^K$. We claim that the following holds true:

1°) Let $\mathcal{E} := \bigcup_{\ell \in \mathcal{L}_x} E_x[x]$, so that the $p_{A(x)}$-probability of the event $\mathcal{E}$, the observations stemming from $x$, is at most

$$2\delta L = \epsilon$$

(see (3.17), (3.95)). Assume that $\omega^K \notin \mathcal{E}$. Then $\bar{L}(\omega^K) \leq L^*$, and only two cases are possible:

A. True estimating procedure does not terminate due to arriving at disagreement. In this case $L^* = \bar{L}(\omega^K)$ and the trajectories of the ideal and the true procedures are identical (same localizers and essential steps, same output segments, etc.), and, in particular, $f(x) \in \Delta$, or

B. True estimating procedure terminates due to arriving at a disagreement. Then $\Delta_x = \Delta_x^*$ for $\ell < \bar{L}$, and $f(x) \in \Delta$. 


In view of A and B the \( p_{A(x)} \)-probability of the event \( f(x) \in \tilde{\Delta} \) is at least \( 1 - \epsilon \), as claimed in Proposition 3.4.

To prove (I), note that the actions at step \( \ell \) in ideal and 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 \( \bar{\omega}^K \notin \mathcal{E} \), and let \( \ell \leq L^* \) be such that \( \Delta_{\ell-1} = \Delta_{\ell-1}^* \), whence also \( u_\ell = u_\ell^*, c_\ell = c_\ell^* \) and \( v_\ell = v_\ell^* \). Assume that \( \ell \) is essential (given that \( \Delta_{\ell-1} = \Delta_{\ell-1}^* \), this may happen if and only if \( \ell \) is \( ^* \)-essential as well). Then either

C. At step \( \ell \) true procedure terminates due to disagreement, in which case \( f(x) \in \tilde{\Delta} \), or

D. At step \( \ell \) there was no disagreement, in which case \( \Delta_\ell \) as given by (3.16) is identical to \( \Delta_\ell^* \) as given by the ideal counterpart of (3.16) in the case of \( \Delta_{\ell-1} = \Delta_{\ell-1}^* \), that is, by the rule

\[
\Delta^* = \begin{cases} 
[c_\ell, b_{\ell-1}], & f(x) > c_\ell, \\
[a_{\ell-1}, c_\ell], & f(x) \leq c_\ell.
\end{cases}
\] (3.96)

To verify (!!), let \( \bar{\omega}^K \) and \( \ell \) 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 \( \Delta_{\ell,rg} = \Delta_{\ell,rg}^* \). Consider first the case when true estimation procedure terminates by disagreement at step \( \ell \), so that \( T^K_{\Delta_{\ell,rg}^*}(\bar{\omega}^K) \neq T^K_{\Delta_{\ell,rg}^*}(\bar{\omega}^K) \). When assuming that \( f(x) < u_\ell = u_\ell^* \), the relation \( \bar{\omega}^K \notin \mathcal{E}[x] \) combines with (3.94) to imply that \( T^K_{\Delta_{\ell,rg}^*}(\bar{\omega}^K) = T^K_{\Delta_{\ell,rg}^*}(\bar{\omega}^K) = \text{left} \), which under disagreement is impossible. Assuming \( f(x) > v_\ell = v_\ell^* \), the same argument results in \( T^K_{\Delta_{\ell,rg}^*}(\bar{\omega}^K) = T^K_{\Delta_{\ell,rg}^*}(\bar{\omega}^K) = \text{right} \), which again is impossible. We conclude that in the case in question \( u_\ell \leq f(x) \leq v_\ell \), i.e., \( f(x) \in \tilde{\Delta} \), as claimed in C. C is proved.

Now, suppose that there was a consensus at the step \( \ell \) in true estimation procedure. Because \( \bar{\omega}^K \notin \mathcal{E}[x] \) this can happen in the following four cases:

1. \( T^K_{\Delta_{\ell,rg}^*}(\bar{\omega}^K) = \text{left} \) and \( f(x) \leq u_\ell = u_\ell^* \),
2. \( T^K_{\Delta_{\ell,rg}^*}(\bar{\omega}^K) = \text{left} \) and \( u_\ell < f(x) \leq c_\ell = c_\ell^* \),
3. \( T^K_{\Delta_{\ell,rg}^*}(\bar{\omega}^K) = \text{right} \) and \( c_\ell < f(x) < v_\ell = v_\ell^* \),
4. \( T^K_{\Delta_{\ell,rg}^*}(\bar{\omega}^K) = \text{right} \) and \( v_\ell \leq f(x) \).

Due to consensus at the step \( \ell \), in situations 1 and 2 (3.16) says that \( \Delta_\ell = [a_{\ell-1}, c_\ell] \), which combines with (3.96) and \( v_\ell = v_\ell^* \) to imply that \( \Delta_\ell = \Delta_\ell^* \). Similarly, in situations 3 and 4, due to consensus at the step \( \ell \), (3.16) implies that \( \Delta_\ell = [c_\ell, b_{\ell-1}] \), which combines with \( u_\ell = u_\ell^* \) and (3.96) to imply that \( \Delta_\ell = \Delta_\ell^* \). D is proved.

3.6.3.2 Proof of Proposition 3.4.ii

There is nothing to prove when \( \frac{b_0 - a_0}{2} \leq \tilde{\rho} \), since in this case the estimate \( \frac{a_0 + b_0}{2} \) which does not use observations at all is \( (\tilde{\rho}, 0) \)-reliable. From now on we assume that \( b_0 - a_0 > 2\tilde{\rho} \), implying that \( L \) is a positive integer.

1°. Observe, first, that if \( a \) and \( b \) 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 \( K \) observations, which decides upon the hypotheses \( H_1, H_2 \), stating that the observations are drawn from \( p_{A(x)} \) with \( x \in Z_{i,\geq}^{b,\geq} \) (\( H_1 \)) or with \( x \in Z_{j,\leq}^{a,\leq} \) (\( H_2 \)) with risk at most \( \epsilon \). Indeed, it suffices to consider the test which accepts \( H_1 \) and rejects \( H_2 \) when \( f(\omega^K) \geq \frac{a+b}{2} \) and accepts \( H_2 \) and rejects \( H_1 \) otherwise.

2°. With parameters of Bisection chosen according to (3.19), by already proved Proposition 3.4.i, we have

\[ E. \text{ For every } x \in X, \text{ the } p_{A(x)} \text{-probability of the event } f(x) \in \Delta, \text{ } \Delta \text{ being the output segment of our Bisection, is at least } 1 - \epsilon. \]

3°. We claim also that

F.1. Every segment \( \Delta = [a,b] \) with \( b - a > 2\rho \) and lower-feasible \( a \) is \( \delta \)-good (right),

F.2. Every segment \( \Delta = [a,b] \) with \( b - a > 2\rho \) and upper-feasible \( b \) is \( \delta \)-good (left),

F.3. Every \( \kappa \)-maximal \( \delta \)-good (left or right) segment has length at most \( 2\rho + \kappa = \hat{\rho} \).

As a result, for every essential step \( \ell \), the lengths of the segments \( \Delta_{\ell,rg} \) and \( \Delta_{\ell,ll} \) do not exceed \( \hat{\rho} \).

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 \( \omega_k \) are sampled from \( p_{A(x)} \) for some \( x \in Z_{i,\geq}^{b,\geq} \), resp., from some \( x \in Z_{j,\leq}^{a,\leq} \), can be decided upon with risk \( \leq \epsilon \), implying, same as in the proof of Proposition 2.25, that

\[ \epsilon_{ij}\Delta \leq [2\sqrt{\epsilon(1-\epsilon)}]^{1/K}. \]

Hence, taking into account that the column and the row sizes of \( E_{\Delta,x} \) do not exceed \( NI \),

\[ \sigma_{\Delta,x} \leq NI \max_{i,j} \epsilon_{ij}^{K} \Delta \leq NI[2\sqrt{\epsilon(1-\epsilon)}]^{K/K} \leq \frac{\epsilon}{2L} = \delta \]

(we have used (3.19)), that is, \( \Delta \) indeed is \( \delta \)-good (right).

4°. Let us fix \( x \in X \) and consider a trajectory of Bisection, the observation being drawn from \( p_{A(x)} \). The output \( \Delta \) of the procedure is given by one of the following options:

1. At some step \( \ell \) of Bisection, the process terminated according to rules in 2b or 2c. In the first case, the segment \( [\ell, b_{\ell-1}] \) has lower-feasible left endpoint and is not \( \delta \)-good (right), implying by F.1 that the length of this segment (which is \( 1/2 \) of the length of \( \Delta = \Delta_{\ell-1} \)) is \( \leq 2\rho \), so that the length \( |\Delta| \) of \( \Delta \) is at most \( 4\rho \leq 2\hat{\rho} \). The same conclusion, by a completely similar argument, holds true if the process terminated at step \( \ell \) according to rule 2c.

2. At some step \( \ell \) of Bisection, the process terminated due to disagreement. In this case, by F.3, we have \( |\Delta| \leq 2\hat{\rho} \).

3. Bisection terminated at step \( L \), and \( \Delta = \Delta_L \). In this case, termination clauses in rules 2b, 2c and 2d were never invoked, clearly implying that \( |\Delta_s| \leq |\Delta_{s-1}|/2 \), \( 1 \leq s \leq L \), and thus \( |\Delta| = |\Delta_L| \leq 2^{-L}|\Delta_0| \leq 2\hat{\rho} \) (see (3.19)).
Thus, we have \(|\Delta| \leq 2\tilde{\rho}\), implying that whenever the signal \(x \in X\) underlying observations and the output segment \(\Delta\) are such that \(f(x) \in \Delta\), the error of the Bisection estimate (which is the midpoint of \(\Delta\)) is at most \(\tilde{\rho}\). Invoking \(E\), we conclude that Bisection estimate is \((\tilde{\rho}, \epsilon)\)-reliable. \(\square\)

### 3.6.4 Proof of Proposition 3.14

Let us fix \(\epsilon \in (0, 1)\). Setting

\[
\rho_K = \frac{1}{2} \left[ \hat{\Psi}^+_{+K}(\bar{h}, \bar{H}) + \hat{\Psi}^-_{-K}(\bar{h}, \bar{H}) \right]
\]

and invoking Corollary 3.13, all we need to prove is that in the case of A.1-3 one has

\[
\limsup_{K \to \infty} \left[ \hat{\Psi}^+_{+K}(\bar{h}, \bar{H}) + \hat{\Psi}^-_{-K}(\bar{h}, \bar{H}) \right] \leq 0. \quad (3.97)
\]

To this end, note that in our current situation, (3.48) and (3.52) simplify to

\[
\Phi(h, H; Z) = -\frac{1}{2} \ln \det(I - \Theta_1^{1/2} H \Theta_1^{1/2}) + \frac{1}{2} \text{Tr} \left( B^T \left[ \frac{H}{h} \right] + [H, h] \text{Tr} \left[ \Theta^{-1}^{-1} [H, h] B \right] \right),
\]

\[
\hat{\Psi}^+_{+K}(h, H) = \inf_{\alpha} \left\{ \max_{Z \in Z} \left[ \alpha \Phi(h/\alpha, H/\alpha; Z) - \text{Tr}(QZ) + K^{-1} \alpha \ln(2/\epsilon) \right] : \alpha > 0, -\gamma \alpha \Theta_{-1}^{-1} \preceq H \preceq \gamma \alpha \Theta_{-1}^{-1} \right\},
\]

\[
\hat{\Psi}^-_{-K}(h, H) = \inf_{\alpha} \left\{ \max_{Z \in Z} \left[ \alpha \Phi(-h/\alpha, -H/\alpha; Z) + \text{Tr}(QZ) + K^{-1} \alpha \ln(2/\epsilon) \right] : \alpha > 0, -\gamma \alpha \Theta_{-1}^{-1} \preceq H \preceq \gamma \alpha \Theta_{-1}^{-1} \right\}.
\]

Hence

\[
\left[ \hat{\Psi}^+_{+K}(\bar{h}, \bar{H}) + \hat{\Psi}^-_{-K}(\bar{h}, \bar{H}) \right] \leq \inf_{\alpha} \left\{ \max_{Z_1, Z_2 \in Z} \left[ \alpha \Phi(h/\alpha, H/\alpha; Z_1) - \text{Tr}(QZ_1) \right. \right.
\]

\[
+ \Phi(-h/\alpha, -H/\alpha; Z_1) + \text{Tr}(QZ_2) + 2K^{-1} \alpha \ln(2/\epsilon) : \alpha > 0, -\gamma \alpha \Theta_{-1}^{-1} \preceq H \preceq \gamma \alpha \Theta_{-1}^{-1} \right\}
\]

\[
= \inf_{\alpha} \max_{Z_1, Z_2 \in Z} \left\{ -\frac{1}{2} \alpha \ln \det \left( I - \Theta_1^{1/2} H \Theta_1^{1/2} \right) + 2K^{-1} \alpha \ln(2/\epsilon) + \text{Tr}(Q[Z_2 - Z_1]) + \frac{1}{2} \left[ \alpha \text{Tr} \left( Z_1 Q(h/\alpha, H/\alpha) \right) + \alpha \text{Tr} \left( Z_2 Q(-h/\alpha, -H/\alpha) \right) \right] : \alpha > 0, -\gamma \alpha \Theta_{-1}^{-1} \preceq H \preceq \gamma \alpha \Theta_{-1}^{-1} \right\}
\]
\[
\inf_{\alpha} \max_{Z_1, Z_2 \in \mathcal{Z}} \left\{ -\frac{1}{2}\alpha \ln \text{Det} \left( I - [\Theta_1^{1/2} \hat{H} \Theta_1^{1/2}]^2/\alpha^2 \right) + 2K^{-1}\alpha \ln(2/\epsilon) + \frac{1}{2} \text{Tr} \left( Z_1 B^T [\hat{H}, \hat{h}]^T [\alpha \Theta_{*}^{-1} - \hat{H}]^{-1} [\hat{H}, \hat{h}] B \right) + \frac{1}{2} \text{Tr} \left( Z_2 B^T [\hat{H}, \hat{h}]^T [\alpha \Theta_{*}^{-1} + \hat{H}]^{-1} [\hat{H}, \hat{h}] B \right) + \text{Tr}(Q[\bar{Z}_2 - Z_1]) + \frac{1}{2} \text{Tr}([\bar{Z}_1 - Z_2] B^T B) \right\} : \quad (3.98)
\]

By (3.57) we have \( \frac{1}{2} B^T \left[ \frac{\hat{H}}{h} \right] h B = B^T [C^T Q C + J] B \), where the only nonzero entry, if any, in \((d + 1) \times (d + 1)\) matrix \( J \) is in the cell \((d + 1, d + 1)\). By definition of \( B \), see (3.48), the only nonzero element, if any, in \( \bar{J} = B^T JB \) is in the cell \((m + 1, m + 1)\), and we conclude that

\[
\frac{1}{2} \text{Tr}([Z_1 - Z_2] B^T \left[ \frac{\hat{H}}{h} \right] h B) = \text{Tr}([Z_1 - Z_2] Q) + \text{Tr}([Z_1 - Z_2] J) = \text{Tr}([Z_1 - Z_2] Q),
\]

implying that the quantity \( T(Z_1, Z_2) \) in (3.98) is zero, provided \( Z_1, Z_2 \in \mathcal{Z} \). Consequently, (3.98) becomes

\[
\left[ \hat{\Psi}_{+, K}(\bar{h}, \bar{H}) + \hat{\Psi}_{-, K}(\bar{h}, \bar{H}) \right] \leq \inf_{\alpha} \max_{Z_1, Z_2 \in \mathcal{Z}} \left\{ -\frac{1}{2}\alpha \ln \text{Det} \left( I - [\Theta_1^{1/2} \hat{H} \Theta_1^{1/2}]^2/\alpha^2 \right) + 2K^{-1}\alpha \ln(2/\epsilon) + \frac{1}{2} \text{Tr} \left( Z_1 B^T [\hat{H}, \hat{h}]^T [\alpha \Theta_{*}^{-1} - \hat{H}]^{-1} [\hat{H}, \hat{h}]^T B \right) + \frac{1}{2} \text{Tr} \left( Z_2 B^T [\hat{H}, \hat{h}]^T [\alpha \Theta_{*}^{-1} + \hat{H}]^{-1} [\hat{H}, \hat{h}] B \right) : \alpha > 0, -\gamma \alpha \Theta_{*}^{-1} \preceq \bar{H} \preceq \gamma \alpha \Theta_{*}^{-1} \right\}.
\]

Now, for appropriately selected independent of \( K \) real \( c \), for \( \alpha \) allowed by (3.99) and all \( Z_1, Z_2 \in \mathcal{Z} \) we have (recall that \( \mathcal{Z} \) is bounded)

\[
\frac{1}{2} \text{Tr} \left( Z_1 B^T [\hat{H}, \hat{h}]^T [\alpha \Theta_{*}^{-1} - \hat{H}]^{-1} [\hat{H}, \hat{h}] B \right) + \frac{1}{2} \text{Tr} \left( Z_2 B^T [\hat{H}, \hat{h}]^T [\alpha \Theta_{*}^{-1} + \hat{H}]^{-1} [\hat{H}, \hat{h}] B \right) \leq c/\alpha,
\]

along with

\[
-\frac{1}{2}\alpha \ln \text{Det} \left( I - [\Theta_1^{1/2} \hat{H} \Theta_1^{1/2}]^2/\alpha^2 \right) \leq c/\alpha.
\]

Therefore, given \( \delta > 0 \), we can find \( \alpha = \alpha_\delta > 0 \) large enough to ensure that

\[
-\gamma \alpha_\delta \Theta_{*}^{-1} \preceq \bar{H} \preceq \gamma \alpha_\delta \Theta_{*}^{-1} \quad \text{and} \quad 2c/\alpha_\delta \leq \delta,
\]

which combines with (3.99) to imply that

\[
\left[ \hat{\Psi}_{+, K}(\bar{h}, \bar{H}) + \hat{\Psi}_{-, K}(\bar{h}, \bar{H}) \right] \leq \delta + 2K^{-1}\alpha_\delta \ln(2/\epsilon),
\]
and (3.97) follows.
Chapter Four

Signal Recovery by Linear Estimation

OVERVIEW

In this chapter we consider several variations of one of the most basic problems of high-dimensional statistics — signal recovery. In its simplest form the problem is as follows: given positive definite $m \times m$ matrix $\Gamma$, $m \times n$ matrix $A$, $\nu \times n$ matrix $B$, and indirect noisy observation

$$\omega = Ax + \xi \quad [\xi \sim \mathcal{N}(0, \Gamma)] \quad (4.1)$$

of unknown "signal" $x$ known to belong to a given convex compact subset $\mathcal{X}$ of $\mathbb{R}^n$, we want to recover the vector $Bx \in \mathbb{R}^\nu$ of $x$. We focus first on the case where the quality of a candidate recovery $\omega \mapsto \hat{x}(\omega)$ is quantified by its worst-case, over $x \in \mathcal{X}$, expected $\|\cdot\|_2$-error, that is, by the risk

$$\text{Risk}[\hat{x}(\cdot)|\mathcal{X}] = \sup_{x \in \mathcal{X}} \sqrt{\mathbb{E}_{\xi \sim \mathcal{N}(0, \Gamma)} \{\|\hat{x}(Ax + \xi) - Bx\|_2^2\}}. \quad (4.2)$$

The simplest and the most studied type of recovery is affine one: $\hat{x}(\omega) = H^T \omega + h$; assuming $\mathcal{X}$ symmetric w.r.t. the origin, we lose nothing when passing from affine estimates to linear ones – those of the form $\hat{x}_H(\omega) = H^T \omega$. 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., [6, 59, 82, 151, 152, 194, 203, 204] 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., [79, 87, 104, 122, 195, 226, 234] and references therein; a typical result of this type states that when recovering $x \in \mathcal{X}$ from direct observation $\omega = x + \sigma \xi$, $\xi \sim \mathcal{N}(0, I_m)$, where $\mathcal{X}$ is an ellipsoid of the form

$$\{x \in \mathbb{R}^n : \sum_j j^{2\alpha} x_j^2 \leq L^2\},$$

or the box

$$\{x \in \mathbb{R}^n : j^{\alpha} |x_j| \leq L, j \leq n\},$$

with fixed $L < \infty$ and $\alpha > 0$, the ratio of the risk of a properly selected linear estimate to the minimax risk

$$\text{Risk}_{\text{opt}}[\mathcal{X}] := \inf_{\hat{x}(\cdot)} \text{Risk}[\hat{x}|\mathcal{X}] \quad (4.3)$$

(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 $n$, $\alpha$ and $L$ being...
fixed. Similar “near-optimality” results are known for “diagonal” case, where \( \mathcal{X} \) is the above ellipsoid/box and \( A, B, \Gamma \) are diagonal matrices. To the best of our knowledge, the only “general” (that is, not imposing severe restrictions on how the geometries of \( \mathcal{X}, A, B, \Gamma \) are linked to each other) result on optimality of linear estimates is due to D. Donoho who proved \[65\] 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 primary goal of this chapter is to establish rather general results on near-optimality of properly built linear estimates as compared to all possible estimates. Results of this type are bound to impose some restrictions on \( \mathcal{X} \), since there are cases (e.g., the case 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:

\[
\mathcal{X} = \{ x \in \mathbb{R}^n : x^T R_k x \leq 1, 1 \leq k \leq K \} \quad [R_k \succeq 0, \sum_k R_k > 0] \quad (4.4)
\]

in particular, \( \mathcal{X} \) can be a symmetric w.r.t. the origin compact polytope given by \( 2K \) linear inequalities \(-1 \leq r^T_k x \leq 1, 1 \leq k \leq K \), or, equivalently, \( \mathcal{X} = \{ x : x^T (r_k r_k^T) x \leq 1, k \leq K \} \). Another instructive example is a set of the form

\[
\mathcal{X} = \{ x : \| S x \|_p \leq L \}, \text{ where } p \geq 2 \text{ and } S \text{ is a matrix with trivial kernel.}
\]

It should be stressed than while imposing some restrictions on \( \mathcal{X} \), we require nothing from \( A, B, \text{ and } \Gamma \), aside of positive definiteness of the latter matrix. Our main result (Proposition 4.5) states, in particular, that with \( \mathcal{X} \) given by \((4.4)\) and with arbitrary \( A \) and \( B \), the risk of properly selected linear estimate \( \hat{x}_H \), with both \( H_*, \) and the risk efficiently computable, satisfies the bound

\[
\text{Risk}[\hat{x}_{H_*}, \mathcal{X}] \leq O(1) \sqrt{\ln(K + 1)} \text{Risk}_{\text{opt}}[\mathcal{X}],
\]

where \( \text{Risk}_{\text{opt}}[\mathcal{X}] \) 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” \( \mathcal{X}, A, B, \Gamma \). 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 \[195\], to bound from below the minimax risk via Bayesian risk associated with properly selected Gaussian prior.\(^1\)

The main body of the chapter originates from \[137, 136\] and is organized as follows.

- Section 4.1 presents basic results on Conic Programming and Conic Duality – the principal optimization tools utilized in all subsequent constructions and proofs.

---

\(^1\)[87, 195] 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 \( \mathcal{X} = \{ x \in \mathbb{R}^n : \sum_j j^{2\alpha} x_j^2 \leq L^2 \} \) 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 \((*)\).
• 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 4.2.2.2). We discuss also the “expressive abilities” of the family of sets (we call them ellitopes) to which our main result applies.

• In Section 4.3 we extend the results of previous section from ellitopes to their “matrix analogs” — spectratopes in the role of signal sets, passing simultaneously from the norm \( \| \cdot \|_2 \) in which the recovery error is measured to arbitrary spectratopic norms — those for which the unit ball of the conjugate norm is a spectratope. In addition, we allow for observation noise to have nonzero mean and to be non-Gaussian.

• Section 4.4 adjusts our preceding results on linear estimation to the case where the signals to be recovered possess stochastic components.

• Finally, Section 4.5 deals with "uncertain-but-bounded" observation noise, that is, noise selected “by nature,” perhaps in an adversarial fashion, from a given bounded set.

4.1 PRELIMINARIES: EXECUTIVE SUMMARY ON CONIC PROGRAMMING

4.1.1 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 that a cone is a convex set. We call a cone \( K \) regular, if it is closed, pointed (that is, does not contain lines passing through the origin, or, equivalently, \( K \cap [-K] = \{0\} \)) and possesses a nonempty interior.

Given a cone \( K \subset E \), we can associate with it its dual cone \( K^* \) defined as

\[
K^* = \{ y \in E : (y, x) \geq 0 \ \forall x \in K \};
\]

where \( (\cdot, \cdot) \) is inner product on \( E \).

It is immediately seen that \( K^* \) 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 \( \mathbf{L}_+^d = \{ x \in \mathbf{R}^d : x_d \geq \sqrt{\sum_{i=1}^{d-1} x_i^2} \} \);
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

$$\langle A, B \rangle = \text{Tr}(AB^T) = \text{Tr}(AB) = \sum_{i,j=1}^{d} A_{ij}B_{ij}, \quad A, B \in S^d.$$ 

All listed so far cones are self-dual.

4. Let $\| \cdot \|$ be a norm on $\mathbb{R}^n$. The set $\{ [x; t] \in \mathbb{R}^n \times \mathbb{R} : t \geq \|x\| \}$ is a regular cone, and the dual cone is $\{ [y; \tau] : \|y\|_* \leq \tau \}$, where

$$\|y\|_* = \max_x \{ x^T y : \|x\| \leq 1 \}$$

is the norm on $\mathbb{R}^n$ conjugate to $\| \cdot \|$.

Additional useful for the sequel example of a regular cone is the conic hull of a convex compact set defined as follows. Let $T$ be a convex compact set with a nonempty interior in Euclidean space $E$. We can associate with $T$ its closed conic hull

$$T = \text{cl} \left\{ [t; \tau] \in E^+ = E \times \mathbb{R} : \tau > 0, t/\tau \in T \right\}.$$ 

It is immediately seen that $T$ is a regular cone, and that to get this cone, one should add to the convex set $K^o(T)$ the origin of $E^+$. It is also clear that one can “see $T$ in $T$:” – $T$ is nothing but the cross-section of the cone $T$ by the hyperplane $\tau = 1$ in $E^+ = \{ [t; \tau] \}$:

$$T = \{ t \in E : [t; 1] \in T \}$$

It is easily seen that the cone $T_*$ dual to $T$ is given by

$$T_* = \{ [g; s] \in E^+ : s \geq \phi_T(-g) \},$$

where

$$\phi_T(g) = \max_{t \in T} \langle g, t \rangle$$

is the support function of $T$.

4.1.2 Conic problems and their duals

Given regular cones $K_i \subset E_i$, $1 \leq i \leq m$, consider optimization problem of the form

$$\text{Opt}(P) = \min \left\{ \langle c, x \rangle : \begin{array}{l} A_i x - b_i \in K_i, \\ R x = r \end{array}, \quad i = 1, ..., m \right\}, \quad (P)$$

where $x \mapsto A_i x - b_i$ 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 $A_i x - b_i \in K_i$ 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, i.e., with $A_i \bar{x} - b_i \in \text{int} K_i$, 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 – lower bounds on the optimal value $\text{Opt}(P)$. 
of the primal problem \((P)\). Linear aggregation of constraints works as follows: let us equip every conic constraint \(A_i x - b_i \in K_i\) 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 \(y_i \in K_i^*, 1 \leq i \leq m\), \(z\) be Lagrange multipliers. By the definition of the dual cone and due to \(A_i x - b_i \in K_i\), \(y_i \in K_i^*\), we have
\[
\langle y_i, A_i x \rangle \geq \langle y_i, b_i \rangle, 1 \leq i \leq m
\]
and of course
\[
z^T Rx \geq r^T z.
\]
Summing up all resulting inequalities, we arrive at the scalar linear inequality
\[
\left\langle R^* z + \sum_i A_i^* y_i, x \right\rangle \geq r^T z + \sum_i \langle b_i, y_i \rangle \tag{!}
\]
where \(A_i^*\) are the conjugates to \(A_i\): \(\langle y, A_i x \rangle_{E_i} \equiv \langle A_i^* y, x \rangle_{E_i}\), 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
\[
\text{Opt}(D) = \max_{y_i: z} \left\{ r^T z + \sum_i \langle b_i, y_i \rangle : y_i \in K_i^*, 1 \leq i \leq m \right\} \tag{D}
\]
of maximizing this lower bound on \(\text{Opt}(P)\).

The relations between the primal and the dual conic problems are the subject of the standard Conic Duality Theorem as follows:

**Theorem 4.1.** [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 \(\text{Opt}(D) \leq \text{Opt}(P)\)
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.

\(^2\)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.
Remark 4.2. 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, $A_i\bar{x} - b_i \in \text{int } K_i$ 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 [184, Section 7.1.3].

4.1.3 Schur Complement Lemma
The following simple fact is extremely useful:

**Lemma 4.3.** [Schur Complement Lemma] Symmetric block matrix

$$A = \begin{bmatrix} P & QT \\ Q & R \end{bmatrix}$$

with $R \succ 0$ is positive (semi)definite if and only if the matrix $P - QT R^{-1} Q$ is so.

**Proof.** With $u, v$ of the same sizes as $P$, respectively, $R$, we have

$$\min_v [u; v]^T A [u; v] = u^T [P - QT R^{-1} Q] u$$

(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 $[P - QT R^{-1} Q]$ is nonnegative everywhere (since the latter quadratic form is obtained from the former one by partial minimization). \hfill $\Box$

4.2 NEAR-OPTIMAL LINEAR ESTIMATION FROM GAUSSIAN OBSERVATIONS

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 $\mathcal{X} \subset \mathbb{R}^n$ via noisy observation

$$\omega = Ax + \xi, \quad \xi \sim \mathcal{N}(0, \Gamma),$$

(4.5)

where $\xi$ is the observation noise. A candidate estimate in this case is a (Borel) function $\hat{x}(:) : \mathbb{R}^m \to \mathbb{R}^\nu$, and the performance of such estimate in what follows will be quantified by the Euclidean risk $\text{Risk}[\hat{x}, \mathcal{X}]$ defined by (4.2).
4.2.1.1 Ellitopes

From now on we assume that \( \mathcal{X} \subset \mathbb{R}^n \) is a set given by

\[
\mathcal{X} = \{ x \in \mathbb{R}^n : \exists (y \in \mathbb{R}^{\bar{n}}, t \in T) : x = Py, y^T R_k y \leq t_k, 1 \leq k \leq K \}, \tag{4.6}
\]

where

- \( P \) is an \( n \times \bar{n} \) matrix,
- \( R_k \geq 0 \) are \( \bar{n} \times \bar{n} \) matrices with \( \sum_k R_k > 0 \),
- \( T \) is a nonempty computationally tractable convex compact subset of \( \mathbb{R}^K_+ \) intersecting the interior of \( \mathbb{R}^K_+ \) and such that \( T \) is monotone, meaning that the relations \( 0 \leq \tau \leq t \) and \( t \in T \) imply that \( \tau \in T \).  

Note that under our assumptions \( \text{int} \ T \neq \emptyset \).

In the sequel, we refer to a set of the form (4.6) with data \( [P, \{R_k, 1 \leq k \leq K\}, T] \) satisfying just formulated assumptions as to an ellitope, and to (4.6) – 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):

- when \( K = 1 \), \( T = [0, 1] \) and \( R_1 > 0 \), \( \mathcal{X} \) is the ellipsoid \( \{ x : x^T R_1 x \leq 1 \} \);
- when \( K \geq 1 \), \( T = \{ t \in \mathbb{R}^K : 0 \leq t_k \leq 1, k \leq K \} \), and \( \mathcal{X} \) is the intersection

  \[
  \bigcap_{1 \leq k \leq K} \{ x : x^T R_k x \leq 1 \}
  \]

  of centered at the origin ellipsoids/elliptic cylinders. In particular, when \( U \) is a \( K \times n \) matrix of rank \( n \) with rows \( u_k^T \), \( 1 \leq k \leq K \), and \( R_k = u_k u_k^T \), \( \mathcal{X} \) is the symmetric w.r.t. the origin polytope \( \{ x : \|Ux\|_\infty \leq 1 \} \);
- when \( U, u_k \) and \( R_k \) are as in the latter example and \( T = \{ t \in \mathbb{R}^K_+ : \sum_k t_k/p/2 \leq 1 \} \)

  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 4.6.

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

\footnote{The latter relation is “for free” – given a nonempty convex compact set \( \mathcal{T} \subset \mathbb{R}^K_+ \), the right hand side of (4.6) remains intact when passing from \( \mathcal{T} \) to its “monotone hull” \( \{ \tau \in \mathbb{R}^K_+ : \exists t \in T : \tau \leq t \} \) which already is a monotone convex compact set.}
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 $\hat{x}(\cdot) : \mathbb{R}^m \to \mathbb{R}^\nu$; given observation (4.5), we recover $w = Bx$ as $\hat{x}(\omega)$. In the sequel, we quantify the quality of an estimate by its worst-case, over $x \in X$, expected $\| \cdot \|^2_2$ recovery error:

$$\text{Risk}[\hat{x}|X] = \sup_{x \in X} \left[ E_{\xi \sim N(0, \Gamma)} \left\{ \| \hat{x}(Ax + \xi) - Bx \|^2_2 \right\} \right]^{1/2}$$

and define the optimal, or the minimax, risk as

$$\text{Risk}_{\text{opt}}[X] = \inf_{\hat{x}} \text{Risk}[\hat{x}|X], \quad (4.7)$$

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 $\omega$ estimate

$$\hat{x}_H(\omega) = H^T \omega \quad (4.8)$$

with properly selected efficiently computable matrix $H$ is near-optimal in terms of its risk.

Our first observation is that when $X$ is the ellitope (4.6), replacing matrices $A$ and $B$ with $AP$ and $BP$, respectively, we pass from the initial estimation problem of interest to the transformed problem, where the signal set is

$$\tilde{X} = \{ y \in \mathbb{R}^n : \exists t \in T : y^T R_k y \leq t_k, \ 1 \leq k \leq K \},$$

and we want to recover $[BP]y$, $y \in \tilde{X}$, 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,

$$\text{Risk}[\hat{x}|X] = \sup_{y \in \tilde{X}} \left[ E_{\xi} \left\{ \| \hat{x}([AP]y + \xi) - [BP]y \|^2_2 \right\} \right]^{1/2}.$$

Therefore, to save notation, from now on, unless explicitly stated otherwise, we assume that matrix $P$ is identity, so that $X$ is the basic ellitope

$$X = \{ x \in \mathbb{R}^n : \exists t \in T : x^T R_k x \leq t_k, \ 1 \leq k \leq K \}. \quad (4.9)$$

We assume in the sequel that $B \neq 0$, since otherwise one has $Bx = 0$ for all $x \in X$, 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 (4.8), we may be interested in the estimate with the smallest risk, that is, the estimate associated with a $\nu \times m$ matrix $H$ which is an optimal solution to the optimization problem

$$
\min_{H} \{ R(H) := \text{Risk}^2[\hat{x}_H|\mathcal{X}] \}
$$

We have

$$
R(H) = \max_{x \in \mathcal{X}} E_{\xi} \{ \| H^T \omega - Bx \|_2^2 \} = E_{\xi} \{ \| H^T \xi \|_2^2 \} + \max_{x \in \mathcal{X}} \| H^T Ax - Bx \|_2^2
$$

$$
= \text{Tr}(H^T \Gamma H) + \max_{x \in \mathcal{X}} x^T (H^T A - B)^T (H^T A - B)x.
$$

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 $\phi_T$ be the support function of $T$:

$$
\phi_T(\lambda) = \max_{t \in T} \lambda^T t : \mathbf{R}^K \rightarrow \mathbf{R}.
$$

Observe that whenever $\lambda \in \mathbf{R}^K_+$ and $H$ are such that

$$
[B - H^T A]^T [B - H^T A] \preceq \sum_k \lambda_k R_k,
$$

(4.10)

for $x \in \mathcal{X}$ it holds

$$
\| Bx - H^T Ax \|_2^2 \leq \phi_T(\lambda).
$$

(4.11)

Indeed, in the case of (4.10) and with $x \in \mathcal{X}$, there exists $t \in T$ such that $x^T R_k x \leq t_k$ for all $t$, and consequently vector $t$ with the entries $t_k = x^T R_k x$ also belongs to $T$, whence

$$
\| Bx - H^T Ax \|_2^2 = x^T [B - H^T A] x \leq \sum_k \lambda_k x^T R_k x = \lambda^T \bar{t} \leq \phi_T(\lambda),
$$

which combines with (4.9) to imply (4.11).

From (4.11) it follows that if $H$ and $\lambda \geq 0$ are linked by (4.10), then

$$
\text{Risk}^2[\hat{x}_H|\mathcal{X}] = \max_{x \in \mathcal{X}} E \{ \| Bx - H^T (Ax + \xi) \|_2^2 \}
$$

$$
= \text{Tr}(H^T \Gamma H) + \max_{x \in \mathcal{X}} \| B - H^T A \|_2^2
$$

$$
\leq \text{Tr}(H^T \Gamma H) + \phi_T(\lambda).
$$

We see that the efficiently computable convex function

$$
\hat{R}(H) = \inf_{\lambda} \left\{ \text{Tr}(H^T \Gamma H) + \phi_T(\lambda) : (B - H^T A)^T (B - H^T A) \preceq \sum_k \lambda_k R_k, \lambda \geq 0 \right\}
$$
(which clearly is well defined due to compactness of $T$ combined with $\sum_k R_k > 0$)
is an upper bound on $R(H)$.

Note that by Schur Complement Lemma the matrix inequality $(B - H^T A)^T(B - H^T A) \preceq \sum_k \lambda_k R_k$ is equivalent to
the linear in $H$, $\lambda$ matrix inequality
\[
\begin{bmatrix}
\sum_k \lambda_k R_k & B^T - A^T H \\
B - H^T A & I_{\nu}
\end{bmatrix} \succeq 0.
\]

We have arrived at the following result:

**Proposition 4.4.** In the situation of this section, the risk of the “presumably good” linear estimate $\hat{x}_H(\omega) = H^T \omega$ yielded by an optimal solution $(H_s, \lambda_s)$ to the
(clearly solvable) convex optimization problem
\[
\begin{aligned}
\text{Opt} &= \min_{H, \lambda} \left\{ \text{Tr}(H^T \Gamma H) + \phi_T(\lambda) : (B - H^T A)^T(B - H^T A) \preceq \sum \lambda_k R_k, \lambda \succeq 0 \right\} \\
&= \min_{H, \lambda} \left\{ \text{Tr}(H^T \Gamma H) + \phi_T(\lambda) : \begin{bmatrix}
\sum_k \lambda_k R_k & B^T - A^T H \\
B - H^T A & I_{\nu}
\end{bmatrix} \succeq 0, \lambda \succeq 0 \right\}
\end{aligned}
\]
is upper-bounded by $\sqrt{\text{Opt}}$.

### 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 a given time $t_0$, $0 < t_0 < t_1$.

Physics, after suitable discretization of spatial variables, offers the following model of the situation. We represent the distribution of temperature at time $t$ as 
\[(2N - 1) \times (2N - 1)\text{ matrix } U(t) = [u_{ij}(t)]_{i,j=1}^{2N-1},\]
where $u_{ij}(t)$ is the temperature, at time $t$, at the point
\[P_{ij} = (p_i, p_j), \; p_k = k/N - 1, \; 1 \leq i, j \leq 2N - 1\]
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
\[\sum_{i=1}^{2N-1} \phi_k^2(p_i) = 1; \; \text{note that } \phi_k(\pm 1) = 0.\]
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

---

4It is well known that when $K = 1$ (i.e., $X$ is an ellipsoid), the above bounding scheme is exact: $R(\cdot) \equiv \hat{R}(\cdot)$. For more complicated $X$'s, $\hat{R}(\cdot)$ could be larger than $R(\cdot)$, although the ratio $\hat{R}(\cdot)/R(\cdot)$ is bounded by $O(\log(K))$, see Section 4.2.3.
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 becomes extremely simple, just
\[ \frac{d}{dt} x_{k\ell}(t) = -(\omega_k^2 + \omega_\ell^2) x_{k\ell}(t) \Rightarrow x_{k\ell}(t) = \exp\{-(\omega_k^2 + \omega_\ell^2)t\} x_{k\ell}. \]

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:
  \[ \mathcal{X} = \{ x \in \mathbb{R}^{(2N-1)\times(2N-1)} : \|x\|_2^2 \leq R^2 \}. \]
  \[ (4.13) \]

- Let the measurements of the temperature at time \( t_1 \) be taken along the points \( P_{i(\nu),j(\nu)}, 1 \leq \nu \leq m \), and let them be affected by a \( \mathcal{N}(0, \sigma^2 I_m) \)-noise, so that our observation is
  \[ \omega = A(x) + \xi, \quad \xi \sim \mathcal{N}(0, \sigma^2 I_m). \]
  Here \( x \mapsto A(x) \) is the linear mapping from \( \mathbb{R}^{(2N-1)\times(2N-1)} \) into \( \mathbb{R}^m \) given by
  \[ [A(x)]_{i,\ell} = \sum_{k,\ell=1}^{2N-1} e^{-\left(\omega_k^2 + \omega_\ell^2\right)t_1} \phi_k(p_{i(\nu)}) \phi_\ell(p_{j(\nu)}) x_{k\ell}(0). \]
  \[ (4.14) \]

- 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 \( \mathbb{R}^{(2N-1)\times(2N-1)} \) into \( \mathbb{R}^{(2K-1)\times(2K-1)} \) is given by
  \[ [B(x)]_{ij} = \sum_{k,\ell=1}^{2N-1} e^{-\left(\omega_k^2 + \omega_\ell^2\right)t_0} \phi_k(r_i) \phi_\ell(r_j) x_{k\ell}(0). \]

**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 (4.14), 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 also very small, provided that $t_0$ is not too small, implying that there is no necessity to recover well high frequency components, unless they are huge. Our linear estimate, roughly speaking, seeks for the best tradeoff between these two opposite phenomena, utilizing (4.13) 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 $63 \times 63$ square grid, and we want to recover the temperatures at time 0.01 along $11 \times 11$ square grid). We used $R = 15$, that is,

$$X = \{[x_{k\ell}]_{k,\ell=1}^{63} : \sum_{k,\ell} x_{k\ell}^2 \leq 225\},$$

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 $t_0$ resulting from a randomly selected signal $x \in X$ along with the recovery $\hat{U}(t_0)$ of $U_*$ by the optimal linear estimate and the naive “least squares” recovery $\tilde{U}(t_0)$ of $U_*$. The latter is defined as $B(x_*)$, where $x_*$ is the least squares recovery of signal underlying observation $\omega$:

$$x = x_*(\omega) := \arg\min_x \| A(x) - \omega \|_2.$$ 

Notice the dramatic difference in performances of the “naive least squares” and the optimal linear estimate.
4.2.2.2 Near-optimality of $\hat{x}_{H}$.

**Proposition 4.5.** The efficiently computable linear estimate $\hat{x}_{H}(\omega) = H^{T}_x \omega$ yielded by an optimal solution to the optimization problem (4.12) is nearly optimal in terms of its risk:

$$\text{Risk}[\hat{x}_{H}, \mathcal{X}] \leq \sqrt{\text{Opt}} \leq 64\sqrt{45 \ln 2 (\ln K + 5 \ln 2)} \text{Risk}_{\text{opt}}[\mathcal{X}],$$  \hspace{1cm} (4.15)

where the minimax optimal risk $\text{Risk}_{\text{opt}}[\mathcal{X}]$ is given by (4.7).

For proof, see Section 4.8.5. Note that the “nonoptimality factor” in (4.15) depends logarithmically on $K$ and is completely independent on what are $A$, $B$, $\Gamma$ and the “details” $R_k$, $\mathcal{T}$, see (4.9), specifying ellitope $\mathcal{X}$.

4.2.2.3 Relaxing the symmetry requirement

Sets $\mathcal{X}$ of the form (4.6) – we called them ellitopes – are symmetric w.r.t. the origin convex compact sets 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 (4.6) that the convex compact set $\mathcal{X}$ known to contain the signals $x$ underlying observations (4.5) can be “sandwiched” by two known to us and similar to each other, with coefficient $\alpha \geq 1$, ellitopes:

$$\{x \in \mathbb{R}^n : \exists(y \in \mathbb{R}^n, t \in \mathcal{T}) : x = Py \& y^T R_k y \leq t_k, 1 \leq k \leq K\} \subset \mathcal{X} \subset \alpha \mathcal{X},$$

with $R_k$ and $\mathcal{T}$ possessing the properties postulated in Section 4.2.1.1. Let Opt and $H_s$ be the optimal value and optimal solution of the optimization problem (4.12) associated with the data $R_1, \ldots, R_K$, $\mathcal{T}$ and matrices $\bar{A} = AP$, $\bar{B} = BP$ in the role of $A$, $B$, respectively. It is immediately seen that the risk $\text{Risk}[\hat{x}_{H}, \mathcal{X}]$ of the linear estimate $\hat{x}_{H}(\omega)$ is at most $\alpha \sqrt{\text{Opt}}$. On the other hand, we have $\text{Risk}_{\text{opt}}[\mathcal{X}] \leq \text{Risk}_{\text{opt}}[\mathcal{X}]$, and by Proposition 4.5 also $\sqrt{\text{Opt}} \leq O(1) \sqrt{\ln(2K)} \text{Risk}_{\text{opt}}[\mathcal{X}]$. Taken together, these relations imply that

$$\text{Risk}[\hat{x}_{H}, \mathcal{X}] \leq O(1) \alpha \sqrt{\ln(2K)} \text{Risk}_{\text{opt}}[\mathcal{X}].$$  \hspace{1cm} (4.16)

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 $\alpha$ 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 $\{x : (x - c_k)^T R_k (x - c_k) \leq 1\}$, $R_k \succeq 0$, concentric or not, is bounded and has a nonempty interior, $\mathcal{X}$ can be approximated by an ellipsoid within the factor $\alpha = K + 2\sqrt{K}$.\footnote{namely, setting $F(x) = -\sum_{k=1}^{K} \ln(1 - (x - c_k)^T R_k (x - c_k)) : \text{int} \mathcal{X} \to \mathbb{R}$ and denoting by $\bar{x}$ the analytic center $\arg \min_{x \in \text{int} \mathcal{X}} F(x)$, one has

$$\{x : (x - \bar{x})^T F''(\bar{x}) (x - \bar{x}) \leq 1\} \subset \mathcal{X} \subset \{x : (x - \bar{x})^T F''(\bar{x}) (x - \bar{x}) \leq K + 2\sqrt{K}\}.$$}

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 (4.16) with
O(1)K in the role of α.

4.2.2.4 Comments

Note that bound (4.16) 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 $\mathbb{R}^n$ 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 $A = B = I$, $\Gamma = \sigma^2I$, in the range $\sigma \leq n\sigma^2 \leq 1$ of values of $n$ and $\sigma$, we have

$$\text{Risk}_{\text{opt}}[\mathcal{X}] \approx \sqrt{\sigma}, \quad \text{Risk}_{\text{opt}}[\hat{x}_H|\mathcal{X}] = O(1)\sqrt{n\sigma},$$

with $\approx$ meaning “up to logarithmic in $n/\sigma$ factor.” In other words, for large $n\sigma$ linear estimates indeed are significantly (albeit not to the full extent of (4.16)) outperformed by nonlinear ones.

Another “bad for linear estimates” situation suggested by (4.15) is the one where the description (4.6) of $\mathcal{X}$, albeit possible, requires a very large value of $K$. Here again (4.15) reflects to some extent the reality: when $\mathcal{X}$ is the unit $\|\cdot\|_1$ ball in $\mathbb{R}^n$, (4.6) takes place with $K = 2^{n-1}$; consequently, the factor at $\text{Risk}_{\text{opt}}[\mathcal{X}]$ in the right hand side of (4.15) becomes at least $\sqrt{n}$. On the other hand, with $A = B = I$, $\Gamma = \sigma^2I$, in the range $\sigma \leq n\sigma^2 \leq 1$ of values of $n$, $\sigma$, the risks $\text{Risk}_{\text{opt}}[\mathcal{X}]$, $\text{Risk}_{\text{opt}}[\hat{x}_H|\mathcal{X}]$ are basically the same as in the case of $\mathcal{X}$ being the perfect simplex inscribed into the unit sphere in $\mathbb{R}^n$, and linear estimates indeed are “heavily non-optimal” when $n\sigma$ is large.

4.2.2.5 How near is “near-optimal:” numerical illustration

The “non-optimality factor” $\theta$ in the upper bound $\sqrt{\text{Opt}} \leq \theta\text{Risk}_{\text{opt}}[\mathcal{X}]$ from Proposition 4.5, while logarithmic, seems to be unpleasantly large. On closer inspection, one can get numerically less conservative bounds on non-optimality factors. Here are some illustrations. In the six experiments to be reported, we used $n = m = \nu = 32$ and $\Gamma = \sigma^2I_m$. In the first triple of experiments, $\mathcal{X}$ was the ellipsoid

$$X = \{x \in \mathbb{R}^{32} : \sum_{j=1}^{32} j^2x_j^2 \leq 1\},$$

that is, $P$ was the identity, $K = 1$, $R_1 = \sum_{j=1}^{32} j^2e_j^Te_j^T$ ($e_j$ are basic orths), and $T = [0,1]$. In the second triple of experiments, $\mathcal{X}$ was the box circumscribed around the above ellipsoid:

$$X = \{x \in \mathbb{R}^{32} : j|x_j| \leq 1, 1 \leq j \leq 32\}$$

$$[P = I, K = 32, R_k = k^2e_ke_k^T, k \leq K, T = [0,1]^K].$$

In these experiments, $B$ was the identity matrix, and $A$ was a common for all experiments randomly rotated matrix with singular values $\lambda_j$, $1 \leq j \leq 32$, forming a geometric progression, with $\lambda_1 = 1$ and $\lambda_{32} = 0.01$. Experiments in a triple differed by the values of $\sigma$ (0.01, 0.001, 0.0001).

The results of the experiments are presented in Table 4.1, where, as above, $\sqrt{\text{Opt}}$ is the given by (4.12) upper bound on the risk $\text{Risk}[\hat{x}_H|\mathcal{X}]$ of recovering $Bx = x$, $x \in X$, by the linear estimate yielded by (4.8) and (4.12), and LwB is the lower bound on $\text{Risk}_{\text{opt}}[\mathcal{X}]$ computed via the techniques outlined in Exercise
Table 4.1: Performance of linear estimates (4.8), (4.12), $m = n = 32$, $B = I$.

4.22 (we skip the details). Whatever might be your attitude to the “reality” as reflected by the data in Table 4.1, this reality is much better than the theoretical upper bound on $\theta$ appearing in (4.15).

4.2.3 Byproduct on semidefinite relaxation

We are about to present an important by its own right byproduct of the reasoning underlying Proposition 4.5. This byproduct is not directly related to Statistics; it relates to the quality of the standard semidefinite relaxation. Specifically, given a quadratic from $x^TCx$ and an ellitope $X$ represented by (4.6), consider the problem

$$\text{Opt}^\ast = \max_{x \in X} x^TCx = \max_y \{ y^TP^TCPy : \exists t \in T : y^TR_ky \leq t_k, k \leq K \}.$$ (4.17)

This problem can be NP-hard (this is already so when $X$ is the unit box and $C$ a general-type positive semidefinite matrix); however, Opt admits an efficiently computable upper bound given by semidefinite relaxation as follows: whenever $\lambda \geq 0$ is such that

$$P^TCP \preceq \sum_{k=1}^K \lambda_k R_k,$$

for $y \in \tilde{X} := \{ y : \exists t \in T : y^TR_ky \leq t_k, k \leq K \}$ we clearly have

$$[Py]^TCPy \leq \sum_k \lambda_k y^TR_ky \leq \phi_T(\lambda)$$

where the last $\leq$ is due to the fact that the vector with the entries $y^TR_ky$, $1 \leq k \leq K$, belongs to $T$. As a result, the efficiently computable quantity

$$\text{Opt} = \min_{\lambda} \left\{ \phi_T(\lambda) : \lambda \geq 0, P^TCP \preceq \sum_{k} \lambda_k R_k \right\}$$ (4.18)

is an upper bound on $\text{Opt}^\ast$. We have the following

**Proposition 4.6.** Let $C$ be a symmetric $n \times n$ matrix and $X$ be given by ellitopic representation (4.6), and let $\text{Opt}^\ast$ and $\text{Opt}$ be given by (4.17) and (4.18). Then

$$\frac{\text{Opt}}{3 \ln(\sqrt{3}K)} \leq \text{Opt}^\ast \leq \text{Opt.}$$ (4.19)
4.3 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.

4.3.1 Spectratopes: definition and examples

We call a set $X \subset \mathbb{R}^n$ a basic spectratope, if it admits simple spectratopic representation – representation of the form

$$X = \{ x \in \mathbb{R}^n : \exists t \in T : R_k^2[x] \preceq t_k I_{d_k}, 1 \leq k \leq K \} \quad (4.20)$$

where

S.1. $R_k[x] = \sum_{i=1}^{n} x_i R^{ki}$ are symmetric $d_k \times d_k$ matrices linearly depending on $x \in \mathbb{R}^n$ (i.e., “matrix coefficients” $R^{ki}$ belong to $\mathbb{S}^n$).

S.2. $T \in \mathbb{R}^K_+$ is the set with the same properties as in the definition of an ellitope, that is, $T$ is a convex compact subset of $\mathbb{R}^K_+$ which contains a positive vector and is monotone:

$$0 \leq t' \leq t \in T \Rightarrow t' \in T.$$

S.3. Whenever $x \neq 0$, it holds $R_k[x] \neq 0$ for at least one $k \leq K$.

An immediate observation is as follows:

**Remark 4.7.** By Schur Complement Lemma, the set (4.20) given by data satisfying S.1-2 can be represented as

$$X = \{ x \in \mathbb{R}^n : \exists t \in T : \begin{bmatrix} R_k[x] & I_{d_k} \\ R_k^2[x] & t_k I_{d_k} \end{bmatrix} \succeq 0, 1 \leq k \leq K \}.$$

By the latter representation, $X$ 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 $X \subset \mathbb{R}^\nu$ is a set represented as linear image of a basic spectratope:

$$X = \{ x \in \mathbb{R}^\nu : \exists (y \in \mathbb{R}^n, t \in T) : x = Py, R_k^2[y] \preceq t_k I_{d_k}, 1 \leq k \leq K \}, \quad (4.21)$$

where $P$ is a $\nu \times n$ matrix, and $R_k[\cdot], T$ are as in S.1-3.

We associate with a basic spectratope (4.20), S.1-3 the following entities:

1. The size

$$D = \sum_{k=1}^{K} d_k;$$
2. Linear mappings

\[ Q \mapsto R_k[Q] = \sum_{i,j} Q_{ij} R^{ki} R^{kj} : S^n \to S^{dk} \]

As is immediately seen, we have

\[ R_k[xx^T] \equiv R^2_k[x], \]  

(4.22)

implying that \( R_k[Q] \geq 0 \) whenever \( Q \geq 0 \), whence \( R_k[\cdot] \) is \( \preceq \)-monotone:

\[ Q' \succeq Q \Rightarrow R_k[Q'] \succeq R_k[Q]. \]  

(4.23)

Besides this, we have

\[ Q \succeq 0 \Rightarrow E_{\xi \sim N(0, Q)} \{ R^2_k[\xi] \} = E_{\xi \sim N(0, Q)} \{ R^2_k[\xi^T] \} = R^2_k[Q], \]  

(4.24)

where the first equality is given by (4.22).

3. Linear mappings \( \Lambda_k \mapsto R^*_k[\Lambda_k] : S^{dk} \to S^n \) given by

\[ [R^*_k[\Lambda_k]]_{ij} = \frac{1}{2} \text{Tr}(\Lambda_k[R^{ki} R^{kj} + R^{kj} R^{ki}]), 1 \leq i, j \leq n. \]  

(4.25)

It is immediately seen that \( R^*_k[\cdot] \) is the conjugate of \( R_k[\cdot] \):

\[ \langle \Lambda_k, R_k[Q] \rangle_F = \text{Tr}(\Lambda_k R_k[Q]) = \text{Tr}(R^*_k[\Lambda_k]Q) = \langle R^*_k[\Lambda_k], Q \rangle_F, \]  

(4.26)

where \( \langle A, B \rangle_F = \text{Tr}(AB) \) is the Frobenius inner product of symmetric matrices. Besides this, we have

\[ \Lambda_k \succeq 0 \Rightarrow R^*_k[\Lambda_k] \geq 0. \]  

(4.27)

Indeed, \( R^*_k[\Lambda_k] \) is linear in \( \Lambda_k \), so that it suffices to verify (4.27) for dyadic matrices \( \Lambda_k = ff^T \); for such a \( \Lambda_k \), (4.25) reads

\[ (R^*_k[ff^T])_{ij} = [R^{ki} f]^T [R^{kj} f], \]

that is, \( R^*_k[ff^T] \) is a Gram matrix and as such is \( \succeq 0 \). Another way to arrive at (4.27) is to note that when \( \Lambda_k \succeq 0 \) and \( Q = xx^T \), the first quantity in (4.26) is nonnegative by (4.22), and therefore (4.26) states that \( x^T R^*_k[\Lambda_k] x \geq 0 \) for every \( x \), implying \( R^*_k[\Lambda_k] \succeq 0 \).

4. The linear space \( \Lambda^K = S^{d_1} \times \ldots \times S^{d_K} \) of all ordered collections \( \Lambda = \{ \Lambda_k \in S^{dk} \}_{k \leq K} \) along with the linear mapping

\[ \Lambda \mapsto \lambda[\Lambda] := [\text{Tr}(\Lambda_1); \ldots; \text{Tr}(\Lambda_K)] : \Lambda^K \to R^K. \]

4.3.1.1 Examples of spectratopes

Example: Ellitopes. Every ellitope

\[ X = \{ x \in R^r : \exists (y \in R^n, t \in T) : x = Py, y^T R_k y \leq t_k, k \leq K \} \]

\[ [R_k \succeq 0, \sum_k R_k \succ 0] \]
is a spectratope as well. Indeed, let \( R_k = \sum_{j=1}^{p_k} r_{kj}^T r_{kj} \), \( p_k = \text{Rank}(R_k) \), be a dyadic representation of the positive semidefinite matrix \( R_k \), so that
\[
y^T R_k y = \sum_j (r_{kj}^T y)^2 \quad \forall y,
\]
and let
\[
\hat{T} = \{ \{ t_{kj} \geq 0, 1 \leq j \leq p_k, 1 \leq k \leq K \} : \exists t \in T : \sum_j t_{kj} \leq t_k \},
\]
\[
R_{kj}[y] = r_{kj}^T y \in S^1 = R.
\]
We clearly have
\[
X = \{ x \in \mathbb{R}^\nu : \exists \{ t_{kj} \} \in \hat{T}, y : x = P y, R_{kj}[y] \leq t_{kj} I_k \forall k, j \}
\]
and the right hand side is a legitimate spectratopic representation of \( X \).

**Example:** “Matrix box.” Let \( L \) be a positive definite \( d \times d \) matrix. Then the “matrix box”
\[
X = \{ X \in S^d : -L \preceq X \preceq L \} = \{ X \in S^d : -I_d \preceq L^{-1/2} X L^{-1/2} \preceq I_d \}
\]
is a basic spectratope (augment \( R_1[\cdot] := R[\cdot] \) with \( K = 1, T = [0, 1] \)). As a result, a bounded set \( X \subset \mathbb{R}^\nu \) given by a system of “two-sided” Linear Matrix Inequalities, specifically,
\[
X = \{ x \in \mathbb{R}^\nu : \exists t \in T : -\sqrt{t_k} L_k \preceq S_k[x] \preceq \sqrt{t_k} L_k, k \leq K \}
\]
where \( S_k[x] \) are symmetric \( d_k \times d_k \) matrices linearly depending on \( x, L_k > 0 \) and \( T \) satisfies S.2, is a basic spectratope:
\[
X = \{ x \in \mathbb{R}^\nu : \exists t \in T : R_k^2[x] \leq t_k I_{d_k}, k \leq K \} \quad [R_k[x] = L_k^{-1/2} S_k[x] L_k^{-1/2}]
\]
Same as ellitopes, spectratopes admit fully algorithmic calculus, see Section 4.6.

### 4.3.2 Semidefinite relaxation on spectratopes

Now let us extend to our current situation Proposition 4.6. The extension reads as follows:

**Proposition 4.8.** Let \( C \) be a symmetric \( n \times n \) matrix and \( X \) be given by spectratopic representation
\[
X = \{ x \in \mathbb{R}^n : \exists y \in \mathbb{R}^\mu, t \in T : x = P y, R_{kj}[y] \leq t_{kj} I_{d_k}, k \leq K \}, \tag{4.28}
\]
let
\[
\text{Opt}_* = \max_{x \in X} x^T C x
\]
and
\[
\text{Opt} = \min_{\Lambda = \{ \Lambda_k \}_{k \leq K}} \{ \phi_T(\lambda[\Lambda]) : \Lambda_k \succeq 0, P^T C P \preceq \sum_k R_k[\Lambda_k] \}
\]
\[
\lambda[\Lambda] = [\text{Tr}(\Lambda_1); \ldots; \text{Tr}(\Lambda_K)] \tag{4.29}
\]
Then (4.29) is solvable, and

\[ \text{Opt}_* \leq \text{Opt} \leq 2 \max[\ln(2D), 1] \text{Opt}_* \leq \sum_k d_k. \] (4.30)

Let us verify the easy and instructive part of the proposition, namely, the left inequality in (4.30); remaining claims will be proved in Section 4.8.3.

The left inequality in (4.30) is readily given by the following

**Lemma 4.9.** Let \( X \) be spectratope (4.28) and \( Q \in S^n \). Whenever \( \Lambda_k \in S_{d_k}^{+} \) satisfy

\[ PQP \preceq \sum_k R_k^*[\Lambda_k], \]

for all \( x \in X \) we have

\[ x^T Q x \leq \phi_T(\lambda[A]), \quad \lambda[A] = [\text{Tr}(\Lambda_1); \ldots; \text{Tr}(\Lambda_K)]. \]

**Proof of the lemma:** Let \( x \in X \), so that for some \( t \in T \) and \( y \) it holds

\[ x = Py, \quad R_k^2[y] \preceq t_k I_{d_k} \forall k \leq K \]

Consequently,

\[ x^T Q x = y^T P^T Q P y \leq y^T \sum_k R_k^*[\Lambda_k]y = y^T \sum_k \text{Tr}(R_k^*[\Lambda_k][yy^T]) \]

\[ = \sum_k \text{Tr}(\Lambda_k R_k[yy^T]) \quad \text{by (4.26)} \]

\[ = \sum_k \text{Tr}(\Lambda_k R_k^2[y]) \quad \text{by (4.22)} \]

\[ \leq \sum_k t_k \text{Tr}(\Lambda_k I_{d_k}) \quad \text{[since } \Lambda_k \preceq 0 \text{ and } R_k^2[y] \preceq t_k I_{d_k}] \]

\[ \leq \phi_T(\lambda[A]) \]

**4.3.3 Linear estimates beyond ellitopic signal sets and \( \| \cdot \|_2 \)-risk**

In Section 4.2, 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 \( X \) in the case when the (squared) risk is defined as the worst, w.r.t. \( x \in X \), expected squared Euclidean norm \( \| \cdot \|_2 \) of the recovery error. We are about to extend these results to the case when \( X \) is a spectratope, and the norm used to measure the recovery error, while not being completely arbitrarily, is not necessarily \( \| \cdot \|_2 \). Besides this, in what follows we also relax our assumptions on observation noise.

**4.3.3.1 Situation and goal**

We consider the problem of recovering the image \( Bx \in R^\nu \) of a signal \( x \in R^n \) known to belong to a given spectatope

\[ X = \{ x \in R^n : \exists t \in T : R_k^2[x] \preceq t_k I_{d_k}, 1 \leq k \leq K \} \]

from noisy observation

\[ \omega = Ax + \xi, \] (4.31)

where \( A \) is a known \( m \times n \) matrix, and \( \xi \) is random observation noise.
Observation noise. In typical signal processing applications, the distribution of noise is fixed and is a part of the data of the estimation problem. In order to cover some applications (e.g., the one in Section 4.3.3.7), we allow for “ambiguous” noise distributions; all we know is that this distribution belongs to a family $\mathcal{P}$ of Borel probability distributions on $\mathbb{R}^m$ associated with a given convex compact subset $\Pi$ of the interior of the cone $\mathbb{S}_+^m$ 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 $\Pi$:

$$P \in \mathcal{P} \Rightarrow \exists Q \in \Pi : \text{Var}[P] := E_{\xi \sim P}\{\xi^T\xi\} \preceq Q.$$ (4.32)

The actual distribution of noise in (4.31) is somehow selected from $\mathcal{P}$ by nature (and may, e.g., depend on $x$).

In the sequel, for a probability distribution $P$ on $\mathbb{R}^m$ we write $P \ll \Pi$ to express the fact that the matrix of second moments of $P$ is $\succeq$-dominated by a matrix from $\Pi$:

$$\{P \ll \Pi\} \Leftrightarrow \{\exists \Theta \in \Pi : \text{Var}[P] \preceq \Theta\}.$$ (4.33)

Quantifying risk. Given $\Pi$ and a norm $\| \cdot \|$ on $\mathbb{R}^\nu$, we quantify the quality of a candidate estimate $\hat{\xi}(\cdot) : \mathbb{R}^m \to \mathbb{R}^\nu$ by its $(\Pi, \| \cdot \|)$-risk on $X$ defined as

$$\text{Risk}_{\Pi, \| \cdot \|}[\hat{\xi}]|X| = \sup_{x \in X, P \ll \Pi} E_{\xi \sim P}\{\|\hat{\xi}(Ax + \xi) - Bx\|\}.$$

Goal. As before, our focus is on linear estimates – estimates of the form

$$\hat{\xi}_H(\omega) = H^T \omega$$

given by $m \times \nu$ matrices $H$. Our goal is to demonstrate that under some restrictions on the signal domain $X$, “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.

4.3.3.2 Assumptions

Preliminaries: conjugate norms. Recall that a norm $\| \cdot \|$ on a Euclidean space $\mathcal{E}$, e.g., on $\mathbb{R}^k$, 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

$$\langle x, y \rangle \leq \|x\| \|y\|_* \ \forall x, y.$$ (4.33)

It is well known that taken twice, norm conjugation recovers the initial norm: $\langle \| \cdot \|_*\rangle_*$ is exactly $\| \cdot \|; \ \text{in other words,}$

$$\|x\| = \max_y \{\langle x, y \rangle : \|y\|_* \leq 1\}.$$
The standard examples are the conjugates to the standard $\ell_p$-norms on $E = \mathbb{R}^k$, $p \in [1, \infty]$; 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 (4.33) is called Hölder inequality—an extension of the Cauchy-Schwartz inequality dealing with the case $\| \cdot \| = \| \cdot \|_{1^*} = \| \cdot \|_2$.

**Assumptions.** From now on we make the following assumptions:

**Assumption A:** The unit ball $B_*$ of the norm $\| \cdot \|_*$ conjugate to the norm $\| \cdot \|$ in the formulation of our estimation problem is a spectratope:

$$B_* = \{ z \in \mathbb{R}^\nu : \exists y \in \mathcal{Y} : z = My \},$$

$$\mathcal{Y} := \{ y \in \mathbb{R}^q : \exists r \in \mathcal{R} : S_\ell^2[y] \preceq r I_{\ell f}, 1 \leq \ell \leq L \},$$

where the right hand side data are as required in a spectratopic representation. (4.34)

Note that Assumption A is satisfied when $\| \cdot \| = \| \cdot \|_p$ with $p \in [1, 2]$; in this case,

$$B_* = \{ u \in \mathbb{R}^\nu : \| u \|_{p^*} \leq 1 \}, \quad p^* = \frac{p}{p-1} \in [2, \infty],$$

so that $B_*$ is an ellitope, see Section 4.2.1.1, and thus is a spectratope. Another potentially useful example of norm $\| \cdot \|$ which obeys Assumption A is the nuclear norm $\| V \|_{Sh, 1}$ on the space $\mathbb{R}^\nu = \mathbb{R}^{p \times q}$ of $p \times q$ matrices—the sum of singular values of a matrix $V$. In this case the conjugate norm is the spectral norm $\| \cdot \| = \| \cdot \|_{2,2}$ on $\mathbb{R}^\nu = \mathbb{R}^{p \times q}$, and the unit ball of the latter norm is a spectratope:

$$\{ X \in \mathbb{R}^{p \times q} : \| X \| \leq 1 \} = \{ X : \exists t \in \mathcal{T} = [0, 1] : R^2[X] \preceq t I_{p+q} \},$$

$$R[X] = \left[ \begin{array}{c|c} X^T \end{array} \right].$$

Besides Assumption A, we make

**Assumption B:** The signal set $\mathcal{X}$ is a basic spectratope:

$$\mathcal{X} = \{ x \in \mathbb{R}^n : \exists t \in \mathcal{T} : R^2_k[x] \preceq t_k I_{d_k}, 1 \leq k \leq K \},$$

where the right hand side data are as required in a spectratopic representation.

**Note:** Similar to what we have observed in Section 4.2.1.3 in the case of ellitopes, the situation where the signal set is a general type spectratope can be straightforwardly reduced to the one where $\mathcal{X}$ is a basic spectratope.

In addition we make the following regularity assumption:

**Assumption R:** All matrices from $\Pi$ are positive definite.
4.3.3.3 Building linear estimate

Let $H \in \mathbb{R}^{m \times \nu}$. We clearly have

$$
\text{Risk}_{\Pi, \|\cdot\|}[\hat{X}_H(\cdot)|\mathcal{X}] = \sup_{x \in \mathcal{X}, P \in \Pi} E_{\xi \sim P} \{\|B - H^T A\|_2^\nu \|x - H^T \xi\|\} \\
\leq \sup_{x \in \mathcal{X}} \|B - H^T A\|_2^\nu + \sup_{P \in \Pi} E_{\xi \sim P} \{\|H^T \xi\|\}
$$

where

$$
\|V\|_{\mathcal{X}, \|\cdot\|} = \max_{x \in \mathcal{X}} \{\|Vx\| : x \in \mathcal{X}\} : \mathbb{R}^{\nu \times n} \rightarrow \mathbb{R},
$$

$$
\Psi_{\Pi}(H) = \sup_{P \in \Pi} E_{\xi \sim P} \{\|H^T \xi\|\}.
$$

Same as in Section 4.2.2, we need to derive efficiently computable convex upper bounds on the norm $\| \cdot \|_{\mathcal{X}, \|\cdot\|}$ and the function $\Psi_{\Pi}$, which by themselves, while being convex, can be difficult to compute.

4.3.3.4 Upper-bounding $\| \cdot \|_{\mathcal{X}, \|\cdot\|}$

With Assumptions $A$, $B$ in force, consider the spectratorpe $Z := \mathcal{X} \times \mathcal{Y} = \{[x; y] \in \mathbb{R}^\nu \times \mathbb{R}^\nu : \exists s = [t; r] \in \mathcal{T} \times \mathcal{R} : \}

\begin{aligned}
& R^2_{x}[x] \leq t_k I_d, \quad 1 \leq k \leq K, S^2_i[y] \geq r_I f_i, \quad 1 \leq I \leq L \\
& = \{w = [x; y] \in \mathbb{R}^\nu \times \mathbb{R}^\nu : \exists s = [t; r] \in \mathcal{T} \times \mathcal{R} : U^2_{i}[w] \leq s_I I_d, \quad 1 \leq i \leq I = K + L \}
\end{aligned}

with $U_i[\cdot]$ readily given by $R_k[\cdot]$ and $S_i[\cdot]$. Given a $\nu \times n$ matrix $V$ and setting

$$
W[V] = \frac{1}{2} \left[ \begin{array}{c} V^T M \\ M^T V \end{array} \right]
$$

we have

$$
\|V\|_{\mathcal{X}, \|\cdot\|} = \max_{x \in \mathcal{X}} \|Vx\| = \max_{x \in \mathcal{X}, y \in \mathcal{Y}} z^T V x = \max_{x \in \mathcal{X}, y \in \mathcal{Y}} y^T M^T V x = \max_{w \in Z} w^T W[V] w.
$$

Applying Proposition 4.8, we arrive at the following

**Corollary 4.10.** In the just defined situation, the efficiently computable convex function

$$
\|V\|_{\mathcal{X}, \|\cdot\|}^* = \min_{\Lambda, \Upsilon} \left\{ \phi_\Upsilon(\lambda[\Lambda]) + \phi_\Lambda(\lambda[\Upsilon]) : \right\}

\begin{align*}
\Lambda &= \{\Lambda_k \in S^d_{+} \}_{k \in K}, \\
\Upsilon &= \{\Upsilon_{\ell} \in S^d_{+} \}_{\ell \in L}, \\
& \left[ \sum_{k} R^*_{a}[\Lambda_k] \text{ } \frac{1}{2} \left[ V^T M \right] \left[ \sum_{\ell} S^*_{\ell}[\Upsilon_{\ell}] \right] \right] \geq 0
\end{align*}

\begin{align*}
\phi_\Lambda(\lambda) &= \max_{t \in T} \lambda^T t, \quad \phi_\Upsilon(\lambda) = \max_{r \in R} \lambda^T r, \\
& [R^*_{a}[\Lambda_k]]_{ij} = \frac{1}{2} \text{Tr}(\Lambda_k [R^k_i R^k_j + R^k_j R^k_i]), \text{ where } R_k[x] = \sum_{i} x_i R^k_i, \\
& [S^*_{\ell}[\Upsilon_{\ell}]]_{ij} = \frac{1}{2} \text{Tr}(\Upsilon_{\ell} [S^\ell_i S^\ell_j + S^\ell_j S^\ell_i]), \text{ where } S_i[y] = \sum_{j} y_j S^i_j
\end{align*}

is a norm on \( \mathbb{R}^{\nu \times n} \), and this norm is a tight upper bound on \( \| \cdot \|_{\mathcal{X}, \| \cdot \|_r} \), namely,
\[
\forall V \in \mathbb{R}^{\nu \times n} : \| V \|_{\mathcal{X}, \| \cdot \|_r} \leq \| V \|_{\mathcal{X}, \| \cdot \|_r}^+ \leq 2 \max \{ \ln(2D), 1 \} \| V \|_{\mathcal{X}, \| \cdot \|_r},
\]
where \( D = \sum_k d_k + \sum_{\ell} f_{\ell} \).

### 4.3.3.5 Upper-bounding \( \Psi_\Pi(\cdot) \)

The next step is to derive an efficiently computable convex upper bound on the function \( \Psi_\Pi \) stemming from a norm obeying Assumption B. The underlying observation is as follows:

**Lemma 4.11.** Let \( V \) be a \( m \times \nu \) matrix, \( Q \in \mathbb{S}_+^m \), and \( P \) be a probability distribution on \( \mathbb{R}^m \) with \( \text{Var}[P] \preceq Q \). Let, further, \( \| \cdot \| \) be a norm on \( \mathbb{R}^\nu \) with the unit ball \( B^* \) of the conjugate norm \( \| \cdot \|_* \) given by (4.34). Finally, let \( \Upsilon = \{ \Upsilon_\ell \in \mathbb{S}_{f_\ell}^+ \}_{\ell \leq L} \) and a matrix \( \Theta \in \mathbb{S}_m \) satisfy the constraint
\[
\frac{\Theta}{2} M V \top \left[ \frac{1}{2} V M \sum_{\ell \leq L} \Upsilon_\ell \right] \succeq 0 \tag{4.37}
\]
(for notation, see (4.34), (4.36)). Then
\[
\mathbb{E}_{\eta \sim P} \{ \| V^T \eta \| \} \leq \text{Tr}(Q \Theta) + \phi_R(\lambda[\Upsilon]). \tag{4.38}
\]

**Proof** is immediate. In the case of (4.37), we have
\[
\| V^T \xi \| = \max_{z \in B} \xi^T V^T z \leq \max_{y \in \mathbb{Y}} \xi^T \Theta \xi + \sum_{\ell \leq L} y^T S_{\ell}^* [\Upsilon_\ell] y \tag{by (4.37)}
\]
\[
= \max_{y \in \mathbb{Y}} \xi^T \Theta \xi + \sum_{\ell \leq L} \text{Tr}(S_{\ell}^* [\Upsilon_\ell] y y^T) \tag{by (4.22) and (4.26)}
\]
\[
= \xi^T \Theta \xi + \max_{y \in \mathbb{Y}} \left\{ \sum_{r \in R} \text{Tr}(\Upsilon_r S_{\ell}^2[y]) : S_{\ell}^2[y] \preceq r \ell I_{f_{\ell}}, \ell \leq L, r \in \mathbb{R} \right\} \tag{by (4.34)}
\]
\[
\leq \xi^T \Theta \xi + \max_{r \in \mathbb{R}} \sum_{\ell \leq L} \text{Tr}(\Upsilon_\ell) r_{\ell} \tag{by (4.34)}
\]
\[
\leq \xi^T \Theta \xi + \phi_R(\lambda[\Upsilon]).
\]

Taking expectation of both sides of the resulting inequality w.r.t. distribution \( P \) of \( \xi \) and taking into account that \( \text{Tr}(\text{Var}[P \Theta]) \leq \text{Tr}(Q \Theta) \) due to \( \Theta \succeq 0 \) (by (4.37)) and \( \text{Var}[P] \preceq Q \), we get (4.38). \( \square \)

Note that when \( P = \mathcal{N}(0, Q) \), the smallest upper bound on \( \mathbb{E}_{\eta \sim P} \{ \| V^T \eta \| \} \) which can be extracted from Lemma 4.11 (this bound is efficiently computable) is tight, see Lemma 4.17 below.

An immediate consequence of the bound of the bound in Lemma (4.11) is

**Corollary 4.12.** Let
\[
\Gamma(\Theta) = \max_{Q \in \Pi} \text{Tr}(Q \Theta) \tag{4.39}
\]
\[ \overline{\Psi}_\Pi(H) = \min_{(T_\ell)_{\ell \leq L}, \Theta \in S^n} \left\{ \frac{1}{2} \overline{\Theta} + \phi_R(\lambda[T]) : \sum_{\ell} S_\ell T_{\ell} \Theta \geq 0 \right\} \] (4.40)

Then \( \overline{\Psi}_\Pi(\cdot) : R^{m \times \nu} \rightarrow R \) is an efficiently computable convex upper bound on \( \Psi_\Pi(\cdot) \).

Indeed, given Lemma 4.11, 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 4.44 stating, in particular, that the optimization problem in (4.40) is feasible, combined with the fact that the objective is coercive on the feasible set (i.e., is not bounded from above along every unbounded sequence of feasible solutions).

Remark 4.13. When \( \Upsilon = \{T_\ell\}_{\ell \leq L}, \Theta \) is a feasible solution to the right hand side problem in (4.40) and \( s > 0 \), the pair \( \Upsilon' = \{sT_\ell\}_{\ell \leq L}, \Theta' = s^{-1}\Theta \) also is a feasible solution. Since \( \phi_R(\cdot) \) and \( \Gamma(\cdot) \) are positive homogeneous of degree 1, we conclude that \( \overline{\Psi}_\Pi \) is in fact the infimum of the function

\[ 2\sqrt{\Gamma(\Theta)} \phi_R(\lambda[T]) = \inf_{s > 0} \left( s^{-1} \Gamma(\Theta) + s \phi_R(\lambda[T]) \right) \]

over \( \Upsilon, \Theta \) satisfying the constraints of the problem (4.40).

In addition, for every feasible solution \( \Upsilon = \{T_\ell\}_{\ell \leq L}, \Theta \) to (4.40) with \( \mathcal{M}[\Upsilon] := \sum_{\ell} S_\ell^T T_{\ell} \Theta \geq 0 \), the pair \( \Upsilon, \Theta = \frac{1}{2} \Gamma M M^{-1} [\Upsilon] M^T \Theta \) is feasible for the problem as well, and \( 0 \leq \Theta \leq \Theta \) (Schur Complement Lemma), so that \( \Gamma(\Theta) \leq \Gamma(\Theta) \). As a result,

\[ \overline{\Psi}_\Pi(H) = \inf_{\Upsilon} \left\{ \frac{1}{2} \Gamma(\Theta) + \phi_R(\lambda[T]) : \Upsilon = \{T_\ell \in S^n_{\leq L}, \mathcal{M}[\Upsilon] \geq 0 \} \right\} \] (4.41)

Illustration. Suppose that \( \|u\| = \|u\|_p \) with \( p \in [1, 2] \), and let us apply the just described scheme for upper-bounding \( \Psi_\Pi \), assuming \( \{Q\} \subset \Pi \subset \{S \in S^n_{\geq L} : S \preceq Q\} \) for some given \( Q \succ 0 \), so that \( \Gamma(\Theta) = \text{Tr}(Q \Theta) \), \( \Theta \succeq 0 \). The unit ball of the norm conjugate to \( \|\cdot\|_1 \), that is, the norm \( \|\cdot\|_q, q = \frac{2}{p-1} \in [2, \infty) \), is the basic spectratope (in fact, ellitope)

\[ B_\star = \{y \in R^d : \exists r \in R := \{R^p_\star : \|r\|_q/2 \leq 1\} : \sum_{\ell} S_{\ell}[y] r_{\ell} \leq r, 1 \leq \ell \leq L = \nu\}, \]

As a result, \( \Upsilon \)'s from Remark 4.13 are collections of \( \nu \) positive semidefinite \( 1 \times 1 \) matrices, and we can identify them with \( \nu \)-dimensional nonnegative vectors \( v \), resulting in \( \lambda[T] = v \) and \( \mathcal{M}[\Upsilon] = \text{Diag}\{v\} \). Furthermore, for \( \text{nonnegative} \ v \) we clearly have \( \phi_R(v) = \|v\|_{p/(2-p)} \), so the optimization problem in (4.41) now reads

\[ \overline{\Psi}_\Pi(H) = \inf_{v \in R^\nu} \left\{ \frac{1}{2} \text{Tr}(V \text{Diag}^{-1}\{v\} V^T) + \|v\|_{p/(2-p)} : v > 0 \right\} \quad [V = Q^{1/2}H], \]
and when setting $a_{\ell} = \|\text{Col}_\ell [V]\|_2$, (4.41) becomes

$$
\Psi_\Pi (H) = \inf_{\nu > 0} \left\{ \frac{1}{4} \sum_{\ell} \frac{a_{\ell}^2}{\nu_{\ell}} + \|\nu\|_{p/(2-p)} \right\}.
$$

This results in $\Psi_\Pi (H) = \| [a_1; \ldots; a_p] \|_p$. Recalling what $a_\ell$ and $V$ are, we end up with

$$
\forall P, \text{Var} [P] \preceq Q : \quad E_{\xi \sim P} \{ \| H^T \xi \| \} \leq \Psi_\Pi (H) := \left\| \| \text{Row}_1 [H^T Q^{1/2}] \|_2 ; \ldots ; \| \text{Row}_p [H^T Q^{1/2}] \|_2 \right\|_p.
$$

This result is quite transparent and could be easily obtained straightforwardly. Indeed, when $\text{Var} [P] \preceq Q$, and $\xi \sim P$, the vector $\zeta = H^T \xi$ clearly satisfies $E \{ \zeta_2^2 \} \leq \sigma_2^2 := \| \text{Row}_i [H^T Q^{1/2}] \|_2^2$, implying, due to $p \in [1, 2]$, that $E \{ \sum_i |\zeta_i|^p \} \leq \sum_i \sigma_i^p$, whence $E \{ \|\zeta\|_p \} \leq \| [\sigma_1; \ldots; \sigma_p] \|_p$.

### 4.3.3.6 Putting things together

An immediate outcome of Corollaries 4.10, 4.12 is the following recipe for building “presumably good” linear estimate:

**Proposition 4.14.** In the situation of Section 4.3.3.1 and under Assumptions $A$, $B$, and $R$ (see Section 4.3.3.2) consider the convex optimization problem (for notation, see (4.36) and (4.39))

$$
\begin{align*}
\text{Opt} = \min_{H, A, \Lambda, Y, \Theta} \left\{ &\phi_T (\Lambda [\Lambda]) + \phi_\kappa (\lambda [Y]) + \phi_\kappa (\lambda [Y']) + \Gamma (\Theta) : \\
&\Lambda = \{ \Lambda_k \geq 0, k \leq K \}, \quad Y = \{ Y_\ell \geq 0, \ell \leq L \}, \quad Y' = \{ Y'_\ell \geq 0, \ell \leq L \}, \quad \Theta = \{ \Theta_k \} , \quad \Theta_k = \frac{1}{2} M^T H M \right\} \quad (4.42) \end{align*}
$$

The problem is solvable, and the $H$-component $H_*$ of its optimal solution yields linear estimate $\hat{x}_{H_*} (\omega) = H^T \omega$ such that

$$
\text{Risk}_{\Pi, \|\cdot\|} [\hat{x}_{H_*} (\cdot) | X] \leq \text{Opt}. 
$$

Note that the only claim in Proposition 4.14 which is not an immediate consequence of Corollaries 4.10 and 4.12 is that problem (4.42) is solvable; this fact is readily given by the feasibility of the problem (by Lemma 4.44) and the coerciveness of the objective on the feasible set (recall that $\Gamma (\Theta)$ is coercive on $S^p_+$ due to $\Pi \subset \text{int} S^p_+$ and that $y \mapsto My$ is an onto mapping, since $B_*$ is full-dimensional).

### 4.3.3.7 Illustration: covariance matrix estimation

Suppose that we observe a sample

$$
\eta^T = \{ \eta_k = A \xi_k \}_{k \leq T} \quad (4.44)
$$

where $A$ is a given $m \times n$ matrix, and $\xi_1, \ldots, \xi_T$ are sampled, independently of each other, from zero mean Gaussian distribution with unknown covariance matrix $\varphi$.
known to satisfy
\[ \gamma \vartheta_* \preceq \vartheta \preceq \vartheta_* \]  \tag{4.45}
where \( \gamma \geq 0 \) and \( \vartheta_* > 0 \) are given. Our goal is to recover \( \vartheta \), and the norm on \( \mathbb{S}^n \) in which recovery error is measured satisfies Assumption A.

**Processing the problem.** We can process the just outlined problem as follows.

1. We represent the set \( \{ \vartheta \in \mathbb{S}_+^n : \gamma \vartheta_* \preceq \vartheta \preceq \vartheta_* \} \) as the image of the matrix box
\[ \mathcal{V} = \{ v \in \mathbb{S}^n : \| v \|_{2,2} \leq 1 \} \]  \[ || \cdot ||_{2,2} : \text{spectral norm} \]
under affine mapping, specifically, we set
\[ \vartheta_0 = 1 + \gamma^2 \vartheta_* \]
and treat the matrix
\[ v = \sigma^{-1} \vartheta_*^{-1/2} (\vartheta - \vartheta_0) \vartheta_*^{-1/2} \]
\[ \Leftrightarrow \vartheta = \vartheta_0 + \sigma \vartheta_*^{1/2} v \vartheta_*^{1/2} \]
as the signal underlying our observations. Note that our a priori information on \( \vartheta \) reduces to \( v \in \mathcal{V} \).

2. We pass from observations \( \eta_k \) to “lifted” observations \( \eta_k \eta_k^T \in \mathbb{S}^m \), so that
\[ \mathbb{E}\{ \eta_k \eta_k^T \} = \mathbb{E}\{ A \xi_k \xi_k^T A^T \} = A \vartheta A^T = A \left( \underbrace{\vartheta_0 + \sigma \vartheta_*^{1/2} v \vartheta_*^{1/2}}_{\vartheta_0[v]} \right) A^T, \]
and treat as “actual” observations the matrices
\[ \omega_k = \eta_k \eta_k^T - A \vartheta_0 A^T. \]

We have
\[ \omega_k = \mathcal{A} v + \zeta_k \]
with \( \mathcal{A} v = \sigma \vartheta_*^{1/2} v \vartheta_*^{1/2} A^T \) and \( \zeta_k = \eta_k \eta_k^T - A \vartheta_0[v] A^T. \) \tag{4.46}

Observe that random matrices \( \zeta_1, \ldots, \zeta_T \) are i.i.d. with zero mean and covariance mapping \( \mathcal{Q}[v] \) (that of random matrix-valued variable \( \zeta = \eta \eta^T - \mathbb{E}\{ \eta \eta^T \}, \eta \sim \mathcal{N}(0, A \vartheta_0[v] A^T) \)).

3. Let us \( \succeq \)-upper-bound the covariance mapping of \( \zeta \). Observe that \( \mathcal{Q}[v] \) is a symmetric linear mapping of \( \mathbb{S}^m \) into itself given by
\[ \langle h, \mathcal{Q}[v] h \rangle = \mathbb{E}\{ \langle h, \zeta \rangle^2 \} = \mathbb{E}\{ \langle h, \eta \eta^T \rangle^2 \} - \langle h, \mathbb{E}\{ \eta \eta^T \} \rangle^2, \ h \in \mathbb{S}^m. \]

---

\[ ^8 \text{In our current considerations, we need to operate with linear mappings acting from } \mathbb{S}^p \text{ to } \mathbb{S}^q. \text{ We treat } \mathbb{S}^k \text{ as Euclidean space equipped with the Frobenius inner product } \langle u, v \rangle = \text{Tr}(uv) \text{ and denote linear mappings from } \mathbb{S}^p \text{ into } \mathbb{S}^q \text{ by capital calligraphic letters, like } \mathcal{A}, \mathcal{Q}, \text{ etc. Thus, } \mathcal{A} \text{ in (4.46) denotes the linear mapping which, on closer inspection, maps matrix } v \in \mathbb{S}^p \text{ into the matrix } \mathcal{A} v = A[\vartheta_0[v] - \vartheta_0][A^T] \text{.} \]
Given \( v \in \mathcal{V} \), let us set \( \theta = \theta[v] \), so that \( 0 \leq \theta \leq \theta_* \), and let \( \mathcal{H}(h) = \theta^{1/2} A^T h A \theta^{1/2} \). We have

\[
\langle h, Q[v] h \rangle = E_{\xi \sim N(0, \theta)} \{ \text{Tr}^2(h A \xi \xi^T A^T) \} - \text{Tr}^2(h E_{\xi \sim N(0, \theta)} \{ A \xi \xi^T A^T \}) = E_{\chi \sim N(0, h \theta)} \{ \text{Tr}^2(h \theta^{1/2} \chi \chi^T \theta^{1/2} A^T) \} - \text{Tr}^2(h \theta A \theta A^T)
\]

Thus,

\[
E_{\chi \sim N(0, h \theta)} \{ (\chi^T \mathcal{H}(h) \chi)^2 \} - \text{Tr}^2(\mathcal{H}(h)) = E_{\chi \sim N(0, h \theta)} \{ (\chi^T \mathcal{H}(\xi) \chi)^2 \} - (\sum_i \lambda_i)^2 = E_{\chi \sim N(0, 1)} \{ (\sum_i \lambda_i \chi_i^2)^2 \} - (\sum_i \lambda_i)^2 = \sum_i \lambda_i^2 - (\sum_i \lambda_i)^2 = 2 \sum_i \lambda_i^2 = 2 \text{Tr}(\mathcal{H}(h))^2.
\]

Thus,

\[
\langle h, Q[v] h \rangle = 2 \text{Tr}(\mathcal{H}(h))^2 = 2 \text{Tr}(\theta^{1/2} A^T h A \theta A^T \theta^{1/2}) \leq 2 \text{Tr}(\theta^{1/2} A^T h A \theta A \theta A^T) \leq 2 \text{Tr}(\theta A^T h A \theta A^T) \leq 2 \text{Tr}(\theta A^T h A \theta A^T).
\]

We conclude that

\[
\forall v \in \mathcal{V} : Q[v] \preceq Q, \quad \langle c, Q h \rangle = 2 \text{Tr}(\theta A^T h A \theta A^T c A), c, h \in S^m. \quad (4.47)
\]

4. To continue, we need to set some additional notation to be used when operating with Euclidean spaces \( S^p \), \( p = 1, 2, \ldots \)

- We denote \( \bar{p} = \frac{p(p+1)}{2} = \dim S^p \), \( \mathcal{I}_p = \{ (i, j) : 1 \leq i \leq j \leq p \} \), and for \( i, j \in \mathcal{I}_p \) set

\[
e_p^{ij} = \begin{cases} e_i e_j^T, & i = j \\ \frac{1}{\sqrt{2}} [e_i e_j^T + e_j e_i^T], & i < j \end{cases}
\]

where \( e_i \) are the standard basic orths in \( \mathbb{R}^p \). Note that \( \{e_p^{ij} : (i, j) \in \mathcal{I}_p \} \) is the standard orthonormal basis in \( S^p \). Given \( v \in S^p \), we denote by \( X^p(v) \) the vector of coordinates of \( v \) in this basis:

\[
X^p_{ij}(v) = \text{Tr}(v e_p^{ij}) = \begin{cases} v_{ii}, & i = j \\ \sqrt{2} v_{ij}, & i < j \end{cases}, \quad (i, j) \in \mathcal{I}_p.
\]

Similarly, for \( x \in \mathbb{R}^{\bar{p}} \), we index the entries in \( x \) by pairs \( ij \), \( i, j \in \mathcal{I}_p \), and set \( V^p(x) = \sum_{(i, j) \in \mathcal{I}_p} x_{ij} e_p^{ij} \), so that \( v \mapsto X^p(v) \) and \( x \mapsto V^p(x) \) are inverse to each other linear norm-preserving maps identifying the Euclidean spaces \( S^p \) and \( \mathbb{R}^{\bar{p}} \) (recall that the inner products on these spaces are, respectively, the Frobenius and the standard one).

- Recall that \( \mathcal{V} \) is the matrix box \( \{ v \in S^n : v^2 \preceq I_n \} = \{ v \in S^n : \exists t \in \mathcal{T} := [0, 1] : v^2 \preceq t I_n \} \). We denote by \( \mathcal{X} \) the image of \( \mathcal{V} \) under the mapping \( X^n \):

\[
\mathcal{X} = \{ x \in R^{\bar{p}} : \exists t \in \mathcal{T} : R^2[x] \preceq t I_n \}, \quad R[x] = \sum_{(i, j) \in \mathcal{I}_n} x_{ij} e_n^{ij}, \quad \bar{n} = \frac{1}{2} n(n+1).
\]

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

\[
    w_k = X^n(\omega_k) = X^n(AV^n(x)) + z_k, \quad z_k = X^n(\zeta_k).
\]

Note that \( z_k \in \mathbb{R}^n \), \( 1 \leq k \leq T \), are zero mean i.i.d. random vectors with covariance matrix \( Q[x] \) satisfying, in view of (4.47), the relation

\[
    Q[x] \preceq Q, \quad \text{where } Q_{ij,kl} = 2\text{Tr}(\partial_x A^T e_{ij} A \partial_x A^T e_{kl} A), \quad (i,j) \in \mathcal{I}_m, (k,\ell) \in \mathcal{I}_m.
\]

Our goal is to estimate \( \bar{\theta}[v] - \bar{\theta}[0] \), or, what is the same, to recover

\[
    \bar{B}x := X^n(\bar{\theta}[V^n(x)] - \bar{\theta}[0]).
\]

We assume that the norm in which the estimation error is measured is “transferred” from \( \mathbb{S}^n \) to \( \mathbb{R}^k \); we denote the resulting norm on \( \mathbb{R}^k \) by \( \| \cdot \| \) and assume that the unit ball \( B_x \) of the conjugate norm \( \| \cdot \|_* \) is given by spectratopic representation:

\[
    \{ u \in \mathbb{R}^n : \|u\| \leq 1 \} = \{ u \in \mathbb{R}^n : \exists y \in \mathcal{Y} : u = My \},
\]

\[
    \mathcal{Y} := \{ y \in \mathbb{R}^n : \exists r \in \mathcal{R} : S^2_r \|y\| \leq r \ell I_{\ell \ell}, 1 \leq \ell \leq L \}. \tag{4.48}
\]

The formulated description of the estimation problem fit the premises of Proposition 4.14, specifically:

- the signal \( x \) underlying our observation \( w^{(T)} = [w_1; ...; w_T] \) is known to belong to basic spectratope \( \mathcal{X} \in \mathbb{R}^n \), and the observation itself is of the form

\[
    w^{(T)} = \mathcal{A}^{(T)} x + z^{(T)}, \quad \mathcal{A}^{(T)} = \left[ \mathcal{A}_1; ...; \mathcal{A}_T \right], \quad z^{(T)} = \left[ z_1; ...; z_T \right];
\]

- the noise \( z^{(T)} \) is zero mean, and its covariance matrix is \( \preceq Q_T := \text{Diag}\{Q, ..., Q\} \), which allows to set \( \Pi = \{Q_T\} \);
- our goal is to recover \( \bar{B}x \), and the norm \( \| \cdot \| \) in which the recovery error is measured satisfies (4.48).

Proposition 4.14 supplies the linear estimate

\[
    \hat{x}(w^{(T)}) = \sum_{k=1}^{T} H^T_{xk} w_k,
\]

of \( \bar{B}x \) with \( H_x = [H_1; ...; H_xT] \) stemming from the optimal solution to the convex optimization problem

\[
    \text{Opt} = \min_{H = [H_1; ...; H_T], \Lambda, \mathcal{T}} \left\{ \text{Tr}(\Lambda) + \phi_{\mathcal{R}}(\Lambda | \mathcal{T}) + \bar{V}(Q_T) (H_1, ..., H_T) : \right. \\
    \left. \begin{array}{c}
    \Lambda \in \mathbb{S}^n_+ , \quad \mathcal{T} = \{ \mathcal{T}_\ell \geq 0, \ell \leq L \}, \\
    \left[\frac{\mathcal{R}^* \Lambda}{\frac{1}{2} M^T [B - (\sum_k H_k)^T \mathcal{A}]} \right] \geq 0 \\
    \frac{1}{2} [B - (\sum_k H_k)^T \mathcal{A}] \geq 0 \\
    \sum_\ell S^2_r |\mathcal{T}_\ell| \geq 0
    \end{array} \right\} \tag{4.49}
\]
where
\[ R^*[A] \in \mathbf{S}^n : (R^*[A])_{ij,k\ell} = \text{Tr}(\Lambda e_n^{ij} e_n^{k\ell}) \], \((i,j) \in \mathcal{I}_n, (k,\ell) \in \mathcal{I}_n,\]
and, cf. (4.40),
\[
\Psi(Q_T)(H_1,\ldots,H_T) = \min_{\Theta} \left\{ \text{Tr}(Q_T \Theta) + \phi_R(\lambda[\Theta']) : \Theta \in \mathbf{S}^{mT}, \Theta' = \{\Theta'_\ell \geq 0, \ell \leq L\}, \right. \\
- \left. \begin{bmatrix}
\frac{1}{2}[M' H'_1,\ldots,M' H'_T] & \frac{1}{2}[H_1 M' ;\ldots; H_T M'] \\
\frac{1}{2}[H_1 M' ;\ldots; H_T M'] & \sum_{\ell} S'_\ell [\Theta'_\ell]
\end{bmatrix} \right\} \geq 0 \}
\]

5. Evidently, the function \( \Psi(Q_T)([H_1,\ldots,H_T]) \) remains intact when permuting \( H_1,\ldots,H_T \); with this in mind, it is clear that permuting \( H_1,\ldots,H_T \) and keeping intact \( \Lambda \) and \( \Gamma \) is a symmetry of (4.49) — such a transformation maps feasible set onto itself and preserves the value of the objective. Since (4.49) is convex and solvable, it follows that there exists an optimal solution to the problem with
\( H_1 = \ldots = H_T = H \). On the other hand,
\[
\Psi(Q_T)(H,\ldots,H) = \min_{\Theta} \left\{ \text{Tr}(Q_T \Theta) + \phi_R(\lambda[\Theta']) : \Theta \in \mathbf{S}^{mT}, \Theta' = \{\Theta'_\ell \geq 0, \ell \leq L\}, \right. \\
- \left. \begin{bmatrix}
\frac{1}{2}[M' H'_1,\ldots,M' H'_T] & \frac{1}{2}[H_1 M' ;\ldots; H_T M'] \\
\frac{1}{2}[H_1 M' ;\ldots; H_T M'] & \sum_{\ell} S'_\ell [\Theta'_\ell]
\end{bmatrix} \right\} \geq 0 \}
\]
\[
= \inf_{\Theta} \left\{ \text{Tr}(Q_T \Theta) + \phi_R(\lambda[\Theta']) : \Theta \in \mathbf{S}^{mT}, \Theta' = \{\Theta'_\ell \geq 0, \ell \leq L\}, \right. \\
- \left. \begin{bmatrix}
\frac{1}{2}[H_1 M' ;\ldots; H_T M'] & \frac{1}{2}[H_1 M' ;\ldots; H_T M'] \\
\frac{1}{2}[H_1 M' ;\ldots; H_T M'] & \sum_{\ell} S'_\ell [\Theta'_\ell]
\end{bmatrix} \right\} \geq 0 \}
\]
\[
= \inf_{\Theta} \left\{ \phi_R(\lambda[\Theta']) + \frac{T}{4} \text{Tr}(QHM [\sum_{\ell} S'_\ell [\Theta'_\ell]]^{-1} M'TH^T) : \Theta' = \{\Theta'_\ell \geq 0, \ell \leq L\} \right\}
\]
due to \( Q_T = \text{Diag}\{Q,\ldots,Q\} \), and we arrive at
\[
\Psi(Q_T)(H,\ldots,H) = \min_{G} \left\{ \frac{1}{2} \text{Tr}(QG) + \phi_R(\lambda[\Theta']) : \Theta' = \{\Theta'_\ell \geq 0, \ell \leq L\}, \right. \\
- \left. G \in \mathbf{S}^n, \begin{bmatrix}
\frac{1}{2}[H_1 M' ;\ldots; H_T M'] & \frac{1}{2}[H_1 M' ;\ldots; H_T M'] \\
\frac{1}{2}[H_1 M' ;\ldots; H_T M'] & \sum_{\ell} S'_\ell [\Theta'_\ell]
\end{bmatrix} \right\} \geq 0 \}
\]
(4.50)

(we have used Schur Complement Lemma combined with the fact that \( \sum_{\ell} S'_\ell [\Theta'_\ell] > 0 \) whenever \( \Theta'_\ell > 0 \) for all \( \ell \), see Lemma 4.44). In view of the above observations, when replacing variables \( H \) and \( G \) with
\( H = TH \) and \( G = T^2G \), respectively, problem (4.49), (4.50) becomes

\[
\text{Opt} = \min_{H, \Lambda, Y, \Upsilon, \Upsilon'} \left\{ \text{Tr}(\Lambda) + \phi_R(\Lambda|Y) + \phi_R(\Lambda'|Y') + \frac{1}{T}\text{Tr}(QG) : \right. \\
\left. \begin{array}{c}
\Lambda \in S_n^+, Y = \{Y_{\ell} \geq 0, \ell \leq L\}, Y' = \{Y'_{\ell} \geq 0, \ell \leq L\}, \\
\frac{1}{2}M^T(B - H^T\Lambda) \geq 0, \\
\frac{1}{2}M^T[B - A^T\Lambda] M \geq 0, \\
\frac{1}{2}M^T[H^T] M \geq 0, \\
\frac{1}{2}M^T[H'] M \geq 0, \\
\end{array} \right\}, \tag{4.51}
\]

and the estimate

\[
\hat{x}(w^T) = \frac{1}{T}H^T \sum_{k=1}^{T} w_k
\]

brought about by an optimal solution to (4.51) satisfies \( \text{Risk}_{R, H, ||\cdot||}\{\hat{x}|X\} \leq \text{Opt} \) where \( \Pi = \{Q_T\} \).

### 4.3.3.8 Estimation from repeated observations

Consider the special case of the situation from Section 4.3.3.1 where observation \( \omega \) in (4.31) is a \( T \)-element sample \( \omega = [\tilde{\omega}_1; ...; \tilde{\omega}_T] \) with components

\[
\tilde{\omega}_t = \tilde{A}x + \xi_t, t = 1, ..., T
\]

and \( \xi_t \) are i.i.d. observation noises with zero mean distribution \( \tilde{P} \) satisfying \( \tilde{P} \ll \tilde{\Pi} \) for some convex compact set \( \tilde{\Pi} \subset \text{int} S_n^+ \). In other words, we are in the situation where

\[
\begin{aligned}
A &\in [\tilde{A}; ...; \tilde{A}] \in \mathbb{R}^{m \times n} \text{ for some } \tilde{A} \in \mathbb{R}^{\tilde{m} \times n} \text{ and } m = T\tilde{m}, \\
\Pi &\in \{Q = \text{Diag}\{\tilde{Q}, ..., \tilde{Q}\}, \tilde{Q} \in \tilde{\Pi}\}.
\end{aligned}
\]

The same argument as used in item 5 of Section 4.3.3.7 above justifies the following

**Proposition 4.15.** In the situation in question and under Assumptions \( A, B, R \) the linear estimate of \( Bx \) yielded by an optimal solution to problem (4.42) can be found as follows. Consider the convex optimization problem

\[
\text{Opt} = \min_{\tilde{H}, \Lambda, Y, \Upsilon, \Upsilon'} \left\{ \phi_T(\Lambda|\Lambda) + \phi_R(\Lambda|Y) + \phi_R(\Lambda'|Y') + \frac{1}{T}\text{Tr}(\tilde{Q}\Theta) : \right. \\
\left. \begin{array}{c}
\Lambda \in \{\Lambda_k \geq 0, k \leq K\}, Y = \{Y_{\ell} \geq 0, \ell \leq L\}, Y' = \{Y'_{\ell} \geq 0, \ell \leq L\}, \\
\frac{1}{2}M^T[B - H^T\Lambda] \geq 0, \\
\frac{1}{2}M^T[B - A^T\Lambda] M \geq 0, \\
\frac{1}{2}M^T[H^T] M \geq 0, \\
\frac{1}{2}M^T[H'] M \geq 0, \\
\end{array} \right\}, \tag{4.52}
\]

where

\[
\Gamma(\Theta) = \max_{Q \in \Pi} \text{Tr}(\tilde{Q}\Theta).
\]

The problem is solvable, and the estimate in question is yielded by the \( \tilde{H} \)-component.
The provided by Proposition 4.14 upper bound on the risk \( \text{Risk}_{\Pi, \| \cdot \|} [\hat{x}(\cdot)|\mathcal{X}] \) of this estimate is equal to \( \text{Opt} \).

The advantage of this result as compared to what is stated under the circumstances by Proposition 4.14 is that the sizes of optimization problem (4.52) are independent of \( T \).

4.3.3.9 Near-optimality in Gaussian case

The risk of the linear estimate \( \hat{x}_{H_*}(\cdot) \) constructed in (4.42), (4.43) 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 \( \Pi \). Formally, the minimax risk is defined as

\[
\text{Risk}_{\Pi, \| \cdot \|} [\mathcal{X}] = \sup_{Q \in \Pi} \inf_{\hat{x}} \left[ \sup_{x \in \mathcal{X}} \mathbb{E}_{\xi \sim \mathcal{N}(0, Q)} \{ \| Bx - \hat{x}(Ax + \xi) \| \} \right] \tag{4.53}
\]

where the infimum is taken over all estimates.

**Proposition 4.16.** Under the premise and in the notation of Proposition 4.14, we have

\[
\text{Risk}_{\Pi, \| \cdot \|} [\mathcal{X}] \geq \frac{\text{Opt}}{64 \sqrt{(2 \ln F + 10 \ln 2)(2 \ln D + 10 \ln 2)}}. \tag{4.54}
\]

where

\[
D = \sum_k d_k, \quad F = \sum_\ell f_\ell. \tag{4.55}
\]

Thus, the upper bound \( \text{Opt} \) on the risk \( \text{Risk}_{\Pi, \| \cdot \|} [\hat{x}_{H_*}|\mathcal{X}] \) of the presumably good linear estimate \( \hat{x}_{H_*} \) yielded by an optimal solution to optimization problem (4.42) is within logarithmic in the sizes of spectratopes \( \mathcal{X} \) and \( B_* \), factor of the Gaussian minimax risk \( \text{Risk}_{\Pi, \| \cdot \|} [\mathcal{X}] \).

For the proof, see Section 4.8.5. The key component of the proof is the following important by its own right fact (for proof, see Section 4.8.4):

**Lemma 4.17.** Let \( Y \) be an \( N \times \nu \) matrix, let \( \| \cdot \| \) be a norm on \( \mathbb{R}^\nu \) such that the unit ball \( \mathcal{B}_* \) of the conjugate norm is the spectrapo \( \mathcal{N}(0, Q) \) for some positive semidefinite \( N \times N \) matrix \( Q \). Then the best upper bound on \( \psi_Q(Y) := \mathbb{E}\{\| Y^T \zeta \| \} \) yielded by Lemma 4.11, that is, the optimal value \( \text{Opt}[Q] \) in the convex optimization problem (cf. (4.40))

\[
\text{Opt}[Q] = \min_{\Theta, \Upsilon} \left\{ \phi_R(\lambda[\Upsilon]) + \text{Tr}(Q\Theta) : \Upsilon = \{ \Upsilon_\ell \geq 0, 0 \leq \ell \leq L \}, \Theta \in S^N,  \frac{\Theta}{2} M^T Y^T \left[ \sum_{\ell} S_{T}[\Upsilon_\ell] \right] \geq 0 \right\}. \tag{4.56}
\]
(for notation, see Lemma 4.11 and (4.36)) satisfies the identity
\[
\forall (Q \succeq 0) : \quad \text{Opt}[Q] = \text{Opt}[\Omega] := \min_{G, \Gamma = \{\Upsilon_\ell | \ell \leq L\}} \left\{ \phi_\mathcal{R}(\lambda[\Gamma]) + \text{Tr}(G) : G \succeq 0, \frac{1}{2}G^{1/2}\mathbf{Y}^TQ^{1/2} + \frac{1}{2}G^{1/2}\mathbf{Y}^TQ^{1/2} \succeq 0 \right\},
\]
and is a tight bound on \(\psi_Q(Y)\), namely,
\[
\psi_Q(Y) \leq \text{Opt}[Q] \leq 22\sqrt{2\ln F + 10\ln 2} \psi_Q(Y),
\]
where \(F = \sum \ell f_\ell\) is the size of the spectratope (4.34).

Besides this, for all \(\kappa \geq 1\) one has
\[
\text{Prob}_\mathbb{C} \left\{ \|Y^T\zeta\| \geq \frac{\text{Opt}[Q]}{4\kappa} \right\} \geq \beta_\kappa := 1 - \frac{e^{3/8}}{2} - 2Fe^{-\kappa^2/2}. \tag{4.59}
\]

In particular, when selecting \(\kappa = \sqrt{2\ln F + 10\ln 2}\), we obtain
\[
\text{Prob}_\mathbb{C} \left\{ \|Y^T\zeta\| \geq \frac{\text{Opt}[Q]}{4\sqrt{2\ln F + 10\ln 2}} \right\} \geq 0.2100 > \frac{3}{16}. \tag{4.60}
\]

### 4.4 Linear Estimates of Stochastic Signals

In the recovery problem considered so far in this chapter, 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” \(\eta\) is random with (partly) known probability distribution \(P_\eta\), 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

\[
\begin{align*}
y_{t+1} &= P_t y_t + \eta_t + u_t, \\
\omega_t &= C_t y_t + \xi_t, \quad 1 \leq t \leq T,
\end{align*}
\]

where \(y_t, \eta_t, u_t\) are, respectively, the state, the random “process noise,” and the deterministic “uncertain but bounded” disturbance affecting the system at time \(t\), \(\omega_t\) is the output – it is what we observe at time \(t\), and \(\xi_t\) is the observation noise. We assume that the matrices \(P_t, C_t\) are known in advance. Note that the trajectory

\[
y = [y_1; \ldots; y_T]
\]

of the states depends not only on the trajectories of process noises \(\eta_t\) and disturbances \(u_t\), but also on the initial state \(y_1\), which can be modeled as a realization of either the initial noise \(\eta_0\), or the initial disturbance \(u_0\). When \(u_t \equiv 0, y_1 = \eta_0\) and
the random vectors \( \{ \eta_t, 0 \leq t \leq T, \xi_t, 1 \leq t \leq T \} \) are independent of each other zero mean Gaussian, (4.61) is the model underlying the celebrated Kalman filter [140, 141, 167, 168].

Now, given model (4.61), 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 \( \eta_0 \) of \( \eta \), or the component \( u_0 \) of \( u \)), and our “full observation” becomes

\[
\omega = [\omega_1; \ldots; \omega_T] = A[\eta; u] + \xi, \quad \xi = [\xi_1; \ldots; \xi_T].
\]

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 the observation. For example, when speaking about (4.61), the goal could be to recover \( y_{T+1} \) (“forecast”).

We arrive at the following estimation problem:

Given noisy observation

\[
\omega = Ax + \xi \in \mathbb{R}^m
\]

of signal \( x = [\eta; u] \) with random component \( \eta \in \mathbb{R}^p \) and deterministic component \( u \) known to belong to a given set \( \mathcal{X} \subset \mathbb{R}^q \), we want to recover the image \( Bx \in \mathbb{R}^\nu \) of the signal. Here \( A \) and \( B \) are given matrices, \( \eta \) is independent of \( \xi \), and we have a priori (perhaps, incomplete) information on the probability distribution \( P_\eta \) of \( \eta \), 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 \( \xi \) 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 \( \hat{x}(\omega) \) by taking expectation of the norm \( \| \hat{x}(A[\eta; u] + \xi) - B[\eta; u]\| \) of the error over both \( \xi \) and \( \eta \) and then taking supremum of the result over the allowed distributions of \( \eta, \xi \) and over \( u \in \mathcal{X} \):

\[
\text{Risk}_\| \| \left[ \hat{x} \right] = \sup_{u \in \mathcal{X}} \sup_{P_\xi \in \mathcal{P}_\xi, P_\eta \in \mathcal{P}_\eta} \mathbb{E}_{[\eta; \xi]} \sim P_\xi \times P_\eta \{ \| \hat{x}(A[\eta; u] + \xi) - B[\eta; u]\| \}.
\]

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

\[
\text{Risk}_{\text{Euc}} \left[ \hat{x} \right] = \left[ \sup_{u \in \mathcal{X}} \sup_{P_\xi \in \mathcal{P}_\xi, P_\eta \in \mathcal{P}_\eta} \mathbb{E}_{[\eta; \xi]} \sim P_\xi \times P_\eta \{ \| \hat{x}(A[\eta; u] + \xi) - B[\eta; u]\|_2^2 \} \right]^{1/2}.
\]

Our next goal is to show that as far as the design of “presumably good” linear estimates \( \hat{x}(\omega) = H^T \omega \) is concerned, the techniques developed so far can be straightforwardly extended to the case of signals with random component.
4.4.1 Minimizing Euclidean risk

For the time being, assume that $P_\xi$ is comprised of all probability distributions $P$ on $\mathbb{R}^m$ with zero mean and covariance matrices $\text{Cov}[P] = E_\xi P \{ \xi \xi^T \}$ running through a computationally tractable convex compact subset $Q_\xi \subset \text{int} S^m_+$, and $P_\eta$ is comprised of all probability distributions $P$ on $\mathbb{R}^p$ with zero mean and covariance matrices running through a computationally tractable convex compact subset $Q_\eta \subset \text{int} S^p_+$. Let, in addition, $\mathcal{X}$ be a basic spectratope:

$$\mathcal{X} = \{ x \in \mathbb{R}^q : \exists t \in \mathcal{T} : R_k^2 |x| \leq t_k I_d, \ k \leq K \}$$

with our standard restrictions on $\mathcal{T}$ and $R_k[\cdot]$. Let us derive efficiently solvable convex optimization problem “responsible” for presumably good, in terms of its Euclidean risk, linear estimate.

For a linear estimate $H^T \omega$, $u \in \mathcal{X}$, $P_\xi \in P_\xi$, $P_\eta \in P_\eta$, denoting by $Q_\xi$ and $Q_\eta$, the covariance matrices of $P_\xi$ and $P_\eta$, and partitioning $A$ as $A = [A_n, A_u]$ and $B = [B_n, B_u]$ according to the partition $x = [\eta, u]$, we have

$$\begin{align*}
E_{\xi, \eta} \sim P_\xi \times P_\eta \{ &\|H^T (A \eta; u) + \xi - B \eta; u\|^2 \} \\
= &E_{\xi, \eta} \sim P_\xi \times P_\eta \{ \|H^T A_n - B_n \eta + H^T \xi + [H^T A_u - B_u u] u\|^2 \} \\
= &u^T [B_n - H^T A_u]^T [B_n - H^T A_u] u + E_{\xi \sim P_\xi} \{ \text{Tr}(H^T \xi \xi^T H) \} \\
&+ E_{\eta \sim P_\eta} \{ \text{Tr}(B_n - H^T A_n) \eta \eta^T (B_n - H^T A_n)^T \} \\
= &u^T [B_n - H^T A_u]^T [B_n - H^T A_u] u + \text{Tr}(H^T Q_\xi H) \\
&+ \text{Tr}(B_n - H^T A_n) Q_\eta (B_n - H^T A_n)^T).
\end{align*}$$

Hence, the squared Euclidean risk of the linear estimate $\hat{x}_H(\omega) = H^T \omega$ is

$$\begin{align*}
\text{Risk}^2_{\text{Eucl}}[\hat{x}_H] &= \Phi(H) + \Psi_\xi(H) + \Psi_\eta(H), \\
\Phi(H) &= \max_{u \in \mathcal{X}} u^T [B_n - H^T A_u]^T [B_n - H^T A_u] u, \\
\Psi_\xi(H) &= \max_{Q \in Q_\xi} \text{Tr}(H^T Q H), \\
\Psi_\eta(H) &= \max_{Q \in Q_\eta} \text{Tr}(B_n - H^T A_n) Q (B_n - H^T A_n)^T).
\end{align*}$$

Functions $\Psi_\xi$ and $\Psi_\eta$ are convex and efficiently computable, function $\Phi(H)$, by Proposition 4.8, admits efficiently computable convex upper bound

$$\begin{align*}
\overline{\Phi}(H) &= \min_{\lambda} \left\{ \phi_{\mathcal{T}}(\lambda[A]) : \Lambda = \{ A_k \geq 0, k \leq K \} \right\}, \\
[B_n - H^T A_u]^T [B_n - H^T A_u] &\preceq \sum_k R_k^* [A_k]
\end{align*}$$

which is tight within the factor $2 \max[\ln(2 \sum_k d_k), 1]$ (see Proposition 4.8). Thus, the efficiently solvable convex problem yielding presumably good linear estimate is

$$\text{Opt} = \min_{H} \left[ \overline{\Phi}(H) + \Psi_\xi(H) + \Psi_\eta(H) \right] ;$$

the Euclidean risk of the linear estimate $H^T \omega$ yielded by the optimal solution to the problem is upper-bounded by $\sqrt{\text{Opt}}$ and is within factor $\sqrt{2 \max[\ln(2 \sum_k d_k), 1]}$ of the minimal Euclidean risk achievable with linear estimates.
4.4.2 Minimizing $\| \cdot \|$-risk

Now let $P_\varepsilon$ be comprised of all probability distributions $P$ on $\mathbb{R}^m$ with matrices of second moments $\text{Var}[P] = E_{\eta \sim P}[\xi^T]$ running through a computationally tractable convex compact subset $Q_{\varepsilon} \subset \mathbb{R}^m$, and $P_\eta$ be comprised of all probability distributions $P$ on $\mathbb{R}^p$ with matrices of second moments $\text{Var}[P]$ running through a computationally tractable convex compact subset $Q_{\eta} \subset \mathbb{R}^p$. Let, same as above, $X$ be a basic spectrapo:

$$X = \{ x \in \mathbb{R}^n : \exists t \in \mathcal{T} : R_k^x [t] \preceq t_k I_{d_k}, k \leq K \},$$

and let $\| \cdot \|$ be such that the unit ball $B_*$ of the conjugate norm $\| \cdot \|_*$ is a spectrapo:

$$B_* = \{ y : \| y \|_* \leq 1 \} = \{ y \in \mathbb{R}^p : \exists (r \in \mathcal{R}, z \in \mathbb{R}^N) : y = M z, S_0^T [z] \preceq r I_{r}, \ell \leq L \},$$

with our standard restrictions on $\mathcal{T}, \mathcal{R}, R_k[\cdot]$ and $S_0[\cdot]$. Here the efficiently solvable convex optimization problem “responsible” for presumably good, in terms of its risk $\text{Risk}_{\| \cdot \|}$, linear estimate can be built as follows.

For a linear estimate $H^T \omega, u \in X, P_\varepsilon \in P_\varepsilon, P_\eta \in P_\eta$, denoting by $Q_{\varepsilon}$ and $Q_{\eta}$ the matrices of second moments of $P_\varepsilon$ and $P_\eta$, and partitioning $A$ as $A = [A_\eta, A_u]$ and $B = [B_\eta, B_u]$ according to the partition $x = [\eta; u]$, we have

$$E_{[\xi; \eta] \sim P_\varepsilon \times P_\eta} \{ \| H^T (A_\eta; u) + \xi \| - B(\eta; u) \}$$

$$= E_{[\xi; \eta] \sim P_\varepsilon \times P_\eta} \{ \| H^T A_\eta - B_\eta \eta + H^T \xi + [H^T A_u - B_u u] \| \}$$

$$\leq \| [B_u - H^T A_u u] \| + E_{\xi \sim P_\varepsilon} \{ \| H^T \xi \| \} + E_{\eta \sim P_\eta} \{ \| [B_\eta - H^T A_\eta \eta] \| \}.$$ 

It follows that for a linear estimate $\hat{x}_H(\omega) = H^T \omega$ one has

$$\text{Risk}_{\| \cdot \|}[\hat{x}_H] \leq \Phi(H) + \Psi_\varepsilon(H) + \Psi_\eta(H),$$

$$\Phi(H) = \max_{u \in X} \| [B_u - H^T A_u u] \|,$n

$$\Psi_\varepsilon(H) = \sup_{P_\varepsilon \in P_\varepsilon} E_{\xi \sim P_\varepsilon} \{ \| H^T \xi \| \},$$

$$\Psi_\eta(H) = \sup_{P_\eta \in P_\eta} E_{\eta \sim P_\eta} \{ \| [B_\eta - H^T A_\eta \eta] \| \}. \text{ As was shown in Section 4.3.3.3, the functions } \Phi, \Psi_\varepsilon, \Psi_\eta \text{ admit efficiently computable upper bounds as follows (for notation, see Section 4.3.3.3):}$$

$$\Phi(H) \leq \Phi(H) := \min_{A,Y} \left\{ \phi_T(\lambda[A]) + \phi_R(\lambda[Y]) : \right.$$  

$$A = \{ A_k \geq 0, k \leq K \}, Y = \{ Y_\ell \geq 0, \ell \leq L \}$$

$$\left[ \sum_k R_k^x [A_k] \right] \left[ \frac{1}{2} M^T [B_u - H^T A_u] M \right] \sum_{\ell} S_\ell [Y_\ell] \geq 0 \right\};$$

$$\Psi_\varepsilon(H) \leq \Psi_\varepsilon(H) := \min_{Y,G} \left\{ \phi_T(\lambda[Y]) + \max_{Q \in Q_\varepsilon} \text{Tr}(GQ) : Y = \{ Y_\ell \geq 0, \ell \leq L \}, \right.$$  

$$\left[ \frac{1}{2} G M^T H^T \right] \sum_{\ell} S_\ell [Y_\ell] \geq 0 \right\};$$

$$\Psi_\eta(H) \leq \Psi_\eta(H) := \min_{Y,G} \left\{ \phi_T(\lambda[Y]) + \max_{Q \in Q_\eta} \text{Tr}(GQ) : Y = \{ Y_\ell \geq 0, \ell \leq L \}, \right.$$  

$$\left[ \frac{1}{2} G M^T [B_\eta - H^T A_\eta] M \right] \sum_{\ell} S_\ell [Y_\ell] \geq 0 \right\};$$

and these bounds are reasonably tight (for details on tightness, see Proposition 4.8).
and Lemma 4.17). As a result, to get a presumably good linear estimate, one needs to solve the efficiently solvable convex optimization problem

$$\text{Opt} = \min_H \left[ \Phi(H) + \Psi_\xi(H) + \Psi_\eta(H) \right].$$

The linear estimate \(\hat{x}_{H^*} = H^T \omega\) yielded by an optimal solution \(H^*\) to this problem admits the risk bound

$$\text{Risk}_{\|\cdot\|}[\hat{x}_{H^*}] \leq \text{Opt}.$$ 

Note that the above derivation did not use independence of \(\xi\) and \(\eta\).

4.5 LINEAR ESTIMATION UNDER 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 “mixed” noise.

4.5.1 Uncertain-but-bounded noise

Consider estimation problem where one, given observation

$$\omega = Ax + \eta$$ (4.62)

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 \(\eta\) 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 \(\omega \mapsto \hat{x}(\omega)\) is

$$\text{Risk}_{\mathcal{H},\|\cdot\|}[\hat{x}|\mathcal{X}] = \sup_{x \in \mathcal{X}, \eta \in \mathcal{H}} \|Bx - \hat{x}(Ax + \eta)\|$$

(“\(\mathcal{H}\)-risk”).

We are about to prove that when \(\mathcal{X}, \mathcal{H}\) and the unit ball \(\mathcal{B}\) of the norm \(\|\cdot\|\), conjugate to \(\|\cdot\|\) are spectratopes, which we assume from now on, an efficiently computable linear estimate is near-optimal in terms of its \(\mathcal{H}\)-risk.

Our initial observation is that in this case the model (4.62) reduces straightforwardly to the model without observation noise. 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 \(\omega\) is \(y = [x; \eta] \in \mathcal{Y}\):

$$\omega = Ax + \eta = \tilde{A}y, \tilde{A} = [A, I_\nu],$$
while the entity to be recovered is

$$Bx = \hat{B}y, \quad \hat{B} = [B, 0_{\nu \times m}].$$

With these conventions, the $H$-risk of a candidate estimate $\hat{x}(\cdot) : \mathbb{R}^m \to \mathbb{R}^\nu$ becomes the quantity

$$\text{Risk}_\|\cdot\| \{\hat{x}|\mathcal{X} \times H\} = \sup_{y=[x;\eta] \in \mathcal{X} \times H} \|\hat{B}y - \hat{x}(\mathcal{A}y)\|,$$

and 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

(1) The problem of interest is to recover the linear image $Bx \in \mathbb{R}^\nu$ of an unknown signal $x$ known to belong to a given spectratope $\mathcal{X}$ (which, as always, we can assume w.l.o.g. to be basic) from (noiseless) observation $\omega = Ax \in \mathbb{R}^m$.

The risk of a candidate estimate is defined as

$$\text{Risk}_\|\cdot\| \{\hat{x}|\mathcal{X}\} = \sup_{x \in \mathcal{X}} \|Bx - \hat{x}(Ax)\|,$$

where $\|\cdot\|$ is a given norm with a spectratope $\mathcal{B}_*$, see (4.34), as the unit ball of the conjugate norm:

$$\mathcal{X} = \{x \in \mathbb{R}^n : \exists \ell \in \mathcal{T} : R_k^2[x] \preceq t_k I_d_k, k \leq K\},$$

$$\mathcal{B}_* = \{z \in \mathbb{R}^\nu : \exists y \in \mathcal{Y} : z = M y\},$$

$$\mathcal{Y} := \{y \in \mathbb{R}^q : \exists \ell \in \mathcal{R} : S_{\ell}[y] \preceq r \ell I_f, 1 \leq \ell \leq L\},$$

with our standard restrictions on $\mathcal{T}, \mathcal{R}$ and $R_k[\cdot], S_{\ell}[\cdot]$.

### 4.5.1.1 Building linear estimate

Let us build a presumably good linear estimate. For a linear estimate $\hat{x}_H(\omega) = H^T\omega$, we have

$$\text{Risk}_\|\cdot\| \{\hat{x}_H|\mathcal{X}\} = \max_{x \in \mathcal{X}} \|(B - H^T A)x\| = \max_{[u;x] \in \mathcal{B}_* \times \mathcal{X}} |u; x|^T \left[ \frac{1}{2} (B - H^T A)^T \frac{1}{2} (B - H^T A) \right] |u; x|.$$

Applying Proposition 4.8, we arrive at the following

**Proposition 4.18.** In the situation of this section, consider the convex optimization problem

$$\text{Opt}_\# = \min_{H,T=(\mathcal{T}_\ell)\Lambda=(\Lambda_k)} \left\{ \phi_\mathcal{T}(\lambda[\mathcal{T}]) + \phi_\mathcal{R}(\lambda[\Lambda]) : \mathcal{T}_\ell \succeq 0, \Lambda_k \succeq 0, \forall (\ell,k) \right\},$$

where

$$\sum_{k=1}^K R_k^2[\Lambda_k] \left[ \frac{1}{2} (B - H^T A)^T M \right] \left[ \sum_{\ell} S_{\ell}[\mathcal{T}_\ell] \right] \succeq 0.$$  (4.64)
where $R_k^*[], S^*_\ell[]$ are induced by $R_k[], S_\ell[],$ respectively, as explained in Section 4.3.1. The problem is solvable, and the risk of the linear estimate $\hat{x}_H(\cdot)$ yielded by the $H$-component of an optimal solution does not exceed $\text{Opt}_\#$.

For proof, see Section 4.8.6.1.

4.5.1.2 Near-optimality

**Proposition 4.19.** The linear estimate $\hat{x}_H$ yielded by Proposition 4.18 is near-optimal in terms of its risk:

$$\text{Risk}_{\| \cdot \|}[\hat{x}_H|X] \leq \text{Opt}_\# \leq O(1) \ln(D) \text{Risk}_{\text{opt}}[X], \quad D = \sum_k d_k + \sum_\ell f_\ell,$$

(4.65)

where $\text{Risk}_{\text{opt}}[X]$ is the minimax optimal risk:

$$\text{Risk}_{\text{opt}}[X] = \inf_{\hat{x}} \text{Risk}_{\| \cdot \|}[\hat{x},X]$$

with $\inf$ taken w.r.t. all Borel estimates.

**Remark 4.20.** When $X$ and $B_*$ are ellitopes rather than spectratopes:

$$X = \{ x \in \mathbb{R}^n : \exists t \in T : x^TR_kx \leq t_k, k \leq K \},$$

$B_* := \{ u \in \mathbb{R}^v : \| u \|_* \leq 1 \}$

$$= \{ u \in \mathbb{R}^v : \exists r \in \mathbb{R}, z : u = Mz, z^TS_\ell z \leq r_\ell, \ell \leq L \}$$

$[R_k \geq 0, \sum_k R_k \succ 0, S_\ell \geq 0, \sum_\ell S_\ell \succ 0]$

problem (4.64) becomes

$$\text{Opt}_\# = \min_{H,\lambda,\mu} \left\{ \phi_R(\mu) + \phi_T(\lambda) : \lambda \geq 0, \mu \geq 0, \right.$$

$$\left. \frac{1}{2} \lambda_k R_k \right\}$$

and (4.65) can be strengthened to

$$\text{Risk}_{\| \cdot \|}[\hat{x}_H,|X] \leq \text{Opt}_\# \leq O(1) \ln(K + L) \text{Risk}_{\text{opt}}[X].$$

For proofs, see Section 4.8.6.

4.5.1.3 Nonlinear estimation

Uncertain-but-bounded model of observation error makes it easy to point out an efficiently computable near-optimal nonlinear estimate. Indeed, in the situation described in the beginning of Section 4.5.1, let us assume that the range of observation error $\eta$ is

$$\mathcal{H} = \{ \eta \in \mathbb{R}^m : \| \eta \|_{(m)} \leq \sigma \},$$

where $\| \cdot \|_{(m)}, \sigma > 0$ are a given norm on $\mathbb{R}^m$ and a given error bound, and let us measure the recovery error by a given norm $\| \cdot \|_{(\nu)}$ on $\mathbb{R}^\nu$. We can immediately point out a (nonlinear) estimate optimal within factor 2, in terms of its $\mathcal{H}$-risk, namely, estimate $\hat{x}_*$ as follows:
Given \( \omega \), we solve the feasibility problem

\[
\text{find } x \in \mathcal{X} : \|Ax - \omega\|_{(m)} \leq \sigma. \quad (F[\omega])
\]

Let \( x_\omega \) be a feasible solution, we set \( \hat{x}_\omega(\omega) = Bx_\omega \).

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 \( \omega \). 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 4.21.** The estimate \( \hat{x}_\omega \) is optimal within factor 2:

\[
\text{Risk}_H[\hat{x}_\omega] \leq \text{Opt}_*_\omega := \sup_{x,y} \{ \|Bx - By\|_{(\nu)} : x, y \in \mathcal{X}, \|A(x - y)\|_{(m)} \leq 2\sigma \}
\]

\[
\leq 2 \text{Risk}_{\text{opt, } H} \quad (4.66)
\]

where \( \text{Risk}_{\text{opt, } H} \) is the infimum of \( H\)-risk over all estimates.

The proof of Proposition is the subject of Exercise 4.28.

### 4.5.1.4 Quantifying risk

Note that Proposition 4.21 does not impose restrictions on \( \mathcal{X} \) and the norms \( \| \cdot \|_{(m)}, \| \cdot \|_{(\nu)} \).

The only - but essential – shortcoming of the estimate \( \hat{x}_\omega \) is that we do not know, in general, what is its \( H\)-risk. From (4.66) it follows that this risk is tightly (namely, within factor 2) upper-bounded by \( \text{Opt}_*_\omega \), but this quantity, being the maximum of a convex function over some domain, can be difficult to compute. Aside from a handful of special cases where this difficulty does not arise, there is a generic situation when \( \text{Opt}_*_\omega \) can be tightly upper-bounded by efficient computation. This is the situation where \( \mathcal{X} \) is the spectratope defined in (4.63), \( \| \cdot \|_{(m)} \) is such that the unit ball of this norm is a basic spectratope:

\[
B_{(m)} := \{ u : \|u\|_{(m)} \leq 1 \} = \{ u \in \mathbb{R}^n : \exists p \in \mathcal{P} : Q_j^2[u] \leq p_j I_{e_j}, 1 \leq j \leq J \},
\]

and the unit ball of the norm \( \| \cdot \|_{(\nu)} \) _conjugate_ to the norm \( \| \cdot \|_{(\nu)} \) is a spectratope:

\[
B_{(\nu)}^* := \{ v \in \mathbb{R}^n : \|v\|_{(\nu)*} \leq 1 \}
= \{ v : \exists (w \in \mathbb{R}^n, r \in \mathcal{R}) : v = Mw, S_j^2[w] \leq r_\ell I_{f_\ell}, 1 \leq \ell \leq L \},
\]

with the usual restrictions on \( \mathcal{P}, \mathcal{R}, Q_j[:], \) and \( S_j[\cdot] \).

**Proposition 4.22.** In the situation in question, consider convex optimization problem

\[
\text{Opt} = \min_{\Lambda = \{ \Lambda_{k,j} \leq K \}, \Sigma = \{ \Sigma_{j} \leq L \}} \left\{ \phi_T(\Lambda_{[A]} + \phi_T(\Lambda_{[T]})) + \sigma^2 \phi_T(\Lambda_{[\Sigma]}) + \phi_T(\Lambda_{[\Sigma]}) : \right. \]

\[
\left. \left[ \sum_{k} S_j^2[T_{k}] M^T B \right] \geq 0 \right\} \quad (4.67)
\]
where $R^*_{k}[\cdot]$ is associated with mapping $x \mapsto R_k[x]$ according to (4.25), $S^*_\ell[\cdot]$ and $Q^*_j[\cdot]$ are associated in the same fashion with mappings $w \mapsto S_\ell[w]$ and $u \mapsto Q_j[u]$, respectively, and $\phi_T, \phi_R$ and $\phi_P$ are the support functions of the corresponding sets $T$, $R$ and $P$.

The optimal value in (4.67) is an efficiently computable upper bound on the quantity $\text{Opt}_\#$ defined in (4.66), and this bound is tight within factor

$$2 \max[\ln(2D), 1], \quad D = \sum_k d_k + \sum_\ell f_\ell + \sum_j e_j.$$  

Proof of Proposition is the subject of Exercise 4.29.

4.5.2 Mixed noise

So far, we have considered separately the cases of random and uncertain-but-bounded observation noises in (4.31). Note that both these observation schemes are covered by the following “mixed” scheme:

$$\omega = Ax + \xi + \eta,$$

where, as above, $A$ is a given $m \times n$ matrix, $x$ us unknown deterministic signal known to belong to a given signal set $\mathcal{X}$, $\xi$ is random noise with distribution known to belong to a family $\mathcal{P}$ of Borel probability distributions on $\mathbb{R}^m$ satisfying (4.32) for a given convex compact set $\Pi \subset \text{int} S^+_m$, and $\eta$ is “uncertain-but-bounded” observation error known to belong to a given set $\mathcal{H}$. As before, our goal is to estimate $Bx \in \mathbb{R}^\nu$ via observation $\omega$. In our present setting, given a norm $\|\cdot\|$ on $\mathbb{R}^\nu$, we can quantify the performance of a candidate estimate $\omega \mapsto \hat{x}(\omega) : \mathbb{R}^m \to \mathbb{R}^\nu$ by its risk

$$\text{Risk}_{\Pi, \mathcal{H}, \|\cdot\|}[\hat{x}|\mathcal{X}] = \sup_{x \in \mathcal{X}, \mathcal{P} \in \Pi, \eta \in \mathcal{H}} \mathbb{E}_{\xi \sim \mathcal{P}} \{\|Bx - \hat{x}(Ax + \xi + \eta)\|\}.$$  

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 4.5 to eliminate the observation noise. 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 $\bar{A}x^+ = Ax + \eta, \bar{B}x^+ = Bx$. With these conventions, the “mixed” observation scheme reduces to

$$\omega = \bar{A}x^+ + \xi,$$

and for every candidate estimate $\hat{x}(\cdot)$ it clearly holds

$$\text{Risk}_{\Pi, \mathcal{H}, \|\cdot\|}[\hat{x}|\mathcal{X}] = \text{Risk}_{\Pi, \|\cdot\|}[\hat{x}|\mathcal{X}^+] ,$$

so that we find ourselves in the situation of Section 4.3.3.1. Assuming that $\mathcal{X}$ and $\mathcal{H}$ are spectratopes, so is $\mathcal{X}^+$, meaning that all results of Section 4.3.3 on building presumably good linear estimates and their near-optimality are applicable to our present setup.
4.6 CALULUS 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 their “ellitopic versions.”

- Intersection \( \mathcal{X} = \bigcap_{i=1}^{I} \mathcal{X}_{i} \) of ellitopes

\[ \mathcal{X}_{i} = \{ x \in \mathbb{R}^{n_{i}} : \exists (y^{i} \in \mathbb{R}^{n_{i}}, t^{i} \in T_{i}) : x = P_{i}y^{i} \land [y^{i}]^{T}R_{ik}y^{i} \leq t^{i}_{k}, 1 \leq k \leq K_{i} \} \]

is an ellitope. Indeed, this is evident when \( \mathcal{X} = \{0\} \). Assuming \( \mathcal{X} \neq \{0\} \), we have

\[ \mathcal{Y} = \{ [y^{1}; ..., y^{I}] \in \mathbb{R}^{n_{1}+...+n_{I}} : P_{i}y^{i} = P_{i}y^{i}, 2 \leq i \leq I \} \]

(note that \( \mathcal{Y} \) can be identified with \( \mathbb{R}^{n} \) with a properly selected \( \bar{n} > 0 \)).

- Direct product \( \mathcal{X} = \prod_{i=1}^{I} \mathcal{X}_{i} \) of ellitopes

\[ \mathcal{X}_{i} = \{ x^{i} \in \mathbb{R}^{n_{i}} : \exists (y^{i} \in \mathbb{R}^{n_{i}}, t^{i} \in T_{i}) : x^{i} = P_{i}y^{i}, 1 \leq i \leq I \land [y^{i}]^{T}R_{ik}y^{i} \leq t^{i}_{k}, 1 \leq k \leq K_{i} \} \]

is an ellitope:

\[ \mathcal{X} = \left\{ [x^{1}; ..., x^{I}] \in \mathbb{R}^{n_{1}} \times ... \times \mathbb{R}^{n_{I}} : \exists \left( y = [y^{1}; ..., y^{I}] \in \mathbb{R}^{n_{1}+...+n_{I}} \right) \right. \\
\left. \quad t = (t^{1}, ..., t^{I}) \in T = T_{1} \times ... \times T_{I} \right\} \\
\quad \left. x = P_{i}y^{i}, 1 \leq i \leq I \land [y^{i}]^{T}R_{ik}y^{i} \leq t^{i}_{k}, 1 \leq k \leq K_{i} \right\} \]

- The linear image \( \mathcal{Z} = \{ Rx : x \in \mathcal{X}, R \in \mathbb{R}^{p \times n} \) of an ellitope

\[ \mathcal{X} = \{ x \in \mathbb{R}^{n} : \exists (y \in \mathbb{R}^{n}, t \in T) : x = Py \land y^{T}R_{k}y \leq t_{k}, 1 \leq k \leq K \} \]

is an ellitope:

\[ \mathcal{Z} = \{ z \in \mathbb{R}^{p} : \exists (y \in \mathbb{R}^{n}, t \in T) : z = [RP]y \land y^{T}R_{k}y \leq t_{k}, 1 \leq k \leq K \}. \]

- The inverse linear image \( \mathcal{Z} = \{ z \in \mathbb{R}^{q} : Rz \in \mathcal{X}, R \in \mathbb{R}^{n \times q} \) of an ellitope

\( \mathcal{X} = \{ x \in \mathbb{R}^{n} : \exists (y \in \mathbb{R}^{n}, t \in T) : x = Py \land y^{T}R_{k}y \leq t_{k}, 1 \leq k \leq K \} \) under linear mapping \( z \mapsto Rz : \mathbb{R}^{q} \rightarrow \mathbb{R}^{n} \) is an ellitope, provided that the mapping is an embedding: \( \ker R = \{0\} \). Indeed, setting \( E = \{ y \in \mathbb{R}^{n} : Py \in \text{Im}R \} \), we get a linear subspace in \( \mathbb{R}^{n} \). If \( E = \{0\} \), \( \mathcal{Z} = \{0\} \) is an ellitope; if \( E \neq \{0\} \), we have

\[ \mathcal{Z} = \{ z \in \mathbb{R}^{q} : \exists (y \in E, t \in T) : z = Py \land y^{T}R_{k}y \leq t_{k}, 1 \leq k \leq K \}, \]

\( \bar{P} = \Pi P \), where \( \Pi : \text{Im}R \rightarrow \mathbb{R}^{q} \) is the inverse of \( z \mapsto Rz : \mathbb{R}^{q} \rightarrow \text{Im}R \)

(\( E \) can be identified with some \( \mathbb{R}^{k} \), and \( \Pi \) is well defined since \( R \) is an embed-
The arithmetic sum $X = \{ x = \sum_{i=1}^{I} x^i : x^i \in X_i, 1 \leq i \leq I \}$, of ellitopes $X_i$ is an ellitope, with representation readily given by those of $X_1, \ldots, X_I$. Indeed, $X$ is the image of $X_1 \times \ldots \times X_I$ under the linear mapping $[x^1; \ldots; x^I] \mapsto x^1 + \ldots + x^I$, and taking direct products and images under linear mappings preserve ellitopes.

**“$S$-product.”** Let $X_i = \{ x^i \in \mathbb{R}^{n_i} : \exists (y^i, t^i) \in T_i : x^i = P_i y^i, 1 \leq i \leq I \}$ be ellitopes, and let $S$ be a convex compact set in $\mathbb{R}^I_+$ which intersects the interior of $\mathbb{R}^I_+$ and is monotone: $0 \leq s' \leq s \in S$ implies $s' \in S$. We associate with $S$ the set

$$S^{1/2} = \{ s \in \mathbb{R}^I_+ : [s^2; \ldots; s_I^2] \in S \}$$

of entrywise square roots of points from $S$; clearly, $S^{1/2}$ is a convex compact set. $X_i$ and $S$ specify the $S$-product of the sets $X_i$, $i \leq I$, defined as the set

$$Z = \{ z = [z^1; \ldots; z^I] : \exists (s \in S^{1/2}, x^i \in X_i, i \leq I) : z^i = s_i x^i, 1 \leq i \leq I \},$$

or, equivalently,

$$Z = \{ z = [z^1; \ldots; z^I] : \exists (r \in \mathbb{R}^{I_1}, y^1, \ldots, y^I) :$$

$$z_i = P_i y_i \forall i \leq I, |y|^T R_{ik} y^i \leq t_k \forall (i \leq I, k \leq K_i),$$

$$\mathcal{R} = \{ [r^1; \ldots; r^I] \geq 0 : \exists (s \in S^{1/2}, t^i \in T_i) : r^i = s_i^2 t^i \forall i \leq I \}. $$

We claim that $Z$ 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 $\mathbb{R}^{K_1+\ldots+K_I}$ containing a strictly positive vector (all this is evident), and that $\mathcal{R}$ is convex. To verify convexity, let $T_i = \text{cl}\{ [t^i; \tau_i] : \tau_i > 0, t^i/\tau_i \in T_i \}$ be the conic hulls of $T_i$’s. We clearly have

$$\mathcal{R} = \{ [r^1; \ldots; r^I] : \exists s \in S^{1/2} : [r^1; s_1^2] \in T_i, i \leq I \}$$

$$= \{ [r^1; \ldots; r^I] : \exists s \in S : [r^i; s_i] \in T_i, i \leq I \},$$

where the concluding equality is due to the origin of $S^{1/2}$. The concluding set in the above chain clearly is convex, and we are done.

As an example, consider the situation where the ellitopes $X_i$ posses nonempty interiors and thus can be thought of as unit balls of norms $\|\cdot\|_i$ on the respective spaces $\mathbb{R}^{n_i}$, and let $S = \{ s \in \mathbb{R}^I_+ : \|s\|_{p/2} \leq 1 \}$, where $p \geq 2$. In this situation, $S^{1/2} = \{ s \in \mathbb{R}^I_+ : \|s\|_p \leq 1 \}$, whence $Z$ is the unit ball of the “block $p$-norm”

$$\| [z^1; \ldots; z^I] \| = \| [\|z^1\|_1; \ldots; \|z^I\|_I] \|_p.$$ 

Note also that the usual direct product of $I$ ellitopes is their $S$-product, with $S = [0, 1]^I$.

**“$S$-weighted sum.”** Let $X_i \subset \mathbb{R}^{n_i}$ be ellitopes, $1 \leq i \leq I$, and let $S \subset \mathbb{R}^I_+$, $S^{1/2}$ be the same as in the previous rule. Then the $S$-weighted sum of the sets $X_i$,
defined as
\[ \mathcal{X} = \{ x : \exists (s \in S^{1/2}, x^i \in \mathcal{X}_i, i \leq I) : x = \sum_i s_i x^i \} \]
is an ellitope. Indeed, the set in question is the image of the $S$-product of $\mathcal{X}_i$ under the linear mapping $[z^1; \ldots; z^I] \mapsto z^1 + \ldots + z^I$, and taking $S$-products and linear images preserves the property to be an ellitope.

It should be stressed that the outlined “calculus rules” are fully algorithmic: representation (4.6) of the result of an operation is readily given by the representations (4.6) of the operands.

### 4.7 EXERCISES FOR CHAPTER 4

#### 4.7.1 Linear estimates vs. Maximum Likelihood

**Exercise 4.1.**

Consider the problem posed in the beginning of Chapter 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 $\mathcal{X} \subset \mathbb{R}^n$, we want to recover $Bx$.

Let us consider the case where matrix $A$ is square and invertible, $B$ 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
\[ \hat{x}_{\text{ML}}(\omega) = \arg\min_{y \in \mathcal{X}} \| \omega - Ay \|_2 \quad (\text{ML}) \]
– 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 estimation problem is: how good is the ML estimate?

The goal of this exercise is to show 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 Chapter 4 do result in efficiently computable near-optimal linear estimate.

To justify the claim, investigate the risk (4.2) of the ML estimate in the case where
\[ \mathcal{X} = \left\{ x \in \mathbb{R}^n : x_1^2 + \epsilon^2 \sum_{i=2}^n x_i^2 \leq 1 \right\} \& A = \text{Diag}\{1, \epsilon^{-1}, \ldots, \epsilon^{-1}\} \]
$\epsilon$ and $\sigma$ are small, and $n$ is large, so that $\sigma^2(n-1) \geq 2$. 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 $\{256, 1024, 2048\}$, $\epsilon = \sigma$ run through $\{0.01; 0.05; 0.1\}$, and signal $x$ is generated as

$$x = [\cos(\phi); \sin(\phi)\epsilon\zeta],$$

where $\phi \sim \text{Uniform}[0, 2\pi]$ and random vector $\zeta$ is independent of $\phi$ and is distributed uniformly on the unit sphere in $\mathbb{R}^{n-1}$.

4.7.2 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 the story is as follows:

We consider the Measurement Design version of signal recovery in Gaussian o.s., specifically, we are allowed to use observations

$$\omega = A_q x + \sigma \xi \quad [\xi \sim \mathcal{N}(0, I_m)]$$

where

$$A_q = \text{Diag}\{\sqrt{q_1}, \sqrt{q_2}, ..., \sqrt{q_m}\} A,$$

with a given $A \in \mathbb{R}^{m \times n}$ and vector $q$ which we can select in a given convex compact set $Q \subset \mathbb{R}^m_+$. The signal $x$ underlying the observation is known to belong to a given ellitope $X$. Your goal is to select $q \in Q$ and a linear recovery $\omega \mapsto G^T \omega$ of the image $Bx$ of $x \in X$, with given $B$, resulting in the smallest worst-case, over $x \in X$, expected $\|\cdot\|_2^2$ recovery risk. Modify, according to this goal, problem (4.12). Is it possible to end up with a tractable problem? Work out in full details the case when $Q = \{q \in \mathbb{R}^m_+ : \sum_i q_i = m\}$.

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 $\rho_i$ of $i$-th segment known to belong to the interval $[\mu - \delta_i, \mu + \delta_i]$.

The bar is lightened from the left end; when light passes through a segment with density $\rho$, light’s intensity is reduced by factor $e^{-\alpha \rho}$. The light intensity 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 $\ell_i$ at the right endpoint of the $i$-th segment during time $q_i$; the result $z_i$ of the measurement is $\ell_i e^{\sigma \xi_i/\sqrt{q_i}}$, 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 (in the latter case, we assume that 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 $X$ be a basic ellitope in $\mathbb{R}^n$:

$$X = \{ x \in \mathbb{R}^n : \exists t \in T : x^T S_k x \leq t_k, 1 \leq k \leq K \}$$

with our usual restrictions on $S_k$ and $T$. Let, further, $m$ be a given positive integer, and $x \mapsto Bx : \mathbb{R}^n \to \mathbb{R}^\nu$ be a given linear mapping. Consider the Measurement Design problem where you are looking for a linear recovery $\omega \mapsto \hat{x}_H(\omega) := H^T \omega$ of $Bx$, $x \in X$, from observation

$$\omega = Ax + \sigma \xi$$

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 (4.12), this problem can be posed as

$$\text{Opt} = \min_{H, \lambda, A} \left\{ \sigma^2 \text{Tr}(H^T H) + \phi_T(\lambda) : \sum_k \lambda_k S_k \begin{bmatrix} B^T - A^T H \\ H \end{bmatrix} I_\nu \succeq 0, \| A \| \leq 1, \lambda \geq 0 \right\},$$

(4.68)

where $\| \cdot \|$ stands for the spectral norm. The objective in this problem is the (upper bound on the) squared risk $\text{Risk}_2^2[\hat{x}_H | X]$, 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, the value of the variable fixed in a round being 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 (4.68). You may restrict your experimentation to the case where the sizes $m, n, \nu$ are quite moderate, in the range of tens, and $X$ is either the box $\{ x : j^{2\gamma} x_j^2 \leq 1, 1 \leq j \leq n \}$, or the ellipsoid $\{ x : \sum_{j=1}^n j^{\gamma} x_j^2 \leq 1 \}$, where $\gamma$ is a nonnegative parameter (try $\gamma = 0, 1, 2, 3$). As for $B$, you can generate it at random, or enforce $B$ to have prescribed singular values, say, $\sigma_j = j^{-\theta}$, $1 \leq j \leq \nu$, and randomly selected system of singular vectors.

2. Identify cases where a globally optimal solution to (4.68) is easy to find and use this information 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 is,” try to improve it.
3. Modify (4.68) and your experimentation to cover the cases where the constraint \( \|A\| \leq 1 \) on the sensing matrix is replaced with one of the following:
   - \( \| \text{Row}_i[A] \|_2 \leq 1, 1 \leq i \leq m, \)
   - \( |A_{ij}| \leq 1 \) 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 \( U \in \mathbb{R}^{\nu \times \nu} \) and \( V \in \mathbb{R}^{n \times n} \) and diagonal \( \nu \times n \) matrix \( D \) with nonnegative and nonincreasing diagonal entries.\(^9\)

   These entries are uniquely defined by \( G \) and are called singular values \( \sigma_i(G) \), \( 1 \leq i \leq \min[\nu, n] \). Singular values admit characterization similar to variational characterization of eigenvalues of a symmetric matrix, see, e.g., [15, Section A.7.3]:

   \[ \sigma_i(G) = \min_{E \in \mathcal{E}_i} \max_{\|e\|_2 = 1} \|Ge\|_2, \ 1 \leq i \leq \min[\nu, n], \tag{4.69} \]

   where \( \mathcal{E}_i \) is the family of all subspaces in \( \mathbb{R}^n \) of codimension \( i - 1 \).

   \[ \sigma_i(G) \geq \sigma_{i+k}(G'), \ 1 \leq i \leq \min[\nu, n], \]

   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 \( \|G\|_{\text{Sh}, p} = \|\sigma(G)\|_p \) 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 \( \|\cdot\|_{\text{Sh}, q} \), with \( \frac{1}{p} + \frac{1}{q} = 1 \). An easy and important consequence of Corollary 4.24 is the following fact:

   \[ \text{Corollary 4.25. Given a } \nu \times n \text{ matrix } G, \text{ an integer } k, 0 \leq k \leq \min[\nu, n], \text{ and } p \in [1, \infty], \ (\text{one of}) \ the \ best \ approximations \ of \ G \ in \ the \ Shatten \ p \text{-norm among } \nu \times n \text{ matrices of rank } \leq k \ text{ is obtained from } G \text{ by zeroing our all but } k \text{ largest singular values, that is, the matrix } G^k = \sum_{i=1}^{k} \sigma_i(G)\text{Col}_i[U]\text{Col}_i^T[V], \text{ where } G = UDV^T \text{ is the singular value decomposition of } G. \]

   Prove Theorem 4.23 and Corollaries 4.24 and 4.25.

5. Consider the Measurement Design problem (4.68) in the case when \( X \) is an ellipsoid:

   \[ X = \left\{ x \in \mathbb{R}^n : \sum_{j=1}^{n} x_i^2/a_j^2 \leq 1 \right\}, \]

   \[ \text{We say that a rectangular matrix } D \text{ is diagonal if all entries } D_{ij} \text{ in } D \text{ with } i \neq j \text{ are zeros.} \]

\(^9\)We say that a rectangular matrix \( D \) is diagonal if all entries \( D_{ij} \) in \( D \) with \( i \neq j \) are zeros.
$A$ is an $m \times n$ matrix of spectral norm not exceeding 1, and there is no noise in observations: $\sigma = 0$. 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 $\sigma$ is small.

4.7.3 Around semidefinite relaxation

Exercise 4.5.

Let $\mathcal{X}$ be an ellitope:

$$\mathcal{X} = \{ x \in \mathbb{R}^n : \exists (y \in \mathbb{R}^N, t \in \mathcal{T}) : x = Py, y^T S_k y \leq t_k, 1 \leq k \leq K \}$$

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:

$$\mathcal{X} = \{ x \in \mathbb{R}^n : \exists (y \in \mathbb{R}^N, t^+ \in \mathcal{T}^+) : x = Py, [s_{kj}^T x]^2 \leq t^+_k, 1 \leq k \leq K, 1 \leq j \leq r_k \}$$

with $T^+ = \{ t^+_k = \{ t^+_k \geq 0 \} : \exists t \in \mathcal{T} : \sum_{j=1}^{r_k} t^+_k \leq t_k, 1 \leq k \leq K \}$

If now $C$ is a symmetric $n \times n$ matrix and $\text{Opt} = \max_{x \in \mathcal{X}} x^T C x$, we have

$$\text{Opt}_s \leq \text{Opt}_e := \min_{\lambda = \{ \lambda_k \in \mathbb{R}_+ \}} \{ \phi_T (\lambda) : P^T C P \preceq \sum_k \lambda_k S_k \}$$

$$\text{Opt}_s \leq \text{Opt}_+ := \min_{\Lambda = \{ \Lambda_{kj} \in \mathbb{R}_+ \}} \{ \phi_T (\Lambda) : P^T C P \preceq \sum_{k,j} \Lambda_{kj} s_{kj} s_{kj}^T \}$$

where the first relation is yielded by ellitopic representation of $\mathcal{X}$ and Proposition 4.6, and the second, on closer inspection (carry this inspection out!) – by the spectratopic representation of $\mathcal{X}$ and Proposition 4.8.

Prove that $\text{Opt}_e = \text{Opt}_s$.

Exercise 4.6.

Proposition 4.6 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

$$\text{Opt} = \max_{x \in X} [f(x) := x^T Ax + 2b^T x + c],$$

$$X = \{ x : \exists (y, t \in \mathcal{T}) : x = Py + p, y^T S_k y \leq t_k, 1 \leq k \leq K \}$$

with our standard assumptions on $S_k$ and $\mathcal{T}$.

Note: $X$ is centered at $p$, and a natural upper bound on $\text{Opt}$ is

$$\text{Opt} \leq f(p) + \text{Opt},$$

where $\text{Opt}$ is an upper bound on the quantity

$$\text{Opt} = \max_{x \in X} [f(x) - f(p)].$$

What you are interested to upper-bound, is the ratio $\text{Opt}/\text{Opt}$. 

Exercise 4.7.
[estimating Kolmogorov widths of spectratope/ellitope]

4.7.A. Preliminaries: Kolmogorov and Gelfand widths. Let $\mathcal{X}$ be a convex compact set in $\mathbb{R}^n$, and let $\|\cdot\|$ be a norm on $\mathbb{R}^n$. Given a linear subspace $E$ in $\mathbb{R}^n$, let

$$\text{dist}_{\|\cdot\|}(x, E) = \min_{z \in E} \|x - z\| : \mathbb{R}^n \to \mathbb{R}_+$$

be the $\|\cdot\|$-distance from $x$ to $E$. The quantity

$$\text{dist}_{\|\cdot\|}(\mathcal{X}, E) = \max_{x \in \mathcal{X}} \text{dist}_{\|\cdot\|}(x, E)$$

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 $E_m$ the family of all linear subspaces in $\mathbb{R}^m$ of dimension $m$, the quantity

$$\delta_m(\mathcal{X}, \|\cdot\|) = \min_{E \in E_m} \text{dist}_{\|\cdot\|}(\mathcal{X}, E)$$

can be viewed as the best achievable quality of approximation, measured in $\|\cdot\|$, of vectors from $\mathcal{X}$ by vectors from an $m$-dimensional linear subspace of $\mathbb{R}^n$. This quantity is called $m$-th Kolmogorov width of $\mathcal{X}$ w.r.t. $\|\cdot\|$.

Observe that one has

$$\text{dist}_{\|\cdot\|}(x, E) = \max_{\xi} \{\xi^T x : \|\xi\| \leq 1, \xi \in E^\perp\}$$

where $E^\perp$ is the orthogonal complement to $E$.

1) Prove (4.70).

Hint: Represent $\text{dist}_{\|\cdot\|}(x, E)$ as the optimal value in a conic problem on the cone $K = \{\{x; t\} : t \geq \|x\|\}$ and use Conic Duality Theorem.

Now consider the case when $\mathcal{X}$ is the unit ball of some norm $\|\cdot\|$. In this case (4.70) combines with the definition of Kolmogorov width to imply that

$$\delta_m(\mathcal{X}, \|\cdot\|) = \min_{E \in E_m} \text{dist}_{\|\cdot\|}(\mathcal{X}, E) = \min_{E \in E_m} \max_{x \in \mathcal{X}} \max_{y \in E^+, \|y\| \leq 1} y^T x = \min_{E \in E_m} \max_{y \in E^+, \|y\| \leq 1} \|y\|_{\mathcal{X}, *},$$

where $\|\cdot\|_{\mathcal{X}, *}$ is the norm conjugate to $\|\cdot\|_{\mathcal{X}}$. Note that when $\mathcal{Y}$ is a convex compact set in $\mathbb{R}^n$ and $\|\cdot\|$ is a norm on $\mathbb{R}^n$, the quantity

$$d_m(\mathcal{Y}, \|\cdot\|) = \min_{F \in E_{n-m}} \max_{y \in \mathcal{Y} \cap F} |y|$$

has a name – it is called the $m$-th Gelfand width of $\mathcal{Y}$ taken w.r.t. $\|\cdot\|$. “Duality relation” (4.71) states that
CHAPTER 4

When $X$, $Y$ are the unit balls of respective norms $\| \cdot \|_X$, $\| \cdot \|_Y$, for every $m < n$ the $m$-th Kolmogorov width of $X$ taken w.r.t. $\| \|_Y$, is the same as $m$-th Gelfand width of $Y$ taken w.r.t. $\| \|_X$.

The goal of the remaining part of the 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 $X \subset \mathbb{R}^n$. In the sequel we assume that

- $X$ is a spectratope:
  $$X = \{ x \in \mathbb{R}^n : \exists (t \in T, u) : x = Pu, R_k^2[u] \preceq t_k I_d, k \leq K \};$$
- The unit ball $B_*$ of the norm conjugate to $\| \cdot \|$ is a spectratope:
  $$B_* = \{ y : \| y \|_* \leq 1 \} = \{ y \in \mathbb{R}^n : \exists (r \in \mathcal{R}, z) : y = Mz, S_\ell^2[z] \preceq r_\ell I_f, \ell \leq L \};$$

with our usual restrictions on $T, \mathcal{R}$ and $R_k[\cdot]$ and $S_\ell[\cdot]$.

4.7.B. Simple case: $\| \cdot \| = \| \cdot \|_2$. We start with the simple case where $\| \cdot \| = \| \cdot \|_2$, so that $B_*$ is the ellitope $\{ y : y^T y \leq 1 \}$.

Let $D = \sum d_k$ be the size of the spectratope $X$, and let
$$\kappa = 2 \max[\ln(2D), 1].$$

Given integer $m < n$, consider convex optimization problem

$$\text{Opt}(m) = \min_{\Lambda = \{ \Lambda_k, k \leq K \}, Y} \left\{ \phi_T(\lambda[\Lambda]) : \Lambda_k \succeq 0 \forall k, 0 \preceq Y \preceq I_n, \sum S^*_k[\Lambda_k] \succeq P^TYP, \text{Tr}(Y) = n - m \right\}. \quad (P_m)$$

2) Prove the following

**Proposition 4.26.** Whenever $1 \leq \mu \leq m < n$, one has

$$\text{Opt}(m) \leq 2\delta_m^2(X, \| \cdot \|_2) \& \delta_m^2(X, \| \cdot \|_2) \leq \frac{m + 1}{m + 1 - \mu} \text{Opt}(\mu). \quad (4.72)$$

Moreover, the above upper bounds on $\delta_m(X, \| \cdot \|_2)$ are “constructive”, meaning that an optimal solution to $(P_\mu)$, $\mu \leq m$, can be straightforwardly converted into a linear subspace $E^{m,\mu}$ of dimension $m$ such that

$$\text{dist}_{\| \cdot \|_2}(X, E^{m,\mu}) \leq \sqrt{\frac{m + 1}{m + 1 - \mu} \text{Opt}(\mu)}.$$

Finally, $\text{Opt}(\mu)$ is nonincreasing in $\mu < n$.

4.7.C. General case. Now consider the case when both $X$ and the unit ball $B_*$ of the norm conjugate to $\| \cdot \|$ are spectratopes. As we 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:

(1) Giveen \( m < n \), let \( Y \) be an orthoprojector of \( \mathbb{R}^n \) of rank \( n - m \), and let collections \( \Lambda = \{ \Lambda_k \geq 0, k \leq K \} \) and \( \Upsilon = \{ \Upsilon_\ell \geq 0, \ell \leq L \} \) satisfy the relation

\[
- \frac{1}{2} \sum_k \| \Lambda_k \|_1 \left\| \frac{1}{2} P^T Y M \right\|_2 \sum_\ell S_\ell \| \Upsilon_\ell \|_1 \geq 0.
\]

(4.73)

Then

\[
\text{dist}_{\| \cdot \|} (X, \text{Ker} Y) \leq \phi_T (\lambda[\Lambda]) + \phi_R (\lambda[\Upsilon]).
\]

(4.74)

As a result,

\[
\delta_m (X, \| \cdot \|) \leq \text{dist}_{\| \cdot \|} (X, \text{Ker} Y) \leq \hat{\text{Opt}} := \min_{\Lambda_k \geq 0 \forall k, \Upsilon_\ell \geq 0 \forall \ell,}
\]

\[
\left[ \sum_k \| \Lambda_k \|_1 \left\| \frac{1}{2} P^T Y M \right\|_2 \sum_\ell S_\ell \| \Upsilon_\ell \|_1 \right] \geq 0.
\]

(4.75)

4) Prove the following statement:

(2) Let \( m, n, Y \) be as in (1). Then

\[
\delta_m (X, \| \cdot \|) \leq \text{dist}_{\| \cdot \|} (X, \text{Ker} Y)
\]

\[
\leq \text{Opt} := \min_{\Lambda_k \geq 0 \forall k, \Upsilon_\ell \geq 0 \forall \ell,}
\]

\[
\left[ \sum_k \| \Lambda_k \|_1 \left\| \frac{1}{2} P^T Y M \right\|_2 \sum_\ell S_\ell \| \Upsilon_\ell \|_1 \right] \geq 0.
\]

(4.76)

and \( \hat{\text{Opt}} \leq \text{Opt} \), with \( \text{Opt} \) given by (4.75).

Statements (1), (2) suggest the following policy for upper-bounding the Kolmogorov width \( \delta_m (X, \| \cdot \|) \):

A. First, we select an integer \( \mu \), \( 1 \leq \mu < n \), and solve the convex optimization problem

\[
\min_{\Lambda_k, \Upsilon_\ell, Y} \left\{ \phi_T (\lambda[\Lambda]) + \phi_R (\lambda[\Upsilon]) : 0 \preceq Y \preceq I, \text{Tr}(Y) = n - \mu, \lambda[\Lambda] \geq 0 \forall k, \Upsilon_\ell \geq 0 \forall \ell, \right. \]

\[
\left. \left[ \sum_k \| \Lambda_k \|_1 \left\| \frac{1}{2} P^T Y M \right\|_2 \sum_\ell S_\ell \| \Upsilon_\ell \|_1 \right] \geq 0 \right\}.
\]

(P\(\mu\))

B. Next, we take the \( Y \)-component \( Y^\mu \) of the optimal solution to (P\(\mu\)) 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 \( Y^\mu \) intact and replace \( m \) smallest eigenvalues with zeros, and all remaining eigenvalues with ones.
C. Finally, we solve the convex optimization problem

\[
\text{Opt}_{m,\mu} = \min_{\lambda,Y,\nu} \left\{ \phi_T(\lambda[A]) + \phi_R(\lambda[Y]) : \nu \geq 0, A = \{A_k \geq 0, k \leq K\}, Y = \{Y_\ell \geq 0, \ell \leq L\}, \right. \\
\left. \frac{1}{2} \sum_k R_\nu^2[A_k] + \frac{1}{2} P^T M \right\} \geq 0 \quad (P^{m,\mu})
\]

By (!!), $\text{Opt}_{m,\mu}$ is an upper bound on the Kolmogorov width $\delta_m(\mathcal{X}, \| \cdot \|)$ (and in fact also on $\text{dist}_{\| \cdot \|}((\mathcal{X}, \text{Ker} Y))$.

Observe all the complications we encounter when 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. Note that Proposition 4.26 gives both a lower bound $\sqrt{\text{Opt}(m)/\kappa}$ on the $m$-th Kolmogorov width of $\mathcal{X}$ w.r.t. $\| \cdot \|_2$, and a family of upper bounds $\sqrt{\frac{m+1}{m+1-\mu}}\text{Opt}(\mu)$, $1 \leq \mu \leq m$, 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 $\epsilon$ and $k$ it holds $\delta_k(\mathcal{X}, \| \cdot \|_2) \leq \epsilon$, then $\text{Opt}(k) \leq \kappa \epsilon^2$ by Proposition 4.26 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

\[
\text{dist}_{\| \cdot \|_2}(\mathcal{X}, E^{m,k}) \leq \sqrt{\frac{2k+1}{k+1}} \text{Opt}(k) \leq \sqrt{2} \text{Opt}(k) \leq \sqrt{2} \kappa \epsilon.
\]

Thus, if $\mathcal{X}$ may somehow be approximated by a $k$-dimensional subspace within $\| \cdot \|_2$-accuracy $\epsilon$, we can efficiently get approximation of “nearly the same quality” ($\sqrt{2} \kappa \epsilon$ instead of $\epsilon$; recall that $\kappa$ is just logarithmic in $D$) and “nearly the same dimension” ($2k$ instead of $k$).

Neither of these options is preserved when passing from the Euclidean norm to a general one: in the latter case, we do not have lower bounds on Kolmogorov widths, and have no understanding of how tight our upper bounds are.

Now – two concluding questions:

5) Why in step A of the above bounding scheme we utilize statement (!) rather than less conservative (since $\widehat{\text{Opt}} \leq \text{Opt}$) statement (!!!)?

6) Implement the scheme numerically and run experiments.

Recommended setup:

- Given $\sigma > 0$ and positive integers $n$ and $\kappa$, let $f$ be a function of continuous argument $t \in [0, 1]$ satisfying the smoothness restriction $|f^{(\kappa)}(t)| \leq \sigma^k$, $0 \leq t \leq 1$, $k = 0, 1, 2, ..., \kappa$. Specify $\mathcal{X}$ as the set of $n$-dimensional vectors $x$ obtained by restricting $f$ onto $n$-point equidistant grid $\{t_i = i/n\}_{i=1}^n$. To this end, translate the description on $f$ into a bunch of two-sided linear constraints on $x$:

\[
|d_{(i)}^T[x_i; x_{i+1}; \ldots; x_{i+k}]| \leq \sigma^k, \quad 1 \leq i \leq n-k, \quad 0 \leq k \leq \kappa,
\]

where $d_{(k)} \in \mathbb{R}^{k+1}$ is the vector of coefficients of finite-difference approximation, with resolution $1/n$, of the $k$-th derivative:

\[
\begin{align*}
d_{(0)} &= 1, & d_{(1)} &= n[-1; 1], & d_{(2)} &= n^2[1; -2; 1], \\
d_{(3)} &= n^3[-1; 3; -3; 1], & d_{(4)} &= n^4[1; -4; 6; -4; 1], & \ldots
\end{align*}
\]
• 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 4.8.**

[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

\[
\text{Opt}_X(B) = \max_{x \in X} x^T B x, \quad [B \in \mathbb{S}^n]
\]

where \( X \subset \mathbb{R}^n \) 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 \( \mathcal{U} \subset \mathbb{S}^n \) such that for every \( x \in X \) it holds \( xx^T \in \mathcal{U} \); in this case, we refer to \( \mathcal{U} \) as to a set matching \( X \) (equivalent wording: "\( \mathcal{U} \) matches \( X \)). Given such a set \( \mathcal{U} \), the optimal value in the convex optimization problem

\[
\overline{\text{Opt}}_{\mathcal{U}}(B) = \max_{U \in \mathcal{U}} \text{Tr}(BU)
\]

is an efficiently computable convex upper bound on \( \text{Opt}_X(B) \).

Given \( \mathcal{U} \) matching \( X \), we can pass from \( \mathcal{U} \) to the conic hull of \( \mathcal{U} \), to the set

\[
U[\mathcal{U}] = \text{cl}\{(U, \mu) \in \mathbb{S}^n \times \mathbb{R}_+ : \mu > 0, U/\mu \in \mathcal{U}\}
\]

which, as it is immediately seen, is a closed convex cone contained in \( \mathbb{S}^n \times \mathbb{R}_+ \). The only point \((U, \mu)\) in this cone with \( \mu = 0 \) has \( U = 0 \) (since \( \mathcal{U} \) is compact), and

\[
\mathcal{U} = \{U : (U, 1) \in \mathcal{U}\} = \{U : \exists \mu \leq 1 : (U, \mu) \in \mathcal{U}\},
\]

so that the definition of \( \overline{\text{Opt}}_{\mathcal{U}}(B) \) can be rewritten equivalently as

\[
\overline{\text{Opt}}_{\mathcal{U}}(B) = \min_{U, \mu} \{\text{Tr}(BU) : (U, \mu) \in U, \mu \leq 1\}.
\]

The question, of course, is where to take a set \( \mathcal{U} \) matching \( X \), and the answer depends on what we know about \( X \). For example, when \( X \) is a basic ellitope:

\[
X = \{x \in \mathbb{R}^n : \exists t \in \mathcal{T} : x^T S_k x \leq t_k, k \leq K\}
\]

with our usual restrictions on \( \mathcal{T} \) and \( S_k \), it is immediately seen that

\[
x \in X \Rightarrow xx^T \in \mathcal{U} = \{U \in \mathbb{S}^n : U \succeq 0, \exists t \in \mathcal{T} : \text{Tr}(US_k) \leq t_k, k \leq K\}.
\]

Similarly, when \( X \) is a basic spectratope:

\[
X = \{x \in \mathbb{R}^n : \exists t \in \mathcal{T} : S_k^2[x] \preceq t_k I_{d_k}, k \leq K\}
\]

with our usual restrictions on \( \mathcal{T} \) and \( S_k[\cdot] \), it is immediately seen that

\[
x \in X \Rightarrow xx^T \in \mathcal{U} = \{U \in \mathbb{S}^n : U \succeq 0, \exists t \in \mathcal{T} : S_k[U] \preceq t_k I_{d_k}, k \leq K\}.
\]

One can verify that the semidefinite relaxation bounds on the maximum of a quadratic form on an ellitope/spectratope \( X \) derived in Sections 4.2.3 (for elli-
4.8.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 $\mathbb{R}^n$ (recall that this is a norm $p(x)$ which depends solely on $\text{abs}[x]$, where $\text{abs}[x]$ is the vector comprised of the magnitudes of entries in $x$). We can convert $p(\cdot)$ into the norm $p^+(\cdot)$ on the space $S^n$ as follows:

\[ p^+(U) = p([p(\text{Col}_1[U]); \ldots; p(\text{Col}_n[U])]) \quad [U \in S^n] \]

1.1) Prove that $p^+$ indeed is a norm on $S^n$, 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 $B_{p(\cdot)} = \{x : p(x) \leq 1\}$ of the norm $p$, the set

\[ \mathcal{U}_{p(\cdot)} = \{U \in S^n : U \succeq 0, p^+(U) \leq 1\} \]

is matching $\mathcal{X}$. If, in addition,

\[ \mathcal{X} \subset \{x : p(x) \leq 1, Px = 0\}, \quad (4.78) \]

then the set

\[ \mathcal{U}_{p(\cdot), P} = \{U \in S^n : U \succeq 0, p^+(U) \leq 1, PU = 0\} \]

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 $p(x) \leq 1$, the vector $[x_1^2; \ldots; x_n^2]$ belongs to $\mathcal{D}$; in the sequel we call such $\mathcal{D}$ square-dominating $p(\cdot)$. For example, when $p(\cdot) = \|\cdot\|_r$, we can take

\[ \mathcal{D} = \left\{ \begin{array}{ll}
\{y \in \mathbb{R}^n_+ : \sum_i y_i \leq 1\}, & r \leq 2 \\
\{y \in \mathbb{R}^n_+ : \|y\|_{r/2} \leq 1\}, & r > 2 \end{array} \right. . \]

Prove that in this situation the above construction can be refined: whenever $\mathcal{X}$ satisfies $(4.78)$, the set

\[ \mathcal{U}_{p(\cdot), P}^\mathcal{D} = \{U \in S^n : U \succeq 0, p^+(U) \leq 1, PU = 0, \text{dg}(U) \in \mathcal{D}\} \]

\[ [\text{dg}(U) = [U_{11}; U_{22}; \ldots; U_{nn}]] \]

matches $\mathcal{X}$.

Note: in the sequel, we suppress $P$ in the notation $\mathcal{U}_{p(\cdot), P}$ and $\mathcal{U}_{p(\cdot), P}^\mathcal{D}$ 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

\[ p^+(U) = \|U\|_r := \left\{ \begin{array}{ll}
\left(\sum_{i,j} |U_{ij}|^r\right)^{1/r}, & 1 \leq r < \infty, \\
\max_{i,j} |U_{ij}|, & r = \infty \end{array} \right. . \]
1.4) Let $X = \{x \in \mathbb{R}^n : \|x\|_1 \leq 1\}$ and $p(x) = \|x\|_1$, so that $X \subset \{x : p(x) \leq 1\}$, and
\[
\text{Conv}\{[x]^2 : x \in X\} \subset D = \left\{y \in \mathbb{R}^n_+ : \sum_i y_i = 1\right\}.
\] (4.79)

What are the bounds $\text{Opt}_{U p}(\cdot)(B)$ and $\text{Opt}_{D p}(\cdot)(B)$? Is it true that the former (the latter) of the bounds is precise? Is it true that the former (the latter) bound is precise when $B \succeq 0$?

1.5) Let $X = \{x \in \mathbb{R}^n : \|x\|_2 \leq 1\}$ and $p(x) = \|x\|_2$, so that $X \subset \{x : p(x) \leq 1\}$ and (4.79) holds true. What are the bounds $\text{Opt}_{U p}(\cdot)(B)$ and $\text{Opt}_{D p}(\cdot)(B)$? Is the former (the latter) bound?

1.6) Let $X \subset \mathbb{R}^n_+$ be closed, convex, bounded, and with a nonempty interior. Verify that the set
\[
X^+ = \{x \in \mathbb{R}^n : \exists y \in X : \text{abs}[x] \leq y\}
\]
is the unit ball of an absolute norm $p_X$, and this is the largest absolute norm $p(\cdot)$ such that $X \subset \{x : p(x) \leq 1\}$. Derive from this observation that the norm $p_X(\cdot)$ is the best (i.e., resulting in the least conservative bounding scheme) among absolute norms which allow to upper-bound $\text{Opt}_X(B)$ via the construction from item 1.2.

4.8.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 $X_\ell \subset \mathbb{R}^{n_\ell}$, $0 \in X_\ell$, along with matching $X_\ell$ convex compact sets $U_\ell \subset S^{n_\ell}$ giving rise to the closed convex cones
\[
U_\ell = \text{cl}\{(U_\ell, \mu_\ell) \in S^{n_\ell} \times \mathbb{R}^+ : \mu_\ell > 0, \mu_\ell^{-1} U_\ell \in U_\ell\}.
\]

We denote by $\vartheta_\ell(\cdot)$ the Minkowski functions of $X_\ell$:
\[
\vartheta_\ell(y_\ell) = \inf\{t : t > 0, t^{-1} y_\ell \in X_\ell\} : \mathbb{R}^{n_\ell} \to \mathbb{R} \cup \{+\infty\};
\]

note that $X_\ell = \{y_\ell : \vartheta_\ell(y_\ell) \leq 1\}$;
- $n_\ell \times n$ matrices $A_\ell$ such that $\sum_\ell A_\ell^T A_\ell \succ 0$.

On the top of it, we are given a monotone convex set $T \subset \mathbb{R}^L_+$ intersecting the interior of $\mathbb{R}^L_+$. These data specify the convex set
\[
\mathcal{X} = \{x \in \mathbb{R}^n : \exists t \in T : \vartheta_\ell^2(A_\ell x) \leq t_\ell, \ell \leq L\}
\] (*)&

2.1) Prove the following

**Lemma 4.27.** In the situation in question, the set
\[
\mathcal{U} = \{U \in S^n : U \succeq 0 \& \exists t \in T : (A_\ell U A_\ell^T, t_\ell) \in U_\ell, \ell \leq L\}
\]
is a closed and bounded convex set which matches $\mathcal{X}$. As a result, the efficiently
2.4) Consider the “direct product” case where $X \subset \mathbb{R}^n$ is a nonempty convex compact set, $P$ is $m \times n$ matrix, and $U$ matches $X$, then the set $V = \{ V = PUP^T : U \in U \}$ matches $Y = \{ y : \exists x \in X : y = Px \}$.

2.3) Prove that if $X \subset \mathbb{R}^n$ is a nonempty convex compact set, $P$ is $n \times m$ matrix of rank $m$, and $U$ matches $X$, then the set $V = \{ V \geq 0 : PVP^T \in \mathcal{U} \}$ matches $Y = \{ y : Py \in X \}$.

2.2) Prove that if $X \subset \mathbb{R}^n$ is a nonempty convex compact set, $P$ is $m \times n$ matrix, and $U$ matches $X$, then the set $V = \{ V = PUP^T : U \in \mathcal{U} \}$ matches $Y = \{ y : \exists x \in X : y = Px \}$.

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 $X$ and claim that this set matches item 2.1 is as follows: given the sets $U$ and $U'$, the setup under consideration. In the direct product case, the construction from item 2.1 is as follows: given the sets $U$ matching $X$, we build the set

$$U = \{ U = [U^{\ell_1 \ell_2}] \in \mathcal{U} : \ell, \ell' \leq L \}$$

and claim that this set matches $X$. 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 $U$ are built via absolute norms. Namely, let $p_\ell(\cdot) : \mathbb{R}^{n_\ell} \to \mathbb{R}_+, \ell \leq L$, be absolute norms, let sets $D_\ell$ be square-dominating $p_\ell(\cdot)$,

$$X_\ell \subset \hat{X}_\ell = \{ x^\ell \in \mathbb{R}^n : P_\ell x_\ell = 0, p_\ell(x^\ell) \leq 1 \},$$

and let

$$\mathcal{U}_\ell = \{ U \in S^{n_\ell} : U \geq 0, P_\ell U = 0, p_\ell(U) \leq 1, \text{dg}(U) \in D_\ell \}.$$

In this case the above construction results in

$$U = \left\{ U = [U^{\ell_1 \ell_2}] \in \mathcal{U} : \ell, \ell' \leq L \right\}$$

and let

$$U = \{ U = [U^{\ell_1 \ell_2}] \in \mathcal{U} : \ell, \ell' \leq L \}.$$

Now let

$$p([x^1; \ldots; x^L]) = \max[p_1(x^1), \ldots, p_L(x^L)] : \mathbb{R}^{n_1} \times \ldots \times \mathbb{R}^{n_L} \to \mathbb{R},$$

so that $p$ is an absolute norm and

$$X \subset \{ x = [x^1; \ldots; x^L] : p(x) \leq 1, P_\ell x^\ell = 0, \ell \leq L \}.$$

Prove that in fact the set

$$\mathcal{U} = \left\{ U = [U^{\ell_1 \ell_2}] \in \mathcal{U} : \ell, \ell' \leq L \right\}$$

is an upper bound on

$$\text{Opt}_X(B) = \max_{x \in X} x^T B x.$$
matches \( \mathcal{X} \), and that we always have \( \mathcal{U} \subseteq \mathcal{U} \). Verify that in general this inclusion is strict.

4.8.C Illustration: Nullspace property revisited. Recall sparsity-oriented signal recovery via \( \ell_1 \) minimization from Chapter 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

\[
\hat{w} \in \text{Argmin} \{ \|z\|_1 : Az = y \}.
\]

We called matrix \( A \) \( s \)-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:

\[
\text{Opt} := \max_z \{ \|z\|_{(s)} : z \in \text{Ker} A, \|z\|_1 \leq 1 \} < 1/2,
\]

where \( \|z\|_{(k)} \) is the sum of the \( k \) largest entries in the vector \( \text{abs}[z] \). A verifiable sufficient condition for \( s \)-goodness is

\[
\widehat{\text{Opt}} := \min_H \max_{j} \|\text{Col}_j[I - H^T A]\|_{(s)} < \frac{1}{2}, \quad (4.80)
\]

the reason being that, as it is immediately seen, \( \widehat{\text{Opt}} \) is an upper bound on \( \text{Opt} \) (see Proposition 1.9 with \( q = 1 \)).

An immediate observation is that \( \text{Opt} \) is nothing but the maximum of quadratic form over an appropriate convex compact set. Specifically, let

\[
\mathcal{X} = \{ [u; v] \in \mathbb{R}^n \times \mathbb{R}^n : Au = 0, \|u\|_1 \leq 1, \sum_i |v_i| \leq s, \|v\|_\infty \leq 1 \},
\]

\[
B = \begin{bmatrix} \frac{1}{2} I_n \end{bmatrix}.
\]

Then

\[
\text{Opt}_\mathcal{X}(B) = \max_{[u; v] \in \mathcal{X}} \{ u; v \}^T B \{ u; v \}
\]

\[
= \max_{u, v} \{ u^T v : Au = 0, \|u\|_1 \leq 1, \sum_i |v_i| \leq s, \|v\|_\infty \leq 1 \}
\]

\[
= \max_u \{ \|u\|_{(s)} : Au = 0, \|u\|_1 \leq 1 \}
\]

\[
= \text{Opt},
\]

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

\[
V = \{ v \in \mathbb{R}^n : \sum_i |v_i| \leq s, \|v\|_\infty \leq 1 \}
\]

are exactly the vectors with at most \( s \) nonzero entries, the nonzero entries being \( \pm 1 \); as a result

\[
\forall (z \in \mathbb{R}^n) : \max_{v \in V} z^T v = \|z\|_{(s)}.
\]
Now, $V$ is the unit ball of the absolute norm

$$r(v) = \min \{ t : \|v\|_1 \leq st, \|v\|_\infty \leq t \},$$

so that $X$ is contained in the unit ball $B$ of the absolute norm on $\mathbb{R}^{2n}$ specified as

$$p([u; v]) = \max \{ \|u\|_1, r(v) \} \quad [u, v \in \mathbb{R}^n],$$

i.e.,

$$X = \{ [u; v] : p([u, v]) \leq 1, Au = 0 \}.$$

As a result, whenever $x = [u; v] \in X$, the matrix

$$U = xx^T = \begin{bmatrix} U_{11} = uu^T & U_{12} = uv^T \\ U_{21} = vu^T & U_{22} = vv^T \end{bmatrix}$$

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

$$U = \{ U = \begin{bmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{bmatrix} \in S^{2n} : U \succeq 0, p^+(U) \leq 1, AU_{11} = 0, AU_{12} = 0 \}$$

(which clearly is a nonempty convex compact set) matches $X$. As a result, the efficiently computable quantity

$$\text{Opt} = \max_{U \in \mathcal{U}} \text{Tr}(BU)$$

is an upper bound on $\text{Opt}$. As a result, the verifiable condition

$$\text{Opt} < 1/2$$

is sufficient for $s$-goodness of $A$.

Now goes the concluding part of the exercise:

3.1) Prove that $\hat{\text{Opt}} \leq \text{Opt}$, so that (4.81) is less conservative than (4.80).

**Hint:** Apply Conic Duality to verify that

$$\hat{\text{Opt}} = \max_{V} \left\{ \text{Tr}(V) : V \in \mathbb{R}^{n \times n}, AV = 0, \sum_{i=1}^{n} r(\text{Col}_i[V^T]) \leq 1 \right\}. \quad (4.82)$$

3.2) Run simulations with randomly generated Gaussian matrices $A$ and play with different values of $s$ to compare $\hat{\text{Opt}}$ and $\text{Opt}$. To save time, you can use toy sizes $m, n$, say, $m = 18, n = 24$. 

4.7.4 Around Propositions 4.4 and 4.14

4.7.4.1 Optimizing linear estimates on convex hulls of unions of spectratopes

Exercise 4.9.

Let

- $X_1, \ldots, X_J$ be spectratopes in $\mathbb{R}^n$:
  
  $$X_j = \{ x \in \mathbb{R}^n : \exists (y \in \mathbb{R}^{N_j}, t \in T_j) : x = P_j y, R^2_{kj}[y] \preceq t_k I_{d_k}, k \leq K_j \}, 1 \leq j \leq J$$

- $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{\nu \times n}$ be given matrices,
- $\| \cdot \|$ be a norm on $\mathbb{R}^\nu$ such that the unit ball $B_*^*$ of the conjugate norm $\| \cdot \|_*$ is a spectratope:
  
  $$B_* := \{ u : \| u \|_* \leq 1 \} = \{ u \in \mathbb{R}^\nu : \exists (z \in \mathbb{R}^N, r \in \mathbb{R}) : u = M z, S^2_{\ell}[z] \preceq r \ell I_{r\ell}, \ell \leq L \}$$
  
  $$S_{\ell}[z] = \sum_{i=1}^N z_i S_{\ell i}$$

- $\Pi$ be a convex compact subset of the interior of the positive semidefinite cone $S^+_m$,

with our standard restrictions on $R_{kj}[\cdot], S_{\ell}[\cdot], T_j$ and $\mathcal{R}$. Let, further,

$$X = \text{Conv} \left( \bigcup_j X_j \right)$$

be the convex hull of the union of spectratopes $X_j$. Consider the situation where, given observation

$$\omega = Ax + \xi$$

of unknown signal $x$ known to belong to $X$, we want to recover $Bx$. We assume that the matrix of second moments of noise is $\succeq$-dominated by a matrix from $\Pi$, and quantify the performance of a candidate estimate $\hat{x}(\cdot)$ by its $\| \cdot \|_*$-risk

$$\text{Risk}_{\Pi, \| \cdot \|_*} [\hat{x}, X] = \sup_{x \in X} \sup_{P \in \Pi} \mathbb{E}_{\xi \sim P} \{ \| Bx - \hat{x}(Ax + \xi) \| \}$$

where $P \subset \Pi$ means that the matrix $\text{Var}[P] = \mathbb{E}_{\xi \sim P} \{ \xi \xi^T \}$ of second moments of distribution $P$ is $\succeq$-dominated by a matrix from $\Pi$.

Prove the following

**Proposition 4.28.** In the situation in question, consider convex optimization problem

$$\text{Opt} = \min_{H, \Theta, \Lambda^j, \Upsilon^j, \Upsilon} \left\{ \max_{j} \left[ \phi_{\mathcal{T}_j}(\lambda[\Lambda^j]) + \phi_{\mathcal{R}}(\lambda[\Upsilon^j]) \right] + \phi_{\mathcal{R}}(\lambda[\Upsilon^j]) + \Gamma_{\Pi}(\Theta) : \right\}$$
\[ \sum_{j} \Lambda^j = \{ \Lambda^j \geq 0, j \leq K_j, j \leq J, \} \]

\[ \frac{1}{2} \sum_{j} \Lambda^j \left( A^j \right) \geq 0, j \leq J, \]

\[ \left( \sum_{j} \rho^j \Lambda^j \right) \geq 0, j \leq J, \]

\[ \left( \frac{1}{2} \sum_{j} H^j \right) \geq 0, j \leq J, \]

where, as usual,

\[ \phi_T (\lambda) = \max_{t \in T_j} t^T \lambda, \]

\[ \phi_R (\lambda) = \max_{r \in R_j} r^T \lambda, \]

\[ \Gamma_H (\Theta) = \max_{Q : Q \in \Pi} \text{Tr} (Q \Theta), \]

\[ \lambda [U_1, ..., U_s] = \left[ \text{Tr} (U_1); ..., \text{Tr} (U_s) \right], \]

\[ S^*_T : S^T \to S^N : S^*_T [U] = \left[ \text{Tr} (S^{lp} U S^{jq}) \right]_{p,q \leq N}, \]

\[ R^*_k : S^{dk} \to S^{N^j} : R^*_k [U] = \left[ \text{Tr} (R^{kp} U R^{jq}) \right]_{p,q \leq N_j}. \]

Problem (4.83) is solvable, and \( H \)-component \( H^* \) of its optimal solution gives rise to linear estimate \( \hat{x}_{H^*} (\omega) = H^T \omega \) such that

\[ \text{Risk}_{H, \Pi, \| \cdot \|} (\hat{x}_{H^*} | \mathcal{X}) \leq \text{Opt}. \] (4.84)

Moreover, the estimate \( \hat{x}_{H^*} \) is near-optimal among linear estimates:

\[ \text{Opt} \leq O(1) \ln (D + F) \text{Risk}_{\Pi, \text{Lin}} \]

\[ \left[ D = \max_{j} \sum_{k \leq K_j} d_{kj}, \ F = \sum_{\ell \leq L} f_{\ell} \right] \] (4.85)

where

\[ \text{Risk}_{\Pi, \text{Lin}} = \inf_{H} \sup_{x \in \mathcal{X}, Q : Q \in \Pi} \mathbb{E}_{\xi \sim \mathcal{N}(0, Q)} \left\{ \| Bx - H^T (Ax + \xi) \| \right\} \]

is the best risk attainable by linear estimates in the current setting under zero mean Gaussian observation noise.

It should be stressed that convex hull of unions of spectrata not necessarily is a spectratope, and that Proposition 4.28 states that the linear estimate stemming from (4.83) is near-optimal only among linear, not among all estimates (the latter might indeed not be the case).

4.7.4.2 Recovering nonlinear vector-valued functions

Exercise 4.10.

Consider the situation as follows: We are given a noisy observation

\[ \omega = Ax + \xi_x \]

of the linear image \( Ax \) of an unknown signal \( x \) known to belong to a given spectratope \( \mathcal{X} \subset \mathbb{R}^n \); here \( \xi_x \) is the observation noise with distribution \( P_x \) which can depend on \( x \). Similarly to Section 4.3.3, we assume that we are given a computationally tractable convex compact set \( \Pi \subset \text{int} \mathbb{S}_m^+ \) such that for every \( x \in \mathcal{X} \), \( \text{Var} [P_x] \preceq \Theta \) for some \( \Theta \in \Pi \), cf. (4.32). We want to recover the value \( f(x) \) of a given vector-valued function \( f : \mathcal{X} \to \mathbb{R}^\nu \), and we measure the recovery error in a
given norm $|\cdot|$ on $\mathbb{R}^\nu$.

### 4.10.A. Preliminaries and Main observation

Let $\|\cdot\|$ be a norm on $\mathbb{R}^n$, and $g(\cdot): \mathcal{X} \to \mathbb{R}^\nu$ 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

$$|g(x) - g(y)| \leq L\|x - y\| \quad \forall (x, y) \in \mathcal{X};$$

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 $\mathbb{R}^n$ appropriate for $f$, is $f$ is Lipschitz continuous with constant 1 w.r.t. $\|\cdot\|, |\cdot|$. Our immediate observation is as follows:

**Observation 4.29.** 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 $\hat{x}(\omega)$ of $x$ via $\omega$ such that $\hat{x}(\cdot) \in \mathcal{X}$ induces the “plug-in” estimate

$$\hat{f}(\omega) = f(\hat{x}(\omega))$$

of $f(x)$, and the $\|\cdot\|$-risk

$$\text{Risk}_{\|\cdot\|}[\hat{x}|\mathcal{X}] = \sup_{x \in \mathcal{X}} E_{\xi \sim P_x} \{ \| \hat{x}(Ax + \xi) - x \| \}$$

of estimate $\hat{x}$ upper-bounds the $|\cdot|$-risk

$$\text{Risk}_{|\cdot|}[\hat{f}|\mathcal{X}] = \sup_{x \in \mathcal{X}} E_{\xi \sim P_x} \{ \| \hat{f}(Ax + \xi) - f(x) \| \}$$

of the induced by $\hat{x}$ estimate $\hat{f}$:

$$\text{Risk}_{|\cdot|}[\hat{f}|\mathcal{X}] \leq \text{Risk}_{\|\cdot\|}[\hat{x}|\mathcal{X}].$$

When $f$ is defined and Lipschitz continuous with constant 1 w.r.t. $\|\cdot\|, |\cdot|$ on the entire $\mathbb{R}^n$, this conclusion remains valid without the assumption that $\hat{x}$ is $\mathcal{X}$-valued.

### 4.10.B. Consequences

Observation 4.29 suggests the following simple approach to solving the estimation problem we started with: assuming that we have at our disposal a norm $\|\cdot\|$ on $\mathbb{R}^n$ such that

- $\|\cdot\|$ is appropriate for $f$, and
- $\|\cdot\|$ is good, goodness meaning that the unit ball $B_*$ 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 4.3.3 to build a near-optimal, in terms of its $\|\cdot\|$-risk, linear estimate of $x$ via $\omega$, and convert this estimate in an estimate of $f(x)$. By the above 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 correction: in general, the linear estimate
\( \tilde{x}(\cdot) \) yielded by Proposition 4.14 (same as any nontrivial – not identically zero – linear estimate) is not guaranteed to take values in \( \mathcal{X} \), which is, in general, required for Observation 4.29 to be applicable. This correction is easy: it is enough to convert \( \tilde{x} \) into the estimate \( \hat{x} \) defined by

\[
\hat{x}(\omega) \in \text{Argmin}_{u \in \mathcal{X}} \|u - \tilde{x}(\omega)\|.
\]

This transformation preserves efficient computability of the estimate, and ensures that the corrected estimate takes its values in \( \mathcal{X} \); at the same time, “correction” \( \tilde{x} \mapsto \hat{x} \) nearly preserves the \( \|\cdot\| \)-risk:

\[
\text{Risk}_{\|\cdot\|}[\hat{x}|\mathcal{X}] \leq 2\text{Risk}_{\|\cdot\|}[\tilde{x}|\mathcal{X}]. \tag{*}
\]

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 (*)

4.10.C. How to select \( \|\cdot\| \). When implementing the outlined approach, the major question is how to select a norm \( \|\cdot\| \) appropriate for \( f \). The best choice would be to select the smallest among the norms appropriate for \( f \) (such a norm does exist under mild assumptions), because the smaller \( \|\cdot\| \), the smaller 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 \( f \); second, our approach 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) overcome.

Example: 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 none 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, such as \( f(x) = \|x-c\| \), 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

\[
f(x) = \sup_{\alpha \in A} \left\{ \sup_{c \in C_\alpha} \inf \|x-c\| \right\}.
\]

In such case, 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 consider 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 \( \mathbb{R}^n \) 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 \( \mathcal{X}^o = \text{int} \mathcal{X} \), and \( f \) is Lipschitz continuous with constant \( 1 \) w.r.t. a norm \( \| \cdot \| \) if and only if
\[
\| f'(x) \| \| \| \| \to \| \| f \| \leq 1
\]
whenever \( x \in \mathcal{X}^o \) is such that the derivative (a.k.a. Jacobian) of \( f \) at \( x \) exists; here \( \| Q \| \| \| \to \| \| f \| \) is the matrix norm of a \( n \times n \) matrix \( Q \) induced by the norms \( \| \cdot \| \) on \( \mathbb{R}^n \) and \( | \cdot | \) on \( \mathbb{R}^q \):
\[
\| Q \| \| \| \to \| \| f \| := \max_{\| x \| \leq 1} | Qx | = \max_{\| y \| \leq 1} y^T Qx = \max_{\| y \| \leq 1} x^T Q^T y = \| Q^T \| \| \| \to \| \| f, \| |\|,
\]
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
\[
B_{f,*} = \text{cl Conv} \{ z : \exists (x \in \mathcal{X}^o, y, | y |_* \leq 1) : z = [f'(x)]^Ty \},
\]
where \( \mathcal{X}^o \) is the set of all \( x \in \mathcal{X}^o \) where \( f'(x) \) exists. Geometrically, \( B_{f,*} \) is the closed convex hull of the union of all images of the unit ball \( B_* \) 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
\[
\| u \| \geq \| u \|_f := \max_{z \in B_{f,*}} z^Tu.
\]

Check that \( \| u \|_f \) is a norm, provided that \( B_{f,*} \) (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 \( \mathbb{R}^n \) for which the unit ball of the conjugate norm is a spectratope containing \( B_{f,*} \).

**Example.** Consider the case of componentwise quadratic \( f \):
\[
f(x) = \left[ \frac{1}{2} x^T Q_1 x; \frac{1}{2} x^T Q_2 x; \ldots; \frac{1}{2} x^T Q_n x \right] \quad [Q_i \in \mathbb{S}^n]
\]
and \( |u| = \| u \|_q \) with \( q \in [1, 2] \). \(^{10}\) In this case
\[\mathcal{B}_* = \{ u \in \mathbb{R}^n : \| u \|_p \leq 1 \}, \quad p = \frac{q}{q-1} \in [2, \infty], \quad \text{and} \ f'(x) = \left[ x^T Q_1; x^T Q_2; \ldots; x^T Q_n \right].\]

Setting \( \mathcal{S} = \{ s \in \mathbb{R}_+^n : \| s \|_{p/2} \leq 1 \} \) and
\[
\mathcal{S}^{1/2} = \{ s \in \mathbb{R}_+^n : [s_1^2; \ldots; s_n^2] \in \mathcal{S} \} = \{ s \in \mathbb{R}_+^n : \| s \|_p \leq 1 \},
\]

\(^{10}\)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 two distinct problems 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).
the set
\[ \mathcal{Z} = \{ [f'(x)]^T u : x \in \mathcal{X}, u \in \mathcal{B} \} \]
is contained in the set
\[ \mathcal{Y} = \{ y \in \mathbb{R}^n : \exists (s \in \mathcal{S}^{1/2}, x^i \in \mathcal{X}, i \leq \nu) : y = \sum_i s_i Q_i x_i \} \].

Set \( \mathcal{Y} \) is a spectratope with spectratopic representation readily given by that of \( \mathcal{X} \). Indeed, \( \mathcal{Y} \) is nothing but the \( \mathcal{S} \)-sum of the spectratopes \( Q_i \mathcal{X} \), \( i = 1, \ldots, \nu \), see Section 4.10. As a result, we can use the spectratope \( \mathcal{Y} \) (when \( \text{int} \mathcal{Y} \neq \emptyset \)) or the arithmetic sum of \( \mathcal{Y} \) with a small Euclidean ball (when \( \text{int} \mathcal{Y} = \emptyset \)) as the unit ball of the norm conjugate to \( \| \cdot \| \), thus ensuring, by 2), that \( \| \cdot \| \) is appropriate for \( f \).

We then can use \( \| \cdot \| \) in order to build an estimate of \( f(\cdot) \).

3.1) For illustration, work out the problem of recovering the value of a scalar quadratic form
\[ f(x) = x^T M x, \quad M = \text{Diag}\{ i^\alpha, i = 1, \ldots, n \} \quad [\nu = 1, |\cdot| \text{ is the absolute value}] \]
from noisy observation
\[ \omega = Ax + \sigma \eta, \quad A = \text{Diag}\{ i^\beta, i = 1, \ldots, n \}, \quad \eta \sim \mathcal{N}(0, I_n) \]
of a signal \( x \) known to belong to the ellipsoid
\[ \mathcal{X} = \{ x \in \mathbb{R}^n : \| P x \|_2 \leq 1 \}, \quad P = \text{Diag}\{ i^\gamma, i = 1, \ldots, n \}, \]
where \( \alpha, \beta, \gamma \) are given reals satisfying
\[ \alpha - \gamma - \beta < -1/2. \]

You could start with the simplest unbiased estimate
\[ \tilde{x}(\omega) = [1^{-\beta} \omega_1; 2^{-\beta} \omega_2; \ldots; n^{-\beta} \omega_n] \]
of \( x \).

3.2) Work out the problem of recovering the norm
\[ f(x) = \| M x \|_p, \quad M = \text{Diag}\{ i^\alpha, i = 1, \ldots, n \}, \quad p \in [1, 2], \]
from observation (4.86) with
\[ \mathcal{X} = \{ x : \| P x \|_r \leq 1 \}, \quad P = \text{Diag}\{ i^\gamma, i = 1, \ldots, n \}, \quad r \in [2, \infty]. \]

4.7.4.3 Suboptimal linear estimation

Exercise 4.11.
[recovery of large-scale signals] Consider the problem of estimating the image \( B x \in \mathbb{R}^\nu \) of signal \( x \in \mathcal{X} \) from observation
\[ \omega = Ax + \sigma \xi \in \mathbb{R}^m \]
in the simplest case where \( \mathcal{X} = \{ x \in \mathbb{R}^n : x^T S x \leq 1 \} \) is an ellipsoid (so that \( S > 0 \)), the recovery error is measured in \( \| \cdot \|_2 \), and \( \xi \sim \mathcal{N}(0, I_m) \). In this case, Problem (4.12) to solve when building “presumably good linear estimate” reduces to

\[
\text{Opt} = \min_{H, \lambda} \left\{ \lambda + \sigma^2 \| H \|_F^2 : \begin{bmatrix} \lambda S & B^T - A^T H \\ B - H^T A & 0 \end{bmatrix} \succeq 0 \right\}, \tag{4.87}
\]

where \( \| \cdot \|_F \) is the Frobenius norm of a matrix. An optimal solution \( H_\ast \) to this problem results in the linear estimate \( \hat{x}_{H_\ast}(\omega) = H_\ast^T \omega \) satisfying the risk bound

\[
\text{Risk}[\hat{x}_{H_\ast} | \mathcal{X}] := \max_{x \in \mathcal{X}} \sqrt{\mathbb{E}\left[ \| B x - H_\ast^T (A x + \sigma \xi) \|_2^2 \right]} \leq \sqrt{\text{Opt}}.
\]

Now, (4.87) is an efficiently solvable convex optimization problem. However, when the sizes \( m, n \) of the problem are large, solving the problem by standard optimization techniques could become prohibitively time-consuming. The goal of what follows is to develop relatively cheap computational technique for finding a good enough suboptimal solution to (4.87). In the sequel, we assume that \( A \neq 0 \), otherwise (4.87) is trivial.

1) Prove that problem (4.87) can be reduced to similar problem with diagonal positive semidefinite matrix \( A \), the reduction requiring several singular value decompositions and multiplications of matrices of the same sizes as those of \( A, B \) and \( S \).

2) By item 1, we can assume from the very beginning that \( S = I \) and \( A = \text{Diag}\{\alpha_1, ..., \alpha_n\} \) with \( 0 \leq \alpha_1 \leq \alpha_2 \leq ... \leq \alpha_n \). Passing in (4.87) from variables \( \lambda, H \) to variables \( \tau = \sqrt{\lambda}, G = H^T \), the problem becomes

\[
\text{Opt} = \min_{G, \tau} \left\{ \tau^2 + \sigma^2 \| G \|_F^2 : \| B - G A \| \leq \tau \right\}, \tag{4.88}
\]

where \( \| \cdot \| \) is the spectral norm. Now consider the construction as follows:

- Consider a partition \( \{1, ..., n\} = I_0 \cup I_1 \cup \ldots \cup I_K \) 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 \( \alpha_j/\alpha_i, i, j \in I_k, \) do not exceed \( \theta > 1 \) (\( \theta \) is the parameter of our construction), while
  (c) for \( 1 \leq k < k' \leq K, \) the ratios \( \alpha_j/\alpha_i, i \in I_k, j \in I_{k'}, \) are \( \theta. \)

The recipe for building the partition is self-evident, and we clearly have

\[
K \leq \ln(\pi/\alpha) / \ln(\theta) + 1,
\]

where \( \pi \) is the largest of \( \alpha_i, \) and \( \alpha \) is the smallest of those \( \alpha_i, \) which are positive.

- For \( 1 \leq k \leq K, \) we denote by \( i_k \) the first index in \( I_k, \) set \( \alpha_k = \alpha_{i_k}, \) \( n_k = \text{Card} I_k, \) and define \( A_k \) as \( n_k \times n_k \) diagonal matrix with diagonal entries \( \alpha_i, \) \( i \in I_k. \)

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:
\[
\begin{align*}
\text{Opt}_k^*(\tau) &= \min_{G_k \in \mathbb{R}^{n \times n}} \left\{ \|G_k\|_F^2 : \|B_k - G_k A_k\| \leq \tau \right\} \quad (P_k^*[\tau]) \\
\text{Opt}_k^*(\tau) &= \min_{G_k \in \mathbb{R}^{n \times n}} \left\{ \|G_k\|_F^2 : \|B_k - \alpha^k G_k\| \leq \tau \right\} \quad (P_k[\tau])
\end{align*}
\]

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
\[
\text{Opt}_k^*(\tau) \leq \text{Opt}_k(\tau) \leq \theta^2 \text{Opt}_k^*(\tau);
\]

2.2) Problems \((P_k[\tau])\) are easy to solve: if \( B_k = U_k D_k V_k^T \) is the 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
\[
\hat{G}_k[\tau] = [\alpha^k]^{-1} U_k D_k[\tau] V_k^T,
\]
where \( D_k[\tau] \) is the diagonal matrix obtained from \( D_k \) by truncating the diagonal entries \( \sigma_{k\ell} \mapsto [\sigma_{k\ell} - \tau]_+ \) (from now on, \( a_+ = \max[a, 0], a \in \mathbb{R} \)). The optimal value in \((P_k[\tau])\) is
\[
\text{Opt}_k(\tau) = [\alpha^k]^{-2} \sum_{\ell=1}^{\nu_k} [\sigma_{k\ell} - \tau]^2_+.
\]

2.3) If \((\tau, G)\) is a feasible solution to \((4.88)\) then \( \tau \geq \bar{\tau} := \|B_0\| \), and the matrices \( G_k, 1 \leq k \leq K \), are feasible solutions to problems \((P_k^*[\tau])\), implying that
\[
\sum_{k} \text{Opt}_k^*(\tau) \leq \|G\|_F^2.
\]

And vice versa: if \( \tau \geq \bar{\tau}, G_k, 1 \leq k \leq K \), are feasible solutions to problems \((P_k^*[\tau])\), and
\[
K_+ = \left\{ K, \begin{array}{c} K + 1, \quad I_0 = \emptyset \end{array} \right\}
\]
then the matrix \( G = [0_{\nu \times n_0}, G_1, \ldots, G_K] \) and \( \tau_+ = \sqrt{K_+} \tau \) form a feasible solution to \((4.88)\).

Extract from these observations that if \( \tau_* \) is an optimal solution to the convex optimization problem
\[
\min_{\tau} \left\{ \theta^2 \tau^2 + \sigma^2 \sum_{k=1}^{K} \text{Opt}_k(\tau) : \tau \geq \bar{\tau} \right\} \quad (4.89)
\]
and \( G_{k,*} \) are optimal solutions to the problems \((P_k[\tau_*])\), then the pair
\[
\hat{\tau} = \sqrt{K_+} \tau_*, \quad \hat{G} = [0_{\nu \times n_0}, G_{1,*}, \ldots, G_{K,*}] \quad \quad \quad \quad \quad [G_{k,*} = \alpha^k G_{k,*} A_k^{-1}]$

is a feasible solution to (4.88), 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 $\text{Opt}$ of this problem. As a result, $\hat{G}$ gives rise to a linear estimate with risk on $X$ which is within the factor $\max[\sqrt{K_+}, \theta]$ of the risk $\sqrt{\text{Opt}}$ of the “presumably good” linear estimate yielded by an optimal solution to (4.87).

Notice that

- After carrying out singular value decompositions of matrices $B_k$, $1 \leq k \leq K$, specifying $\tau_*$ 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 $\alpha/\alpha$ of $A$.

Your next task is as 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 linear estimate yielded by the optimal solution to (4.12)\textsuperscript{11} 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 (4.12) 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)\textsuperscript{11}
- range of $\sigma$: \{1.0, 0.01, 0.0001\}
- generation of $S$, $A$, $B$: generate the matrices at random according to

$$
S = U_S \text{Diag}\{1, 2, ..., n\} U_S^T, \quad A = U_A \text{Diag}\{\mu_1, ..., \mu_n\} V_A^T, \quad B = U_B \text{Diag}\{\mu_1, ..., \mu_n\} V_B^T,
$$

where $U_S, U_A, V_A, U_B, V_B$ are random orthogonal $n \times n$ matrices, and $\mu_i$ form a geometric progression with $\mu_1 = 0.01$ and $\mu_n = 1$.

You could run the above construction for several values of $\theta$ and select the best, in terms of its risk bound, of the resulting suboptimal estimates.

\textbf{4.11.A. Simple case.} There is a trivial case where (4.88) is really easy; this is the case where the right orthogonal factors in the singular value decompositions of $A$ and $B$ are the same, that is, when

$$
B = W F V^T, \quad A = U D V^T
$$

with orthogonal $n \times n$ matrices $W, U, V$ and diagonal $F, D$. This, at the first glance,

\textsuperscript{11}When $X$ is an ellipsoid, semidefinite relaxation bound on the maximum of a quadratic form over $x \in X$ is exact, so that we are in the case when an optimal solution to (4.12) yields the best, in terms of risk on $X$, linear estimate.
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 \( x \in X \) on \( x \). In this situation, setting \( W^T H^T U = G \), problem (4.88) becomes

\[
\text{Opt} = \min_G \left\{ \| F - GD \|^2 + \sigma^2 \| G \|^2_F \right\}.
\]

Now goes the concluding part of the exercise:

4) Prove that in the situation in question an optimal solution \( G^* \) to (4.90) can be selected to be diagonal, with diagonal entries \( \gamma_i, 1 \leq i \leq n \), yielded by the optimal solution to the optimization problem

\[
\text{Opt} = \min_{\gamma} \left\{ f(G) := \max_{i \leq n} (\phi_i - \gamma_i \delta_i)^2 + \sigma^2 \sum_{i=1}^n \gamma_i^2 \right\} \quad [\phi_i = F_{ii}, \delta_i = D_{ii}]
\]

**Exercise 4.12.**

[image reconstruction – follow-up to Exercise 4.11] A grayscale image can be represented by an \( m \times n \) matrix \( x = [x_{pq}] \) \( 0 \leq p < m, \ 0 \leq q < n \) with entries in the range \([-x, x]\), \( x = 255 / 2 \).

Taking picture can be modeled as observing in noise the 2D convolution \( x \ast \kappa \) of image \( x \) with known blurring kernel \( \kappa = [\kappa_{uv}] \) \( 0 \leq u < 2\mu, \ 0 \leq v < 2\nu \), so that the observation is the random matrix

\[
\omega = [\omega_{rs}] = \sum_{0 \leq u \leq 2\mu, 0 \leq v \leq 2\nu} x_{pq} \kappa_{uv} + \sigma \xi_{rs} \quad [0 \leq r < m + 2\mu, \ 0 \leq s < n + 2\nu]
\]

where independent of each other random variables \( \xi_{rs} \sim N(0, 1) \) form the observation noise.\(^{13}\) Our goal is to build a presumably good linear estimate of \( x \) via \( \omega \), the recovery error being measured in \( \| \cdot \|_2 \). To apply the techniques developed in Section 4.2.2, we need to cover the set of signals \( x \) allowed by our a priori assumptions by an ellitope \( X \), and then solve the associated optimization problem (4.12). The difficulty, however, is that this problem is really high-dimensional – with \( 256 \times 256 \) 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 difficult 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.

**4.12.A. The construction.** Our key observation is that when passing from representations of \( x \) and \( \omega \) “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

\[
\alpha = 2\mu + m, \ \beta = 2\nu + n,
\]

\(^{12}\)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 \( \pi \), of a matrix with entries in \([-\pi, \pi]\).

\(^{13}\)Be careful, everywhere in this exercise indexing of elements of 2D arrays starts from 0, and not from 1!
let \( F_{\alpha,\beta} \) be the 2D discrete Fourier Transform – a linear mapping from the space \( \mathbb{C}^{\alpha \times \beta} \) onto itself given by

\[
[F_{\alpha,\beta}y]_{rs} = \frac{1}{\sqrt{\alpha \beta}} \sum_{0 \leq p < \alpha, 0 \leq q < \beta} y_{pq} \exp \left\{ -\frac{2\pi ir}{\alpha} - \frac{2\pi is}{\beta} \right\},
\]

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 \( x \in \mathbb{R}^{m \times n} \), setting \( x^+ = \begin{bmatrix} x & 0_{2\mu \times n} & 0_{2\mu \times 2\nu} \\ 0_{2\nu \times n} & 0_{2\nu \times 2\mu} \end{bmatrix} \in \mathbb{R}^{\alpha \times \beta} \), we have

\[
F_{\alpha,\beta}(x * \kappa) = \chi \cdot [F_{\alpha,\beta}x^+] \tag{14}
\]

with easy-to-compute \( \chi \). Now, let \( \delta \) be another \( (2\mu + 1) \times (2\nu + 1) \) kernel, with the only nonzero entry, equal to 1, in the position \((\mu, \nu)\) (recall that indices are enumerated starting from 0); then

\[
F_{\alpha,\beta}(x * \delta) = \theta \cdot [F_{\alpha,\beta}x^+] \tag{15}
\]

with easy-to-compute \( \theta \).

Now consider the auxiliary estimation problem as follows:

Given \( R > 0 \) and noisy observation

\[
\hat{\omega} = \chi \cdot \hat{x} + \sigma \sqrt{\alpha \beta} F_{\alpha,\beta} \zeta \tag{16}
\]

[\( \zeta = [\xi_{rs}] \) with independent \( \xi_{rs} \sim \mathcal{N}(0, 1) \)],

of signal \( \hat{x} \in \mathbb{C}^{\alpha \times \beta} \) known to satisfy \( \| \hat{x} \|_2 \leq R \), we want to recover the matrix \( \theta \cdot \hat{x} \), the error being measured in the Frobenius norm \( \| \cdot \|_2 \),

Treating signals \( \hat{x} \) and noises \( \eta \) 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 \( \Theta z \) of signal \( z \) known to belong to the centered at the origin Euclidean ball \( \mathcal{Z}_R \) of radius \( R \) in \( \mathbb{C}^{\alpha \beta} \), from noisy observation

\[
\zeta = Az + \sigma \eta,
\]

where \( \Theta \) and \( A \) are diagonal matrices with complex entries, and \( \eta \) 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 \( \tilde{z} = H \zeta \) are concerned, we lose nothing when restricting ourselves with diagonal matrices \( H = \text{Diag}\{h\} \), 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

\[
R^2 \max_{\ell, \mu, \nu} |\Theta_{\ell \ell} - h_{\ell \ell} A_{\ell \ell}|^2 + \sigma^2 \sum_{\ell, \mu, \nu} |h_{\ell \ell}|^2.
\]

\[14\]Here \( \chi = \sqrt{\alpha \beta} F_{\alpha,\beta} \kappa^+ \), where \( \kappa^+ \) is the \( \alpha \times \beta \) matrix with \( \kappa \) as its \((2\mu + 1) \times (2\nu + 1)\) upper-left block and zeros outside this block.
Coming back to the initial setting of our auxiliary estimation problem, we conclude that the best linear recovery of \( \theta \cdot \hat{x} \) via \( \hat{\omega} \) is given by

\[
\hat{z} = h \cdot \hat{\omega},
\]

where \( h \) is an optimal solution to the optimization problem

\[
\text{Opt} = \min_{h \in C^{\alpha \times \beta}} \left\{ R^2 \max_{r,s} |\theta_{rs} - h_{rs} \chi_{rs}|^2 + \sigma^2 \sum_{r,s} |h_{rs}|^2 \right\},
\]

and the \( \| \cdot \|_2 \)-risk

\[
\text{Risk}_R[\hat{z}] = \max_{\|\hat{z}\|_2 \leq R} \mathbb{E} \left\{ \|\theta \cdot \hat{x} - h \cdot [\chi \cdot \hat{x} + \sigma \eta]\|_2 \right\}
\]

of this estimate does not exceed \( \sqrt{\text{Opt}} \).

Now goes your first task:

1.1) Prove that the above \( h \) induces the estimate

\[
\hat{w}(\omega) = F_{\alpha,\beta}^{-1} [h \cdot [F_{\alpha,\beta}\omega]]
\]

of \( x \ast \delta, x \in X_R = \{x \in \mathbb{R}^{m \times n} : \|x\|_2 \leq R\} \), via observation \( \omega = x \ast \kappa + \sigma \xi \), with risk

\[
\text{Risk}_R[\hat{w}] = \max_{x \in \mathbb{R}^{m \times n} : \|x\|_2 \leq R} \mathbb{E} \left\{ \|x \ast \delta - \hat{w}(x \ast \kappa + \sigma \xi)\|_2 \right\}
\]

not exceeding \( \sqrt{\text{Opt}} \). Note that \( x \) itself is nothing but a block in \( x \ast \delta \); observe also that in order for \( X_R \) to cover all images we are interested in, it suffices to take \( R = \sqrt{mn} \).

1.2) Prove that finding optimal solution to (4.91) is easy – the problem is in fact one-dimensional!

1.3) What are the sources, if any, of the conservatism of the estimate \( \hat{w} \) we have built as compared to the linear estimate given by an optimal solution to (4.12)?

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

\[
\|x \ast \kappa\|_2 \leq R_L,
\]

where \( \kappa_L \in \mathbb{R}^{(2\nu+1) \times (2\nu+1)} \), along with a priori upper bounds \( u_{rs} \) on the magnitudes of Fourier coefficients of \( x^+ \):

\[
|\left[F_{\alpha,\beta}x^+\right]_{rs}| \leq u_{rs}, \ 0 \leq r < \alpha, 0 \leq s < \beta.
\]

4.12.B. Mimicking Total Variation constraints. For an \( m \times n \) image \( x \in \mathbb{R}^{m \times n} \), its (anisotropic) total variation is defined as the \( \ell_1 \) norm of the “discrete gradient field” of \( x \):

\[
\text{TV}(x) = \sum_{p=0}^{m-1} \sum_{q=0}^{n-1} |x_{p+1,q} - x_{p,q}| + \sum_{p=0}^{m-1} \sum_{q=0}^{n-1} |x_{p,q+1} - x_{p,q}|.
\]

\[
\text{TV}_a(x) = \text{TV}(x) - \text{TV}_b(x)
\]
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. 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 speculate that bounds on $\text{TV}_{a,b}(x)$ can be somehow mimicked by bounds on the energy of two convolutions: one with kernel $\kappa_a \in \mathbb{R}^{(2\mu+1) \times (2\nu+1)}$ with the only nonzero entries $[\kappa_a]_{\mu,\nu} = -1, [\kappa_a]_{\mu+1,\nu} = 1$, and the other one with kernel $\kappa_b \in \mathbb{R}^{(2\mu+1) \times (2\nu+1)}$ with the only nonzero entries $[\kappa_b]_{\mu,\nu} = -1, [\kappa_b]_{\mu,\nu+1} = 1$(recall that the indices start from 0, and not from 1). Note that $x \ast \kappa_a$ and $x \ast \kappa_b$ are “discrete partial derivatives” of $x \ast \delta$.

For a small library of grayscale $m \times n$ images $x$ we dealt with, experiment shows that, in addition to the energy constraint $\|x\|_2 \leq R = \sqrt{mn}$, the images satisfy the constraints

$$\|x \ast \kappa_a\|_2 \leq \gamma_2 R, \|x \ast \kappa_b\|_2 \leq \gamma_2 R \quad (*)$$

with small $\gamma_2$, e.g., $\gamma_2 = 0.25$. In addition, it turns out that the $\infty$-norms of the Fourier transforms of $x \ast \kappa_a$ and $x \ast \kappa_b$ 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

$$\|\mathcal{F}_{\alpha\beta}[x \ast \kappa_a]\|_\infty \leq \gamma_\infty R, \quad \|\mathcal{F}_{\alpha\beta}[x \ast \kappa_b]\|_\infty \leq \gamma_\infty R, \quad \|\{z_{rs}\}_{r,s}\|_\infty = \max_{r,s} |z_{rs}| \quad (**)$$

with $\gamma_\infty = 0.01$.

Now, relations $(**)$ read

$$\max |\omega^a_{rs}, |\omega^b_{rs}| |\mathcal{F}_{\alpha\beta}(x^+)|_{rs} | \leq \gamma_\infty R \forall r, s$$

with easy-to-compute $\omega^a$, $\omega^b$, and in addition $|\mathcal{F}_{\alpha\beta}(x^+)|_{rs} \leq R$ due to $\|\mathcal{F}_{\alpha\beta}x^+\|_2 = \|x^+\|_2 \leq R$. We arrive at the bounds

$$|\mathcal{F}_{\alpha\beta}(x^+)|_{rs} \leq \min\left[1, 1/|\omega^a_{rs}|, 1/|\omega^b_{rs}| \right] R \forall r, s.$$

on the magnitudes of entries in $\mathcal{F}_{\alpha\beta}x^+$, and can utilize item 1.4 to incorporate these bounds, along with relations $(*)$.

Here is your next task:

2) Write software implementing the outlined deblurring and denoising image reconstruction routine and run numerical experiments.

---

Note that from $(*)$ it follows that $(**)$ holds with $\gamma_\infty = \gamma_2$, while with our empirical $\gamma$’s, $\gamma_\infty$ is 25 times smaller than $\gamma_2$. 

Recommended kernel $\kappa$: set $\mu = \lfloor m/32 \rfloor$, $\nu = \lfloor n/32 \rfloor$, start with
\[
\kappa_{uv} = \begin{cases} 
1 & \text{if } u = \mu, v = \nu \\
\Delta, & \text{if } 0 \leq u \leq 2\mu, 0 \leq v \leq 2\nu, \\
0, & \text{otherwise}
\end{cases}
\]
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 $\Delta$, the smaller is $\min_{r,s} |\chi_{rs}|$ (note that when decreasing the magnitudes of $\chi_{rs}$, we increase the optimal value in (4.91)).

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 (\ast) and (\ast\ast).

On the top of it, you can compare your results with what is given by \textit{"\ell_1-minimization recovery} described as follows:

As we remember from item 4.12.A, our problem of interest can be equivalently reformulated as recovering the image $\Theta z$ of a signal $z \in \mathbb{C}^{\alpha\beta}$ from noisy observation $\hat{\omega} = Az + \sigma \eta$, where $\Theta$ and $A$ are diagonal matrices, and $\eta$ is the zero mean complex Gaussian noise with unit covariance matrix. In other words, the entries $\eta_\ell$ in $\eta$ 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” $\epsilon$, say, $\epsilon = 0.1$, we can easily point out the smallest “confidence radius” $\rho$ such that for $\zeta \sim N(0, \frac{1}{2}I_2)$ it holds $\text{Prob}_{\epsilon\eta}\{\|\zeta\|_2 > \rho\} \leq \frac{\epsilon}{\alpha\beta}$, implying that for every $\ell$ it holds
\[
\text{Prob}_{\eta}\{|\hat{\omega}_\ell - A\hat{z}_\ell| > \sigma \rho\} \leq \frac{\epsilon}{\alpha\beta},
\]
and therefore
\[
\text{Prob}_{\eta}\{\|\hat{\omega} - Az\|_\infty > \sigma \rho\} \leq \epsilon.
\]

We now can easily find the smallest, in $\| \cdot \|_1$, vector $\hat{z} = \hat{z}(\omega)$ which is “compatible with our observation,” that is, satisfies the constraint
\[
\|\hat{\omega} - Az\|_\infty \leq \sigma \rho,
\]
and take $\Theta \hat{z}$ as the estimate of the “entity of interest” $\Theta z$ (cf. Regular \textit{\ell_1} recovery from Section 1.2.3).

Note that this recovery needs no a priori information on $z$.

\textbf{Exercise 4.13.}

\textit{[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 $\eta_i \sim N(0, 1)$ are independent across $i$ observation noises. Usually, a priori restrictions on $f$ are \textit{smoothness assumptions} – existence of $\kappa$ continuous derivatives satisfying a priori upper bounds
\[
\left( \int_0^1 |f^{(k)}(t)|^{p_k} \, dt \right)^{1/p_k} \leq L_k, 0 \leq k \leq \kappa,
\]
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, and the primary
emphasis of classical studies here is on how the minimax optimal (i.e., the best, over all estimates) risk goes to 0 as the number of observations \( n \) goes to infinity, what are near-optimal estimates, etc. Many of these studies deal with \textit{periodic case} – one where \( f \) can be extended onto the entire real axis as \( \kappa \) times continuously differentiable periodic, with period 1, 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 first established for the latter, usually extends 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 regression functions.

\textbf{4.13.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 \( \Gamma \) on the circumference. These restrictions form the usual \( n \)-dimensional coordinate space \( \mathbb{R}^n \); it is convenient to index the entries in \( f \in \mathbb{R}^n \) starting from 0 rather than from 1. We equip \( \mathbb{R}^n \) with two linear operators:

- \textit{Cyclic shift} (in the sequel – just \textit{shift}) \( \Delta \):

  \[
  \Delta \cdot [f_0; f_1; \ldots; f_{n-2}; f_{n-1}] = [f_{n-1}; f_0; f_1; \ldots; f_{n-2}],
  \]

  and

- \textit{Derivative} \( D \):

  \[
  D = n[I - \Delta];
  \]

  Treating \( f \in \mathbb{R}^n \) as a restriction of a function \( F \) on \( C \) onto \( \Gamma \), \( Df \) is the finite-difference version of the first order derivative of the function, and the norms

  \[
  |f|_p = n^{-1/p}\|f\|_p, \quad p \in [1, \infty],
  \]

  are the discrete versions of \( L_p \)-norms of \( F \).

Next, we associate with \( \chi \in \mathbb{R}^n \) the operator \( \sum_{i=0}^{n-1} \chi_i \Delta^i \); the image of \( f \in \mathbb{R}^n \) under this operator is denoted \( \chi \star f \) and is called (cyclic) \textit{convolution} of \( \chi \) and \( f \).

The problem we focus on is as follows:

Given are:

- \textit{smoothness data} represented by a nonnegative integer \( \kappa \) and two collections: \( \{L_\iota > 0 : 0 \leq \iota \leq \kappa\} \), \( \{p_\iota \in [2, \infty], 0 \leq \iota \leq \kappa\} \). The smoothness data specify the set

  \[
  \mathcal{F} = \{f \in \mathbb{R}^n : |f|_{p_\iota} \leq L_\iota, 0 \leq \iota \leq \kappa\}
  \]

  of signals we are interested in (this is the discrete analog of \textit{periodic Sobolev ball} – the set of \( \kappa \) times continuously differentiable functions on \( C \) with derivatives of orders up to \( \kappa \) bounded, in integral \( p_\iota \)-norms, by given quantities \( L_\iota \));

- two vectors \( \alpha \in \mathbb{R}^n \) (sensing kernel) and \( \beta \in \mathbb{R}^n \) (decoding kernel);

- positive integer \( \sigma \) (noise intensity) and a real \( q \in [1, 2] \).
These data define the estimation problem as follows: given noisy observation

\[ \omega = \alpha \ast f + \sigma \eta \]

of unknown signal \( f \) known to belong to \( \mathcal{F} \), where \( \eta \in \mathbb{R}^n \) is a random observation noise, we want to recover \( \beta \ast f \) in norm \( \| \cdot \|_q \).

Our only assumption on the noise is that

\[ \text{Var}[\eta] := \mathbb{E} \{ \eta \eta^T \} \preceq I_n. \]

The risk of a candidate estimate \( \hat{f} \) is defined as

\[ \text{Risk}_r[\hat{f}|\mathcal{F}] = \sup_{f \in \mathcal{F}, \eta : \text{Var}[\eta] \preceq I_n} \mathbb{E}_\eta \{ |\beta \ast f - \hat{f}(\alpha \ast f + \sigma \eta)|_q \}. \]

Here is the exercise:

1) Check that the situation in question fits the framework of Section 4.3.3 and figure out what, under the circumstances, reduces the optimization problem (4.42) responsible for the presumably good linear estimate \( \hat{f}_H(\omega) = H^T \omega \).

2) Prove that in the case in question the linear estimate yielded by an appropriate optimal solution to (4.42) is just the cyclic convolution

\[ \hat{f}(\omega) = h \ast \omega \]

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 – the problem where \( \alpha \ast x \equiv \beta \ast x \equiv x \) and \( \eta \sim \mathcal{N}(0, I_n) \), and compare numerically the computed risks of your estimates with the classical results on the limits of performance in recovering smooth univariate regression functions. According to those results, in the situation in question and under natural assumption that \( L_\iota \) are nondecreasing in \( \iota \), the minimax optimal risk, up to a factor depending solely on \( \kappa \), is

\[ (\sigma^2/n) \frac{\kappa^2}{\kappa+1} L_\kappa \frac{\kappa^2}{\kappa+1}. \]

4.7.4.4 Probabilities of large deviations in linear estimation under sub-Gaussian noise


The goal of Exercise is to derive bounds for probabilities of large deviations for estimates built in Proposition 4.14.

1) Prove the following fact:

**Lemma 4.30.** Let \( \Theta, Q \in S_+^n \), with \( Q \succ 0 \), and let \( \xi \) be sub-Gaussian, with parameters \( (\mu, S) \), random vector, where \( \mu \) and \( S \) satisfy \( \mu \mu^T + S \preceq Q \). Setting \( \rho = \text{Tr}(\Theta Q) \), we have

\[ \mathbb{E}_\xi \left\{ \exp \left( \frac{1}{2\rho} \xi^T \Theta \xi \right) \right\} \leq \sqrt{2} \exp \{ 1/4 \}. \]

(4.92)
As a result, for \( t > 0 \) it holds
\[
\Pr \left\{ \sqrt{\xi^T \Theta \xi} \geq t \sqrt{p} \right\} \leq \sqrt{2} \exp \{ \frac{1}{4} \} \exp \{ -t^2 / 8 \}, \quad t \geq 0.
\] (4.93)

**Hint:** Use the same trick as in the proof of Lemma 2.53.

2) Recall that (proof of) Proposition 4.14 states that in the situation of Section 4.3.3.1 and under Assumptions \( \mathbf{A}, \mathbf{B}, \mathbf{R} \), for every feasible solution \( (H, \Lambda, \Upsilon, \Theta') \) to the optimization problem\(^{16}\)
\[
\text{Opt} = \min_{H, \Lambda, \Upsilon, \Theta'} \left\{ \phi_T (\lambda[\Lambda]) + \phi_R (\lambda[\Upsilon]) : \right. \\
\left. \begin{array}{c}
\Lambda = \{ \Lambda_k \geq 0, k \leq K \}, \quad \Upsilon = \{ \Upsilon_k \geq 0, k \leq L \}, \\
\Upsilon' = \{ \Upsilon_k' \geq 0, k \leq L \}, \\
\frac{1}{2} M^1 |B - H^T A| \preceq 0, \\
\Theta \preceq 0, \\
\frac{1}{2} \Lambda^T |H M| \preceq 0 \\
\end{array} \right\},
\] (4.94)

one has
\[
\max_{x \in \mathcal{X}} \| (B - H^T A) x \| \leq A & \quad \max_{P : \text{Var}[P] \leq \Pi} \mathbb{E}_{\xi \sim P} \left\{ \| H^T \xi \| \right\} \leq B,
\] (4.95)

implying that the linear estimate \( \hat{x}_H (\omega) = H^T \omega \) satisfies the risk bound
\[
\text{Risk}_{H, \| \cdot \|} [\hat{x}_H (\cdot) | \mathcal{X}] \leq A + B.
\] (4.96)

Prove the following

**Proposition 4.31.** Let \( H, \Lambda, \Upsilon, \Theta' \) be a feasible solution to (4.94), and let \( \hat{x}_H (\omega) = H^T \omega \). Let, further, \( P \) be sub-Gaussian probability distribution on \( \mathbb{R}^m \), with parameters \( (\mu, S) \) satisfying
\[
\mu \mu^T + S \prec \Pi,
\]
and, finally, let \( x \in \mathcal{X} \). Then

(i) One has
\[
\mathbb{E}_{\xi \sim P} \left\{ \| B x - \hat{x}_H (A x + \xi) \| \right\} \leq A_* + B_*,
\]
\[
A_* = A_*(\Lambda, \Upsilon) := 2 \sqrt{\phi_T \left( \lambda[\Lambda] \right) \phi_R \left( \lambda[\Upsilon] \right)} \leq A(\Lambda, \Upsilon) := \phi_T (\lambda[\Lambda]) + \phi_R (\lambda[\Upsilon])
\]
\[
B_* = B_*(\Theta, \Upsilon') := 2 \sqrt{\Gamma_{\Pi}(\Theta) \phi_R (\lambda[\Upsilon'])} \leq B(\Theta, \Upsilon') := \Gamma_{\Pi}(\Theta) + \phi_R (\lambda[\Upsilon'])
\]

(ii) For every \( \epsilon \in (0, 1) \) one has
\[
\Pr_{\xi \sim P} \left\{ \xi : \| B x - \hat{x}_H (A x + \xi) \| > A_* + \theta_\epsilon B_* \right\} \leq \epsilon
\] (4.97)

where \( \theta_\epsilon = 2 \sqrt{2 \ln(\sqrt{2e}^1/4 / \epsilon)} \).

\(^{16}\)For notation, see Section 4.3.3.1, (4.36), and (4.39). For reader’s convenience, we recall part of this notation: for a probability distribution \( P \) on \( \mathbb{R}^m \), \( \text{Var}[P] = \mathbb{E}_{\xi \sim P} \left( \xi^T \xi \right) \). II is a convex compact subset of \( \text{int} \mathcal{S}^m_+ \), \( Q \prec II \) means that \( Q \leq Q' \) for some \( Q' \in II \), and \( \Gamma_{\Pi}(\Theta) = \max_{Q \in II} \text{Tr}(\Theta Q) \).
3) Suppose we are given observation \( \omega = Ax + \xi \) of unknown signal \( x \) known to belong to a given spectratope \( \mathcal{X} \subset \mathbb{R}^n \) and want to recover the signal. We quantify the error of a recovery \( \hat{x} \) by \( \max_{k \leq K} \| B_k (\hat{x} - x) \|_{(k)} \), where \( B_k \in \mathbb{R}^{\nu_k \times n} \) are given matrices, and \( \| \cdot \|_{(k)} \) are given norms on \( \mathbb{R}^{\nu_k} \) (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). We also assume, same as in item 2, that observation noise \( \xi \) is independent of signal \( x \) and is sub-Gaussian with sub-Gaussianity parameters \( \mu, S \) satisfying \( \mu \mu^T + S \preceq Q \), for some given matrix \( Q \succ 0 \). Finally, we suppose that the unit balls of the norms conjugate to the norms \( \| \cdot \|_{(k)} \) are spectratopes. In this situation, Proposition 4.14 provides us with \( K \) efficiently computable linear estimates \( \hat{x}_k(\omega) = H^T_k \omega : \mathbb{R}^{\dim \omega} \rightarrow \mathbb{R}^{\nu_k} \) along with upper bounds \( \text{Opt}_k \) on their risks \( \max_{x \in \mathcal{X}} \mathbb{E} \{ \| B_k x - \hat{x}_k (Ax + \xi) \|_{(k)} \} \).

Think how, given reliability tolerance \( \epsilon \in (0, 1) \), aggregate these linear estimates into a single estimate \( \hat{x}(\omega) : \mathbb{R}^{\dim \omega} \rightarrow \mathbb{R}^n \) such that for every \( x \in \mathcal{X} \), the probability of the event

\[
\| B_k (\hat{x}(Ax + \xi) - x) \|_{(k)} \leq \theta \text{Opt}_k, 1 \leq k \leq K,
\]

is at least \( 1 - \epsilon \), for some moderate (namely, logarithmic in \( K \) and \( 1/\epsilon \)) “assembling price” \( \theta \).

Exercise 4.15.
Prove that if \( \xi \) is uniformly distributed on the unit sphere \( \{ x : \| x \|_2 = 1 \} \) in \( \mathbb{R}^n \), then \( \xi \) is sub-Gaussian with parameters \( (0, \frac{1}{n} I_n) \).

4.7.4.5 Linear recovery under signal-dependent noise

Exercise 4.16.
Consider the situation as follows: we observe a realization \( \omega \) of \( m \)-dimensional random vector

\[ \omega = Ax + \xi_x, \]

where

- \( x \) is unknown signal belonging to a given signal set \( \mathcal{X} \), specifically, spectratope (which, as usual, we can assume to be basic):
  \[ \mathcal{X} = \{ x \in \mathbb{R}^n : \exists t \in \mathcal{T} : R_k^2[x] \preceq t_k I_{d_k}, k \leq K \} \]
  with standard restrictions on \( \mathcal{T} \) and \( R_k[.] \);
- \( \xi_x \) is the observation noise with distribution which can depend on \( x \); all we know is that
  \[ \text{Var}[\xi_x] := \mathbb{E} \{ \xi_x \xi_x^T \} \preceq C[x], \]
  where the entries of symmetric matrix \( C[x] \) are quadratic in \( x \). We assume in the sequel that signals \( x \) belong to the subset
  \[ \mathcal{X}_C = \{ x \in \mathcal{X} : C[x] \succeq 0 \} \]
  of \( \mathcal{X} \);
- Our goal is to recover \( Bx \), with given \( B \in \mathbb{R}^{\nu \times n} \), in a given norm \( \| \cdot \| \) such that
the unit ball $B_*$ of the conjugate norm is a spectrahope:

$$B_* = \{ u : \|u\|_* \leq 1 \} = \mathcal{M} \mathcal{V}, \mathcal{V} = \{ v : \exists r \in \mathcal{R} : S^2_I[v] \preceq r_\ell I_{f_\ell}, \ell \leq L \}.$$ 

We quantify the performance of a candidate estimate $\hat{x}(\omega) : \mathbb{R}^m \to \mathbb{R}^n$ by the risk

$$\text{Risk}_{\|\cdot\|}(|\hat{x}|,\mathcal{A}_C) = \sup_{x \in \mathcal{A}_C} \sup_{\xi : \text{Var}[\xi] \subseteq \mathcal{C}[x]} \mathbb{E} \{ \| Bx - \hat{x}(Ax + \xi_x) \| \}.$$ 

1) Utilize semidefinite relaxation in order to build, in a computationally efficient fashion, a “presumably good” linear estimate, specifically, prove the following

**Proposition 4.32.** In the situation in question, for $G \in \mathbb{S}^m$ let us define $\alpha_0[G] \in \mathbb{R}, \alpha_1[G] \in \mathbb{R}^n, \alpha_2[G] \in \mathbb{S}^n$ from the identity

$$\text{Tr}(\mathcal{C}[x]G) = \alpha_0[G] + \alpha_1^T[G]x + x^T \alpha_2[G]x \forall (x \in \mathbb{R}^n, G \in \mathbb{S}^m),$$

so that $\alpha_x[G]$ are affine in $G$. Consider convex optimization problem

$$\text{Opt} = \min_{H, \mu, D, \Lambda, \Upsilon, G} \left\{ \mu + \phi_T(\lambda[\Lambda]) + \phi_R(\lambda[\Upsilon]) + \phi_R(\lambda[\Upsilon']) : \Lambda = \{ \Lambda_k \in \mathbb{S}^k, k \leq K \}, \Upsilon = \{ \Upsilon_\ell \in \mathbb{S}^\ell, \ell \leq L \}, \Upsilon' = \{ \Upsilon'_\ell \in \mathbb{S}^{\ell'}, \ell \leq L \}, D \in \mathbb{S}^n \right\}$$

$$= \min \left[ \begin{array}{cccc}
\alpha_0[G] & \frac{1}{2} \alpha_1^T[G] & \frac{1}{2} \alpha_2 \left[ B^T - A^T H M \right] \\
\frac{1}{2} \alpha_1^T[G] & \alpha_2^T & -\frac{1}{2} \alpha_1^T[D] \\
\frac{1}{2} \alpha_2 \left[ B^T - A^T H M \right] & -\frac{1}{2} \alpha_1^T[D] & \sum_k \mathcal{R}_k \left[ \Lambda_k \right] - \frac{1}{2} \alpha_2^T[D] \\
\end{array} \right] \geq 0$$

$$= \min \left[ \begin{array}{ccc}
\left[ \mathcal{R}_k \left[ \Lambda_k \right] \right] & \text{Tr}(\Lambda_k \frac{1}{2} [R^{k_i} R^{k_j} + R^{k_j} R^{k_i}]), \text{ where } R_k \in \mathbb{S}^n \\
\left[ S^2_I[\Upsilon_\ell] \right] & \text{Tr}(\Upsilon_\ell \frac{1}{2} [S^{f_i} S^{f_j} + S^{f_j} S^{f_i}] \text{, where } S^2_I[v] = \sum_{j, k} \mathcal{S}^{f_j s_k} \\
\text{Tr}(\lambda[\mathcal{Z}_i, i \leq I]) & = \max_{x \in A} q^T s \\
\end{array} \right] \right.$$ 

Whenever $H, \mu, D, \Lambda, \Upsilon, \Upsilon'$ and $G$ are feasible for the problem, one has

$$\text{Risk}_{\|\cdot\|}(|\hat{x}|,\mathcal{A}_C) \leq \mu + \phi_T(\lambda[\Upsilon]) + \phi_R(\lambda[\Upsilon]) + \phi_R(\lambda[\Upsilon'])$$

where $\hat{x}(\omega) = H^T \omega$.

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 $\mathcal{Y} = \{ y \in \mathbb{R}^n : a \leq y \leq \sigma \}$, where $0 \leq a < \sigma$. You want to recover $y$ in $\| \cdot \|_p$-norm with given $p \in [1, 2]$.

**4.7.5** Signal recovery in Discrete and Poisson observation schemes

**Exercise 4.17.**
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 $\omega^K = (\omega_1, \ldots, \omega_K)$ with independent identically distributed components $\omega_k$ taking values in $m$-element set. As always, we encode the points from this $m$-element set by the standard basic orths $e_1, \ldots, e_m$ in $\mathbb{R}^m$.
- The (common for all $k$) probability distribution of $\omega_k$ is $Ax$, where $x$ is unknown “signal” – $n$-dimensional probabilistic vector known to belong to a closed convex subset $X$ of $n$-dimensional probabilistic simplex $\Delta_n = \{x \in \mathbb{R}^n : x \geq 0, \sum_i x_i = 1\}$, 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 $\hat{x}(\omega^K) : \mathbb{R}^{mK} \rightarrow \mathbb{R}^\nu$ by its risk

$$\text{Risk}_{\|\cdot\|}[\hat{x} | X] = \sup_{x \in X} E_{\omega^K \sim [Ax] \times \ldots \times [Ax]} \{\|Bx - \hat{x}(\omega^K)\|\},$$

where $\|\cdot\|$ is a given norm on $\mathbb{R}^\nu$.

We use **linear estimates** – estimates of the form

$$\hat{x}_H(\omega^K) = H^T \left[ \frac{1}{K} \sum_{k=1}^{K} \omega_k \right], \quad (4.98)$$

where $H \in \mathbb{R}^{m \times \nu}$.

1) In the main body of Chapter 4, $X$ 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 where $X$ can be “heavily asymmetric,” which, in general, can make “genuinely affine” estimates preferable. Show that in the case in question, we still lose nothing when restricting ourselves ro linear, rather than affine, estimates.

**4.17.A. Observation scheme revisited.** When observation $\omega^K$ stems from a signal $x \in \Delta_n$, we have

$$\bar{\omega}_K[\omega^K] = Ax + \xi_x,$$

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

$$Q[x] = \text{Diag}\{Ax\} - [Ax][Ax]^T,$$
and derive from this fact that the covariance matrix of $\xi$ is

$$Q_K[x] = \frac{1}{K}Q[x].$$

Setting

$$\Pi = \Pi_X = \left\{ Q = \frac{1}{K}\text{Diag}\{Ax\} : x \in \mathcal{X} \right\},$$

check that $\Pi_X$ 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 $Q \in \Pi$.

4.17.B. Upper-bounding risk of a linear estimate. We can upper-bound the risk of a linear estimate $\hat{x}_H$ as follows:

$$\text{Risk}_{\|\cdot\|}[\hat{x}_H] = \sup_{x \in \mathcal{X}} \mathbb{E}_{\xi \sim \mathcal{X}} \left\{ \left\| Bx - H^T\hat{\omega}_K[\omega^K] \right\| \right\}
= \sup_{x \in \mathcal{X}} \mathbb{E}_{\xi \sim \mathcal{X}} \left\{ \left\| [Bx - H^T A]x - H^T \xi x \right\| \right\}
\leq \sup_{x \in \mathcal{X}} \left\{ \left\{ B - H^T A \right\|x \} + \sup_{\xi : \text{Cov}\{\xi\} \in \Pi_X} \mathbb{E}_{\xi} \left\{ \left\| H^T \xi \right\| \right\} \right\}.
$$

As in the main body of Chapter 4, we intend to build a “presumably good” linear estimate by minimizing over $H$ the sum of efficiently computable upper bounds $\Phi(H)$ on $\Phi(H)$ and $\Psi^X(H)$ on $\Psi^X(H)$.

Assuming from now on that the unit ball $\mathcal{B}_\ast$ of the norm conjugate to $\|\cdot\|$ is a spectratope:

$$\mathcal{B}_\ast := \{ u : \|u\|_\ast \leq 1 \} = \{ u : \exists r \in \mathcal{R}, y : u = My, S^2_r[y] \preceq r_\ell I, \ell \leq L \}$$
with our usual restrictions on $\mathcal{R}$ and $S_\ell$, we can take as $\mathcal{B}^X(\cdot)$ the function (4.40).

For the sake of simplicity, we from now assume that $\mathcal{X}$ is cut off $\Delta_n$ by linear inequalities:

$$\mathcal{X} = \{ x \in \Delta_n : Gx \leq g, Ex = e \}
\quad [G \in \mathbb{R}^{p \times n}, E \in \mathbb{R}^{q \times n}]$$

Observe that replacing $G$ with $G - g1_n^T$ and $E$ with $E - e1_n^T$, we reduce the situation to that where all linear constraints are homogeneous, that is,

$$\mathcal{X} = \{ x \in \Delta_n : Gx \leq 0, Ex = 0 \},$$

and this is what we assume from now on. Setting

$$F = [G; E; -E] \in \mathbb{R}^{(p+2q) \times n},$$
we have also

$$\mathcal{X} = \{ x \in \Delta_n :Fx \leq 0 \}.$$

Suppose that $\mathcal{X}$ is nonempty. Finally, in addition to what was already assumed about the norm $\|\cdot\|$, let us also suppose 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?).

Our next task is to efficiently upper-bound $\Phi(\cdot)$. 
4.17.C. Bounding $\Phi$, simple case. We start with the simple case where there are no linear constraints (formally, $G$ and $E$ are zero matrices), in this case bounding $\Phi$ is straightforward:

3) Prove that in the simple case $\Phi$ 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 $\mathbb{R}^n$.

4.17.D. Lagrange upper bound on $\Phi$.

4) Observing that when $\mu \in \mathbb{R}^{p+2q}$, the function

$$
\| (B - H^T A)x \| - \mu^T Fx
$$

of $x$ is convex in $x \in \Delta_n$, and overestimates $\| (B - H^T A)x \|$ everywhere on $X$, conclude that the efficiently computable convex function

$$
\Phi_L(H) = \min_{\mu} \max_{i \leq n} \{ \| (B - H^T A)g_i \| - \mu^T Fg_i : \mu \geq 0 \}
$$

upper-bounds $\Phi(H)$. In the sequel, we call this function Lagrange upper bound on $\Phi$.

4.17.E. Basic upper bound on $\Phi$. For vectors $u$ and $v$ of the same dimension, say, $k$, let $\text{Max}[u, v]$ stand for the entrywise maximum of $u, v$:

$$
[\text{Max}[u, v]]_i = \max[u_i, v_i],
$$

and let

$$
[u]_+ = \text{Max}[u, 0_k],
$$

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 $\Lambda$ is entrywise nonnegative. Prove that whenever $x \in X$, one has

$$
\| (B - H^T A)x \| \leq B(x, H, \Lambda_+, \Lambda_-)
:= \min_t \{ \| t \| : t \geq \text{Max} \{ [(B - H^T A)x - \Lambda_+ Fx]_+, [- (B - H^T A)x - \Lambda_- Fx]_+ \} \}
$$

and that $B(x, H, \Lambda_+, \Lambda_-)$ is convex in $x$.

5.2) Derive from 5.1 that whenever $\Lambda_\pm$ are as in 5.1, one has

$$
\Phi(H) \leq B^+(H, \Lambda_+, \Lambda_-) := \max_{i \leq n} B(g_i, H, \Lambda_+, \Lambda_-),
$$

where, as in item 3, $g_1, ..., g_n$ are the standard basic orths in $\mathbb{R}^n$. Conclude that

$$
\Phi(H) \leq \Phi_B(H) = \inf_{\Lambda_\pm} \left\{ B^+(H, \Lambda_+, \Lambda_-) : \Lambda_\pm \in \mathbb{R}^{\nu \times (p+2q)}_+ \right\}
$$

and that $\Phi_B$ is convex and real-valued. In the sequel we refer to $\Phi_B(\cdot)$ as to Basic upper bound on $\Phi(\cdot)$. 
4.17.F. Sherali-Adams upper bound on $\Phi$. Let us apply the approach we used in Chapter 1, Section 1.3.2, when deriving verifiable sufficient conditions for $s$-goodness, see p. 21. Specifically, setting

$$W = \begin{bmatrix} G & I \end{bmatrix},$$

let us introduce the slack variable $z \in \mathbb{R}^p$ and rewrite the description of $X$ as

$$X = \{ x \in \Delta_n : \exists z \geq 0 : W[x; z] = 0 \},$$

so that $X$ is the projection of the polyhedral set $X^+ = \{ [x; z] : x \in \Delta_n, z \geq 0, W[x; z] = 0 \}$ on the $x$-space. Projection of $X^+$ on the $z$-space is a nonempty (since $X$ is so) and clearly bounded subset of the nonnegative orthant $\mathbb{R}^p_+$, and we can in many ways cover $Z$ by the simplex

$$\Delta[\alpha] = \{ z \in \mathbb{R}^p : z \geq 0, \sum_i \alpha_i z_i \leq 1 \},$$

where all $\alpha_i$ are positive.

6.1) Let $\alpha > 0$ be such that $Z \subset \Delta[\alpha]$. Prove that

$$X^+ = \{ [x; z] : W[x; z] = 0, [x; z] \in \text{Conv}\{v_{ij} = [g_i; h_j], 1 \leq i \leq n, 0 \leq j \leq p\} \},$$

where $g_i$ are the standard basic orhts in $\mathbb{R}^n$, $h_0 = 0 \in \mathbb{R}^p$, and $\alpha_j h_j, 1 \leq j \leq p$, are the standard basic orhts in $\mathbb{R}^p$.

6.2) Derive from 5.1 that the efficiently computable convex function

$$\Phi_{SA}(H) = \inf_C \max_{i,j} \left\{ \| (B - H^T A)g_i + C^T Wv_{ij} \| : C \in \mathbb{R}^{(p+q)\times \nu} \right\},$$

is an upper bound on $\Phi(H)$. In the sequel, we refer to $\Phi_{SA}(H)$ as to Sherali-Adams bound [211].

4.17.G. Combined bound. We can combine the above bounds, specifically, as follows:

7) Prove that the efficiently computable convex function

$$\Phi_{LBS}(H) = \inf_{(\Lambda_\pm, C_\pm, \mu, \mu_+)} \max_{i,j} G_{ij}(H, \Lambda_\pm, C_\pm, \mu, \mu_+),$$

where

$$G_{ij}(H, \Lambda_\pm, C_\pm, \mu, \mu_+) := -\mu^T Fg_i + \mu_+^T Wv_{ij} + \min_t \left\{ \| t \| : t \geq \text{Max} \left\{ \| (B - H^T A - \Lambda_\pm F)g_i + C_\pm^T Wv_{ij} \|, \| (-B + H^T A - \Lambda_\pm F)g_i + C_\pm^T Wv_{ij} \| \right\} \right\},$$

$$R = \{ (\Lambda_\pm, C_\pm, \mu, \mu_+) : \Lambda_\pm \in \mathbb{R}^{p+x(p+2q)}, C_\pm \in \mathbb{R}^{(p+q)\times \nu}, \mu \in \mathbb{R}^{p+2q}, \mu_+ \in \mathbb{R}^{p+q} \}$$

is an upper bound on $\Phi(H)$, and that this Combined bound is at least as good as any of Lagrange, Basic, or Sherali-Adams bounds.
**4.17. How to select $\alpha$?** A shortcoming of the Sherali-Adams and the combined upper bounds on $\Phi$ is the presence of a “degree of freedom” – the positive vector $\alpha$. Intuitively, we would like to select $\alpha$ to make the simplex $\Delta[\alpha] \supset Z$ to be “as small as possible.” It is unclear, however, what “as small as possible” is in our context, not to speak of how to select the required $\alpha$ after we agree how we measure the “size” of $\Delta[\alpha]$. It turns out, however, that we can select efficiently $\alpha$ resulting in the smallest volume $\Delta[\alpha]$.

8) Prove that minimizing the volume of $\Delta[\alpha] \supset Z$ in $\alpha$ reduces to solving the following convex optimization problem:

$$\inf_{\alpha,u,v} \left\{ -\sum_{s=1}^{p} \ln(\alpha_s) : 0 \leq \alpha \leq -v, E^T u + G^T v \leq 1 \right\} \quad (*)$$

9) Run numerical experiments to evaluate the quality of the above bounds. It makes sense to generate problems where we know in advance the actual value of $\Phi$, e.g., to take $\mathcal{X} = \{ x \in \Delta_n : x \geq a \}$ (a) with $a \geq 0$ such that $\sum_i a_i \leq 1$. 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

$$\text{Opt} = \min_{H,Y,\Theta} \left\{ \overline{\Phi}(H) + \phi_\mathcal{X}(\lambda[Y]) + \Gamma_\mathcal{X}(\Theta) : Y = \{ Y_\ell \geq 0, \ell \leq L \}, \Theta \right\} \quad (4.99)$$

where

$$\Gamma_\mathcal{X}(\Theta) = \frac{1}{K} \max_{x \in \mathcal{X}} \text{Tr}(\text{Diag}(Ax)\Theta),$$

see Corollary 4.12, with the actual $\Phi$ (which is available for our $\mathcal{X}$), or the upper bounds on $\Phi$ (Lagrange, Basic, Sherali-Adams, and Combined) in the role of $\overline{\Phi}$. Note that it may make sense to test 7 bounds rather than just 4. Indeed, with additional constraints on the optimization variables in (a), 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 (a) the constraints $A_\pm = 0$.

**Exercise 4.18.**

The exercise to follow deals with recovering discrete probability distributions in Wasserstein norm.

**Wasserstein distance** between probability distributions is extremely popular in today Statistics; it is defined as follows.\(^{17}\) Consider discrete random variables taking values in finite observation space $\Omega = \{ 1, 2, \ldots, n \}$ which is equipped with the metric

\(^{17}\)The distance we consider stems from the Wasserstein 1-distance between discrete probability distributions. This is a particular case of the general Wasserstein $p$-distance between (not necessarily discrete) probability distributions.
1) Let \( p, q \) be two probability distributions. Prove that

\[
W(p, q) = \min \left\{ \sum_i d_{ij} x_{ij} : x_{ij} \geq 0, \sum_j x_{ij} = p_i, \sum_i x_{ij} = q_j \forall 1 \leq i, j \leq n \right\}
\]

(4.100)

In other words, one may think of \( p \) and \( q \) as of two distributions of unit mass on the points of \( \Omega \), and consider the mass transport problem of redistributing the mass assigned to points by distribution \( p \) to get the distribution \( q \). Denoting by \( x_{ij} \) the mass moved from point \( i \) to point \( j \), constraints \( \sum_j x_{ij} = p_i \) say that the total mass taken from point \( i \) is exactly \( p_i \), constraints \( \sum_i p_{ij} = q_j \) say that as the result of transportation, the mass at point \( j \) will be exactly \( q_j \), and the constraints \( x_{ij} \geq 0 \) reflect the fact that transport of a negative mass is forbidden. Assuming that the cost of transporting a mass \( \mu \) from point \( i \) to point \( j \) is \( d_{ij} \mu \), the Wasserstein distance \( W(p, q) \) between \( p \) and \( q \) is the cheapest transportation plan which converts \( p \) into \( q \).

As compared to other natural distances between discrete probability distributions, like \( \|p - q\|_1 \), the advantage of the Wasserstein distance is that it allows to model the situation (indeed arising in some applications) where the effect, measured in terms of intended application, of changing probability masses of points from \( \Omega \) is small when the probability mass of a point is redistributed among close points.\(^{19}\)

Now goes the first part of the exercise:

1) Let \( p, q \) be two probability distributions. Prove that

\[
W(p, q) = \max_{f \in \mathbb{R}^n} \left\{ \sum_i f_i (p_i - q_i) : |f_i - f_j| \leq d_{ij} \forall i, j \right\}
\]

(4.101)

Treating vector \( f \in \mathbb{R}^n \) as a function on \( \Omega \), the value of the function at a point \( i \in \Omega \) being \( f_i \), (4.101) admits a very transparent interpretation: Wasserstein distance \( W(p, q) \) between probability distributions \( p, q \) is the maximum of inner products of \( p - q \) and Lipschitz continuous, with constant 1 w.r.t. the metric \( d \), functions \( f \) on \( \Omega \). When shifting \( f \) by a constant, the inner product remains intact (since \( p - q \) is a vector with zero sum of entries). Therefore, denoting by

\[
D = \max_{i,j} d_{ij}
\]

the \( d \)-diameter of \( \Omega \), we have

\[
W(p, q) = \max_f \left\{ f^T (p - q) : |f_i - f_j| \leq d_{ij}, |f_i| \leq D/2 \forall i, j \right\},
\]

(4.102)

the reason being that every Lipschitz continuous, with constant 1 w.r.t. metric \( d \), function \( f \) on \( \Omega \) can be shifted by a constant to ensure \( \|f\|_\infty \leq D/2 \) (look what happens when the shift ensures that \( \min_i f_i = -D/2 \)).

\(^{18}\)namely, positivity: \( d_{ij} = d_{ji} \geq 0 \), with \( d_{ii} = 0 \) if and only if \( i = j \), and the triangle inequality: \( d_{ik} \leq d_{ij} + d_{jk} \) for all triples \( i, j, k \).

\(^{19}\)In fact, Wasserstein distance share this property with some other distances between distributions used in Probability Theory, such as Skorohod, or Prokhorov, or Ky Fan distances. What makes Wasserstein distance so “special” is its representation (4.100) as the optimal value of a Linear Programming problem, responsible for efficient computational handling of this distance.
Representation (4.102) shows that the Wasserstein distance is generated by a norm on $\mathbb{R}^n$: for all probability distributions on $\Omega$ one has

$$W(p, q) = ||p - q||_W,$$

where $|| \cdot ||_W$ is the Wasserstein norm on $\mathbb{R}^n$ given by

$$||x||_W = \max_{f \in B_*} f^T x,$$

$$B_* = \{ u \in \mathbb{R}^n : u^T S_{ij} u \leq 1, 1 \leq i \leq j \leq n \},$$

$$S_{ij} = \begin{cases} d_{ij}^{-2}[e_i - e_j][e_i - e_j]^T, & 1 \leq i < j \leq n, \\ 4D^{-2}e_i e_i^T, & 1 \leq i = j \leq n, \end{cases}$$

(4.103)

where $e_1, ..., e_n$ are the standard basic orths in $\mathbb{R}^n$.

2) Let us equip $n$-element set $\Omega = \{1, ..., d\}$ with the metric $d_{ij} = \begin{cases} 2, & i \neq j \\ 0, & i = j \end{cases}$. What is the associated Wasserstein norm?

Note that the set $B_*$ in (4.103) is the unit ball of the norm conjugate to $|| \cdot ||_W$, and as we see, this set is a basic ellitope. As a result, the estimation machinery developed in Chapter 4 is well-suited for recovering discrete probability distributions in Wasserstein norm. This observation motivates the concluding part of the exercise:

3) Consider the situation as follows: Given $m \times n$ column-stochastic matrix $A$ and $\nu \times n$ column-stochastic matrix $B$, we observe $K$ independent of each other samples $\omega_k, 1 \leq k \leq K$, drawn from discrete probability distribution $Ax \in \Delta_m$ (as always, $\Delta_\nu \subset \mathbb{R}^\nu$ is the probabilistic simplex in $\mathbb{R}^\nu$), $x \in \Delta_n$ being unknown "signal" underlying observations; realizations of $\omega_k$ are identified with respective vertices $f_1, ..., f_m$ of $\Delta_m$. Our goal is to use the observations to estimate the distribution $Bx \in \Delta_\nu$. We are given a metric $d$ on the set $\Omega_\nu = \{1, 2, ..., \nu\}$ of indices of entries in $Bx$, and measure the recovery error in the associated with $d$ Wasserstein norm $|| \cdot ||_W$.

Build explicit convex optimization problem responsible for "presumably good" linear recovery of the form

$$\tilde{x}_H = \frac{1}{K} H^T \sum_{k=1}^{K} \omega_k.$$

Exercise 4.19.

[follow-up to Exercise 4.17] In Exercise 4.17, we have built a "presumably good" linear estimate $\tilde{x}_H(\cdot)$, see (4.98), yielded by the $H$-component $H_*$ of an optimal solution to problem (4.99). The optimal value Opt in this problem is an upper bound on the risk $\text{Risk}_{|| \cdot ||}[\tilde{x}_H, \mathcal{X}]$ (here and in what follows we use the same notation and impose the same assumptions as in Exercise 4.17). Recall that $\text{Risk}_{|| \cdot ||}$ 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 consider here, cf. Exercise 4.14.

1) Prove the following
**Lemma 4.33.** Let $Q \in S^m_m$, let $K$ be a positive integer, and let $p \in \Delta_m$. Let, further, $\omega^K = (\omega_1, ..., \omega_K)$ be i.i.d. random vectors, with $\omega_k$ taking the value $e_j$ ($e_1, ..., e_m$ are the standard basic orths in $R^m$) with probability $p_j$. Finally, let $\xi_k = \omega_k - E\{\omega_k\} = \omega_k - p$, and $\hat{\xi} = \frac{1}{K} \sum_{k=1}^{K} \xi_k$. Then for every $\epsilon \in (0, 1)$ it holds
\[
\text{Prob}\left\{ \left\| \hat{\xi} \right\|_2^2 \leq \frac{12 \ln(2m/\epsilon)}{K} \right\} \geq 1 - \epsilon.
\]
Hint: use the classical Bernstein inequality:

**Bernstein inequality:** Let $X_1, ..., X_K$ be independent zero mean random variables taking values in $[-M, M]$, and let $\sigma_k^2 = E\{X_k^2\}$. Then for every $t \geq 0$ one has
\[
\text{Prob}\left\{ \sum_{k=1}^{K} X_k \geq t \right\} \leq \exp \left\{ -\frac{t^2}{2 \left( \sum_k \sigma_k^2 + \frac{1}{3} Mt \right) } \right\}.
\]

2) Consider the situation described in Exercise 4.17 with $X = \Delta_n$, specifically,

- Our observation is a sample $\omega^K = (\omega_1, ..., \omega_K)$ with i.i.d. components $\omega_k \sim A x$, where $X \in \Delta_n$ is unknown $n$-dimensional probabilistic vector, $A$ is an $m \times n$ stochastic matrix (nonnegative matrix with unit column sums), and $\omega \sim A x$ means that $\omega$ is random vector taking value $e_i$ ($e_i$ are standard basic orths in $R^m$) with probability $[A x]_i$, $1 \leq i \leq m$;
- Our goal is to recover $B x$ in a given norm $\| \cdot \|$; here $B$ is a given $\nu \times n$ matrix.
- We assume that the unit ball $B_*$ of the norm $\| \cdot \|_*$ conjugate to $\| \cdot \|$ is a spectratope:
\[
B_* = \{ u = My, y \in Y \}, \quad Y = \{ y \in R^N : \exists r \in R : S^2_f[y] \preceq r I_{f_1}, \ell \leq L \}.
\]

Our goal is to build a presumably good linear estimate
\[
\hat{x}_H(\omega^K) = H^T \hat{\omega}[\omega^K], \quad \hat{\omega}[\omega^K] = \frac{1}{K} \sum_k \omega_k.
\]

Prove the following

**Proposition 4.34.** Let $H, \Theta, \Upsilon$ be a feasible solution to the convex optimization problem
\[
\min_{H, \Theta, \Upsilon} \left\{ \Phi(H) + \phi_{\mathcal{R}}(\lambda | \Upsilon) \right\}/K : \Upsilon = \{ \Upsilon_\ell \succeq 0, \ell \leq L \} \left[ \frac{\Theta}{H^T M} \right] \left( \frac{1}{2} M^T H^T \right) \succeq 0 \} \text{ (4.104)}
\]

where
\[
\Phi(H) = \max_{j \leq n} \| \text{Col}_j [B - H^T A] \|, \quad \Gamma(\Theta) = \max_{x \in \Delta_n} \text{Tr} (\text{Diag} \{ Ax \} \Theta).
\]

Then
3) Look what happens when \( \nu = \kappa \), seller draws the \( \kappa \)-th experiment, is the item, if any, bought by customer, and we make statistical inferences from these observations. As a result, our observation scheme can be formalized as follows. Let \( \Omega, B, P, \) and \( n \) be the set of all subsets of the \( n \)-element set, so that \( \Omega \) is of cardinality \( N = 2^n \). The population of customers induces a probability distribution \( p \) on \( \Omega \): for \( \omega \in \Omega \), \( p_{\omega} \) is the fraction of customers with the preference set being \( \omega \); we refer to \( p \) as to the preference distribution. An outcome of a single experiment can be represented by a pair \( (\omega, B) \), where \( B \in S_{m,n} \) is the offer used in the experiment, and \( \omega \) 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 \( \Omega = \{ (\omega, B) \} \) 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 following observation scheme:

\[
\forall (x \in \Delta_n) :
\begin{align*}
E \{ \| x - \hat{x}_1(\omega^K) \|_p \} & \leq n^{\frac{1}{2} - \frac{1}{2}} K^{-\frac{1}{2}} \Phi(\lambda[\hat{Y}]) || x ||_{S_{kn,\infty}} \\
\text{Prob} \{ \| x - \hat{x}_1(\omega^K) \|_p \leq 2\sqrt{3 \ln(2n/\epsilon)} n^{\frac{1}{2} - \frac{1}{2}} K^{-\frac{1}{2}} \} & \geq 1 - \epsilon
\end{align*}
\]

Exercise 4.20.

[follow-up to Exercise 4.17] 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 \( S = \{ 1, \ldots, n \} \) of the items. A customer has private preferences represented by a subset \( P \) of \( \mathcal{S} - \) 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 making offers to \( K \) customers. Specifically, the seller draws the \( k \)-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 \( S_{m,n} \) of all \( m \)-item offers. What is observed in the \( k \)-th experiment, is the item, if any, bought by customer, and we want to make statistical inferences from these observations.

The outlined observation scheme can be formalized as follows. Let \( S \) be the set of all subsets of the \( n \)-element set, so that \( S \) is of cardinality \( N = 2^n \). The population of customers induces a probability distribution \( p \) on \( S \): for \( \omega \in S \), \( p_{\omega} \) is the fraction of customers with the preference set being \( \omega \); we refer to \( p \) as to the preference distribution. An outcome of a single experiment can be represented by a pair \( (\omega, B) \), where \( B \in S_{m,n} \) is the offer used in the experiment, and \( \omega \) 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 \( \Omega = \{ (\omega, B) \} \) 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 following observation scheme:

\[
\begin{align*}
\mathbb{E}_{\omega,K} \left\{ \| Bx - \hat{x}_H(\omega^K) \| \right\} & \leq \Phi(H) + 2K^{-1/2} \sqrt{\phi_R(\lambda[\hat{Y}]} \Gamma(\Theta)
\leq \Phi(H) + \phi_R(\lambda[\hat{Y}]) + \Phi(H) + \Gamma(\Theta)/K
\end{align*}
\]
columns $A_p$ indexed by $P \in S$. When a customer is drawn at random from the uniform distribution on the population of customers, the distribution of the outcome clearly is $A_p$, where $p$ is the (unknown) preference distribution. Our inferences should be based on $K$-element sample $\omega^K = (\omega_1, \ldots, \omega_K)$, with $\omega_1, \ldots, \omega_K$ drawn, independently of each other, from the distribution $A_p$.

Now we can pose various inference problems, e.g., that of estimating $p$. We, however, intend to focus on a simpler problem – one of recovering $A_p$. In terms of our story, this makes sense: when we know $A_p$, 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 $S_{m,n}$) to something more rewarding.

Now goes the exercise:

1. Use the results of Exercise 4.17 to build “presumably good” linear estimate

$$\hat{x}_H(\omega^K) = H^T \left[ \frac{1}{K} \sum_{k=1}^{K} \omega_k \right]$$

of $A_p$ (as always, we encode observations $\omega$, which are elements of $M$-element set $\Omega$, by standard basic orths in $\mathbb{R}^M$). 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 $\omega \sim A_p$).

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 4.21.**

[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 $\omega \in \mathbb{R}_+^m$ 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 $\mathbb{R}_+^n$. 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 $\Delta_n$ by a collection of linear equality and inequality constraints. The assumption $\mathcal{X} \subset \Delta_n$ 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 $\mathcal{X}$. Introducing slack variable $s$ given by $\sum_i \alpha_i x_i + s = 1$ and passing from signal $x$ to the new signal $[\alpha_1 x_1; \ldots; \alpha_n x_n; s]$, after a straightforward modification of matrices $A$ and $B$, we arrive at the situation where $\mathcal{X}$ is a subset of the probabilistic simplex.

Our goal in the sequel is to build a presumably good linear estimate $\hat{x}_H(\omega) = H^T \omega$ of $Bx$. Same as in Exercise 4.17, we start with upper-bounding the risk of a

---

20For 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.
linear estimate. When representing

$$\omega = Ax + \xi_x,$$

we arrive at zero mean observation noise \(\xi_x\) 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 \(\hat{x}_H(\cdot)\) in the same way as in Exercise 4.17. Specifically, denoting by \(\Pi_X\) the set of all diagonal matrices \(\text{Diag}\{Ax\}\), \(x \in X\) and by \(P_{i,x}\) the Poisson distribution with parameter \([Ax]_i\), we have

$$\text{Risk}_{\|\cdot\|} \left[ \hat{x}_H|X \right] = \sup_{x \in X} \mathbb{E}_{\omega \sim P_{1,x} \times \cdots \times P_{m,x}} \left\{ \|Bx - H^T\omega\| \right\}$$

$$\leq \sup_{x \in X} \mathbb{E}_{\xi} \left\{ \|B - H^TAx - H^T\xi_x\| \right\} + \sup_{\xi \in \Pi_X} \mathbb{E}_{\xi} \left\{ \|H^T\xi\| \right\}.$$

In order to build a presumably good linear estimate, it suffices to build efficiently computable convex in \(H\) upper bounds \(\Phi(H)\) on \(\Phi(H)\) and \(\Psi^X(H)\) on \(\Psi^X(H)\). and then take as \(H\) an optimal solution to the convex optimization problem

$$\text{Opt} = \min_{H} \left[ \Phi(H) + \Psi^X(H) \right].$$

Same as in Exercise 4.17, assume from now on that \(\|\cdot\|\) is an absolute norm, and the unit ball \(B_*\) of the conjugate norm is a spectratope:

$$B_* := \{ u : \|u\|_* \leq 1 \} = \{ u : \exists r \in \mathcal{R}, y : u = My, S^2_y[y] \preceq rI, \ell \leq L \}$$

Observe that

- In order to build \(\Phi\), we can use exactly the same techniques as those developed in Exercise 4.17. Indeed, as far as building \(\Phi\) is concerned, the only difference with the situation of Exercise 4.17 is that in the latter, \(A\) was column-stochastic matrix, while now \(A\) is just entrywise nonnegative matrix. Note, however, that when upper-bounding \(\Phi\) in Exercise 4.17, we never used the fact that \(A\) is column-stochastic.

- In order to upper-bound \(\Psi^X\), we can use the bound (4.40) of Exercise 4.17.

The bottom line is that in order to build a presumably good linear estimate, we need to solve the convex optimization problem

$$\text{Opt} = \min_{H,T,\Theta} \left\{ \Phi(H) + \phi_{\mathcal{R}}(\lambda[T]) + \Gamma_X(\Theta) : \mathcal{Y}_{\ell} \geq 0, \ell \leq L \right\} \text{(P)}$$

where

$$\Gamma_X(\Theta) = \max_{x \in X} \text{Tr}(\text{Diag}\{Ax\}\Theta)$$

(cf. problem (4.99)) with \(\Phi\) yielded by a whatever construction from Exercise 4.17, e.g., the least conservative Combined upper bound on \(\Phi\).

What in our present situation differs significantly from the situation of Exercise 4.17, are the bounds on probabilities of large deviations (for Discrete o.s., estab-
lished in Exercise 4.19), and the goal of what follows is to establish these bounds for Poisson Imaging.

Here is what you are supposed to do:

1. Let \( \omega \) be an \( m \)-dimensional random vector with independent entries \( \omega_i \sim \text{Poisson}(\mu_i) \), and let \( \mu = [\mu_1; \ldots; \mu_m] \). Prove that whenever \( h \in \mathbb{R}^m \), \( \gamma > 0 \), and \( \delta \geq 0 \), one has

\[
\ln \left( \frac{\text{Prob}\{ h^T \omega > h^T \mu + \delta \} }{\text{Prob}\{ h^T \omega > h^T \mu \}} \right) \leq \sum_i \left[ \exp \{ \gamma h_i \} - 1 \right] \mu_i - \gamma h^T \mu - \gamma \delta. \quad (4.107)
\]

2. Taking for granted (or see, e.g., [175]) that \( e^x - x - 1 \leq \frac{x^2}{2(1 - x/3)} \) when \( |x| < 3 \), prove that in the situation of item 1 one has for \( t > 0 \):

\[
0 \leq \gamma < \frac{3}{\| h \|_{\infty}} \Rightarrow \ln \left( \frac{\text{Prob}\{ h^T \omega > h^T \mu + t \} }{\text{Prob}\{ h^T \omega > h^T \mu \}} \right) \leq \frac{\gamma^2 \sum_i h_i^2 \mu_i}{2(1 - \gamma \| h \|_{\infty}/3)} - \gamma t. \quad (4.108)
\]

Derive from the latter fact that

\[
\text{Prob}\{ h^T \omega > h^T \mu + \delta \} \leq \exp \left\{ -\frac{\delta^2}{2\sum_i h_i^2 \mu_i + \| h \|_{\infty}/3} \right\}, \quad (4.109)
\]

and conclude that

\[
\text{Prob}\{ |h^T \omega - h^T \mu| > \delta \} \leq 2 \exp \left\{ -\frac{\delta^2}{2\sum_i h_i^2 \mu_i + \| h \|_{\infty}/3} \right\}. \quad (4.110)
\]

3. Extract from (4.110) the following

**Proposition 4.35.** In the situation and under the assumptions of Exercise 4.21, let \( \text{Opt} \) be the optimal value, and \( H, \mathbf{Y}, \Theta \) be a feasible solution to problem \( (P) \). Whenever \( x \in X \) and \( \epsilon \in (0, 1) \), denoting by \( P_x \) the distribution of observations stemming from \( x \) (i.e., the distribution of random vector \( \omega \) with independent entries \( \omega_i \sim \text{Poisson}(\| A x \|_i) \)), one has

\[
\mathbb{E}_{\omega \sim P_x} \{ \| B x - \hat{x}_H(\omega) \| \} \leq \mathcal{F}(H) + 2\sqrt{\phi_R(\lambda[\mathbf{Y}])} \text{Tr}(\text{Diag}(A x) \Theta) \leq \mathcal{F}(H) + \phi_R(\lambda[\mathbf{Y}]) + \Gamma_X(\Theta) \quad (4.111)
\]

and

\[
\text{Prob}_{\omega \sim P_x} \{ \| B x - \hat{x}_H(\omega) \| \leq \mathcal{F}(H) \} \geq 1 - \epsilon. \quad (4.112)
\]

**Note** that in the case of \( [A x]_i \geq 1 \) for all \( x \in X \) and all \( i \) we have \( \text{Tr}(\Theta) \leq \text{Tr}(\text{Diag}(A x) \Theta) \), so that in this case the \( P_x \)-probability of the event

\[
\{ \omega : \| B x - \hat{x}_H(\omega) \| \leq \mathcal{F}(H) + O(1) \ln(2m/\epsilon) \sqrt{\phi_R(\lambda[\mathbf{Y}])} \Gamma_X(\Theta) \}
\]

is at least \( 1 - \epsilon \).

### 4.7.6 Numerical lower-bounding minimax risk

**Exercise 4.22.**
4.22.A. Motivation. From the theoretical viewpoint, the results on near-optimality of presumably good linear estimates stated in Propositions 4.5, 4.16 seem to be quite strong and general. This being said, for a practically oriented user the “nonoptimality factors” arising in these propositions can be too large to make any practical sense. This 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 4.5, 4.16. 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.

4.22.B. Cramer-Rao bound. Consider the situation as follows: we are given

- an observation space \( \Omega \) equipped with reference measure \( \Pi \), basic examples being (A) \( \Omega = \mathbb{R}^m \) with Lebesgue measure \( \Pi \), and (B) (finite or countable) discrete set \( \Omega \) with counting measure \( \Pi \);
- a convex compact set \( \Theta \subset \mathbb{R}^k \) and a family \( P = \{ p(\omega, \theta) : \theta \in \Theta \} \) of probability densities, taken w.r.t. \( \Pi \).

Our goal is, given an observation \( \omega \sim p(\cdot, \theta) \) stemming from unknown \( \theta \) known to belong to \( \Theta \), to recover \( \theta \). We quantify the risk of a candidate estimate \( \hat{\theta} \) as

\[
\text{Risk}[\hat{\theta}|\Theta] = \sup_{\theta \in \Theta} \left( E_{\omega \sim p(\cdot, \theta)} \left\{ \| \hat{\theta}(\omega) - \theta \|^2 \right\} \right)^{1/2},
\]

and define the “ideal” minimax risk as

\[
\text{Risk}_{\text{opt}} = \inf_{\hat{\theta}} \text{Risk}[\hat{\theta}],
\]

the infimum being taken w.r.t. all estimates, or, which is the same, all bounded estimates (indeed, passing from a candidate estimate \( \hat{\theta} \) to the projected estimate \( \hat{\theta}_{\Theta}(\omega) = \arg\min_{\theta \in \Theta} \| \hat{\theta}(\omega) - \theta \|_2 \) will only reduce the estimate risk).

The Cramer-Rao inequality [60, 202], which we intend to use,\(^{21}\) is a certain relation between the covariance matrix of a bounded estimate and its bias; this relation is valid under mild regularity assumptions on the family \( P \), specifically, as follows:

1) \( p(\omega, \theta) > 0 \) for all \( \omega \in \Omega, \theta \in U \), and \( p(\omega, \theta) \) is differentiable in \( \theta \), the with \( \nabla_{\theta} p(\omega, \theta) \) continuous in \( \theta \in \Theta \);

\(^{21}\)As a matter of fact, the classical Cramer-Rao inequality dealing with unbiased estimates is not sufficient for our purposes “as is.” What we need to build bounds for the minimax risk is a “bias enabled” version of this inequality. Such inequality may be developed using Bayesian argument [98, 229]. To the best of our knowledge, the version of the inequality we discuss here first appeared in [188].
2) The Fisher Information matrix

\[ \mathcal{I}(\theta) = \int_{\Omega} \frac{\nabla p(\omega, \theta) [\nabla p(\omega, \theta)]^T}{p(\omega, \theta)} \Pi(d\omega) \]

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

\[ ||\nabla p(\omega, \theta)||_2 \leq M(\omega) \quad \forall \omega \in \Omega, \theta \in \Theta. \]

The derivation of the Cramer-Rao bound is as follows. Let \( \hat{\theta}(\omega) \) be a bounded estimate, and let

\[ \phi(\theta) = [\phi_1(\theta); \ldots; \phi_k(\theta)] = \int_{\Omega} \hat{\theta}(\omega)p(\omega, \theta)\Pi(d\omega) \]

be the expected value of the estimate. By item 3, \( \phi(\theta) \) is differentiable on \( \Theta \), with the Jacobian \( \phi'(\theta) = \left[ \frac{\partial \phi_i(\theta)}{\partial \theta_j} \right]_{i,j \leq k} \) given by

\[ \phi'(\theta) = \int_{\Omega} \hat{\theta}(\omega)h^T \nabla p(\omega, \theta)\Pi(d\omega), \quad h \in \mathbb{R}^k. \]

Besides this, recalling that \( \int_{\Omega} p(\omega, \theta)\Pi(d\omega) \equiv 1 \) and invoking item 3, we have

\[ \int_{\Omega} h^T \nabla p(\omega, \theta)\Pi(d\omega) = 0, \text{ whence, in view of the previous identity,} \]

\[ \phi'(\theta)h = \int_{\Omega} [\hat{\theta}(\omega) - \phi(\theta)]h^T \nabla p(\omega, \theta)\Pi(d\omega), \quad h \in \mathbb{R}^k. \]

Therefore for all \( g, h \in \mathbb{R}^k \) we have

\[ [g^T \phi'(\theta)h]^2 = \left[ \int_{\Omega} g^T(\hat{\theta} - \phi(\theta)) [h^T \nabla p(\omega, \theta)/p(\omega, \theta)]p(\omega, \theta)\Pi(d\omega) \right]^2 \]

\[ \leq \left[ \int_{\Omega} g^T(\hat{\theta} - \phi(\theta)) [\hat{\theta} - \phi(\theta)]^T g \Pi(d\omega) \right] \times \left[ \int_{\Omega} [h^T \nabla p(\omega, \theta)/p(\omega, \theta)]^2 p(\omega, \theta)\Pi(d\omega) \right] \]

\[ \text{[by the Cauchy Inequality]} \]

\[ = [g^T \text{Cov}_{\hat{\theta}}(\theta)g] [h^T \mathcal{I}(\theta)h], \]

where \( \text{Cov}_{\hat{\theta}}(\theta) \) is the covariance matrix \( \mathbb{E}_{\omega \sim p(\cdot, \theta)} \left\{ [\hat{\theta}(\omega) - \phi(\theta)][\hat{\theta}(\omega) - \phi(\theta)]^T \right\} \) of \( \hat{\theta}(\omega) \) induced by \( \omega \sim p(\cdot, \theta) \). We have arrived at the inequality

\[ [g^T \text{Cov}_{\hat{\theta}}(\theta)g] [h^T \mathcal{I}(\theta)h] \geq [g^T \phi'(\theta)h]^2 \quad \forall (g, h \in \mathbb{R}^k, \theta \in \Theta). \quad (*) \]

For \( \theta \in \Theta \) fixed, let \( \mathcal{J} \) be a positive definite matrix such that \( \mathcal{J} \geq \mathcal{I}(\theta) \), whence by \((*)\) it holds

\[ [g^T \text{Cov}_{\hat{\theta}}(\theta)g] [h^T \mathcal{J}h] \geq [g^T \phi'(\theta)h]^2 \quad \forall (g, h \in \mathbb{R}^k). \quad (***) \]

For \( g \) fixed, the maximum of the right hand side quantity in \((***)\) over \( h \) satisfying \( h^T \mathcal{J}h \leq 1 \) is \( g^T \phi'(\theta)\mathcal{J}^{-1}[\phi'(\theta)]^T g \), and we arrive at the Cramer-Rao inequality

\[ \forall (\theta \in \Theta, \mathcal{J} \geq \mathcal{I}(\theta), \mathcal{J} > 0) : \text{Cov}_{\hat{\theta}}(\theta) \leq [\phi'(\theta)]^T \mathcal{J}^{-1}[\phi'(\theta)]^T \quad (4.114) \]
which holds true for every bounded estimate $\hat{\theta}(\cdot)$. Note also that for every $\theta \in \Theta$ and every bounded estimate $x$ we have

$$\text{Risk}^2[\hat{\theta}] \geq \mathbb{E}_{\omega \sim p(\cdot, \theta)} \left\{ [\hat{\theta}(\omega) - \theta]^2 \right\} = \mathbb{E}_{\omega \sim p(\cdot, \theta)} \left\{ [\hat{\theta}(\omega) - \phi(\theta)]^2 + \|\phi(\theta) - \theta\|_2^2 \right\} - 2 \mathbb{E}_{\omega \sim p(\cdot, \theta)} \left[ (\hat{\theta}(\omega) - \phi(\theta))^T (\phi(\theta) - \theta) \right] = 0$$

Hence, in view of (4.114), for every bounded estimate $\hat{\theta}$ it holds

$$\forall (J > 0 : J \succeq \mathcal{I}(\theta) \forall \theta \in \Theta) : \text{Risk}^2[\hat{\theta}] \geq \sup_{\theta \in \Theta} \left[ \text{Tr}(\phi'(\theta)J^{-1} [\phi'(\theta)]^T) + \|\phi(\theta) - \theta\|^2_2 \right],$$

(4.115)

The fact that we considered the risk of estimating “the entire” $\theta$ rather than a given vector-valued function $f(\theta) : \Theta \to \mathbb{R}^p$ plays no special role, and in fact the Cramer-Rao inequality admits the following modification yielded by a completely similar reasoning:

**Proposition 4.36.** In the situation described in item 4.22.B and under assumptions 1) – 3) of this item, let $f(\cdot) : \Theta \to \mathbb{R}^p$ be a bounded Borel function, and let $\hat{f}(\omega)$ be a bounded estimate of $f(\omega)$ via observation $\omega \sim p(\cdot, \theta)$. Then, setting for $\theta \in \Theta$

$$\phi(\theta) = \mathbb{E}_{\omega \sim p(\cdot, \theta)} \left\{ \hat{f}(\theta) \right\},$$

$$\text{Cov}_{\hat{f}}(\theta) = \mathbb{E}_{\omega \sim p(\cdot, \theta)} \left\{ (\hat{f}(\omega) - \phi(\theta))[\hat{f}(\omega) - \phi(\theta)]^T \right\},$$

one has

$$\forall (\theta \in \Theta, J \succeq \mathcal{I}(\theta), J > 0) : \text{Cov}_{\hat{f}}(\theta) \succeq \phi'(\theta)J^{-1} [\phi'(\theta)]^T.$$

As a result, for

$$\text{Risk}[\hat{f}] = \sup_{\theta \in \Theta} \left[ \mathbb{E}_{\omega \sim p(\cdot, \theta)} \left\{ \|\hat{f}(\omega) - f(\theta)\|^2_2 \right\} \right]^{1/2}$$

it holds

$$\forall (J > 0 : J \succeq \mathcal{I}(\theta) \forall \theta \in \Theta) : \text{Risk}^2[\hat{f}] \geq \sup_{\theta \in \Theta} \left[ \text{Tr}(\phi'(\theta)J^{-1} [\phi'(\theta)]^T) + \|\phi(\theta) - f(\theta)\|^2_2 \right]$$

Now goes the first part of the exercise:

1) Derive from (4.115) the following

**Proposition 4.37.** In the situation of item 4.22.B, let

- $\Theta \subset \mathbb{R}^k$ be $\| \cdot \|_2$-ball of radius $r > 0$, 

the family $\mathcal{P}$ be such that $I(\theta) \preceq J$ for some $J > 0$ and all $\theta \in \Theta$.

Then the minimax optimal risk satisfies the bound

$$\text{Risk}_{\text{opt}} \geq \frac{rk}{r \sqrt{\text{Tr}(J)} + k}.$$  

(4.116)

In particular, when $J = \alpha^{-1} I_k$, we have

$$\text{Risk}_{\text{opt}} \geq \frac{r \sqrt{\alpha k}}{r + \sqrt{\alpha k}}.$$  

(4.117)

Hint. Assuming w.l.o.g. that $\Theta$ is centered at the origin, and given a bounded estimate $\hat{\theta}$ with risk $R$, let $\phi(\theta)$ be associated with the estimate via (4.115). 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 $R$ by $\max_{\theta \in \Theta} \|\phi(\theta) - \theta\|_2$, see (4.115). In the case of (b), lower-bound $R^2$ by $\max_{\theta \in \Theta} \text{Tr}(\phi'(\theta) J^{-1} [\phi'(\theta)]^T)$, see (4.115), and use Gauss 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 program, you might find useful the following fact (prove it!)

**Lemma 4.38.** Let $\Phi$ be an $n \times n$ matrix, and $J$ be a positive definite $n \times n$ matrix. Then

$$\text{Tr}(\Phi J^{-1} \Phi^T) \geq \frac{\text{Tr}(\Phi)^2}{\text{Tr}(J)}.$$  

4.22.C. Application to signal recovery. Proposition 4.37 allows to build computation-based lower risk bounds in the signal recovery problem considered in Section 4.2, in particular, the problem where one wants to recover the linear image $Bx$ of unknown signal $x$ known to belong to a given ellitope $X = \{x \in \mathbb{R}^n : \exists t \in T : x^T S_{\ell} x \leq t, \ell \leq L\}$

(with our usual restriction on $S_{\ell}$ and $T$) 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 (4.113).\footnote{In fact, the approach to be developed can be applied to signal recovery problems involving Discrete/Poisson observation schemes and different from $\| \cdot \|_2$ norms used to measure the recovery error, signal-dependent noises, etc.} 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.\footnote{This assumption is nonrestrictive. Indeed, when $B \in \mathbb{R}^{\nu \times n}$ 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 $\text{Ker} B = \{0\}$, which, for small enough perturbation, again keeps our estimation problem basically intact. It remains to note that when $\text{Ker} B = \{0\}$ we can replace $\mathbb{R}^\nu$ with the image space of $B$, which again does not affect the estimation problem we are interested in.} Under this
assumption, setting
\[ \mathcal{Y} = B^{-1}\mathcal{X} = \{ y \in \mathbb{R}^n : \exists t \in T : y^T[B^{-1}]^TS_tB^{-1}y \leq t, \ell \leq L \} \]
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 \( \mathcal{X} \). Note that in our new situation \( A \) is replaced with \( \bar{A} \), \( \mathcal{X} \) with \( \mathcal{Y} \), and \( B \) is the unit matrix \( I_n \). For the sake of simplicity, we assume from now on that \( A \) (and therefore \( \bar{A} \)) has trivial kernel. Finally, let \( \bar{S}_t \geq S_t \) be close to \( S_k \) positive definite matrices, e.g., \( \bar{S}_t = S_t + 10^{-100}I_n \). Setting \( \bar{S}_t = [B^{-1}]^T\bar{S}_tB^{-1} \) and
\[ \bar{\mathcal{Y}} = \{ y \in \mathbb{R}^n : \exists t \in T : y^T\bar{S}_ty \leq t, \ell \leq L \}, \]
we get \( \bar{S}_t \succeq 0 \) and \( \bar{\mathcal{Y}} \subset \mathcal{Y} \). Therefore, 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 \( \text{Risk}_{\text{opt}} \) corresponding to our original problem of interest.

Now assume that we can point out a \( k \)-dimensional linear subspace \( E \) in \( \mathbb{R}^n \) and positive reals \( r, \gamma \) such that

(i) the centered at the origin \( \| \cdot \|_2 \)-ball \( \Theta = \{ \theta \in E : \| \theta \|_2 \leq r \} \) is contained in \( \bar{\mathcal{Y}} \);
(ii) The restriction \( \bar{A}_E \) of \( \bar{A} \) onto \( E \) satisfies the relation
\[ \text{Tr}(\bar{A}_E^*\bar{A}_E) \leq \gamma \]
\( (\bar{A}_E : \mathbb{R}^m \rightarrow E \) is the conjugate of the linear map \( \bar{A}_E : E \rightarrow \mathbb{R}^m \).\)

Consider the auxiliary estimation problem obtained from the (reformulated) problem of interest by replacing the signal set \( \bar{\mathcal{Y}} \) with \( \Theta \). Since \( \Theta \subset \bar{\mathcal{Y}} \), the minimax risk in the auxiliary problem is a lower bound on the minimax risk \( \text{Risk}_{\text{opt}} \) we are interested in. On the other hand, the auxiliary problem is nothing but the problem of recovering parameter \( \theta \in \Theta \) from observation \( \omega \sim \mathcal{N}(\bar{A}\theta, \sigma^2I) \), which is nothing but a special case of the problem considered in item 4.22.B. As it is immediately seen, the Fisher Information matrix in this problem is independent of \( \theta \) and is \( \sigma^{-2}\bar{A}_E^*\bar{A}_E \):
\[ e^T\mathcal{I}(\theta)e = \sigma^{-2}e^T\bar{A}_E^*\bar{A}_Ee, e \in E. \]
Invoking Proposition 4.37, we arrive at the lower bound on the minimax risk in the auxiliary problem (and thus – in the problem of interest as well):
\[ \text{Risk}_{\text{opt}} \geq \frac{r\sigma k}{r\sqrt{\gamma} + \sigma k}. \tag{4.118} \]
The resulting risk bound depends on \( r, k, \gamma \) and is the larger the smaller is \( \gamma \) and the larger are \( k \) and \( r \).

**Lower-bounding** \( \text{Risk}_{\text{opt}} \). In order to make the just outlined bounding scheme to give its best, we need a mechanism which allows to generate \( k \)-dimensional “disks” \( \Theta \subset \bar{\mathcal{Y}} \) along with associated quantities \( r, \gamma \). In order to design such a mechanism, it is convenient to represent \( k \)-dimensional linear subspaces of \( \mathbb{R}^n \) 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 \( \Theta_P \) of the radius \( r = r_P \), where \( r_P \) is the largest \( \rho \) such that the set \( \{ y \in \text{Im}P : y^TPy \leq \rho^2 \} \) is contained in \( \bar{\mathcal{Y}} \) (“condition \( \mathcal{C}(r) \)”), and we...
can equip the disk with $\gamma$ satisfying (ii) if and only if
$$\text{Tr}(PA^T \bar{A}P) \leq \gamma,$$
or, which is the same (recall that $P$ is orthogonal projector)
$$\text{Tr}(\bar{A}P\bar{A}^T) \leq \gamma$$
(4.119)
(“condition $\mathcal{D}(\gamma)$”). Now, when $P$ is a nonzero orthogonal projector, the simplest
sufficient condition for the validity of $C(r)$ is the existence of $t \in \mathcal{T}$ such that
$$\forall (y \in \mathbb{R}^n, \ell \leq L) : y^T P S_\ell P y \leq t \ell r^{-2} y^T P y,$$
or, which is the same,
$$\exists s : r^2 s \in \mathcal{T} \& P S_\ell P \preceq s \ell P, \ell \leq L.$$ (4.120)
Let us rewrite (4.119), (4.120) as a system of linear matrix inequalities. This is
what you are supposed to do:

2.1) Prove the following simple fact:

Observation 4.39. Let $Q$ be a positive definite, $R$ be a nonzero positive semidefinite matrix, and let $s$ be a real. Then
$$RQR \preceq sR$$
if and only if
$$sQ^{-1} \succeq R.$$
2.2) Extract from the above observation the conclusion as follows. Let $\mathcal{T}$ be the conic hull of $\mathcal{T}$:
$$\mathcal{T} = \text{cl}\{[s; \tau] : \tau > 0, s/\tau \in \mathcal{T}\} = \{[s; \tau] : \tau > 0, s/\tau \in \mathcal{T}\} \cup \{0\}.$$
Consider the system of constraints
$$[s; \tau] \in \mathcal{T} \& s \ell S_\ell^{-1} \succeq P, \ell \leq L \& \text{Tr}(\bar{A}P\bar{A}^T) \leq \gamma,$$\label{eq:system}
in variables $[s; \tau], k, \gamma$ and $P$. Every feasible solution to this system gives rise to
a $k$-dimensional Euclidean subspace $E \subset \mathbb{R}^n$ (the image space of $P$) such that
the centered at the origin Euclidean ball $\Theta$ in $E$ of radius
$$r = 1/\sqrt{\tau}$$
taken along with $\gamma$ satisfy conditions (i) - (ii). Consequently, such feasible solution yields the lower bound
$$\text{Risk}_{\text{opt}} \geq \psi_{\sigma,k}(\gamma, \tau) := \frac{\sigma_k}{\sqrt{\tau} + \sigma \sqrt{\tau} k}$$
on the minimax risk in the problem of interest.

Ideally, to utilize item 2.2 to lower-bound $\text{Risk}_{\text{opt}}$, we should look through $k =$
1, ..., n and maximize for every \( k \) the lower risk bound \( \psi_{\sigma,k}(\gamma, \tau) \) under constraints (\#), thus arriving at the problem

\[
\min_{[s; \tau], \gamma, P} \left\{ \frac{\gamma}{k^2} + \sigma^2 \tau : \begin{array}{l}
[s; \tau] \in T \land s_t \tilde{S}_t^{-1} \succeq P, \ell \leq L \land \text{Tr}(AP\tilde{A}^T) \leq \gamma, \\
P \text{ is an orthogonal projector of rank } k.
\end{array} \right\} (P_k)
\]

This problem seems to be computationally intractable, since the constraints of \((P_k)\) include the nonconvex restriction on \( P \) to be a projector of rank \( k \). A natural convex relaxation of this constraint is

\[
0 \preceq P \preceq I_n, \quad \text{Tr}(P) = k.
\]

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 a convex “proxy” of \((P_k)\) – the problem

\[
\min_{[s; \tau], \gamma, P} \left\{ \frac{\gamma}{k^2} + \sigma^2 \tau : \begin{array}{l}
[s; \tau] \in T, 0 \preceq P \preceq I_n, \text{Tr}(P) = k, \\
s_t \tilde{S}_t^{-1} \succeq P, \ell \leq L, \text{Tr}(AP\tilde{A}^T) \leq \gamma
\end{array} \right\} (P[k])
\]

\( k = 1, \ldots, n \). Problem \((P[k])\) clearly is solvable, and the \( P \)-component \( P^{(k)} \) of its optimal solution gives rise to a collection of orthogonal projectors \( P^{(k)}_\kappa \), \( \kappa = 1, \ldots, n \) obtained from \( P^{(k)} \) by “rounding” – to get \( P^{(k)}_\kappa \), we replace the \( \kappa \) leading eigenvalues of \( P^{(k)} \) with ones, and the remaining eigenvalues – with zeros, while keeping the eigenvectors intact. We can now for every \( \kappa = 1, \ldots, n \) fix the \( P \)-variable in \((P_k)\) as \( P^{(k)}_\kappa \) and solve the resulting problem in the remaining variables \([s; \tau]\) and \( \gamma \), which is easy – with \( P \) fixed, the problem clearly reduces to minimizing \( \tau \) under the convex constraints

\[
s_t \tilde{S}_t^{-1} \succeq P, \ell \leq L, [s; \tau] \in T
\]

on \([s; \tau]\). As a result, for every \( k \in \{1, \ldots, n\} \), we get \( n \) lower bounds on Risk\(_{\text{opt}}\), 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 program numerically and compare the lower bound on the minimax risk with the upper risk bounds of presumably good linear estimates yielded by Proposition 4.4.

Recommended setup:

- Sizes: \( m = n = \nu = 16 \)
- \( A, B : B = I_n, A = \text{Diag}\{a_1, \ldots, a_n\} \) with \( a_i = i^{-\alpha} \) and \( \alpha \) running through \( \{0, 1, 2\} \);
- \( \mathcal{X} = \{x \in \mathbb{R}^n : x^T S_{\ell} x \leq 1, \ell \leq L\} \) (i.e., \( T = [0, 1]^L \)) with randomly generated \( S_{\ell}\),
- Range of \( L \): \( \{1, 4, 16\} \). For \( L \) in this range, you can generate \( S_{\ell}, \ell \leq L \), as \( S_{\ell} = R_{\ell} R_{\ell}^T \) with \( R_{\ell} = \text{randn}(n, p) \), where \( p = \lfloor n/L \rfloor \).
- Range of \( \sigma \): \( \{1, 0.1, 0.01, 0.001, 0.0001\} \)

Exercise 4.23. [follow-up to Exercise 4.22]
1) Prove the following version of Proposition 4.37:

**Proposition 4.40.** In the situation of item 4.22.B and under Assumptions 1) – 3) from this item, let

- $\| \cdot \|$ be a norm on $\mathbb{R}^k$ such that
  $$\| \theta \|_2 \leq \kappa \| \theta \| \forall \theta \in \mathbb{R}^k$$
- $\Theta \subset \mathbb{R}^k$ be $\| \cdot \|$-ball of radius $r > 0$,
- the family $\mathcal{P}$ be such that $I(\theta) \preceq J$ for some $J \succ 0$ and all $\theta \in \Theta$.

Then the minimax optimal risk

$$\text{Risk}_{\text{opt}, \| \cdot \|} = \inf_{\hat{\theta}} \left( \sup_{\theta \in \Theta} \mathbb{E}_{\omega \sim p(\cdot, \theta)} \left\{ \| \theta - \hat{\theta}(\omega) \|_2^2 \right\} \right)^{1/2}$$

of recovering parameter $\theta \in \Theta$ from observation $\omega \sim p(\cdot, \theta)$ in the norm $\| \cdot \|$ satisfies the bound

$$\text{Risk}_{\text{opt}, \| \cdot \|} \geq \frac{r \kappa}{r \kappa \sqrt{\text{Tr}(J)} + k}, \quad (4.121)$$

In particular, when $J = \alpha^{-1} I_k$, we get

$$\text{Risk}_{\text{opt}, \| \cdot \|} \geq \frac{r \sqrt{\alpha k}}{r \kappa + \sqrt{\alpha k}}, \quad (4.122)$$

2) Apply Proposition 4.40 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 $\theta$ known to belong to $\Theta = \{ \theta \in \mathbb{R}^k : \| \theta \|_p \leq r \}$ with given $A$, Ker $A = \{ 0 \}$, $p \in [2, \infty]$, $r > 0$, we want to recover $\theta$ in $\| \cdot \|_p$.

- 2.2) Given indirect observation $\omega = L\theta R + \sigma \xi$, where $\theta$ is unknown $\mu \times \nu$ matrix known to belong to the Shatten norm ball $\Theta \subset \mathbb{R}^{\mu \times \nu} : \| \theta \|_{\text{Sh}, p} \leq r$, we want to recover $\theta$ in $\| \cdot \|_{\text{Sh}, p}$. Here $L \in \mathbb{R}^{m \times \mu}$, Ker $L = \{ 0 \}$ and $R \in \mathbb{R}^{n \times \nu}$, Ker $R^T = \{ 0 \}$ are given matrices, $p \in [2, \infty]$, and $\xi$ is random Gaussian $m \times n$ matrix (i.e., the entries in $\xi$ are independent of each other $\mathcal{N}(0, 1)$ random variables).

- 2.3) Given a $K$-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 $\theta \in \mathbb{S}^n$ 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 $\theta$ in the spectral norm.

Exercise 4.24.

[More on Cramer-Rao risk bound] Let us fix $\mu \in (1, \infty)$ and a norm $\| \cdot \|$ on $\mathbb{R}^k$, and let $\| \cdot \|_*$ be the norm conjugate to $\| \cdot \|$, and $\mu_* = \frac{\mu}{\mu - 1}$. Assume that we are in the situation of item 4.22.B and under assumptions 1) and 3) from this item; as about assumption 2) we now replace it with the assumption that the quantity

$$I_{\| \cdot \|, \mu_*}(\theta) := \left[ \mathbb{E}_{\omega \sim p(\cdot, \theta)} \left\{ \| \nabla_{\theta} p(\omega, \theta) \|_{\mu_*}^\mu \right\} \right]^{1/\mu_*}$$
is well defined and bounded on $\Theta$; in the sequel, we set
\[ I_{\| \cdot \|, \mu} = \sup_{\theta \in \Theta} I_{\| \cdot \|, \mu} (\theta). \]

1) Prove the following variant of Cramer-Rao risk bound:

**Proposition 4.41.** In the situation described in the beginning of item 4.22.D, let $\Theta \subset \mathbb{R}^k$ 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

\[ \text{Risk}_{\| \cdot \|, \mu} := \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \left[ \mathbb{E}_{\omega \sim p(\cdot, \theta)} \left\{ \| \hat{\theta}(\omega) - \theta \|^\mu \right\} \right]^{1/\mu} \geq \frac{r^k}{\text{rk}_0}, \]

\[ I_{\| \cdot \|, \mu} = \max_{\theta \in \Theta} \left[ I_{\| \cdot \|, \mu} (\theta) := \mathbb{E}_{\omega \sim p(\cdot, \theta)} \left\{ \| \nabla \ln(p(\omega, \theta)) \|^\mu \right\} \right]^{1/\mu}. \]

**Example I: Gaussian case, estimating shift.** Let $\mu = 2$, and let $p(\omega, \theta) = \mathcal{N}(A\theta, \sigma^2 I_m)$ with $A \in \mathbb{R}^{m \times k}$. Then

\[ \nabla \ln(p(\omega, \theta)) = \sigma^{-2} A^T (\omega - A\theta) \Rightarrow \int \| \nabla \ln(p(\omega, \theta)) \|^2 p(\omega, \theta) d\omega = \sigma^{-4} \int \| A^T (\omega - A\theta) \|^2 p(\omega, \theta) d\omega \]

\[ = \sigma^{-4} \int \frac{1}{(2\pi\sigma)^m} \| A^T \omega \|^2 \exp\left\{-\frac{1}{2\sigma^2} \| A^T \omega \|^2 \right\} d\omega \]

\[ = \sigma^{-4} \int \frac{1}{(2\pi)^{m/2}} \| A^T \|_2^2 \exp\left\{-\frac{1}{2\sigma^2} \| A^T \|_2^2 \right\} d\omega \]

whence
\[ I_{\| \cdot \|, 2} = \sigma^{-1} \left[ \mathbb{E}_{\xi \sim \mathcal{N}(0, I_m)} \left\{ \| A^T \xi \|_2^2 \right\} \right]^{1/2}. \]

Consequently, assuming $\Theta$ to be a $\| \cdot \|$-ball of radius $r$ in $\mathbb{R}^k$, lower bound (4.123) becomes

\[ \text{Risk}_{\| \cdot \|, \mu} (\Theta) \geq \frac{r^k}{\text{rk}_0} \geq \frac{r^k}{\text{rk}_0 \gamma_{\| \cdot \|}(A) + k} = \frac{r\sigma k}{r\gamma_{\| \cdot \|}(A) + \sigma k}. \]

**The case of direct observations.** To see “how it works,” consider the case $m = k$, $A = I_k$ of direct observations, and let $\Theta = \{ \theta \in \mathbb{R}^k : \| \theta \| \leq r \}$. Then

- We have $\gamma_{\| \cdot \|}(I_k) \leq O(1) \sqrt{\ln(k)}$, whence the $\| \cdot \|_1$-risk bound is
  \[ \text{Risk}_{\| \cdot \|_1} (\Theta) \geq O(1)\frac{r\sigma k}{r\sqrt{\ln(k)} + \sigma k}; \quad [\Theta = \{ \theta \in \mathbb{R}^k : \| \theta - a \|_1 \leq r \}] \]

- We have $\gamma_{\| \cdot \|}(I_k) = \sqrt{k}$, whence the $\| \cdot \|_2$-risk bound is
  \[ \text{Risk}_{\| \cdot \|_2} (\Theta) \geq \frac{r\sigma \sqrt{k}}{r + \sigma \sqrt{k}}; \quad [\Theta = \{ \theta \in \mathbb{R}^k : \| \theta - a \|_2 \leq r \}] \]
• We have $\gamma_{\|\cdot\|_\infty}(I_k) \leq O(1)^k$, whence the $\|\cdot\|_\infty$-risk bound is
\[
\text{Risk}_{\text{opt},\|\cdot\|_\infty}[\Theta] \geq O(1) \frac{r \sigma}{r + \sigma}.
\]
In fact, the above examples are essentially covered by the following

Observation 4.42. Let $\|\cdot\|$ be a norm on $\mathbb{R}^k$, and let
\[
\Theta = \{ \theta \in \mathbb{R}^k : \|\theta-a\|_\infty \leq r \}\]

be the associated minimax risk.

Assume that the norm $\|\cdot\|$ is absolute and symmetric w.r.t permutation of coordinates. Then
\[
\text{Risk}_{\text{opt},\|\cdot\|}[\Theta] \geq \frac{r \sigma k}{2 \sqrt{\ln(e^k)} r \alpha_\# + \sigma k}, \quad \alpha_\# = \|1; \ldots; 1\|_*.
\]

Here is the concluding part of the exercise:

2) Prove the observation and compare the lower risk bound it yields with the $\|\cdot\|_\infty$-risk of the “plug-in” estimate $\hat{\chi}(\omega) \equiv \omega$.

Example II: Gaussian case, estimating covariance. Let $\mu = 2$, let $K$ be a positive integer, and let our observation $\omega$ be a collection of $K$ i.i.d. samples $\omega_t \sim N(0, \sigma^2 I_k)$, $1 \leq t \leq K$, with unknown $\theta$ known to belong to a given convex compact subset $\Theta$ of the positive semidefinite cone $S^+_n$. Given $\omega_1, \ldots, \omega_K$, we want to recover $\theta$ in the Shatten norm $\|\cdot\|_{\text{Sh},s}$ with $s \in [1, \infty]$. Our estimation problem is covered by the setup of Exercise 4.22 with $P$ 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 $N(0, \theta)$. We have
\[
\nabla_\theta \ln(p(\omega, \theta)) = \frac{1}{2} \sum_t \nabla_\theta \ln(g(\omega_t, \theta)) = \frac{1}{2} \sum_t \left[ \theta^{-1} \omega_t \omega_t^T \theta^{-1} - \theta^{-1} \right] = \frac{1}{2} \theta^{-1/2} \left[ \sum_t \left[ (\theta^{-1/2} \omega_t) \theta^{-1/2} \omega_t^T - I_n \right] \right] \theta^{-1/2}
\]

With some effort [146] it can be proved that when
\[
K \geq n,
\]
which we assume from now on, for independent across $t$ random vectors $\xi_1, \ldots, \xi_K$ sampled from the standard Gaussian distribution $N(0, I_n)$ for every $u \in [1, \infty]$ one
We can now use Proposition 4.41 to lower-bound the minimax risk has

\[ \left( \mathbb{E} \left( \left\| \sum_{t=1}^{K} \xi_t^T - I_n \right\|_{2, u}^2 \right) \right)^{1/2} \leq C n^{\frac{1}{2} + \frac{1}{2}} \sqrt{K} \]  \tag{4.127}

with appropriate absolute constant \( C \). Consequently, for \( \theta \in \Theta \) and all \( u \in [1, \infty] \) we have

\[
\begin{align*}
\mathbb{E}_{\omega \sim \mathcal{P}(\cdot, \theta)} \left\{ \| \nabla_{\theta} \ln(p(\omega, \theta)) \|_{2, u}^2 \right\} &= \frac{1}{n} \mathbb{E}_{\omega \sim \mathcal{P}(\cdot, \theta)} \left\{ \| \theta^{-1/2} \left[ \sum_t \left[ \theta^{-1/2} \omega_t \right]^T - I_n \right] \theta^{-1/2} \|_{2, u}^2 \right\} \\
&= \frac{1}{n} \mathbb{E}_{\xi \sim \mathcal{P}(\cdot, I_n)} \left\{ \| \theta^{-1/2} \left[ \sum_t \xi_t^T - I_n \right] \theta^{-1/2} \|_{2, u}^2 \right\} \text{[by (4.126)]} \\
&\leq \frac{1}{n} \| \theta^{-1/2} \|_{2, \infty} \mathbb{E}_{\xi \sim \mathcal{P}(\cdot, I_n)} \left\{ \| \sum_t [\xi_t^T - I_n] \|_{2, u}^2 \right\} \text{[setting \( \theta^{-1/2} \omega_t = \xi_t \)]} \\
&\leq \frac{1}{n} \| \theta^{-1/2} \|_{2, \infty} C n^{\frac{1}{2} + \frac{1}{2}} \sqrt{K} \leq \frac{C}{2} \| \theta^{-1} \|_{2, \infty} n^{\frac{1}{2} + \frac{1}{2}} \sqrt{K}. \tag{4.128}
\end{align*}
\]

Now assume that \( \Theta \) is \( \| \cdot \|_{2, \infty} \)-ball of radius \( r < 1 \) centered at \( I_n \):

\[ \Theta = \{ \theta \in \mathbb{S}^n : \| \theta - I_n \|_{2, \infty} \leq r \}. \tag{4.129} \]

In this case the estimation problem from Example II is the scope of Proposition 4.41, and the quantity \( I_{\| \cdot \|_{2, \infty}} \) as defined in (4.123) can be upper-bounded as follows:

\[
\begin{align*}
I_{\| \cdot \|_{2, \infty}} &= \max_{\theta \in \Theta} \left[ \mathbb{E}_{\omega \sim \mathcal{P}(\cdot, \theta)} \left\{ \| \nabla_{\theta} \ln(p(\omega, \theta)) \|_{2, u}^2 \right\} \right]^{1/2} \\
&\leq O(1) n^{\frac{1}{2} + \frac{1}{2}} \sqrt{K} \max_{\theta \in \Theta} \| \theta^{-1} \|_{2, \infty} \quad \text{[see (4.128)]} \\
&\leq O(1) n^{\frac{1}{2} + \frac{1}{2}} \frac{1}{1 - r} \sqrt{K}.
\end{align*}
\]

We can now use Proposition 4.41 to lower-bound the minimax risk \( \| \cdot \|_{2, \infty} \)-risk, thus arriving at

\[
\text{Risk}_{\text{opt.}, \| \cdot \|_{2, \infty}}[\Theta] \geq O(1) n^{(1-r)r} \sqrt{K n^2 \frac{1}{r} + n(1-r)} \tag{4.130}
\]

(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 \( \| \cdot \|_{2, \infty} \)-risk of the “plug-in” estimate

\[ \hat{\theta}(\omega) = \frac{1}{K} \sum_{t=1}^{K} \omega_t \theta^T. \]
Assuming \( \theta \in \Theta \), we have
\[
E_{\omega \sim p(\cdot, \theta)} \left\{ \| K\hat{\theta}(\omega) - \theta \|_{S_{h,s}}^2 \right\} = E_{\omega \sim p(\cdot, \theta)} \left\{ \| \sum_t [\omega_t^T \omega_t - \theta_t^2] \|_{S_{h,s}}^2 \right\}
\]
\[
= E_{\omega \sim p(\cdot, \theta)} \left\{ \| \theta^{1/2} \left[ \sum_t [\theta^{-1/2} \omega_t \theta^{1/2} \omega_t^T - I_n] \right] \theta^{1/2} \|_{S_{h,s}}^2 \right\}
\]
\[
= E_{\xi \sim p(\cdot, I_n)} \left\{ \| \theta^{1/2} \left[ \sum_t [\xi_t \xi_t^T - I_n] \right] \theta^{1/2} \|_{S_{h,s}}^2 \right\}
\]
\[
\leq \| \theta^{1/2} \|_{S_{h,\infty}} E_{\xi \sim p(\cdot, I_n)} \left\{ \| \sum_t [\xi_t \xi_t^T - I_n] \|_{S_{h,s}}^2 \right\}
\]
\[
\leq \| \theta^{1/2} \|_{S_{h,\infty}}^4 \left[ C n^{1/2} + \sqrt{K} \right]^2, \quad \text{[see (4.127)]}
\]
and we arrive at
\[
\text{Risk}_{\|\cdot\|_{S_{h,s}} \left\{ \hat{\theta} \vert \Theta \right\}} \leq O(1) \max_{\theta \in \Theta} \| \theta \|_{S_{h,\infty}} n^{1/2} \sqrt{K}. \]

In the case of (4.129), the latter bound becomes
\[
\text{Risk}_{\|\cdot\|_{S_{h,s}} \left\{ \hat{\theta} \vert \Theta \right\}} \leq O(1) \max_{\theta \in \Theta} \| \theta \|_{S_{h,\infty}} n^{1/2} \sqrt{K}. \tag{4.131}
\]

For the sake of simplicity, assume that \( r \) in (4.129) 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 (4.130) on the minimax \( \| \cdot \|_{S_{h,s}} \)-risk reads
\[
\text{Risk}_{\text{opt.-ll}} \| \cdot \|_{S_{h,s}} \left\{ \Theta \right\} \geq O(1) \min \left\{ n^{1/2} \sqrt{K}, 1 \right\}.
\]

When \( K \) is “large:” \( K \geq n^{1+\frac{3}{2}} \), this lower bound matches, within an absolute constant factor, the upper bound (4.131) on the risk of the plug-in estimate, so that the latter estimate is near-optimal. When \( K < n^{1+\frac{3}{2}} \), the lower risk bound becomes \( O(1) \), so that here a nearly optimal estimate is the trivial estimate \( \hat{\theta}(\omega) \equiv I_n \).

### 4.7.7 Around \( S \)-Lemma

**\( S \)-Lemma** is a classical result of extreme importance in Semidefinite Optimization. Basically, the lemma states that when the ellipsoid \( \mathcal{X} \) in Proposition 4.6 is an ellipsoid, (4.19) can be strengthen to \( \text{Opt} = \text{Opt}_* \). In fact, \( S \)-Lemma is even stronger:

**Lemma 4.43.** [\( S \)-Lemma] Consider two quadratic forms \( f(x) = x^T Ax + 2a^T x + \alpha \) and \( g(x) = x^T Bx + 2b^T x + \beta \) 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
\[
\begin{bmatrix}
\lambda B - A \\
\lambda B - a^T \\
\lambda \beta - \alpha
\end{bmatrix} \geq 0
\]
in scalar variable $\lambda$ has a nonnegative solution.

Proof of $S$-Lemma can be found, e.g., in [15, 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

$$X = \{x \in \mathbb{R}^n : \exists t \in T : x^T S_k x \leq t k, 1 \leq k \leq K \}$$

where, as always, $S_k$ are positive semidefinite matrices with positive definite sum, and $T$ is a computationally tractable convex compact subset in $\mathbb{R}_k^+$ such that $t \in T$ implies $t' \in T$ whenever $0 \leq t' \leq t$ and $T$ contains a positive vector.

Exercise 4.25.

Let us associate with ellitope $X$ given by (4.132) the sets

$$X = \text{Conv}\{xx^T : x \in X\},$$

$$\hat{X} = \{Y \in \mathbb{S}^n : Y \succeq 0, \exists t \in T : \text{Tr}(S_k Y) \leq t k, 1 \leq k \leq K \},$$

so that $X, \hat{X}$ are convex compact sets containing the origin, and $\hat{X}$ is computationally tractable along with $T$. Prove that

1. When $K = 1$, we have $X = \hat{X}$;
2. We always have $X \subset \hat{X} \subset 3 \ln(\sqrt{3}K) X$.


For $x \in \mathbb{R}^n$ let $Z(x) = [x; 1][x; 1]^T$, $Z^o[x] = \begin{bmatrix} xx^T & x \\ x^T & 1 \end{bmatrix}$. Let

$$C = \begin{bmatrix} 1 \\ \vdots \end{bmatrix},$$

and let us associate with ellitope $X$ given by (4.132) the sets

$$X^+ = \text{Conv}\{Z^o[x] : x \in X\},$$

$$\hat{X}^+ = \{Y = \begin{bmatrix} U & u \\ u^T & 1 \end{bmatrix} \in \mathbb{S}^{n+1} : Y + C \succeq 0, \exists t \in T : \text{Tr}(S_k U) \leq t k, 1 \leq k \leq K \},$$

so that $X^+, \hat{X}^+$ are convex compact sets containing the origin, and $\hat{X}^+$ is computationally tractable along with $T$. Prove that

1. When $K = 1$, we have $X^+ = \hat{X}^+$;
2. We always have $X^+ \subset \hat{X}^+ \subset 3 \ln(\sqrt{3}(K + 1)) X^+$.

4.7.8 Miscellaneous exercises

Exercise 4.27.

Let $X \subset \mathbb{R}^n$ be a convex compact set, let $b \in \mathbb{R}^n$, 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 $B x = b^T x$ 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 $\rho$ such that

$$\text{Prob}\{\xi : b^T x - f^T (Ax + \sigma \xi) > \rho\} \leq \epsilon/2 \& \text{Prob}\{\xi : b^T x - f^T (Ax + \sigma \xi) < \rho\} \leq \epsilon/2 \forall x \in X.$$  

Pose the problem as convex optimization problem and consider in details the case where $X$ is the box $\{x \in \mathbb{R}^n : a_j |x_j| \leq 1, 1 \leq j \leq n\}$, where $a_j > 0$ for all $j$.

Exercise 4.28.

Prove Proposition 4.21.

Exercise 4.29.

Prove Proposition 4.22.

4.8 PROOFS

4.8.1 Preliminaries

4.8.1.1 Technical lemma

Lemma 4.44. Given basic spectrapo \(\mathcal{X} = \{x \in \mathbb{R}^n : \exists t \in \mathcal{T} : R_k^2 [x] \leq t_k I_{d_k}, 1 \leq k \leq K\}\) (4.133)

and a positive definite $n \times n$ matrix $Q$ and setting $\Lambda_k = \mathcal{R}_k [Q]$ (for notation, see Section 4.3.1), we get a collection of positive semidefinite matrices, and $\sum_k \mathcal{R}_k^* [\Lambda_k]$ is positive definite.

As a corollaries,

(i) whenever $M_k$, $k \leq K$, are positive definite matrices, the matrix $\sum_k \mathcal{R}_k^* [M_k]$ is positive definite;

(ii) the set $Q_T = \{Q \succeq 0 : \mathcal{R}_k [Q] \preceq T I_{d_k}, k \leq K\}$ is bounded for every $T$.

Proof. Let us prove the first claim, Assuming the opposite, we would be able to find a nonzero vector $y$ such that $\sum_k y^T \mathcal{R}_k^* [\Lambda_k] y < 0$, whence

$$0 \geq \sum_k y^T \mathcal{R}_k^* [\Lambda_k] y = \sum_k \text{Tr}(\mathcal{R}_k^* [\Lambda_k] [yy^T]) = \sum_k \text{Tr}(\Lambda_k \mathcal{R}_k [yy^T])$$

(we have used (4.26), (4.22)). Since $\Lambda_k = \mathcal{R}_k [Q] \succeq 0$ due to $Q \succeq 0$, see (4.23), it follows that $\text{Tr}(\Lambda_k \mathcal{R}_k [yy^T]) = 0$ for all $k$. Now, the linear mapping $\mathcal{R}_k [\cdot]$ is $\succeq$-monotone, and $Q$ is positive definite, implying that $Q \succeq r_k yy^T$ for some $r_k > 0$, whence $\Lambda_k \succeq r_k \mathcal{R}_k [yy^T]$, and therefore $\text{Tr}(\Lambda_k \mathcal{R}_k [yy^T]) = 0$ implies that $\text{Tr}(\mathcal{R}_k^2 [yy^T]) = 0$, that is, $\mathcal{R}_k [yy^T] = \mathcal{R}_k^2 [y] = 0$. Since $\mathcal{R}_k [\cdot]$ takes values in $S^{d_k}$, we get $\mathcal{R}_k [y] = 0$ for all $k$, which is impossible due to $y \neq 0$ and property S.3, see Section 4.3.1.

To verify (i), note that 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 (4.27), 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 $\sum_k \mathcal{R}_k^* [\Lambda_k]$.

To verify (ii), assume, on the contrary to what should be proved, that $Q_T$ is unbounded. Since $Q_T$ is closed and convex, it must possess a nonzero recessive
direction, that is, there should exist nonzero positive semidefinite matrix $D$ such that $R_k[D] \preceq 0$ for all $k$. Selecting positive definite matrices $M_k$, the matrices $R_k^*[M_k]$ are positive semidefinite (see Section 4.3.1), and their sum $S$ is positive definite by (i). We have

$$0 \geq \sum_k \text{Tr}(R_k[D]M_k) = \sum_k \text{Tr}(DR_k^*[M_k]) = \text{Tr}(DS),$$

where the first inequality is due to $M_k \succeq 0$, and the first equality is due to (4.26). The resulting inequality is impossible due to $0 \neq D \succeq 0$ and $S \succ 0$, which is a desired contradiction. 

4.8.1.2 Noncommutative Khintchine Inequality

We will use deep result from Functional Analysis ("Noncommutative Khintchine Inequality") due to Lust-Piquard [171], Pisier [196] and Buchholz [34], see [224, Theorem 4.6.1]:

**Theorem 4.45.** Let $Q_i \in S^n$, $1 \leq i \leq I$, and let $\xi_i$, $i = 1, \ldots, I$, be independent Rademacher ($\pm 1$ with probabilities $1/2$) or $\mathcal{N}(0,1)$ random variables. Then for all $t \geq 0$ one has

$$\text{Prob} \left\{ \left\| \sum_{i=1}^{I} \xi_i Q_i \right\| \geq t \right\} \leq 2n \exp \left\{ -\frac{t^2}{2v_Q} \right\},$$

where $\| \cdot \|$ is the spectral norm, and $v_Q = \left\| \sum_{i=1}^{I} Q_i \right\|$. We need the following immediate consequence of the theorem:

**Lemma 4.46.** Given spectratope (4.20), let $Q \in S^n_+$ be such that

$$R_k^*[Q] \preceq \rho t_k I_{d_k}, \quad 1 \leq k \leq K,$$

(4.134)

for some $t \in T$ and some $\rho \in (0,1)$. Then

$$\text{Prob}_{\xi \sim \mathcal{N}(0,Q)} \{ \xi \notin \mathcal{X} \} \leq \min \left[ 2De^{-\frac{D}{2D}}, 1 \right], \quad D := \sum_{k=1}^{K} d_k.$$  

**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

$$\sum_{i} [\bar{R}_{ki}]^2 = E_{\eta \sim \mathcal{N}(0,I_n)} \{ \bar{R}_{ki}^2[\eta] \} = E_{\xi \sim \mathcal{N}(0,Q)} \{ R_{ki}^2[\xi] \} = R_k^*[Q] \preceq \rho t_k I_{d_k}$$

due to (4.24). Hence, by Theorem 4.45

$$\text{Prob}_{\xi \sim \mathcal{N}(0,Q)} \{ \| R_k[\xi] \|^2 \geq t_k \} = \text{Prob}_{\eta \sim \mathcal{N}(0,I_n)} \{ \| \bar{R}_k[\eta] \| \geq t_k \} \leq 2d_k e^{-\frac{t_k}{2D}}.$$
We conclude that

$$\text{Prob}_{\xi \sim \mathcal{N}(0, Q)} \{ \xi \notin \mathcal{X} \} \leq \text{Prob}_{\xi \sim \mathcal{N}(0, Q)} \{ \exists k : \| R_k \xi \|^2 > t_k \} \leq 2D e^{-\frac{1}{2\rho}}.$$  \hfill \Box

The elliptic version of Lemma 4.46 is as follows:

**Lemma 4.47.** Given ellitope (4.9), let $Q \in \mathbb{S}^n_+$ be such that

$$\text{Tr}(R_k Q) \leq \rho t_k, \ 1 \leq k \leq K,$$

(4.135)

for some $t \in T$ and some $\rho \in (0, 1]$. Then

$$\text{Prob}_{\xi \sim \mathcal{N}(0, Q)} \{ \xi \notin \mathcal{X} \} \leq 2K \exp \left\{ -\frac{1}{3\rho} \right\}.$$

**Proof.** Observe that if $P \in \mathbb{S}^n_+$ satisfies $\text{Tr}(R) \leq 1$, we have

$$\mathbb{E}_{\eta \sim \mathcal{N}(0, I_n)} \left\{ \exp \left\{ \frac{1}{3} \eta^T P \eta \right\} \right\} \leq \sqrt{3}. \quad (4.136)$$

Indeed, we lose nothing when assuming that $P = \text{Diag}\{\lambda_1, ..., \lambda_n\}$ with $\lambda_i \geq 0$, $\sum \lambda_i \leq 1$. In this case

$$\mathbb{E}_{\eta \sim \mathcal{N}(0, I_n)} \left\{ \exp \left\{ \frac{1}{3} \eta^T P \eta \right\} \right\} = f(\lambda) := \mathbb{E}_{\eta \sim \mathcal{N}(0, I_n)} \left\{ \exp \left\{ \frac{1}{3} \sum \lambda_i \eta_i^2 \right\} \right\}.$$

Function $f$ is convex, so that its maximum on the simplex $\{ \lambda \geq 0 : \sum \lambda_i \leq 1 \}$ is achieved at a vertex, that is,

$$f(\lambda) \leq \mathbb{E}_{\eta \sim \mathcal{N}(0, 1)} \left\{ \exp \left\{ \frac{1}{3} \eta^2 \right\} \right\} = \sqrt{3}; \quad (4.136) \text{ is proved.}$$

Note that (4.136) implies that

$$\text{Prob}_{\eta \sim \mathcal{N}(0, I_n)} \left\{ \eta : \eta^T P \eta > s \right\} < \sqrt{3} \exp \left\{ -s/3 \right\}, \ s \geq 0. \quad (4.137)$$

Now let $Q$ and $t$ satisfy Lemma’s premise. Setting $\xi = Q^{1/2} \eta$, $\eta \sim \mathcal{N}(0, I_n)$, for $k \leq K$ such that $t_k > 0$ we have

$$\xi^T R_k \xi = \rho t_k \eta^T P_k \eta, \ P_k := [\rho t_k]^{-1} Q^{1/2} R_k Q^{1/2} \geq 0 \text{ & } \text{Tr}(P_k) = [\rho t_k]^{-1} \text{Tr}(Q R_k) \leq 1,$$

so that

$$\text{Prob}_{\xi \sim \mathcal{N}(0, Q)} \left\{ \xi : \xi^T R_k \xi > \rho s t_k \right\} = \text{Prob}_{\eta \sim \mathcal{N}(0, I_n)} \left\{ \eta^T P_k \eta > s \right\} < \sqrt{3} \exp \left\{ -s/3 \right\}, \quad (4.138)$$

where the inequality is due to (4.137). Relation (4.138) was established for $k$ with $t_k > 0$; it is trivially true when $t_k = 0$, since in this case $Q^{1/2} R_k Q^{1/2} = 0$ due to $\text{Tr}(Q R_k) \leq 0$ and $R_k, Q \in \mathbb{S}^n_+$. Setting $s = 1/\rho$, we get from (4.138) that

$$\text{Prob}_{\eta \sim \mathcal{N}(0, Q)} \left\{ \xi^T R_k \xi > t_k \right\} \leq \sqrt{3} \exp \left\{ -\frac{1}{3\rho} \right\}, \ k \leq K,$$

and (4.137) follows due to the union bound.  \hfill \Box
4.8.1.3 Anderson’s Lemma

Below we use a simply-looking, but by far nontrivial, fact

Anderson’s Lemma [4]. Let \( f \) be a nonnegative even (\( f(x) \equiv f(-x) \)) summable function on \( \mathbb{R}^N \) such that the level sets \( \{ x : f(x) \geq t \} \) are convex for all \( t \) and let \( X \subset \mathbb{R}^N \) be a symmetric w.r.t. the origin closed convex set. Then for every \( y \in \mathbb{R}^N \)

\[
\int_{X+ty} f(z) \, dz
\]

is a nonincreasing function of \( t \geq 0 \). In particular, if \( \zeta \) is a zero mean \( N \)-dimensional Gaussian random vector, then for every \( y \in \mathbb{R}^N \)

\[
\text{Prob}\{\zeta \notin y + X\} \geq \text{Prob}\{\zeta \notin X\}
\]

Hence, for every norm \( \| \cdot \| \) on \( \mathbb{R}^N \) it holds

\[
\text{Prob}\{\zeta : \|\zeta - y\| > \rho\} \geq \text{Prob}\{\zeta : \|\zeta\| > \rho\} \quad \forall (y \in \mathbb{R}^N, \rho \geq 0).
\]

4.8.2 Proof of Proposition 4.6

1°. We need the following

Lemma 4.48. Let \( S \) be a positive semidefinite \( N \times N \) matrix with trace \( \leq 1 \) and \( \xi \) be \( N \)-dimensional Rademacher random vector (i.e., the entries in \( \xi \) are independent and take values \( \pm 1 \) with probabilities \( 1/2 \)). Then

\[
\mathbb{E}\left\{ \exp\left\{ \frac{1}{3} \xi^T S \xi \right\} \right\} \leq \sqrt{3},
\]

implying that

\[
\text{Prob}\{\xi^T S \xi > s\} \leq \sqrt{3}\exp\{-s/3\}, \quad s \geq 0.
\]

Proof. Let \( S = \sum \sigma_i |h^i|^T h^i \) be the eigenvalue decomposition of \( S \), so that \( |h^i|^T h^i = 1 \), \( \sigma_i \geq 0 \), and \( \sum \sigma_i \leq 1 \). The function

\[
F(\sigma_1, \ldots, \sigma_n) = \mathbb{E}\left\{ e^{\frac{1}{2} \sum \sigma_i \xi^T [h^i]^T \xi} \right\}
\]

is convex on the simplex \( \{ \sigma \geq 0, \sum \sigma_i \leq 1 \} \) and thus attains it maximum over the simplex at a vertex, implying that for some \( f = h^i \), \( f^T f = 1 \), it holds

\[
\mathbb{E}\{e^{\frac{1}{2} \xi^T S \xi}\} \leq \mathbb{E}\{e^{\frac{1}{2} (f^T \xi)^2}\}.
\]
Let $\xi \sim \mathcal{N}(0, 1)$ be independent of $\xi$. We have
\[
\begin{aligned}
E_\xi \left\{ \exp\left( \frac{1}{3} (f^T \xi)^2 \right) \right\} &= E_\xi \left\{ E_\xi \left\{ \exp\left( \sqrt{2/3} f^T \xi \right) \right\} \right\} \\
&= E_\xi \left\{ \exp\left( \sqrt{2/3} f^T \xi \right) \right\} \\
&= E_\xi \left\{ \prod_{j=1}^N \exp\left( \sqrt{2/3} \xi f_j \right) \right\} \\
&= E_\xi \left\{ \exp\left( \frac{1}{3} \xi^2 \right) \right\} \leq \sqrt{3}.
\end{aligned}
\]
\[\square\]

2'. The right inequality in (4.19) has been justified in Section 4.2.3. To prove the left inequality in (4.19), let $T$ be the closed conic hull of $T$ (see Section 4.1.1), and let us consider the conic problem
\[
\text{Opt}_\# = \max_{Q,t} \left\{ \text{Tr}(P^T C P Q) : Q \succeq 0, \text{Tr}(Q R_k) \leq t_k \forall k, [t;1] \in T \right\}.
\]
We claim that
\[
\text{Opt} = \text{Opt}_\#.
\]
Indeed, (4.139) clearly is a strictly feasible and bounded conic problem, so that its optimal value is equal to the optimal value of its conic dual (Conic Duality Theorem). Taking into account that the cone $T$, dual to $T$ is $\{[g;s] : s \geq \phi_T(-g)\}$, see Section 4.1.1, we therefore get
\[
\begin{aligned}
\text{Opt}_\# &= \min_{\lambda, [g,s], L} \left\{ \text{Tr}((\sum_k \lambda_k R_k - L)Q) - \sum_k \lambda_k [g_k + g_k] t_k = \text{Tr}(P^T C P Q) \forall (Q,t), \text{Tr}(L) \geq 0 \right\} \\
&= \min_{\lambda, [g,s], L} \left\{ \text{Tr}(P^T C P R) = \text{Tr}(Q^T P^T C P) \forall (Q,t), \text{Tr}(L) \geq 0 \right\} \\
&= \min_{\lambda} \left\{ \phi_T(\lambda) : \sum_k \lambda_k R_k \succeq \text{Tr}(P^T C P), \lambda \geq 0 \right\} = \text{Opt},
\end{aligned}
\]
as claimed.

3'. With Lemma 4.48 and (4.140) at our disposal, we can now complete the proof of Proposition 4.6 by adjusting the technique from [187]. Specifically, problem (4.139) clearly is solvable; let $Q_*, t^*$ be an optimal solution to the problem. Next, let us set $R_* = Q_*^{1/2}$, $C = R_* P^T C P R_*$, let $C = U D U^T$ be the eigenvalue decomposition of $C$, and let $R_k = U^T R_k R_k U$. Observe that
\[
\begin{aligned}
\text{Tr}(C) &= \text{Tr}(R_* P^T C P R_*) = \text{Tr}(Q_* P^T C P) = \text{Opt}_\# = \text{Opt}, \\
\text{Tr}(R_k) &= \text{Tr}(R_* R_k R_*) = \text{Tr}(Q_* R_k) \leq t_k^*.
\end{aligned}
\]
Now let $\xi$ be Rademacher random vector. For $k$ with $t_k^* > 0$, applying Lemma 4.48 to matrices $R_k/t_k^*$, we get for $s > 0$
\[
\text{Prob} \left\{ \xi^T R_k \xi > s t_k^* \right\} \leq \sqrt{3} \exp\left\{ -s/3 \right\};
\]
if $k$ is such that $t_k^* = 0$, we have $\text{Tr}(R_k) = 0$, that is, $R_k = 0$ (since $R_k \succeq 0$), and (4.141) holds true as well. Now let
\[
s_* = 3 \ln(\sqrt{3}K),
\]
so that \( \sqrt{3} \exp\{-s/3\} < 1/K \) when \( s > s_* \). The latter relation combines with (4.141) to imply that for every \( s > s_* \) there exists a realization \( \xi \) of \( \xi \) such that

\[
\xi^T \bar{R}_k \xi \leq s t_k^* \forall k.
\]

Let us set \( \bar{y} = \frac{1}{\sqrt{3}} R_* U \xi \). Then

\[
\bar{y}^T R_k \bar{y} = s^{-1} \xi^T U^T R_* R_k R_* U \xi = s^{-1} \xi^T \bar{R}_k \xi \leq t_k^* \forall k
\]

implying that \( \bar{x} := P \bar{y} \in \mathcal{X} \), and

\[
\bar{x}^T C \bar{x} = s^{-1} \xi^T U^T R_* P^T C P R_* U \xi = s^{-1} \xi^T D \xi = s^{-1} \text{Tr}(D) = s^{-1} \text{Tr}(\bar{C}) = s^{-1} \text{Opt}.
\]

Thus, \( \text{Opt}_* := \max_{x \in \mathcal{X}} x^T C x \geq s^{-1} \text{Opt} \) whenever \( s > s_* \), which implies the left inequality in (4.19). \( \square \)

### 4.8.3 Proof of Proposition 4.8

Proof follows the lines of the proof of Proposition 4.6. First, passing from \( C \) to the matrix \( \bar{C} = P^T C P \), the situation clearly reduces to the one where \( P = I \). To save notation, in the rest of the proof we assume that \( P \) is the identity.

Second, from Lemma 4.44 and the fact that the level sets of \( \phi_T(\cdot) \) on the non-negative orthant are bounded (since \( T \) contains a positive vector) it immediately follows that problem (4.29) is feasible with bounded level sets of the objective, so that the problem is solvable. The left inequality in (4.30) was proved in Section 4.3.2. Thus, all we need is to prove the right inequality in (4.30).

1°. Let \( \mathbf{T} \) be the closed conic hull of \( T \) (see Section 4.1.1). Consider the conic problem

\[
\text{Opt}_\# := \max_{Q,t} \{ \text{Tr}(CQ) : Q \succeq 0, R_k[Q] \preceq t_k I_{d_k} \forall k \}, [t; 1] \in \mathbf{T} \}.
\]

(4.142)

This problem clearly is strictly feasible; by Lemma 4.44, the feasible set of the problem is bounded, so the problem is solvable. We claim that

\[
\text{Opt}_\# = \text{Opt}.
\]

Indeed, (4.142) is a strictly feasible and bounded conic problem, so that its optimal value is equal to the one in its conic dual, that is,

\[
\text{Opt}_\# = \min_\Lambda \left\{ \text{Tr}\left(\sum_k \mathcal{R}_k^\Lambda[\Lambda_k] - L[Q]\right) - \sum_k \text{Tr}(\Lambda_k) + g_k [t_k] \right\}_{\Lambda = \Lambda_k \in \mathcal{K}, [g; s], L}^{\forall \Lambda_n \geq 0 \forall k, L \succeq 0, s \geq \phi_T(-g)}
\]

\[
= \min_{\Lambda, [g; s], L} \left\{ s : \sum_k \mathcal{R}_k^\Lambda[\Lambda_k] - L = C, g = -\lambda[\Lambda], \right\}_{\forall \Lambda_k \geq 0 \forall k, L \succeq 0, s \geq \phi_T(-g)}
\]

\[
= \min_{\Lambda} \left\{ \phi_T(\lambda[\Lambda]) : \sum_k \mathcal{R}_k^\Lambda[\Lambda_k] \geq C, \Lambda_k \geq 0 \forall k \right\} = \text{Opt},
\]

as claimed.

2°. Problem (4.142), as we already know, is solvable; let \( Q_*, t^* \) be an optimal
solution to the problem. Next, let us set $R_\ast = Q_\ast^{1/2}, \hat{C} = R_\ast C R_\ast$, and let $\hat{C} = U D U^T$ be the eigenvalue decomposition of $\hat{C}$, so that the matrix $D = U^T R_\ast C R_\ast U$ is diagonal, and the trace of this matrix is $\text{Tr}(R_\ast C R_\ast) = \text{Tr}(CQ_\ast) = \text{Opt}_\# = \text{Opt}$.

Now let $V = R_\ast U$, and let $\xi = V \eta$, where $\eta$ is $n$-dimensional random Rademacher vector (independent entries taking values $\pm 1$ with probabilities $1/2$). We have

$$\xi^T C \xi = \eta^T [V^T C V] \eta = \eta^T [U^T R_\ast C R_\ast U] \eta = \eta^T D \eta \equiv \text{Tr}(D) = \text{Opt} \quad (4.143)$$

(recall that $D$ is diagonal) and

$$E_{\xi} \{ \xi \xi^T \} = E_{\eta} \{ V \eta \eta^T V \} = V V^T = R_\ast U U^T R_\ast = R_\ast^2 = Q_\ast.$$

From the latter relation,

$$E_{\xi} \{ R_k^2 [\xi] \} = E_{\eta} \{ R_k [\xi \xi^T] \} = R_k [E_{\xi} \{ \xi \xi^T \}] = R_k [Q_\ast] \leq t_k^2 I_{d_k}, \, 1 \leq k \leq K. \quad (4.144)$$

On the other hand, with properly selected symmetric matrices $\bar{R}^{kj}$ we have

$$R_k [Vy] = \sum_i y_i \bar{R}^{ki}$$

identically in $y \in \mathbb{R}^n$, whence

$$E_{\xi} \{ R_k^2 [\xi] \} = E_{\eta} \{ R_k^2 [V \eta] \} = E_{\eta} \left\{ \left[ \sum_i \eta_i \bar{R}^{ki} \right]^2 \right\} = \sum_{i,j} E_{\eta} \{ \eta_i \eta_j \} \bar{R}^{ki} \bar{R}^{kj} = \sum_i (\bar{R}^{ki})^2.$$

This combines with (4.144) to imply that

$$\sum_i (\bar{R}^{ki})^2 \leq t_k^2 I_{d_k}, \, 1 \leq k \leq K. \quad (4.145)$$

3°. Let us fix $k \leq K$. Assuming $t_k^* > 0$ and applying Theorem 4.45, we derive from (4.145) that

$$\text{Prob}\{ \eta : \| R_k [\eta] \|^2 > t_k^*/\rho \} < 2 d_k e^{-\frac{t_k^*}{\rho}},$$

and recalling the relation between $\xi$ and $\eta$, we arrive at

$$\text{Prob}\{ \xi : \| R_k [\xi] \|^2 > t_k^*/\rho \} < 2 d_k e^{-\frac{t_k^*}{\rho}} \quad \forall \rho \in (0, 1). \quad (4.146)$$

Note that when $t_k^* = 0$ (4.145) implies $\bar{R}^{ki} = 0$ for all $i$, so that $R_k [\xi] \equiv \hat{R}_k [\eta] \equiv 0$, and (4.146) holds for those $k$ as well.

Now let us set $\rho = \frac{1}{2 \max_{i} (2D_{ii})^{1/2}}$. For this $\rho$, the sum over $k \leq K$ of the right hand sides in inequalities (4.146) is $\leq 1$, implying that there exists a realization $\xi$ of $\xi$ such that

$$\| R_k [\xi] \|^2 \leq t_k^*/\rho, \quad \forall k,$$

or, equivalently,

$$\bar{x} := \rho^{1/2} \xi \in \mathcal{X}$$

(recall that $P = I$), implying that

$$\text{Opt}_\ast := \max_{x \in \mathcal{X}} x^T C x \geq \bar{x}^T C \bar{x} = \rho^{T} C \xi = \rho \text{Opt}$$
(the concluding equality is due to (4.143)), and we arrive at the right inequality in (4.30).

4.8.4 Proof of Lemma 4.17

1°. Let us verify (4.57). When \( Q \succ 0 \), passing from variables \((\Theta, \Upsilon)\) in problem (4.56) to the variables \((G = Q^{1/2}\Theta Q^{1/2}, \Upsilon)\), the problem becomes exactly the optimization problem in (4.57), implying that \( \text{Opt}[Q] = \bar{\text{Opt}}[Q] \) when \( Q \succ 0 \). As it is easily seen, both sides in this equality are continuous in \( Q \succeq 0 \), and (4.57) follows.

2°. Let us prove (4.59). Setting \( \zeta = Q^{1/2}\eta \) with \( \eta \sim \mathcal{N}(0, I_N) \) and \( Z = Q^{1/2}Y \), to justify (4.59) we have to show that when \( \kappa \geq 1 \) one has

\[
\bar{\delta} = \frac{\text{Opt}[Q]}{4\kappa} \Rightarrow \text{Prob}_\eta\{\|Z^T\eta\| \geq \bar{\delta}\} \geq \beta_\kappa := 1 - \frac{e^{3/8}}{2} - 2Fe^{-\kappa^2/2}, \quad (4.147)
\]

where (cf. (4.57))

\[
\text{Opt}[Q] = \min_{\Theta, \Upsilon = (Y, \ell \leq L)} \left\{ \phi_R(\lambda[\Upsilon]) + \text{Tr}(\Theta) : \begin{bmatrix} \Theta \\ \frac{1}{2}M^T Z^T \end{bmatrix} \begin{bmatrix} \frac{1}{2}ZM \\ \sum_{r} S_r[\Upsilon \ell] \end{bmatrix} \succeq 0 \right\}. \quad (4.148)
\]

Justification of (4.147) is as follows.

2.1°. Let us represent \( \text{Opt}[Q] \) as the optimal value of a conic problem. Setting \( K = \text{cl}\{[r; s] : s > 0, r/s \in \mathcal{R}\} \), we ensure that

\[
\mathcal{R} = \{r : [r; 1] \in K\}, \quad K_* = \{[g; s] : s \geq \phi_R(-g)\},
\]

where \( K_* \) is the cone dual to \( K \). Consequently, (4.148) reads

\[
\text{Opt}[Q] = \min_{\Theta, \Upsilon, \theta} \left\{ \begin{array}{c}
\Upsilon_\ell \succeq 0, 1 \leq \ell \leq L \\
\frac{1}{2}M^T Z^T \sum_{r} S_r[\Upsilon \ell] \succeq 0
\end{array} \right\}. \quad (P)
\]

2.2°. Now let us prove that there exist a matrix \( W \in S_+^L \) and \( r \in \mathcal{R} \) such that

\[
S_\ell[W] \preceq r_\ell I_f, \ell \leq L, \quad (4.149)
\]

and

\[
\text{Opt}[Q] \leq \sum_1 \sigma_1(ZMW^{1/2}), \quad (4.150)
\]

where \( \sigma_1(\cdot) \geq \sigma_2(\cdot) \geq \ldots \) are singular values.
To get the announced result, let us pass from problem \((P)\) to its conic dual. Applying Lemma 4.44 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 \(\text{Opt}[Q]\). Let us build \((D)\). Denoting by \(\Lambda_\ell \geq 0, \ell \leq L\), \(\begin{bmatrix} G & -R \\ -R^T & W \end{bmatrix} \succeq 0\), \([r; \tau] \in K\) 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:

\[
\text{Tr}(\Theta G) + \text{Tr}(W \sum_\ell \mathbf{S}_\ell^*[\Upsilon_\ell]) + \sum_\ell \text{Tr}(\Lambda_\ell \Upsilon_\ell) - \sum_\ell r_\ell \text{Tr}(\Upsilon_\ell) + \theta \tau \geq \text{Tr}(ZMR^T).
\]

To get the dual problem, we impose on the Lagrange multipliers, in addition to the initial conic constraints like \(\Lambda_\ell \geq 0, 1 \leq \ell \leq L\), the restriction that the left hand side in the aggregated constraint, identically in \(\Theta, \Upsilon_\ell\) and \(\theta, \tau\), is equal to the objective of \((P)\), that is,

\[
G = I, \mathbf{S}_\ell[W] + \Lambda_\ell - r_\ell I_{f_\ell} = 0, 1 \leq \ell \leq L, \tau = 1,
\]

and maximize, under the resulting restrictions, the right-hand side of the aggregated constraint. After immediate simplifications, we arrive at

\[
\text{Opt}[Q] = \max_{W, r, \tau, \ell} \left\{ \text{Tr}(ZMR^T) : W \succeq R^T R, r \in R, \mathbf{S}_\ell[W] \preceq r_\ell I_{f_\ell}, 1 \leq \ell \leq L \right\}
\]

(note that \(r \in R\) is equivalent to \([r; 1] \in K\), and \(W \succeq R^T R\) is the same as \(\begin{bmatrix} I & -R \\ -R^T & W \end{bmatrix} \succeq 0\)). Now, to say that \(R^T R \preceq W\) is exactly the same as to say that \(R = SW^{1/2}\) with the spectral norm \(\|S\|_{2,2}\) of \(S\) not exceeding 1, so that

\[
\text{Opt}[Q] = \max_{W, r, \ell} \left\{ \text{Tr}(ZM[S^{1/2}]^T) : W \succeq 0, \|S\|_{2,2} \leq 1, r \in R, \mathbf{S}_\ell[W] \preceq r_\ell I_{f_\ell}, 1 \leq \ell \leq L \right\}.
\]

We can immediately eliminate the \(S\)-variable, using the well-known fact that for a \(p \times q\) matrix \(J\) it holds

\[
\max_{S \in \mathbb{R}^{p \times q}, \|S\|_{2,2} \leq 1} \text{Tr}(JS^T) = \|J\|_{S_{1,1}},
\]

where \(\|J\|_{S_{1,1}}\) is the nuclear norm (the sum of singular values) of \(J\). We arrive at

\[
\text{Opt}[Q] = \max_{W, r} \left\{ \|ZM[S^{1/2}]_{\text{sh},1}\|_{S_{1,1}} : r \in R, W \succeq 0, \mathbf{S}_\ell[W] \preceq r_\ell I_{f_\ell}, 1 \leq \ell \leq L \right\}.
\]

The resulting problem clearly is solvable, and its optimal solution \(W\) ensures the target relations \((4.149)\) and \((4.150)\).

2.3. Given \(W\) satisfying \((4.149)\) and \((4.150)\), let \(UJV = W^{1/2}M^T Z^T\) be the singular value decomposition of \(W^{1/2}M^T Z^T\), so that \(U\) and \(V\) are, respectively, \(q \times q\) and \(N \times N\) orthogonal matrices, \(J\) is \(q \times N\) matrix with diagonal \(\sigma = [\sigma_1; \ldots; \sigma_p]\), \(p = \min[q, N]\), and zero off-diagonal entries; the diagonal entries \(\sigma_i, 1 \leq i \leq p\) are the singular values of \(W^{1/2}M^T Z^T\), or, which is the same, of \(ZMW^{1/2}\). Therefore, by
(4.150) we have
\[ \sum_{i} \sigma_i \geq \text{Opt}[Q]. \] (4.151)

Now consider the following construction. Let \( \eta \sim \mathcal{N}(0, I_N) \); we denote by \( v \) the vector comprised of the first \( p \) entries in \( V\eta \); note that \( v \sim \mathcal{N}(0, I_p) \), since \( V \) is orthogonal. We then augment, if necessary, \( v \) by \( q - p \) independent of each other and of \( \eta \sim \mathcal{N}(0,1) \) random variables to obtain a \( q \)-dimensional random vector \( v' \sim \mathcal{N}(0, I_q) \), and set \( \chi = Uv' \). Because \( U \) is orthogonal we also have \( \chi \sim \mathcal{N}(0, I_q) \).

Observe that
\[ \chi^T W^{1/2} M^T Z^T \eta = \chi^T UJV\eta = \left[ \sum_{i=1}^{p} \sigma_i \right] \chi^T v_i^2. \] (4.152)

To continue we need two simple observations.

(i) One has
\[ \alpha := \text{Prob} \left\{ \sum_{i=1}^{p} \sigma_i v_i^2 < \frac{1}{4} \sum_{i=1}^{p} \sigma_i \right\} \leq \frac{e^{3/8}}{2} \approx 0.7275... . \] (4.153)

The claim is evident when \( \sigma := \sum_{i} \sigma_i = 0 \). Now let \( \sigma > 0 \), and let us apply the Cramer bounding scheme. Namely, given \( \gamma > 0 \), consider the random variable
\[ \omega = \exp \left\{ \frac{1}{4}\gamma \sum_{i} \sigma_i - \gamma \sum_{i} \sigma_i v_i^2 \right\}. \]

Note that \( \omega > 0 \) a.s., and is \( >1 \) when \( \sum_{i=1}^{p} \sigma_i v_i^2 < \frac{1}{4} \sum_{i=1}^{p} \sigma_i \), so that \( \alpha \leq \mathbb{E}\{\omega\} \), or, equivalently, thanks to \( v \sim \mathcal{N}(0, I_p) \),
\[ \ln(\alpha) \leq \ln(\mathbb{E}\{\omega\}) = \frac{1}{4}\gamma \sum_{i} \sigma_i + \sum_{i} \ln(\mathbb{E}\{\exp\{-\gamma \sigma_i v_i^2\}\}) \]
\[ \leq \frac{1}{4}\gamma \sigma - \frac{1}{2} \sum_{i} \ln(1 + 2\gamma \sigma_i). \]

Function \( -\sum_{i} \ln(1 + 2\gamma \sigma_i) \) is convex in \( \sigma_1; \ldots; \sigma_p \geq 0 \), 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
\[ \ln(\alpha) \leq \frac{1}{4}\gamma \sigma - \frac{1}{2} \ln(1 + 2\gamma \sigma). \]

Minimizing the right hand side in \( \gamma > 0 \), we arrive at (4.153).

(ii) Whenever \( \kappa \geq 1 \), one has
\[ \text{Prob}\{\|MW^{1/2}\chi\|_\ast > \kappa\} \leq 2 F \exp\{-\kappa^2/2\}, \] (4.154)

with \( F \) given by (4.55).

Indeed, setting \( \rho = 1/\kappa^2 \leq 1 \) and \( \omega = \sqrt{r}W^{1/2} \chi \), we get \( \omega \sim \mathcal{N}(0, \rho W) \). Let us apply Lemma 4.46 to \( Q = \rho W, \mathcal{R} \) 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
\[ \mathcal{Y} := \{ y : \exists r \in \mathcal{R} : S^2\ell[y] \leq r I_{\ell}, \ell \leq L \}, \]

we have \( \mathcal{S}_\ell[Q] = \rho \mathcal{S}_\ell[W] \leq \rho r I_{\ell}, \ell \leq L, \) with \( r \in \mathcal{R} \) (see (4.149)), so we are
under the premise of Lemma 4.46 (with \(Y\) in the role of \(X\) and thus with \(F\) in the role of \(D\)). Applying the lemma, we conclude that

\[
\text{Prob}\left\{\chi : \kappa^{-1}W^{1/2}\chi \notin Y\right\} \leq 2F \exp\left\{-1/(2\rho)\right\} = 2F \exp\{-\kappa^2/2\}.
\]

Recalling that \(B_* = MY\), we see that \(\text{Prob}\left\{\chi : \kappa^{-1}MW^{1/2}\chi \notin B_*\right\}\) is indeed upper-bounded by the right hand side of (4.154), and (4.154) follows.

2.4°. Now, for \(\kappa \geq 1\), let

\[
E_*^\kappa = \left\{(\chi, \eta) : \|MW^{1/2}\chi\|_* \leq \kappa, \sum_i \sigma_i v_i^2 \geq \frac{1}{4} \sum_i \sigma_i \right\},
\]

and let \(E_*^{\kappa^+} = \{\eta : \exists \chi : (\chi, \eta) \in E_*^\kappa\}\). For \(\eta \in E_*^{\kappa^+}\) there exists \(\chi\) such that \((\chi, \eta) \in E_*^\kappa\), so that

\[
\kappa\|Z^T\eta\| \geq \|MW^{1/2}\chi\|_*\|Z^T\eta\| \geq \chi^T W^{1/2} M^T Z^T \eta = \sum_i \sigma_i v_i^2 \geq \frac{1}{4} \sum_i \sigma_i \geq \frac{1}{4}\text{Opt}[Q]
\]

(we have used (4.152) and (4.151)). Thus,

\[
\eta \in E_*^\kappa \Rightarrow \|Z^T\eta\| \geq \frac{\text{Opt}[Q]}{4\kappa}.
\]

On the other hand, due to (4.153) and (4.154), for our random \((\chi, \eta)\) it holds

\[
\text{Prob}\{E_*^\kappa\} \geq 1 - \frac{e^{\kappa^2/8}}{2} - 2F e^{-\kappa^2/2} = \beta_*
\]

and the marginal distribution of \(\eta\) is \(\mathcal{N}(0, I_N)\), implying that

\[
\text{Prob}_{\eta \sim \mathcal{N}(0, I_N)}\{\eta \in E_*^{\kappa^+}\} \geq \beta_*.
\]

(4.147) is proved.

3°. As it was explained in the beginning of item 2°, (4.147) is exactly the same as (4.59). The latter relation clearly implies (4.60) which, in turn, implies the right inequality in (4.58).

4.8.5 Proofs of Propositions 4.5, 4.16, 4.19

Below, we focus on the proof of Proposition 4.16; Propositions 4.5, 4.19 will be derived from it in Sections 4.8.5.2, 4.8.6.2, respectively.

4.8.5.1 Proof of Proposition 4.16

In what follows, we use the assumptions and the notation of Proposition 4.16.

1°. Let

\[
\Phi(H; \Lambda, \Upsilon, \Upsilon'; \Theta) = \phi_T(\lambda[\Lambda]) + \phi_R(\lambda[\Upsilon]) + \phi_R(\lambda[\Upsilon']) + \text{Tr}(Q\Theta) : \mathcal{M} \times \Pi \rightarrow \mathbb{R},
\]
where

\[ \mathcal{M} = \left\{ (H, \Lambda, Y, Y', \Theta) : \begin{array}{l}
\Lambda = \{ \Lambda_k \geq 0, k \leq K \}, \\
Y = \{ Y_\ell \geq 0, \ell \leq L \}, \\
Y' = \{ Y'_\ell \geq 0, \ell \leq L \} \\
\frac{1}{2} M^T H \frac{1}{2} M^T [B - H^T A] \Theta \sum_k \Lambda_k \geq 0, \\
\frac{1}{2} M^T H \sum_k \Lambda_k \geq 0 \\
\frac{1}{2} M^T H \sum_k \Lambda_k \geq 0 \\
\Theta \sum_k \Lambda_k \geq 0 \end{array} \right\} \]

Looking at (4.42), we see immediately that the optimal value Opt in (4.42) is nothing but

\[
\text{Opt} = \min_{(H, \Lambda, Y, Y', \Theta) \in \mathcal{M}} \Phi(H, \Lambda, Y, Y', \Theta) := \max_{Q \in \Pi} \Phi(H, \Lambda, Y, Y', \Theta; Q). \tag{4.155}
\]

Note that sets \( \mathcal{M} \) and \( \Pi \) are closed and convex, \( \Pi \) is compact, and \( \Phi \) is a continuous convex-concave function on \( \mathcal{M} \times \Pi \). In view of these observations, the fact that \( \Pi \subset \text{int} \mathbb{S}^n_+ \) combines with the Sion-Kakutani Theorem to imply that \( \Phi \) possesses saddle point \((H_*, \Lambda_*, Y_*, Y'_*, \Theta_*, Q_*)\) (min in \((H, \Lambda, Y, Y', \Theta)\), max in \(Q\)) on \( \mathcal{M} \times \Pi \), whence \( \text{Opt} \) is the saddle point value of \( \Phi \) by (4.155). We conclude that for properly selected \( Q_* \in \Pi \) it holds

\[
\text{Opt} = \min_{(H, \Lambda, Y, Y', \Theta) \in \mathcal{M}} \Phi(H, \Lambda, Y, Y', \Theta; Q_*)
\]

\[
= \min_{H, \Lambda, Y, Y', \Theta} \left\{ \phi_r(\lambda[A]) + \phi_R(\lambda[Y]) + \phi_R(\lambda[Y']) + \text{Tr}(Q, \Theta) : \begin{array}{l}
\Lambda = \{ \Lambda_k \geq 0, k \leq K \}, \\
Y = \{ Y_\ell \geq 0, \ell \leq L \}, \\
Y' = \{ Y'_\ell \geq 0, \ell \leq L \} \\
\frac{1}{2} M^T [B - H^T A] \sum_k \Lambda_k \geq 0, \\
\frac{1}{2} M^T H \sum_k \Lambda_k \geq 0 \\
\Theta \sum_k \Lambda_k \geq 0 \end{array} \right\}
\]

\[
= \min_{H, \Lambda, Y, Y', \Theta} \left\{ \phi_r(\lambda[A]) + \phi_R(\lambda[Y]) + \phi_R(\lambda[Y']) + \text{Tr}(G) : \begin{array}{l}
\Lambda = \{ \Lambda_k \geq 0, k \leq K \}, \\
Y = \{ Y_\ell \geq 0, \ell \leq L \}, \\
Y' = \{ Y'_\ell \geq 0, \ell \leq L \} \\
\frac{1}{2} M^T [B - H^T A] \sum_k \Lambda_k \geq 0, \\
\frac{1}{2} M^T H \sum_k \Lambda_k \geq 0 \\
G \sum_k \Lambda_k \geq 0 \end{array} \right\}
\]

\[
= \min_{H, \Lambda, Y, Y', \Theta} \left\{ \phi_r(\lambda[A]) + \phi_R(\lambda[Y]) + \phi_R(\lambda[Y']) + \bar{\Psi}(H) : \begin{array}{l}
\Lambda = \{ \Lambda_k \geq 0, k \leq K \}, \\
Y = \{ Y_\ell \geq 0, \ell \leq L \} \\
\frac{1}{2} M^T [B - H^T A] \sum_k \Lambda_k \geq 0 \end{array} \right\}, \tag{4.156}
\]

where

\[
\bar{\Psi}(H) := \min_{G, Y'} \left\{ \phi_R(\lambda[Y']) + \text{Tr}(G) : Y' = \{ Y'_\ell \geq 0, \ell \leq L \}, \\
\frac{1}{2} M^T H \sum_k \Lambda_k \geq 0 \right\}
\]

and Opt is given by (4.42), and the equalities are due to (4.56) and (4.57).
From now on we assume that the noise $\xi$ in observation (4.31) is $\xi \sim \mathcal{N}(0,Q_\eta)$. We also assume that $B \neq 0$, since otherwise the conclusion of Proposition 4.16 is evident.

2'. $\epsilon$-risk. In Proposition 4.16, we are speaking about $\| \cdot \|$-risk of an estimate -- the maximal, over signals $x \in \mathcal{X}$, expected norm $\| \cdot \|$ of the error in recovering $Bx$: what we need to prove that the minimax optimal risk $\text{RiskOpt}_\|\cdot\|_\|\| [\mathcal{X}]$ as given by (4.53) can be lower-bounded by a quantity “of order of” $\text{Opt}$. To this end, of course, it suffices to build such a lower bound for the quantity

$$\text{RiskOpt}_\|\cdot\| := \inf_{\hat{x}(\cdot)} \mathbb{E}_{x \sim \mathcal{N}(0,Q_\eta)} \left\{ \| Bx - \hat{x}(Ax + \xi) \| \right\},$$

since this quantity is a lower bound on $\text{RiskOpt}_\|\cdot\|_\|\|$. Technically, it is more convenient to work with the $\epsilon$-risk defined in terms of “$\| \cdot \|$-confidence intervals” rather than in terms of the expected norm of the error. Specifically, in the sequel we will heavily use the minimax $\epsilon$-risk defined as

$$\text{RiskOpt}_{\|\cdot\|,\epsilon} := \inf_{\hat{x},\rho} \left\{ \rho : \Pr_{x \sim \mathcal{N}(0,Q_\eta)} \{ \| Bx - \hat{x}(Ax + \xi) \| > \rho \} \leq \epsilon \ \forall x \in \mathcal{X} \right\},$$

where $\hat{x}$ in the infimum runs through the set of all Borel estimates. When $\epsilon \in (0,1)$ is once and forever fixed (in the sequel, we use $\epsilon = \frac{1}{2}$) we can use $\epsilon$-risk to lower-bound $\text{RiskOpt}_\|\cdot\|$, since by evident reasons

$$\text{RiskOpt}_\|\cdot\| \geq \epsilon \cdot \text{RiskOpt}_{\|\cdot\|,\epsilon}. \quad (4.157)$$

Consequently, all we need in order to prove Proposition 4.16 is to lower-bound $\text{RiskOpt}_{\|\cdot\|,\bar{\epsilon}}$ by a “not too small” multiple of $\text{Opt}$, and this is our current objective.

3'. Let $W$ be a positive semidefinite $n \times n$ matrix, let $\eta \sim \mathcal{N}(0,W)$ be random signal, and let $\xi \sim \mathcal{N}(0,Q_\eta)$ be independent of $\eta$: vectors $(\eta,\xi)$ induce random vector

$$\omega = A\eta + \xi \sim \mathcal{N}(0,AWA^T + Q_\eta).$$

Consider the Bayesian version of the estimation problem where given $\omega$ we are interested to recover $B\eta$. Recall that, because $[\omega;B\eta]$ is zero mean Gaussian, the conditional expectation $\mathbb{E}_{\omega} [B\eta]$ of $B\eta$ given $\omega$ is linear in $\omega$: $\mathbb{E}_{\omega} [B\eta] = \tilde{H}^T \omega$ for some $\tilde{H}$ depending on $W$ only.\(^{24}\) Therefore, denoting by $\Pi^\omega$ conditional, $\omega$ given, probability distribution, for any $\rho > 0$ and estimate $\tilde{x}(\cdot)$ one has

$$\Pr_{\eta,\xi} \{\| B\eta - \tilde{x}(A\eta + \xi) \| \geq \rho \} = \mathbb{E}_{\omega} \{ \Pr_{\omega} \{ \| B\eta - \tilde{x}(\omega) \| \geq \rho \} \} \geq \mathbb{E}_{\omega} \{ \Pr_{\omega} \{ \| B\eta - \mathbb{E}_{\omega} [B\eta] \| \geq \rho \} \} = \Pr_{\eta,\xi} \{ \| B\eta - \tilde{H}^T (A\eta + \xi) \| \geq \rho \},$$

with the inequality given by the Anderson Lemma as applied to the shift of the Gaussian distribution $P_{\omega}$ by its mean. Applying the Anderson Lemma again we

\(^{24}\)We have used the following standard fact [168]: let $\zeta = [\omega;\eta] \sim \mathcal{N}(0,S)$, the covariance matrix of the marginal distribution of $\omega$ being nonsingular. Then the conditional, $\omega$ given, distribution of $\eta$ is Gaussian with the mean linearly depending on $\omega$ and covariance matrix independent of $\omega$. 
get
\[ \mathrm{Prob}_{\eta,\xi}(\|B\eta - \bar{H}^T(A\eta + \xi)\| \geq \rho) = \mathbb{E}_\xi \{ \mathrm{Prob}_{\eta}(\| (B - \bar{H}^T A)\eta - \bar{H}^T \xi\| \geq \rho) \} \] 
\[ \geq \mathrm{Prob}_{\eta}(\| (B - \bar{H}^T A)\eta\| \geq \rho), \]
and, by “symmetric” reasoning,
\[ \mathrm{Prob}_{\eta,\xi}(\|B\eta - \bar{H}^T(A\eta + \xi)\| \geq \rho) \geq \mathrm{Prob}_{\xi}(\|\bar{H}^T \xi\| \geq \rho). \]

We conclude that for any \( \hat{x}(\cdot) \)
\[ \mathrm{Prob}_{\eta,\xi}(\|B\eta - \hat{x}(\omega)\| \geq \rho) \geq \max \left\{ \mathrm{Prob}_{\eta}(\| (B - \bar{H}^T A)\eta\| \geq \rho), \mathrm{Prob}_{\xi}(\|\bar{H}^T \xi\| \geq \rho) \right\}. \] (4.158)

4°. Let \( H \) be \( m \times \nu \) matrix. Applying Lemma 4.17 to \( N = m, Y = \bar{H}, Q = Q_* \), we get from (4.59)
\[ \mathrm{Prob}_{\xi \sim \mathcal{N}(0,Q_*)}(\|\bar{H}^T \xi\| \geq [4\kappa]^{-1}\bar{\Psi}(\bar{H})) \geq \beta_* \forall \kappa \geq 1, \] (4.159)
where \( \bar{\Psi}(H) \) is defined by (4.156). Similarly, applying Lemma 4.17 to \( N = n, Y = (B - \bar{H}^T A)^T, Q = W \), we obtain
\[ \mathrm{Prob}_{\eta \sim \mathcal{N}(0,W)}(\| (B - \bar{H}^T A)\eta\| \geq [4\kappa]^{-1}\bar{\Psi}(W,\bar{H})) \geq \beta_* \forall \kappa \geq 1, \] (4.160)
where
\[ \bar{\Psi}(W,\bar{H}) = \min_{\mathcal{T}=(\mathcal{T}_\ell,\ell \leq L),\Theta} \left\{ \mathrm{Tr}(W\Theta) + \phi_R(\lambda|\mathcal{T}) : \mathcal{T}_\ell \succeq 0 \forall \ell, \right. \]
\[ \left. \left[ \frac{1}{2} M^T [B - \bar{H}^T A]^T \right] - \frac{\lambda}{2} M^T \sum_{\ell} S_\ell [\mathcal{T}_\ell] \succeq 0 \right\}. \] (4.161)

Let us put \( \rho(W,\bar{H}) = [8\kappa]^{-1}(\bar{\Psi}(\bar{H}) + \bar{\Psi}(W,\bar{H})) \); when combining (4.160) with (4.159) we conclude that
\[ \max \left\{ \mathrm{Prob}_{\eta}(\| (B - \bar{H}^T A)\eta\| \geq \rho(W,\bar{H})), \mathrm{Prob}_{\xi}(\|\bar{H}^T \xi\| \geq \rho(W,\bar{H})) \right\} \geq \beta_* \]
and the same inequality holds if \( \rho(W,\bar{H}) \) is replaced with the smaller quantity
\[ \bar{\rho}(W) = [8\kappa]^{-1} \inf_{\bar{H}} (\bar{\Psi}(\bar{H}) + \bar{\Psi}(W,\bar{H})). \]

Now, the latter bound combines with (4.158) to imply the following result:

**Lemma 4.49.** Let \( W \) be a positive semidefinite \( n \times n \) matrix, and \( \kappa \geq 1 \). Then for any estimate \( \hat{x}(\cdot) \) of \( B\eta \) given observation \( \omega = A\eta + \xi \), where \( \eta \sim \mathcal{N}(0,W) \) is independent of \( \xi \sim \mathcal{N}(0,Q_*) \), one has
\[ \mathrm{Prob}_{\eta,\xi} \left\{ \| B\eta - \hat{x}(\omega) \| \geq [8\kappa]^{-1} \inf_{\bar{H}} (\bar{\Psi}(\bar{H}) + \bar{\Psi}(W,\bar{H})) \right\} \geq \beta_* = 1 - \frac{e^{3/8}}{2} - 2Fe^{-\kappa^2/2} \]
where \( \bar{\Psi}(H) \) and \( \bar{\Psi}(W,\bar{H}) \) are defined, respectively, by (4.156) and (4.161).
In particular, for

$$\kappa = \bar{\kappa} := \sqrt{2 \ln F + 10 \ln 2} \quad (4.162)$$

it holds

$$\text{Prob}_{\eta, \xi}\{\|B\eta - \hat{x}(\omega)\| \geq [8\bar{\kappa}]^{-1} \inf_H [\Psi(H) + \overline{\Phi}(W, H)]\} > \frac{3}{16}.$$ 

5°. For $0 < \kappa \leq 1$, let us set

(a) $W_\kappa = \{W \in S^n_+ : \exists t \in T : R_k[W] \geq \kappa t_k I_{d_k}, 1 \leq k \leq K\},$

(b) $Z = \left\{ (T = \{T_\ell, \ell \leq L\}, \Theta, H) : \begin{bmatrix} \Theta \\ \frac{1}{2} M^T [B - H^T A] \\ \sum_{\ell} S_\ell [T_\ell] \end{bmatrix} \succeq 0 \right\}.$

Note that $W_\kappa$ is a nonempty convex and compact (by Lemma 4.44) set such that $W_\kappa = \kappa W_1$, and $Z$ is a nonempty closed convex set. Consider the parametric saddle point problem

$$\text{Opt}(\kappa) = \max_{W \in W_\kappa} \inf_{(\Upsilon, \Theta, H) \in Z} \left[ E(W; \Upsilon, \Theta, H) := \text{Tr}(W \Theta) + \phi_R(\lambda[\Upsilon]) + \Psi(H) \right]. \quad (4.163)$$

This problem is convex-concave; utilizing the fact that $W_\kappa$ is compact and contains positive definite matrices, it is immediately seen that the Sion-Kakutani theorem ensures the existence of a saddle point whenever $\kappa \in (0, 1]$. We claim that

$$0 < \kappa \leq 1 \Rightarrow \text{Opt}(\kappa) \geq \sqrt{\kappa} \text{Opt}(1). \quad (4.164)$$

Indeed, $Z$ is invariant w.r.t. scalings

$$(\Upsilon = \{\Upsilon_\ell, \ell \leq L\}, \Theta, H) \mapsto ((\theta \Upsilon := \{\theta \Upsilon_\ell, \ell \leq L\}, \theta^{-1} \Theta, H), \quad [\theta > 0].$$

When taking into account that $\phi_R(\lambda[\theta \Upsilon]) = \theta \phi_R(\lambda[\Upsilon])$, we get

$$E(W) := \inf_{(\Upsilon, \Theta, H) \in Z} E(W; \Upsilon, \Theta, H) = \inf_{(\Upsilon, \Theta, H) \in Z} \inf_{\theta > 0} E(W; \theta \Upsilon, \theta^{-1} \Theta, H) = \inf_{(\Upsilon, \Theta, H) \in Z} \left[ 2 \sqrt{\text{Tr}(W \Theta)} \phi_R(\lambda[\Upsilon]) + \Psi(H) \right].$$

Because $\Psi$ is nonnegative we conclude that whenever $W \succeq 0$ and $\kappa \in (0, 1]$, one has

$$E(\kappa W) \geq \sqrt{\kappa} E(W).$$

This combines with $W_\kappa = \kappa W_1$ to imply that

$$\text{Opt}(\kappa) = \max_{W \in W_\kappa} E(W) = \max_{W \in W_1} E(\kappa W) \geq \sqrt{\kappa} \max_{W \in W_1} E(W) = \sqrt{\kappa} \text{Opt}(1),$$

and (4.164) follows.

6°. We claim that

$$\text{Opt}(1) = \text{Opt}, \quad (4.165)$$

where Opt is given by (4.42) (and, as we have seen, by (4.156) as well). Note that (4.165) combines with (4.164) to imply that

$$0 < \kappa \leq 1 \Rightarrow \text{Opt}(\kappa) \geq \sqrt{\kappa} \text{Opt}. \quad (4.166)$$
Verification of (4.165) is given by the following computation. By the Sion-Kakutani Theorem,

\[
\text{Opt}(1) = \max_{W \in W_1} \inf_{(\tau, \theta, H) \in \mathcal{Z}} \{ \text{Tr}(W \Theta) + \phi_\mathcal{R}(\lambda|\mathcal{T}) + \mathcal{V}(H) \} \\
= \inf_{(\tau, \theta, H) \in \mathcal{Z}} \max_{W \in W_1} \{ \text{Tr}(W \Theta) + \phi_\mathcal{R}(\lambda|\mathcal{T}) + \mathcal{V}(H) \} \\
= \inf_{(\tau, \theta, H) \in \mathcal{Z}} \left\{ \mathcal{V}(H) + \phi_\mathcal{R}(\lambda|\mathcal{T}) + \max_{W} \left\{ \text{Tr}(\Theta W) : W \geq 0, \exists t \in \mathcal{T} : R_k[W] \leq t_kI_{d_k}, k \leq K \right\} \right\} \\
= \inf_{(\tau, \theta, H) \in \mathcal{Z}} \left\{ \mathcal{V}(H) + \phi_\mathcal{R}(\lambda|\mathcal{T}) + \max_{W,t} \left\{ \text{Tr}(\Theta W) : W \geq 0, [t; 1] \in \mathcal{T}, R_k[W] \leq t_kI_{d_k}, k \leq K \right\} \right\},
\]

where \( \mathcal{T} \) is the closed conic hull of \( \mathcal{T} \). On the other hand, using Conic Duality combined with the fact that \( \mathcal{T}_* = \{ [g; s] : s \geq \phi_\mathcal{T}(-g) \} \) we obtain

\[
\max_{W,t} \{ \text{Tr}(\Theta W) : W \geq 0, [t; 1] \in \mathcal{T}, R_k[W] \leq t_kI_{d_k}, k \leq K \}
\]

\[
= \min_{Z, [g; s], \Lambda = \Lambda_k} \left\{ \begin{array}{l}
\forall (W \in \mathcal{S}^n, t \in \mathbb{R}^K) \\
Z \geq 0, [g; s] \in \mathcal{T}_*, \Lambda_k \geq 0, k \leq K \\
\text{Tr}(ZW) - g^Tt + \sum_k \text{Tr}(R^*_k[\Lambda_k]W) - \sum_k t_k \text{Tr}(\Lambda_k) = \Theta \n\end{array} \right\}
\]

\[
= \min_{Z, [g; s], \Lambda = \Lambda_k} \left\{ \begin{array}{l}
Z \geq 0, s \geq \phi_\mathcal{T}(-g), \Lambda_k \geq 0, k \leq K \\
\Theta = \sum_k R^*_k[\Lambda_k] - Z, g = -\lambda[\Lambda] \n\end{array} \right\}
\]

\[
= \min_{\Lambda} \left\{ \phi_\mathcal{T}(\lambda[\Lambda]) : \Lambda = \{ \Lambda_k \geq 0, k \leq K \}, \Theta \leq \sum_k R^*_k[\Lambda_k] \right\},
\]

and we arrive at

\[
\text{Opt}(1) = \inf_{\tau, \theta, H, \Lambda} \left\{ \mathcal{V}(H) + \phi_\mathcal{R}(\lambda|\mathcal{T}) + \phi_\mathcal{T}(\lambda[\Lambda]) : \right\}
\]

\[
\mathcal{Y} = \{ \tau_{\ell} \geq 0, \ell \leq L \}, \Lambda = \{ \Lambda_k \geq 0, k \leq K \}, \\
\Theta \leq \sum_k R^*_k[\Lambda_k], \\
\begin{bmatrix} \frac{1}{2}M^T[B - H^T]M & \frac{1}{2}M^T[B - H^T]M \end{bmatrix} \geq 0
\]

\[
= \inf_{\tau, H, \Lambda} \left\{ \mathcal{V}(H) + \phi_\mathcal{R}(\lambda|\mathcal{T}) + \phi_\mathcal{T}(\lambda[\Lambda]) : \right\}
\]

\[
\mathcal{Y} = \{ \tau_{\ell} \geq 0, \ell \leq L \}, \Lambda = \{ \Lambda_k \geq 0, k \leq K \} \\
\begin{bmatrix} \sum_k R^*_k[\Lambda_k] & \sum_k R^*_k[\Lambda_k] \end{bmatrix} \geq 0
\]

\[
= \text{Opt} \ [\text{see (4.156)}].
\]
a saddle point solution to the saddle point problem (4.163). Then, by (4.166),
\[
\sqrt{\kappa} \text{Opt} \leq \text{Opt}(\kappa) = \inf_{(\Upsilon, \Theta, H) \in \mathcal{Z}} \left\{ \text{Tr}(W_{\kappa} \Theta) + \phi_{\mathcal{R}}(\lambda[\Upsilon]) + \Psi(H) \right\}
\]
(4.167)
On the other hand, when applying Lemma 4.46 to \( Q = W_{\kappa} \) and \( \rho = \kappa \), we obtain, in view of relations \( 0 < \kappa \leq 1, W_{\kappa} \in \mathcal{W} \kappa \),
\[
\delta(\kappa) := \text{Prob}_{\zeta \sim N(0, I_n)} \{ W_{\kappa}^{1/2} \zeta \notin \mathcal{X} \} \leq 2D e^{-\frac{1}{4}}, \quad (4.168)
\]
with \( D \) given by (4.55). In particular, when setting \( \bar{\kappa} = \frac{1}{2} \ln D + 10 \ln 2 \) (4.169)
we obtain \( \delta_{\kappa} \leq \frac{1}{16} \). Therefore,
\[
\text{Prob}_{\eta \sim N(0, W_{\kappa})} \{ \eta \notin \mathcal{X} \} \leq \frac{1}{16}. \quad (4.170)
\]
Now let
\[
\varrho_* := \frac{\text{Opt}}{8 \sqrt{(2 \ln F + 10 \ln 2)(2 \ln D + 10 \ln 2)}} \quad (4.171)
\]
All we need in order to achieve our goal, that is, to justify (4.54), is to show that
\[
\text{RiskOpt}_{\frac{1}{8}} \geq \varrho_*, \quad (4.172)
\]
since given the latter relation, (4.54) will be immediately given by (4.157) as applied with \( \epsilon = \frac{1}{8} \).
To prove (4.172), assume, on the contrary to what should be proved, that the \( \frac{1}{8} \)-risk is \( < \varrho_* \), and let \( \hat{x}(\cdot) \) be an estimate with \( \frac{1}{8} \)-risk \( \varrho' < \varrho_* \). We can utilize \( \hat{x} \) to estimate \( B\eta \), in the Bayesian problem of recovering \( B\eta \) from observation \( \omega = A\eta + \xi \), \( (\eta, \xi) \sim N(0, \Sigma) \) with \( \Sigma = \text{Diag} \{ W_{\bar{\kappa}}, Q_* \} \). From (4.170) we conclude that
\[
\text{Prob}(\eta, \xi) \sim N(0, \Sigma) \{ \| B\eta - \hat{x}(A\eta + \xi) \| \geq \varrho' \} \leq \frac{1}{8} + \frac{1}{16} = \frac{3}{16} \quad (4.173)
\]
On the other hand, by (4.167) we have
\[
\inf_H \left\{ \Psi(W_{\bar{\kappa}}, H) + \Psi(H) \right\} = \text{Opt}(\bar{\kappa}) \geq \sqrt{\bar{\kappa}} \text{Opt} = [8\varrho] \varrho_*
\]
with \( \varrho \) given by (4.162), so by Lemma 4.49, for any estimate \( \hat{x}(\cdot) \) of \( B\eta \) via observation \( \omega = A\xi + \xi \) it holds
\[
\text{Prob}(\eta, \xi) \{ \| B\eta - \hat{x}(A\eta + \xi) \| \geq \varrho_* \} \geq \beta_{\kappa} > 3/16;
\]
in particular, this relation should hold true for \( \hat{x}(\cdot) \equiv \hat{x}(\cdot) \), but the latter is impossible: the \( \frac{1}{16} \)-risk of \( \hat{x} \) is \( \leq \varrho' < \varrho_* \), see (4.173). \( \Box \)
4.8.5.2 Proof of Proposition 4.5

We shall extract Proposition 4.5 from the following result, meaningful by its own right (it can be considered as “ellitopic refinement” of Proposition 4.16):

**Proposition 4.50.** Consider recovery of the linear image \( Bx \in \mathbb{R}^\nu \) of unknown signal \( x \) known to belong to a given signal set \( X \subset \mathbb{R}^n \) from noisy observation

\[
\omega = Ax + \xi \in \mathbb{R}^m \quad [\xi \sim \mathcal{N}(0, \Gamma), \, \Gamma > 0]
\]

the recovery error being measured in norm \( \| \cdot \| \) on \( \mathbb{R}^\nu \). Assume that \( X \) and the unit ball \( B^* \) of the norm \( \| \cdot \|^* \) conjugate to \( \| \cdot \| \) are ellitopes:

\[
X = \{ x \in \mathbb{R}^n : \exists t \in T : x^T R_k x \leq t_k, \, k \leq K \}
\]

\[
B^* = \{ y \in \mathbb{R}^\nu : \exists (r \in \mathbb{R}, y) : u = My, y^T S_\ell y \leq r_\ell, \, \ell \leq L \} \quad (4.174)
\]

with our standard restrictions on \( T, \, R, \, R_k \) and \( S_\ell \) (as always, we lose nothing when assuming that the ellitope \( X \) is basic).

Consider the optimization problem

\[
\text{Opt}_\# = \min_{\Theta, H, \lambda, \mu, \mu'} \left\{ \phi_T(\lambda) + \phi_R(\mu) + \phi_R(\mu') + \text{Tr}(\Gamma \Theta) : \lambda \geq 0, \mu \geq 0, \mu' \geq 0 \right\}
\]

\[
\begin{bmatrix}
\frac{1}{2} M^T |B - H^T A|^T M \\
\Theta \\
\frac{1}{2} M^T H^T \\
\sum_\ell \mu_\ell S_\ell
\end{bmatrix} \succeq 0
\]

\[
(4.175)
\]

The problem is solvable, and the linear estimate \( \hat{x}_H(\omega) = H^T \omega \) yielded by the \( H \)-component of an optimal solution to the problem satisfies the risk bound

\[
\text{Risk}_{\Gamma, \| \cdot \|}[\hat{x}_H, X] = \max_{x \in X} \mathbb{E}_{\xi \sim \mathcal{N}(0, \Gamma)} \{ \| Bx - \hat{x}_H(Ax + \xi) \| \} \leq \text{Opt}_\#.
\]

Furthermore, the estimate \( \hat{x}_H(\cdot) \) is near-optimal:

\[
\text{Opt}_\# \leq 64 \sqrt{(3 \ln K + 15 ln 2)(3 \ln L + 15 \ln 2)} \text{RiskOpt}, \quad (4.176)
\]

where \( \text{RiskOpt} \) is the minimax optimal risk

\[
\text{RiskOpt} = \inf_{\hat{x}} \sup_{x \in X} \mathbb{E}_{\xi \sim \mathcal{N}(0, \Gamma)} \{ \| Bx - \hat{x}(Ax + \xi) \| \},
\]

the infimum being taken w.r.t. all estimates.

**Proposition 4.50 ⇒ Proposition 4.5:** Clearly, the situation considered in Proposition 4.5 is a particular case of the setting of Proposition 4.50, namely, the case where \( B_* \) is the standard Euclidean ball, \( B_* = \{ u \in \mathbb{R}^\nu : u^T u \leq 1 \} \). In this case,
problem (4.175) reads

\[ \text{Opt}_{\#} = \min_{\Theta, H, \lambda, \mu, \mu'} \left\{ \phi_T(\lambda) + \mu + \mu' + \text{Tr}(\Gamma \Theta) : \begin{array}{l}
\lambda \geq 0, \mu \geq 0, \mu' \geq 0 \\
[\frac{1}{2}B - H^T A]^T \frac{1}{2} [B - H^T A] \geq \lambda \Theta \\
\frac{1}{2} H^T \frac{1}{2} H \geq \mu \Theta \\
\end{array} \right\} \]

\[ = \min_{\Theta, H, \lambda, \mu, \mu'} \left\{ \phi_T(\lambda) + \mu + \mu' + \text{Tr}(\Gamma \Theta) : \begin{array}{l}
\lambda \geq 0, \mu \geq 0, \mu' \geq 0 \\
\mu \left[ \sum_k \lambda_k R_k \right] \geq \frac{1}{4} (B - H^T A)^T [B - H^T A] \\
\mu' \Theta \geq \frac{1}{4} H H^T \end{array} \right\} \]

[Schur Complement Lemma]

\[ = \min_{\chi, H} \left\{ \sqrt{\phi_T(\chi)} + \sqrt{\text{Tr}(H^T H)} : \begin{array}{l}
\chi \geq 0, \left[ \sum_k \chi^k R_k \right] \left[ B - H^T A \right]^T \geq 0 \\
\end{array} \right\}. \]

(by eliminating $\mu, \mu'$; note that $\phi_T(\cdot)$ is positively homogeneous of degree 1)

Comparing the resulting representation of $\text{Opt}_{\#}$ with (4.12), we see that the upper bound $\sqrt{\text{Opt}}$ on the risk of the linear estimate $\hat{x}_H$, appearing in (4.15) is $\leq \text{Opt}_{\#}$. Combining this observation with (4.176) and the evident relation

\[ \text{Risk}_{\text{Opt}} = \inf_{\hat{x}} \sup_{x \in X} E_{x \sim N(0, \Gamma)} \left\{ \| Bx - \hat{x}(Ax + \xi) \|_2 \right\} \leq \inf_{\hat{x}} \sqrt{\sup_{x \in X} E_{x \sim N(0, \Gamma)} \left\{ \| Bx - \hat{x}(Ax + \xi) \|_2^2 \right\}} = \text{Risk}_{\text{opt}}, \]

(recall that we are in the case of $\| \cdot \| = \| \cdot \|_2$), we arrive at (4.15) and thus justify Proposition 4.5.

**Proof of Proposition 4.50.** It is immediately seen that problem (4.175) is nothing but problem (4.42) in the case when the spectratopes $\mathcal{X}, \mathcal{B}_s$ and the set $\Pi$ participating in Proposition 4.14 are, respectively, the ellitopes given by (4.174), and the singleton $\{ \Gamma \}$. Thus, Proposition 4.50 is, essentially, a particular case of Proposition 4.16. The only refinement in Proposition 4.50 as compared to Proposition 4.16 is the form of the logarithmic “non-optimality” factor in (4.176); similar factor in Proposition 4.16 is expressed in terms of spectratopic sizes $D, F$ of $\mathcal{X}$ and $\mathcal{B}_s$ (the total ranks of matrices $R_k$, $k \leq K$, and $S_k$, $s \leq L$, in the case of (4.174)), while in (4.176) the nonoptimality factor is expressed in terms of ellitopic sizes $K, L$ of $\mathcal{X}$ and $\mathcal{B}_s$. Strictly speaking, to arrive at this (slight – the sizes in question are under logs) refinement, we were supposed to reproduce, with minimal modifications, the reasoning of items 2o – 7o of Section 4.8.5.1, with $\Gamma$ in the role of $Q_\Gamma$, and slightly refine Lemma 4.17 underlying this reasoning. Instead of carrying out this plan literally, we detail “local modifications” to be made in the proof of Proposition 4.16 in order to prove Proposition 4.50. Here are these modifications:

A. The collections of matrices $\Lambda = \{ \Lambda_k \geq 0, k \leq K \}$, $\Upsilon = \{ \Upsilon_\ell \geq 0, \ell \leq L \}$ should be substituted by collections of nonnegative reals $\lambda \in \mathbb{R}_+^K$, resp., $\mu \in \mathbb{R}_+^L$, and
vectors $\lambda, \Lambda$, resp. $\mu$. Expressions like $R_k W$, $R_k^* [\Lambda_k]$, $S^* [\Upsilon]$ should be replaced, respectively, with $\text{Tr}(R_k W)$, $\lambda_k R_k$, $\mu \mu R_k$. Finally, $Q_*$ should be replaced with $\Gamma$, and scalar matrices, like $t_k I_d$, should be replaced with the corresponding reals, like $t_k$.

B. The role of Lemma 4.17 is now played by

**Lemma 4.51.** Let $Y$ be an $N \times \nu$ matrix, let $\| \cdot \|$ be a norm on $\mathbb{R}^\nu$ such that the unit ball $\mathcal{B}_*$ of the conjugate norm is the ellipote:

$$\mathcal{B}_* = \{ y \in \mathbb{R}^\nu : \exists (r \in \mathbb{R}, y) : u = M y, y^T S y \leq r, \ell \leq L \}, \quad (4.174)$$

and let $\zeta \sim \mathcal{N}(0, Q)$ for some positive semidefinite $N \times N$ matrix $Q$. Then the best upper bound on $\psi_Q(Y) := \mathbb{E}\{\|Y^T \zeta\|\}$ yielded by Lemma 4.11, that is, the optimal value $\text{Opt}[Q]$ in the convex optimization problem (cf. (4.40))

$$\text{Opt}[Q] = \min_{\Theta G, \mu} \left\{ \phi_R(\mu) + \text{Tr}(G) : \mu \geq 0, \left[ \frac{\Theta}{2} M^T Y^T \right] \frac{1}{2} Y M \frac{1}{2} \sum_\ell \mu_\ell R_\ell \right\} \geq 0$$

satisfies for all $Q \succeq 0$ the identity

$$\text{Opt}[Q] = \text{Opt}[Q] : \mu \geq 0, \left[ \frac{G}{2} M^T Y^T \right] \frac{1}{2} Q^{1/2} Y M \frac{1}{2} \sum_\ell \mu_\ell R_\ell \geq 0$$

and is a tight bound on $\psi_Q(Y)$. Namely,

$$\psi_Q(Y) \leq \text{Opt}[Q] \leq 22 \sqrt{3 \ln L + 15 \ln 2} \psi_Q(Y),$$

where $L$ is the size of the ellipote $\mathcal{B}_*$, see (4.174). Furthermore, for all $\varkappa \geq 1$ one has

$$\text{Prob}_\varkappa\left\{ \|Y^T \zeta\| \geq \frac{\text{Opt}[Q]}{4 \varkappa} \right\} \geq \beta_\varkappa := 1 - \frac{e^{3/8}}{2} - 2Le^{-\varkappa^2/3}.$$  \hspace{1cm} (4.178)

In particular, when selecting $\varkappa = \sqrt{3 \ln L + 15 \ln 2}$, we obtain

$$\text{Prob}_\varkappa\left\{ \|Y^T \zeta\| \geq \frac{\text{Opt}[Q]}{4 \sqrt{3 \ln L + 15 \ln 2}} \right\} \geq \beta_\varkappa = 0.2100 \geq \frac{3}{16}.$$  

**Proof** of Lemma 4.51 follows the lines of the proof of Lemma 4.17, with Lemma 4.47 substituting Lemma 4.46.

1°. Relation (4.177) can be verified exactly in the same fashion as in the case of Lemma 4.17.

2°. Let us set $\zeta = Q^{1/2} \eta$ with $\eta \sim \mathcal{N}(0, I_N)$ and $Z = Q^{1/2} Y$. Observe that to prove (4.178) is the same as to show that when $\varkappa \geq 1$ one has

$$\bar{\delta} = \frac{\text{Opt}[Q]}{4 \varkappa} \Rightarrow \text{Prob}_\varkappa\{ \|Z^T \eta\| \geq \bar{\delta} \} \geq \beta_\varkappa := 1 - \frac{e^{3/8}}{2} - 2Le^{-\varkappa^2/3}.$$  \hspace{1cm} (4.179)
where

\[
\begin{align*}
\text{Opt}[Q] = & \quad \text{Opt}[Q] := \min_{\Theta, \mu} \phi_{\mathcal{R}}(\mu) + \text{Tr}(\Theta) : \mu \geq 0, \\
& \quad \left[ \frac{1}{2} \Theta M^T Z^T \left\| \sum_{\ell} \mu_{\ell} R_{\ell} \right\| \geq 0 \right].
\end{align*}
\]

(4.180)

Justification of (4.179) goes as follows.

2.1\textsuperscript{o}. Let us represent \( \text{Opt}[Q] \) as the optimal value of a conic problem. Setting

\[
\mathbf{K} = \text{cl}\{(r; s) : s > 0, r/s \in \mathcal{R}\},
\]

we ensure that

\[
\mathcal{R} = \{r : [r; 1] \in \mathbf{K}\}, \quad \mathbf{K}_* = \{[g; s] : s \geq \phi_{\mathcal{R}}(g)\},
\]

where \( \mathbf{K}_* \) is the cone dual to \( \mathbf{K} \). Consequently, (4.180) reads

\[
\text{Opt}[Q] = \min_{\Theta, \mu, \theta} \left\{ \begin{array}{l}
\mu \geq 0 \\
\frac{1}{2} \Theta M^T Z^T \left\| \sum_{\ell} \mu_{\ell} S_{\ell} \right\| \geq 0 \\
\left[ -\mu; \theta \right] \in \mathbf{K}_*
\end{array} \right\}.
\]

(P\textsubscript{E})

2.2\textsuperscript{o}. Now let us prove that there exists matrix \( W \in \mathbf{S}_{++}^R \) and \( r \in \mathcal{R} \) such that

\[
\text{Tr}(WS_{\ell}) \leq r_{\ell}, \quad \ell \leq L,
\]

and

\[
\text{Opt}[Q] \leq \sum_{i} \sigma_i(ZMW^{1/2}),
\]

(4.182)

where \( \sigma_1(\cdot) \geq \sigma_2(\cdot) \geq \ldots \) are singular values.

To get the announced result, let us pass from problem \( P \) to its conic dual. \( P\textsubscript{E} \) clearly is strictly feasible and bounded, so that the dual to \( P\textsubscript{E} \) problem \( D\textsubscript{E} \) is solvable with optimal value \( \text{Opt}[Q] \). Denoting by \( \lambda_{\ell} \geq 0, \ell \leq L, \quad \left[ G \quad -R \right] \geq 0, \quad [r; \tau] \in \mathbf{K} \)

the Lagrange multipliers for the respective constraints in \( P\textsubscript{E} \), and aggregating these constraints, the multipliers being the aggregation weights, we arrive at the following aggregated constraint:

\[
\text{Tr}(\Theta G) + \text{Tr}(W \sum_{\ell} \mu_{\ell} S_{\ell}) + \sum_{\ell} \lambda_{\ell} - \sum_{\ell} r_{\ell} \mu_{\ell} + \theta \tau \geq \text{Tr}(ZMR^T).
\]

To get the dual problem, we impose on the Lagrange multipliers, in addition to the initial constraints, the restriction that the left hand side in the aggregated constraint, identically in \( \Theta, \mu_{\ell} \) and \( \theta \), is equal to the objective of \( P \), that is,

\[
G = I, \quad \text{Tr}(WS_{\ell}) + \lambda_{\ell} - r_{\ell} = 0, \quad 1 \leq \ell \leq L, \quad \tau = 1,
\]

and maximize the right-hand side of the aggregated constraint. After immediate simplifications, we arrive at

\[
\text{Opt}[Q] = \max_{W, r, r} \left\{ \text{Tr}(ZMR^T) : W \succeq R^T R, r \in \mathcal{R}, \text{Tr}(WS_{\ell}) \leq r_{\ell}, \quad 1 \leq \ell \leq L \right\}
\]

(note that \( r \in \mathcal{R} \) is equivalent to \([r; 1] \in \mathbf{K}\), and \( W \succeq R^T R \) is the same as \( \left[ I \quad -R^T \right] \geq 0 \)). Exactly as in the proof of Lemma 4.17, the above representation of \( \text{Opt}[Q] \) implies that

\[
\text{Opt}[Q] = \max_{W, r} \left\{ \|ZMW^{1/2}\|_{\mathcal{S}_{++}^1} : r \in \mathcal{R}, \text{Tr}(WS_{\ell}) \leq r_{\ell}, \quad \ell \leq L \right\}.
\]
The resulting problem clearly is solvable, and its optimal solution $W$ ensures the target relations (4.181) and (4.182).

2.3’. Given $W$ satisfying (4.181) and (4.182), we proceed exactly as in item 2.3’ of the proof of Lemma 4.17, thus arriving at three random vectors $(\chi, \upsilon, \eta)$ with marginal distributions $\mathcal{N}(0, I_q)$, $\mathcal{N}(0, I_r)$, and $\mathcal{N}(0, I_N)$, respectively, such that

$$
\chi^T W^{1/2} M^T Z^T \eta = \sum_{i=1}^p \sigma_i v_i^2.
$$

(4.183)

where $p = \min[q, N]$ and $\sigma_i = \sigma_i(ZM^{1/2})$. Same as in item 3’i of the proof of Lemma 4.17, we have (i)

$$
\alpha := \text{Prob} \left\{ \sum_{i=1}^p \sigma_i v_i^2 < \frac{1}{4} \sum_{i=1}^p \sigma_i \right\} \leq \frac{e^{3/8}}{2} \approx 0.7275... .
$$

(4.184)

The role of item 3’ii in the aforementioned proof is now played by (ii) Whenever $\kappa \geq 1$, one has

$$
\text{Prob}\{ \|MW^{1/2}\| > \kappa \} \leq 2L \exp\{-\kappa^2/3\},
$$

(4.185)

with $L$ as defined in (4.174).

Indeed, setting $\rho = 1/\kappa^2 \leq 1$ and $\omega = \sqrt{\rho}W^{1/2}\chi$, we get $\omega \sim \mathcal{N}(0, \rho W)$. Let us apply Lemma 4.47 to $Q = \rho W$, $R$ in the role of $T$, with $L$ in the role of $K$, and $S_i$’s in the role of $R_k$’s. Denoting

$$
\mathcal{Y} := \{ y : \exists r \in R : y^T S_i y \leq r_i, \ell \leq L \},
$$

we have $\text{Tr}(QS_i) = \rho \text{Tr}(WS_i) = \rho \text{Tr}(WS_i) \leq \rho r_i, \ell \leq L$, with $r \in R$ (see (4.181)), so we are under the premise of Lemma 4.47 (with $\mathcal{Y}$ in the role of $X$ and therefore with $L$ in the role of $K$). Applying the lemma, we conclude that

$$
\text{Prob}\left\{ \chi : \kappa^{-1}W^{1/2}\chi \notin \mathcal{Y} \right\} \leq 2L \exp\{-1/(3\rho)\} = 2L \exp\{-\kappa^2/3\}.
$$

Recalling that $\mathcal{B}_x = M\mathcal{Y}$, we see that $\text{Prob}\{ \chi : \kappa^{-1}MW^{1/2}\chi \notin \mathcal{B}_x \}$ is indeed upper-bounded by the right hand side of (4.185), and (4.185) follows.

With (i), (ii) at our disposal, we complete the proof of Lemma 4.51 in exactly the same way as in items 2.4”, 3” of the proof of Lemma 4.17. \(\square\)

C. As a result of substituting Lemma 4.17 with Lemma 4.51, the counterpart of Lemma 4.49 used in item 4” of the proof of Proposition 4.16 now reads as follows:

**Lemma 4.52.** Let $W$ be a positive semidefinite $n \times n$ matrix, and $\kappa \geq 1$. Then for any estimate $\hat{x}(\cdot)$ of $B\eta$ given observation $\omega = A\eta + \xi$ with independent of each other $\eta \sim \mathcal{N}(0, W)$ and $\xi \sim \mathcal{N}(0, \Gamma)$, one has

$$
\text{Prob}_{\eta, \xi} \left\{ \|B\eta - \hat{x}(\omega)\| \geq [8 \kappa]^{-1} \inf_{H} \|\mathcal{V}(H) + \mathcal{F}(W, H)\| \right\} \geq \beta_\kappa = 1 - \frac{e^{3/8}}{2} - 2L e^{-\kappa^2/3}
$$

where $\mathcal{V}(H)$ and $\mathcal{F}(W, H)$ are defined, respectively, by (4.156) (where $Q_x$ should be set to $\Gamma$) and (4.161).

In particular, for

$$
\kappa = \kappa := \sqrt{3 \ln K + 15 \ln 2}
$$

the latter probability is $> 3/16$. 

D. We substitute reference to Lemma 4.46 in item 7 of the proof with Lemma 4.47, resulting in replacing

• definition of \( \delta(\kappa) \) in (4.168) with

\[
\delta(\kappa) := \text{Prob}_{\zeta \sim \mathcal{N}(0,I_\kappa)}\{W^{1/2}_\kappa \zeta \notin \mathcal{X}\} \leq 3Ke^{-\frac{1}{2}},
\]

• definition (4.169) of \( \bar{\kappa} \) with

\[
\bar{\kappa} = \frac{1}{3\ln K + 15\ln 2},
\]

• and, finally, definition (4.171) of \( \rho_* \) with

\[
\rho_* := \text{Opt} \frac{8}{3\ln L + 15\ln 2)(3\ln K + 15\ln 2)}.
\]

4.8.6 Proofs of Propositions 4.18, 4.19, and justification of Remark 4.20

4.8.6.1 Proof of Proposition 4.18

The only claim of the proposition which is not an immediate consequence of Proposition 4.8 is that problem (4.64) is solvable; let us justify this claim. Let \( F = \text{Im} A \).

Clearly, feasibility of a candidate solution \((H, \Lambda, \Upsilon)\) to the problem depends solely on the restriction of the linear mapping \( z \mapsto H^Tz \) 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 \( \mathbb{R}^m \) 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 \( a \in \mathbb{R} \) form a compact set, so that the latter problem (and thus – the original one) indeed is solvable. \( \square \)

4.8.6.2 Proof of Proposition 4.19

We are about to derive Proposition 4.19 from Proposition 4.16. Observe that in the situation of the latter Proposition, setting formally \( \Pi = \{0\} \), problem (4.42) becomes problem (4.64), so that Proposition 4.19 looks as the special case \( \Pi = \{0\} \) of Proposition 4.16. However, the premise of the latter Proposition forbids specializing \( \Pi \) as \( \{0\} \) – this would violate the regularity assumption \( \Gamma \) which is part of the premise. The difficulty, however, can be easily resolved. Assume w.l.o.g. that the image space of \( A \) is the entire \( \mathbb{R}^m \) (otherwise we could from the very beginning replace \( \mathbb{R}^m \) with the image space of \( A \)), and let us pass from our current noiseless recovery problem of interest (1), see Section 4.5.1, to its “noisy modification,” the differences with (1) being

• noisy observation \( \omega = Ax + \sigma \xi, \sigma > 0, \xi \sim \mathcal{N}(0, I_m); \)

• risk quantification of a candidate estimate \( \hat{x}(\cdot) \) according to

\[
\text{Risk}^2_{\|\cdot\|} [\hat{x}(Ax + \sigma \xi)|\mathcal{X}] = \sup_{x \in \mathcal{X}} \mathbb{E}_{\xi \sim \mathcal{N}(0, I_m)} \{\|Bx - \hat{x}(Ax + \sigma \xi)\| \},
\]
the corresponding minimax optimal risk being
\[
\text{RiskOpt}_{\|\cdot\|}^\sigma [\mathcal{X}] = \inf_{\hat{x}(\cdot)} \text{Risk}^\sigma_{\|\cdot\|} [\hat{x}(Ax + \sigma \xi) | \mathcal{X}]
\]

Proposition 4.16 does apply to the modified problem – it suffices to specify \( \Pi \) as \( \{ \sigma^2 I_m \} \). According to this proposition, the quantity
\[
\text{Opt}^\sigma = \min_{H, \Lambda, \Theta, \Upsilon, \Upsilon'} \left\{ \phi_T(\lambda[\Lambda]) + \phi_R(\lambda[\Upsilon]) + \phi_R(\lambda[\Upsilon']) + \sigma^2 \text{Tr}(\Theta) : \Lambda = \left\{ \Lambda_k \succeq 0, k \leq K \right\}, \Upsilon = \left\{ \Upsilon_{\ell} \succeq 0, \ell \leq L \right\}, \Upsilon' = \left\{ \Upsilon'_{\ell} \succeq 0, \ell \leq L \right\}, \Theta \right\}
\]
satisfies the relation
\[
\text{Opt}^\sigma \leq O(1) \ln(D) \text{RiskOpt}_{\|\cdot\|}^\sigma [\mathcal{X}] \tag{4.186}
\]
with \( D \) defined in (4.65). Looking at problem (4.64) we immediately conclude that \( \text{Opt}_{\#} \leq \text{Opt}[\sigma] \). Thus, all we need in order to extract the target relation (4.65) from (4.186) is to prove that the minimax optimal risk \( \text{Risk}_{\text{opt}}[\mathcal{X}] \) defined in Proposition 4.19 satisfies the relation
\[
\liminf_{\sigma \to +0} \text{RiskOpt}_{\|\cdot\|}^\sigma [\mathcal{X}] \leq \text{Risk}_{\text{opt}}[\mathcal{X}] \tag{4.187}
\]
To prove this relation, let us fix \( r > \text{Risk}_{\text{opt}}[\mathcal{X}] \), so that for some Borel estimate \( \hat{x}(\cdot) \) it holds
\[
\sup_{x \in \mathcal{X}} \| Bx - \hat{x}(Ax) \| < r. \tag{4.188}
\]
Were we able to ensure that \( \hat{x}(\cdot) \) is bounded and continuous, we would be done, since in this case, due to compactness of \( \mathcal{X} \), it clearly holds
\[
\liminf_{\sigma \to +0} \text{RiskOpt}_{\|\cdot\|}^\sigma [\mathcal{X}] \leq \liminf_{\sigma \to +0} \sup_{x \in \mathcal{X}} \mathbb{E}_{\xi \sim \mathcal{N}(0, I_m)} \{ \| Bx - \hat{x}(Ax + \sigma \xi) \| \}
\leq \sup_{x \in \mathcal{X}} \| Bx - \hat{x}(Ax) \| < r,
\]
and since \( r > \text{Risk}_{\text{opt}}[\mathcal{X}] \) is arbitrary, (4.187) would follow. Thus, all we need to do is to verify that given Borel estimate \( \hat{x}(\cdot) \) satisfying (4.188), we can update it into a bounded and continuous estimate satisfying the same relation. Verification is as follows:

1. Setting \( \beta = \max_{x \in \mathcal{X}} \| Bx \| \) and replacing estimate \( \hat{x} \) with its truncation

\[
\tilde{x}(\omega) = \begin{cases} 
\hat{x}(\omega), & \| \hat{x}(\omega) \| \leq 2\beta \\
0, & \text{otherwise}
\end{cases}
\]

for any \( x \in \mathcal{X} \) we only reduce the norm of the recovery error. At the same time, \( \tilde{x} \) is Borel and bounded. Thus, we lose nothing when assuming in the rest of the proof that \( \hat{x}(\cdot) \) is Borel and bounded.
2. For $\epsilon > 0$, let $\hat{x}_\epsilon(\omega) = (1 + \epsilon)\hat{x}(\omega/(1 + \epsilon))$ and let $\mathcal{X}_\epsilon = (1 + \epsilon)\mathcal{X}$. Observe that

$$\sup_{x \in \mathcal{X}_\epsilon} \|Bx - \hat{x}_\epsilon(Ax)\| = \sup_{y \in \mathcal{X}} \|B[1 + \epsilon]y - \hat{x}_\epsilon(A[1 + \epsilon]y)\|$$

$$= \sup_{y \in \mathcal{X}} \|B[1 + \epsilon]y - [1 + \epsilon]\hat{x}(Ay)\| = [1 + \epsilon] \sup_{y \in \mathcal{X}} \|By - \hat{x}(Ay)\|,$$

implying, in view of (4.188), that for small enough positive $\epsilon$ we have

$$\bar{r} := \sup_{x \in \mathcal{X}_\epsilon} \|Bx - \hat{x}_\epsilon(Ax)\| < r. \quad (4.189)$$

3. Finally, let $A^\dagger$ be the pseudoinverse of $A$, so that $AA^\dagger z = z$ for every $z \in \mathbb{R}^m$ (recall that the image space of $A$ is the entire $\mathbb{R}^m$). Given $\rho > 0$, let $\theta_\rho(\cdot)$ be a nonnegative smooth function on $\mathbb{R}^m$ with integral 1 such that $\theta_\rho$ vanishes outside of the ball of radius $\rho$ centered at the origin, and let

$$\hat{x}_{\epsilon,\rho}(\omega) = \int_{\mathbb{R}^m} \hat{x}_\epsilon(\omega - z)\theta_\rho(z)dz$$

be the convolution of $\hat{x}_\epsilon$ and $\theta_\rho$. Since $\hat{x}_\epsilon(\cdot)$ is Borel and bounded, this convolution is a well-defined smooth function on $\mathbb{R}^m$. Because $\mathcal{X}$ contains a neighbourhood of the origin, for all small enough $\rho > 0$, all $z$ from the support of $\theta_\rho$ and all $x \in \mathcal{X}$ the point $x - A^\dagger z$ belongs to $\mathcal{X}_\epsilon$. For such $\rho$ and any $x \in \mathcal{X}$ we have

$$\|Bx - \hat{x}_\epsilon(Ax - z)\| = \|Bx - \hat{x}_\epsilon(A[x - A^\dagger z])\|$$

$$\leq \|BA^\dagger z\| + \|B[x - A^\dagger z] - \hat{x}_\epsilon(A[x - A^\dagger z])\|$$

$$\leq C\rho + \bar{r}$$

with properly selected constant $C$ independent of $\rho$ (we have used (4.189); note that for our $\rho$ and $x$, $x - A^\dagger z \in \mathcal{X}_\epsilon$). We conclude that for properly selected $r' < r$, $\rho > 0$ and all $x \in \mathcal{X}$ we have

$$\|Bx - \hat{x}_\epsilon(Ax - z)\| \leq r' \forall (z \in \text{supp } \theta_\rho),$$

implying, by construction of $\hat{x}_{\epsilon,\rho}$, that

$$\forall (x \in \mathcal{X}) : \|Bx - \hat{x}_{\epsilon,\rho}(Ax)\| \leq r' < r.$$

The resulting estimate $\hat{x}_{\epsilon,\rho}$ is the continuous and bounded estimate satisfying (4.188) we were looking for.

4.8.6.3 Justification of Remark 4.20

Justification of Remark is given by repeating word by word the proof of Proposition 4.19, with Proposition 4.50 in the role of Proposition 4.16.
Chapter Five

Signal Recovery Beyond Linear Estimates

OVERVIEW

In this chapter, same as in Chapter 4, we focus on signal recovery. In contrast to the previous chapter, on our agenda now are

- a special kind of nonlinear estimation – *polyhedral estimate* (Section 5.1), an alternative to linear estimates which were our subject in Chapter 4. We demonstrate that as applied to the same estimation problem as in Chapter 4 – recovery of unknown signal via noisy observation of a linear image of the signal, polyhedral estimation possesses the same attractive properties as linear estimation, that is, efficient computability and near-optimality, provided the signal set is an ellitope/spectratope. Besides this, we show that properly built polyhedral estimates are near-optimal in several special cases where linear estimates could be heavily suboptimal.

- recovering signals from noisy observations of *nonlinear* images of the signal. Specifically, we consider signal recovery in *generalized linear models*, where the expected value of an observation is a known nonlinear transformation of the signal we want to recover, in contrast to observation model (4.1) where this expectation is linear in the signal.

5.1 POLYHEDRAL ESTIMATION

5.1.1 Motivation

The estimation problem we were considering so far is as follows:

*We want to recover the image $Bx \in \mathbb{R}^\nu$ of unknown signal $x$ known to belong to signal set $\mathcal{X} \subset \mathbb{R}^n$ from a noisy observation $\omega = Ax + \xi_x \in \mathbb{R}^m$, where $\xi_x$ is observation noise (index $x$ in $\xi_x$ indicates that the distribution $P_x$ of the observation noise may depend on $x$). Here $\mathcal{X}$ is a given nonempty convex compact set, and $A$ and $B$ are given $m \times n$ and $\nu \times n$ matrices; in addition, we are given a norm $\cdot || \cdot$ on $\mathbb{R}^\nu$ in which the recovery error is measured.*

We have seen that if $\mathcal{X}$ is an ellitope/spectratope then, under reasonable assumptions on observation noise and $\cdot || \cdot$, an appropriate efficiently computable linear in $\omega$ estimate is near-optimal. Note that the ellitopic/spectratopic structure of $\mathcal{X}$ is crucial here. What follows is motivated by the desire to build an alternative estimation scheme which works beyond the ellitopic/spectratopic case, where linear estimates can become “heavily nonoptimal.”
Motivating example. 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:

$$\mathcal{X} = \{x \in \mathbb{R}^n : \sum_i |x_i| \leq 1\}.$$ 

In this situation, building the optimal, in terms of the worst-case, over $\mathcal{X}$ expected squared risk linear estimate $\hat{x}_H(\omega) = H^T \omega$ is extremely simple:

$$\text{Risk}^2[\hat{x}_H, \mathcal{X}] := \max_{x \in \mathcal{X}} \mathbb{E}\left\{\|\hat{x}_H(\omega) - Bx\|_2^2\right\}$$

$$= \max_{x \in \mathcal{X}} \left\{\|I - H^T x\|_2^2 + \sigma^2 \text{Tr}(HH^T)\right\}$$

$$= \max_{i \leq n} \|\text{Col}_i[I - H^T]\|_2^2 + \sigma^2 \text{Tr}(HH^T).$$

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 attainable with linear estimates squared risk is

$$R^2 = \min_h [(1 - h)^2 + \sigma^2 nh^2] = \frac{n \sigma^2}{1 + n \sigma^2}.$$ 

On the other hand, consider a nonlinear estimate $\hat{x}(\omega)$ as follows. Given observation $\omega$, specify $\hat{x}(\omega)$ as an optimal solution to the optimization problem

$$\text{Opt}(\omega) = \min_{y \in \mathcal{X}} \|y - \omega\|_{\infty}.$$ 

Note that for every $\rho > 0$ the probability for the true signal to satisfy $\|x - \omega\|_{\infty} \leq \rho \sigma$ ("event $E$") is at least $1 - 2n \exp\{-\rho^2/2\}$, and if this event happens, then both $x$ and $\hat{x}$ belong to the box $\{y : \|y - \omega\|_{\infty} \leq \rho \sigma\}$, implying that $\|x - \hat{x}\|_{\infty} \leq 2 \rho \sigma$. In addition, we always have $\|x - \hat{x}\|_2 \leq \|x - \hat{x}\|_1 \leq 2$, since $x \in \mathcal{X}$ and $\hat{x} \in \mathcal{X}$. We therefore have

$$\|x - \hat{x}\|_2 \leq \sqrt{\|x - \hat{x}\|_{\infty}} \|x - \hat{x}\|_1 \leq \left\{ \begin{array}{ll} 2 \sqrt{\rho \sigma}, & \omega \in E, \\ 2, & \omega \notin E, \end{array} \right.$$ 

whence

$$\mathbb{E}\left\{\|\hat{x} - x\|_2^2\right\} \leq 4 \rho \sigma + 8n \exp\{-\rho^2/2\}. \quad (*)$$ 

Assuming $\sigma \leq 2n \exp\{-1/2\}$ and specifying $\rho$ 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 $\hat{x}(\omega)$ it holds

$$\text{Risk}^2[\hat{x}, \mathcal{X}] \leq 8 \sqrt{\ln(2n/\sigma)} \sigma.$$ 

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 $\sigma$ 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 $\frac{\Omega(1)}{\sigma \sqrt{\ln(1/\sigma)}}$ – the factor which is “by far” worse than the nonoptimality factor in the case of ellitope/spectratope $\mathcal{X}$. 
The construction of the nonlinear estimate \( \hat{x} \) which we have built\(^1\) admits a natural extension yielding what we shall call \textit{polyhedral estimate}, and our present goal is to design and to analyse presumably good estimates of this type.

### 5.1.2 Generic polyhedral estimate

A generic polyhedral estimate is as follows:

Given the data \( A \in \mathbb{R}^{m \times n}, B \in \mathbb{R}^{\nu \times n}, X \subset \mathbb{R}^n \) of our recovery problem (where \( X \) is a computationally tractable convex compact set) and a “reliability tolerance” \( \epsilon \in (0, 1) \), we specify somehow positive integer \( N \) along with \( N \) linear forms \( h^T_\ell x \) on the space \( \mathbb{R}^m \) where observations live. These forms define linear forms \( g^T_\ell x := h^T_\ell Ax \) on the space of signals \( \mathbb{R}^n \). Assuming that the observation noise \( \xi_x \) is zero mean for every \( x \in X \), the "plug-in" estimates \( h^T_\ell \omega \) are unbiased estimates of the forms \( g^T_\ell x \). Assume that vectors \( h_\ell \) are selected in such a way that

\[
\forall (x \in X) : \text{Prob}\{ |h^T_\ell \xi_x | > 1 \} \leq \epsilon/N. \quad (5.1)
\]

In this situation, setting \( H = [h_1,....,h_N] \) (in the sequel, \( H \) is referred to as \textit{contrast matrix}), we can ensure that whatever be signal \( x \in X \) underlying our observation \( \omega = Ax + \xi_x \), the observable vector \( H^T \omega \) satisfies the relation

\[
\text{Prob} \{ \| H^T \omega - H^T Ax \|_\infty > 1 \} \leq \epsilon. \quad (5.2)
\]

With the polyhedral estimation scheme, we act as if all information about \( x \) contained in our observation \( \omega \) were represented by \( H^T \omega \), and we estimate \( Bx \) by \( B\bar{x} \), where \( \bar{x} = \bar{x}(\omega) \) is a (whatever) vector from \( X \) compatible with this information, specifically, such that \( \bar{x} \) is solves the feasibility problem

\[
\text{find } \bar{x} \in X \text{ such that } \| H^T \omega - H^T Ax \|_\infty \leq 1.
\]

Note that this feasibility problem with positive probability can be unsolvable; all we know in this respect is that the latter probability if \( \leq 1 - \epsilon \), since by construction the true signal \( x \) underlying observation \( \omega \) is with probability \( 1 - \epsilon \) a feasible solution. In other words, such \( \bar{x} \) is not always well defined. To circumvent this difficulty, let us define \( \bar{x} \) as

\[
\bar{x} \in \text{Argmin} \{ \| H^T \omega - H^T Au \|_\infty : u \in X \} \quad (5.3)
\]

so that \( \bar{x} \) always is well defined and belongs to \( X \), and estimate \( Bx \) by \( B\bar{x} \). Thus,

\[\text{a polyhedral estimate is specified by } m \times N \text{ contrast matrix } H = [h_1,....,h_N] \text{ with columns } h_\ell \text{ satisfying (5.1) and is as follows: given observation } \omega, \text{ we build } \bar{x} = \bar{x}(\omega) \in X \text{ according to (5.3) and estimate } Bx \text{ by } \bar{x}^H(\omega) = B\bar{x}(\omega).\]

The rationale behind polyhedral estimation scheme is the desire to reduce complex

\(^1\)In fact, this estimate is nearly optimal under the circumstances in a meaningful range of values of \( n \) and \( \sigma \).
estimating problems to those of estimating linear forms. To the best of our knowledge, this approach was first used in [190] (see also [182, Chapter 2]) in connection with recovering from direct observations (restrictions on regular grids of) multivariate functions from Sobolev balls. Recently, the ideas underlying the results of [190] have been taken up in the MIND estimator of [107], then applied to multiple testing in [200].

$(\epsilon, \|\cdot\|)$-risk. Given a desired “reliability tolerance” $\epsilon \in (0, 1)$, it is convenient to quantify the performance of polyhedral estimate by its $(\epsilon, \|\cdot\|)$-risk

$$ \text{Risk}_{\epsilon,\|\cdot\|}(\hat{x}(\cdot)|\mathcal{X}) = \inf \{ \rho : \text{Prob} \left( \| B x - \hat{x}(A x + \xi) \| > \rho \right) \leq \epsilon \forall x \in \mathcal{X} \}, \quad (5.4) $$

that is, the worst, over $x \in \mathcal{X}$, size of “$(1 - \epsilon)$-reliable $\| \cdot \|$-confidence interval” associated with the estimate $\hat{x}(\cdot)$.

An immediate observation is as follows:

**Proposition 5.1.** In the situation in question, denoting by $\mathcal{X}_\epsilon = \frac{1}{2}(\mathcal{X} - \mathcal{X})$ the symmetrization of $\mathcal{X}$, given a contrast matrix $H = [h_1, \ldots, h_N]$ with columns satisfying (5.1) the quantity

$$ \Re[H] = \max_z \left\{ \| B z \| : \| H^T A z \|_\infty \leq 2, z \in 2\mathcal{X}_\epsilon \right\} \quad (5.5) $$

is an upper bound on the $(\epsilon, \|\cdot\|)$-risk of the polyhedral estimate $\hat{x}^H(\cdot)$:

$$ \text{Risk}_{\epsilon,\|\cdot\|}(\hat{x}^H|\mathcal{X}) \leq \Re[H]. $$

**Proof** is immediate. Let us fix $x \in \mathcal{X}$, and let $\mathcal{E}$ be the set of all realizations of $\xi_x$ such that $\| H^T \xi_x \|_\infty \leq 1$, so that $P_x(\mathcal{E}) \geq 1 - \epsilon$ by (5.2). Let us fix a realization $\xi \in \mathcal{E}$ of the observation noise, and let $\omega = A x + \xi, \hat{x} = \bar{x}(A x + \xi)$. Then $u = x$ is a feasible solution to the optimization problem (5.3) with the value of the objective $\leq 1$, implying that the value of this objective at the optimal solution $\bar{x}$ to the problem is $\leq 1$ as well, so that $\| H^T A(x - \bar{x}) \|_\infty \leq 2$. Besides this, $z = x - \bar{x} \in 2\mathcal{X}_\epsilon$. We see that $z$ is a feasible solution to (5.5), whence $\| B [x - \bar{x}] \| = \| B x - \hat{x}^H(\omega) \| \leq \Re[H]$. It remains to note that the latter relation holds true whenever $\omega = A x + \xi$ with $\xi \in \mathcal{E}$, and the $P_x$-probability of the latter inclusion is at least $1 - \epsilon$, whatever be $x \in \mathcal{X}$. \qed

**What is ahead.** In what follows our focus will be on the following questions pertinent to the design of polyhedral estimates:

1. Given the data of our estimation problem and a tolerance $\delta \in (0, 1)$, how to find a set $\mathcal{H}_\delta$ of vectors $h \in \mathbb{R}^m$ satisfying the relation

$$ \forall(x \in \mathcal{X}) : \text{Prob} \left\{ h^T \xi_x \geq 1 \right\} \leq \delta. \quad (5.6) $$

With our approach, after the number $N$ of columns in a contrast matrix has been selected, we choose the columns of $H$ from $\mathcal{H}_\delta$, with $\delta = \epsilon/N$, $\epsilon$ being a given reliability tolerance of the estimate we are designing. Thus, the problem of constructing sets $\mathcal{H}_\delta$ arises, the larger $\mathcal{H}_\delta$, the better.

2. The upper bound $\Re[H]$ on the $(\epsilon, \|\cdot\|)$-risk of the polyhedral estimate $\hat{x}^H$ is, in general, difficult to compute – this is the maximum of a convex function over a computationally tractable convex set. Thus, similarly to the case of linear
estimates, we need techniques for computationally efficient upper bounding of \( R[h] \).

3. With “raw materials” (sets \( \mathcal{H}_\delta \)) and efficiently computable upper bounds on the risk of candidate polyhedral estimates at our disposal, how to design the best, in terms of (the upper bound on) its risk, polyhedral estimate?

We are about to consider these questions one by one.

5.1.3 Specifying sets \( \mathcal{H}_\delta \) for basic observation schemes

To specify reasonable sets \( \mathcal{H}_\delta \) we need to make some assumptions on the distributions of observation noises we want to handle. In the sequel we restrict ourselves with 3 special cases as follows:

- **sub-Gaussian case**: For every \( x \in \mathcal{X} \), the observation noise \( \xi_x \) is sub-Gaussian with parameters \((0, \sigma^2 I_m)\), where \( \sigma > 0 \), i.e. \( \xi_x \sim \mathcal{SG}(0, \sigma^2 I_m) \).
- **Discrete case**: \( \mathcal{X} \) is a convex compact subset of the probabilistic simplex \( \Delta_n = \{ x \in \mathbb{R}^n : x \geq 0, \sum_i x_i = 1 \} \), \( A \) is a column-stochastic matrix, and
  \[ \omega = \frac{1}{K} \sum_{k=1}^K \zeta_k \]
with independent across \( k \leq K \) random vectors \( \zeta_k \), \( \zeta_k \) taking value \( e_i \) with probability \( |Ax|_i, i = 1, \ldots, m, e_i \) being the basic orths in \( \mathbb{R}^m \).
- **Poisson case**: \( \mathcal{X} \) is a convex compact subset of the nonnegative orthant \( \mathbb{R}_+^n \), \( A \) is entrywise nonnegative, and the observation \( \omega \) stemming from \( x \in \mathcal{X} \) is random vector with independent across \( i \) entries \( \omega_i \sim \text{Poisson}(|Ax|_i) \).

The associated sets \( \mathcal{H}_\delta \) can be built as follows.

5.1.3.1 Sub-Gaussian case

When \( h \in \mathbb{R}^n \) is deterministic and \( \xi \) is sub-Gaussian with parameters \( 0, \sigma^2 I_m \), we have
\[
\text{Prob}\{ |h^T \xi| > 1 \} \leq 2 \exp\left\{ -\frac{1}{2\sigma^2||h||_2^2} \right\}.
\]
Indeed, when \( h \neq 0 \) and \( \gamma > 0 \), we have
\[
\text{Prob}\{ h^T \xi > 1 \} \leq \exp(\gamma) E \{ \exp\{\gamma h^T \xi\} \} \leq \exp\left\{ \frac{1}{2} \sigma^2 \gamma ||h||_2^2 - \gamma \right\}.
\]
Minimizing the resulting bound in \( \gamma > 0 \), we get \( \text{Prob}\{ h^T \xi > 1 \} \leq \exp\left\{ -\frac{1}{2||h||_2^2} \right\} \).

Minimizing the resulting bound in \( \gamma > 0 \), we get \( \text{Prob}\{ h^T \xi > 1 \} \leq \exp\left\{ -\frac{1}{2||h||_2^2} \right\} \).

Consequently
\[
\pi_G(h) := \sigma \sqrt{2 \ln(2/\delta)} ||h||_2 \leq 1 \Rightarrow \text{Prob}\{ h^T \xi > 1 \} \leq \delta,
\]
and we can set
\[
\mathcal{H}_\delta = \mathcal{H}_\delta^G := \{ h : \pi_G(h) \leq 1 \}.
\]
5.1.3.2 Discrete case

Given \( x \in \mathcal{X} \), setting \( \mu = Ax \) and \( \eta_k = \zeta_k - \mu \), we get

\[
\omega = Ax + \frac{1}{K} \sum_{k=1}^{K} \eta_k.
\]

Given \( h \in \mathbb{R}^m \),

\[
h^T \xi_x = \frac{1}{K} \sum_k h^T \eta_k.
\]

Random variables \( \chi_1, \ldots, \chi_K \) are independent zero mean and clearly satisfy

\[
\mathbb{E} \{ \chi_k^2 \} \leq \sum_i [Ax]_i h_i^2, \quad |\chi_k| \leq 2 \| h \|_{\infty}.
\]

When applying Bernstein’s inequality we get (cf. Exercise 4.19)

\[
\text{Prob}\{ |h^T \xi_x| > t \} \leq 2 \exp \left\{ -\frac{t^2}{2 \sum_i [Ax]_i h_i^2 + \frac{4}{3} \| h \|_{\infty}^2} \right\}.
\]

(5.7)

Setting

\[
\pi_D(h) = \sqrt{\vartheta_D} \max_{x \in \mathcal{X}} \sum_i [Ax]_i h_i^2 + \varrho_D \| h \|_{\infty},
\]

\[
\vartheta_D = 2 \sqrt{\frac{\ln(2/\delta)}{K}}, \quad \varrho_D = \frac{8 \ln(2/\delta)}{3K},
\]

after a completely straightforward computation, we conclude from (5.7) that

\[
\pi_D(h) \leq 1 \Rightarrow \text{Prob}\{ |h^T \xi_x| > 1 \} \leq \delta, \forall x \in \mathcal{X}.
\]

Thus, in the Discrete case we can set

\[
\mathcal{H}_\delta = \mathcal{H}_D^\delta := \{ h : \pi_D(h) \leq 1 \}.
\]

5.1.3.3 Poisson case

In the Poisson case, for \( x \in \mathcal{X} \), setting \( \mu = Ax \), we have

\[
\omega = Ax + \xi_x, \quad \xi_x = \omega - \mu.
\]

It turns out that for every \( h \in \mathbb{R}^m \) one has

\[
\forall t \geq 0 : \text{Prob}\{ |h^T \xi_x| \geq t \} \leq 2 \exp \left\{ -\frac{t^2}{2 \sum_i h_i^2 \mu_i + \frac{4}{3} \| h \|_{\infty} t} \right\}.
\]

(5.8)

---

\(^2\)The classical Bernstein inequality states that if \( X_1, \ldots, X_K \) are independent zero mean scalar random variables with finite variances \( \sigma_k^2 \) such that \( |X_k| \leq M \) a.s., then for every \( t > 0 \) one has

\[
\text{Prob}\{ X_1 + \ldots + X_K > t \} \leq \exp \left\{ -\frac{t^2}{2 \sum_k \sigma_k^2 + \frac{4}{3} Mt} \right\}.
\]
(for verification, see Exercise 4.21 or Section 5.4.1). As a result, we conclude via a straightforward computation that setting
\[
\pi_P(h) = \frac{\sqrt{\vartheta_P^2 \max_{x \in X} \sum_i |A_x| h_i^2 + \varrho_P^2 \|h\|_\infty^2}}{\vartheta_P^2 \max_{x \in X} \sum_i [Ax]_i h_i^2 + \varrho_P^2 \|h\|_\infty^2},
\]
we ensure that
\[
\pi_P(h) \leq 1 \Rightarrow \text{Prob}\{\|h^T \xi_x\| > 1\} \leq \delta, \; \forall x \in \mathcal{X}.
\]
Thus, in the Poisson case we can set
\[
\mathcal{H}_\delta = \mathcal{H}_\delta^P := \{h : \pi_P(h) \leq 1\}.
\]

5.1.4 Efficient upper-bounding of $\mathcal{R}[H]$ and contrast design, I.

The scheme for upper-bounding $\mathcal{R}[H]$ to be presented in this section (an alternative, completely different, scheme will be presented in Section 5.1.5) is inspired by our motivating example. Note that there is a special case of (5.5) where $\mathcal{R}[H]$ is easy to compute – the case where $\|\cdot\|$ is the uniform norm $\|\cdot\|_\infty$, whence
\[
\mathcal{R}[H] = \hat{\mathcal{R}}[H] := 2 \max_{i \leq \nu} \max_x \{\text{Row}^T_i [B] x : x \in \mathcal{X}, \|H^T A x\|_\infty \leq 1\}
\]
is the maximum of $\nu$ efficiently computable convex functions. It turns out that when $\|\cdot\| = \|\cdot\|_\infty$, it is not only easy to compute $\mathcal{R}[H]$, but to optimize this risk bound in $H$ as well.\(^3\) These observations underly the forthcoming developments in this section: under appropriate assumptions, we bound the risk of a polyhedral estimate with contrast matrix $H$ via the efficiently computable quantity $\hat{\mathcal{R}}[H]$ and then show that the resulting risk bounds can be efficiently optimized w.r.t. $H$. We shall also see that in some “simple for analytical analysis” situations, like that of the example, the resulting estimates are nearly minimax optimal.

5.1.4.1 Assumptions

We stay within the setup introduced in Section 5.1.1 which we augment with the following assumptions:

**A.1.** $\|\cdot\| = \|\cdot\|_r$ with $r \in [1, \infty].$

**A.2.** We have at our disposal a sequence $\gamma = \{\gamma_i > 0, i \leq \nu\}$ and $\rho \in [1, \infty]$ such that the image of $\mathcal{X}$ under the mapping $x \mapsto B x$ is contained in the “scaled $\|\cdot\|_\rho$-ball”
\[
\mathcal{V} = \{y \in \mathbb{R}^\nu : \|\text{Diag}\{\gamma\} y\|_\rho \leq 1\}.
\]

\(^3\)On closer inspection, in the situation considered in the motivating example the $\|\cdot\|_\infty$-optimal contrast matrix $H$ is proportional to the unit matrix, and the quantity $\hat{\mathcal{R}}[H]$ can be easily translated into an upper bound on, say, $\|\cdot\|_2$-risk of the associated polyhedral estimate.
Proposition 5.2. In the situation described in Section 5.1.1, let us assume that Assumptions A.1-2 hold. Let \( \epsilon \in (0, 1) \) and let a positive real \( N \geq \nu \) be given; let also \( \pi(\cdot) \) be a norm on \( \mathbb{R}^m \) such that

\[
\forall (h : \pi(h) \leq 1, x \in \mathcal{X}) : \text{Prob}\{ |h^T \xi | > 1 \} \leq \epsilon/N.
\]

Next, let a matrix \( H = [H_1, ..., H_\nu] \) with \( H_\ell \in \mathbb{R}^{m \times m_\ell}, m_\ell \geq 1, \) and positive reals \( \varsigma_\ell, \ell \leq \nu, \) satisfy the relations

\[
\begin{align*}
(a) \quad & \pi(\text{Col}_j[H]) \leq 1, 1 \leq j \leq N; \\
(b) \quad & \max_x \{ B^T_\ell x : x \in \mathcal{X}_s, \| H^T_\ell Ax \|_\infty \leq \varsigma_\ell, 1 \leq \ell \leq \nu \} \leq \varsigma_\ell, 1 \leq \ell \leq \nu.
\end{align*}
\]

(5.10)

Then the quantity \( \mathfrak{R}[H] \) as defined in (5.5) can be upper-bounded as follows:

\[
\mathfrak{R}[H] \leq \Psi(\varsigma) := 2\max_w \{ \| [w_1/\gamma_1 ; \ldots ; w_\nu/\gamma_\nu] \|_\rho : \| w \|_\rho \leq 1, 0 \leq w_\ell \leq \gamma \varsigma_\ell, \ell \leq \nu \},
\]

(5.11)

which combines with Proposition 5.1 to imply that

\[
\text{Risk}_{\epsilon,\| \cdot \| \rightarrow \| \cdot \|} [\hat{x}_H|\mathcal{X}] \leq \Psi(\varsigma).
\]

(5.12)

Function \( \Psi \) is nondecreasing on the nonnegative orthant and is easy to compute.

Proof. Let \( z = 2\bar{z} \) be a feasible solution to (5.5), thus \( \bar{z} \in \mathcal{X} \) and \( \| H^T A\bar{z} \|_\infty \leq 1. \) Let \( y = B\bar{z}, \) so that \( y \in \mathcal{Y} \) (see (5.9)) due to \( \bar{z} \in \mathcal{X} \) and A.2. Then \( \| \text{Diag} \{ \gamma \} y \|_\rho \leq 1. \) Besides this, by (5.10.b) relations \( \bar{z} \in \mathcal{X} \) and \( \| H^T A\bar{z} \|_\infty \leq 1 \) combine with the symmetry of \( \mathcal{X} \) w.r.t. the origin to imply that

\[
|y_\ell| = |B^T_\ell \bar{z}| \leq \varsigma_\ell, \ell \leq \nu.
\]

Taking into account that \( \| \cdot \| = \| \cdot \|_r \) by A.1, we see that

\[
\mathfrak{R}[H] = \max_{y} \{ \| B\bar{z} \|_r : z \in 2\mathcal{X}_s, \| H^T A\bar{z} \|_\infty \leq 2 \}
\leq 2\max_y \{ \| y \|_r : |y_\ell| \leq \varsigma_\ell, \ell \leq \nu, \text{ and } \| \text{Diag} \{ \gamma \} y \|_\rho \leq 1 \}
\leq 2\max_w \{ \| [w_1/\gamma_1 ; \ldots ; w_\nu/\gamma_\nu] \|_r : \| w \|_\rho \leq 1, 0 \leq w_\ell \leq \gamma \varsigma_\ell, \ell \leq \nu \},
\]

as stated in (5.11).

It is evident that \( \Psi \) is nondecreasing on the nonnegative orthant. Computing \( \Psi \) can be carried out as follows:

1. When \( r = \infty, \) we need to compute \( \max_{\ell \leq \nu} \max_w \{ w_\ell/\gamma_\ell : \| w \|_\rho \leq 1, 0 \leq w_\ell \leq \gamma \varsigma_\ell, j \leq \nu \} \) so that evaluating \( \Psi \) reduces to solving \( \nu \) simple convex optimization problems;
2. When \( \rho = \infty, \) we clearly have \( \Psi(\varsigma) = \| \bar{w}_1/\gamma_1 ; \ldots ; \bar{w}_\nu/\gamma_\nu \|_r, \bar{w}_\ell = \min[1, \gamma \varsigma_\ell]; \)
3. When \( 1 \leq r, \rho < \infty, \) passing from variables \( w_\ell \) to variables \( u_\ell = w_\ell^\rho, \) we get

\[
\Psi^r(\varsigma) = 2^r \max_u \left\{ \sum_\ell \gamma_\ell^{-r} u_\ell^{r/\rho} : \sum_\ell u_\ell \leq 1, 0 \leq u_\ell \leq (\gamma \varsigma_\ell)^{\rho} \right\}.
\]

When \( r \leq \rho, \) optimization problem on the right hand side is the easily solvable problem of maximizing a simple concave function over a simple convex compact set. When \( \infty > r > \rho, \) this problem can be solved by Dynamic Programming. \( \Box \)
Comment. When we want to recover $Bx$ in $\|\cdot\|_\infty$ (i.e., we are in the case of $r = \infty$), under the premise of Proposition 5.2 we clearly have $\Psi(\varsigma) \leq \max_{\ell} \varsigma_\ell$, resulting in the bound
\[
\text{Risk}_{\epsilon,\|\cdot\|_\infty} \left[ \mathbb{E}^H |\mathcal{X}| \right] \leq 2 \max_{\ell} \varsigma_\ell.
\]
Note that this bound in fact does not require Assumption A.2 (since it is satisfied for any $\rho$ with large enough $\gamma_i$'s).

5.1.4.3 Specifying contrasts

Risk bound (5.12) allows for an easy design of contrast matrices. Recalling that $\Psi$ is monotone on the nonnegative orthant, all we need is to select $h_\ell$'s satisfying (5.10) and resulting in the smallest possible $\varsigma_\ell$'s, which is what we are about to do now.

Preliminaries. Given a vector $b \in \mathbb{R}^m$ and a norm $s(\cdot)$ on $\mathbb{R}^m$, consider convex-concave saddle point problem
\[
\text{Opt} = \inf_{g \in \mathbb{R}^m} \max_{x \in \mathcal{X}_s} \{ \phi(g, x) := [b - A^T g]^T x + s(g) \} \quad (SP)
\]
along with the induced primal and dual problems
\[
\text{Opt}(P) = \inf_{g \in \mathbb{R}^m} \left[ \bar{\phi}(g) := \max_{x \in \mathcal{X}_s} \phi(g, x) \right] = \inf_{g \in \mathbb{R}^m} \left[ s(g) + \max_{x \in \mathcal{X}_s} [b - A^T g]^T x \right], \quad (P)
\]
and
\[
\text{Opt}(D) = \max_{x \in \mathcal{X}_s} \left[ \bar{\phi}(g) := \inf_{g \in \mathbb{R}^m} \phi(g, x) \right] = \max_{x \in \mathcal{X}_s} \left[ \inf_{g \in \mathbb{R}^m} [b^T x - [A^T g + s(g)]] \right] = \max_{x \in \mathcal{X}_s} [b^T x : x \in \mathcal{X}_s, g(Ax) \leq 1] \quad (D)
\]
where $g(\cdot)$ is the norm conjugate to $s(\cdot)$ (we have used the evident fact that $\inf_{g \in \mathbb{R}^m} [f^T g + s(g)]$ is either $-\infty$ or 0 depending on whether $q(f) > 1$ or $q(f) \leq 1$).

Since $\mathcal{X}_s$ is compact, we have $\text{Opt}(P) = \text{Opt}(D) = \text{Opt}$ by Sion-Kakutani Theorem. Besides this, $(D)$ is solvable (evident) and $(P)$ is solvable as well, since $\bar{\phi}(g)$ is continuous due to the compactness of $\mathcal{X}_s$ and $\bar{\phi}(g) \geq s(g)$, so that $\bar{\phi}(\cdot)$ has bounded level sets. Let $\bar{g}$ be an optimal solution to $(P)$, $\bar{x}$ be an optimal solution to $(D)$, and let $\bar{h}$ be $s(\cdot)$-unit normalization of $\bar{g}$, so that $s(\bar{h}) = 1$ and $\bar{g} = s(\bar{g}) \bar{h}$. Now let us make the following observation:

Observation 5.3. In the situation in question, we have
\[
\max_{x} \{|b^T x : x \in \mathcal{X}_s, |\bar{h}^T Ax| \leq 1\} \leq \text{Opt}. \quad (5.13)
\]
In addition, for any matrix $G = [g_1, ..., g_M] \in \mathbb{R}^{m \times M}$ with $s(g_j) \leq 1$, $j \leq M$, one has
\[
\max_{x} \{|b^T x : x \in \mathcal{X}_s, ||G^T Ax||_\infty \leq 1\} = \max_{x} \{|b^T x : x \in \mathcal{X}_s, ||G^T Ax||_\infty \leq 1\} \geq \text{Opt}. \quad (5.14)
\]

Proof. Let $x$ be a feasible solution to the problem in (5.13). Replacing, if
necessary, \( x \) with \(-x\), we can assume that \(|b^T x| = b^T x\). We now have

\[
|b^T x| = b^T x = [g^T Ax - s(\bar{g})] + [\bar{b} - A^T \bar{g}]^T x + s(\bar{g}) \leq \varphi(\bar{g}) = \text{Opt}(P) \leq \text{Opt}(P) + s(\bar{g}) [\bar{h}^T Ax - s(\bar{g})] \leq 1
\]

\[
\leq \text{Opt}(P) = \text{Opt},
\]

as claimed in (5.13). Now, the equality in (5.14) is due to the symmetry of \( \mathcal{X} \) w.r.t. the origin. To verify the inequality in (5.14), note that \( \bar{x} \) satisfies the relations \( \bar{x} \in \mathcal{X} \) and \( q(A \bar{x}) \leq 1 \), implying, due to the fact that the columns of \( G \) are of \( s(\cdot) \)-norm \( \leq 1 \), that \( \bar{x} \) is a feasible solution to the optimization problems in (5.14). As a result, the second quantity in (5.14) is at least \( b^T \bar{x} = \text{Opt}(D) = \text{Opt} \), and (5.14) follows.

**Comment.** Note that problem \( (P) \) has a very transparent origin. In the situation of Section 5.1.1, assume that our goal is, to estimate, given observation \( \omega = Ax + \xi \), the value at \( x \in \mathcal{X} \) of the linear function \( b^T x \), and we want to use for this purpose affine in \( \omega \) estimate \( \hat{g}(\omega) = g^T \omega + \gamma \). Given \( \epsilon \in (0, 1) \), how to construct a presumably good, in terms of its \( \epsilon \)-risk, estimate? Let us show that a meaningful answer is yielded by optimal solution to \( (P) \). Indeed, we have

\[
b^T x - \hat{g}(Ax + \xi) = [b - A^T g]^T x - \gamma - g^T \xi.
\]

Assume that we have at our disposal a norm \( s(\cdot) \) on \( \mathbb{R}^m \) such that

\[
\forall (h \in \mathbb{R}^m, s(h) \leq 1, x \in \mathcal{X}) : \text{Prob}\{\xi_x : |h^T \xi_x| > 1\} \leq \epsilon,
\]

or, which is the same,

\[
\forall (g \in \mathbb{R}^m, x \in \mathcal{X}) : \text{Prob}\{\xi_x : |g^T \xi_x| > s(g)\} \leq \epsilon.
\]

Then we can safely upper-bound the \( \epsilon \)-risk of a candidate estimate \( \hat{g}(\cdot) \) by the quantity

\[
\rho = \max_{x \in \mathcal{X}} [b - A^T g]^T x - \gamma + s(g).
\]

Observe that for \( g \) fixed, the minimal, over \( \gamma \), bias is nothing but

\[
M(g) := \max_{x \in \mathcal{X}} |b - A^T g| x.
\]

Postponing verification of this claim, here is the conclusion:

*in the present setting, problem \( (P) \) is nothing but the problem of building the best, in terms of the upper bound \( \rho \) on \( \epsilon \)-risk, affine estimate of linear function \( b^T x \).*

It remains to justify the above claim, which is immediate: on one hand, for all \( u \in \mathcal{X}, v \in \mathcal{X} \) we have

\[
B(g, \gamma) \geq [b - A^T g]^T u - \gamma, \quad B(g, \gamma) \geq -[b - A^T g]^T v + \gamma
\]
implying that
\[ B(g, \gamma) \geq \frac{1}{2} [b - A^T g]^T [u - v] \forall (u \in \mathcal{X}, v \in \mathcal{X}), \]
same as \( B(g, \gamma) \geq M(g) \). On the other hand, let
\[ M_+ (g) = \max_{x \in \mathcal{X}} [b - A^T g]^T x, \quad M_- (g) = - \min_{x \in \mathcal{X}} [b - A^T g]^T x, \]
so that \( M(g) = \frac{1}{2} [M_+(g) + M_-(g)] \). Setting \( \tilde{\gamma} = \frac{1}{2} [M_+(g) - M_-(g)] \), we have
\[
\begin{align*}
\max_{x \in \mathcal{X}} [b - A^T g]^T x - \tilde{\gamma} &= M_+(g) - \tilde{\gamma} = \frac{1}{2} [M_+(g) + M_-(g)] = M(g), \\
\min_{x \in \mathcal{X}} [b - A^T g]^T x - \tilde{\gamma} &= -M_-(g) + \tilde{\gamma} = -\frac{1}{2} [M_+(g) + M_-(g)] = -M(g).
\end{align*}
\]
That is, \( B(g, \tilde{\gamma}) = M(g) \). Combining these observations, we arrive at \( \min_{\gamma} B(g, \gamma) = M(g) \), as claimed. \( \square \)

**Contrast design.** Proposition 5.2, Observation 5.3 allow for a straightforward solution of the associated contrast design problem, at least in the case of sub-Gaussian, Discrete, and Poisson observation schemes. Indeed, in these cases, when designing contrast matrix with Gaussian, Discrete, and Poisson observation schemes, it follows that we lose nearly nothing when assuming that \( N \geq \nu \). Let us act as follows:

We set \( N = \nu \), specify \( \bar{\pi} (\cdot) \) as the norm \( (\pi_G, \pi_D, \pi_P) \) associated with the observation scheme (sub-Gaussian, or Discrete, or Poisson) in question and \( \delta = \epsilon/\nu \). We solve \( \nu \) convex optimization problems

\[
\begin{align*}
\text{Opt}_{\ell} &= \min_{g \in \mathbb{R}^m} \{ \bar{\phi}_{\ell} (g) := \max_{x \in \mathcal{X}, s} \phi_{\ell} (g, x) \} \\
\phi_{\ell} (g, x) &= [B_{\ell} - A^T g]^T x + \bar{\pi} (g).
\end{align*}
\]

Next, we convert optimal solution \( g_{\ell} \) to \( (P_\ell) \) into vector \( h_{\ell} \in \mathbb{R}^m \) by representing \( g_{\ell} = \bar{\pi} (g_{\ell}) h_{\ell} \) with \( \bar{\pi} (h_{\ell}) = 1 \), and set \( H_{\ell} = h_{\ell} \). As a result, we obtain \( m \times \nu \) contrast matrix \( H = [h_1, ..., h_\nu] \) which, taken along with \( N = \nu \), quantities
\[
\varsigma_{\ell} = \text{Opt}_{\ell}, \quad 1 \leq \ell \leq \nu,
\]
and with \( \pi (\cdot) \equiv \bar{\pi} (\cdot) \), in view of the first claim in Observation 5.3 as applied with \( s (\cdot) \equiv \bar{\pi} (\cdot) \) satisfies the premise of Proposition 5.2.

Consequently, by Proposition 5.2 we have
\[
\text{Risk}_{\mathcal{X}, \| \cdot \|} [\hat{\pi}^H | \mathcal{X}] \leq \Psi ([\text{Opt}_1; ..., \text{Opt}_\nu]). \tag{5.16}
\]

**Comment.** Optimality of the outlined contrast design for sub-Gaussian, or Discrete, or Poisson observation scheme stems, within the framework set by Proposition 5.2, from the second claim of Observation 5.3 which states that when \( N \geq \nu \) and the columns of \( m \times N \) contrast matrix \( H = [H_1, ..., H_\nu] \) belong to the set \( \mathcal{H}_{c/N} \) associated with the observation scheme in question. I.e., the norm \( \pi (\cdot) \) in Proposition is the norm \( \pi_G, \pi_D, \) or \( \pi_P \) associated with \( \delta = \epsilon/N \), the quantities \( \varsigma_{\ell} \) participating in (5.10.b) cannot be less than \( \text{Opt}_{\ell} \).
Indeed, the norm $\pi(\cdot)$ from Proposition 5.2 is $\geq$ the norm $\bar{\pi}(\cdot)$ participating in $(P_\ell)$ (because the value $\epsilon/N$ in the definition of $\pi(\cdot)$ is at most $\frac{\epsilon}{\nu}$), implying, by (5.10.a), that the columns of matrix $H$ obeying the premise of Proposition satisfy the relation $\bar{\pi}(\text{Col}_j[H]) \leq 1$. Invoking the second part of Observation 5.3 with $s(\cdot) \equiv \bar{\pi}(\cdot)$, $b = B_\ell$, and $G = H_\ell$, and taking (5.10.b) into account we conclude that $\varsigma_\ell \geq \text{Opt}_\ell$ for all $\ell$, as claimed.

Since the bound on the risk of a polyhedral estimate offered by Proposition 5.2 is the better the less are $\varsigma_\ell$'s, we see that as far as this bound is concerned, the outlined design procedure is the best possible, provided $N \geq \nu$.

An attractive feature of the contrast design we have just presented is that it is completely independent of the entities participating in assumptions A.1-2 – these entities affect theoretical risk bounds of the resulting polyhedral estimate, but not the estimate itself.

5.1.4.4 Illustration: diagonal case

Let us consider the diagonal case of our estimation problem, where

- $X = \{ x \in \mathbb{R}^n : \| Dx \|_\rho \leq 1 \}$, where $D$ is a diagonal matrix with positive diagonal entries $D_{\ell \ell} =: d_\ell$,
- $m = \nu = n$, and $A$ and $B$ are diagonal matrices with diagonal entries $0 < A_{\ell \ell} =: a_\ell, 0 < B_{\ell \ell} =: b_\ell$,
- $\| \cdot \|_r = \| \cdot \|_2$,
- We consider sub-Gaussian case, that is, observation noise $\xi_x$ is $(0, \sigma^2 I_n)$-sub-Gaussian for every $x \in X$.

Let us implement the approach developed in Sections 5.1.4.1 – 5.1.4.3.

1. Given reliability tolerance $\epsilon$, we set
   \[ \delta = \epsilon/n, \quad \vartheta_G := \sigma \sqrt{2 \ln(2/\delta)} = \sigma \sqrt{2 \ln(2n/\epsilon)} \],
   \[ \text{and} \quad H = H_G^{\delta} = \{ h \in \mathbb{R}^n : \pi_G(h) := \vartheta_G \| h \|_2 \leq 1 \} \]

2. We solve $\nu = n$ convex optimization problems $(P_\ell)$ associated with $\bar{\pi}(\cdot) \equiv \pi_G(\cdot)$, which is immediate: the resulting contrast matrix is $H = \vartheta_G^{-1} I_n$, and
   \[ \text{Opt}_\ell = \varsigma_\ell := b_\ell \min\{ \vartheta_G/a_\ell, 1/d_\ell \} \].

Risk analysis. The $(\epsilon, \| \cdot \|_r)$-risk of the resulting polyhedral estimate $\hat{x}(\cdot)$ can be bounded by Proposition 5.2. Note that setting $\gamma_\ell = d_\ell/b_\ell, 1 \leq \ell \leq n$, we meet assumptions A.1-2, and the above choice of $H, N = n$ and $\varsigma_\ell$ satisfies the premise of Proposition 5.2. By this proposition,

\[ \text{Risk}_{\epsilon, \| \cdot \|_r} [\hat{x}^H | X] \leq \Psi := 2 \max \{ \| \tilde{w}_1/\gamma_1, \ldots, \tilde{w}_n/\gamma_n \|_r : \| \tilde{w}_\rho \|_r \leq 1, 0 \leq \tilde{w}_\ell \leq \gamma_\ell \varsigma_\ell \} \].

(5.19)
Let us work out what happens in the simple case where

\[(a) \quad 1 \leq \rho \leq r < \infty, \quad (b) \quad a_{\ell}/d_{\ell} \text{ and } b_{\ell}/a_{\ell} \text{ are nonincreasing in } \ell.\]  

(5.20)

**Proposition 5.4.** In the just defined simple case, let \(n = n\) when

\[\sum_{\ell=1}^{n} (\vartheta_{G} d_{\ell}/a_{\ell})^\rho \leq 1,\]

otherwise let \(n\) be the smallest integer such that

\[\sum_{\ell=1}^{n} (\vartheta_{G} d_{\ell}/a_{\ell})^\rho > 1,\]

with \(\vartheta_{G}\) given by (5.17). Then for the contrast matrix \(H = \vartheta_{G}^{-1} I_n\) one has

\[\text{Risk}_{\epsilon,\|\cdot\|_r} \left[ \hat{x}^H |X\right] \leq \Psi \leq 2 \left[ \sum_{\ell=1}^{n} (\vartheta_{G} b_{\ell}/a_{\ell})^\rho \right]^{1/r}\]

**Proof.** Consider the optimization problem specifying \(\Psi\) in (5.19). Setting \(\theta = r/\rho \geq 1\), let us pass in this problem from variables \(w_{\ell}\) to variables \(z_{\ell} = w_{\ell}^{\rho}\), so that

\[\Psi^r = 2^r \max_{\tilde{z}} \left\{ \sum_{\ell} z_{\ell}^{\theta} (b_{\ell}/d_{\ell})^r : \sum_{\ell} z_{\ell} \leq 1, 0 \leq z_{\ell} \leq (d_{\ell}/b_{\ell})^\rho \right\} \leq 2^r \Gamma,\]

where

\[\Gamma = \max_{\tilde{z}} \left\{ \sum_{\ell} z_{\ell}^{\theta} (b_{\ell}/d_{\ell})^r : \sum_{\ell} z_{\ell} \leq 1, 0 \leq z_{\ell} \leq \chi_{\ell} := (\vartheta_{G} d_{\ell}/a_{\ell})^\rho \right\}\]

(we have used (5.18)). Note that \(\Gamma\) is the optimal value in the problem of maximizing a convex (since \(\theta \geq 1\)) function \(\sum_{\ell} z_{\ell}^{\theta} (b_{\ell}/d_{\ell})^r\) over a bounded polyhedral set, so that the maximum is attained at an extreme point \(\tilde{z}\) of the feasible set. By the standard characterization of extreme points, the (clearly nonempty) set \(I'\) of positive entries in \(\tilde{z}\) is as follows: denoting by \(I'\) the set of indexes \(\ell \in I\) such that \(z_{\ell}\) is on its upper bound \(z_{\ell} = \chi_{\ell}\), the cardinality \(|I'|\) of \(I'\) is at least \(|I| - 1\). Since

\[\sum_{\ell \in I'} \tilde{z}_{\ell} = \sum_{\ell \in I'} \chi_{\ell} \leq 1 \quad \text{and} \quad \chi_{\ell} \text{ are nondecreasing in } \ell \text{ by (5.20.b), we conclude that}\]

\[\sum_{\ell = 1}^{\ell = 1} \chi_{\ell} \leq 1,\]

implying that \(|I'| < n\) provided that \(n < n\), so that in this case \(|I| \leq n\); and of course \(|I| \leq n\) when \(n = n\). Next, we have

\[\Gamma = \sum_{\ell \in I} z_{\ell}^{\theta} (b_{\ell}/d_{\ell})^r \leq \sum_{\ell \in I} (b_{\ell}/d_{\ell})^r = \sum_{\ell \in I} \vartheta_{G} b_{\ell}/a_{\ell})^r,\]

and since \(b_{\ell}/a_{\ell}\) is nonincreasing in \(\ell\) and \(|I| \leq n\), the latter quantity is at most
\[\sum_{\ell=1}^{n} \left( \vartheta_G b_{\ell} / a_{\ell} \right)^r.\]

**Application.** Consider the “standard case” \([71, 73]\) where
\[0 < \sqrt{\ln(2n/\epsilon)} \sigma \leq 1, a_{\ell} = \ell^{-\alpha}, b_{\ell} = \ell^{-\beta}, d_{\ell} = \ell^\kappa\]
with \(\beta \geq \alpha \geq 0, \kappa \geq 0\) and \((\beta - \alpha)r < 1\). In this case for large \(n\), namely,
\[n \geq c \vartheta_G \frac{1}{\alpha + \beta + r + \kappa + 1} [\vartheta_G = \sigma \sqrt{2 \ln(2n/\epsilon)}]\]  (5.21)
(here and in what follows, the factors denoted \(c\) and \(C\) depend solely on \(\alpha, \beta, \kappa, r, \rho\))
we get
\[n \leq C \vartheta_G \frac{1}{\alpha + \beta + r + \kappa + 1},\]
resulting in
\[\text{Risk}_{\epsilon, \|\cdot\|r}[\hat{x}|X] \leq C \vartheta_G \frac{\beta + \kappa + 1}{\alpha + \beta + r + \kappa + 1}.\]  (5.22)

Setting \(x = D^{-1}y, \bar{\alpha} = \alpha + \kappa, \bar{\beta} = \beta + \kappa\) and treating \(y\), rather than \(x\), as the signal underlying the observation, we obtain the estimation problem which is similar to the original one in which \(\alpha, \beta, \kappa\) and \(X\) are replaced, respectively, with \(\bar{\alpha}, \bar{\beta}, \bar{\kappa} = 0\) and \(Y = \{y : \|y\|_\rho \leq 1\}\), and \(A, B\) replaced with \(\bar{A} = \text{Diag}\{\ell^{-\bar{\alpha}}, \ell \leq n\}, \bar{B} = \text{Diag}\{\ell^{-\bar{\beta}}, \ell \leq n\}\). When \(n\) is large enough, namely, \(n \geq \sigma^{-\frac{1}{\beta + 1}}\), \(Y\) contains the “coordinate box”
\[\bar{Y} = \{x : |x_\ell| \leq m^{-1/\rho}, m/2 \leq \ell \leq m, x_\ell = 0 \text{ otherwise}\}\]
of dimension \(\geq m/2\), where
\[m \geq c \sigma^{-\frac{1}{\beta + 1}}.\]

Observe that for all \(y \in \bar{Y}\), \(\|\bar{A}y\|_2 \leq C m^{-\bar{\alpha}} \|y\|_2\), and \(\|\bar{B}y\|_r \geq c m^{-\bar{\beta}} \|y\|_r\). This observation, when combined with the Fano inequality, implies (cf. \([79]\)) that for \(\epsilon \ll 1\) the minimax optimal w.r.t. the family of all Borel estimates \((\epsilon, \|\cdot\|_r)\)-risk on the signal set \(\bar{X} = D^{-1}\bar{Y} \subset X\) is at least
\[c \sigma^{\frac{\beta + 1}{\alpha + \beta + r + \kappa + 1}}.\]

In other words, in this situation, the upper bound (5.22) on the risk of the polyhedral estimate is within a logarithmic in \(n/\epsilon\) factor from the minimax risk. In particular, without surprise, in the case of \(\beta = 0\) the polyhedral estimates attain well known optimal rates \([71, 107]\).

### 5.1.5 Efficient upper-bounding of \(\mathcal{R}[H]\) and contrast design, II.

#### 5.1.5.1 Outline

In this section we develop an alternative approach to the design of polyhedral estimates which resembles in many aspects the approach to building linear estimates from Chapter 4. Recall that the principal technique underlying the design of a presumably good linear estimate \(\hat{x}_H(\omega) = H^T \omega\) was upper-bounding of maximal risk of the estimate – the maximum of a quadratic form, depending on \(H\) as a parameter, over the signal set \(X\), and we were looking for a bounding scheme
allowing to efficiently optimize the bound in $H$.

The design of a presumably good polyhedral estimate also reduces to minimizing the optimal value in a parametric maximization problem (5.5) over the contrast matrix $H$. However, while design of presumably good linear estimate reduces to unconstrained minimization, to conceive a polyhedral estimate we need to minimize bound $\mathcal{R}[H]$ on the estimation risk under the restriction on the contrast matrix $H$ — the columns $h_\ell$ of this matrix should satisfy condition (5.1). In other words, in the case of polyhedral estimate the “design parameter” affects the constraints of the optimization problem rather than the objective.

**Our strategy** can be outlined as follows. Let us denote by

$$B_* = \{u \in \mathbb{R}^\nu : \|u\|_* \leq 1\}$$

the unit ball of the norm $\|\cdot\|_*$, conjugate to the norm $\|\cdot\|$ in the formulation of the estimation problem in Section 5.1.2. Assume we have at our disposal a technique for bounding quadratic forms on the set $B_* \times \mathcal{X}_s$, so that we have at our disposal an efficiently computable convex function $\mathcal{M}(\lambda)$ on $\mathbb{S}^{\nu+n}$ such that

$$\mathcal{M}(\lambda) \geq \max_{[u; z] \in B_* \times \mathcal{X}_s} [u; z]^T M[u; z] \forall M \in \mathbb{S}^{\nu+n}. \quad (5.23)$$

Note that the upper bound $\mathcal{R}[H]$, as defined in (5.5), on the risk of a candidate polyhedral estimate $\hat{x}^H$ is nothing but

$$\mathcal{R}[H] = 2 \max_{[u; z]} \left\{ [u; z]^T \left[ \begin{array}{c} \frac{1}{2} B^T \\ \frac{1}{2} B \end{array} \right] [u; z] : \begin{array}{l}
  u \in B_*, z \in \mathcal{X}_s, \\
  z^T A^T h_\ell h_\ell^T A z \leq 1, \ell \leq N
\end{array} \right\}. \quad (5.24)$$

Given $\lambda \in \mathbb{R}_+^N$, the constraints $z^T A^T h_\ell h_\ell^T A z \leq 1$ in (5.24) can be aggregated to yield the quadratic constraint

$$z^T A^T \Theta \lambda A z \leq \mu_\lambda, \quad \Theta \lambda = H \text{Diag}\{\lambda\} H^T, \mu_\lambda = \sum_\ell \lambda_\ell.$$

Observe that for every $\lambda \geq 0$ we have

$$\mathcal{R}[H] \leq 2 \mathcal{M}\left( \begin{array}{c} \frac{1}{2} B^T \\ -A^T \Theta \lambda A \end{array} \right)_{B_+} + 2 \mu_\lambda. \quad (5.25)$$

Indeed, let $[u; z]$ be a feasible solution to the optimization problem (5.24) specifying $\mathcal{R}[H]$. Then

$$[u; z]^T B_+ [u; z] = [u; z]^T B_+[\Theta \lambda][u; z] + z^T A^T \Theta \lambda A z;$$

the first term in the right hand side is $\leq \mathcal{M}(B_+[\Theta \lambda])$ since $[u; z] \in B_* \times \mathcal{X}_s$, and the second term in the right hand side, as we have already seen is $\leq \mu_\lambda$, and (5.25) follows.
Now assume that we have at our disposal a computationally tractable cone

\[ \mathbf{H} \subset \mathbb{S}_+^N \times \mathbb{R}_+ \]

satisfying the following assumption:

**C.** Whenever \((\Theta, \mu) \in \mathbf{H}\), we can efficiently find an \(n \times N\) matrix \(H = [h_1, ..., h_N]\) and a nonnegative vector \(\lambda \in \mathbb{R}_+^N\) such that

\[
\begin{align*}
(a) & \quad h_\ell \text{ satisfies } (5.1), \quad 1 \leq \ell \leq N, \\
(b) & \quad \Theta = H \text{Diag}\{\lambda\} H^T, \\
(c) & \quad \sum_\ell \lambda_\ell \leq \mu.
\end{align*}
\]

(5.26)

The following simple observation is crucial to what follows:

**Proposition 5.5.** Consider the estimation problem posed in Section 5.1.1, and let efficiently computable convex function \(\mathcal{M}\) and computationally tractable closed convex cone \(\mathbf{H}\) satisfy (5.23) and Assumption C, respectively. Consider the convex optimization problem

\[
\text{Opt} = \min_{\tau, \Theta, \mu} \left\{ 2\tau + 2\mu : (\Theta, \mu) \in \mathbf{H}, \mathcal{M}(B_+[\Theta]) \leq \tau \right\}
\]

with \(B_+[\Theta] = \left[ \frac{1}{2}B^T + \frac{1}{2}A^T \Theta A \right]\) (5.27)

Given a feasible solution \((\tau, \Theta, \mu)\) to this problem, by C we can efficiently convert it to \((H, \lambda)\) such that \(H = [h_1, ..., h_N]\) with \(h_\ell\) satisfying (5.1) and \(\lambda \geq 0\) with \(\sum_\ell \lambda_\ell \leq \mu\). We have

\[ \mathfrak{R}[H] \leq 2\tau + 2\mu, \]

whence the \((\epsilon, \| \cdot \|)\)-risk of the polyhedral estimate \(\hat{x}^H\) satisfies the bound

\[ \text{Risk}_{\epsilon, \| \cdot \|}[\hat{x}^H|\mathcal{X}] \leq 2\tau + 2\mu. \]

(5.28)

Consequently, we can efficiently construct polyhedral estimates with \((\epsilon, \| \cdot \|)\)-risk arbitrarily close to \(\text{Opt}\) (and with risk exactly \(\text{Opt}\), provided problem (5.27) is solvable).

**Proof** is readily given by the reasoning preceding the proposition. Indeed, with \(\tau, \Theta, \mu, H, \lambda\) as in the premise of the proposition, the columns \(h_\ell\) of \(H\) satisfy (5.1) by C, implying, by Proposition 5.1, that \(\text{Risk}_{\epsilon, \| \cdot \|}[\hat{x}^H|\mathcal{X}] \leq \mathfrak{R}[H]\). Besides this, C says that for our \(H, \lambda\) it holds \(\Theta = \Theta_\lambda\) and \(\mu_\lambda \leq \mu\), so that (5.25) combines with the constraints of (5.27) to imply that \(\mathfrak{R}[H] \leq 2\tau + 2\mu\), and (5.28) follows by Proposition 5.1.

The approach to the design of polyhedral estimates we develop in this section amounts to reducing the construction of the estimate (i.e., construction of the contrast matrix \(H\)) to finding (nearly) optimal solutions to (5.27). Implementing this approach requires devising techniques for constructing cones \(\mathbf{H}\) along with efficiently computable functions \(\mathcal{M}(\cdot)\) satisfying (5.23). These tasks are the subjects of the sections to follow.
5.1.5.2 Specifying cones $H$

We specify cones $H$ in the case when the number $N$ of columns in the candidate contrast matrices is $m$ and under the following assumption on the given reliability tolerance $\epsilon$ and observation scheme in question:

\textbf{D}. There is a computationally tractable convex compact subset $Z \subset \mathbb{R}_+^m$ intersecting $\text{int} \mathbb{R}_+^m$ such that the norm $\pi(\cdot)$

$$
\pi(h) = \sqrt{\max_{z \in Z} \sum_i z_i h_i^2}
$$

induced by $Z$ satisfies the relation

$$
\pi(h) \leq 1 \Rightarrow \text{Prob}\{|h^T \xi x| > 1\} \leq \frac{\epsilon}{m} \forall x \in X.
$$

Note that condition D is satisfied for sub-Gaussian, Discrete, and Poisson observation schemes: according to the results of Section 5.1.3,

- in the sub-Gaussian case, it suffices to take $Z = \{2\sigma^2 \ln(2m/\epsilon)[1; \ldots; 1]\};$
- in the Discrete case, it suffices to take $Z = 4 \ln(2m/\epsilon) A \mathcal{X} + 64 \ln^2(2m/\epsilon) \Delta_m,$
  where $A\mathcal{X} = \{Ax : x \in \mathcal{X}\}$, $\Delta_m = \{y \in \mathbb{R}^m : y \geq 0, \sum_i y_i = 1\}$
- in the Poisson case, it suffices to take $Z = 2 \ln(2m/\epsilon) A \mathcal{X} + \frac{16}{9} \ln^2(2m/\epsilon) \Delta_m,$
  with $A\mathcal{X}$ and $\Delta_m$ as above.

Note that in all these cases $Z$ only “marginally” – logarithmically – depends on $\epsilon$ and $m$.

Under Assumption D, the cone $H$ can be built as follows:

- When $Z$ is a singleton: $Z = \{\bar{z}\}$, so that $\pi(\cdot)$ is scaled Euclidean norm, we set
  $$
  H = \left\{(\Theta, \mu) \in \mathbb{S}_+^m \times \mathbb{R}_+ : \mu \geq \sum_i \bar{z}_i \Theta_{ii}\right\}.
  $$

Given $(\Theta, \mu) \in H$, the $m \times m$ matrix $H$ and $\lambda \in \mathbb{R}_+^m$ are built as follows: setting $S = \text{Diag}\{\sqrt{\bar{z}_1}, \ldots, \sqrt{\bar{z}_m}\}$, we compute the eigenvalue decomposition of the matrix $S\Theta S$:

$$
S\Theta S = U\text{Diag}\{\lambda\}U^T,
$$

where $U$ is orthonormal, and set $H = S^{-1}U$, thus ensuring $\Theta = H\text{Diag}\{\lambda\}H^T$. Since $\mu \geq \sum_i \bar{z}_i \Theta_{ii}$, we have $\sum_i \lambda_i = \text{Tr}(S\Theta S) \leq \mu$. Finally, a column $h$ of $H$ is
of the form $S^{-1}f$ with $\| \cdot \|_2$-unit vector $f$, implying that

$$
\pi(h) = \sqrt{\sum_i \tilde{z}_i [S^{-1}f]_i^2} = \sqrt{\sum_i f_i^2} = 1,
$$

so that $h$ satisfies (5.1) by $D$.

- When $Z$ is not a singleton, we set

$$
\phi(r) = \max_{z \in Z} z^T r, \quad \kappa = 6 \ln(2 \sqrt{3} m^2), \quad H = \{(\Theta, \mu) \in S_{m}^{+} \times R_{m}^{+} : \mu \geq \kappa \phi(dg(\Theta))\},
$$

where $dg(Q)$ is the diagonal of a (square) matrix $Q$. Note that $\phi(r) > 0$ whenever $r \geq 0, r \neq 0$, since $Z$ contains a positive vector.

The justification of this construction and the efficient (randomized) algorithm for converting a pair $(\Theta, \mu) \in H$ into $(H, \lambda)$ satisfying, when taken along with $(\Theta, \mu)$, the requirements of $C$ are given by the following

**Lemma 5.6.** Let norm $\pi(\cdot)$ satisfy $D$.

(i) Whenever $H$ is an $m \times m$ matrix with columns $h_\ell$ satisfying $\pi(h_\ell) \leq 1$ and $\lambda \in R_{m}^{+}$, we have

$$
(\Theta_{\lambda} = H \text{Diag}\{\lambda\} H^T, \mu = \kappa \sum_i \lambda_i) \in H.
$$

(ii) Given $(\Theta, \mu) \in H$ with $\Theta \neq 0$, we find decomposition $\Theta = QQ^T$ with $m \times m$ matrix $Q$, fix an orthonormal $m \times m$ matrix $V$ with magnitudes of entries not exceeding $\sqrt{2/m}$ (e.g., the orthonormal scaling of the matrix of the cosine transform). When $\mu > 0$, we set $\lambda = \frac{\mu}{m}[1; \ldots; 1] \in R_{m}^{+}$ and consider the random matrix

$$
H_\chi = \sqrt{\frac{m}{\mu}} Q \text{Diag}\{\chi\} V,
$$

where $\chi$ is the $m$-dimensional Rademacher random vector. We have

$$
H_\chi \text{Diag}\{\lambda\} H_\chi^T = \Theta, \lambda \geq 0, \sum_i \lambda_i = \mu. \quad (5.30)
$$

Moreover, the probability of the event

$$
\pi(\text{Col}[H_\chi]) \leq 1 \forall \ell \leq m \quad (5.31)
$$

is at least $1/2$. Thus, generating independent samples of $\chi$ and terminating with $H = H_\chi$ when the latter matrix satisfies (5.31), we with probability 1 terminate with $(H, \lambda)$ satisfying $C$, and the probability for the outlined procedure to terminate in course of the first $M = 1, 2, \ldots$ steps is at least $1 - 2^{-M}$.

When $\mu = 0$, we have $\Theta = 0$ (since $\mu = 0$ implies $\phi(dg(\Theta)) = 0$, which with $\Theta \succeq 0$ is possible only when $\Theta = 0$); thus, when $\mu = 0$, we set $H = 0_{m \times m}$ and $\lambda = 0_{m \times 1}$.

Note that the lemma states, essentially, that the cone $H$ is a tight, up to loga-
Cones compatible with convex sets. Given a nonempty convex compact set \( \mathcal{Y} \subseteq \mathbb{R}^N \), we say that a cone \( \mathbf{Y} \) is compatible with \( \mathcal{Y} \), if

- \( \mathbf{Y} \) is a closed convex computationally tractable cone contained in \( \mathbf{S}^N_+ \times \mathbb{R}_+ \).
- one has \( \forall (V, \tau) \in \mathbf{Y} : \max_{y \in \mathcal{Y}} y^T V y \leq \tau \) \hspace{1cm} (5.32)
- \( \mathbf{Y} \) contains a pair \((V, \tau)\) with \( V \succ 0 \).
- relations \((V, \tau) \in \mathbf{Y}\) and \((V', \tau') \geq \tau\) imply that \((V, \tau') \in \mathbf{Y} \).

We call a cone \( \mathbf{Y} \) sharp, if \( \mathbf{Y} \) is a closed convex cone contained in \( \mathbf{S}^N_+ \times \mathbb{R}_+ \) and such that the only pair \((V, \tau) \in \mathbf{Y}\) with \( \tau = 0 \) is the pair \((0, 0)\), or, equivalently, a sequence \( \{(V_i, \tau_i) \in \mathbf{Y}, i \geq 1\} \) is bounded if and only if the sequence \( \{\tau_i, i \geq 1\} \) is bounded.

Note that whenever the linear span of \( \mathcal{Y} \) is the entire \( \mathbb{R}^N \), every compatible with \( \mathcal{Y} \) cone is sharp.

Observe that if \( \mathcal{Y} \subseteq \mathbb{R}^N \) is a nonempty convex compact set and \( \mathbf{Y} \) is a cone compatible with a shift \( \mathcal{Y} - a \) of \( \mathcal{Y} \), then \( \mathbf{Y} \) is compatible with \( \mathcal{Y}_a \).

Indeed, when shifting a set \( \mathcal{Y} \), its symmeterization \( \frac{1}{2}[\mathcal{Y} - \mathcal{Y}] \) remains intact, so that we can assume that \( \mathbf{Y} \) is compatible with \( \mathcal{Y} \). Now let \((V, \tau) \in \mathbf{Y}\) and \( y, y' \in \mathcal{Y} \). We have

\[ [y - y']^T V [y - y'] + [y + y']^T V [y + y'] = 2[y^T V y + |y']^T V y] \leq 4\tau, \]

whence for \( z = \frac{1}{2}[y - y'] \) it holds \( z^T V z \leq \tau \). Since every \( z \in \mathcal{Y}_a \) is of the form \( \frac{1}{2}[y - y'] \) with \( y, y' \in \mathcal{Y} \), the claim follows.

Note that the claim can be “nearly inverted” if \( 0 \in \mathcal{Y} \) and \( \mathbf{Y} \) is compatible with \( \mathcal{Y}_a \), then the “widening” of \( \mathbf{Y} \) – the cone \( \mathbf{Y}^+ = \{(V, \tau) : (V, \tau/4) \in \mathbf{Y}\} \) is compatible with \( \mathcal{Y} \) (evident, since when \( 0 \in \mathcal{Y} \), every vector from \( \mathcal{Y} \) is proportional, with coefficient 2, to a vector from \( \mathcal{Y}_a \)).
Constructing functions $\mathcal{M}$. The role of compatibility in our context becomes clear from the following observation:

**Proposition 5.7.** In the situation described in Section 5.1.1, assume that we have at our disposal cones $X$ and $U$ compatible, respectively, with $X_*$ and with the unit ball

$$B_* = \{ v \in \mathbb{R}^\nu : \| u \|_* \leq 1 \}$$

of the norm $\| \cdot \|_*$ conjugate to the norm $\| \cdot \|$. Given $M \in S^{\nu+n}$, let us set

$$\mathcal{M}(M) = \inf_{X,t,U,s} \{ t + s : (X,t) \in X, (U,s) \in U, \text{Diag}\{U,X\} \succeq M \} \quad (5.33)$$

Then $\mathcal{M}$ is real-valued efficiently computable convex function on $S^{\nu+n}$ such that (5.23) takes place: for every $M \in S^{\nu+n}$ it holds

$$\mathcal{M}(M) \geq \max_{[u;z] \in B_* \times X_*} [u;z]^T M[u;z].$$

In addition, when $X$ and $U$ are sharp, the infimum in (5.33) is achieved.

**Proof** is immediate. Given that the objective of the optimization problem specifying $\mathcal{M}(M)$ is nonnegative on the feasible set, the fact that $\mathcal{M}$ is real-valued is equivalent to problem’s feasibility, and the latter is readily given by the fact that $X$ is a cone containing a pair $(X,t)$ with $X \succ 0$ and similarly for $U$. Convexity of $\mathcal{M}$ is evident. To verify (5.23), let $(X,t,U,s)$ form a feasible solution to the optimization problem in (5.33). When $[u;z] \in B_*$ we have

$$[u;z]^T M[u;z] \leq u^T U u + z^T X z \leq s + t,$$

where the first inequality is due to the $\succeq$-constraint in (5.33), and the second is due to the fact that $U$ is compatible with $B_*$, and $X$ is compatible with $X_*$. Since the resulting inequality holds true for all feasible solutions to the optimization problem in (5.33), (5.23) follows. Finally, when $X$ and $U$ are sharp, (5.33) is a feasible conic problem with bounded level sets of the objective and as such is solvable. \qed

### 5.1.5.4 Putting things together

The following statement combining the results Propositions 5.7 and 5.5 summarizes our second approach to the design of polyhedral estimate.

**Proposition 5.8.** In the situation of Section 5.1.1, assume that we have at our disposal cones $X$ and $U$ compatible, respectively, with $X_*$ and with the unit ball $B_*$ of the norm conjugate to $\| \cdot \|$. Given reliability tolerance $\epsilon \in (0,1)$ along with a positive integer $N$ and a computationally tractable cone $H$ satisfying Assumption $C$, consider (clearly feasible) convex optimization problem

$$\text{Opt} = \min_{\Theta,\mu,X,t,U,s} \left\{ f(t,s,\mu) := 2(t + s + \mu) : \begin{array}{c} (\Theta,\mu) \in H, (X,t) \in X, (U,s) \in U \\ \begin{bmatrix} U & \frac{1}{2}B \\ \frac{1}{2}B^T & A^T \Theta A + X \end{bmatrix} \succeq 0 \end{array} \right\} \quad (5.34)$$

Let $\Theta,\mu,X,t,U,s$ be a feasible solution to (5.34). Invoking $C$, we can convert, in
a computationally efficient manner, \((\Theta, \mu)\) into \((H, \lambda)\) such that the columns of the 
m \times N contrast matrix \(H\) satisfy (5.1), \(\Theta = H \text{Diag}\{\lambda\} H^T\), and \(\mu \geq \sum_i \lambda_i\). The 
\((\epsilon, \|\cdot\|)\)-risk of the polyhedral estimate \(\hat{x}^H\) satisfies the bound

\[
\text{Risk}_{\epsilon, \|\cdot\|}[\hat{x}^H|\mathcal{X}] \leq f(t, s, \mu). 
\]

(5.35)

In particular, we can build, in a computationally efficient manner, polyhedral estimates with risks arbitrarily close to Opt (and with risk Opt, provided that (5.34) is solvable).

**Proof.** Let \(\Theta, \mu, X, t, U, s\) form a feasible solution to (5.34). By the semidefinite constraint in (5.34) we have

\[
0 \preceq \begin{bmatrix}
U & \frac{1}{2}B \\
-\frac{1}{2}B^T & A^T \Theta A + X
\end{bmatrix} = \text{Diag}\{U, X\} - \begin{bmatrix}
-\frac{1}{2}B^T & \frac{1}{2}B \\
\frac{1}{2}B & -A^T \Theta A
\end{bmatrix} =: M,
\]

whence for the function \(M\) defined in (5.33) one has

\[
M(M) \leq t + s.
\]

Since \(M\), by Proposition 5.7, satisfies (5.23), invoking Proposition 5.5 we arrive at

\[
\mathcal{R}[H] \leq 2(\mu + M(M)) \leq f(t, s, \mu).
\]

By Proposition 5.1 this implies the target relation (5.35). \(\square\)

5.1.5.5 Compatibility: basic examples and calculus

Our approach to the design of polyhedral estimates utilizing the recipe described in Proposition 5.8 relies upon our ability to equip convex “sets of interest” (in our context, these are the symmeterization \(X_s\) of the signal set and the unit ball \(B_*\) of the norm conjugate to the norm \(\|\cdot\|\)) with compatible cones.\(^5\) Below, we discuss two principal sources of such cones, namely (a) spectratopes/ellitopes, and (b) absolute norms. More examples of compatible cones can be constructed using a “compatibility calculus.” Namely, let us assume that we are given a finite collection of convex sets (operands) and apply to them some basic operation, such as taking the intersection, or arithmetic sum, direct or inverse linear image, or convex hull of the union. It turns out that cones compatible with the results of such operations can be easily (in a fully algorithmic fashion) obtained from the cones compatible with the operands; see Section 5.1.8 for principal calculus rules.

In view of Proposition 5.8, the larger are the cones \(X\) and \(U\) compatible with \(X_s\) and \(B_*\), the better – the wider is the optimization domain in (5.34) and, consequently, the less is (the best) risk bound achievable with the recipe presented in the proposition. Given convex compact set \(\mathcal{Y} \in \mathbb{R}^N\), the “ideal” – the largest – candidate to the role of the cone compatible with \(\mathcal{Y}\) would be

\[
\mathcal{Y}^* = \{(V, \tau) \in S_+^N \times \mathbb{R}_+: \tau \geq \max_{y \in \mathcal{Y}} y^T V y\}.
\]

\(^5\)Recall that we already know how to specify the second element of the construction, the cone \(H\).
However, this cone is typically intractable, therefore, we look for "as large as possible" tractable inner approximations of $Y^*$. 

5.1.5.5.A. Cones compatible with ellitopes/spectratopes are readily given by semidefinite relaxation. Specifically, when

$$\mathcal{Y} = \{ y \in \mathbb{R}^N : \exists (r \in \mathcal{R}, z \in \mathbb{R}^K) : y = Mz, R_{\ell}^2[z] \leq r_{\ell}I_{d_\ell}, \ell \leq L \}$$

with our standard restrictions on $\mathcal{R}$, invoking Proposition 4.8 it is immediately seen that the set

$$\mathcal{Y} = \{(V, \tau) \in \mathbb{S}_+^{\ell} \times \mathbb{R}_+ : \exists \Lambda = \{ \Lambda_{\ell} \in \mathbb{S}_+^{d_\ell}, \ell \leq L \} : \phi_\mathcal{R}(\lambda[\Lambda]) \leq \tau, M^TV \mathcal{M} \leq \sum_{\ell} \mathcal{R}^*[\Lambda_{\ell}], \}$$

(5.36)

is a closed convex cone which is compatible with $\mathcal{Y}$; here, as usual,

$$[\mathcal{R}[\Lambda_{\ell}]]_{ij} = \text{Tr}(R_{\ell}^i \Lambda_{\ell} R_{\ell}^j), \lambda[\Lambda] = [\text{Tr}(\Lambda_1) ; \ldots ; \text{Tr}(\Lambda_L)], \phi_\mathcal{R}(\lambda) = \max_{r \in \mathcal{R}} r^T \lambda.$$

Similarly, when $\mathcal{Y}$ is an ellitope:

$$\mathcal{Y} = \{ y \in \mathbb{R}^N : \exists (r \in \mathcal{R}, z \in \mathbb{R}^K) : y = Mz, z^T R_{\ell} z \leq r_{\ell}, \ell \leq L \}$$

with our standard restrictions on $R_{\ell}$, invoking Proposition 4.6, the set

$$\mathcal{Y} = \{(V, \tau) \in \mathbb{S}_+^{\ell} \times \mathbb{R}_+ : \exists \lambda \in \mathbb{R}_+^\ell : M^T \mathcal{V} \mathcal{M} \leq \sum_{\ell} \lambda_{\ell} R_{\ell}, \phi_\mathcal{R}(\lambda) \leq \tau \}$$

(5.37)

is a closed convex cone which compatible with $\mathcal{Y}$. In both cases, $\mathcal{Y}$ is sharp, provided that the image space of $M$ is the entire $\mathbb{R}^N$. 

Note that in both these cases $\mathcal{Y}$ is a reasonably tight inner approximation of $Y^*$: whenever $(V, \tau) \in Y^*$, we have $(V, \theta \tau) \in \mathcal{Y}$, with a moderate $\theta$ (specifically, $\theta = O(1) \ln (2 \sum_{\ell} d_\ell)$ in the spectratopic, and $\theta = O(1) \ln (2L)$ in the ellitopic case, see Propositions 4.8, 4.6, respectively).

5.1.5.5.B. Compatibility via absolute norms.

Preliminaries. Recall that a norm $p(\cdot)$ on $\mathbb{R}^N$ is called absolute, if $p(x)$ is a function of the vector $\text{abs}[x] := [\text{abs}[x_1] ; \ldots ; \text{abs}[x_N]]$ of the magnitudes of entries in $x$. It is well known that an absolute norm $p$ is monotone on $\mathbb{R}_+^N$, so that $\text{abs}[x] \leq \text{abs}[x']$ implies that $p(x) \leq p(x')$, and that the norm

$$p_*(x) = \max_{y : p(y) \leq 1} x^Ty$$

conjugate to $p(\cdot)$ is absolute along with $p$.

Let us say that an absolute norm $r(\cdot)$ fits an absolute norm $p(\cdot)$ on $\mathbb{R}^N$, if for every vector $x$ with $p(x) \leq 1$ the entrywise square $[x]^2 = [x_1^2 ; \ldots ; x_N^2]$ of $x$ satisfies $r([x]^2) \leq 1$. For example, the largest norm $r(\cdot)$ which fits the absolute norm $p(\cdot) = \| \cdot \|_s$, $s \in [1, \infty]$, is

$$r(\cdot) = \begin{cases} \| \cdot \|_1, & 1 \leq s \leq 2 \\ \| \cdot \|_{s/2}, & s \geq 2 \end{cases}$$
An immediate observation is that an absolute norm \( p(\cdot) \) on \( \mathbb{R}^N \) can be "lifted" to a norm on \( S^N \), specifically, the norm
\[
p^+(Y) = p([p(\text{Col}_1[Y]); \ldots; p(\text{Col}_N[Y])]) : S^N \rightarrow \mathbb{R}_+,
\]
where \( \text{Col}_j[Y] \) is \( j \)th column in \( Y \). It is immediately seen that when \( p \) is an absolute norm, the right hand side in (5.38) indeed is a norm on \( S^N \) satisfying the identity
\[
p^+(xx^T) = p^2(x), \ x \in \mathbb{R}^N.
\]

**Absolute norms and compatibility.** Our interest in absolute norms is motivated by the following immediate

**Observation 5.9.** Let \( p(\cdot) \) be an absolute norm on \( \mathbb{R}^N \), and \( r(\cdot) \) be another absolute norm which fits \( p(\cdot) \), both norms being computationally tractable. These norms give rise to the computationally tractable and sharp closed convex cone

\[
P = P_{p(\cdot), r(\cdot)} = \left\{ (V, \tau) \in S^N_+ \times \mathbb{R}_+ : \exists (W \in S^N, w \in \mathbb{R}^N) : \right. \\
V \preceq W + \text{Diag}\{w\} \\
\left\{ p^+(W) + r_* (w) \leq \tau \right\}
\]

where \( [p^+]_* \) is the norm on \( S^N \) conjugate to the norm \( p^+(\cdot) \), and \( r_* (\cdot) \) is the norm on \( \mathbb{R}^N \) conjugate to the norm \( r(\cdot) \), and this cone is compatible with the unit ball of the norm \( p(\cdot) \) (and thus – with any convex compact subset of this ball).

Verification is immediate. The fact that \( P \) is a computationally tractable and closed convex cone is evident. Now let \( (V, \tau) \in P \), so that \( V \succeq 0 \) and \( V \preceq W + \text{Diag}\{w\} \) with \( [p^+]_* (W) + r_* (w) \leq \tau \). For \( x \) with \( p(x) \leq 1 \) we have
\[
x^T V x ~ \leq ~ x^T [W + \text{Diag}\{w\}] x = \text{Tr}(W [xx^T]) + w^T [x]^2
\]
\[
\leq ~ p^+(xx^T) [p^+]_*(W) + r([x]^2) r_* (w) = p^2(x) [p^+]_* (W) + r_* (w) \\
\leq ~ [p^+]_* (W) + r_* (w) \leq \tau
\]

(we have used (5.40)), whence \( x^T V x \leq \tau \) for all \( x \) with \( p(x) \leq 1 \).

Let us look at the proposed construction in the case where \( p(\cdot) = \|\cdot\|_s, \ s \in [1, \infty] \), and let \( r(\cdot) = \|\cdot\|_{\bar{s}} \), \( \bar{s} = \max[s/2, 1] \). Setting \( s_* = \frac{s}{s-1}, \ \bar{s}_* = \frac{\bar{s}}{\bar{s}-1} \), we clearly have
\[
[p^+]_* (W) = \|W\|_{s_*} := \left\{ \left( \sum_{i,j} |W_{ij}|^{s_*} \right)^{1/s_*} : \right. \\
\left. \max_{i,j} |W_{ij}|, \ s_* < \infty \right\}, \ r_* (w) = \|w\|_{\bar{s}_*},
\]
resulting in
\[
P^* := P_{\|\cdot\|_s, \|\cdot\|_{\bar{s}}} = \left\{ (V, \tau) : V \in S^N_+, \exists (W \in S^N, w \in \mathbb{R}^N) : \\
V \preceq W + \text{Diag}\{w\} \\
\|W\|_{s_*} + \|w\|_{\bar{s}_*} \leq \tau \right\},
\]

By Observation 5.9, \( P^* \) is compatible with the unit ball of \( \|\cdot\|_s \)-norm on \( \mathbb{R}^N \) (and therefore with every closed convex subset of this ball).
When $s = 1$, that is, $s_* = \bar{s}_* = \infty$, (5.42) results in

$$P^1 = \left\{ (V, \tau) : V \succeq 0, \exists (W \in S^N, w \in \mathbb{R}^N_+) : V \preceq W + \text{Diag}\{w\}, \|W\|_\infty + \|w\|_\infty \leq \tau \right\} = \{ (V, \tau) : V \succeq 0, \|V\|_\infty \leq \tau \},$$

(5.43)

and it is easily seen that the situation is as good as it could be, namely,

$$P^1 = \left\{ (V, \tau) : V \succeq 0, \max \|x\|_1 \leq 1 \right\} = \left\{ (V, \tau) : V \succeq 0, \|x^T V x\| \leq \tau \right\}.$$

It can be shown (see Section 5.4.3) that when $s \in [2, \infty]$, and so $\bar{s}_* = \frac{2}{s-2}$, (5.42) results in

$$P^s = \left\{ (V, \tau) : V \succeq 0, \exists (w \in \mathbb{R}^N_+) : V \preceq \text{Diag}\{w\} \& \|w\|_{\frac{s}{s-2}} \leq \tau \right\}.$$

(5.44)

Note that

$$P^2 = \left\{ (V, \tau) : V \succeq 0, \|V\|_2 \leq \tau \right\}$$

and this is exactly the largest cone compatible with the unit Euclidean ball.

When $s \geq 2$, the unit ball $Y$ of the norm $\|\cdot\|_s$ is an ellitope:

$$\{ y \in \mathbb{R}^N : \|y\|_s \leq 1 \} = \{ y \in \mathbb{R}^N : \exists (t \geq 0, \|t\|_s \leq 1) : y^T R t y := y^2 \leq t \ell, \ell \leq L = N \},$$

so that one of the cones compatible with $Y$ is given by (5.37) with the identity matrix in the role of $M$. It comes as no surprise that, as it is immediately seen, the latter cone is nothing but the cone (5.44).

5.1.5.6 Near-optimality of polyhedral estimate in the spectratopic sub-Gaussian case

As an instructive application of the approach developed so far, consider the special case of the estimation problem stated in Section 5.1.1, where

1. The signal set $\mathcal{X}$ and the unit ball $\mathcal{B}_*$ of the norm conjugate to $\|\cdot\|$ are spectratopes:

$$\mathcal{X} = \{ x \in \mathbb{R}^N : \exists t \in T : R_t^* [x] \preceq t_k I_{d_k}, 1 \leq k \leq K \},$$

$$\mathcal{B}_* = \{ z \in \mathbb{R}^N : \exists y \in Y : z = M y \},$$

$$Y := \{ y \in \mathbb{R}^N : \exists \tau \in \mathbb{R} : S_\ell^* [y] \preceq \tau I_{L}, 1 \leq \ell \leq L \},$$

(cf. Assumptions A, B in Section 4.3.3.2; as always, we lose nothing assuming spectrotape $\mathcal{X}$ to be basic).

2. For every $x \in \mathcal{X}$, observation noise $\xi_x$ is sub-Gaussian, i.e., $\xi_x \sim \mathcal{G}(0, \sigma^2 I_m)$.

We are about to show that in the present situation, the polyhedral estimate constructed in Sections 5.1.5.2–5.1.5.4, i.e., yielded by the efficiently computable (high accuracy near-) optimal solution to the optimization problem (5.34) is near-optimal in the minimax sense.

Given reliability tolerance $\epsilon \in (0, 1)$, the recipe for constructing $m \times m$ contrast matrix $H$ as presented in Proposition 5.8 is as follows:

- Set

$$Z = \{ \vartheta^2 [1; \ldots; 1] \}, \vartheta = \sigma \kappa, \kappa = \sqrt{2 \ln(2m/\epsilon)},$$

where

$$\vartheta^2 = \sigma \kappa, \kappa = \sqrt{2 \ln(2m/\epsilon)},$$

and
and utilize the construction from Section 5.1.5.2, thus arriving at the cone

$$\mathbf{H} = \{ (\Theta, \mu) \in S^n_k \times \mathbb{R}_+ : \sigma^2 \kappa^2 \text{Tr}(\Theta) \leq \mu \}$$

satisfying the requirements of Assumption C.

• Specify the cones \( \mathbf{X} \) and \( \mathbf{U} \) compatible with \( \mathcal{X}_s = \mathcal{X} \), and \( \mathcal{B}_s \), respectively, according to (5.36).

The resulting problem (5.34), after immediate straightforward simplifications, reads

$$\text{Opt} = \min_{\Theta, U, \Lambda, \Upsilon} \left\{ \frac{1}{2} \left[ \phi_R(\lambda[\Upsilon]) + \phi_T(\lambda[\Lambda]) + \sigma^2 \kappa^2 \text{Tr}(\Theta) \right] : \begin{array}{c} U \geq 0, \ U \geq 0, \ \Lambda = \{ \Lambda_k \geq 0, k \leq K \}, \\ \Upsilon = \{ \Upsilon_\ell \geq 0, \ \ell \leq L \}, \\ \Lambda \mathbf{M}^T \Upsilon \leq \sum_\ell S_\ell^*[\Upsilon_\ell], \end{array} \right\}$$

where, as always,

$$[\mathbf{R}_k]\{\Lambda_k\}_{ij} = \text{Tr}(\mathbf{R}_k \Lambda_k \mathbf{R}^T_k), \quad [\mathbf{R}_k]\{x\} = \sum_i x_i \mathbf{R}_k i,$$

$$[\mathbf{S}_\ell]\{\Upsilon_\ell\}_{ij} = \text{Tr}(\mathbf{S}_\ell \Upsilon_\ell \mathbf{S}^T_\ell), \quad [\mathbf{S}_\ell]\{u\} = \sum_\ell u_\ell \mathbf{S}_\ell i.$$
with the cone
\[ U = P^2 = \{(U, \tau) : U \geq 0, \|U\|_{2,2} \leq \tau\} \]
and \( X \) — with the cone
\[ X = P^1 = \{(X, t) : X \geq 0, \|X\|_\infty \leq t\}, \]
(note that both cones are the largest w.r.t. inclusion cones compatible with the respective sets). The corresponding problem (5.34) reads
\[
\text{Opt} = \min_{\Theta, X, U} \left\{ 2 \left( \kappa^2 \sigma^2 \text{Tr}(\Theta) + \max_i X_{ii} + \|U\|_{2,2} \right) : \begin{bmatrix} \Theta & U \\ \frac{1}{2} I_n & \Theta + X \end{bmatrix} \succeq 0 \right\}
\]
\[
\text{Opt} = \min_{\Theta, X} \left\{ 2 \left( \kappa^2 \sigma^2 \text{Tr}(\Theta) + \max_i X_{ii} + \tau \right) : \begin{bmatrix} \Theta & \tau I_n \\ \frac{1}{2} I_n & \Theta + X \end{bmatrix} \succeq 0 \right\}
\] (5.46)

Observe that every \( n \times n \) matrix of the form \( Q = EP \), where \( E \) is diagonal with diagonal entries \( \pm 1 \), and \( P \) is a permutation matrix, induces a symmetry \((\Theta, X, \tau) \rightarrow (Q\Theta Q^T, QXQ^T, \tau)\) of the second optimization problem in (5.46), that is, a transformation which maps the feasible set onto itself and keeps the objective intact. Since the problem is convex and solvable, we conclude that it has an optimal solution which remains intact under the symmetries in question, i.e., solution with scalar matrices \( \Theta = \theta I_n \) and \( X = u I_n \). As a result,
\[
\text{Opt} = \min_{\theta \geq 0, u \geq 0, \tau} \left\{ 2 \left( \kappa^2 \sigma^2 \kappa \theta + u + \tau \right) : \tau (\theta + u) \geq \frac{1}{\tau} \right\} = 2 \min \left[ \kappa^2 \sigma \sqrt{n} \right] \] (5.47)

A similar derivation shows that the value \( \text{Opt} \) remains intact if we replace the set \( \mathcal{X} = \{ x : \|x\|_1 \leq 1 \} \) with \( \mathcal{X} = \{ x : \|x\|_s \leq 1 \}, \ s \in [1, 2] \), and the cone \( X = P^1 \) with \( X = P^s \), see (5.42). Since the \( \Theta \)-component of an optimal solution to (5.46) can be selected to be scalar, the contrast matrix \( H \) we end up with can be selected to be the unit matrix. An unpleasant observation is that when \( s < 2 \), the quantity \( \text{Opt} \) given by (5.47) “heavily overestimates” the actual risk of the polyhedral estimate with \( H = I_n \). Indeed, the analysis of this estimate in Section 5.1.4 results in the risk bound (up to a logarithmic in \( n \) factor) \( \min[\sigma^{1-s/2}, \sigma \sqrt{n}] \), what can be much less than \( \text{Opt} = 2 \min \left[ \kappa \sigma \sqrt{n} \right] \), e.g., in the case of large \( n \), and \( \sigma \sqrt{n} = O(1) \).

5.1.6 Assembling estimates: contrast aggregation

A good news is that whenever the approaches to the design of polyhedral estimates presented in Sections 5.1.4 and 5.1.5 are applicable, they can be utilized simultaneously. The underlying observation is that

(!) In the problem setting described in Section 5.1.2, a collection of \( K \) candidate polyhedral estimates can be assembled into a single polyhedral estimate with the (upper bound on the) risk, as given by Proposition 5.1, being nearly the minimum of the risks of estimates we aggregate.

Indeed, given an observation scheme (that is, collection of probability distributions \( P_x \) of noises \( \xi_x, x \in \mathcal{X} \)), assume we have at our disposal norms \( \pi_\delta(h) : \mathbb{R}^m \rightarrow \mathbb{R} \) parameterized by \( \delta \in (0, 1) \) such that \( \pi_\delta(h) \), for every \( h \), is the larger the less is \( \delta \),
and

\[ \forall (x \in \mathcal{X}, \delta \in (0,1), h \in \mathbb{R}^m) : \pi_\delta(h) \leq 1 \implies \text{Prob}_{\xi \sim P_x} \{ \xi : |h^T \xi| > 1 \} \leq \delta. \]

Assume also (as is indeed the case in all our constructions) that we ensure (5.1) by imposing on the columns \( h_k \) of an \( m \times N \) contrast matrix \( H \) the restrictions \( \pi_{\epsilon/N}(h_k) \leq 1 \).

Now suppose that given risk tolerance \( \epsilon \in (0,1) \), we have generated somehow \( K \) candidate contrast matrices \( H_k \in \mathbb{R}^{m \times N_k} \) such that

\[ \pi_{\epsilon/N_k}(\text{Col}_j[H_k]) \leq 1, j \leq N_k, \]

so that the \((\epsilon, \| \cdot \|)-\text{risk}\) of the polyhedral estimate yielded by the contrast matrix \( H_k \) does not exceed

\[ R_k = \max_x \{ \|Bx\| : x \in 2\mathcal{X}, \|H_k^T Ax\|_\infty \leq 2 \}. \]

Let us combine the contrast matrices \( H_1, \ldots, H_K \) into a single contrast matrix \( H \) with \( N = N_1 + \ldots + N_K \) columns by normalizing the columns of the concatenated matrix \( [H_1, \ldots, H_K] \) to have \( \pi_{\epsilon/N}(\cdot) \) norms equal to 1, so that

\[ H = [H_1, \ldots, H_K], \quad \text{Col}_j[H_k] = \theta_{jk} \text{Col}_j[H_k] \quad \forall (k \leq K, j \leq N_k) \]

with

\[ \theta_{jk} = \frac{1}{\pi_{\epsilon/N}(\text{Col}_j[H_k])} \geq \vartheta_k := \min_{h \neq 0} \frac{\pi_{\epsilon/N_k}(h)}{\pi_{\epsilon/N}(h)}, \]

where the concluding \( \geq \) is due to \( \pi_{\epsilon/N_k}(\text{Col}_j[H_k]) \leq 1 \). We claim that in terms of \((\epsilon, \| \cdot \|)-\text{risk}\), the polyhedral estimate yielded by \( H \) is “almost as good” as the best of the polyhedral estimates yielded by the contrast matrices \( H_1, \ldots, H_K \), specifically,\(^6\)

\[ R[H] := \max_x \{ \|Bx\| : x \in 2\mathcal{X}, \|H^T Ax\|_\infty \leq 2 \} \leq \min_k \vartheta_k^{-1} R_k. \]

The justification is readily given by the following observation: when \( \vartheta \in (0,1) \), we have

\[ R_{k,\vartheta} := \max_x \{ \|Bx\| : x \in 2\mathcal{X}, \|H_{k}^T Ax\|_\infty \leq 2/\vartheta \} \leq R_k/\vartheta. \]

Indeed, when \( x \) is a feasible solution to the maximization problem specifying \( R_{k,\vartheta} \), \( \vartheta x \) is a feasible solution to the problem specifying \( R_k \), implying that \( \vartheta \|Bx\| \leq R_k \).

It remains to note that we clearly have \( R[H] \leq \min_k R_{k,\vartheta} \).

The bottom line is that the just described aggregation of contrast matrices \( H_1, \ldots, H_K \) into a single contrast matrix \( H \) results in a polyhedral estimate which in terms of upper bound \( R[\cdot] \) on its \((\epsilon, \| \cdot \|)-\text{risk}\) is, up to factor \( \vartheta = \max_k \vartheta_k^{-1} \), not worse than the best of the \( K \) estimates yielded by the original contrast matrices. Consequently, if \( \pi_\delta(\cdot) \) grows slowly as \( \delta \) decreases, the “price” \( \vartheta \) of assembling the original estimates is quite moderate. For example, in our basic cases (sub-Gaussian, Discrete, and Poisson), \( \vartheta \) is logarithmic in \( \max_k N_k^{-1}(N_1 + \ldots + N_K) \), and \( \vartheta = 1 + o(1) \) as \( \epsilon \to +0 \) for \( K, N_1, \ldots, N_K \) fixed.

\(^6\)This is the precise “quantitative expression” of the observation (!).
5.1.7 Numerical illustration

We are about to illustrate the numerical performance of polyhedral estimates by comparing it to the performance of a “presumably good” linear estimate. Our setup is deliberately simple: the signal set \( \mathcal{X} \) is just the unit box \( \{ x \in \mathbb{R}^n : \| x \|_\infty \leq 1 \} \), \( B \in \mathbb{R}^{n \times n} \) is “numerical double integration:” for a \( \delta > 0 \),
\[
B_{ij} = \begin{cases} 
\delta^2(i - j + 1), & j \leq i \\
0, & j > i , 
\end{cases}
\]
so that, modulo boundary effects, is the second order finite difference derivative of \( w = Bx \):
\[
x_i = \frac{w_i - 2w_{i-1} + w_{i-2}}{\delta^2}, \quad 2 < i \leq n,
\]
and \( Ax \) is comprised of \( m \) randomly selected entries of \( Bx \). The observation is
\[
\omega = Ax + \xi, \; \xi \sim \mathcal{N}(0, \sigma^2 I_m).
\]
and the recovery norm is \( \| \cdot \|_2 \). In other words, we want to recover a restriction of twice differentiable function of one variable on the \( n \)-point regular grid on the segment \( \Delta = [0, n\delta] \) from noisy observations of this restriction taken along \( m \) randomly selected points of the grid. A priori information on the function is that the magnitude of its second order derivative does not exceed 1.

Note that in the considered situation both linear estimate \( \hat{x}_H \) yielded by Proposition 4.14 and polyhedral estimate \( \hat{x}_H \) yielded by Proposition 5.7, are near-optimal in the minimax sense in terms of their \( \| \cdot \|_2 \)-, resp., \( (\epsilon, \| \cdot \|_2) \)-risk.

In the experiments reported in Figure 5.1, we used \( n = 64 \), \( m = 32 \), and \( \delta = 4/n \) (i.e., \( \Delta = [0, 4] \)); the reliability parameter for the polyhedral estimate was set to \( \epsilon = 0.1 \). For different noise levels \( \sigma = \{ 0.1, 0.01, 0.001, 0.0001 \} \) we generate 20 random signals \( x \) from \( \mathcal{X} \) and record the \( \| \cdot \|_2 \)-recovery errors of the linear and the polyhedral estimates. In addition to testing the nearly optimal polyhedral estimate \( PolyI \) yielded by Proposition 5.7, we also record the performance of the polyhedral estimate \( PolyII \) yielded by the construction from Section 5.1.4. The observed \( \| \cdot \|_2 \)-recovery errors of the three estimates are plotted in Figure 5.1.

All three estimates exhibit similar empirical performance in these simulations. However, when the noise level becomes small, polyhedral estimates seem to outperform the linear one. In addition, the estimate \( PolyII \) seems to “work” better than or, at the very worst, similarly to \( PolyI \) in spite of the fact that in the situation in question the estimate \( PolyI \), in contrast to \( PolyII \), is provably near-optimal.

5.1.8 Calculus of compatibility

The principal rules of the calculus of compatibility are as follows (verification of the rules is straightforward and is therefore skipped):

1. [passing to a subset] When \( \mathcal{Y}' \subset \mathcal{Y} \) are convex compact subsets of \( \mathbb{R}^N \) and a cone \( \mathcal{Y} \) is compatible with \( \mathcal{Y} \), the cone is compatible with \( \mathcal{Y}' \) as well.
2. [finite intersection] Let cones \( \mathcal{Y}_j \) be compatible with convex compact sets \( \mathcal{Y}_j \subset \mathbb{R}^N, \; j = 1, \ldots, J \). Then the cone
\[
\mathcal{Y} = cl\{(V, \tau) \in \mathbb{S}^N_+ \times \mathbb{R}_+ : \exists((V_j, \tau_j) \in \mathcal{Y}_j, j \leq J) : V \succeq \sum_j V_j, \sum_j \tau_j \leq \tau \}
\]
is compatible with $\mathcal{Y} = \bigcap_j \mathcal{Y}_j$. The closure operation can be skipped when all cones $\mathcal{Y}_j$ are sharp, in which case $\mathcal{Y}$ is sharp as well.

3. [convex hulls of finite union] Let cones $\mathcal{Y}_j$ be compatible with convex compact sets $\mathcal{Y}_j \subset \mathbb{R}^N$, $j = 1, \ldots, J$, and let there exist $(V, \tau)$ such that $V > 0$ and

$$(V, \tau) \in \mathcal{Y} := \bigcap_j \mathcal{Y}_j.$$

Then $\mathcal{Y}$ is compatible with $\mathcal{Y} = \text{Conv}\{\bigcup_j \mathcal{Y}_j\}$ and, in addition, is sharp provided that at least one of $\mathcal{Y}_j$ is sharp.

4. [direct product] Let cones $\mathcal{Y}_j$ be compatible with convex compact sets $\mathcal{Y}_j \subset \mathbb{R}^{N_j}$, $j = 1, \ldots, J$. Then the cone

$$\mathcal{Y} = \{(V, \tau) \in S_+^{N_1 + \cdots + N_J} \times \mathbb{R}_+: \exists (V_j, \tau_j) \in \mathcal{Y}_j : V \preceq \text{Diag}(V_1, \ldots, V_J) \& \tau \geq \sum_j \tau_j\}$$

is compatible with $\mathcal{Y} = \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_J$. This cone is sharp, provided that all $\mathcal{Y}_j$ are so.
5. [linear image] Let cone \( \mathbf{Y} \) be compatible with convex compact set \( \mathcal{Y} \subset \mathbb{R}^N \), let \( A \) be a \( K \times N \) matrix, and let \( \mathbf{Z} = A\mathbf{Y} \). The cone 
\[
\mathbf{Z} = \text{cl}\{ (V, \tau) \in S^K_+ \times \mathbb{R}_+ : \exists U \succeq A^TVA : (U, \tau) \in \mathbf{Y} \}
\]
is compatible with \( \mathcal{Z} \). The closure operation can be skipped whenever \( \mathbf{Y} \) is either sharp, or complete, completeness meaning that \( (V, \tau) \in \mathbf{Y} \) and \( 0 \preceq V' \preceq V \) imply that \( (V', \tau) \in \mathbf{Y} \). The cone \( \mathbf{Z} \) is sharp, provided \( \mathbf{Y} \) is so and the rank of \( A \) is \( K \).

6. [inverse linear image] Let cone \( \mathbf{Y} \) be compatible with convex compact set \( \mathcal{Y} \subset \mathbb{R}^N \), let \( A \) be a \( N \times K \) matrix with trivial kernel, and let \( \mathbf{Z} = A^{-1}\mathbf{Y} := \{ z \in \mathbb{R}^K : Az \in \mathcal{Y} \} \). The cone 
\[
\mathbf{Z} = \text{cl}\{ (V, \tau) \in S^K_+ \times \mathbb{R}_+ : \exists U : A^TUA \succeq V & (U, \tau) \in \mathbf{Y} \}
\]
is compatible with \( \mathcal{Z} \). The closure operations can be skipped whenever \( \mathbf{Y} \) is sharp, in which case \( \mathbf{Z} \) is sharp as well.

7. [arithmetic summation] Let cones \( \mathbf{Y}_j \) be compatible with convex compact sets \( \mathcal{Y}_j \subset \mathbb{R}^N \), \( j = 1, ..., J \). Then the arithmetic sum \( \mathbf{Y} = \mathcal{Y}_1 + ... + \mathcal{Y}_J \) of the sets \( \mathcal{Y}_j \) can be equipped with compatible cone readily given by the cones \( \mathbf{Y}_j \); this cone is sharp, provided all \( \mathbf{Y}_j \) are so.

Indeed, the arithmetic sum of \( \mathcal{Y}_j \) is the linear image of the direct product of \( \mathcal{Y}_j \)'s under the mapping \( [y^1; ...; y^J] \mapsto y^1 + ... + y^J \), and it remains to combine rules 4 and 5; note the cone yielded by rule 4 is complete, so that when applying rule 5, the closure operation can be skipped.

5.2 RECOVERING SIGNALS FROM NONLINEAR OBSERVATIONS BY STOCHASTIC OPTIMIZATION

The “common denominator” of all estimation problems considered so far in this chapter is that what we observed was obtained by adding noise to the linear image of the unknown signal to be recovered. In this section we consider the problem of signal estimation in the case where the observation is obtained by adding noise to a nonlinear transformation of the signal.

5.2.1 Problem’s setting

Motivating example for what follows is provided by logistic regression model, where

- the unknown signal to be recovered is a vector \( x \) known to belong to a given signal set \( \mathcal{X} \subset \mathbb{R}^n \), which we assume to be a nonempty convex compact set;
- our observation 
\[
\omega^K = \{ \omega_k = (\eta_k, y_k), 1 \leq k \leq K \}
\]
stemming from a signal \( x \) is as follows:

- the regressors \( \eta_1, ..., \eta_K \) are i.i.d. realizations of \( n \)-dimensional random vector \( \eta \) with distribution \( Q \) independent of \( x \) and such that \( Q \) possesses finite and positive definite matrix \( \mathbf{E}_{\eta^\sim Q}\{\eta\eta^T\} \) of second moments;
the labels \( y_k \) are generated as follows: \( y_k \) is independent of the "history" \( \eta_1, \ldots, \eta_{k-1}, y_1, \ldots, y_{k-1} \) Bernoulli random variable, and the conditional, given \( \eta_k \), probability for \( y_k \) to be 1 is \( \phi(\eta_k^T x) \), where

\[
\phi(s) = \frac{\exp\{s\}}{1 + \exp\{s\}}.
\]

In this model, the standard (and very well studied) approach to estimating the signal \( x \) underlying observations is to use Maximum Likelihood (ML) estimate: the logarithm of the conditional, given \( \eta_k \), probability to get the observed labels as a function of a candidate signal \( z \) is

\[
\ell(z, \omega^K) = \sum_{k=1}^K \left[ y_k \ln(\phi(\eta_k^T z)) + (1 - y_k) \ln(1 - \phi(\eta_k^T z)) \right],
\]

and the ML estimate of the "true" signal \( x \) underlying our observation \( \omega^K \) is obtained by maximizing the log-likelihood \( \ell(z, \omega^K) \) over \( z \in \mathcal{X} \):

\[
\hat{x}_{\text{ML}}(\omega^K) \in \text{Argmax}_{z \in \mathcal{X}} \ell(z, \omega^K),
\]

which is a convex optimization problem.

The problem we intend to consider (referred as \textit{generalized linear model} (GLM) in Statistics) can be viewed as a natural generalization of the just presented logistic regression and is as follows:

Our observation depends on unknown signal \( x \) known to belong to a given convex compact set \( \mathcal{X} \subset \mathbb{R}^n \) and is

\[
\omega^K = \{ \omega_k = (\eta_k, y_k), 1 \leq k \leq K \}
\]

with \( \omega_k, 1 \leq k \leq K \), which are i.i.d. realizations of a random pair \( (\eta, y) \) with the distribution \( P_x \) such that

- the regressor \( \eta \) is a random \( n \times m \) matrix with some independent of \( x \) probability distribution \( Q \);
- the label \( y \) is an \( m \)-dimensional random vector such that the conditional, given \( \eta \), distribution of \( y \) induced by \( P_x \) has the expectation \( f(\eta^T x) \):

\[
\mathbb{E}_{y|\eta}^x \{ y \} = f(\eta^T x),
\]

where \( \mathbb{E}_{y|\eta}^x \) is the conditional, \( \eta \) given, distribution of \( y \) stemming from the distribution \( P_x \) of \( \omega = (\eta, y) \), and \( f(\cdot) : \mathbb{R}^m \to \mathbb{R}^m \) ("link function") is a given mapping.

Note that the logistic regression model corresponds to the case where \( m = 1 \), \( f(s) = \frac{\exp\{s\}}{1 + \exp\{s\}} \), and \( y \) takes values 0,1, with the conditional, \( \eta \) given, probability to take value 1 equal to \( f(\eta^T x) \).
Another example is provided by the model
\[ y = f(\eta^T x) + \xi, \]
where \( \xi \) is independent of \( \eta \) random vector with zero mean, say, \( \xi \sim \mathcal{N}(0, \sigma^2 I_m) \). Note that in the latter case the ML estimate of the signal \( x \) underlying observations is
\[
\hat{x}_{\text{ML}}(\omega^K) \in \text{Argmin}_{z \in \mathcal{X}} \sum_k \|y_k - f(\eta_k^T z)\|_2^2.
\]
(5.52)

In contrast to what happens with logistic regression, now the optimization problem – “Nonlinear Least Squares” – responsible for the ML estimate typically is nonconvex and can be computationally difficult.

We intend to impose on the data of the estimation problem we have just described (namely, on \( \mathcal{X}, f(\cdot) \), and the distributions \( P_x, x \in \mathcal{X} \), of the pair \( (\eta, y) \)) assumptions which allow to reduce our estimation problem to a problem with convex structure — a strongly monotone variational inequality represented by stochastic oracle. At the end of the day, this will lead to a consistent estimate of the signal, with explicit “finite sample” accuracy guarantees.

5.2.2 Assumptions

**Preliminaries: monotone vector fields.** A monotone vector field on \( \mathbb{R}^m \) is a single-valued everywhere defined mapping \( g(\cdot) : \mathbb{R}^m \to \mathbb{R}^m \) which possesses the monotonicity property
\[
[g(z) - g(z')]^T[z - z'] \geq 0 \quad \forall z, z' \in \mathbb{R}^m.
\]

We say that such a field is monotone with modulus \( \kappa \geq 0 \) on a closed convex set \( Z \subset \mathbb{R}^m \), if
\[
[g(z) - g(z')]^T[z - z'] \geq \kappa \|z - z'\|_2^2 \quad \forall z, z' \in Z,
\]
and say that \( g \) is strongly monotone on \( Z \) if the modulus of monotonicity of \( g \) on \( Z \) is positive. It is immediately seen that for a monotone vector field which is continuously differentiable on a closed convex set \( Z \) with a nonempty interior, the necessary and sufficient condition for being monotone with modulus \( \kappa \) on the set is
\[
d^T f'(z)d \geq \kappa d^T d \quad \forall (d \in \mathbb{R}^m, z \in Z).
\]
(5.53)

Basic examples of monotone vector fields are:

- gradient fields \( \nabla \phi(x) \) of continuously differentiable convex functions of \( m \) variables or, more generally, the vector fields \( \{\nabla_x \phi(x, y); -\nabla_y \phi(x, y)\} \) stemming from continuously differentiable functions \( \phi(x, y) \) which are convex in \( x \) and concave in \( y \);

- “diagonal” vector fields \( f(x) = [f_1(x_1); f_2(x_2); \ldots; f_m(x_m)] \) with monotonically nondecreasing univariate components \( f_i(\cdot) \). If, in addition, \( f_i(\cdot) \) are continuously differentiable with positive first order derivatives, then the associated field \( f \) is strongly monotone on every compact convex subset of \( \mathbb{R}^m \), the monotonicity modulus depending on the subset.

Monotone vector fields on \( \mathbb{R}^n \) admit simple calculus which includes, in particular, the following two rules:
I. [affine substitution of argument]: If \( f(\cdot) \) is monotone vector field on \( \mathbb{R}^m \) and \( A \) is an \( n \times m \) matrix, the vector field
\[
g(x) = Af(A^T x + a)
\]
is monotone on \( \mathbb{R}^n \); if, in addition, \( f \) is monotone with modulus \( \kappa \geq 0 \) on a closed convex set \( Z \subset \mathbb{R}^m \) and \( X \subset \mathbb{R}^n \) is closed, convex, and such that \( A^T x + a \in Z \) whenever \( x \in X \), \( g \) is monotone with modulus \( \sigma^2 \kappa \) on \( X \), where \( \sigma \) is the minimal singular value of \( A \).

II. [summation]: If \( S \) is a Polish space, \( f(x,s) : \mathbb{R}^m \times S \to \mathbb{R}^m \) is a Borel vector-valued function which is monotone in \( x \) for every \( s \in S \) and \( \mu(ds) \) is a Borel probability measure on \( S \) such that the vector field
\[
F(x) = \int_S f(x,s) \mu(ds)
\]
is well defined for all \( x \), then \( F(\cdot) \) is monotone. If, in addition, \( X \) is a closed convex set in \( \mathbb{R}^m \) and \( f(\cdot,s) \) is monotone on \( X \) with Borel in \( s \) modulus \( \kappa(s) \) for every \( s \in S \), then \( F \) is monotone on \( X \) with modulus \( \int_S \kappa(s) \mu(ds) \).

**Assumptions.** In what follows, we make the following assumptions on the ingredients of the estimation problem posed in Section 5.2.1:

- **A.1.** The vector field \( f(\cdot) \) is continuous and monotone, and the vector field
\[
F(z) = \mathbb{E}_{\eta \sim Q} \{ \eta f(\eta^T z) \}
\]
is well defined (and therefore is monotone along with \( f \) by I, II);

- **A.2.** The signal set \( \mathcal{X} \) is a nonempty convex compact set, and the vector field \( F \) is monotone with positive modulus \( \kappa \) on \( \mathcal{X} \);

- **A.3.** For properly selected \( M < \infty \) and every \( x \in \mathcal{X} \) it holds
\[
\mathbb{E}_{(\eta,y) \sim P} \{ \| \eta y \|_2^2 \} \leq M^2. \tag{5.54}
\]

A simple sufficient condition for the validity of Assumptions A.1-3 with properly selected \( M < \infty \) and \( \kappa > 0 \) is as follows:

- The distribution \( Q \) of \( \eta \) has finite moments of all orders, and \( \mathbb{E}_{\eta \sim Q} \{ \eta \eta^T \} > 0 \);

- \( f \) is continuously differentiable, and \( d^T f'(z) d > 0 \) for all \( d \neq 0 \) and all \( z \). Besides this, \( f \) is of polynomial growth: for some constants \( C \geq 0 \) and \( p \geq 0 \) and all \( z \) one has \( \| f(z) \|_2 \leq C(1 + \| z \|_p^p) \).

Verification of sufficiency is straightforward.

The principal observation underlying the construction we are about to discuss is as follows.

**Proposition 5.11.** With Assumptions A.1-3 in force, let us associate with a pair \((\eta,y) \in \mathbb{R}^{n \times m} \times \mathbb{R}^m\) the vector field
\[
G_{(\eta,y)}(z) = \eta f(\eta^T z) - \eta y : \mathbb{R}^n \to \mathbb{R}^n. \tag{5.55}
\]
Then for every $x \in \mathcal{X}$ we have

$$
E_{(\eta,y) \sim P_x} \left\{ G_{(\eta,y)}(z) \right\} = F(z) - F(x) \quad \forall z \in \mathbb{R}^n \quad (a)
$$

$$
\left\| F(z) \right\|_2 \leq M \quad \forall z \in \mathcal{X} \quad (b)
$$

$$
E_{(\eta,y) \sim P_x} \left\{ \left\| G_{(\eta,y)}(z) \right\|_2^2 \right\} \leq 4M^2 \quad \forall z \in \mathcal{X} \quad (c)
$$

**Proof** is immediate. Indeed, let $x \in \mathcal{X}$. Then

$$
E_{(\eta,y) \sim P_x} \{ \eta y \} = E_{\eta \sim Q} \left\{ E_{\eta} \{ \eta f(\eta^T x) \} \right\} = F(x)
$$

(we have used (5.51) and the definition of $F$), whence,

$$
E_{(\eta,y) \sim P_x} \{ G_{(\eta,y)}(z) \} = E_{(\eta,y) \sim P_x} \{ \eta f(\eta^T z) - \eta y \} = E_{(\eta,y) \sim P_x} \{ \eta f(\eta^T z) \} - F(x)
$$

$$
= E_{\eta \sim Q} \{ \eta f(\eta^T z) \} - F(x) = F(z) - F(x),
$$

as stated in (5.56.a). Besides this, for $x, z \in \mathcal{X}$, taking into account that the marginal distribution of $\eta$ induced by $P_x$ is $Q$, we have

$$
E_{(\eta,y) \sim P_x} \{ \left\| \eta f(\eta^T z) \right\|_2 \} = E_{\eta \sim Q} \left\{ \left\| \eta f(\eta^T z) \right\|_2^2 \right\}
$$

$$
= E_{\eta \sim Q} \left\{ \left\| E_{y \sim P_{\eta}^z} \{ \eta y \} \right\|_2^2 \right\} \quad [\text{since } E_{y \sim P_{\eta}^z} \{ y \} = f(\eta^T z)]
$$

$$
\leq E_{\eta \sim Q} \left\{ E_{y \sim P_{\eta}^z} \{ \left\| \eta y \right\|_2^2 \right\} \right\} \quad [\text{by Jensen’s inequality}]
$$

$$
= E_{(\eta,y) \sim P_x} \{ \left\| \eta y \right\|_2^2 \} \leq M^2 \quad [\text{by A.3 due to } z \in \mathcal{X}].
$$

This combines with the relation $E_{(\eta,y) \sim P_x} \{ \left\| \eta y \right\|_2^2 \} \leq M^2$ given by A.3 due to $x \in \mathcal{X}$ to imply (5.56.b) and (5.56.c).

**Consequences.** Our goal is to recover the signal $x \in \mathcal{X}$ underlying observations (5.50), and under assumptions A.1-3, $x$ is a root of the monotone vector field

$$
G(z) = F(z) - F(x), \quad F(z) = E_{\eta \sim Q} \{ \eta f(\eta^T z) \};
$$

(5.57)

we know that this root belongs to $\mathcal{X}$, and this root is unique because $G(\cdot)$ is strongly monotone on $\mathcal{X}$ along with $F(\cdot)$. Now, finding a root, known to belong to a given convex compact set $\mathcal{X}$, of a strongly monotone on this set vector field $G$ is known to be a computationally tractable problem, provided we have access to an “oracle” which, given on input a point $z \in \mathcal{X}$, returns the value $G(z)$ of the field at the point. The latter is not exactly the case in the situation we are interested in: the field $G$ is the expectation of a random field:

$$
G(z) = E_{(\eta,y) \sim P_x} \{ \eta f(\eta^T z) - \eta y \},
$$

and we do not know a priori what is the distribution over which the expectation is taken. However, we can sample from this distribution – the samples are exactly the observations (5.50), and we can use these samples to approximate somehow $G$ and use this approximation to approximate the signal $x$.\footnote{Whatever simple, the observation expressed by Proposition 5.11 and the resulting course of actions seem to be new. In retrospect, one can recognize unperceived ad hoc utilization of this approach in Perceptron and Isotron algorithms, see [1, 2, 29, 62, 114, 138, 139, 207] and references therein.} Two standard imple-
mentations of this idea are Sample Average Approximation (SAA) and Stochastic Approximation (SA). We are about to consider these two techniques as applied to the situation we are in.

5.2.3 Estimating via Sample Average Approximation

The idea underlying SAA is quite transparent: given observations (5.50), let us approximate the field of interest \( G \) with its empirical counterpart

\[
G_{\omega,K}(z) = \frac{1}{K} \sum_{k=1}^{K} \left[ \eta_k f(\eta_k^T z) - \eta_k y_k \right].
\]

By the Law of Large Numbers, as \( K \to \infty \), the empirical field \( G_{\omega,K} \) converges to the field of interest \( G \), so that under mild regularity assumptions, when \( K \) is large, \( G_{\omega,K} \), with overwhelming probability, will be uniformly on \( X \) close to \( G \). Due to strong monotonicity of \( G \), this would imply that a set of “near-zeros” of \( G_{\omega,K} \) on \( X \) will be close to the zero \( x \) of \( G \), which is nothing but the signal we want to recover. The only question is how we can consistently define a “near-zero” of \( G_{\omega,K} \) on \( X \).

A convenient in our context notion of a “near-zero” is provided by the concept of a weak solution to a variational inequality with monotone operator, defined as follows (we restrict the general definition to the situation of interest):

Let \( X \subset \mathbb{R}^n \) be a nonempty convex compact set, and \( H(z) : X \to \mathbb{R}^n \) be a monotone (i.e., \([H(z) - H(z')]^T [z - z'] \geq 0 \) for all \( z, z' \in X \)) vector field.

A vector \( z_* \in X \) is called a weak solution to the variational inequality (VI) associated with \( H, X \) when

\[
H^T(z)[z - z_*] \geq 0 \quad \forall z \in X.
\]

Let \( X \subset \mathbb{R}^n \) be a nonempty convex compact set and \( H \) be monotone on \( X \). It is well known that

- The VI associated with \( H, X \) (let us denote it \( VI(H, X) \)) always has a weak solution. It is clear that if \( \bar{z} \in X \) is a root of \( H \), then \( \bar{z} \) is a weak solution to \( VI(H, X) \).
- When \( H \) is continuous on \( X \), every weak solution \( \bar{z} \) to \( VI(H, X) \) is also a strong solution, meaning that

\[
H^T(\bar{z})(z - \bar{z}) \geq 0 \quad \forall z \in X. \quad (5.58)
\]

Indeed, (5.58) clearly holds true when \( z = \bar{z} \). Assuming \( z \neq \bar{z} \) and setting \( z_t = \bar{z} + t(z - \bar{z}), 0 < t \leq 1 \), we have \( H^T(z_t)(z_t - \bar{z}) \geq 0 \) (since \( \bar{z} \) is a weak solution), whence \( H^T(z_t)(z - \bar{z}) \geq 0 \) (since \( z - \bar{z} \) is a positive multiple of \( z_t - \bar{z} \)). Passing to limit as \( t \to +0 \) and invoking the continuity of \( H \), we get \( H^T(\bar{z})(z - \bar{z}) \geq 0 \), as claimed.

- When \( H \) is the gradient field of a continuously differentiable convex function on
\(\mathcal{X}\) (such a field indeed is monotone), weak (or, which in the case of continuous \(H\) is the same, strong) solutions to \(VI(H, \mathcal{X})\) are exactly the minimizers of the function on \(\mathcal{X}\).

Note also that a strong solution to \(VI(H, \mathcal{X})\) with monotone \(H\) always is a weak one: if \(\bar{z} \in \mathcal{X}\) satisfies \(H^T(\bar{z})(z - \bar{z}) \geq 0\) for all \(z \in \mathcal{X}\), then \(H(z)^T(z - \bar{z}) \geq 0\) for all \(z \in \mathcal{X}\), since by monotonicity \(H^T(\bar{z})(z - \bar{z}) \geq H^T(\bar{z})(z - \bar{z})\).

In the sequel, we heavily exploit the following simple and well known fact:

**Lemma 5.12.** Let \(\mathcal{X}\) be a convex compact set, and \(H\) be a monotone vector field on \(\mathcal{X}\) with monotonicity modulus \(\kappa > 0\), i.e.

\[
\forall z, z' \in \mathcal{X} \ [H(z) - H(z')]^T[z - z'] \geq \kappa \|z - z\|_2^2.
\]

Further, let \(\bar{z}\) be a weak solution to \(VI(H, \mathcal{X})\). Then the weak solution to \(VI(H, \mathcal{X})\) is unique. Besides this,

\[
H^T(z)[z - \bar{z}] \geq \kappa \|z - \bar{z}\|_2^2 \quad \forall z \in \mathcal{X}.
\]  
\[
(5.59)
\]

**Proof:** Under the premise of Lemma, let \(z \in \mathcal{X}\) and let \(\bar{z}\) be a weak solution to \(VI(H, \mathcal{X})\) (recall that it does exist). Setting \(z_t = \bar{z} + t(z - \bar{z})\), for \(t \in (0, 1)\) we have

\[
H^T(z_t)[z - z_t] \geq H^T(z_t)[z - z_t] + \kappa \|z - z_t\|_2^2 \geq \kappa \|z - z_t\|_2^2,
\]

where the first \(\geq\) is due to strong monotonicity of \(H\), and the second \(\geq\) is due to the fact that \(H^T(z_t)[z - z_t]\) is proportional, with positive coefficient, to \(H^T(z_t)[z_t - z]\), and the latter quantity is nonnegative since \(\bar{z}\) is a weak solution to the VI in question. We end up with \(H^T(z)[z - z_t] \geq \kappa \|z - z_t\|_2^2\); passing to limit as \(t \to +0\), we arrive at \((5.59)\). To prove uniqueness of a weak solution, assume that aside of the weak solution \(\bar{z}\) there exists a weak solution \(\tilde{z}\) distinct form \(\bar{z}\) and, let us set \(z' = \frac{1}{2}[\bar{z} + \tilde{z}]\). Since both \(\bar{z}\) and \(\tilde{z}\) are weak solutions, both the quantities \(H^T(z')[z' - \bar{z}]\) and \(H^T(z')[z' - \bar{z}]\) should be nonnegative, and because the sum of these quantities is 0, both of them are zero. Thus, when applying \((5.59)\) to \(z = z'\), we get \(z' = \bar{z}\), whence \(\tilde{z} = \bar{z}\) as well. \(\Box\)

Now let us come back to the estimation problem under consideration. Let Assumptions A.1-3 hold, so that vector fields \(G_{\eta_k}(\cdot)(\omega)\) defined in \((5.55)\), and therefore vector field \(G_{\nu_k}(\cdot)\) are continuous and monotone. When using the SAA, we compute a weak solution \(\bar{z}(\omega^K)\) to \(VI(G_{\nu_k}, \mathcal{X})\) and treat it as the SAA estimate of signal \(x\) underlying observations \((5.50)\). Since the vector field \(G_{\nu_k}(\cdot)\) is monotone with efficiently computable values, provided that so is \(f\), computing (a high accuracy approximation to) a weak solution to \(VI(G_{\nu_k}, \mathcal{X})\) is a computationally tractable problem (see, e.g., [186]). Moreover, utilizing the techniques from [30, 201, 217, 209, 210], under mild additional to A.1-3 regularity assumptions one can get non-asymptotical upper bound on, say, the expected \(\|\cdot\|_2^2\)-error of the SAA estimate as a function of the sample size \(K\) and find out the rate at which this bound converges to 0 as \(K \to \infty\); this analysis, however, goes beyond our scope.

Let us specify the SAA estimate in the logistic regression model. In this case
we have $f(u) = (1 + e^{-u})^{-1}$, and

$$G_{(\eta_k,y_k)}(z) = \left[ \frac{\exp\{\eta_k^T z\}}{1 + \exp\{\eta_k^T z\}} - y_k \right] \eta_k,$$

$$G_{\omega^K}(z) = \frac{1}{K} \sum_{k=1}^{K} \left[ \frac{\exp\{\eta_k^T z\}}{1 + \exp\{\eta_k^T z\}} - y_k \right] \eta_k$$

$$= \frac{1}{K} \nabla_z \left[ \sum_k \left( \ln \left(1 + \exp\{\eta_k^T z\} \right) - y_k \eta_k^T z \right) \right].$$

In other words, $G_{\omega^K}(z)$ is proportional, with negative coefficient $-1/K$, to the gradient field of the log-likelihood $\ell(z,\omega^K)$, see (5.48). As a result, in the case in question weak solutions to $\text{VI}(G_{\omega^K},\mathcal{X})$ are exactly the maximizers of the log-likelihood $\ell(z,\omega^K)$ over $z \in \mathcal{X}$, that is, for the logistic regression the SAA estimate is nothing but the Maximum Likelihood estimate $\hat{x}_{\text{ML}}(\omega^K)$ as defined in (5.49).

On the other hand, in the “nonlinear least squares” example described in Section 5.2.1 with (for the sake of simplicity, scalar) monotone $f(\cdot)$ the vector field $G_{\omega^K}(\cdot)$ is given by

$$G_{\omega^K}(z) = \frac{1}{K} \sum_{k=1}^{K} \left[ f(\eta_k^T z) - y_k \right] \eta_k$$

which is different (provided that $f$ is nonlinear) from the gradient field

$$2 \sum_{k=1}^{K} f'(\eta_k^T z) \left[ f(\eta_k^T z) - y_k \right] \eta_k$$

of the minus log-likelihood appearing in (5.52). As a result, in this case the ML estimate (5.52) is, in general, different from the SAA estimate (and, in contrast to the ML, the SAA estimate is easy to compute).

---

10This phenomenon is specific for the logistic regression model. The equality of the SAA and the ML estimates in this case is due to the fact that the logistic sigmoid $f(s) = \exp\{s\}/(1+\exp\{s\})$ “happens” to satisfy the identity $f'(s) = f(s)(1-f(s))$. When replacing the logistic sigmoid with $f(s) = \phi(s)/(1+\phi(s))$ with differentiable monotonically nondecreasing positive $\phi(\cdot)$, the SAA estimate becomes the weak solution to $\text{VI}(\Phi,\mathcal{X})$ with

$$\Phi(z) = \sum_{k} \left[ \frac{\phi(\eta_k^T z)}{1 + \phi(\eta_k^T z)} - y_k \right] \eta_k.$$

On the other hand, the gradient field of the minus log-likelihood

$$- \sum_k \left[ y_k \ln(f(\eta_k^T z)) + (1 - y_k) \ln(1 - f(\eta_k^T z)) \right]$$

we need to minimize when computing the ML estimate is

$$\Psi(z) = \sum_{k} \phi'(\eta_k^T z) \left[ \frac{\phi(\eta_k^T z)}{1 + \phi(\eta_k^T z)} - y_k \right] \eta_k.$$

When $k > 1$ and $\phi$ is not an exponent, $\Phi$ and $\Psi$ are “essentially different,” so that the SAA estimate typically will differ from the ML one.
5.2.4 Stochastic Approximation estimate

The Stochastic Approximation (SA) estimate stems from a simple algorithm – Subgradient Descent – for solving variational inequality VI($G, X$). Were the values of the vector field $G(\cdot)$ available, one could approximate a root $x \in X$ of this VI using the recurrence

$$z_k = \text{Proj}_X[z_{k-1} - \gamma_k G(z_{k-1})], \ k = 1, 2, ..., K,$$

where

- $\text{Proj}_X[z]$ is the metric projection of $\mathbb{R}^n$ onto $X$:
  $$\text{Proj}_X[z] = \arg\min_{u \in X} \|z - u\|_2;$$
- $\gamma_k > 0$ are given stepsizes;
- the initial point $z_0$ is an arbitrary point of $X$.

It is well known that under Assumptions A.1-3 this recurrence with properly selected stepsizes and started at a point from $X$ allows to approximate the root of $G$ (in fact, the unique weak solution to VI($G, X$)) to a whatever high accuracy, provided $K$ is large enough. However, we are in the situation when the actual values of $G$ are not available; the standard way to cope with this difficulty is to replace in the above recurrence the “unobservable” values $G(z_{k-1})$ of $G$ with their unbiased random estimates $G(\eta_k, y_k)(z_{k-1})$. This modification gives rise to Stochastic Approximation (coming back to [143]) – the recurrence

$$z_k = \text{Proj}_X[z_{k-1} - \gamma_k G(\eta_k, y_k)(z_{k-1})], \ 1 \leq k \leq K,$$

where $z_0$ is a once and forever chosen point from $X$, and $\gamma_k > 0$ are deterministic stepsizes.

Next item on our agenda is the (well known) convergence analysis of SA under assumptions A.1-3. To this end observe that $z_k$ are deterministic functions of the initial fragments $\omega^k = \{\omega_t, 1 \leq t \leq k\} \sim \prod_{x} P_x \times \ldots \times P_x$ of our sequence of observations $\omega^K = \{\omega_k = (\eta_k, y_k), 1 \leq k \leq K\}$: $z_k = Z_k(\omega^k)$. Let us set

$$D_k(\omega^k) = \frac{1}{2} \|Z_k(\omega^k) - x\|_2^2 = \frac{1}{2} \|z_k - x\|_2^2, \ \ d_k = E_{\omega^k \sim P_x} \{D_k(\omega^k)\},$$

where $x \in X$ is the signal underlying observations (5.50). Note that, as it is well known, the metric projection onto a closed convex set $X$ is contracting:

$$\forall (z \in \mathbb{R}^n, u \in X) : \|\text{Proj}_X[z] - u\|_2 \leq \|z - u\|_2.$$

Consequently, for $1 \leq k \leq K$ it holds

$$D_k(\omega^k) = \frac{1}{2} \|\text{Proj}_X[z_{k-1} - \gamma_k G_{\omega_k}(z_{k-1})] - x\|_2^2$$

$$\leq \frac{1}{2} \|z_{k-1} - \gamma_k G_{\omega_k}(z_{k-1}) - x\|_2^2$$

$$= \frac{1}{2} \|z_{k-1} - x\|_2^2 - \gamma_k G_{\omega_k}(z_{k-1})^T(z_{k-1} - x) + \frac{1}{2} \gamma_k^2 \|G_{\omega_k}(z_{k-1})\|_2^2.$$
keeping in mind relations (5.56) along with the fact that $z_{k-1} \in \mathcal{X}$, we get
\[ d_k \leq d_{k-1} - \gamma_k \mathbb{E}_{\omega_{k-1} \sim P_{z_{k-1}}^{x}} \{ G(z_{k-1})^T (z_{k-1} - x) \} + 2\gamma_k^2 M^2. \] 
(5.61)
Recalling that we are in the case where $G$ is strongly monotone on $\mathcal{X}$ with modulus $\kappa > 0$, $x$ is the weak solution VI$(G, \mathcal{X})$, and $z_{k-1}$ takes values in $\mathcal{X}$, invoking (5.59), the expectation in (5.61) is at least $2\kappa d_k$, and we arrive at the relation
\[ d_k \leq (1 - 2\kappa \gamma_k) d_{k-1} + 2\gamma_k^2 M^2. \] 
(5.62)
We put
\[ S = \frac{2M^2}{\kappa}, \quad \gamma_k = \frac{1}{\kappa(k+1)}. \] 
(5.63)
Let us verify by induction in $k$ that for $k = 0, 1, ..., K$ it holds
\[ d_k \leq (k+1)^{-1} S. \] 
(\text{*}_k)

**Base** $k = 0$. Let $D$ stand for the $\| \cdot \|_2$-diameter of $\mathcal{X}$, and $z_+ \in \mathcal{X}$ be such that $\| z_+ - z_- \|_2 = D$. By (5.56) we have $\| F(z) \|_2 \leq M$ for all $z \in \mathcal{X}$, and by strong monotonicity of $G(\cdot)$ on $\mathcal{X}$ we have
\[ [G(z_+) - G(z_-)]^T [z_+ - z_-] = [F(z_+) - F(z_-)][z_+ - z_-] \geq \kappa \| z_+ - z_- \|_2^2 = \kappa D^2. \]
By Cauchy inequality, the left hand side in the concluding $\geq$ is at most $2MD$, and we get
\[ D \leq \frac{2M}{\kappa}, \]
whence $S \geq D^2/2$. On the other hand, due to the origin of $d_0$ we have $d_0 \leq D^2/2$. Thus, $(\text{*}_0)$ holds true.

**Inductive step** $(\text{*}_{k-1}) \Rightarrow (\text{*}_k)$. Now assume that $(\text{*}_{k-1})$ holds true for some $k$, $1 \leq k \leq K$, and let us prove that $(\text{*}_k)$ holds true as well. Observe that $\kappa \gamma_k = (k+1)^{-1} \leq 1/2$, so that
\[ d_k \leq d_{k-1} (1 - 2\kappa \gamma_k) + 2\gamma_k^2 M^2 \ [\text{by } (5.62)] \]
\[ \leq \frac{S}{k} (1 - 2\kappa \gamma_k) + 2\gamma_k^2 M^2 \ [\text{by } (\text{*}_{k-1}) \text{ and due to } \kappa \gamma_k \leq 1/2] \]
\[ = \frac{S}{k} \left( 1 - \frac{2}{k+1} \right) + \frac{S}{(k+1)^2} \left( \frac{k-1}{k} + \frac{1}{k+1} \right) \leq \frac{S}{k+1}, \]
so that $(\text{*}_k)$ holds true. Induction is complete.

Recalling that $d_k = \frac{1}{2} \mathbb{E} \{ \| z_k - x \|_2^2 \}$, we arrive at the following

**Proposition 5.13.** **Under Assumptions A.1-3 and with the stepsizes**
\[ \gamma_k = \frac{1}{\kappa(k+1)}, \quad k = 1, 2, ... \] 
(5.64)

**for every signal** $x \in \mathcal{X}$ **the sequence of estimates** $\hat{x}_k(\omega^k) = z_k$ **given by the SA**
recurrence (5.60) and \( \omega_k = (\eta_k, y_k) \) defined in (5.50) obeys the error bound

\[
E_{\omega^k \sim P_{\omega^k}} \{ \| \tilde{x}_k(\omega^k) - x \|_2^2 \} \leq \frac{4M^2}{2^2(k+1)}, \quad k = 0, 1, \ldots
\]

(5.65)

\( P_{\omega} \) being the distribution of \((\eta, y)\) stemming from signal \( x \).

5.2.5 Numerical illustration

To illustrate the above developments, we present here the results of some numerical experiments. Our deliberately simplistic setup is as follows:

- \( \mathcal{X} = \{ x \in \mathbb{R}^n : \| x \|_2 \leq 1 \} \);
- the distribution \( Q \) of \( \eta \) is \( \mathcal{N}(0, I_n) \);
- \( f \) is the monotone vector field on \( \mathbb{R} \) given by one of the following four options:
  - A. \( f(s) = \exp\{s\}/(1 + \exp\{s\}) \) ("logistic sigmoid");
  - B. \( f(s) = s \) ("linear regression");
  - C. \( f(s) = \max[s, 0] \) ("hinge function");
  - D. \( f(s) = \min[1, \max[s, 0]] \) ("ramp sigmoid");
- conditional, given \( \eta \), distribution of \( y \) induced by \( P_x \) is
  - Bernoulli distribution with probability \( f(\eta^T x) \) of outcome 1 in the case of A (i.e., A corresponds to the logistic model),
  - Gaussian distribution \( \mathcal{N}(f(\eta^T x), I_n) \) in cases B – D.

Note that when \( m = 1 \) and \( \eta \sim \mathcal{N}(0, I_n) \), one can easily compute the field \( F(z) \). Indeed, we have \( \forall z \in \mathbb{R}^n \setminus \{0\} \):

\[
\eta = \frac{zz^T}{\| z \|_2^2} \eta + \left( I - \frac{zz^T}{\| z \|_2^2} \right) \eta_{\perp},
\]

and due to the independence of \( \eta^T z \) and \( \eta_{\perp} \),

\[
F(z) = E_{\eta \sim \mathcal{N}(0, I)} \{ \eta f(\eta^T z) \} = E_{\eta \sim \mathcal{N}(0, I)} \left\{ \frac{zz^T}{\| z \|_2^2} f(\eta^T z) \right\} = \frac{z}{\| z \|_2} E_{\zeta \sim \mathcal{N}(0, 1)} \{ \zeta f(\| z \|_2\zeta) \},
\]

so that \( F(z) \) is proportional to \( z/\| z \|_2 \) with proportionality coefficient

\[
h(\| z \|_2) = E_{\zeta \sim \mathcal{N}(0, 1)} \{ \zeta f(\| z \|_2\zeta) \}.
\]

In Figure 5.2 we present the plots of the function \( h(t) \) for the situations A – D, same as the dependencies of the moduli of strong monotonicity of the corresponding mappings \( F \) in a centered at the origin \( \| \cdot \|_2 \)-ball of radius \( R \) on \( R \). The dimension \( n \) in all experiments was set to 100, and the number of observations \( K \) was 400, 1000, 4000, 10000, and 40000. For each combination of parameters we ran 10 simulations for signals \( x \) underlying observations (5.50) drawn randomly from the uniform distribution on the unit sphere (the boundary of \( \mathcal{X} \)).
Figure 5.2: Left: functions $h$; right: moduli of strong monotonicity of the operators $F(\cdot)$ in $\{z : \|z\|_2 \leq R\}$ as functions of $R$. Dashed lines – case A (logistic sigmoid), solid lines – case B (linear regression), dash-dotted lines – case C (hinge function), dotted line – case D (ramp sigmoid).

In each experiment, we computed the SAA and the SA estimates (note that in the cases A and B the SAA estimate is the Maximum Likelihood estimate as well). The SA stepsizes $\gamma_k$ were selected according to (5.64) with “empirically tuned” $\kappa$. Namely, given observations $\omega_k = (\eta_k, y_k)$, $k \leq K$, see (5.50), we used them to build the SA estimate in two stages:
— at tuning stage, we generate a random “training signal” $x' \in \mathcal{X}$ and then generate labels $y'_k$ as if $x'$ were the actual signal. For instance, in the case of A, $y'_k$ is assigned value 1 with probability $f(\eta_k^T x')$ and value 0 with complementary probability. After “training signal” and associated labels are generated, we run on the resulting artificial observations SA with different values of $\kappa$, compute the accuracy of the resulting estimates, and select the value of $\kappa$ resulting in the best recovery;
— at execution stage, we run SA on the actual data with stepsizes (5.64) specified by $\kappa$ found at the tuning stage.

The results of some numerical experiments are presented in Figure 5.3.

Note that the CPU time for SA includes both tuning and execution stages. The conclusion from these experiments is that as far as estimation quality is concerned, the SAA estimate marginally outperforms the SA, while being significantly more time consuming. Note also that the observed in our experiments dependence of recovery errors on $K$ is consistent with the convergence rate $O(1/\sqrt{K})$ established by Proposition 5.13.

**Comparison with Nonlinear Least Squares.** Observe that in the case $m = 1$ of scalar monotone field $f$, the SAA estimate yielded by our approach as applied to observation $\omega^K$ is the minimizer of the convex function

$$H_{\omega^K}(z) = \frac{1}{K} \sum_{k=1}^{K} \left[ v(\eta_k^T z) - y_k \eta_k^T z \right], \ v(r) = \int_0^r f(s)ds.$$ 

on the signal set $\mathcal{X}$. When $f$ is the logistic sigmoid, $H_{\omega^K}(\cdot)$ is exactly the convex loss function leading to the ML estimate in the logistic regression model. As we have already mentioned, this is not the case for a general GLM. Consider, e.g., the vectors fields $F(\cdot)$ we are interested in analytically, but this would be boring and conservative.
situation where the regressors and the signals are reals, the distribution of regressor $\eta$ is $\mathcal{N}(0, 1)$, and the conditional, given $\eta$, distribution of $y$ is $\mathcal{N}(f(\eta x), \sigma^2)$, with $f(s) = \arctan(s)$. In this situation the ML estimate stemming from observation $\omega^K$ is the minimizer on $X$ of the function

$$M_{\omega^K}(z) = \frac{1}{K} \sum_{k=1}^{K} [y_k - \arctan(\eta_k z)]^2.$$  \hspace{1cm} (5.66)

The latter function is typically nonconvex and can be multi-extremal. For example, when running simulations\textsuperscript{12} we from time to time observe the situation similar to that presented in Figure 5.4.

Of course, in our toy situation of scalar $x$ existence of several local minima of $M_{\omega^K}(\cdot)$ is not an issue — we can easily compute the ML estimate by brute force search along a dense grid. What to do in the multidimensional case – this is another question. We could also add that in the simulations which led to Figure 5.4 both the SAA and the ML estimates exhibited nearly the same performance in terms of the estimation error: in 1000 experiments, the median of the observed recovery errors was 0.969 for the ML, and 0.932 for the SAA estimate. When increasing the number of observations to 1000, the empirical median, taken over 1000 simulations, of recovery errors became 0.079 for the ML, and 0.085 for the SAA estimate.

5.2.6 ”Single-observation” case

Let us look at the special case of our estimation problem where the sequence $\eta_1, ..., \eta_K$ of regressors in (5.50) is deterministic. At the first glance, this situation goes beyond our setup, where the regressors should be i.i.d. drawn from

\textsuperscript{12}In these simulations, the “true” signal $x$ underlying observations was drawn from $\mathcal{N}(0, 1)$, the number $K$ of observations also was random with uniform distribution on $\{1, ..., 20\}$, and $X = [-20, 20]$, $\sigma = 3$ were used.
some distribution \( Q \). We can, however, circumvent this “contradiction” by saying that we are now in the single-observation case with the regressor being the matrix \([\eta_1, \ldots, \eta_K]\) and \( Q \) being a degenerate distribution supported at a singleton. Specifically, consider the case where our observation is

\[
\omega = (\eta, y) \in \mathbb{R}^{n \times mK} \times \mathbb{R}^{mK}
\]  

(5.67)

\((m, n, K\) are given positive integers), and the distribution \( P_x \) of observation stemming from a signal \( x \in \mathbb{R}^n \) is as follows:

- \( \eta \) is a given independent of \( x \) deterministic matrix;
- \( y \) is random, and the distribution of \( y \) induced by \( P_x \) is with mean \( \phi(\eta^T x) \), where \( \phi : \mathbb{R}^{mK} \to \mathbb{R}^{mK} \) is a given mapping.

As an instructive example connecting our current setup with the previous one, consider the case where \( \eta = [\eta_1, \ldots, \eta_K] \) with \( n \times m \) deterministic “individual regressors” \( \eta_k \), \( y = [y_1; \ldots; y_K] \) with random “individual labels” \( y_k \in \mathbb{R}^m \) conditionally independent, given \( x \), across \( k \), and such that the induced by \( x \) expectations of \( y_k \) are \( f(\eta_k^T x) \) for some \( f : \mathbb{R}^m \to \mathbb{R}^m \). We set \( \phi([u_1; \ldots; u_K]) = [f(u_1); \ldots; f(u_K)] \). The resulting “single observation” model is a natural analogy of the \( K \)-observation model considered so far, the only difference being that the individual regressors now form a fixed deterministic sequence rather than being a sample of realizations of some random matrix.

Same as before, our goal is to use observation (5.67) to recover the (unknown) signal \( x \) underlying, as explained above, the distribution of the observation. Formally, we are now in the case \( K = 1 \) of our previous recovery problem where \( Q \) is supported on a singleton \( \{\eta\} \) and can use the constructions developed so far. Specifically,
• The field $F(z)$ associated with our problem (it used to be $E_{\eta \sim Q}\{\eta f(\eta^T z)\}$) is
\[
F(z) = \eta \phi(\eta^T z),
\]
and the vector field $G(z) = F(z) - F(x)$, $x$ being the signal underlying observation (5.67), is
\[
G(z) = E_{(\eta, y) \sim P_z}\{F(z) - \eta y\}
\]
(cf. (5.57)). Same as before, the signal to recover is a zero of the latter field. Note that now the vector field $F(z)$ is observable, and the vector field $G$ still is the expectation, over $P_z$, of an observable vector field:
\[
G(z) = E_{(\eta, y) \sim P_z}\{\eta \phi(\eta^T z) - \eta y\},
\]
cf. Lemma 5.11.
• Assumptions A.1-2 now read
\[
A.1' \quad \text{The vector field } \phi(\cdot) : \mathbb{R}^{mK} \to \mathbb{R}^{mK} \text{ is continuous and monotone, so that } F(\cdot) \text{ is continuous and monotone as well,}
\]
\[
A.2' \quad \mathcal{X} \text{ is a nonempty compact convex set, and } F \text{ is strongly monotone, with modulus } \kappa > 0, \text{ on } \mathcal{X}.
\]
A simple sufficient condition for the validity of the above monotonicity assumptions is positive definiteness of the matrix $\eta \eta^T$ plus strong monotonicity of $\phi$ on every bounded set.
• For our present purposes, it is convenient to reformatulate assumption A.3 in the following equivalent form:
\[
A.3' \quad \text{For properly selected } \sigma \geq 0 \text{ and every } x \in \mathcal{X} \text{ it holds}
\]
\[
E_{(\eta, y) \sim P_z}\{||y - \phi(\eta^T x)||^2\} \leq \sigma^2.
\]
In the present setting, the SAA $\hat{x}(y)$ is the unique weak solution to VI($G_y, \mathcal{X}$), and we can easily quantify the quality of this estimate:

**Proposition 5.14.** In the situation in question, let Assumptions A.1'-3' hold. Then for every $x \in \mathcal{X}$ and every realization $(\eta, y)$ of induced by $x$ observation (5.67) one has
\[
\|\hat{x}(y) - x\|_2 \leq \kappa^{-1} \|\eta[y - \phi(\eta^T x)]\|_2, \quad (5.68)
\]
whence also
\[
E_{(\eta, y) \sim P_z}\{\|\hat{x}(y) - x\|_2^2\} \leq \sigma^2/\kappa^2. \quad (5.69)
\]

**Proof.** Let $x \in \mathcal{X}$ be the signal underlying observation (5.67), and $G(z) = F(z) - F(x)$ be the associated vector field $G$. We have
\[
G_y(z) = F(z) - \eta y = F(z) - F(x) + [F(x) - \eta y] = G(z) - \eta[y - \phi(\eta^T x)] = G(z) - \Delta(x, y).
\]
For $y$ fixed, $\hat{z} = \hat{x}(y)$ is the weak, and therefore the strong (since $G_y(\cdot)$ is continuous) solution to VI($G_y, \mathcal{X}$), implying, due to $x \in \mathcal{X}$, that
\[
0 \leq G_y(\hat{z})[x - \hat{z}] = G^T(\hat{z})[x - \hat{z}] - \Delta^T(x, y)[x - \hat{z}],
\]
whence
\[-G^T(z)[x - \bar{z}] \leq -\Delta^T(x, y)[x - \bar{z}].\]

Besides this, \(G(x) = 0\), whence \(G^T(x)[x - \bar{z}] = 0\), and we arrive at
\[[G(x) - G(\bar{z})]^T [x - \bar{z}] \leq -\Delta^T(x, y)[x - \bar{z}],\]
whence also
\[\kappa \|x - \bar{z}\|^2 \leq -\Delta^T(x, y)[x - \bar{z}].\]

(recall that \(G\), along with \(F\), is strongly monotone with modulus \(\kappa\) on \(\mathcal{X}\) and \(x, \bar{z} \in \mathcal{X}\)). Applying the Cauchy inequality, we arrive at (5.68).

**Example.** Consider the case where \(m = 1\), \(\phi\) is strongly monotone, with modulus \(\kappa_\phi > 0\), on the entire \(\mathbb{R}^K\), and \(\eta\) in (5.67) is drawn from a “Gaussian ensemble” – the columns \(\eta_k\) of the \(n \times K\) matrix \(\eta\) are independent \(\mathcal{N}(0, I_n)\)-random vectors. Assume also that the observation noise is Gaussian:
\[y = \phi(\eta^T x) + \lambda \xi, \quad \xi \sim \mathcal{N}(0, I_K)\]

It is well known that as \(K/n \to \infty\), the minimal singular value of the \(n \times n\) matrix \(\eta \eta^T\) is at least \(O(1)K\) with overwhelming probability, implying that when \(K/n \gg 1\), the typical modulus of strong monotonicity of \(F(\cdot)\) is \(\kappa \geq O(1)K\kappa_\phi\). Furthermore, in our situation, as \(K/n \to \infty\), the Frobenius norm of \(\eta\) with overwhelming probability is at most \(O(1)\sqrt{nK}\). In other words, when \(K/n\) is large, a “typical” recovery problem from the just described ensemble satisfies the premise of Proposition 5.14 with \(\kappa = O(1)K\kappa_\phi\) and \(\sigma^2 = O(\lambda^2 n K)\). As a result, (5.69) reads
\[E_{(n, y) \sim P_x}\{\|\hat{x}(y) - x\|^2\} \leq O(1)\frac{\lambda^2 n}{\kappa^2 \phi K} \quad [K \gg n]\]

It is well known that in the standard case of linear regression, where \(\phi(x) = \kappa_\phi x\), the resulting bound is near-optimal, provided \(\mathcal{X}\) is large enough.

**Numerical illustration:** in the situation described in the example above, we set \(m = 1\), \(n = 100\) and use
\[\phi(u) = \arctan[u] := [\arctan(u_1); \ldots; \arctan(u_K)] : \mathbb{R}^K \to \mathbb{R}^K;\]
the set \(\mathcal{X}\) is the unit ball \(\{x \in \mathbb{R}^n : \|x\|_2 \leq 1\}\). In a particular experiment, \(\eta\) is chosen at random from the Gaussian ensemble as described above, and signal \(x \in \mathcal{X}\) underlying observation (5.67) is drawn at random; the observation noise \(y - \phi(\eta^T x)\) is \(\mathcal{N}(0, \lambda^2 I_K)\). Some typical results (10 simulations for each combination of the samples size and noise variance \(\lambda^2\)) are presented in Figure 5.5.

### 5.3 Exercises for Chapter 5

#### 5.3.1 Estimation by Stochastic Optimization

**Exercise 5.1.**

Consider the following “multinomial” version of logistic regression problem from
Section 5.2.1:
For $k = 1, ..., K$, we observe pairs

$$(\zeta_k, \ell_k) \in \mathbb{R}^n \times \{0, 1, ..., m\}$$

(5.70)

drawn independently of each other from a probability distribution $P_x$ parameterized by an unknown signal $x = [x^1; ...; x^m] \in \mathbb{R}^n \times ... \times \mathbb{R}^n$ as follows:

- The probability distribution of regressor $\zeta$ induced by the distribution $S_x$ of $(\zeta, \ell)$ is a once for ever fixed independent of $x$ distribution $R$ on $\mathbb{R}^n$ with finite second order moments and positive definite matrix $Z = E_{\zeta \sim R}\{\zeta \zeta^T\}$ of second order moments;
- The conditional, $\zeta$ given, probability distribution of label $\ell$ induced by the distribution $S_x$ of $(\zeta, \ell)$ is the distribution of discrete random variable taking value $\iota \in \{0, 1, ..., m\}$ with probability

$$p_\iota = \begin{cases} 
\frac{\exp\{\zeta^T x^\iota\}}{1 + \sum_{i=1}^m \exp\{\zeta^T x^i\}}, & 1 \leq \iota \leq m, \\
\frac{1}{1 + \sum_{i=1}^m \exp\{\zeta^T x^i\}}, & \iota = 0.
\end{cases}$$

Given a nonempty convex compact set $X \in \mathbb{R}^{mn}$ known to contain the (unknown) signal $x$ underlying observations (5.70), we want to recover $x$. Note that the recovery problem associated with the standard logistic regression model is the case $m = 1$ of the just defined problem.

Your task is to process the above recovery problem via the approach developed in Section 5.2 and to compare the resulting SAA estimate with the Maximum Likelihood estimate.

**Exercise 5.2.**

Let

$$H(x) : \mathbb{R}^n \to \mathbb{R}^n$$
be a strongly monotone and Lipschitz continuous on the entire space vector field:
\[
\forall (x, x' \in \mathbb{R}^n) : \quad [H(x) - H(x')]^T[x - x'] \geq \kappa \|x - x'\|_2^2,
\]
\[
\|H(x) - H(x')\|_2 \leq L\|x - x'\|_2
\]
for some \( \kappa > 0 \) and \( L < \infty \).

1.1) Prove that for every \( x \in \mathbb{R}^n \), the vector equation
\[
H(z) = x
\]
in variable \( z \in \mathbb{R}^n \) has unique solution (which we denote \( H^{-1}(x) \)), and that for every \( x, y \in \mathbb{R}^n \) one has
\[
\|H^{-1}(x) - y\|_2 \leq \kappa^{-1}\|x - H(y)\|_2.
\]

1.2) Prove that the vector field
\[
x \mapsto H^{-1}(x)
\]
is strongly monotone with modulus
\[
\kappa_* = \frac{\kappa}{L^2}
\]
and Lipschitz continuous, with constant \( 1/\kappa \) w.r.t. \( \| \cdot \|_2 \), on the entire \( \mathbb{R}^n \).

Let us interpret \( -H(\cdot) \) as a field of “reaction forces” applied to a particle: when particle is in a position \( y \in \mathbb{R}^n \) the reaction force applied to the particle is \( -H(y) \). Next, let us interpret \( x \in \mathbb{R}^n \) as an external force applied to the particle. An equilibrium \( y \) is a point in space where the reaction force \( -H(y) \) compensates the external force, that is, \( H(y) = x \), or, which for our \( H \) is the same, \( y = H^{-1}(x) \).

Note that with this interpretation, strong monotonicity of \( H \) makes perfect sense, implying that the equilibrium in question is stable: when the particle is moved from the equilibrium \( y = H^{-1}(x) \) to a position \( y + \Delta \), the total force acting at the particle becomes \( f = x - H(y + \Delta) \), so that
\[
f^T\Delta = [x - H(y + \Delta)]^T\Delta = [H(y) - H(y + \Delta)]^T[\Delta] \leq -\kappa\Delta^2,
\]
that is, the force is oriented “against” the displacement \( \Delta \) and “wants” to return the particle to the equilibrium position.

Now imagine that we can observe in noise equilibrium \( H^{-1}(x) \) of the particle, the external force \( x \) being unknown, and want to recover \( x \) from our observation. For the sake of simplicity, let the observation noise be zero mean Gaussian, so that our observation is
\[
\omega = H^{-1}(x) + \sigma \xi, \; \xi \sim \mathcal{N}(0, I_n).
\]

2) Verify that the recovery problem we have posed is a special case of the “single observation” recovery problem from Section 5.2.6, with \( \mathbb{R}^n \) in the role of \( \mathcal{X} \) and that the SAA estimate \( \hat{x}(\omega) \) from that section under the circumstances is

\[\text{In Section 5.2.6, } \mathcal{X} \text{ was assumed to be closed, convex, and bounded; a straightforward inspection shows that when the vector field } \phi \text{ is strongly monotone, with some positive modulus, on the entire space, and } \eta \text{ has trivial kernel, all constructions and results of Section 5.2.6 can be extended to the case of an arbitrary closed convex } \mathcal{X} \text{.} \]
just the root of the equation
\[ H^{-1}(\cdot) = \omega, \]
that is,
\[ \hat{x}(\omega) = H(\omega). \]

Prove also that
\[ \mathbb{E}\{\|\hat{x}(\omega) - x\|_2^2\} \leq n\sigma^2 L^2. \] (5.73)

Note that in the situation in question the ML estimate should be the minimizer of
the function
\[ f(z) = \|\omega - H^{-1}(z)\|_2^2, \]
and this minimizer is nothing but \( \hat{x}(\omega) \).

**Exercise 5.3.**

[identification of parameters of a linear dynamic system] Consider the problem as follows:

A deterministic sequence \( x = \{x_t : t \geq -d+1\} \) satisfies linear finite-difference equation
\[ \sum_{i=0}^{d} \alpha_i x_{t-i} = y_t, \ t = 1, 2, \ldots \] (5.74)
of given order \( d \) and is bounded:
\[ |x_t| \leq M_x < \infty, \forall t \geq -d + 1, \]
implying that the sequence \( \{y_t\} \) also is bounded:
\[ |y_t| \leq M_y < \infty, \forall t \geq 1. \]

The vector \( \alpha = [\alpha_0; \ldots; \alpha_d] \) is unknown, all we know is that this vector belongs to a given closed convex set \( X \subset \mathbb{R}^{d+1} \). We have at our disposal observations
\[ \omega_t = x_t + \sigma_x \xi_t, \ -d + 1 \leq t \leq K, \] (5.75)
of the terms in the sequence, with \( \xi_t \sim \mathcal{N}(0,1) \) independent across \( t \), with some given \( \sigma_x \), same as observations
\[ \zeta_t = y_t + \sigma_y \eta_t \] (5.76)
with \( \eta_t \sim \mathcal{N}(0,1) \) independent across \( t \) and independent of \( \{\xi_t\}_t \). Our goal is to recover from these observations the vector \( \alpha \).

**Strategy.** To get the rationale underlying the construction to follow, let us start with the case when there is no observation noise at all: \( \sigma_x = \sigma_y = 0 \). In this case we could act as follows: let us denote
\[ x^t = [x_t; x_{t-1}; x_{t-2}; \ldots; x_{t-d}], \ 1 \leq t \leq K, \]
and rewrite (5.74) as
\[ [x^t]^T \alpha = y_t, \ 1 \leq t \leq K. \]
When setting
\[ A_K = \frac{1}{K} \sum_{t=1}^{K} x^t [x^t]^T, \quad a_K = \frac{1}{K} \sum_{t=1}^{K} y_t x^t, \]
we get
\[ A_K \alpha = a_K. \tag{5.77} \]
Assuming that \( K \) is large and trajectory \( x \) is “rich enough” to ensure that \( A_K \) is nonsingular, we could identify \( \alpha \) by solving the linear system (5.77).

Now, when the observation noise is present, we could try to use the noisy observations of \( x^t \) and \( y_t \) we have at our disposal in order to build good, for large \( K \), empirical approximations to \( A_K \) and \( a_K \), and identify \( \alpha \) by solving the “empirical counterpart” of (5.77). The straightforward way would be to define \( \omega^t \) an “observable version” of \( x^t \):
\[ \omega^t = [\omega_t; \omega_{t-1}; \ldots; \omega_{t-d}] = x^t + \sigma_x [\xi_t; \xi_{t-1}; \ldots; \xi_{t-d}] \]
and to replace \( A_K \), resp., \( a_K \) with
\[ \tilde{A}_t = \frac{1}{K} \sum_{t=1}^{K} \omega^t [\omega^t]^T, \quad \tilde{a}_K = \sum_{t=1}^{K} \zeta_t \omega^t. \]
As far as empirical approximation of \( a_K \) is concerned, this approach works: we have
\[ \tilde{a}_K = a_K + \delta a_K, \quad \delta a_K = \frac{1}{K} \sum_{t=1}^{K} [\sigma_x y_t \xi^t + \sigma_y \eta_t x^t + \sigma_x \sigma_y \eta_t \xi^t]. \]

Since the sequence \( \{y_t\} \) is bounded, the random error \( \delta a_K \) of approximation \( \tilde{a}_K \) of \( a_K \) is small for large \( K \) with overwhelming probability. Indeed, \( \delta a_K \) is the average of \( K \) zero mean random vectors \( \delta_t \) (recall that \( \xi^t \) and \( \eta_t \) are independent and zero mean) with \( \| \delta_t \|_2^2 \leq \sigma_x^2 M_d^2 + \sigma_y^2 M_x^2 + \sigma_x^2 \sigma_y^2 \).

The quality of approximating \( A_K \) with \( \tilde{A}_K \) is essentially worse: setting
\[ \delta A_K = \tilde{A}_K - A_K = \frac{1}{K} \sum_{t=1}^{K} [\sigma_x^2 \xi^t [\xi^t]^T + \sigma_x \xi^T [x^t]^T + \sigma_x x^T [\xi^t]^T] \]
we see that \( \delta A_K \) is the average of \( K \) random matrices \( \Delta_t \) with nonzero mean, namely, the mean \( \sigma_x^2 I_{d+1} \) and as such \( \Delta A_K \) is “large” independently of how large is \( K \). There is, however, a simple way to overcome this difficulty – splitting obser-
\[ \text{We use the elementary inequality } \| \sum_{t=1}^{K} a_t \|_2^2 \leq p \sum_{t=1}^{K} \| a_t \|_2^2. \]
vations $\omega_t$.\(^{15}\)

**Splitting observations.** Let $\theta$ be a random $n$-dimensional vector with unknown mean $\mu$ and known covariance matrix, namely, $\sigma^2 I_n$, and let $\chi \sim \mathcal{N}(0, I_n)$ be independent of $\theta$. Finally, let $\kappa > 0$ be a deterministic real.

1) Prove that setting

$$
\theta' = \theta + \sigma \kappa \chi, \quad \theta'' = \theta - \sigma \kappa^{-1} \chi,
$$

we get two random vectors with mean $\mu$ and covariance matrices $\sigma^2 (1 + \kappa^2) I_n$ and $\sigma^2 (1 + 1/\kappa^2) I_n$, respectively, and these vectors are uncorrelated

$$
E\{[\theta' - \mu][\theta'' - \mu]^T\} = 0.
$$

In view of item 1, let us do as follows: given observations $\{\omega_t\}$ and $\{\zeta_t\}$, let us generate i.i.d. sequence $\{\chi_t \sim \mathcal{N}(0, 1), t \geq -d + 1\}$, so that the sequences $\{\xi_t\}$, $\{\eta_t\}$, and $\{\chi_t\}$ are i.i.d. and independent of each other, and let us set

$$
u_t = \omega_t + \sigma x \chi_t, \quad \nu_t = \omega_t - \sigma x \chi_t.
$$

Note that given the sequence $\{\omega_t\}$ of actual observations, sequences $\{\nu_t\}$ and $\{\nu_t\}$ are observable as well, and that the sequence $\{u_t, v_t\}$ is i.i.d.. Moreover, for all $t$,

$$
E\{u_t\} = E\{v_t\} = x_t, \quad E\{|u_t - x_t|^2\} = 2\sigma^2 x, \quad E\{|v_t - x_t|^2\} = 2\sigma^2 x,
$$

and for all $t$ and $s$

$$
E\{|u_t - x_t|[v_s - x_s]\} = 0.
$$

Now, let us put

$$
u = [u_t; u_{t-1}; \ldots; u_{t-d}], \quad \nu = [v_t; v_{t-1}; \ldots; v_{t-d}],
$$

and let

$$\hat{A}_K = \frac{1}{K} \sum_{t=1}^K u_t[v_t]^T.
$$

2) Prove that $\hat{A}_K$ is a good empirical approximation of $A_K$:

$$
E\{\hat{A}_K\} = A_K, \quad E\{\|\hat{A}_K - A_K\|^2\} \leq \frac{12[d + 1]^2[2d + 3]}{K} \left[ M^2 + \sigma_x^2 \right] \sigma^2_x.
$$

the expectation being taken over the distribution of observation noises $\{\xi_t\}$ and auxiliary random sequence $\{\chi_t\}$.

**Conclusion.** We see that as $K \to \infty$, the differences of typical realizations of $\hat{A}_K - A_K$ and $\tilde{a}_K - a_K$ approach 0. It follows that if the sequence $\{x_t\}$ is “rich enough” to ensure that the minimal eigenvalue of $A_K$ for large $K$ stay bounded

\(^{15}\)The model (5.74)–(5.76) is referred to as Errors in Variables model [84] in the statistical literature or Output Error model in the literature on System Identification [169, 215]. In general, statistical inference for such models is difficult – for instance, parameter estimation problem in such models is ill-posed. The estimate we develop in this exercise can be seen as a special application of the general Instrumental Variables methodology [7, 216, 237].
away from 0, the estimate
\[ \hat{\alpha}_K \in \text{Argmin} \| \hat{A}_K \beta - \tilde{a}_K \|_2^2 \]
will converge in probability to the desired vector \( \alpha \), and we can even say something reasonable about the rate of convergence. To account for a priori information \( \alpha \in X \), we can modify the estimate by setting
\[ \hat{\alpha}_K \in \text{Argmin} \| \hat{A}_K \beta - \tilde{a}_K \|_2^2. \]

Note that the assumption that noises affecting observations of \( x_t \)'s and \( y_t \)'s are independent of each other zero mean \textit{Gaussian} random variables with known dispersions is not that important; we could survive the situation where samples \( \{ (\omega_t - x_t, t > -d) \}, \{ \zeta_t - y_t, t \geq 1 \} \) are zero mean i.i.d., and independent of each other, \textit{with a priori known variance of} \( \omega_t - x_t \) (!). Under this and mild additional assumptions (like finiteness of the fourth moments of \( \omega_t - x_t \) and \( \zeta_t - y_t \)), the results obtained would be similar to those for the Gaussian case.

Now goes the concluding part of the exercise:

3) To evaluate numerically the performance of the proposed identification scheme, run experiments as follows:

- Given an even value of \( d \) and \( \rho \in (0, 1] \), select \( d/2 \) complex numbers \( \lambda_i \) at random on the circle \( \{ z \in \mathbb{C} : |z| = \rho \} \), and build real polynomial of degree \( d \) with roots \( \lambda_i, \lambda_i^* \) (* here stands for complex conjugation). Build a finite-difference equation (5.77) with this polynomial as characteristic polynomial.
- Generate i.i.d. \( \mathcal{N}(0, 1) \) "inputs" \( \{ y_t, t = 1, 2, \ldots \} \), select at random initial conditions \( x_{-d+1}, x_{-d+2}, \ldots, x_0 \) for the trajectory \( \{ x_t \} \) of states (5.77), and simulate the trajectory along with observations \( \omega_t \) of \( x_t \) and \( \zeta_t \) of \( y_t \), with \( \sigma_x, \sigma_y \) being experiment's parameters.
- Look at the performance of the estimate \( \hat{\alpha}_K \) on the simulated data.

Exercise 5.4.

[more on generalized linear models] Consider a generalized linear model as follows: we observe i.i.d. random pairs
\[ \omega_k = (y_k, \zeta_k) \in \mathbb{R} \times \mathbb{R}^{\nu \times \mu}, \quad k = 1, \ldots, K, \]
where the conditional given \( \zeta_k \) expectation of the scalar label \( y_k \) is \( \psi(\zeta_k^T z) \), \( z \) being unknown signal underlying observations. What we know is that \( z \) belongs to a given convex compact set \( Z \subset \mathbb{R}^n \). Our goal is to recover \( z \).

Note that while the estimation problem we have just posed looks similar to those treated in Section 5.2, it cannot be straightforwardly handled via techniques developed in that section unless \( \mu = 1 \). Indeed, these techniques in the case of \( \mu > 1 \) require \( \psi \) to be a monotone vector field on \( \mathbb{R}^\mu \), while our \( \psi \) is just a scalar function on \( \mathbb{R}^\mu \). The goal of the exercise is to show that when
\[ \psi(w) = \sum_{q \in Q} c_q w^q = \sum_{q \in Q} c_q w_1^{q_1} \ldots w_\mu^{q_\mu}, \quad [c_q \neq 0, q \in Q \subset \mathbb{Z}_+^\mu] \]
is an algebraic polynomial (which we assume from now on), one can use lifting to
reduce the situation to that considered in Section 5.2.

The construction is straightforward. Let us associate with algebraic monomial with \( \nu \) variables

\[
z^p := z_1^{p_1} z_2^{p_2} \ldots z_\nu^{p_\nu}
\]

a real variable \( x_p \). For example, monomial \( z_1 z_2 \) is associated with \( x_1, x_1, x_0, \ldots, x_0, z_2 \), etc. For \( q \in \mathcal{Q} \), the contribution of the monomial \( c_q w^q \) into \( \psi(\zeta^T z) \) is

\[
c_q [\text{Col}^T_1 [\zeta] z] q_1 [\text{Col}^T_2 [\zeta] z] q_2 \ldots [\text{Col}^T_p [\zeta] z] q_\nu = \sum_{p \in \mathcal{P}_q} h_{pq}(\zeta) z_1^{p_1} z_2^{p_2} \ldots z_\nu^{p_\nu},
\]

where \( \mathcal{P}_q \) is a properly built set of multi-indices \( p = (p_1, \ldots, p_\nu) \), and \( h_{pq}(\zeta) \) are easily computable functions of \( \zeta \). Consequently,

\[
\psi(\zeta^T z) = \sum_{q \in \mathcal{Q}} \sum_{p \in \mathcal{P}_q} h_{pq}(\zeta) z^p = \sum_{p \in \mathcal{P}} H_p(\zeta) z^p,
\]

with properly selected finite set \( \mathcal{P} \) and readily given functions \( H_p(\zeta) \), \( p \in \mathcal{P} \). We can always take, as \( \mathcal{P} \), the set of all \( \nu \)-entry multi-indices with sum of entries not exceeding \( d \), where \( d \) is the total degree of the polynomial \( \psi \). This being said, the structure of \( \psi \) and/or the common structure, if any, of regressors \( \zeta_k \) can enforce some of the functions \( H_p(\cdot) \) to be identically zero. When this is the case, it makes sense to eliminate the corresponding “redundant” multi-index \( p \) from \( \mathcal{P} \).

Next, consider the mapping \( x[z] \) which maps a vector \( z \in \mathbb{R}^\nu \) into a vector with entries \( x_p[z] = z^p \) indexed by \( p \in \mathcal{P} \), and let us associate with our estimation problem its “lifting” with observations

\[
\mathcal{W}_k = (y_k, \eta_k = \{H_p(\zeta_k), p \in \mathcal{P}\}).
\]

I.e., new observations are deterministic transformations of the actual observations; observe that the new observations still are i.i.d., and the conditional, \( \eta_k \) given, expectation of \( y_k \) is nothing but

\[
\sum_{p \in \mathcal{P}} [\eta_k]_p x_p[z].
\]

In our new situation, the “signal” underlying observations is a vector from \( \mathbb{R}^N \), \( N = \text{Card}(\mathcal{P}) \), the regressors are vectors from the same \( \mathbb{R}^N \), and regression is linear – the conditional, given regressor \( \eta_k \), expectation of the label \( y_k \) is linear function \( \eta_k^T x \) of our new signal. Given convex compact localizer \( Z \) for the “true signal” \( z \), we can in many ways find convex compact localizer \( X \) for \( x = x[z] \). Thus, we find ourselves in the simplest possible case of the situation considered in Section 5.2 (one with scalar \( \phi(s) \equiv s \)), and can apply the estimation procedures developed in this section. Note that in the “lifted” problem the SAA estimate \( \hat{x}(\cdot) \) of the

\[\footnote{Note that factors in the monomial are ordered according to the indices of the variables.}
lifted signal \( x = x[z] \) is nothing but the standard Least Squares:

\[
\hat{x}(\mathbf{z}^K) \in \arg\min_{x \in X} \left\{ \frac{1}{2} x^T \left[ \sum_{k=1}^{K} \eta_k \eta_k^T \right] x - \left[ \sum_{k=1}^{K} y_k \eta_k \right]^T x \right\}
\]

(5.80)

Of course, there is no free lunch, and there are some potential complications:

- It may happen that the matrix \( H = E_{\eta \sim Q} \{\eta \eta^T\} \) (\( Q \) is the common distribution of “artificial” regressors \( \eta_k \) induced by the common distribution of the actual regressors \( \zeta_k \)) is not positive definite, which would make it impossible to recover well the signal \( x[z] \) underlying our transformed observations, whatever large be \( K \);
- Even when \( H \) is positive definite, so that \( x[z] \) can be recovered well, provided \( K \) is large, we still need to recover \( z \) from \( x[z] \), that is, to solve a system of polynomial equations, which can be difficult; besides, this system can have more than one solution.
- Even when the above difficulties can be somehow avoided, “lifting” \( z \rightarrow x[z] \) typically increases significantly the number of parameters to be identified, which, in turn, deteriorates “finite time” accuracy bounds.

Note also that when \( H \) is not positive definite, this still is not the end of the world. Indeed, \( H \) is positive semidefinite; assuming that it has a nontrivial kernel \( L \) which we can identify, a realization \( \eta_k \) of our artificial regressor is orthogonal to \( L \) with probability 1, implying that replacing artificial signal \( x \) with its orthogonal projection onto \( L^\perp \), we almost surely keep the value of the objective in (5.80) intact. Thus, we lose nothing when restricting optimization domain in (5.80) to the orthogonal projection of \( X \) onto \( L^\perp \). Since the restriction of \( H \) onto \( L^\perp \) is positive definite, with this approach, for large enough values of \( K \) we will still get good approximation of the projection of \( x[z] \) onto \( L^\perp \). With luck, this approximation, taken together with the fact that the “artificial signal” we are looking for is not an arbitrary vector from \( X \) – it is of the form \( x[z] \) for some \( z \in Z \) – will allow to get a good approximation of \( z \). Here is the first part of the exercise:

1) Carry out the outlined approach in the situation where

- The common distribution \( \Pi \) of regressors \( \zeta_k \) has density w.r.t. Lebesgue measure on \( \mathbb{R}^{r \times p} \) and possesses finite moments of all orders
- \( \psi(w) \) is a quadratic form, either (case A) homogeneous:
  \[
  \psi(w) = w^T S w \quad [S \neq 0],
  \]
  or (case B) inhomogeneous:
  \[
  \psi(w) = w^T S w + s^T w \quad [S \neq 0, s \neq 0]
  \]
- The labels are linked to the regressors and to the true signal \( z \) by the relation
  \[
  y_k = \psi(\eta_k^T z) + \chi_k,
  \]
  where \( \chi_k \sim \mathcal{N}(0,1) \) are mutually independent and independent from the regressors.
Now goes the concluding part of the exercise, where you are supposed to apply the approach we have developed to the situation as follows:

You are given a DC electric circuit comprised of resistors, that is, connected oriented graph with $m$ nodes and $n$ arcs $\gamma_j = (s_j, t_j)$, $1 \leq j \leq n$, where $1 \leq s_j, t_j \leq m$ and $s_j \neq t_j$ for all $j$; arcs $\gamma_j$ are assigned with known to us resistances $R_j > 0$. At instant $k = 1, 2, \ldots, K$, “the nature” specifies “external currents” (charge flows from the “environment” into the circuit) $s_1, \ldots, s_m$ at the nodes; these external currents specify currents in the arcs and voltages at the nodes, and consequently, the power dissipated by the circuit.

Note that the nature cannot be completely free in generating the external currents: their total should be zero. As a result, all what matters is the vector $s = [s_1; \ldots; s_{m-1}]$ of external currents at the first $m-1$ nodes, due to $s_m \equiv -[s_1 + \ldots + s_{m-1}]$. We assume that the mechanism of generating the vector of external currents at instant $k$, let this vector be denoted $s^k \in \mathbb{R}^{m-1}$, is as follows. There are somewhere $m-1$ sources producing currents $z_1, \ldots, z_{m-1}$. At time $k$ the nature selects a one-to-one correspondence $i \mapsto \pi_k(i)$, $i = 1, \ldots, m-1$, between these sources and the first $m-1$ nodes of the circuit, and “forwards” current $z_{\pi_k(i)}$ to node $i$:

$$s^k_i = z_{\pi_k(i)}, \quad 1 \leq i \leq m-1.$$  

For the sake of definiteness, assume that the permutations $\pi_k$ of $1, \ldots, m-1$, $k = 1, \ldots, K$, are i.i.d. drawn from the uniform distribution on the set of $(m-1)!$ permutations of $m-1$ elements.

Assume that at time instants $k = 1, \ldots, K$ we observe the permutations $\pi_k$ and noisy measurements of the power dissipated at this instant by the circuit; given those observations, we want to recover the vector $z$.

Here is your task:

2) Assuming the noises in the dissipated power measurements to be independent of each other and of $\pi_k$ zero mean Gaussian noises with variance $\sigma^2$, apply to the estimation problem in question the approach developed in item 1 of the exercise and run numerical experiments.

Exercise 5.5.

[shift estimation] Consider the situation as follows: given a continuous vector field $f(u) : \mathbb{R}^m \to \mathbb{R}^m$ which is strongly monotone on bounded subsets of $\mathbb{R}^m$ and a convex compact set $S \subset \mathbb{R}^m$, we observe in noise vectors $f(p - s)$, where $p \in \mathbb{R}^m$ is known to us observation point, and $s \in \mathbb{R}^m$ is unknown to us shift known to belong to $S$. Precisely, assume that our observations are

$$y_k = f(p_k - s) + \xi_k, \quad k = 1, \ldots, K,$$

where $p_1, \ldots, p_K$ is a known to us deterministic sequence, and $\xi_1, \ldots, \xi_K$ are independent of each other $\mathcal{N}(0, \gamma^2 I_m)$ observation noises. Our goal is to recover from observations $y_1, \ldots, y_K$ the shift $s$.

1. Pose the problem as a single-observation version of the estimation problem from Section 5.2
2. Assuming $f$ to be strongly monotone, with modulus $\kappa > 0$, on the entire space, what is the error bound for the SAA estimate?

3. Run simulations in the case of $m = 2$, $S = \{ u \in \mathbb{R}^2 : \| u \|_2 \leq 1 \}$ and

$$f(u) = \begin{bmatrix} 2u_1 + \sin(u_1) + 5u_2 \\ 2u_2 - \sin(u_2) - 5u_1 \end{bmatrix}$$

Note: field $f(\cdot)$ is not potential; this is the monotone vector field associated with the strongly convex-concave function

$$\psi(u) = u_1^2 - \cos(u_1) - u_2^2 - \cos(u_2) + 5u_1 u_2,$$

so that $f(u) = [\frac{\partial}{\partial u_1} \phi(u); -\frac{\partial}{\partial u_2} \phi(u)]$. Compare the actual recovery errors with their theoretical upper bounds.

4. Think what can be done when our observations are

$$y_k = f(Ap_k - s) + \xi_k, 1 \leq k \leq K$$

with known $p_k$, independent across $k$ noises $\xi_k \sim \mathcal{N}(0, \gamma^2 I_2)$, and unknown $A$ and $s$ which we want to recover.

5.4 PROOFS

5.4.1 Proof of (5.8)

Let $h \in \mathbb{R}^m$, and let $\omega$ be random vector with independent across $i$ entries $\omega_i \sim \text{Poisson}(\mu_i)$. Taking into account that $\omega_i$ are independent across $i$, we have

$$\mathbb{E} \{ \exp \{ \gamma h^T \omega \} \} = \prod_i \mathbb{E} \{ \gamma \omega_i \} \exp \{ \exp \{ \gamma \mu_i \} - 1 \} \mu_i \}
= \exp \{ \sum_i \exp \{ \gamma \mu_i \} - 1 \} \mu_i \}
$$

whence by the Chebyshev inequality for $\gamma \geq 0$ it holds

$$\text{Prob} \{ h^T \omega > h^T \mu + t \} = \text{Prob} \{ \gamma h^T \omega > \gamma h^T \mu + \gamma t \}
\leq \mathbb{E} \{ \exp \{ \gamma h^T \omega \} \} \exp \{ -\gamma h^T \mu - \gamma t \}
\leq \exp \{ \sum_i \exp \{ \gamma \mu_i \} - 1 \} \mu_i - \gamma h^T \mu - \gamma t \}, \quad (5.81)$$
Now, for $|s| < 3$, one has $e^s \leq 1 + s + \frac{s^2}{2(1 - s/3)}$ (see, e.g., [175]), which combines with (5.81) to imply that

$$0 \leq \gamma < \frac{3}{\|h\|_{\infty}} \Rightarrow \ln \left( \Pr\{h^T \omega > h^T \mu + t\} \right) \leq -\frac{\gamma^2 \sum \mu_i}{2(1 - \gamma \|h\|_{\infty}/3)} - \gamma t.$$ Minimizing the right hand side in this inequality in $\gamma \in [0, \frac{3}{\|h\|_{\infty}})$, we get

$$\Pr\{h^T \omega > h^T \mu + t\} \leq \exp \left\{ -\frac{t^2}{2\|h\|_{\infty}^2 + \frac{t}{3}} \right\}.$$ This inequality combines with the same inequality applied to $-h$ in the role of $h$ to imply (5.8).

### 5.4.2 Proof of Lemma 5.6

(i): When $\pi(\text{Col}_\ell[H]) \leq 1$ for all $\ell$ and $\lambda \geq 0$, denoting by $[h]^2$ the vector comprised of squares of the entries in $h$, we have

$$\phi(dg(H \text{Diag}([\lambda] H^T)) = \phi(\sum \ell \lambda \ell [\text{Col}_\ell[H]]^2) \leq \sum \ell \lambda \ell \phi([\text{Col}_\ell[H]]^2) = \sum \ell \lambda \ell \pi^2([\text{Col}_\ell[H]]^2) \leq \sum \ell \lambda \ell,$$

implying that $(H^T \text{Diag}([\lambda] H^T, \propto \sum \ell \lambda \ell)$ belongs to $H$.

(ii): Let $\Theta, \mu, Q, V$ be as stated in (ii); there is nothing to prove when $\mu = 0$, thus assume that $\mu > 0$. Let $d = dg(\Theta)$, so that

$$d_i = \sum_j Q^2_{ij} \& \& \phi(d) \leq \mu$$

(5.82)

(the second relation is due to $(\Theta, \mu) \in H$). (5.30) is evident. We have

$$[H_\chi]_{ij} = \sqrt{m/\mu}[G_\chi]_{ij}, G_\chi = Q \text{Diag}\{\chi\}V = \left[ \sum_{k=1}^m Q_{ik} \chi_k V_{kj} \right]_{i,j}.$$ We claim that for every $i$ it holds

$$\forall \gamma > 0 : \Pr\{ [G_\chi]_{ij}^2 > 3\gamma d_i/m \} \leq \sqrt{3} \exp(-\gamma/2).$$

(5.83)

Indeed, let us fix $i$. There is nothing to prove when $d_i = 0$, since in this case $Q_{ij} = 0$ for all $j$ and therefore $[G_\chi]_{ij} = 0$. When $d_i > 0$, by homogeneity in $Q$ it suffices to verify (5.83) when $d_i/m = 1/3$. Assuming that this is the case, let $\eta \sim N(0, 1)$ be independent of $\chi$. We have

$$\mathbb{E}_\eta\{ \exp\{\eta [G_\chi]_{ij}^2\} \} = \mathbb{E}_\eta\{ \prod_k \cosh(\eta Q_{ik} V_{kj}) \} \leq \mathbb{E}_\eta\{ \prod_k \exp\{\frac{1}{2} \eta^2 Q^2_{ik} V^2_{kj}\} \} = \mathbb{E}_\eta\{ \exp\{\frac{1}{2} \eta^2 d_i/m \} \} = \mathbb{E}_\eta\{ \exp\{\eta^2/3\} \} = \sqrt{3},$$

$$\mathbb{E}_\eta\{ \exp\{3\eta^2 d_i/m \} \} \leq \sqrt{3} \exp(-\gamma/2).$$
and
\[
\mathbf{E}_\chi \{ \mathbf{E}_\eta \{ \exp \{ \eta |G_\chi|_{ij} \} \} \} = \mathbf{E}_\chi \{ \exp \{ \frac{1}{2} |G_\chi|_{ij}^2 \} \},
\]
implying that
\[
\mathbf{E}_\chi \{ \exp \{ \frac{1}{2} |G_\chi|_{ij}^2 \} \} \leq \sqrt{3}.
\]
Therefore in the case of \(d_i/m = 1/3\) for all \(s > 0\) it holds
\[
\text{Prob}\{\chi : |G_\chi|_{ij}^2 > s\} \leq \sqrt{3} \exp(-s/2),
\]
and (5.83) follows. Recalling the relation between \(H\) and \(G\), we get from (5.29) that
\[
\forall \gamma > 0 : \text{Prob}\{\chi : |H_\chi|_{ij}^2 \leq 3\gamma d_i/\mu \} \leq \sqrt{3} \exp(-\gamma/2).
\]
By the latter inequality, with \(\approx\) given by (5.29) the probability of the event
\[
\forall i, j : |H_\chi|_{ij}^2 \leq \frac{d_i}{\mu}
\]
is at least 1/2. Let this event take place; in this case we have \(|\text{Col}_i[H_\chi]|^2 \leq \approx d_i/\mu\), whence, by definition of the norm \(\pi(\cdot)\), \(\pi^2(\text{Col}_i[H_\chi]) \leq \approx \phi(d)/\mu \leq 1\) (see the second relation in (5.82)). Thus, the probability of the event (5.31) is at least 1/2. \(\square\)

### 5.4.3 Verification of (5.44)

Given \(s \in [2, \infty]\) and setting \(s = s/2\), \(s_\ast = \frac{s}{2 - \frac{\tau}{s}}\), \(\bar{s}_\ast = \frac{s}{2 - \frac{\tau}{s}}\), we want to prove that
\[
\{(V, \tau) \in S^N_+ \times R_+: \exists (W \in S^N, w \in R^N_+) : V \preceq W + \text{Diag}\{w\} \& \|W\|_{s_\ast} + \|w\|_{\bar{s}_\ast} \leq \tau\}
= \{(V, \tau) \in S^N_+ \times R_+: \exists w \in R^N_+: V \preceq \text{Diag}\{w\}, \|w\|_{\bar{s}_\ast} \leq \tau\}.
\]

To this end it clearly suffices to check that whenever \(W \in S^N\), there exists \(w \in R^N\) satisfying
\[
W \preceq \text{Diag}\{w\}, \|w\|_{\bar{s}_\ast} \leq \|W\|_{s_\ast}.
\]
The latter is equivalent to saying that for any \(W \in S^N\) such that \(\|W\|_{s_\ast} \leq 1\), the conic optimization problem
\[
\text{Opt} = \min_{t, w} \{ t : t \geq \|w\|_{\bar{s}_\ast}, \text{Diag}\{w\} \succeq W \}
\]
is solvable (which is evident) with optimal value \(\leq 1\). To see that the latter indeed is the case, note that the problem clearly is strictly feasible, whence its optimal value is the same as the optimal value in the conic problem
\[
\text{Opt} = \max_P \{ \text{Tr}(PW) : P \succeq 0, \|\text{dg}\{P\}\|_{s_\ast/(\bar{s}_\ast - 1)} \leq 1 \}
= [P_{11}; P_{22}; \ldots; P_{NN}]
\]
dual to (5.84). Since \(\text{Tr}(PW) \leq \|P\|_{s_\ast/(s_\ast - 1)}\|W\|_{s_\ast} \leq \|P\|_{s_\ast/(s_\ast - 1)}\), recalling what \(s_\ast\) and \(\bar{s}_\ast\) are, our task boils down to verifying that when a matrix \(P \succeq 0\) satisfies \(|\text{dg}\{P\}|_{s/2} \leq 1\), one also has \(|P|_s \leq 1\). This is immediate: since \(P\) is positive semidefinite, we have \(|P_{ij}| \leq P_{ii}^{1/2} P_{jj}^{1/2}\), whence, assuming \(s < \infty\),
\[
\|P\|_s^s = \sum_{i,j} |P_{ij}|^s \leq \sum_{i,j} P_{ii}^{s/2} P_{jj}^{s/2} = \left( \sum_{i} P_{ii}^{s/2} \right)^2 \leq 1.
\]
When \( s = \infty \), the same argument leads to
\[
\|P\|_\infty = \max_{i,j} |P_{ij}| = \max_i |P_{ii}| = \|\text{diag}(P)\|_\infty.
\]

### 5.4.4 Proof of Proposition 5.10

1°. Let us consider the optimization problem (4.42) (where one should set \( Q = \sigma^2 I_n \)) which under the circumstances is responsible for building a nearly optimal linear estimate of \( Bx \) yielded by Proposition 4.14, namely,

\[
\text{Opt}_* = \min_{\Theta, H, A, T', T''} \left\{ \frac{1}{2} \|A\|^2 + \lambda \text{Tr} (\Theta) : \begin{array}{l}
\lambda \geq 0, k \leq K, T' = \{ T'_{\ell} \geq 0, \ell \leq L \}, \\
\frac{1}{2} B - H A A^T_{T'} + \sum_k \mathcal{R}_k A_k \end{array} \right\}
\]

\[
\frac{1}{2} B - H A A^T_{T'} + \sum_k \mathcal{R}_k A_k \succeq 0, \quad 0 \leq \ell \leq L.
\]

Let us show that the optimal value \( \text{Opt} \) of (5.45) satisfies
\[
\text{Opt} \leq 2\kappa \text{Opt}_* = 2\sqrt{2 \ln(2m/\epsilon)} \text{Opt}_*.
\]

To this end, observe that the matrices

\[
Q := \begin{bmatrix}
U & \frac{1}{2} B \\
\frac{1}{2} B^T & A^T A + \sum_k \mathcal{R}_k A_k
\end{bmatrix}
\]

and

\[
\begin{bmatrix}
M^T U M & \frac{1}{2} M^T B \\
\frac{1}{2} B^T M & A^T A + \sum_k \mathcal{R}_k A_k
\end{bmatrix} = \begin{bmatrix}
M^T & I_n
\end{bmatrix} Q \begin{bmatrix}
M \\
I_n
\end{bmatrix}
\]

simultaneously are/are not positive semidefinite due to the fact that the image space of \( M \) contains the full-dimensional set \( \mathcal{E}_* \) and thus is the entire \( \mathbb{R}^n \), so that the image space of \( \begin{bmatrix}
M \\
I_n
\end{bmatrix} \) is the entire \( \mathbb{R}^n \times \mathbb{R}^n \). Therefore

\[
\text{Opt} = \min_{\Theta, U, A, T} \left\{ 2 \left[ \frac{1}{2} \lambda \text{Tr}(\Theta) + \lambda \text{Tr}(\Theta) + \sigma^2 \kappa^2 \text{Tr}(\Theta) \right] : \begin{array}{l}
\Theta \geq 0, U \geq 0, A = \{ A_k \geq 0, k \leq K \}, T = \{ T_{\ell} \geq 0, \ell \leq L \}, \\
\frac{1}{2} B - H A A^T_{T'} + \sum_k \mathcal{R}_k A_k \end{array} \right\}
\]

Further, note that if a collection \( \Theta, U, \{ A_k \}, \{ T_{\ell} \} \) is a feasible solution to the latter problem and \( \theta > 0 \), the scaled collection \( \theta \Theta, \theta^{-1} U, \{ \theta A_k \}, \{ \theta^{-1} T_{\ell} \} \) is also a feasible
solution. When optimizing with respect to the scaling, we get

\[
\text{Opt} = \inf_{\Theta, U, A, T} \left\{ 4\sqrt{\phi_R(\lambda[\mathbf{T}])} \left[ \phi_R(\lambda[A]) + \sigma^2 \text{Tr}(\Theta) \right] : \right.
\]

\[
\begin{align*}
\Theta &\succeq 0, U \succeq 0, A = \{A_k \succeq 0, k \leq K\}, 
\mathbf{Y} = \{\mathbf{Y}_\ell \succeq 0, \ell \leq L\}, 
\begin{bmatrix}
M^T U M & \frac{1}{2} M^T B \\
\frac{1}{2} B^T M & A^T \Theta A + \sum_k \mathcal{R}_k'[\Lambda_k]
\end{bmatrix} &\succeq 0, 
M^T U M &\leq \sum_\ell S_\ell'[\mathbf{Y}_\ell] 
\end{align*}
\]

\[
\leq 2\kappa \text{Opt}_+, \quad (5.87)
\]

where (note that $\kappa > 1$)

\[
\text{Opt}_+ = \inf_{\Theta, U, A, T} \left\{ 2\sqrt{\phi_R(\lambda[\mathbf{T}])} \left[ \phi_R(\lambda[A]) + \sigma^2 \text{Tr}(\Theta) \right] : \right.
\]

\[
\begin{align*}
\Theta &\succeq 0, U \succeq 0, A = \{A_k \succeq 0, k \leq K\}, 
\mathbf{Y} = \{\mathbf{Y}_\ell \succeq 0, \ell \leq L\}, 
\begin{bmatrix}
M^T U M & \frac{1}{2} M^T B \\
\frac{1}{2} B^T M & A^T \Theta A + \sum_k \mathcal{R}_k'[\Lambda_k]
\end{bmatrix} &\succeq 0, 
\end{align*}
\]

(5.88)

On the other hand, when strengthening the constraint $\Lambda_k \succeq 0$ of (5.85) to $\Lambda_k > 0$, we still have

\[
\text{Opt}_* = \inf_{\Theta, H, A, T', T''} \left\{ \phi_T(\lambda[A]) + \phi_R(\lambda'[\mathbf{Y}']) + \phi_R(\lambda[\mathbf{Y}'']) + \sigma^2 \text{Tr}(\Theta) : \right.
\]

\[
\begin{align*}
\Lambda &\succeq 0, A = \{A_k > 0, k \leq K\}, 
\mathbf{Y} = \{\mathbf{Y}_\ell \succeq 0, \ell \leq L\}, 
\mathbf{Y}' = \{\mathbf{Y}'_\ell \succeq 0, \ell \leq L\}, 
\mathbf{Y}'' = \{\mathbf{Y}''_\ell \succeq 0, \ell \leq L\}, 
\begin{bmatrix}
\sum_\ell S_\ell'[\mathbf{Y}'_\ell] & \frac{1}{2} M^T H^T \\
\frac{1}{2} [H^T A]^T M & \frac{1}{2} [B - H^T A]^T M \\
\frac{1}{2} A^T H M & A^T \Theta A + \sum_k \mathcal{R}_k'[\Lambda_k]
\end{bmatrix} &\succeq 0, 
\end{align*}
\]

(5.89)

Now let $\Theta, H, A, \mathbf{Y}', \mathbf{Y}''$ be a feasible solution to the latter problem. By the second semidefinite constraint in (5.89) we have

\[
\begin{bmatrix}
\sum_\ell S_\ell'[\mathbf{Y}'_\ell] & \frac{1}{2} M^T H^T A \\
\frac{1}{2} A^T H M & A^T \Theta A + \sum_k \mathcal{R}_k'[\Lambda_k]
\end{bmatrix} = \begin{bmatrix} I & A \end{bmatrix} \begin{bmatrix}
\sum_\ell S_\ell'[\mathbf{Y}'_\ell] & \frac{1}{2} M^T H^T \\
\frac{1}{2} H M & \Theta
\end{bmatrix} \begin{bmatrix} I & A \end{bmatrix} \succeq 0
\]

which combines with the first semidefinite constraint in (5.89) to imply that

\[
\begin{bmatrix}
\sum_\ell S_\ell'[\mathbf{Y}'_\ell + \mathbf{Y}''_\ell] & \frac{1}{2} M^T B \\
\frac{1}{2} B^T M & A^T \Theta A + \sum_k \mathcal{R}_k'[\Lambda_k]
\end{bmatrix} \succeq 0.
\]

Next, by the Schur Complement Lemma (which is applicable due to

\[
A^T \Theta A + \sum_k \mathcal{R}_k'[\Lambda_k] \succeq \sum_k \mathcal{R}_k'[\Lambda_k] > 0,
\]

where the concluding $\succ$ is due to Lemma 4.44 combined with $\Lambda_k > 0$), this relation implies that for

\[
\mathbf{Y}_\ell = \mathbf{Y}'_\ell + \mathbf{Y}''_\ell,
\]
we have
\[ \sum_{\ell} S^*_\ell[\Upsilon_\ell] \geq M^T \left[ \frac{1}{2} B[A^T \Theta A + \sum_k R^*_k[\Lambda_k]]^{-1} B^T \right] M. \]

Using the Schur Complement Lemma again, for the just defined \( U \geq 0 \) we obtain
\[ \begin{bmatrix} M^T U M & \frac{1}{2} M^T B \\ \frac{1}{2} B^T M & A^T \Theta A + \sum_k R^*_k[\Lambda_k] \end{bmatrix} \geq 0, \]
and in addition, by the definition of \( U \),
\[ M^T U M \preceq \sum_{\ell} S^*_\ell[\Upsilon_\ell]. \]

We conclude that
\[ (\Theta, U, \Lambda, \Upsilon := \{ \Upsilon_\ell = \Upsilon'_\ell + \Upsilon''_\ell, \ell \leq L \}) \]
is a feasible solution to optimization problem (5.88) specifying \( \text{Opt}_+ \). The value of the objective of the latter problem at this feasible solution is
\[ 2 \sqrt{\phi_R(\lambda[\Upsilon']) + \phi_R(\lambda[\Upsilon'']) + \phi_T(\lambda[\Lambda]) + \sigma^2 \text{Tr}(\Theta)} \]
\[ \leq \phi_R(\lambda[\Upsilon']) + \phi_R(\lambda[\Upsilon'']) + \phi_T(\lambda[\Lambda]) + \sigma^2 \text{Tr}(\Theta), \]
the concluding quantity in the chain being the value of the objective of problem (5.89) at the feasible solution \( \Theta, H, \Lambda, \Upsilon', \Upsilon'' \) to this problem. Since the resulting inequality holds true for every feasible solution to (5.89), we conclude that \( \text{Opt}_+ \leq \text{Opt}_\ast \), and we arrive at (5.86) due to (5.87).

2°. Now, from Proposition 4.16 we conclude that \( \text{Opt}_\ast \) is within a logarithmic factor of the minimax optimal \((\epsilon, \| \cdot \|)-\text{risk}\) corresponding to the case of Gaussian noise \( \xi_x \sim \mathcal{N}(0, \sigma^2 I_m) \) for all \( x \):
\[ \text{Opt}_\ast \leq \theta_* \text{RiskOpt}_{1/8}, \]
where
\[ \theta_* = 8 \sqrt{(2 \ln F + 10 \ln 2)(2 \ln D + 10 \ln 2)}, \quad F = \sum_{\ell} f_\ell, \quad D = \sum_k d_k. \]
Since the minimax optimal \((\epsilon, \| \cdot \|)-\text{risk}\) clearly only grows when \( \epsilon \) decreases, we conclude that for \( \epsilon \leq 1/8 \) a feasible near optimal solution to (5.45) is minimax optimal within the factor \( 2\theta^\ast \kappa \). \( \square \)
Appendix A

Executive Summary on Efficient Solvability of Convex Optimization Problems

Convex Programming is a “solvable case” in Optimization: under mild computability and boundedness assumptions, a globally optimal solution to a convex optimization problem can be “approximated to whatever high accuracy in reasonable time.” The goal of what follows is to provide a reader with “sufficient for all practical purposes” understanding of what these words mean.\(^1\)

In the sequel we are interested in computational tractability of a convex optimization problem in the form
\[
\text{Opt} = \min_{x \in B^n_R} \{ f_0(x) : f_i(x) \leq 0, \ 1 \leq i \leq m \}, \quad B^n_R = \{ x \in \mathbb{R}^n : \|x\|_2 \leq R \}, \quad (A.1)
\]
We always assume that \(f_i(\cdot), \ 0 \leq i \leq m\), are convex and continuous real-valued functions on \(B^n_R\), and what we are looking for, is an \(\epsilon\)-accurate solution to the problem, that is, a point \(x_\epsilon \in B^n_R\) such that
\[
\begin{align*}
  f_i(x_\epsilon) &\leq \epsilon, \ i = 1, \ldots, m \quad [\epsilon\text{-feasibility}] \\
  f_0(x_\epsilon) &\leq \text{Opt} + \epsilon \quad [\epsilon\text{-optimality}]
\end{align*}
\]
where \(\epsilon > 0\) is a given tolerance. Note that when (A.1) is infeasible, i.e., \(\text{Opt} = +\infty\), every point of \(B^n_R\) is an \(\epsilon\)-optimal (but not necessary \(\epsilon\)-feasible, and thus not necessary \(\epsilon\)-accurate) solution.

We intend to provide two versions of what “efficient solvability” means: “practical” and “scientific.”

“Practical” version of efficient solvability: for most practical purposes, efficient solvability means that we can feed (A.1) to \texttt{cvx}, that is, to rewrite the problem in the form
\[
\text{Opt} = \min_{x,u} \{ c^T[x;u] : A(x,u) \preceq 0 \}, \quad (A.2)
\]
where \(A(x,u)\) is a symmetric matrix which is affine in \([x;u]\).

“Scientific” version of efficient solvability. Let us start with the following basic fact:

(!) Assume that when solving (A.1), we have at our disposal \(R\), a real \(V\) such that
\[
\max_{x \in B^n_R} |f_i(x)| \leq V/2, \quad (A.3)
\]
and have access to \textit{First Order oracle} – a “black box” which, given on input a query point \(x \in B^n_R\) and tolerance \(\delta > 0\), returns \(\delta\)-subgradients of \(f_i\) at

\(^1\)For rigorous treatment of this subject in the context of continuous optimization (this is what we deal with in our book) the reader is referred to [15, Chapter 5].
x, that is, affine functions \( g_{i,x}(\cdot) \), \( 0 \leq i \leq m \), such that
\[
g_{i,x}(y) \leq f_i(y) \quad \forall y \in B^n_R & g_{i,x}(x) \geq f_i(x) - \delta, \quad 0 \leq i \leq m.
\]

Then for every \( \epsilon \in (0, V) \), an \( \epsilon \)-accurate solution to (A.1), or a correct claim that the problem is infeasible, can be found by appropriate algorithm (e.g., the Ellipsoid method) at the cost of at most
\[
N(\epsilon) = \lfloor 2n^2 \ln(2V/\epsilon) \rfloor + 1 \tag{A.4}
\]
subsequent steps, with

- at most one call to First Order oracle per step, the input at the \( t \)-th step being \( (x_t, \epsilon/2) \), with \( x_1 = 0 \) and recursively computed \( x_2, \ldots, x_{N(\epsilon)} \);
- at most \( O(1)n(m+n) \) operations of precise real arithmetic per step needed to update \( x_t \) and the output of the First Order oracle (if the latter was invoked at step \( t \)) into \( x_{t+1} \).

The remaining computational effort when executing the algorithm is just \( O(N(\epsilon)n) \) operations of precise real arithmetic needed to convert the trajectory \( x_1, \ldots, x_{N(\epsilon)} \) and the outputs of the First Order oracle into the result of the computation – either an \( \epsilon \)-accurate solution to (A.1), or a correct infeasibility claim.

The consequences are as follows. Consider a family \( \mathcal{F} \) of convex functions of a given structure, so that every function \( f \) in the family can be identified by a finite-dimensional data vector \( \text{Data}(f) \). For the sake of simplicity, assume that all functions from the family are real-valued (extensions to partially defined functions are straightforward). Examples include (but, of course, not reduce to)

1. affine functions of \( n = 1, 2, \ldots \) variables,
2. convex quadratic functions of \( n = 1, 2, \ldots \) variables,
3. (logarithms of) posynomials – functions of \( n = 1, 2, \ldots \) variables of the form \( \ln(\sum_{i=1}^m \exp\{\phi_i(x)\}) \), each with its own \( m \), with affine \( \phi_i \),
4. functions of the form \( \lambda_{\max}(A_0 + \sum_{i=1}^n x_j A_j) \), where \( A_j, 1 \leq j \leq m \), are \( m \times m \) symmetric matrices, \( \lambda_{\max}(\cdot) \) is the maximal eigenvalue of a symmetric matrix, and \( m, n \) can be arbitrary positive integers

(in all these examples, it is self-evident what the data is). For \( f \in \mathcal{F} \), denoting by \( n(f) \) the dimension of the argument of \( f \), let us call the quantity
\[
\text{Size}(f) = \max\{n(f), \dim \text{Data}(f)\}
\]
the size of \( f \). Let us say that family \( \mathcal{F} \)

- is with polynomial growth, if for all \( f \in \mathcal{F} \) and \( R > 0 \) it holds
\[
V(f, R) := \max_{x \in B^n_R} f(x) - \min_{x \in B^n_R} f(x) \\
\leq \chi\text{Size}(f) + R + \|\text{Data}(f)\|_\infty \chi\text{Size}(f),
\]

here and in what follows \( \chi \)'s stand for positive constants, perhaps different in dif-
ferent places, depending solely on \( F \);

- is polynomially computable, if there exists a code for a Real Arithmetic computer\(^2\) with the following property: whenever \( f \in F \), \( R > 0 \), and \( \delta > 0 \), executing the code on input comprised of Data(\( f \)) augmented by \( \delta \), \( R \), and a query vector \( x \in \mathbb{R}^n(f) \) with \( \|x\|_2 \leq R \), the computer after finitely many operations outputs the coefficients of affine function \( g_{f,x}(\cdot) \) which is a \( \delta \)-subgradient of \( f \), taken at \( x \), on \( B^n_R(f) \):

\[
g_{f,x}(y) \leq f(y) \quad \forall y \in B^n_R(f) \quad \& \quad f(x) \leq g_{f,x}(x) + \delta,
\]

and the number \( N \) of arithmetic operations in this computation is upper-bounded by a polynomial in Size(\( f \)) and “the required number of accuracy digits”

\[
\text{Digits}(f, R, \delta) = \log \left( \frac{\text{Size}(f) + \|\text{Data}(f)\|_\infty + R + \delta^2}{\delta} \right),
\]

that is,

\[
N \leq \chi[\text{Size}(f) + \text{Digits}(f, R, \delta)]^X.
\]

Observe that typical families of convex functions, like those in the above examples are both with polynomial growth and polynomially computable.

In the main body of this book, the words “a convex function \( f \) is efficiently computable” mean that \( f \) belong to a polynomially computable family (what is this family, it always is clear from the context). Similarly, the words “a closed convex set \( X \) is computationally tractable” mean that the convex function \( f(x) = \min_{y \in X} \|y - x\|_2 \) is efficiently computable.

The role of the introduced notions in our context is as follows. Consider problem (A.1) and assume that the functions \( f_i, i = 0, 1, ..., m \) participating in the problem are taken from a family \( F \) which is polynomially computable and with polynomial growth (as is the case when (A.1) is a linear, or second order conic, or semidefinite program). In this situation a particular instance \( P \) of (A.1) is fully specified by its data vector Data(\( P \)) obtained by augmenting the “sizes” \( n, m, R \) of the instance by the concatenation of the data vectors of \( f_0, f_1, ..., f_m \). Similarly to the above, let us define the size of \( P \) as

\[
\text{Size}(P) = \max[n, m, \dim \text{Data}(P)],
\]

so that \( \text{Size}(P) \geq \text{Size}(f_i) \) for all \( i, 0 \leq i \leq m \). Given Data(\( P \)) and \( R \) and invoking the fact that \( F \) is with polynomial growth, we can easily compute \( V \) satisfying (A.3) and such that

\[
V = V(P, R) \leq \chi[\text{Size}(P) + R + \|\text{Data}(P)\|_\infty]^{(\chi\text{Size}(P))} \quad (\text{A.5})
\]

Similarly to the above, we set

\[
\text{Digits}(P, R, \delta) = \log \left( \frac{\text{Size}(P) + \|\text{Data}(P)\|_\infty + R + \delta^2}{\delta} \right),
\]

\(^2\)an idealized computer capable to store reals and to carry out operations of precise real arithmetic – four arithmetic operations, comparison, and computing elementary univariate functions, like \( \sin(s), \sqrt{s}, \) etc.
so that $\text{Digits}(P,R,\delta) \geq \text{Digits}(f_i,R,\delta)$, $0 \leq i \leq m$. Invoking polynomial computability of $\mathcal{F}$, we can implement First Order oracle for problems $P$ of the form (A.1) with $f_i \in \mathcal{F}$ on the Real Arithmetic Computer in such a way that executing the resulting code on input comprised by the data vector $\text{Data}(P)$ augmented by $\delta > 0$, $R$, and a query vector $x \in \mathbf{B}_R^n$, the code will produce $\delta$-subgradients, taken at $x$, of $f_i$, $0 \leq i \leq m$, with the total number $M = M(P,R,\delta)$ of real arithmetic operations in the course of computation upper-bounded by a polynomial in $\text{Size}(P)$ and $\text{Digits}(P,R,\delta)$:

$$M(P,R,\delta) \leq \chi[\text{Size}(P) + \text{Digits}(P,R,\delta)]^\chi.$$ (A.6)

Finally, given, along with $\text{Data}(P)$ and $R$, a desired accuracy $\epsilon > 0$ and assuming w.l.o.g. that $\epsilon < V = V(P,R)$,\(^3\) we can use the above First Order oracle (with $\delta$ set to $\epsilon/2$) in (!) in order to find an $\epsilon$-accurate solution to problem $P$ (or conclude correctly that the problem is infeasible). The number $N$ of steps in this computation, in view of (A.4) and (A.5), is upper-bounded by a polynomial in $\text{Size}(P)$ and $\text{Digits}(P,R,\epsilon)$:

$$N \leq O(1)[\text{Size}(P) + \text{Digits}(P,R,\epsilon)]^\chi,$$

with computational expenses per step stemming from mimicking First Order oracle upper-bounded by a polynomial in $\text{Size}(P)$, $\text{Digits}(P,R,\epsilon)$ (by (A.6)). By (!), the overall “computational overhead” needed to process oracle’s outputs and to generate the result is bounded by another polynomial of the same type. The bottom line is that

When $\mathcal{F}$ is a polynomially computable family of convex functions with polynomial growth and the objective and the constraints $f_i \in \mathcal{F}$, $i \leq m$, in (A.1) belong to $\mathcal{F}$, the overall number of arithmetic operations needed to find an $\epsilon$-approximate solution to (A.1) (or to conclude correctly that (A.1) is infeasible) is, for every $\epsilon > 0$, upper-bounded by a polynomial, depending solely on $\mathcal{F}$, in the size $\text{Size}(P)$ of the instance and the number $\text{Digits}(P,R,\epsilon)$ of accuracy digits in the desired solution.

This is enough general for all our purposes “scientific translation” of the informal claim “an explicit convex problem with computationally tractable objective and constraints is efficiently solvable.”

\(^3\)this indeed is w.l.o.g., since, say, the origin is $V$-accurate solution to $P$
Bibliography


[63] D. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposi-


BIBLIOGRAPHY


[223] A. G. Tartakovsky and V. V. Veeravalli. Asymptotically optimal quickest change detection in distributed sensor systems. *Sequential Analysis,*


Index

O(1), xxii
Diag, xxi
Erfc, 16
ErfcInv, 17
Risk, 41
E\{, \}, E_{\xi-P}\{, \}, xxii
Q_{s}(s,\kappa)-condition, 12
   links with RIP, 24
   tractability when q = \infty, 22
   verifiable sufficient conditions
   for, 20
R^n, R^{m\times n}, xxii
R_+, R^n_+, xxii
S^n, xxii
S^n_+, xxii
A^*, xxii
N(\mu, \Theta), xxii
R_k[\cdot], R^*_k[\cdot], S_l[\cdot], S^*_l[\cdot], ..., 275
dg, xxii
\ell_1 minimization, see Compressed Sensing
Poisson(\cdot), xxii
Uniform(\cdot), xxii
\int_\Omega f(\xi)\Pi(d\xi), xxii
\lambda[\cdot], 276
\geq, \succ, \preceq, \prec, xxii
\xi \sim P, \xi \sim p(\cdot), xxii
s-goodness, 9
\parallel \cdot \parallel_p, xxii
\parallel \cdot \parallel_{2,2}, xxii
Bisection estimate, 198
   near-optimality of, 201
closeness relation, 59
Compressed Sensing, 3–6
   via \ell_1 minimization, 6–17
   imperfect, 11
   validity of, 8
   verifiable sufficient validity conditions, 17–26
   verifiable sufficient validity conditions, limits of performance, 24
   via penalized \ell_1 recovery, 13
   via regular \ell_1 recovery, 13
conditional quantize, 197
cone
dual, 262
Lorentz, 262
regular, 262
semidefinite, 262
conic
   problem, 263
dual of, 263
programming, 262, 265
Conic Duality Theorem, 264
conic hull, 263
contrast matrix, see nullspace property quantification
detector, 64
   affine, 123
   in simple observation schemes, 83
   quadratic, 139
   risks of, 64
   structural properties, 65
ellitope, 265–266
   calculus of, 299–302, 430
estimation
   of \\ N-convex functions, 192–210
   of linear form, 185, 210–221
   from repeated observations, 215–216
   of sub-Gaussianity parameters, 216–221
   of sub-Gaussianity parameters, direct product case, 218
of quadratic form, 221–231
Gaussian case, 221–227
Gaussian case, consistency, 226
Gaussian case, construction, 223
sub-Gaussian case, 227, 231
sub-Gaussian case, construction, 227

family of distributions
regular/simple, 125–132
calculus of, 127
examples of, 125
spherical, 53
cap of, 53

function
$N$-convex, 196
examples of, 197

Gaussian mixtures, 53
generalized linear model, 416
GLM, see generalized linear model

Hellinger affinity, 83

Hypothesis Testing
change detection
via quadratic lifting, 150–157
of multiple hypotheses, 58–64
in simple observation schemes, 88–105
up to closeness, 59, 91
via Euclidean separation, 61–64
via repeated observations, 95
of unions, 88
problem’s setting, 41
sequential, 105–113
test, 42
detector-based, 64
detector-based, limits of performance, 70
detector-based, via repeated observations, 66
deterministic, 42
partial risks of, 45
randomized, 42
simple, 42
total risk of, 45
two-point lower risk bound, 46
via affine detectors, 132–139
via Euclidean separation, 49–58
and repeated observations, 55
majority test, 56
multiple hypotheses case, 61–64
pairwise, 50
via quadratic lifting, 139
Gaussian case, 140–145
sub-Gaussian case, 145–150
via repeated observations, 42

inequality
Cramer-Rao, 349

lemma
on Schur Complement, see Schur Complement Lemma
LMI, xxii
logistic regression, 415–416

matrices
notation, xxi
sensing, 1
MD, see Measurement Design
Measurement Design, 113–123
simple case
discrete o.s., 118
Gaussian o.s., 122
Poisson o.s., 121
Mutual Incoherence, 23

norm
conjugate, 279
Shatten, 305
Wasserstein, 340
Nullspace property, 9, 10
quantification of, 11

o.s., see observation scheme
observation scheme
discrete, 77, 85
Gaussian, 74, 84
Poisson, 75, 85
simple, 72–87
$K$-th power of, 86
definition of, 73
direct product of, 78

PET, see Positron Emission Tomography
Poisson Imaging, 75
polyhedral estimate, 386–415
Positron Emission Tomography, 75
Rademacher random vector, 364
regular data, 124
repeated observations
quasi-stationary, 44
semi-stationary, 43
stationary, 42
Restricted Isometry Property, 20
RIP, see Restricted Isometry Property
risk
\[ \text{Risk}(\mathcal{T}|H_1,\ldots,H_L), \]
\[ \text{Risk}\text{Opt}_{\Pi,\|\cdot\|}[\hat{\mathcal{X}}], \]
\[ \text{Risk}(\hat{\mathcal{Z}}(\cdot)|\mathcal{X}), \]
\[ \text{Risk}^\times, \]
\[ \text{Risk}^{\mathcal{C}}(\mathcal{T}|H_1,\ldots,H_L), \]
\[ \text{Risk}^{\mathcal{C}}_{\mathcal{H}}(\mathcal{T}|H_1,\ldots,H_L), \]
\[ \text{Risk}^{\mathcal{C}}(\hat{g}(\cdot)|G,\mathcal{X},v,A,\mathcal{H},\mathcal{M},\Phi), \]
\[ \text{Risk}_+\phi[\mathcal{P}], \text{Risk}[\phi]\mathcal{P}_1,\mathcal{P}_2], \]
\[ \text{Risk}_{\text{tot}}(\mathcal{T}|H_1,\ldots,H_L), \]
\[ \text{Risk}_{\mathcal{H}}[\hat{x},\mathcal{X}], \]
\[ \text{Risk}_\mathcal{H}[\hat{\mathcal{X}}]\mathcal{X}], \]
\[ \mathcal{C}^{-}, \]
\[ \mathcal{H}^{-}, \]
\[ \epsilon^{-}, \]
\[ \text{Risk}_{\Pi,\|\cdot\|}[\hat{\mathcal{Z}}], \]
\[ \text{Risk}_{\mathcal{H},\|\cdot\|}[\hat{\mathcal{Z}}], \]
in Hypothesis Testing
partial, 45
total, 45
up to closeness, 59
minimax, 266
\[ \epsilon^{-}, \]
of detector, 64
of simple test, 45

SA, see Stochastic Approximation
SAA, see Sample Average Approximation
saddle point
convex-concave saddle point
problem, 79
Sample Average Approximation, 420–422
Schur Complement Lemma, 265
semidefinite relaxation
on ellitope
tightness of, 273
on spectratope
tightness of, 277
signal estimation, see signal recovery
signal recovery
linear, 267
on ellitope, 267–271
on ellitope, near-optimality of, 271–273
on spectratope, 277–291
on spectratope under uncertain-but-bounded noise, 291–297
on spectratope under uncertain-but-bounded noise, near-optimality of, 297
on spectratope, near-optimality of, 277, 290
problem setting, 1, 260
sparsity, s-sparsity, 3
spectratope, 274
calculus of, 299–302, 430
examples of, 276
Stochastic Approximation, 422–425
test, see Hypothesis Testing test
theorem
Sion-Kakutani, 81
vectors
notation, xxi