Course:

Advanced Nonlinear Programming – ISyE 7683

a.k.a.

Descriptive Foundations of Convex Optimization

- **Instructor:** Dr. Arkadi Nemirovski `nemirovs@isye.gatech.edu`,

Office hours (Zoom): Tuesday 10:00-11:45 and in-person by appointment Foreclose 446

You always can contact me by e-mail `nemirovs@isye.gatech.edu` *and by phone* `404-429-1528` *10:00am-9:00pm, except for time of our classes*

- **Teaching Assistant:** None
- **Classes:** Monday & Wednesday 14:00-15:15 ISyE Main 228
- **Lecture Notes, Transparencies**,

course site on Canvas and

  - Lecture Notes `https://www2.isye.gatech.edu/~nemirovs/KKN.pdf`
  - Solution Manual `https://www2.isye.gatech.edu/~nemirovs/KKN_SM.pdf` (password-protected)

- **Grading Policy:**
  - Take Home Final Exam: 100%

Syllabus

♣ **The course** focuses on the descriptive mathematical foundations of Convex Optimization. Convex Optimization is a "solvable case" in Mathematical Programming – under mild computability and boundedness assumptions; one can approximate global solutions to convex optimization problems to whatever high accuracy with reasonable computational effort.

♠ **Mathematics** of Convex optimization is composed of

(a) descriptive foundations (existence of optimal solutions, optimality conditions, duality, etc.),

(b) modeling (techniques allowing for posing the problem of interest as a convex optimization program), and

(c) operational toolbox — algorithms.

Traditionally, Mathematical Programming/Conex Optimization courses focus on algorithms and present descriptive foundations of Optimization at a minimal level sufficient to serve the needs of designing and analyzing the algorithms.

This course is different: its emphasis is on descriptive foundations and to a lesser extent – on modeling, with the algorithmic component reduced to a brief "executive summary" of Interior Point Methods – the state-of-the-art techniques for processing "well-structured" convex programs. Let me start with a brief historical excursion to motivate the emphasis on descriptive foundations.

Linear Programming — the "starting point" and one of the cornerstones of Convex Optimization — was discovered in the late 1940's and from the very beginning was equipped with a mighty solution algorithm, the Dantzig's Simplex method, making LP a "working entity," not a wishful thinking.

For over 40 years, the Simplex method was *the* LP algorithm; those working on other LP computational techniques were considered harmless city lunatics. As a result, the Simplex method was the focus of university courses on LP (some professors were even proud that they extracted the LP theory from this algorithm; in my opinion, this is as meaningful as extracting the first principles of Mechanics from the civil engineering manuals). Well, today, the default setting in all commercial LP solvers is Interior Point methods.

♠ **The morale:** In Applied Math, the "algorithmic toolbox" is the most important component, as far as applications are concerned; at the same time, it is the "most unstable" component: for LP, it was pivoting algorithms during the first ∼50 years, is primarily Interior Point Methods in the last 25-30 years, and nobody knows what will the computational "working horse" of LP in 15 years from now. In contrast, the descriptive component of LP was, for all practical purposes, completed in the early 1950s and has remained intact since then.

♠ I believe that as far as university education is concerned, it is important, along with "ready to use" knowledge that will serve the student well immediately upon graduation, to provide the students with "timeless" knowledge of foundations that will serve them well on the span of their entire professional life. The Pythagoras' Theorem, known as a "practical rule" for at least 3500 years and as a theorem – for about 2300 years, still is as instructive and useful as 2300 years ago; in contrast, how Babylonians, ancient Greeks, and Romans did their Arithmetic, is of absolutely no importance today and is interesting only to those studying the history of civilization.

♠ **The contents**, aside from the "executive summary" on Interior Point methods, represent what is called Convex Analysis (this is the technical name of the descriptive foundations of Mathematical Programming/Convex Optimization), including

• Basic results on the geometry of convex functions and sets (definitions, elementary properties, Caratheodory's, Helly's, Krein-Milman,... theorems, subdifferentials, Legendre transform,...)

- Calculus of convexity
- Convex optimization problems in Mathematical Programming, cone-constrained, and conic forms (definitions, basic properties, duality, optimality conditions,. . .)
- Saddle points and Sion-Kakutani Theorem

To get a complete impression of the descriptive component of the contents, see the course textbook at `https://www.isye.gatech.edu/~nemirovs/KKN.pdf`

♠ **Prerequisites.** Formal prerequisites are the most basic Linear Algebra, Calculus, and Real Analysis. The informal (and crucial!) prerequisite is the basic mathematical culture – the ability to comprehend, create, and enjoy rigorous mathematical reasoning. To give an example, the claim "2 x 2 = 5" does not witness the lack of mathematical culture; this is just a miscalculation. In contrast, the claim "2 x 2 = triangle" (believe me, from time to time, I hear in class something like this) does witness the lack of mathematical culture: one should know that under any circumstances, rain or snow, the product of two reals is a real, and not a triangle or a violin.

♠ To complete my informal Syllabus, let me present here my favorite Mathematical joke:

> A team flying on a balloon lost their orientation. Suddenly a gust of wind brought them closer to the ground, and they cried to a man they saw below: "Hey, where are we?" After a short pause, they got the answer: You are on a balloon." The next gust of wind lifted the balloon, and as their journey continued, the leader of the team said: "This was a mathematician. First, he thought before answering. Second, his answer was absolutely correct. Third, it was absolutely useless."

If you fully agree with the leader, then perhaps this course is not your best choice. As for me, I would say that

- To get a helpful answer, ask a proper question: do not ask, "Where are we?"; ask, "What is our location?"
- A useless answer, by definition, cannot be used and is therefore harmless. In contrast, acting based on a seemingly useful incorrect answer, you can get into trouble. . .

Arkadi Nemirovski

Preface

A man searches for a lost wallet at the place where the wallet was lost.
A wise man searches at a place with enough light...

♣ Where should we search for a wallet? Where is "enough light" – what Optimization can do well?

The most straightforward answer is: we can solve well *convex* optimization problems.

The very existence of what is called Mathematical Programming stemmed from discovery of Linear Programming (George Dantzig, late 1940's) – a modeling methodology accompanied by extremely powerful in practice (although "theoretically bad") computational tool – Simplex Method. Linear Programming, which is a special case of Convex Programming, still underlies the majority of real life applications of Optimization, especially large-scale ones.

♠ When photography was invented in XIX Century, processing pictures was very sophisticated and required skills and training.

• Kodak Company changed the situation completely by offering (1888) centralized processing of films. Their slogan was

You press the button, we do the rest

♠ *In the realm of Mathematical Programming, Convex Optimization is the area most close to this slogan,* with

"pressing the button" = *creating convex optimization model of the problem at hand and feeding this model with necessary data*

♣ The major components of Convex Optimization as a science are

• **Descriptive foundations** – basic properties of convex sets and functions (their geometry, calculus, etc.), classification of convex optimization problems, existence and characterization of optimal solutions, etc.;

• **Modeling** – techniques allowing to recognize optimization problems which can be converted to convex programs, and to carry out this conversion when it is possible;

• **Algorithms** – methods for numerical processing of convex optimization programs.

Note: Design and analysis of convex optimization algorithms is the major activity area in Optimization due to its importance for applications.

However: Algorithms form the most *un*stable component of Optimization

For example,

• Since the birth of Linear and Mathematical Programming (late 1940's) till mid-1990's, the Dantzig's Simplex Method was *the* LP algorithm

• Since mid-1990's, as a result of "Interior Point Revolution," the standard solvers for LP and other well-structured convex problems are Interior Point Path-Following methods (IPM's) which have nothing to do with pivoting LP algorithms, like Simplex Method. Common belief is that Interior Point Revolution and progress in hardware nearly equally contribute to the overall 10^6-fold acceleration of LP solvers

• Since $\approx 2010's$, IPM's have been augmented with the completely different from IPM's First Order algorithms aimed at solving extremely large-scale convex problems which, due to their huge sizes, are beyond the "practical grasp" of IPM's.

Conclusion: in the time scale typical for sciences, the "convex optimization toolbox" rapidly varies, and nobody knows what will be the "optimization working horses" in 15-20 years from now.

In contrast: *Descriptive foundations of Convex Optimization* (basically completed in the mid-1960's) *seem to be "eternal truths" forming a timeless backbone of Optimization, the knowledge destined to underline development of Convex Optimization in the foreseen future and beyond*.

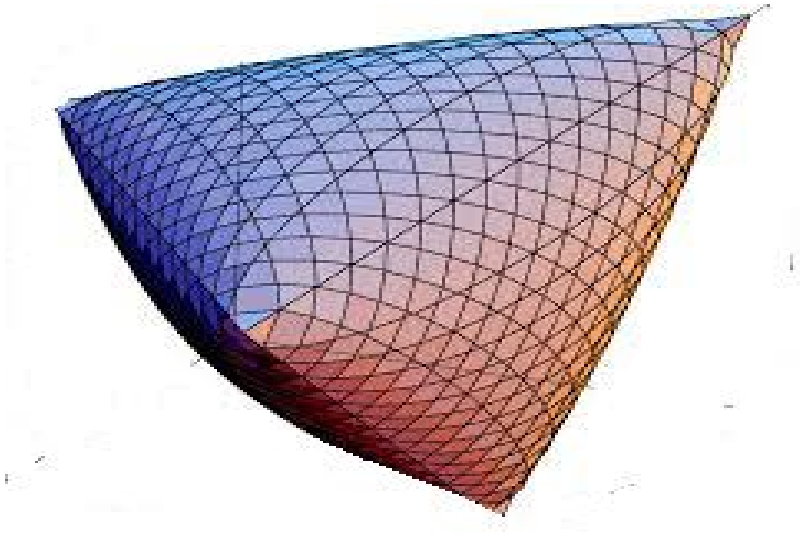
♣ **Traditionally,** university courses, undergraduate and graduate alike, mainly focus on "algorithmic toolbox" of Optimization, presenting (if at all) the foundational knowledge at the bare minimum sufficient to explain algorithms.
With this approach, the emphasis is at providing students with the "ready to use here and now" knowledge.
♠ **Our course is different:** its emphasis is on descriptive foundations of Convex Optimization (technical name: Convex Analysis), and to a lesser extent - on modeling; algorithms will be represented by "executive summary" of IPM's as applied to well-structured convex problems – those of Linear and Semidefinite Programming.
The underlying rationale is to provide listeners with knowledge which will allow them to comprehend and to carry out optimization-related research and, as far as Optimization is concerned, will serve them on the entire span of their professional careers.

Note: While I intend to obey the standards of rigorousness of pure Math, the selection of material has nothing to do with mathematical niceties, intention to be as general as possible, and other attributes of "science for the sake of science;" *the selection of material is motivated by the desire to provide listeners with the knowledge sufficient,* to the best of my professional judgement, *to comprehend, to develop, and to apply convex optimization models and algorithms.*

♠ The contents:

- **Basics of convex sets:**
  - Instructive examples and "calculus" (convexity-preserving operations)
  - Theorems of Caratheodory and Helly
  - Topology of convex sets
  - Descriptive basics of Linear Programming – General Theorem of the Alternative and Linear Programming duality
- **Separation Theorem** and its applications:
  - Extreme points, extreme rays, recessive directions
  - Finite-dimensional Krein-Milman Theorem
  - Geometry of polyhedral sets
- **Basics on convex functions:**
  - Instructive examples and "calculus"
  - Detecting convexity
  - Gradient Inequality and basics on subgradients
  - Maxima and minima
  - Legendre transform and Fenchel Duality
- **Basic theory of Convex Optimization:**
  - Lagrange Duality and Lagrange Duality Theorem for problems in standard, cone-constrained, and conic form
  - Conic Programming and Conic Duality Theorem
  - Saddle points and Sion-Kakutani Theorem
- **"Structure revealing" Conic representations** of convex sets and functions
- **Executive summary on Interior Point methods** for Linear/Conic Quadratic/Semidefinite Programming

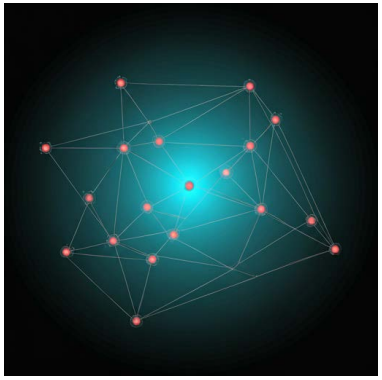
PART I.
Convex Sets



Lecture I.1

Metric Spaces

Definition
Convergence
Separability
Compactness
Continuity



Metric Spaces

♣ Let X be a set. A *metric* (a.k.a. *distance*) on X is a function $d(x,y) : X \times X \to \mathbf{R}$ which is
- *positive*: $d(x,y) > 0$ whenever $x \neq y$ and $d(x,x) = 0$ for all x
- *symmetric*: $d(x,y) = d(y,x)$ $\forall x, y$
- satisfies *Triangle inequality:* $d(x,z) \leq d(x,y) + d(y,z)$ for all x, y, z

♠ A set X equipped with metric is called *metric space*

Note: By Triangle inequality and symmetry, for $x, y, x \in X$ it holds $d(x,z) \leq d(x,y) + d(y,z)$ and $d(y,z) \leq d(y,x) + d(x,z) = d(x,y) + d(x,z)$, implying

Fact I.1 *When d is a metric on X, for all $x, y, z \in X$ it holds*

$$|d(x,z) - d(y,z)| \leq d(x,y).$$

Examples:

- The real line $\mathbf{R}$ equipped with the distance $d(x, y) = |x - y|$
- The space $\mathbf{R}^n$ of n-dimensional column vectors with the *uniform distance*

$$d(x, y) = \|x - y\|_\infty := \max_{i \le n} |x_i - y_i|$$

♠ **Remember**: When speaking about $\mathbf{R}$ and $\mathbf{R}^n$, we refer to the just defined metric spaces, not just to linear spaces!

♠ **Remember:** *Given metric space (X, d), we always treat subsets $Y \subset X$ as metric spaces, the metric being the restriction of metric d onto Y.*

Direct Product of Metric Spaces

♣ Passing from the metric space $\mathbf{R}$ to the metric space $\mathbf{R}^n$ is a special case of the following construction:

♠ **Direct product:** Given metric spaces $(X_1, d_1), ..., (X_m, d_m)$, we can build their *direct product* – the metric space $(X_1 \times X_2 \times ... \times X_m, d)$, where

— the *direct product* $X = X_1 \times X_2 \times ... \times X_m$ of the sets X_k is the set of all k-element ordered collections $(x_1, ..., x_m)$ with $x_k \in X_k$, $1 \leq k \leq m$

— the distance d on X is defined as

$$d((x_1, ..., x_m), (y_1, ..., y_m)) = \max_k d_k(x_k, y_k).$$

Convergence

♣ Let (X, d) be a metric space. A sequence $\{x^i \in X\}_{i\geq 1}$ is called *converging to* $\bar{x} \in X$ (equivalent wording: x^i *converge to* $\bar{x}$ *as* $i \to \infty$ or $\bar{x}$ is the limit of $\{x^i\}_i$), notation:

$$\bar{x} = \lim_{i\to\infty} x^i$$

if the distance from x^i to x converges to 0 as $i \to \infty$, that is, if *for every* $\epsilon > 0$ *and all large enough values of* i *one has* $d(x^i, \bar{x}) \leq \epsilon$.

A sequence $\{x^i \in X\}_i$ which has a limit is called *converging*.

Examples:

- $\lim_{i\to\infty} 1/i = 0$
- A sequence $\{x^i \in \mathbf{R}^n\}_{i\geq 1}$ converges to $\bar{x}$ iff for every $k \leq n$ the sequence $\{x^i_k\}_{i\geq 1}$ of k-th entries in x^i converges, as $i \to \infty$, to the k-th entry $\bar{x}_k$ of $\bar{x}$. More generally, *a sequence* $\{(x^i_1, ..., x^i_m)\}_i$ *of points from the direct product of metric spaces converges iff its "projections on the factors" – the sequences* $\{x^i_k\}_i$ *– converge in the respective factors,* $1 \leq k \leq m$.

Fact I.2 1) *The limit of a sequence, if any, is uniquely defined by the sequence*
2) *A converging sequence* $\{x^i\}$ *is a Cauchy sequence, meaning that* $d(x^i, x^j)$ *converges to zero as* $i, j \to \infty$*: for every* $\epsilon > 0$ *and for all large enough values of* i, j *it holds* $d(x^i, x^j) \leq \epsilon$

♣ Metric space (X, d) is called *complete*, if every Cauchy sequence of points from X converges (i.e., has a limit).

Die ganzen Zahlen hat der liebe Gott gemacht, alles andere ist Menschenwerk
[God made the integers, all else is the work of man] - Leopold Kronecker, 1886

♠

Fact I.3 *[The Most Basic Fact of Analysis]* $\mathbf{R}$ *equipped with the standard distance* $d(x,y) = |x-y|$ *is a complete metric space*

As an immediate corollary, $\mathbf{R}^n$ *is a complete metric space.* In fact, *Direct product of finitely many complete metric spaces if complete.*

Comment: • All we need to count entities are *natural numbers* $0, 1, 2, ...$

• The set $\mathbf{N}$ of natural numbers is equipped with two basic operations: addition and multiplication

— to make addition invertible, we extend $\mathbf{N}$ to the set $\mathbf{Z}$ of integers $\{0, \pm 1, \pm 2, ...\}$

— to make multiplication invertible as well, we next extend $\mathbf{Z}$ to the set $\mathbf{Q}$ of all rational numbers.

• $\mathbf{Q}$ still is too narrow – we need roots (e.g., to measure the diagonal of unit square). The next extension is the set of *algebraic numbers* – real roots of polynomials with rational coefficients.

• Algebraic numbers sufficient for Algebra are unsufficient for Analysis, where completeness is crucial. The real line $\mathbf{R}$ is the smallest complete extension of the set of rational (and of algebraic) numbers.

♠ *The journey from* $\mathbf{N}$ *to* $\mathbf{R}$ *took over 2000 years and was completed at the end of XIX Century*

Closed and open sets

♣ Let a metric space (X, d) be given. A set $Y \subset X$ is called
- *closed*, if the limit of every converging sequence of points from Y belongs to Y
- *open*, if along with every point $\bar{y} \in Y$, Y contains a ball of positive radius centered at $\bar{y}$:

$$\forall \bar{y} \in Y \exists r > 0 : d(y, \bar{y}) \leq r \Rightarrow y \in Y.$$

Examples:
- The empty set is both open and closed
- The segments $[0, 1] := \{x \in \mathbf{R} : 0 \leq x \leq 1\} \subset \mathbf{R}$ is closed;

the interval $(0, 1) := \{x \in \mathbf{R} : 0 < x < 1\} \subset \mathbf{R}$ is open;
the half-segment $[0, 1) := \{x \in \mathbf{R} : 0 \leq x < 1\} \subset \mathbf{R}$ is neither closed nor open.
- The subset $\mathbf{Z} = \{0, \pm 1, \pm 2, ...\}$ of $\mathbf{R}$ are closed, and its complement $\mathbf{R} \backslash \mathbf{Z} = \{x \in \mathbf{R} :$ x is not integer$\}$ is open

Fact I.4 **1)** *A set $Y \subset X$ is closed iff its complement $X \backslash Y$ is open*
2) *The intersection of any family of closed sets, same as the union of finite family of closed sets are closed*
3) *The intersection of finitely many open sets, same as the union of any family of open sets, is open*

Consequences:
• The *closed d-ball of radius $r \geq 0$ centered at $x \in X$* – the set $B_r(x) = \{x' : d(x, x') \leq r\}$ – is closed.
Indeed, the complement of $B_r(x)$ is the set $\{x' : d(x, x') > r$, and this set is open by Fact I.1.
• The *open d-ball of radius $r \geq 0$ centered at $x \in X$* – the set $B'_r(x) = \{x' : d(x, x') < r\}$ – is open (by Triangle inequality).

Fact I.5 *A subset in a complete metric space is closed iff it is complete*

Closure, Interior, Boundary

♣ Let (X, d) be metric space, and Y be a subset of X.

♠ The set $Y \subset X$ is a part of some closed subset of X (e.g., the entire X). As a result, there exists the smallest w.r.t. the inclusion closed set containing Y, specifically, the intersection of all closed sets containing Y. This set is called the *closure* $\mathrm{cl}\, Y$ of Y.

Example: When $X = \mathbf{R}$, one has $\mathrm{cl}\, \{x : 0 \leq x < 1\} = \{x : 0 \leq x \leq 1\}$.

Fact I.6 $\mathrm{cl}\, Y$ *is composed of the limits of all converging sequences of points from* Y.

Why Fact I.6 is not a tautology: clearly, a closed set containing Y contains the set $\overline{Y}$ of limits of converging sequences of points of Y $\Rightarrow$ *to prove that* $\overline{Y} = \mathrm{cl}\, Y$ *is the same as to prove that* $\overline{Y}$ *is closed*, which is not immediately evident: $\overline{Y}$ contains the limits of all converging sequences *from* Y; why it contains the limits of all converging sequences *from* $\overline{Y}$ – the set which can be larger than Y?

♣ Let (X, d) be metric space, and Y be a subset of X.

♠ $Y \subset X$ contains an open subset (e.g., the empty set). As a result, there exists the largest w.r.t. the inclusion open set contained in Y, specifically, the union of all open sets contained in Y. This set is called the *interior* $\operatorname{int} Y$ of Y.

Example: When $X = \mathbf{R}$, one has $\operatorname{int}\{x : 0 \le x < 1\} = \{x : 0 < x < 1\}$.

Fact I.7 $\operatorname{int} Y$ *is composed of all interior points of* Y *– points* y *belonging to* Y *along with* d*-ball* $B_r(y) = \{z : d(y, z) \le r\}$ *of positive (perhaps, small and depending on* y*) radii.*

♠ We clearly have $\operatorname{int} Y \subset Y \subset \operatorname{cl} Y$. The complement $\operatorname{cl} Y \backslash \operatorname{int} Y$ of the interior in the closure is called the *boundary* ∂Y of Y.

Example: with $X = \mathbf{R}$, we have

$$\partial\{x : 0 \le x \le 1\} = \partial\{x : 0 \le x < 1\} = \partial\{x : 0 < x \le 1\} = \partial\{x : 0 < x < 1\} = \{0, 1\}.$$

• $\operatorname{cl} Y$ is composed of all point of X which can be approximated to whatever high accuracy by points from Y – points such that every d-ball of positive radius centered at the point contains points of Y

• $\operatorname{int} Y$ is composed of all points which cannot be approximated to high enough accuracy by points from the complement $X \backslash Y$ of Y in X (what is "high enough," depends on the point)

• boundary ∂Y of Y is composed of all points of X which can be approximated to whatever high accuracy by points from Y and by points from $X \backslash Y$.

Fact I.8 *Consider metric spaces (X_k, d_k), $1 \leq k \leq m$, along with their direct product (X, d), and let $Y_k \subset X_k$, $1 \leq k \leq m$, and $Y = Y_1 \times ... \times Y_m$. Then*

- *Y is closed iff all Y_k are so*
- *Y is open iff all Y_k are so*
- $\mathrm{cl}\, Y = [\mathrm{cl}\, Y_1] \times ... \times [\mathrm{cl}\, Y_m]$
- $\mathrm{int}\, Y = [\mathrm{int}\, Y_1] \times ... \times [\mathrm{int}\, Y_m]$

Countable sets

♣ A set X is called *countable*, if its elements can be assigned *serial numbers* – indexes from $\{1, 2, ...\}$ – in such a way that different points from the set get different indexes.

Examples:

- Finite sets, including $\emptyset$
- Natural numbers $\mathbf{N} = \{1, 2, ...\}$ and integers $\mathbf{Z} = \{0, \pm 1, \pm 2....\}$
- Rational numbers $\mathbf{Q}$.

To "index" $\mathbf{Q}$, look one by one at the finite sets Q^s of rationals p/q with $|p| + |q| \leq s$, $s = 1, 2, ...$ and list first the elements of Q^0, then still unlisted elements of Q^1, then the still unlisted elements of Q^2, and so on.

Facts:

- A nonempty set X is countable iff there exists a sequence $\{x^i\}_i$ of its elements such that every $x \in X$ is a member of the sequence
- A subset of countable set is countable
- The union Q of a sequence $\{Q^s\}_s$ of countable sets Q^s is countable

Indeed, elements of Q can be assigned pairs of indexes s, i, with s being the index of the first of the sets Q^r containing the element, and i being the serial number of the element in Q^s. We can now list the elements of Q as follows: first those with $s + i \leq 1$, next the yet unlisted elements with $s + i \leq 2$, next the yet unlisted elements of with $s + i \leq 2$, and so on.

- Finite direct product $Q = Q_1 \times ... \times Q_m$ of countable sets Q_i is countable.

Indeed, the sets Q^s, $s = 1, 2, ...$, composed of all collections $(x^1, ..., x^m)$, $x_i \in Q_i$, with the sum, over $i \leq m$, of serial numbers of x^i in Q_i not exceeding s is finite, and Q is the union of the sets Q^s over $s = 1, 2, ...$

- $\mathbf{R}$ is *not* countable

Reals from [0.1) are sequences of digits 0,1,...,9, with sequences with all but finitely many terms equal to 9 excluded. Assuming that these reals can be assigned serial numbers 1,2,..., let us build a sequence of zeros and ones as follows: i-th term in this sequence is 0, if i-th digit in i-th real differs from 0, and is 1 otherwise. The resulting sequence is composed of digits 0 and 1 and differs from all our reals, which is a contradiction.

Separable metric space

♣ A metric space (X, d) is called *separable*, if it either is empty, or there exists a countable set $Y \subset X$ such that every $x \in X$ is the limit of a converging sequence of points from Y, or, equivalently, such that $X = \mathrm{cl}\, Y$.

Examples:

• $\mathbf{R}^n$. Indeed, the set $\mathbf{Q}^n$ of n-dimensional vectors is countable (as finite direct product of the countable sets Q of rational numbers), and every $x \in \mathbf{R}^n$ is the limit of a sequence of vectors with rational coordinates.

By similar reason,

• The (finite) direct product of separable metric spaces is separable.

• A subset Z of separable metric space (X, d) (considered as metric space with the metric inherited from (X, d)) is separable, In particular, *any subset of* $\mathbf{R}^n$ *is separable*.

The claim is clearly true when $Z = \emptyset$. Assuming $Z \neq \emptyset$, there exists a countable set Y such that $X = \mathrm{cl}\, Y$. For $s = 1, 2, ...$, let us build set $Z^s \subset Z$ as follows: looking one by one at points from Y (in the order given by the serial numbers of points in the *countable* set Y), we look whether the d-ball of radius $1/s$ centered at the current $y \in Y$ intersects Z. If it is the case, we add to Z^s (which initially is empty) a point $z(y)$ from this intersection, otherwise pass to the next $y \in Y$. As a result, we get a countable set $Z^s \subset Z$ such that every $z \in Z$ is at the d-distance at most $2/s$ from some point of Z^s (namely, any point $z(y)$ generated when processing $y \in Y$ with $d(y, z) < 1/s$; such an y exists, since Y is dense in $X \supset Z$; z is a candidate to the role of $z(y)$, so that $z(y)$ is well defined, and $d(z, z(y)) \leq d(y, z(y)) + d(y, z) \leq 2/s$). The union $W = \cup_s Z^s$ is a countable, along with all Z^s, subset of Z, and every point of Z clearly is the limit of a sequence of points from W.

Compact metric space

♠ A metric space (X, d) be a metric space is called *compact*, if every sequence of points from X contains a converging subsequence.

• *In the sequel, given metric space (X, d), we call a subset Y of X compact, if Y equipped with the metric d reduced to Y is a compact metric space.*

Example: • Finite metric space is compact

Note: • The entire $\mathbf{R}$ is not compact.

Fact I.9 • *Direct product of (finitely many) compact metric spaces is a compact metric space.*

• *When (X, d) is a compact metric space and $Y \subset X$, Y is compact iff Y is closed.*

Fact I.10 *Metric space (X, d) is compact iff*

A. *The space is totally bounded – for every $\epsilon > 0$, X can be covered by finite collection of balls of radius ϵ (equivalently: X admits finite ϵ-net – a finite collection of points such that every point from X is at the distance at most ϵ of some point from the collection)*

B. *The space is complete.*

Fact *Metric space (X, d) is compact iff*

A. *The space is totally bounded – for every $\epsilon > 0$, X can be covered by finite collection of balls of radius ϵ (equivalently: admits finite ϵ-net – a finite collection of points such that every point from X is at the distance at most ϵ of some point from the collection)*

B. *The space is complete.*

Proof: ✓ In one direction: Let (X, d) be a compact metric space, and let us prove that **A** and **B** take place.

A: Given $\epsilon > 0$, take a point $x^1 \in X$; if the ball $B_\epsilon(x^1)$ contains X, $\{x_1\}$ is a finite ϵ-set of X, otherwise we can find a point x^2 with $d(x^1, x^2) > \epsilon$. In the second case, it may happen that $X \in B_\epsilon(x^1) \cup B_\epsilon(x^2)$, so that $\{x^1, x^2\}$ is a finite ϵ-net, otherwise there exists x^3 such that $d(x^i, x^j) > \epsilon$ for $i \neq j$ and $1 \leq i, j \leq 3$. Proceeding in the same fashion, we either terminate with finite ϵ-net, or generate a sequence $\{x^i\}_{i\geq 1}$ with $d(x^i, x^j) > \epsilon$ for all $i \neq j$. The second option is contradictory, since the sequence $\{x^i\}$ clearly does *not* have a converging subsequence, which is impossible.

B: We should prove that a Cauchy sequence $\{x^i\}$ has a limit. This is immediate: by compactness, the sequence has a converging subsequence, and clearly the limit of a converging subsequence of a Cauchy sequence is the limit of this entire sequence as well.

✓ In the opposite direction: assume that **A** and **B** take place, and let us prove compactness, that is, that every sequence $\{x^i\}_{i\geq 1}$ has a converging subsequence. Indeed, by **A**, for every $k = 1, 2, ...$ there exists a finite $(1/k)$-net $\{u^k_\ell\}_{\ell \leq L_k}$. The union X of the L_1 balls $B_1(u^1_\ell)$ contains the entire sequence $\{x^i\}$ – let us call it "sequence 0," so that one of these balls contains its subsequence, let us call it "sequence 1." By similar argument, sequence 1 contains a subsequence, "sequence 2," belonging to one of the L_2 balls $B_{1/2}(u^2_\ell)$, $\ell \leq L_2$. Proceeding in the same fashion, we build a collection of sequences k, $k = 0, 1, 2, ...$, which is nested – the next sequence is a subsequence of the previous one, with sequence k belonging to a d-ball of radius $1/k$. The sequence $\{k$-th term of sequence $k\}_{k\geq 1}$ is a subsequence of sequence 0 (i.e., the sequence $\{x^i\}$) and by construction is a Cauchy, and thus converging by **B**, sequence. Thus, $\{x^i\}$ indeed has a converging subsequence.

Corollary I.1 *A compact metric space (X, d) is separable.*

Indeed, for $s = 1, 2, ...$, X admits a finite $1/s$-net Y^s. The union Y of finite sets $Y^1, Y^2, ...$ is countable, and clearly every $x \in X$ is the limit of a sequence from Y.

Fact I.11 *A set X in $\mathbf{R}^n$ is compact iff it is bounded and closed.*

Proof. ✓In one direction: A compact subset Y in $\mathbf{R}^n$ must be totally bounded, and thus bounded, and complete, and thus closed (as a complete subset of the complete metric space $\mathbf{R}^n$, see Fact I.5)

✓In the opposite direction: Let X be bounded and closed, and let us prove that X is compact. The bounded set $X \subset \mathbf{R}^n$ clearly is totally bounded – assuming that $X \subset B_R(0)$ and given $\epsilon > 0$, we can split $B_R(0)$ into finitely many boxes $\{x : i_k R/N \le x_k \le (i_k+1)R/K, 1 \le k \le N\}$ with integer $i_k \in \{-N, -N+1, ..., N-1\}$ and integer $N \ge R/\epsilon$. Selecting a point in every nonempty intersection of X and a box from the resulting finite family, we get finite ϵ-net in X. Since X is closed and $\mathbf{R}^n$ is complete, X is complete by Fact I.5. Being totally bounded and complete, X is compact by Fact I.13.

Illustrations:

- The following sets are compact:
  $\emptyset$, $[0,1] := \{x \in \mathbf{R} : 0 \le x \le 1\} \subset \mathbf{R}$, $\{x \in \mathbf{R}^n : \|x\|_\infty \le 1\}$
- The following sets are non-compact:
  $\mathbf{R}^n$ $(n \ge 1)$, $[0,1) := \{x \in \mathbf{R} : 0 \le x < 1\} \subset \mathbf{R}$, $\{x \in \mathbf{R}^n : \|x\|_\infty < 1\}$

Fact I.12 *A metric space (X, d) is compact iff from every covering of X by open sets one can extract a finite subcovering.*

Proof. ✓ In one direction: Let the space be compact, and let $\{U_\alpha\}_\alpha$ be a covering of X by open sets. Let us prove that one can select from this covering a finite subcovering.
A. First, let us prove that our claim is true when the given covering is a countable covering $\{U^i\}_{i\geq 1}$. Assuming that we can*not* extract from this covering a finite subcovering, for every k there exists a point $x^k \notin \cup_{i\leq k} U^i$. As X is compact, the resulting sequence $\{x^k\}$ has a subsequence $\{x^{k_j}\}_j$ converging, as $j \to \infty$, to some $\bar{x}$. The point $\bar{x}$ belongs to some of the sets U^i, say, to the set $U^{\bar{k}}$; this set is open, whence $x^{k_j} \to \bar{x} \in U^{\bar{k}}$ implies that $x^k_j \in U^{\bar{k}}$ for all large enough j. This is the desired contradiction – by construction $x^k \notin U^{\bar{k}}$ whenever $k \geq \bar{k}$.
B. It remains to extract the desired result from its just established "countable covering" version. To this end let us make a useful observation:
A separable metric space $X, d(\cdot)$ has a countable base – a countable collection $\{V^\ell\}_\ell$ of open sets such that every open set U in X is the union of all contained in U sets from the collection $\{V^\ell\}$. In particular, every compact set has a countable base.
Indeed, as X is separable, X has a dense countable sub set $\{x_i, i = 1, 2, ...\}$. The collection of all open balls of rational radii, each centered at a point from $\{x^i\}$, is countable, and it clearly is a desired base.
C. Taken together, **A** and **B** immediately imply that from every open covering $\{U_\alpha\}$ of X one can extract a finite subcovering. Indeed, let $\{V^\ell\}$ be a countable base of X, and let $\mathcal{L}$ be the set of indexes of those V^ℓ which are contained each in its own set of the family $\{U_\alpha\}$. As U_α is the union of all sets V^ℓ from the base contained in U_α and $\{U_\alpha\}$ form a covering of X, the countable collection $\{V^\ell\}_{\ell\in\mathcal{L}}$ is an open covering of X. By **A**, this open covering admits a finite subcovering $V^\ell_1, ..., V^\ell_J$; as $V^\ell_j \subset U_{\alpha_j}$ for properly selected α_j, the sets U_{α_j} form the desired finite subcovering of X by sets from the family $\{U_\alpha\}$.

✓ In the opposite direction: Assume that every open covering of X admits selection of a finite subcovering, and let us prove that every sequence $\{x^i \in X\}_{i\geq 1}$ has a converging subsequence. Assuming that the latter is the case, there exists a sequence $\{x^i \in X\}_{i\geq 1}$ with no converging subsequences ⇒ For every $y \in X$ there exists $r > 0$ such that the centered at y open d-ball of radius r does <u>not</u> contain x^i's with large enough values of i. The family of open balls $\{B_y, y \in X\}$ forms an open covering of X, and therefore admits selection of a finite subcovering B_{y^1}, ... B_{y^L}. For every k, x^k belongs to one of these L balls ⇒ *at least one of the balls B_{y^ℓ}, $\ell \leq L$, contains x^k for infinitely many values of index k.* This is a desired contradiction – by construction, every ball B_y contains x^k for finitely many values of k !

Fact I.13 *Let (X, d) be a metric space. The following properties of the space are equivalent to each other:*
(i) (X, d) is compact
(ii) Every open covering of X admits selection of a finite subcovering
(iii) If every finite collection of sets from a family of closed subsets in X has a nonempty intersection, all sets from the family have a nonempty intersection

Indeed, equivalence between (i) and (ii) is Fact I.12. To see the equivalence of (ii) and (iii), note that if $\{F_\alpha\}$ is a collection of closed subsets of X, then the intersection of all sets from the family is empty iff the collection of open subsets $\{X \backslash F_\alpha\}_\alpha$ form an open covering of X.

Continuity

♠ Let (X, d) and (Z, g) be metric spaces and $F : X \to Z$ be a mapping.

Definition F is called *continuous* at a point $\bar{x} \in X$, is for every sequence $\{x^i \in X\}_i$ converging to $\bar{x}$, the sequence $\{F(x^i)\}_i$ converges to $F(\bar{x})$.

- F is called *continuous*, if F is continuous at every point $x \in X$.

Equivalent "$\epsilon - \delta$" definition of continuity: F is continuous at x iff for every $\epsilon > 0$ there exists $\delta > 0$ such that

$$\{x' \in X, d(x, x') \leq \delta\} \Rightarrow g(F(x'), F(x)) \leq \epsilon.$$

Examples:

- Algebraic polynomial of n variables is continuous on $\mathbf{R}^n$.
- A function $F(x) = [F_1(x); ...; F_m(x)] : X \to \mathbf{R}^m$ is continuous iff every one of the real-valued functions $F_k(x)$, $k \leq m$, is continuous.

Fact I.14 *Let (X, d) and (Z, g) be metric spaces and $F : X \to Z$. F is continuous*
— iff the inverse image $F^{-1}(Y) := \{x : F(x) \in Y\}$ of every open subset Y of Z is open, and
— iff the inverse image $F^{-1}(Y)$ of every closed subset Y of Z is closed.

Note: Let (X, d) be a metric space. Then
A. *If a real-valued function $F : X \to \mathbf{R}$ is continuous, then the sublevel sets $\{x : F(x) \leq a\}$ and superlevel sets $\{x : F(x) \geq a\}$ of F are closed for every $a \in \mathbf{R}$* (as inverse images of the closed subsets $\{t : t \leq a\}$, $\{t : t \geq a\}$ of $\mathbf{R}$)*, while the sets $\{x : F(x) < a)$ and $\{x : F(x) > a\}$, $a \in \mathbf{R}$, are open* (as inverse images of open subsets of $\mathbf{R}$).
In particular, *the solution set $\{x \in \mathbf{R}^n : a_\alpha^T x \leq b_\alpha \, \forall \alpha \in \mathcal{A}\}$, of every system of nonstrict linear inequalities is closed* (as the intersection of sublevel sets of continuous functions $a_\alpha^T x$ of x).
B. Vice versa, *If $F : X \to \mathbf{R}$ is a real-valued function such that for every $a \in \mathbf{R}$ both the sublevel set $\{x : F(x) \leq a\}$ and the superlevel set $\{x : F(x) \geq a\}$ are closed* (or, which is the same, the complements to these sets $\{x : F(x) > a\}$ and $\{x : F(x) < a\}$ are closed), *then F is continuous.*

Fact I.15 1) *Let (X,d) be a metric space, and (Z,g) be the direct product of metric spaces (Z_k,d_k), $1\le k\le m$. Let also $F_k: X\to Z_k$ be mappings. The mapping*

$$F(x)=(F_1(x),...,F_m(x)): X\to Z$$

is continuous iff every one of $F_k: X\to Z_k$, $k\le m$, is so.

2) *[continuity of composition] Let (X,d), (Y,e), (Z,h) be metric spaces and $F: X\to Y$, $G: Y\to Z$ be mappings. Let also $\bar{x}\in X$, and $\bar{Y}=F(\bar{x})$. Assume that the mapping F is continuous at $\bar{x}$, and the mapping G is continuous at $\bar{y}$. Then the composite mapping*

$$H(x):=G(F(x)): X\to Z$$

is continuous at $\bar{z}$.

Fact I.16 *Let (X,d) be compact and let $F: X \to Z$ be continuous. Then the image $F(X) = \{y = F(x), x \in X\}$ is a compact subset of Z.*

Indeed, given an open covering $\{U_\alpha\}_\alpha$ of $F(Y)$, the sets $F^{-1}(U_\alpha)$ are open (since F is continuous) and clearly form a covering of $X \Rightarrow$ for properly selected $\alpha_1, ..., \alpha_K$ the sets U_{α_k}, $k \leq K$, form a covering of $Y \Rightarrow$ the open covering $\{U_\alpha\}_\alpha$ of Y admits selection of finite subcovering $\{U_{\alpha_k},\ k \leq K\}$ of $Y \Rightarrow Y$ is compact.

♠ **Corollary** [Weierstrass Theorem] *Let (X,d) be compact metric spaces nd F be continuous real-valued function. Then F is bounded and attains its minimum and maximum on X.*

Indeed, by Fact I.16, $F(X) = \{F(x) : x \in X\}$ is a compact subset of $\mathbf{R} \Rightarrow F(X)$ is nonempty and is closed and bounded by Fact 1.8 $\Rightarrow F$ is bounded along with $F(X)$. and attains its minimum and maximum at X, since a nonempty, closed and bounded set Y of reals contains the smallest and the largest point.

Exercise: Prove the latter claim.

Solution: Let us build a nested sequence of segments Δ_k intersecting Y and with lengths converging to 0 as $k \to \infty$. Namely, $\Delta_0 = [a_0, b_0]$ be a segment containing the nonempty compact (and thus nonempty and bounded!) set Y. Given segment $\Delta_k = [a_k, b_k]$ intersecting Y, we set $c_k = \frac{1}{2}[a_k + b_k]$. Δ_k intersects $Y \Rightarrow$ *at least one of the segments* $[a_k, c_k]$, $[c_k, b_k]$ *intersects* Y. If $[a_k, c_k]$ intersects with Y, we take it as Δ_{k+1}, otherwise we set $\Delta_{k+1} = [c_k, b_k]$.
By construction, the lengths of segments Δ_k go to 0 as $k \to \infty$, and the sequence $\{a_k\}_{k \geq 1}$ is nondecreasing and is bounded from above (by any point $y \in Y$).
Any nondecreasing bounded from above sequence of reals is a Cauchy sequence (why?); since $\mathbf{R}$ is complete, such a sequence converges. In particular, a_k converge to some $\bar{y}$ as $k \to \infty$. By construction $a_k \leq y$ for all $y \in Y$ and all $k \Rightarrow \bar{y} \leq y$ *for all* $y \in Y$. On the other hand, selecting $y_k \in \Delta_k \cap Y$ (the latter intersection is nonempty!), we get $|y_k - a_k| \leq b_k - a_k \to 0$, $k \to \infty$, which combines with $\bar{y} = \lim_{k \to \infty} a_k$ to imply that $\bar{y} = \lim_k y_k$. Since Y is closed, we conclude that $\bar{y} \in Y$. Thus, $\bar{y}$ is the smallest of the reals in Y. Similar reasoning demonstrates that among the reals from Y there exists the largest one.

Uniform Continuity

♣ Let (X,d) and (Z,g) be metric spaces, and let $F: X \to Y$ be a continuous mapping.
♠ Continuity of F means that *for every $x \in X$ and every $\epsilon > 0$ there exists $\delta > 0$ such that*

$$d(y,x) \le \delta \Rightarrow g(F(y),F(x)) \le \epsilon.$$

Given $\epsilon > 0$, the "$\delta > 0$" here is allowed to depend on x and can be arbitrarily small for "bad" choices of x. *Uniform continuity* of F means that the for every $\epsilon > 0$, *a single* $\delta > 0$ "serves" all $x \in X$:

Definition: $F: X \to Y$ is called *uniformly continuous*, if for every $\epsilon > 0$ there exists $\delta > 0$ such that

$$\{x, y \in X, d(x,y) \le \Delta\} \Rightarrow g(F(y),F(x)) \le \epsilon.$$

A continuous mapping not necessarily is uniformly continuous (look at the real-valued function $1/x$ on the positive ray $X = \{x \in R : x > 0\}$).

Fact I.17 *If (X,d) is compact and $F: X \to Y$ is continuous, F is uniformly continuous.*

Proof. Assuming the opposite, there exists $\epsilon > 0$ and a sequence of pairs $\{x^i, y^i\}_i$ such that $d(x^i,y^i) \to 0$ as $i \to \infty$, while $g(F(y^i),F(x^i)) > \epsilon$ for all i. Since (X,d) is compact, $\{x^i\}$ has a converging subsequence $\{x^{i_j}\}_j$. since $d(x^{i_j},y^{i_j}) \to 0$ as $j \to \infty$, we have $\bar{x} := \lim_{j\to\infty} x^{i_j} = \lim_{j\to\infty} y^{i_j}$. Since F is continuous at $\bar{x}$, we conclude that $F(\bar{x}) = \lim_{j\to\infty} F(x^{i_j}) = \lim_{j\to\infty} F(y^{i_j})$, contradicting $g(F(x^{i_j}),F(y^{i_j}) > \epsilon > 0$ for all j.

Norms

♠ A *norm* on $\mathbf{R}^n$ is a real-valued function $\|\cdot\|$ on $\mathbf{R}^n$ with the following properties:

- *positivity*: $\|x\|$ is positive whenever $x \neq 0$ and is zero when $x = 0$
- *positive homogeneity:* $\|\lambda x\| = |\lambda| \|x\|$ for every $x \in \mathbf{R}^n$ and $\lambda \in \mathbf{R}$
- *Triangle inequality:* $\|x + y\| \leq \|x\| + \|y\|$ for all $x, y \in \mathbf{R}^n$.

Note: By Triangle inequality, $\|x\| \leq \|x - y\| + \|y\|$ and $\|y\| \leq \|y - x\| + \|x\|$, which combines with homogeneity to imply

Fact I.18 *For every norm $\|\cdot\|$ on $\mathbf{R}^n$ and all $x, y \in \mathbf{R}^n$ it holds* $|\|x\| - \|y\|| \leq \|x - y\|$ $\forall x, y \in \mathbf{R}^n$

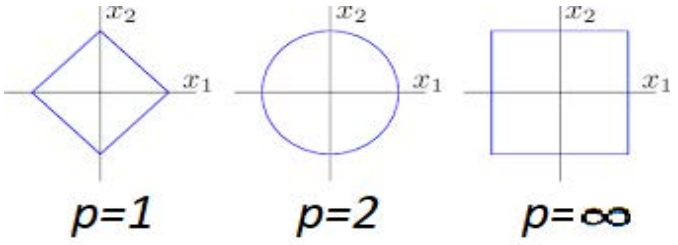
Examples:

- *The uniform (a.k.a. ℓ_∞) norm* $\|x\|_\infty = \max_{k \leq n} |x_k|$
- *The ℓ_1-norm* $\|x\|_1 = \sum_k |x_k|$

The fact that ℓ_1 and ℓ_∞ norms indeed are norms is justified by trivial verification.

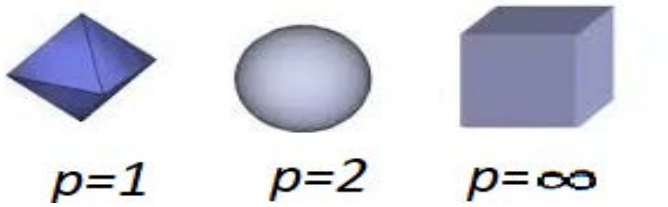
Note: $\|\cdot\|_1$ and $\|\cdot\|_\infty$ are "extremes" of the family of ℓ_p-norms $\|\cdot\|_p$, $1 \leq p \leq \infty$. When $1 < p < \infty$, $\|x\|_p = \left(\sum_k |x_k|^p\right)^{1/p}$.

Note: $\|x\|_\infty = \lim_{p\to\infty} \|x\|_p$.



2D

$p = 1$:rhombus $p = 2$:circle $p = \infty$:square



3D

$p = 1$:octahedron $p = 2$:ball $p = \infty$:cube

Unit $\|\cdot\|_p$-balls $\{x : \|x\|_p \leq 1\}$

$$\|x\|_p = \begin{cases} \left(\sum_k |x_k|^p\right)^{1/p} & , 1 \le p < \infty \\ \max_k |x_k| & , p = \infty \end{cases}$$

Positivity and homogeneity of $\|\cdot\|_p$ are evident. Verifying Triangle inequality for $1 < 0 < \infty$ requires some effort (for $p \neq 2$ postponed till better times).

♠ To see that ℓ_2 (a.k.a. "standard Euclidean") norm $\|x\|_2 = \sqrt{\sum_k x_k^2}$ indeed is a norm

— We start with *Cauchy inequality:* $|x^T y| \le \|x\|_2 \|y\|_2$ *for all* $x, y \in \mathbf{R}^n$

Indeed, the inequality is evident when $y = 0$, and

$$\begin{array}{l} \forall (x, y \in \mathbf{R}^n, y \neq 0): \\ \quad (x - ty)^T (x - ty) \ge 0 \forall t \in \mathbf{R} \Leftrightarrow f(t) := t^2 y^T y - 2t x^T y + x^T x \ge 0 \ \forall t \\ \quad \Rightarrow [x^T y]^2 - [x^T x][y^T y] \le 0 \\ \qquad \text{"the discriminant of everywhere nonnegative quadratic trinomial} \\ \qquad (f \text{ is a trinomial due to } y^T y > 0) \text{ is nonpositive"} \\ \Rightarrow |x^T y| \le \sqrt{x^T x}\sqrt{y^T y} \end{array}$$

— By Cauchy inequality,

$$\|x + y\|_2^2 = x^T x + 2x^T y + y^T y \le x^T x + 2\sqrt{x^T x}\sqrt{y^T y} + y^T y = [\sqrt{x^T x} + \sqrt{y^T y}]^2$$

that is, $\|x + y\|_2^2 \le [\|x\|2 + \|y\|_2]^2$, implying the Triangle inequality.

Note: From the proof of Cauchy inequality it follows that *when* $y \neq 0$, *the inequality* $|x^T y| \le \|x\|_2 \|y\|_2$ *is equality* iff the discriminant of the quadratic trinomial f is zero, that is, iff f has a real root $\bar{t}$, that is *iff* x *is a real multiple of* y.

Norms and Metrics

♣ A norm $\|\cdot\|$ on $\mathbf{R}^n$ induces a distance:

$$d_{\|\cdot\|}(x,y) = \|x-y\|$$

Fact I.19 *$d_{\|\cdot\|}(\cdot,\cdot)$ indeed is a distance. This distance possesses two additional properties:*
— it is shift-invariant: $d_{\|\cdot\|}(x+z,y+z) = d_{\|\cdot\|}(x,y)$ for all $x,y,z \in \mathbf{R}^n$
— it is homogeneous: $d_{\|\cdot\|}(\lambda x,\lambda y) = |\lambda| d_{\|\cdot\|}(x,y)$ for all $x,y \in \mathbf{R}^n$ and all $\lambda \in \mathbf{R}$.
Vice versa, every shift-invariant and homogeneous distance $d(\cdot,\cdot)$ on $\mathbf{R}^n$ is of the form $d_{\|\cdot\|}(\cdot,\cdot)$ for some, uniquely defined by the distance, norm $\|\cdot\|$.

Fact I.20 *Every two norms $\|\cdot\|$, $\|\cdot\|'$ on $\mathbf{R}^n$ are equivalent, meaning that they are within positive constant factors from each other: there exist positive constants c and C such that*

$$\forall x \neq 0 : c \leq \frac{\|x\|}{\|x\|'} \leq C.$$

• *As a corollary, all norm-induced distances on $\mathbf{R}^n$ result in the same notions of convergence, open and closed sets, closure, interior, compactness, etc.*

Note: Fact I.20 is characteristic for *finite-dimensional* linear spaces.

• As an additional corollary, invoking Fact I.19, we arrive at

Fact I.21 *A norm $\|\cdot\|$ on $\mathbf{R}^n$ is continuous (and even uniformly continuous) function.*

Fact *Every two norms $\|\cdot\|$, $\|\cdot\|'$ on $\mathbf{R}^n$ are equivalent, meaning that they are within positive constant factors from each other: there exist positive constants c and C such that*

$$\forall x \neq 0 : c \leq \frac{\|x\|}{\|x\|'} \leq C.$$

Proof. It suffices to verify that every norm $\|\cdot\|$ on $\mathbf{R}^n$ is equivalent to $\|\cdot\|_\infty$. Let $\|\cdot\|$ be a norm on $\mathbf{R}^n$.
A. Denoting by e^k, $k \leq n$, the standard basic orths in $\mathbf{R}^n$, we have

$$\|x\| = \|\sum_k x_k e^k\| \leq \sum_k \|x_k e^k\| = \sum_k |x_k| \|e^k\| \leq \|x\|_\infty \max_k \|e^k\|,$$

that is, $\|x\| \leq C\|x\|_\infty \; \forall x$, with $C = \sum_k \|e^k\|$.
B. As a consequence of **A**, $\|x\|$ *is a continuous function of* x. Indeed, $|\|x\| - \|y\|| \leq \|x - y\| \leq C\|x - y\|_\infty$, with the first inequality given by Fact 1.15.
C, Consider the set $X = \{x \in \mathbf{R}^n : \|x\|_\infty = 1\}$. This set clearly is nonempty, closed and bounded, and is therefore compact. The function $f(*x) = \|x\|$ is continuous on this set and therefore attains its minimum on the set (Weierstrass Theorem). Since f is positive outside of the origin, this minimum c is positive:

$$\|x\|_\infty = 1 \Rightarrow \|x\| \geq c > 0.$$

Both sides in this inequality are of the same homogeneity degree w.r.t.x., implying that

$$\forall x : \|x\| \geq c\|x\|_\infty,$$

Thus,

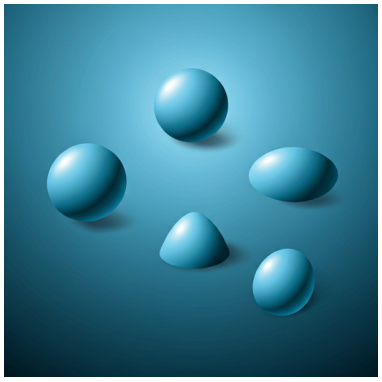
$$\forall x \in \mathbf{R}^n : \|x\| \leq C\|x\|_\infty \;\&\; \|x\| \geq c\|x\|_\infty \qquad [C < \infty, c > 0]$$

$\Rightarrow$ $\|\cdot\|$ is indeed equivalent to $\|\cdot\|_\infty$

Lecture I.2

Convex Sets – First Acquaintance

Definition
Basic examples
Calculus
Caratheodory and Helly Theorems

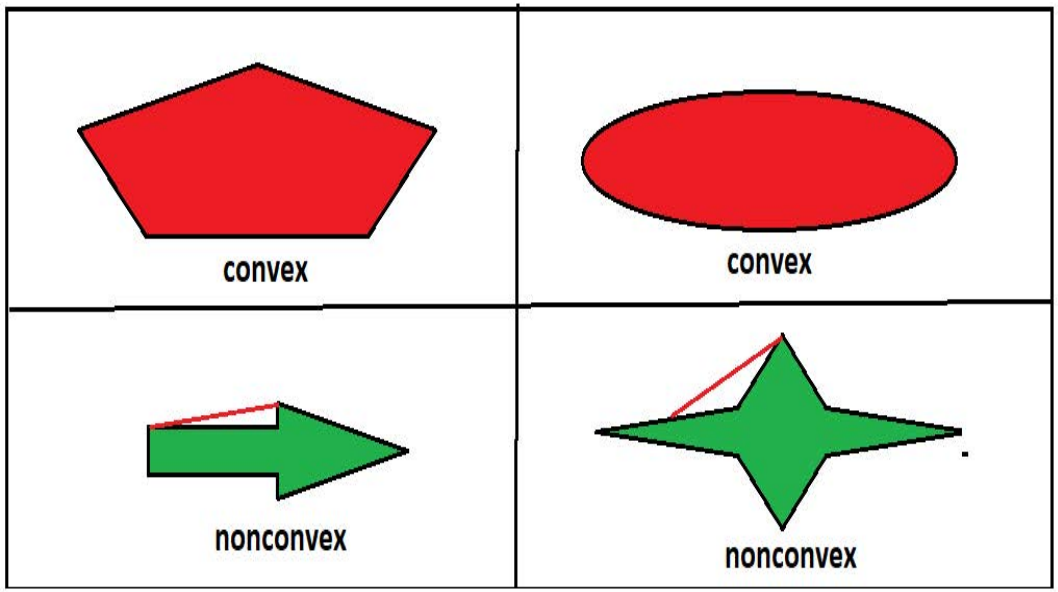


Convex Sets

Definition. A set $X \subset \mathbf{R}^n$ is called *convex*, if X contains, along with every pair x, y of its points, the entire segment $[x, y]$ with the endpoints x, y:

$$x, y \in X \Rightarrow (1-\lambda)x + \lambda y \in X \ \forall \lambda \in [0, 1].$$

Note: when λ runs through $[0, 1]$, the point $(1-\lambda)x + \lambda y \equiv x + \lambda(y-x)$ runs through the segment $[x, y]$.



♣ Immediate examples of convex sets in $\mathbf{R}^n$:

- $\mathbf{R}^n$
- $\emptyset$
- singleton $\{x\}$
- open unit box $\{x \in \mathbf{R}^n : -1 < x_i < 1, i \leq n\}$ and closed unit box $\{x \in \mathbf{R}^n : -1 \leq x_i \leq 1, i \leq n\}$

Remember: *The closure of a convex set is convex* (why?)

Examples of convex sets, I: Affine sets

Definition: *Affine set* M in $\mathbf{R}^n$ is a set which can be obtained as a shift of a *linear subspace* $L \subset \mathbf{R}^n$ by a vector $a \in \mathbf{R}^n$:

$$M = a + L = \{x = a + y : y \in L\} \tag{1}$$

Note: I. The linear subspace L is uniquely defined by affine subspace M and is the set of differences of vectors from M:

$$(1) \Rightarrow L = M - M = \{y = x' - x'' : x', x'' \in M\}$$

II. The shift vector a is *not* uniquely defined by affine subspace M; in (1), one can take as a *every* vector from M (and *only* vector from M):

$$(1) \Rightarrow M = a' + L \ \forall a' \in M.$$

Fact II.1 *[Generic example of affine subspace] The set of solutions of a solvable system of linear equations:*

$$\begin{array}{c} M \text{ is affine subspace in } \mathbf{R}^n \\ \Updownarrow \\ \emptyset \neq M \equiv \{x \in \mathbf{R}^n : Ax = b\} \equiv \underbrace{a}_{Aa=b} + \underbrace{\{x : Ax = 0\}}_{\mathrm{Ker}A} \end{array}$$

♣ By Fact II.1, affine subspace is convex, due to

Fact II.2 *The solution set of an arbitrary (finite or infinite) system of linear inequalities is convex:*

$$X = \{x \in \mathbf{R}^n : a_\alpha^T x \leq b_\alpha, \alpha \in \mathcal{A}\} \Rightarrow X \text{ is convex}$$

In particular, every polyhedral set $\{x : Ax \leq b\}$ *is convex.*

Proof:

$$\begin{aligned}
& x, y \in X, \lambda \in [0,1] \\
\Leftrightarrow \quad & a_\alpha^T x \leq b_\alpha,\ a_\alpha^T y \leq b_\alpha \forall \alpha \in \mathcal{A},\ \lambda \in [0,1] \\
\Rightarrow \quad & \underbrace{\lambda a_\alpha^T x + (1-\lambda) a_\alpha^T y}_{a_\alpha^T[\lambda x + (1-\lambda) y]} \leq \underbrace{\lambda b_\alpha + (1-\lambda) b_\alpha}_{b_\alpha} \ \forall \alpha \in \mathcal{A} \\
\Rightarrow \quad & [\lambda x + (1-\lambda) y] \in X \ \forall \lambda \in [0,1].
\end{aligned}$$

Remark: *Fact II.2 remains valid when part of the nonstrict inequalities* $a_\alpha^T x \leq b_\alpha$ *are replaced with their strict versions* $a_\alpha^T x < b_\alpha$.

Remark: *The solution set*

$$X = \{x : a_\alpha^T x \leq b_\alpha, \alpha \in \mathcal{A}\}$$

of a system of nonstrict inequalities is not only convex, it is closed (why?)

We shall see in the mean time that

• Vice versa, *every closed and convex set* $X \subset \mathbf{R}^n$ *is the solution set of an appropriate countable system of nonstrict linear inequalities:*

$$\begin{array}{c} X \text{ is closed and convex} \\ \Downarrow \\ X = \{x : a_i^T x \leq b_i,\ i = 1, 2, ...\} \end{array}$$

Examples of convex sets, II: Unit balls of norms

Recall that a real-valued function $\|x\|$ on $\mathbf{R}^n$ is called *a norm*, if it possesses the following three properties:

♢ [positivity] $\|x\| \geq 0$ for all x and $\|x\| = 0$ iff $x = 0$;

♢ [homogeneity] $\|\lambda x\| = |\lambda|\|x\|$ for all vectors x and reals λ;

♢ [triangle inequality] $\|x + y\| \leq \|x\| + \|y\|$ for all vectors x, y.

Proposition II.1 *Let $\|\cdot\|$ be a norm on $\mathbf{R}^n$. The unit ball of this norm – the set $\{x : \|x\| \leq 1\}$, same as any other $\|\cdot\|$-ball $\{x : \|x - a\| \leq r\}$, is convex.*

Proof:

$$\begin{array}{rl} & \|x - a\| \leq r, \|y - a\| \leq r, \lambda \in [0,1] \\ \\ \Rightarrow r & \geq \lambda\|x - a\| + (1-\lambda)\|y - a\| = \|\lambda(x - a)\| + \|(1-\lambda)(y - a)\| \\ & \geq \|\lambda(x - a) + (1-\lambda)(y - a)\| = \|[\lambda x + (1-\lambda)y] - a\| \\ \\ \Rightarrow & \|[\lambda x + (1-\lambda)y] - a\| \leq r \ \forall \lambda \in [0,1]. \end{array}$$

Note: By the same argument, *for a norm $\|\cdot\|$, the open $\|\cdot\|$-ball of radius $r \geq 0$ – the set $\{x' \in \mathbf{R}^n : \|x' - x\| < r\}$ – is convex.*

Note: By Fact I.21, a norm $\|\cdot\|$ on $\mathbf{R}^n$ is continuous, implying by Fact I.14 that the $\|\cdot\|$-balls $\{x : \|x - a\| \leq r\}$ are closed, while open $\|\cdot\|$-balls $\{x : \|x - a\| < r\}$ indeed are open.

Fact II.3 *The unit ball B of a norm $\|\cdot\|$ on $\mathbf{R}^n$ remembers the norm:*

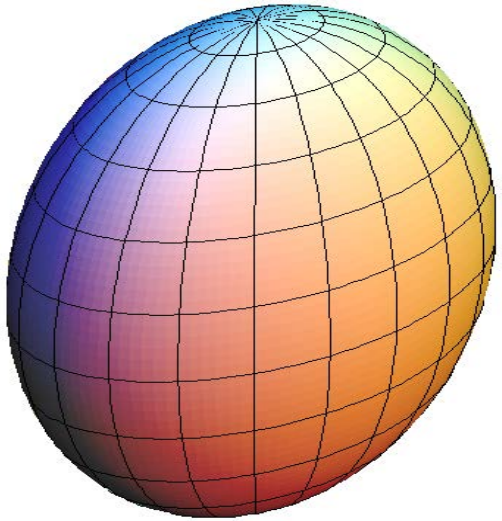
$$\forall x : \|x\| = \inf\{t > 0 : t^{-1}x \in B\}$$

Fact II.4 *[characterization of $\|\cdot\|$-balls] A set V in $\mathbf{R}^n$ is the unit ball of a norm iff V is*
(a) convex and symmetric w.r.t. 0: $V = -V$,
(b) bounded and closed (i.e., is compact), and
(c) satisfies $0 \in \operatorname{int} V$.

Examples of convex sets, III: Ellipsoid

Definition: An ellipsoid in $\mathbf{R}^n$ is a set X given by
$\diamondsuit$ positive definite and symmetric $n \times n$ matrix Q (that is, $Q = Q^T$ and $u^T Q u > 0$ whenever $u \neq 0$),
$\diamondsuit$ *center* $a \in \mathbf{R}^n$,
$\diamondsuit$ *radius* $r > 0$
via the relation

$$X = \{x : (x-a)^T Q (x-a) \leq r^2\}.$$



3D ellipsoid

$$X = \{x : (x-a)^T Q(x-a) \leq r^2\}.$$

Fact II.5 *An ellipsoid is convex.*

Proof. Since Q is symmetric positive definite, by Linear Algebra $Q = (Q^{1/2})^2$ for uniquely defined symmetric positive definite matrix $Q^{1/2}$. Setting $\|x\|_Q = \|Q^{1/2}x\|_2$, we clearly get a norm on $\mathbf{R}^n$ (since $\|\cdot\|_2$ is a norm and $Q^{1/2}$ is nonsingular). We have

$$\begin{aligned}(x-a)^T Q(x-a) &= [(x-a)^T Q^{1/2}][Q^{1/2}(x-a)] \\ &= \|Q^{1/2}(x-a)\|_2^2 = \|x-a\|_Q^2,\end{aligned}$$

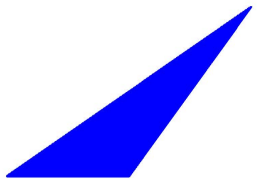
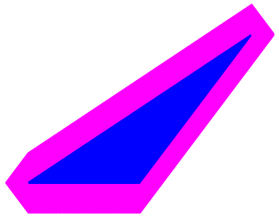
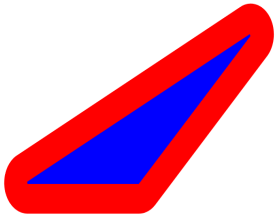
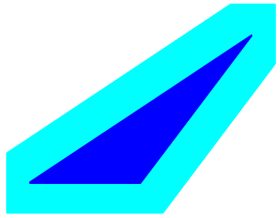
so that X is a $\|\cdot\|_Q$-ball and is therefore a convex set.

Examples of convex sets, IV: ϵ-neighbourhood of convex set

Fact II.6 *Let $\|\cdot\|$ be a norm in $\mathbf{R}^n$ and M be a nonempty convex set in $\mathbf{R}^n$, $\|\cdot\|$ be a norm, and $\epsilon \geq 0$. Then the set*

$$X = \{x : \mathrm{dist}_{\|\cdot\|}(x, M) \equiv \inf_{y \in M} \|x - y\| \leq \epsilon\}$$

is convex.

			
blue: M	magenta and blue: X $\|\cdot\| = \|\cdot\|_1$	red and blue: X $\|\cdot\| = \|\cdot\|_2$	cyan and blue: X $\|\cdot\| = \|\cdot\|_\infty$

Fact II.6 *Let $\|\cdot\|$ be a norm in $\mathbf{R}^n$ and M be a nonempty convex set in $\mathbf{R}^n$, $\|\cdot\|$ be a norm, and $\epsilon \geq 0$. Then the set*

$$X = \{x : \mathrm{dist}_{\|\cdot\|}(x, M) \equiv \inf_{y \in M} \|x - y\| \leq \epsilon\}$$

is convex.

Proof: $x \in X$ if and only if for every $\epsilon' > \epsilon$ there exists $y \in M$ such that $\|x - y\| \leq \epsilon'$. We now have

$$\begin{aligned}
& x, y \in X, \lambda \in [0,1] \\
\Rightarrow \quad & \forall \epsilon' > \epsilon \exists u, v \in M : \|x - u\| \leq \epsilon', \|y - v\| \leq \epsilon' \\
\Rightarrow \quad & \forall \epsilon' > \epsilon \exists u, v \in M : \\
& \qquad \underbrace{\lambda\|x - u\| + (1-\lambda)\|y - v\|}_{\geq \|[\lambda x + (1-\lambda)y] - [\lambda u + (1-\lambda)v]\|} \leq \epsilon' \ \forall \lambda \in [0,1] \\
\Rightarrow \quad & \forall \epsilon' > \epsilon \, \forall \lambda \in [0,1] \exists w = \lambda u + (1-\lambda)v \in M : \\
& \qquad \|[\lambda x + (1-\lambda)y] - w\| \leq \epsilon' \\
\Rightarrow \quad & \lambda x + (1-\lambda)y \in X \ \forall \lambda \in [0,1]
\end{aligned}$$

Convex Combinations and Convex Hulls

Definition: A *convex combination* of m vectors $x_1, ..., x_m \in \mathbf{R}^n$ is their linear combination

$$\sum_i \lambda_i x_i$$

with *nonnegative* coefficients and *unit sum of the coefficients*:

$$\lambda_i \geq 0 \ \forall i, \sum_i \lambda_i = 1.$$

Fact II.7 *A set $X \subset \mathbf{R}^n$ is convex iff it is closed w.r.t. taking convex combinations of its points:*

$$X \text{ is convex}$$
$$\Updownarrow$$
$$x_i \in X, \lambda_i \geq 0, \sum_i \lambda_i = 1 \Rightarrow \sum_i \lambda_i x_i \in X.$$

Proof, $\Rightarrow$: Assume that X is convex, and let us prove by induction in k that every k-term convex combination of vectors from X belongs to X. Base $k = 1$ is evident. Step $k \Rightarrow k+1$: let $x_1, ..., x_{k+1} \in X$ and $\lambda_i \geq 0$, $\sum_{i=1}^{k+1} \lambda_i = 1$; we should prove that $\sum_{i=1}^{k+1} \lambda_i x_i \in X$. Assume w.l.o.g. that $0 \leq \lambda_{k+1} < 1$. Then

$$\sum_{i=1}^{k+1} \lambda_i x_i = (1-\lambda_{k+1}) \Big(\underbrace{\sum_{i=1}^{k} \frac{\lambda_i}{1-\lambda_{k+1}} x_i}_{\in X} \Big) + \lambda_{k+1} x_{k+1} \in X.$$

Proof, $\Leftarrow$: evident, since the definition of convexity of X is nothing but the requirement for every *2-term* convex combination of points from X to belong to X.

Fact II.8 *The intersection* $X = \bigcap_{\alpha \in \mathcal{A}} X_\alpha$ *of an arbitrary family* $\{X_\alpha\}_{\alpha \in \mathcal{A}}$ *of convex subsets of* $\mathbf{R}^n$ *is convex.*

Proof: evident.

Corollary II.1 *Let* $X \subset \mathbf{R}^n$ *be an arbitrary set. Then among convex sets containing* X *(which do exist, e.g.* $\mathbf{R}^n$*) there exists the smallest one, namely, the intersection of all convex sets containing* X*.*

Definition: The smallest convex set containing X is called the *convex hull* $\mathrm{Conv}(X)$ of X.

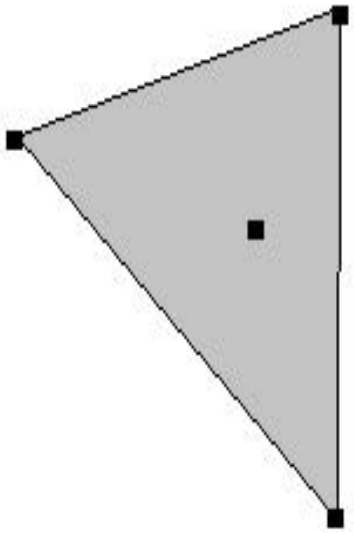
Fact II.9 *[convex hull via convex combinations] For every subset X of $\mathbf{R}^n$, its convex hull* $\mathrm{Conv}(X)$ *is exactly the set $\widehat{X}$ of all convex combinations of points from X.*

Proof. 1) Every convex set which contains X contains every convex combination of points from X as well. Therefore $\mathrm{Conv}(X) \supset \widehat{X}$.
2) It remains to prove that $\mathrm{Conv}(X) \subset \widehat{X}$. To this end, by definition of $\mathrm{Conv}(X)$, it suffices to verify that the set $\widehat{X}$ contains X (evident) and is convex. To see that $\widehat{X}$ is convex, let $x = \sum_i \nu_i x_i$, $y = \sum_i \mu_i x_i$ be two points from $\widehat{X}$ represented as convex combinations of points from X, and let $\lambda \in [0,1]$. We have

$$\lambda x + (1-\lambda) y = \sum_i [\lambda \nu_i + (1-\lambda)\mu_i] x_i,$$

i.e., the left hand side vector is a convex combination of vectors from X.

set X (4 points)



Conv(X) (triangle)

Examples of convex sets, V: simplex

Definition A collection of $m+1$ points x_i, $i = 0, ..., m$, in $\mathbf{R}^n$ is called *affine independent*, if no nontrivial combination of the points *with zero sum of the coefficients* is zero:

$$x_0, ..., x_m \text{ are affine independent}$$
$$\Updownarrow$$
$$\sum_{i=0}^{m} \lambda_i x_i = 0 \ \& \ \sum_i \lambda_i = 0 \Rightarrow \lambda_i = 0, 0 \le i \le m$$

Motivation: Let $X \subset \mathbf{R}^n$ be nonempty.
I. The intersection of all affine subspaces containing X is an affine subspace. This clearly is the *smallest* affine subspace containing X; it is called the *affine span* (or affine hull) $\mathrm{Aff}(X)$ of X.
Compare: *The intersection of all linear subspaces containing X is a linear subspace. This clearly is the smallest linear subspace containing X; it is called the linear span* $\mathrm{Lin}(X)$ *of X.*

II. It is easily seen that the affine span $\mathrm{Aff}(X)$ of X is nothing but the set of all *affine combinations* of points from X, that is, linear combinations *with unit sum of coefficients*:

$$\mathrm{Aff}(X) = \{x = \sum_i \lambda_i x_i : x_i \in X, \sum_i \lambda_i = 1\}.$$

Compare: It is easily seen that the linear span $\mathrm{Lin}(X)$ of X is nothing but the set of all *linear combinations* of points from X:

$$\mathrm{Lin}(X) = \{x = \sum_i \lambda_i x_i, x_i \in X\}$$

III. $m+1$ points $x_0, ..., x_m$ are affinely independent iff every point $x \in \mathrm{Aff}(\{x_0, ..., x_m\})$ of their affine span can be *uniquely represented* as an affine combination of $x_0, ..., x_m$:

$$\sum_i \lambda_i x_i = \sum_i \mu_i x_i \ \& \ \sum_i \lambda_i = \sum_i \mu_i = 1 \Rightarrow \lambda_i \equiv \mu_i$$

Compare:

• Vectors $y_1, ..., y_k$ are called linearly independent if no nontrivial linear combination of these vectors is zero:

$$\sum_i \lambda_i y_i = 0 \Rightarrow \ \lambda_i = 0 \, \forall i$$

• k vectors $y_1, ..., y_k$ are linearly independent iff every point $y \in \mathrm{Lin}(\{y_1, ..., y_k\})$ of their linear span can be *uniquely represented* as a linear combination of $y_1, ..., y_k$:

$$\sum_i \lambda_i y_i = \sum_i \mu_i y_i \ \ \Rightarrow \lambda_i \equiv \mu_i$$

♣ When $x_0, ..., x_m$ are affinely independent, the coefficients λ_i in the representation

$$x = \sum_{i=0}^{m} \lambda_i x_i \qquad [\sum_i \lambda_i = 1]$$

of a point $x \in M = \mathrm{Aff}(\{x_0, ..., x_m\})$ as an affine combination of $x_0, ..., x_m$ are uniquely defined by x and are called the *barycentric coordinates* of $x \in M$ *taken w.r.t. affine basis* $x_0, ..., x_m$ *of* M.

Fact II.10 *Let* $X \subset \mathbf{R}^n$ *be a nonempty set. Then* $M := \mathrm{Aff}(X)$ *has affine basis composed of points from* X. *Moreover, every affinely independent collection of vectors from* X *can be augmented by vectors from* X *to yield an affine basis of* M.

Compare:
Let $X \subset \mathbf{R}^m$. *Then* $L = \mathrm{Lin}(X)$ *admits a linear basis composed of vectors from* X. *Moreover, every linearly independent collection of vectors from* X *can be augmented by vectors from* X *to yield a linear basis of* $\mathrm{Lin}(X)$.

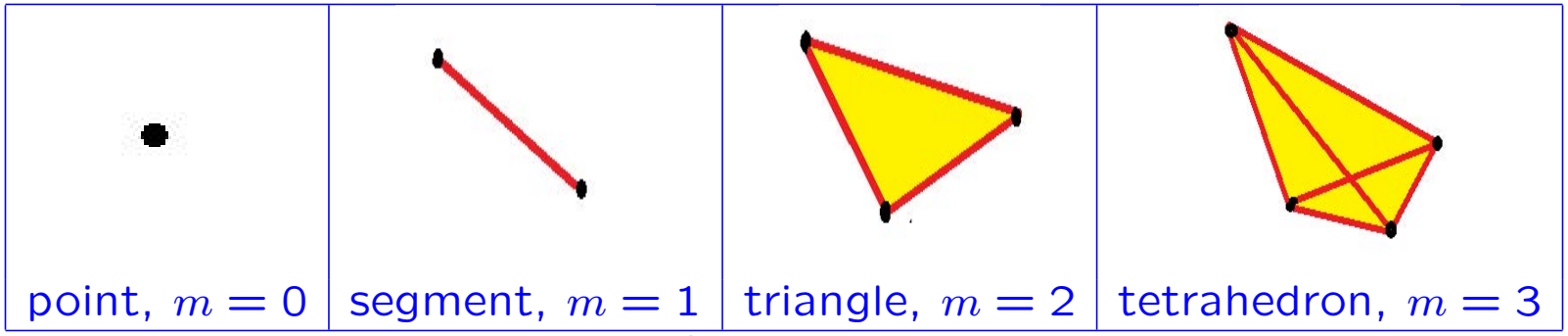
Definition: m-dimensional simplex Δ with vertices $x_0, ..., x_m$ is the convex hull of $m+1$ affine independent points $x_0, ..., x_m$:

$$\Delta = \Delta(x_0, ..., x_m) = \mathrm{Conv}(\{x_0, ..., x_m\}).$$

Examples: A. 2-dimensional simplex is given by 3 points not belonging to a line and is the triangle with vertices at these points.
B. Let $e_1, ..., e_n$ be the standard basic orths in $\mathbf{R}^n$. These n points are affinely independent, and the corresponding $(n-1)$-dimensional simplex is the *standard* (a.k.a *probabilistic*) *simplex* $\Delta_n = \{x \in \mathbf{R}^n : x \geq 0, \sum_i x_i = 1\}$.

C. Adding to $e_1, ..., e_n$ the vector $e_0 = 0$, we get $n+1$ affine independent points. The corresponding n-dimensional simplex is $\Delta_n^+ = \{x \in \mathbf{R}^n : x \geq 0, \sum_i x_i \leq 1\}$.

• *Simplex with vertices $x_0, ..., x_m$ is convex (as a convex hull of a set), and every point from the simplex is a convex combination of the vertices with the coefficients uniquely defined by the point.*



Simplexes

Examples of convex sets, VI: cones

Definition: A nonempty subset K of $\mathbf{R}^n$ is called *conic*, if it contains, along with every point x, the entire ray emanating from the origin and passing through x:

$$\begin{array}{c} K \text{ is conic} \\ \Updownarrow \\ K \neq \emptyset \ \& \ \forall (x \in K, t \geq 0) : tx \in K. \end{array}$$

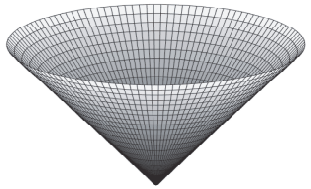
A *convex conic set* is called a *cone*.

Examples: A. *Nonnegative orthant*

$$\mathbf{R}^n_+ = \{x \in \mathbf{R}^n : x \geq 0\}$$

B. *Lorentz cone*

$$\mathbf{L}^n = \{x \in \mathbf{R}^n : x_n \geq \sqrt{x_1^2 + ... + x_{n-1}^2}\}$$



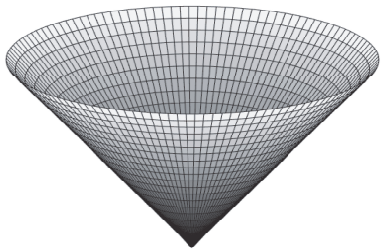
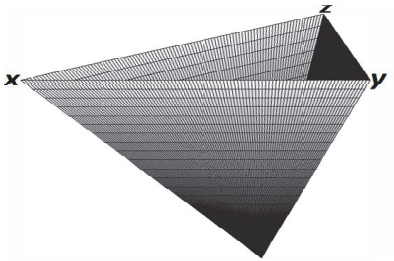
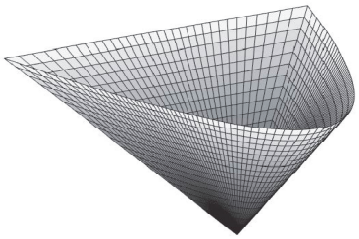
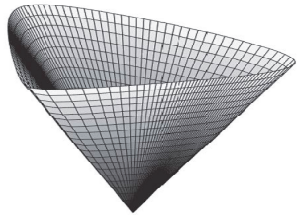
[Boundary of] 3D Lorentz cone $\mathbf{L}^3$

Lorentz cone is the epigraph of the standard Euclidean norm on $\mathbf{R}^{n-1}$:

$$\mathbf{L}^n = \{[x'; x_n] \in \mathbf{R}^{n-1} \times \mathbf{R} : x_n \geq \|x'\|_2\}.$$

It is immediately seen that *the epigraph* $\{[x'; x_n] \in \mathbf{R}^{n-1} \times \mathbf{R} : x_n \geq \|x'\|\}$ *of a norm on* $\mathbf{R}^{n-1}$ *is a cone.*

C. *Semidefinite cone* $\mathbf{S}^n_+$. This cone "lives" in the space $\mathbf{S}^n$ of $n \times n$ symmetric matrices and is composed of all positive semidefinite symmetric $n \times n$ matrices, that is, matrices $A \in \mathbf{S}^n$ producing nonnegative quadratic forms: $x^T Ax \geq 0 \, \forall x$

<table>
<tr>
<td> 
$\{(x,y,z) \in \mathbf{R}^3 : \left[\begin{array}{c|c|c} z & x & y \\ \hline x & z & 0 \\ \hline y & 0 & z \end{array}\right] \succeq 0\}$
3D Lorentz cone $\{z \geq \sqrt{x^2+y^2}\}$</td>
<td> 
$\{(x,y,z) \in \mathbf{R}^3 : \left[\begin{array}{c|c|c} x & 0 & 0 \\ \hline 0 & y & 0 \\ \hline 0 & 0 & z \end{array}\right] \succeq 0\}$
Nonegative orthant $\{x \geq 0, y \geq 0, z \geq 0\}$</td>
</tr>
<tr>
<td> 
$\{(x,y,z) \in \mathbf{R}^3 : \left[\begin{array}{c|c|c} z & x & y \\ \hline x & z & x \\ \hline y & x & z \end{array}\right] \succeq 0\}$</td>
<td> 
random 3D cross-section of $\mathbf{S}^3_+$</td>
</tr>
</table>

[boundaries of] Several 3D cross-sections of $\mathbf{S}^3_+$

D. *The solution set* $\{x : a_\alpha^T x \le 0 \, \forall \alpha \in \mathcal{A}\}$ of an arbitrary (finite or infinite) *homogeneous* system of *nonstrict* linear inequalities is a *closed* cone. In particular, so is a *polyhedral cone* $\{x : Ax \le 0\}$.
Note: Every *closed* cone in $\mathbf{R}^n$ is the solution set of a countable system of nonstrict *homogeneous* linear inequalities.

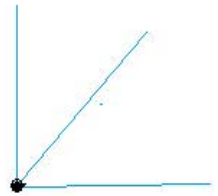
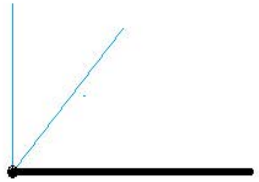
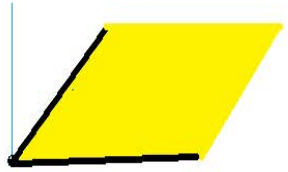
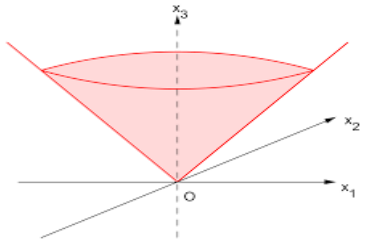
Cones **A** – **D** are closed.

Remember: *The closure of a cone is a cone* (why?)

Fact II.11 *A nonempty subset* K *of* $\mathbf{R}^n$ *is a cone iff*
♢ K *is conic:* $x \in K, t \ge 0 \Rightarrow tx \in K$, *and*
♢ K *is closed w.r.t. addition:*

$$x, y \in K \Rightarrow x + y \in K.$$

Proof, $\Rightarrow$: Let K be convex and $x, y \in K$, Then $\frac{1}{2}(x + y) \in K$ by convexity, and since K is conic, we also have $x + y \in K$. Thus, a convex conic set is closed w.r.t. addition.
Proof, $\Leftarrow$: Let K be conic and closed w.r.t. addition. In this case, a convex combination $\lambda x + (1 - \lambda)y$ of vectors x, y from K is the sum of the vectors λx and $(1 - \lambda)y$; since K is conic, both these vectors belong to K along with x, y. It follows that $\lambda x + (1 - \lambda)y \in K$, since K is closed w.r.t. addition. Thus, a conic set which is closed w.r.t. addition is convex.

			
Origin	Ray of x_1-axis	Quadrant of x_1x_2-plane	Ice-cream cone
$\{x = 0\}$	$\{x_1 \geq 0, x_2 = x_3 = 0\}$	$\{x_1 \geq 0, x_2 \geq 0, x_3 = 0\}$	$\{x_3 \geq \sqrt{x_1^2 + x_2^2}\}$

Several cones in $\mathbf{R}^3$

♣ Cones form an extremely important class of convex sets with properties "parallel" to those of general convex sets. For example,

♢ Intersection of an arbitrary family of cones again is a cone. As a result, *for every nonempty set X, among the cones containing X there exists the smallest cone* Cone (X), *called the conic hull of X.*

By definition Cone $(\emptyset) = \{0\}$

♢ A nonempty set is a cone iff it is closed w.r.t. taking *conic* combinations of its elements (i.e., linear combinations with nonnegative coefficients).

♢ The conic hull of a set $X \subset \mathbf{R}^n$ is exactly the set of all conic combinations of elements of X.

Note: we use the standard convention *sum of vectors $x^i \in \mathbf{R}^n$ taken over an empty set of indexes i has a value, namely, the origin.*

Conic and Perspective Transforms

♣ Let $X \subset \mathbf{R}^n$ be a nonempty convex set.

♠ *Conic transform* $\mathrm{ConeT}(X)$ of X is the conic hull of the set $X^+ = X \times \{1\} = \{[x;1] : x \in X\} \subset \mathbf{R}^{n+1}$.

- $\mathrm{ConeT}(X)$ is a cone living in the half-space $\{[x;t] \in \mathbf{R}^n_x \times \mathbf{R}^1_t : t \geq 0\}$ of $\mathbf{R}^{n+1}$.
- the only point of $\mathrm{ConeT}(X)$ of the form $[x;0]$ is the origin
- *points* $[x;t] \in \mathrm{ConeT}(X)$ *with* $t > 0$ *are exactly the points with* $x/t \in X$*, or, equivalently, the points of the form* $[ty;t]$ *with* $y \in X$*, and,*

$$\mathrm{ConeT}(X) = \{t[x;1] : t \geq 0, x \in X\}$$

Indeed, points $[x;t] \in \mathrm{ConeT}(X)$ with $t > 0$ are exactly the points $[x;t] = \sum_i \lambda_i [x^i;1]$ with $x^i \in X$ and $\lambda_i \geq 0$. For such a point, $\sum_i \lambda_i = t \Rightarrow y := t^{-1} \underbrace{\sum_i \lambda_i x^i}_{=x} \in X$ (X is convex!) *and* $[x;t] = [ty;t]$. Vice versa, when $[x;t] = [ty;t]$ with $y \in X$ and $t > 0$, then $[x;t] = t \underbrace{[y;1]}_{\in X^*} \Rightarrow [x;t] \in \mathrm{Cone}\,(X^+) = \mathrm{ConeT}(X)$.

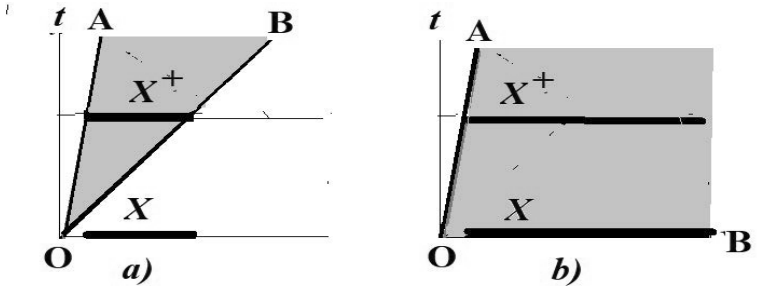
$\Rightarrow$ *Cross-section* $\Pi_t = \{x : [x;t] \in \mathrm{ConeT}(X)\}$ *is*

– tX*, when* $t > 0$

– $\{0\}$*, when* $t = 0$

– $\emptyset$*, when* $t < 0$.

Geometrically: to get ConeT(X), we
— place Z in the hyperplane $t = 0$ of $\mathbf{R}^{n+1} = \mathbf{R}^n_x \times \mathbf{R}^1_t$ and "lift" it along the t-axis to get X^+
— take the union of emanating from the origin rays of $\mathbf{R}^{n+1}$ crossing X^+; thus union is ConeT(X).



Conic transform
a) conic transform of segment X is the angle AOB
b) conic transform of ray X is the angle AOB with relative interior of the ray OB excluded

♠ The "nonzero part" ConeT(X)\{0} of the conic transform of X is called *the perspective transform of X:*

$$\text{Persp}(X) = \text{ConeT}(X)\backslash\{0\} = \{[x;t] : t > 0, t^{-1}x \in X\}$$

The perspective transform of X is convex along with X.

♠ Conic transform of a *closed* nonempty convex set not necessarily is closed (see example b) on the previous plot.

Fact II.12 *Let X be a closed nonempty convex set and $\alpha > 0$. Then*
- *The part of* $\mathrm{ConeT}(X)$ *in the half-space* $\{[x;t] : t \geq \alpha\}$ *is closed*
- $\mathrm{ConeT}(X)$ *is closed iff X is bounded.*

♠ Let $X \subset \mathbf{R}^n$ be nonempty and convex. The *closed conic transform of X* is, by definition, the closure of $\mathrm{ConeT}(X)$.
- $\overline{\mathrm{ConeT}}(\mathrm{cl}\, X) = \overline{\mathrm{ConeT}}(X)$
- When X is closed, the parts of $\mathrm{ConeT}(X)$ and $\overline{\mathrm{ConeT}}(x)$ in the domain $t > 0$ of $\mathbf{R}^{n_x} \times \mathbf{R}^1_t$ are the same.

Examples:
- $X = \{a\} \in \mathbf{R}^n \Rightarrow \mathrm{ConeT}(X) = \overline{\mathrm{ConeT}}(X) = \mathbf{R}_+ \cdot [a;1]$
- X is the unit ball of norm $\|\cdot\| \Rightarrow \mathrm{ConeT}(X) = \overline{\mathrm{ConeT}}(X)$ is the epigraph of $\|\cdot\|$
- $X \subset \mathbf{R}^n$ is a closed cone $\Rightarrow \mathrm{ConeT}(X) = [X \times \{t > 0\}] \cup \{0\}$, $\overline{\mathrm{ConeT}}(X) = X \times \mathbf{R}_+$

Recessive directions and recessive cone

♣ Let X be a *nonempty, convex, and closed* set in $\mathbf{R}^n$.

♠ A vector $d \in \mathbf{R}^n$ is called a *recessive direction of* X, if X contains a ray directed by d:

$$\exists \bar{x} \in X : \bar{x} + td \in X \, \forall t \geq 0$$

Examples:

- $d = 0$ is a recessive direction for every X
- every vector $d = [d_1; d_2] \in \mathbf{R}^2$ with $d_1 \geq 0$ is a recessive direction of the right half-plane $X = \{[x_1; x_2] : x_1 \geq 0\} \subset \mathbf{R}^2$.

Fact II.13 *If X contains a ray directed by d, it contains all parallel rays emanating from points from X:*

$$\exists \bar{x} : \bar{x} + td \in X, \, \forall t \geq 0 \Rightarrow \forall (x \in X, t \geq 0) : x + rd \in X$$

Indeed, let $d \in \mathbf{R}^n$ and $\bar{x} \in X$ be such that $\bar{x} + td \in X$ for all $t \geq 0$, and let $x \in X$. Given $t \geq 0$, the points $\bar{x} + it$ and x belong to the convex set X

$\Rightarrow$ $x^i = (1 - 1/i)x + (1/i)[\bar{x} + itd = [x + (1/i)(\bar{x} - x)] + td \in X$

As $i \to \infty$, $x^i \in X$ converge to $x + td$. Since X is closed we get $x + td \in X$, Q.E.D.

♠ The set $\mathrm{Rec}(X)$ of all recessive directions of a nonempty closed convex set X is called the *recessive cone* of X. From Fact II.13 it follows that

Fact II.14 $\mathrm{Rec}(X)$ *is a closed cone such that*

$$X + \mathrm{Rec}(X) = X.$$

As an immediate corollary,

Fact II.15 *A nonempty closed convex set A contains a line $a + \mathbf{R} \cdot d$ directed by $d \neq 0$ iff $0 \neq \pm d \in \mathrm{Rec}(X)$, and for such a d, every line directed by d and intersecting X is contained in X.*

Note The set of $d \in \mathbf{R}^n$ such that $\{a\} + \mathbf{R} \cdot d \subset X$ for some a is a linear subspace, called the *recessive subspace* of X. Nonzero directions of lines contained in X are exactly the nonzero vectors, if any, from the recessive subspace of X. The recessive subspace of nonempty closed convex set X is the same as the recessive subspace of the cone $\mathrm{Rec}(X)$ and is $\mathrm{Rec}(X) \cap [-\mathrm{Rec}(X)]$.

Examples of recessive cones:

- When X is bounded, Rec$(X) = \{0\}$
- When X is a closed cone, Rec$(X) = X$
- When $X = \{[x; t] \in \mathbf{R}^2 : t \geq x^2\}$, Rec$(X)$ is the ray $\{[0; t] : t \geq 0\}$

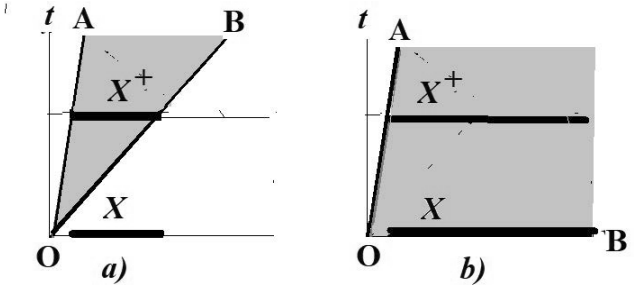
Fact II.16 *The recessive cone of a nonempty polyhedral set $X = \{x : Ax \leq b\}$ is the polyhedral cone $\{d : Ad \leq 0\}$ given by homogeneous versions of the linear inequalities specifying X*

Indeed, $b \geq A[x + td]$ for all $t \geq 0$ iff $b \geq Ax$ & $0 \geq Ad$

Fact II.17 *Let X be a nonempty closed convex set in $\mathbf{R}^n$. Then*
1. *When X is unbounded, the $\|\cdot\|_2$-unit recessive directions of X are exactly the asymptotic directions of X – the $\|\cdot\|_2$-unit vectors $d \in \mathbf{R}^n$ such that $d = \lim_{i\to\infty} x^i/\|x^i\|_2$ for some diverging (i.e., with $\|x^i\|_2 \to \infty$ as $i \to \infty$) sequence $\{x^i\}_i$ of vectors from X.*
2. *X is bounded iff $\mathrm{Rec}(X)$ is trivial: $\mathrm{Rec}(X) = \{0\}$*

Proof 1) ✓ Let X be unbounded. When $d \in \mathrm{Rec}(A)$ is a $\|\cdot\|_2$-unit vector and $x^i = x * id$, $i = 1, 2, ...$ form a diverging sequence, and $\|x^i\|_2^{=1} x^i$ converge to d as $i \to \infty \Rightarrow d$ is an asymptotic direction of X
✓ Vice versa, let d be an asymptotic direction of X, and $\{x^i \in X\}$ be a diverging sequence with $d = \lim_{i\to\infty} x^i/\|x^i\|_2$. For $t \geq 0$ and all but finitely many values of i we have $\|x^i i x^1\|_2 > t \Rightarrow$ vectors for these i $x^i_t = x^1 + t(x^i - x^1)/\|x^i - x^1\|$ are convex combinations of $x^1 \in X$ and $x^i \in X$ and thus belong to X (X is convex!). Since $\{x^i\} - i$ is diverging, $x^i_t \to x^1 + td$ as $i \to \infty \Rightarrow x^1 + td \in X$ (X is closed!) $\Rightarrow d \in \mathrm{Rec}(X)$.
2) When X is bounded, we clearly have $\mathrm{Rec}(X) = \{0\}$. When X is unbounded, we can select a diverging sequence $\{x^i \in X\}_i$; the sequence $x^i/\|x^i\|_2$ of $\|\cdot\|_2$-unit vectors is bounded; passing to a subsequence, we may assume that $x^i/\|x^i\|_2 \to d$, as $i \to \infty$. d is an asymptotic direction of X, and by 1), $d \in \mathrm{Rec}(X)$. Thus, the cone $\mathrm{Rec}(X)$ contains unit vector and this is different from the trivial cone $\{0\}$.

♣ How to visualize the recessive cone?



a) conic transform of segment X is the angle AOB
b) conic transform of ray X is the angle AOB with relative interior of the ray OB excluded

♠ Let $X \subset \mathbf{R}^n$ be a nonempty closed convex set. As we remember, the cross-sections

$$\Pi_t = \{x : [x;t] \in \mathrm{ConeT}(X)\},\ \overline{\Pi}_t = \{x : [x;t] \in \overline{\mathrm{ConeT}}(X)\}$$

for $t \neq 0$ coincide with each other:

$$t \neq 0 \Rightarrow \Pi_t = \overline{\Pi}_t = \begin{cases} tX & , t > 0 \\ \emptyset & , t < 0 \end{cases}$$

and $\Pi_0 = \{0\}$.

Question: *What is* $\overline{\Pi}_0$ *?*

Answer: $\overline{\Pi}_0 = \mathrm{Rec}(X)$.

Indeed, ✓ When $d \in \mathrm{Rec}(X)$, $x \in X$, and $t > 0$, we have $x + t^{-1}d \in X \Rightarrow [tx + d; t] \in \mathrm{ConeT}(X) \Rightarrow [d; 0] = \lim_{t \to +0}[tx + d; t] \in \mathrm{cl}\,\mathrm{ConeT}(X) = \overline{\mathrm{ConeT}}(X) \Rightarrow d \in \overline{\Pi}_0$.

✓ Vice versa, if $[d; 0] \in \overline{\mathrm{ConeT}}(X))$, then $[d, 0] = \lim_{i \to \infty}[y^i; t_i]$ with $t_i > 0$ and $x^i := y^i/t_i \in X$. When $d \neq 0$ the sequence $\{x^i \in X\}_i$ is diverging due to $t_i \to +0$ and $t_i x^i = y^i \to d \neq 0$ as $i \to \infty$, whence also $t_i\|x^i\|_2 \to \|d\|_2 > 0$ as $i \to \infty \Rightarrow x^i/\|x^i\|_2 = [t_i x^i]/[t_i\|x_i\|_2] \to d/\|d\|_2$ as $i \to \infty \Rightarrow d/\|d\|_2$ is an asymptotic direction of $X \Rightarrow d/\|d\|_2 \in \mathrm{Rec}(X)$ (Fact II.17) $\Rightarrow d \in \mathrm{Rec}(X)$, Q.E.D.

♠ Closed conic transform of a nonempty convex set $X \subset \mathbf{R}^n$ is a closed cone which lives in the half-space $\mathbf{R}^n \times \mathbf{R}_+$ of $\mathbf{R}^N_x \times \mathbf{R}_t$ and does *not* belong to the subspace $t = 0$ of the latter space. This observation can be inverted:

Fact II.18 *Let K be a closed cone contained in $\mathbf{R}^n \times \mathbf{R}_+$ and not contained in $\mathbf{R}^n \times \{0\}$. Then K is the closed conic transform of a nonempty closed convex set, specifically, the set*

$$X = \{x \in \mathbf{R}^n : [x; 1] \in K\}.$$

Indeed, under the premise of Fact, K contains a vector with positive last entry, and since K is a cone, it contains a vector with last entry equal to 1 $\Rightarrow$ X is nonempty (and clearly closed along with K. Now, $\mathrm{ConeT}(X)$ is the smallest cone containing $X^+ = X \times \{1\}$, whence $\overline{\mathrm{ConeT}}(X) = \mathrm{cl}\,\mathrm{ConeT}(X)$ is the smallest *closed* cone containing X^+. By construction of X, the closed cone K contains X^+ and therefore $\overline{\mathrm{ConeT}}(X) \subset K$. To prove the inverse inclusion $K \subset \overline{\mathrm{ConeT}}(X)$, note that by construction of X, every point $[x; t] \in K$ with $t > 0$ is a positive multiple of a point from X^+ and therefore belongs to $\mathrm{ConeT}(X)$ and thus to $\overline{\mathrm{ConeT}}(X)$. It remains to note that as K contains a vector $[\bar{x}; \bar{t}]$ with positive $\bar{t}$ and lives in $\mathbf{R}^n \times \mathbf{R}^+$, *every* point $[x; t] \in K$ is the limit, as $i \to \infty$, of the points $[x + \bar{x}/i; t + \bar{t}/i]$ which are vectors from K with positive last entries and thus belong to $\mathrm{ConeT}(X)$. Consequently, $K \subset \mathrm{cl}\,\mathrm{ConeT}(X) = \overline{\mathrm{ConeT}}(X)$. □

Calculus of Convex Sets and Cones

♣ The standard Calculus starts with *"raw materials"* – a handful of simple univariate functions like $\equiv 1$, $\equiv x$, exp, log, sin, etc., for which we compute derivatives "by bare hands" – according to the definition of the derivative. To make Calculus indeed working, these raw materials are augmented by *calculus rules* stating what happens with the derivatives when carrying out operations preserving, under appropriate assumptions, differentiability (summation, multiplication, division, taking superpositions, etc.).
The convex sets, the situation is similar. We already possess a fistful of examples of convex sets. It is time to outline *basic convexity-preserving operations*

Fact II.19 *The following operations preserve convexity of sets:*

1) Taking intersection: *If $X_\alpha \subset \mathbf{R}^n$, $\alpha \in \mathcal{A}$, are convex sets, so is $\bigcap\limits_{\alpha\in\mathcal{A}} X_\alpha$*

- *If all X_α are cones, so is $\bigcap\limits_{\alpha\in\mathcal{A}} X_\alpha$.*

2) Taking direct product: *If $X_\ell \subset \mathbf{R}^{n_\ell}$, $1 \le \ell \le L$, are convex sets, so is the set*

$$\begin{array}{rcl} X & = & X_1 \times ... \times X_L \\ & \equiv & \{x = (x^1, ..., x^L) : x^\ell \in X_\ell, 1 \le \ell \le L\} \\ & := & \mathbf{R}^{n_1+...+n_L} \end{array}$$

- *If all X_ℓ are cones, so is $X_1 \times ... \times X_L$.*

3) Taking weighted sums: *If $X_1, ..., X_L$ are nonempty convex sets in $\mathbf{R}^n$ and λ_1,...,λ_L are reals, then the set*

$$\begin{array}{rcl} & & \lambda_1 X_1 + ... + \lambda_L X_L \\ & := & \{x = \lambda_1 x_1 + ... + \lambda_L x_L : x_\ell \in X_\ell, 1 \le \ell \le L\} \end{array}$$

is convex.

- *If all X_ℓ are cones, so is $\lambda_1 X_1 + ... + \lambda_L X_L$*

4. Taking affine image: *Let $X \subset \mathbf{R}^n$ be convex and $x \mapsto \mathcal{A}(x) = Ax + b$ be an affine mapping from $\mathbf{R}^n$ to $\mathbf{R}^k$. Then the image of X under the mapping – the set*

$$\mathcal{A}(X) := \{y = Ax + b : x \in X\}$$

is convex.

- *If X is a cone and $\mathcal{A}(x) = Ax$ is linear, then $\mathcal{A}(X)$ is a cone.*

5. Taking inverse affine image: *Let $X \subset \mathbf{R}^n$ be convex and $y \mapsto \mathcal{A}(y) = Ay + b$ be an affine mapping from $\mathbf{R}^k$ to $\mathbf{R}^n$. Then the inverse image of X under the mapping – the set*

$$\mathcal{A}^{-1}(X) := \{y : Ay + b \in X\}$$

is convex.

- *If X is a cone and $\mathcal{A}(y) = Ay$ is linear, then $\mathcal{A}^{-1}(X)$ is a cone.*

Illustration: Consider a factory which can utilize at various intensities n types of production processes, consuming k types of resources and producing m types of products. Given the available volumes of resources $r = [r_1; ...; r_k]$ and requested volumes of products $p = [p_1; ...; p_m]$, the management should decide on *production plan* – vector $x = [x_1; ...; x_n]$ of intensities at which the production processes will be used. A production plan $x = [x_1; ...; x_n]$ is feasible if and only if x, r, and d satisfy the system of constraints

$$\begin{array}{rcll} Dx & \geq & d & \text{[demand must be satisfied]} \\ Rx & \leq & r & \text{[resource bounds must be obeyed]} \\ x & \in & X & \text{[technological feasibilty constraints]} \end{array} \qquad (S)$$

Assume that the set X of feasible production plans is convex.
Question: *What is the convexity status of the set of implementable pairs (r, d), that is, the set $\mathcal{RD} = \{(r, d) : \exists x : (x, r, d)$ satisfy $(S)\}$?*

Answer: $\mathcal{RD}$ is convex.

$$\frac{Dx \geq d \quad (a), \quad Rx \leq r \quad (b), \quad x \in X \quad (c)}{\mathcal{RD} = \{(r,p) : \exists x : (x,r,d) \text{ satisfy } \quad (a),(b),(c)\}}$$

Claim: When X is convex, so is $\mathcal{RD}$.
Indeed,

- the set $\mathcal{S}$ of solutions (x,r,d) to the system of linear constraints $(a),(b)$ is polyhedral and thus convex,
- the set $\mathcal{X} = \{(x,r,d) : x \in X\}$ is the direct product of convex sets X, $\mathbf{R}^k_r$ and $\mathbf{R}^m_d$ and thus is convex,

$\Rightarrow$ the set

$$\mathcal{XS} = \mathcal{X} \cap \mathcal{S} = \{(x,r,d) : x,r,d \text{ satisfy } (a),(b) \quad \text{and} \quad x \in X\}$$

is convex as intersection of two convex sets

$\Rightarrow$ the set $\mathcal{RD}$ is convex as the image of the set $\mathcal{XS}$ under the linear mapping $(x,r,d) \mapsto (r,d)$.

Calculus of Closed Convex Sets

♣ Many important results of Convex Analysis require not just convexity, but also *closedness* of the sets involved. This is how Calculus of convex sets changes when we want to preserve not just convexity, but also cloyedness:

♠ "Good" operations: taking intersections, direct products, inverse affine images

Fact II.20 *The following operations preserve convexity and closedness of sets:*

1) Taking intersection: *If $X_\alpha \subset \mathbf{R}^n$, $\alpha \in \mathcal{A}$, are closed convex sets, so is $\bigcap\limits_{\alpha\in\mathcal{A}} X_\alpha$*

2) Taking direct product: *If $X_\ell \subset \mathbf{R}^{n_\ell}$, $1 \le \ell \le L$, are closed convex sets, so is the set*

$$\begin{array}{rcl} X & = & X_1 \times ... \times X_L \\ & \equiv & \{x = (x^1, ..., x^L) : x^\ell \in X_\ell, 1 \le \ell \le L\} \\ & \subset & \mathbf{R}^{n_1+...+n_L} \end{array}$$

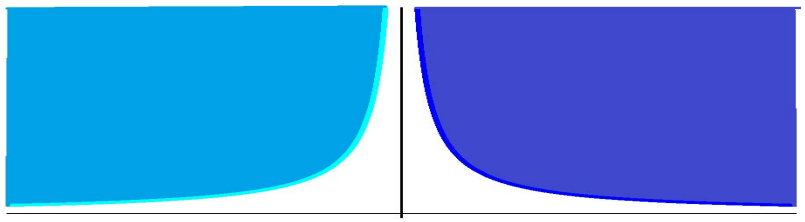
3) Taking inverse affine image: *Let $X \subset \mathbf{R}^n$ be convex and closed, and $y \mapsto \mathcal{A}(y) = Ay + b$ be an affine mapping from $\mathbf{R}^k$ to $\mathbf{R}^n$. Then the inverse image of X under the mapping – the set*

$$\mathcal{A}^{-1}(X) = \{y : Ay + b \in X\}$$

is convex and closed (as the inverse image of a closed set under a continuous mapping).

♠ Problematic operations: taking weighted sums and affine images

A. Taking weighted sums: "As is", a weighted sum of closed nonempty convex sets not necessarily is closed. For example, the sets $X_1 = \{x \in \mathbf{R}^2 : x_1 < 0, x_2 \geq -1/x_1\}$ and $X_2 = \{x \in \mathbf{R}^2 : x_1 > 0, x_2 \geq 1/x_1\}$ are nonempty, closed and convex:



while their sum is a nonclosed set – the interior $\{x \in \mathbf{R}^2 : x_2 > 0\}$ of the upper half-plane. However,

Fact II.21 *If $X_1, ..., X_L$ are nonempty closed sets in $\mathbf{R}^n$, with at most one of the sets unbounded, and λ_1,...,λ_L are reals, then the set*

$$Y = \lambda_1 X_1 + ... + \lambda_L X_L$$

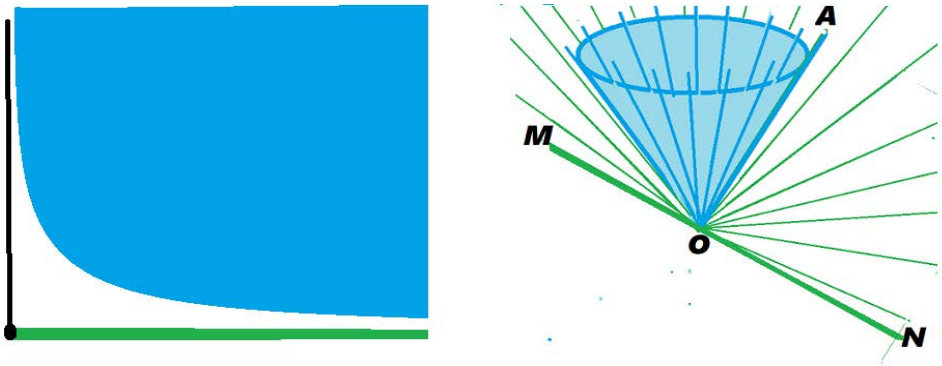
is closed. If, in addition, all X_ℓ is convex, so is Y (this we already know).

Indeed, let $X_1, ..., X_L$ be nonempty and closed, and $X_1, ..., X_{L-1}$ be bounded, and let

$$\{y^i = \textstyle\sum_{\ell=1}^{L} \lambda_\ell x_\ell^i\}_i, \ x_\ell^i \in X_\ell,$$

be a converging sequence of points from Y. To prove that the limit $\bar{y}$ of the sequence belongs to Y, assume w.l.o.g. that $\lambda_L \neq 0$. Since $X_1, ..., X_{L-1}$ are bounded, passing to a subsequence, we can assume that the $L-1$ sequences $\{x_\ell^i\}_i$, $\ell \leq L-1$, converge. Since $\{y^i\}_i$ converges as well and $\lambda_L \neq 0$, the sequence $\{x_L^i\}_i$ also converges $\Rightarrow \bar{y} = \sum_\ell \underbrace{\lim_{i\to\infty} x_\ell^i}_{\in X_\ell = \mathsf{Cl}\, X_\ell} \in Y$.

B. Taking affine images: The affine image of a closed convex set not necessarily is closed, e.g., the projection of the above closed convex set $X_1 = \{x \in \mathbf{R}^2 : x_1 < 0, x_2 \geq -1/x_1\}$ is the non-closed set $\{s \in \mathbf{R} : s < 0\}$.



Examples of closed convex sets with non- closed projections:
Left: The projection of the closed blue set onto the x-axis is the nonnegative ray *with the origin excluded*
Right: The projection of the 3D ice-cream cone onto the 2D plane tangent to the cone along the ray $[OA)$ is half-plane *with all points on the boundary line MN, except for O, excluded.*

However: *Affine image of a closed and bounded convex set X is closed* (as the image of a compact set under continuous mapping), or, more generally (see Fact II.23 below)

Let $X \subset \mathbf{R}^n$ be a nonempty closed convex set and $x \mapsto \mathcal{A}(x) = Ax + b : \mathbf{R}^n \to \mathbf{R}^m$ be an affine mapping such that $\mathrm{Ker}(A) \cap \mathrm{Rec}(X) = \{0\}$. *Then $\mathcal{A}(X)$ is closed.*

Note: We shall eventually see that *as applied to polyhedral (*and thus closed*) operands, all basic convexity-preserving operations preserve polyhedrality, and thus preserve closedness.*

Calculus of Recessive Cones

♣ What happens with recessive cones of nonempty closed convex sets under convexity-preserving operations with the sets?

♠ "Good" operations: taking intersections, direct products, and inverse affine images

Fact II.22 *The following operations with nonempty closed convex sets act naturally on their recessive cones:*

1) Taking intersection: *Let $X_\alpha \subset \mathbf{R}^n$, $\alpha \in \mathcal{A}$, be closed convex sets such that $X := \bigcap_\alpha X_\alpha \neq \emptyset$. Then X is a closed convex set with $\mathrm{Rec}(X) = \bigcap_\alpha \mathrm{Rec}(\alpha)$.*

2) Taking direct product: *If $X_1, ..., X_L$ are nonempty closed convex sets, then so is $X_1 \times ... \times X_L$, and*

$$\mathrm{Rec}(X_1 \times ... \times X_L) = \mathrm{Rec}(X_1) \times ... \times \mathrm{Rec}(X_L).$$

3) Taking inverse affine image: *Let $X \subset \mathbf{R}^n$ be a nonempty closed convex set and $y \to \mathcal{A}(y) := Ay + b : \mathbf{R}^m \to \mathbf{R}^n$ be an affine mapping such that the inverse image $Y = \{y : \mathcal{A}(y) \in X\}$ of X is nonempty. Then Y is a nonempty closed convex set, and $\mathrm{Rec}(Y)$ is the inverse image $\{d : Ad \in \mathrm{Rec}(X)\}$ of $\mathrm{Rec}(X)$ under the linear part $y \mapsto Ax$ of the affine mapping $\mathcal{A}$.*

This is evident.

♠ Problematic operations: taking weighted sums and affine images.

Taking affine images and weighted sums of nonempty closed convex sets acts poorly on the recessive cones of the sets.

A. *When taking affine image $\mathcal{A}(X) = \{y = Ax + b : x \in X\}$ of a nonempty closed convex set $X \subset \mathbf{R}^n$ under affine mapping $\mathcal{A}(x) = Ax + b : \mathbf{R}^n \to \mathbf{R}^m$, the image $A\mathrm{Rec}(X)$ of the recessive cone of X under the linear part of the mapping clearly belongs to the recessive cone of the closure of $\mathcal{A}(X)$* ($\mathcal{A}(X)$ is nonempty and convex, but not necessarily is closed). However. $A\mathrm{Rec}(X)$ can be negligibly small as compared to $\mathrm{Rec}(\mathrm{cl}\,\mathcal{A}(X))$. For example, when $X = \{[x;t] \in \mathbf{R}^2 : t \geq x^2\}$ and $\mathcal{A}([x;t]) = x$, we have $\mathrm{Rec}(X) = \{[0;t] : t \geq 0\}$ $\Rightarrow A\mathrm{Rec}(X) = \{0\}$, while $\mathcal{A}(X) = \mathbf{R} \Rightarrow \mathrm{Rec}(\mathrm{cl}\,\mathcal{A}(X)) = \mathbf{R}$.

However:

Fact II.23 *Let $X \subset \mathbf{R}^n$ be a nonempty closed convex set and $\mathcal{A}(x) = Ax + b : \mathbf{R}^n \to \mathbf{R}^m$ be an affine mapping. Assume that $\mathrm{Ker}A \cap \mathrm{Rec}(X) = \{0\}$. Then $\mathcal{A}(X)$ is closed, and $\mathrm{Rec}(\mathcal{A}(X)) = A\mathrm{Rec}(X)$. In particular, linear image AK of a closed cone K not intersecting $[\mathrm{Ker}A]\backslash\{0\}$ is closed.*

To see that $Y := \mathcal{A}(X)$ is closed, assume that $Y \ni y^i \to \bar{y}$ as $i \to \infty$ and $\bar{y} \notin Y$, and let us lead this assumption to a contradiction. We have $y^i = Ax^i + b$ with $x^i \in X$, and *the sequence $\{x^i\}$ diverges* (since otherwise the limit $\bar{x}$ of a converging subsequence of $\{x^i\}$ were a point from X (X is closed!), and $Y \ni \mathcal{A}(\bar{x}) = \lim_i \mathcal{A}(X^i) = \bar{y}$, which is a contradiction. Since $\{x^i\}$ diverges and $\{\mathcal{A}(x^i)\}$ converges, (every) asymptotic direction h of $\{x^i\}$ satisfies $Ah = 0$ (why) $\Rightarrow \mathrm{Ker}A \cap \mathrm{Rec}(X) \neq \emptyset$, which is a desired contradiction.

Let us prove that $A\mathrm{Rec}(X) = \mathrm{Rec}(Y)$ We already know that $A\mathrm{Rec}(X) \subset \mathrm{Rec}(Y)$. Now let $h \in \mathrm{Rec}(Y)$, and let us prove that $h \in A\mathrm{Rec}(X)$. There is nothing to prove when $h = 0$. Now let $h \neq 0$, and let $y^0 \in Y$. We have $y_i := y^0 + ih = Ax^i + b$ for some $x^i \in X \Rightarrow \{x^i\}$ diverges along with $\{y^i\}$ diverges. Selecting $i_1 < i_2 < ...$ to ensure that $x^{i_s}/\|x^{i_s}\|_2 \to d \in \mathrm{Rec}(X)$ as $i \to \infty$. We have $0 \neq d \in \mathrm{Rec}(X) \Rightarrow Ad \neq 0$

$$\Rightarrow \quad 0 \neq Ad = \lim_s Ax^{i_s}/\|x^i_s\|_2 = \lim_s [y^i - b]/\|x^{i_s}\|_2 = \lim_s \Big[\underbrace{[y^0 - b]/\|x^{i_s}\|_2}_{\to 0} + \frac{i_s}{\|x^{i_s}\|_2} h\Big]$$

$\Rightarrow h$ is positive multiple of $Ad \Rightarrow h \in A\mathrm{Rec}(X)$, Q.E.D.

B. *When taking the sum $X = X_1 + ... + X_L$ of nonempty closed convex sets $X_\ell \subset \mathbf{R}^n$, the sum* $\mathrm{Rec}(X_1) + ... + \mathrm{Rec}(X_L)$ *of their recessive cones clearly belongs to the recessive cone of the closure* $\mathrm{cl}\, X$ *of* X (X is nonempty and convex, but not necessarily is closed). However, $\mathrm{Rec}(X_1) + ... + \mathrm{Rec}(X_L)$ can be a negligibly small part of $\mathrm{Rec}(\mathrm{cl}\, X)$, as is the case when $X_1 = \{[x;t] \in \mathbf{R}^2 : t \geq x^2\}$ and $X_2 = \{[x;t] \in \mathbf{R}^2 : t \leq -x^2\}$, where $\mathrm{Rec}(X_1) = \{[0;t] : t \geq 0\}$, $\mathrm{Rec}(X_2) = \{[0;t] : t \leq 0\}$, whence $\mathrm{Rec}(X_1) + \mathrm{Rec}(X_2) = \{[0;t] : t \in \mathbf{R}\}$, while $X_1 + X_2 = \mathbf{R}^2$ and thus $\mathrm{Rec}(X_1 + X_2) = \mathbf{R}^2$.

However:

Fact II.24 *When $X_1, ..., X_L$ are nonempty closed convex sets in $\mathbf{R}^n$ and at most one of the sets is unbounded, $X := X_1 + ... + X_L$ is a nonempty closed convex set, and* $\mathrm{Rec}(X) = \mathrm{Rec}(X_1) + ... + \mathrm{Rec}(X_L)$.

Indeed, X clearly is nonempty and convex, and is closed by Fact II.21. Assuming that $X_1, ..., X_{L-1}$ are bounded, a diversing sequence $\{x^i \in X\}_i$ is of the sum of $L-1$ bounded, along with X_ℓ, sequences $\{x^i_\ell \in X_\ell\}_i$, $\ell < L$, and diverging sequence $\{y^i \in X_L\}_i \Rightarrow$ asymptotic direction, if any, of $\{x^i\}_i$ is the asymptotic direction of $\{y^i \in X_L\}_i$, implying by Fact II.17.1 that $\mathrm{Rec}(X) \subset \mathrm{Rec}(X_L) = \mathrm{Rec}(X_1) + ... + \mathrm{Rec}(X_L)$. The inverse inclusion $\mathrm{Rec}(X_1) + ... + \mathrm{Rec}(X - L) \subset \mathrm{Rec}(X)$ is stated in **B.**

♠ Situation with recessive cones of affine images and weighted sums somehow improves when assuming the operands to be (V, R)-*sets* defined as follows:

Let $V \subset \mathbf{R}^n$ be a nonempty bounded set in $\mathbf{R}^n$ and $R \subset \mathbf{R}^n$ be a cone. We say that a convex set X is a (V, R)-set, if $X \subset V + R$ and $\{v\} + R \subset X$ for some $v \in V$ (neither one of V, R, X is assumed to be closed).

Example: When R is a cone, every convex subset of ϵ-neighborhood of R is (V, R)-set for properly selected V (e,g., centered at the origin 2ϵ-ball of the norm underlying the neighborhood).

It is immediately seen that (V, R)-sets are well suited for "major part" of our convexity calculus:

Fact II.25

1. *If $X_\ell \subset \mathbf{R}^\ell$, $\ell \leq L$, are (V_ℓ, R_ℓ)-sets, then their direct product $X = X_1 \times ... \times X_L$ is $(V_1 \times ... \times V_L, R_1 \times ... \times R_L)$-set*

2. *If $X \subset \mathbf{R}^n$ is a (V, R)-set and $\mathcal{A}(X) = Ax + b : \mathbf{R}^n \to \mathbf{R}^m$ is an affine mapping, then $\mathcal{A}(X)$ is $(\mathcal{A}(V), AR)$-set.*

3. *If $X_\ell \subset \mathbf{R}^n$, $\ell \leq L$, are (V_ℓ, R_ℓ)-sets and $\lambda_1, ..., \lambda_L$ are reals, the set $X := \lambda_1 X_1 + ... + \lambda_L X_L$ is $(\lambda_1 V_1 + ... + \lambda_L V_L, \lambda_1 R_1 + ... + \lambda_L R_L)$-set.*

Let us make the following observation:

Fact II.26 *Let $X \subset \mathbf{R}^n$ be a (V, R)-set. Then*

$$\mathrm{Rec}(\mathrm{cl}\, X) = \mathrm{cl}\, R.$$

Indeed, under our premise $\overline{X} := \mathrm{cl}\, X$ is nonempty closed and convex (since X is nonempty and convex), and for some $v \in V$ we have $v + R \subset X \Rightarrow R \subset \overline{R} := \mathrm{Rec}(\overline{X})$. All we need is to prove that $\overline{R} = \mathrm{cl}\, R$. To this end it suffices to verify that if $d \in \overline{R}$ is a $\|\cdot\|_2$-unit vector, then $din\mathrm{cl}\, R$. By Fact II.15, there exists a diverging sequence $\{\bar{x}^i \in \overline{X}\}_i$ with asymptotic direction d. As X is dense in $\overline{X}$, we can approximate $\bar{x}_i$ with $x^i \in X$ to convert $\{\bar{x}^i\}_i$ into diverging sequence $\{x^i \in X\}_i$ with the same asymptotic direction d.As X is (V, R)-set,e have $x^i + r^i$ with4 $v^i \in V$ and $r^i \in R$. Since $\{x^i\}_i$ is diverging and V is bounded, the sequence $\{r_i\}_i$ is diverging, and its asymptotic direction is the same d as for $\{x^i\}_i \Rightarrow R \ni r^i/\|r^i\|_2 \to d$ as $i \to \infty \Rightarrow d \in \mathrm{cl}\, R$, Q.E.D.

• Combining Facts II.25, II.26, we get techniques for computing the recessive cones of (the closures of) affine images and weighted sums of (V, R)-sets.

Note: We shall eventually see that when restricting the "calculus of recessive cones" onto polyhedral sets, no difficulties occur.

Nice Topological Properties of Convex Sets

♣ Whenever $X \subset \mathbf{R}^n$, we have

$$\text{int}\, X \subset X \subset \text{cl}\, X$$

In general, the discrepancy between $\text{int}\, X$ and $\text{cl}\, X$ can be pretty large. E.g., let $X \subset \mathbf{R}$ be the set of irrational numbers in $[0,1]$. Then $\text{int}\, X = \emptyset$, $\text{cl}\, X = [0,1]$, so that $\text{int}\, X$ and $\text{cl}\, X$ differ dramatically.

♠ Fortunately, a *convex* set is perfectly well approximated by its closure (and by interior, *if the latter is nonempty*).

Fact II.27 *Let $X \subset \mathbf{R}^n$ be a convex set. Then*
(i) Both $\text{int}\, X$ and $\text{cl}\, X$ are convex
(ii) If $\text{int}\, X$ is nonempty, then $\text{int}\, X$ is dense in $\text{cl}\, X$, density of a set Y in a set X meaning that every point from X can be approximated to whatever high accuracy by points of Y. Formally: *Y is dense in X $\Leftrightarrow$ Every point from X is the limit of a converging sequence of points from Y.*
Moreover,

$$x \in \text{int}\, X,\ y \in \text{cl}\, X \Rightarrow \lambda x + (1-\lambda) y \in \text{int}\, X\ \forall \lambda \in (0,1] \qquad (!)$$

• **Claim (i)**: Let X be convex. Then both $\operatorname{int} X$ and $\operatorname{cl} X$ are convex

Proof. (i) is nearly evident. Indeed, to prove that $\operatorname{int} X$ is convex, note that for every two points $x, y \in \operatorname{int} X$ there exists a common $r > 0$ such that the $\|\cdot\|_2$-balls B_x, B_y of radius r centered at x and y belong to X. Since X is convex, for every $\lambda \in [0,1]$ X contains the set $\lambda B_x + (1-\lambda)B_y$, which clearly is nothing but the ball of the radius r centered at $\lambda x + (1-\lambda)y$. Thus, $\lambda x + (1-\lambda)y \in \operatorname{int} X$ for all $\lambda \in [0,1]$.
Similarly, to prove that $\operatorname{cl} X$ is convex, assume that $x, y \in \operatorname{cl} X$, so that $x = \lim_{i\to\infty} x_i$ and $y = \lim\limits_{i\to\infty} y_i$ for appropriately chosen $x_i, y_i \in X$. Then for $\lambda \in [0,1]$ we have

$$\lambda x + (1-\lambda)y = \lim_{i\to\infty} \underbrace{[\lambda x_i + (1-\lambda)y_i]}_{\in X},$$

so that $\lambda x + (1-\lambda)y \in \operatorname{cl} X$ for all $\lambda \in [0,1]$.

• **Claim (ii)**: Let X be convex and $\operatorname{int} X$ be nonempty. Then $\operatorname{int} X$ is dense in $\operatorname{cl} X$; moreover,

$$x \in \operatorname{int} X,\ y \in \operatorname{cl} X \Rightarrow \lambda x + (1-\lambda)y \in \operatorname{int} X\ \forall \lambda \in (0,1] \qquad (!)$$

Proof. It suffices to prove (!). Indeed, let $\bar{x} \in \operatorname{int} X$ (the latter set is nonempty). Every point $x \in \operatorname{cl} X$ is the limit of the sequence $x_i = \frac{1}{i}\bar{x} + \left(1 - \frac{1}{i}\right)x$. Given (!), all points x_i belong to $\operatorname{int} X$, thus $\operatorname{int} X$ is dense in $\operatorname{cl} X$.
Proof of (!): Let $x \in \operatorname{int} X$, $y \in \operatorname{cl} X$, $\lambda \in (0,1]$. Let us prove that $\lambda x + (1-\lambda)y \in \operatorname{int} X$.
$x \in \operatorname{int} X \Rightarrow \exists r > 0 : B_r(x) := \{y : \|y - x\|_2 \le r\| \subset X$.
$y \in \operatorname{cl} X \Rightarrow y = \lim_{i\to\infty} y_i$ with $y_i \in X$
Let

$$B^i := \lambda B_r(x) + (1-\lambda)y_i = \{z = \underbrace{[\lambda x + (1-\lambda)y_i]}_{z_i} + \lambda h : \|h\|_\infty \le r\} = B_\rho(z_i),\ \rho = \lambda r > 0.$$

$B_r(x) \subset X$, $y_i \in X$, X is convex $\Rightarrow B^i \subset X$.
Since $z_i \to z = \lambda x + (1-\lambda)y$ as $i \to \infty$, the $\|\cdot\|_2$-balls $B^i = B_\rho(z_i)$ for large enough i contain the $\|\cdot\|_2$-ball of radius $\rho/2$ centered at $z \Rightarrow z \in \operatorname{int} X$.

♣ Let X be a convex set. It may happen that $\mathrm{int}\, X = \emptyset$ (e.g., X is a segment in 3D); in this case, interior definitely does not approximate X and $\mathrm{cl}\, X$. What to do?

A natural way to overcome this difficulty is to pass to *relative interior*, which is the interior of X *taken w.r.t. the affine hull* $\mathrm{Aff}(X)$ *of* X rather than w.r.t. $\mathbf{R}^n$. This affine hull, geometrically, is just certain $\mathbf{R}^m$ with $m \leq n$; replacing, if necessary, $\mathbf{R}^n$ with this $\mathbf{R}^m$, we arrive at the situation where $\mathrm{int}\, X$ is nonempty.

Implementation of the outlined idea goes through the following

Definition: [relative interior and relative boundary] Let X be a nonempty convex set and $M = \mathrm{Aff}(X)$. The *relative interior* $\mathrm{rint}\, X$ of X is the set of all points $x \in X$ such that a centered at x ball *in* M of a positive radius is contained in X:

$$\mathrm{rint}\, X = \{x : \exists r > 0 : \{y \in \mathrm{Aff}(X), \|y - x\|_2 \leq r\} \subset X\}.$$

The *relative boundary* $\partial_{\mathrm{r}} X$ of X is, by definition, $\mathrm{cl}\, X \backslash \mathrm{rint}\, X$.

Note: $\mathrm{Aff}(X)$ is closed, so that $\mathrm{rint}\, X \subset X \subset \mathrm{cl}\, X \subset \mathrm{Aff}(X)$ and $\partial_{\mathrm{r}} X \subset \mathrm{Aff}(X)$.

Fact II.28 *Let $X \subset \mathbf{R}^n$ be a nonempty convex set. Then $\mathrm{rint}\, X \neq \emptyset$.*

Thus, replacing, if necessary, the original "universe" $\mathbf{R}^n$ with a smaller *geometrically similar* universe, we can reduce investigating an *arbitrary* nonempty convex set X to the case where this set has a nonempty interior (which is nothing but the *relative* interior of X). In particular, our results for the "full-dimensional" case imply

Fact II.29 *For a nonempty convex set X, both $\mathrm{rint}\, X$ and $\mathrm{cl}\, X$ are convex sets such that*

$$\emptyset \neq \mathrm{rint}\, X \subset X \subset \mathrm{cl}\, X \subset \mathrm{Aff}(X)$$

and $\mathrm{rint}\, X$ is dense in $\mathrm{cl}\, X$ (implying, in particular, that $\mathrm{cl}\, \mathrm{rint}\, X = \mathrm{cl}\, X$). Moreover, whenever $x \in \mathrm{rint}\, X$, $y \in \mathrm{cl}\, X$ and $\lambda \in (0, 1]$, one has

$$\lambda x + (1 - \lambda) y \in \mathrm{rint}\, X.$$

Fact II.28 *The relative interior of a nonempty convex set X is nonempty.*

Proof. By Fact II.10. Let m be the dimension of the linear space parallel to $M = \mathrm{Aff}(X)$, When $m = 0$, X is a singleton, and clearly $\mathrm{rint}\,(X) = X \neq \emptyset$. Now let $m > 0$. By Fact II.10, M admits an affine basis $x^0, x^1, ..., x^m$ composed of vectors from X, so that

$$M = \left\{ x : \exists \lambda_0, ..., \lambda_m : \begin{array}{rcl} \sum_{\ell=0}^m \lambda_\ell x^\ell & = & x \\ \sum_{\ell=0}^m \lambda_\ell & = & 1 \end{array} \right\}.$$

Solution to the system of linear equalities

$$\begin{array}{rcl} \sum_{\ell=0}^m \lambda_\ell x^\ell & = & x \\ \sum_{\ell=0}^m \lambda_\ell & = & 1 \end{array} \qquad (S)$$

in variables λ exists when $x \in M$ and is unique (since $x^0, ..., x^m$ are affinely independent), that is, the vectors $y^\ell = [x^\ell; 1]$, $0 \leq \ell \leq m$, are linearly independent. Therefore by Linear Algebra the solution $\lambda(x)$ to the system of equations

$$\sum_{\ell=0}^m \lambda_\ell y^\ell = [x; 1]$$

(which exists whenever $x \in M$) is an *affine*, and thus continuous, function of $x \in M$
$\Rightarrow$ with $\bar{x} = \frac{1}{m+1} \sum_{\ell=0}^m x^\ell \in M$, all close enough to $\bar{x}$ points from M have their barycentric coordinates positive, and thus belong to X (due to X being convex and $x^\ell \in X$, $\ell \leq L$) $\Rightarrow$ All close enough to $\bar{x}$ points from M belong to X $\Rightarrow$ $\bar{x} \in \mathrm{rint}\, X$ $\Rightarrow$ $\mathrm{rint}\, X \neq \emptyset$. $\bar{x} \in \mathrm{rint}\, X$, Q.E.D.

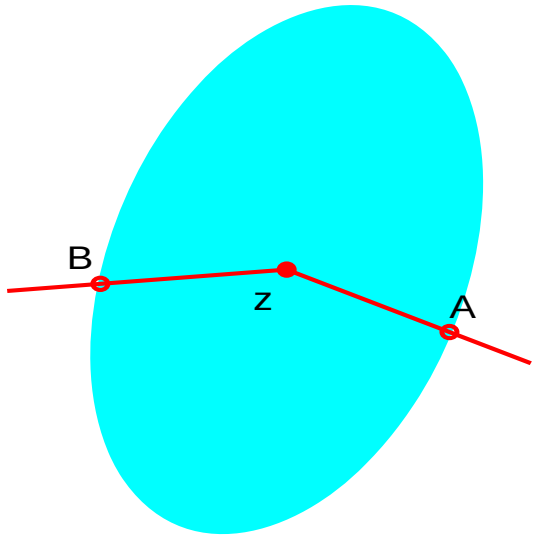
♣ In general, the "gap" between $\operatorname{rint} X$ and $\operatorname{cl} X$ can be dramatic. For example, when X is the set of rational numbers from $[0,1]$, we have $\operatorname{rint} X = \emptyset$, $\operatorname{cl} X = [0,1]$, $\partial_{\mathrm{r}} X = [0.1]$. In contrast, for a *convex* set X, passing from $\operatorname{rint} X$ to cl requires "tiny adjustment" – taking *radial closure*.

♠ Let $X \subset \mathbf{R}^n$ be a nonempty convex set which is not a singleton, let $M = \operatorname{Aff}(X)$, and let $z \in \operatorname{int} X$. Given a nonzero vector e belonging to the linear space L parallel to m, the ray $R_e = \{x + te : t \geq 0\}$ belongs to M, and the quantity

$$t_e = \sup\{t : z + te \in X\}$$

so that $0 < t_e \leq +\infty$ and the points $x + te$
— belong to X when $0 \leq t < t_e$
— do not belong to X when $t > t_e$.
Whether the point $x + t_e e$ belongs or does not belong to X, it depends on e and X
• *To pass from X to* $\operatorname{rint} X$*, it suffices to remove from* the intersections of X with the rays R_e those of the points $x + t_e e$ (for all e with $t_e < \infty$) which belong to X
• *To pass from X to* $\operatorname{cl} X$*, it suffices to add to* the intersections of X with the rays R_e those of the points $x + t_e e$ (for all e with $t_e < \infty$) which do not belong to X

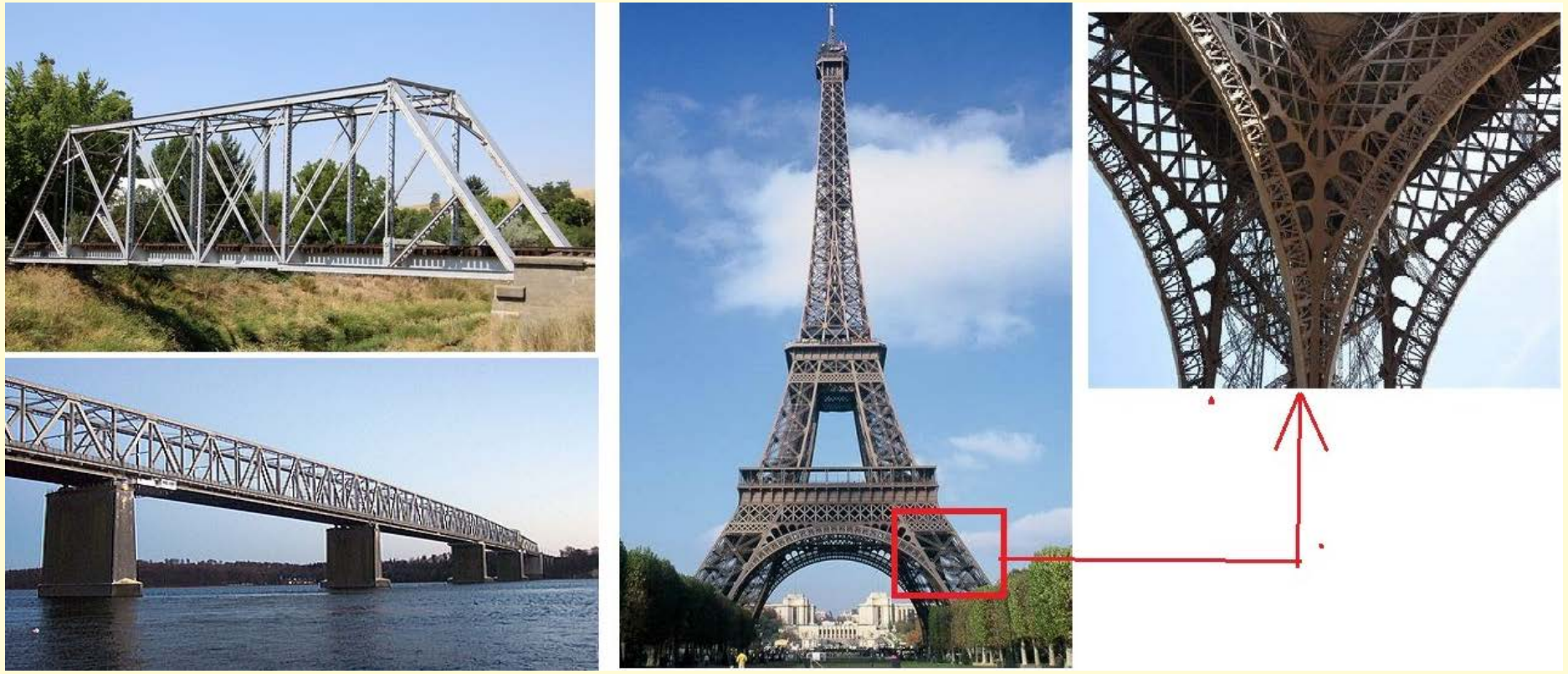
For example, to pass from X to $\mathrm{cl}\, X$, we look at all rays in $\mathrm{Aff}(X)$ emanating from z and add to X all "missing" – not contained in X from the very beginning – boundary points, like A and B, of the intersections of rays with X.



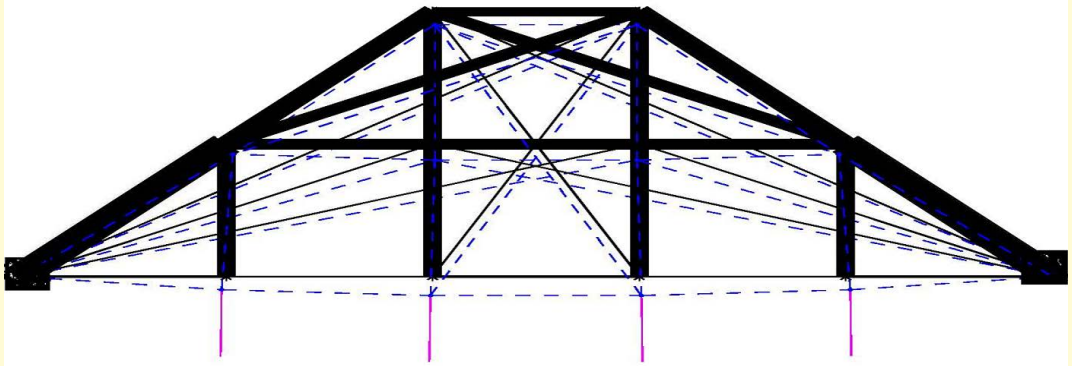
Embedded story: Truss Topology Design, I

♣ In the sequel, we from time to time speak about a particular application of Convex Optimization – **T***russ* **T***opology* **D***esign* allowing to illustrate and "visualize" many (by far not all!) of the abstract constructions we deal with.
We start with building the TTD model.

♠ **Truss** is a mechanical construction composed of thin elastic *bars* connected with each other at *nodes*



Trusses



Static equilibrium of railroad bridge under 4-force load

• When affected by a *load* – collection of external forces acting at the nodes – the construction deformates until the reaction forces caused by extensions/contractions of bars compensate the external load ("static equilibrium.")

• At equilibrium, the truss capacitates certain potential energy – the *compliance*. The compliance measures rigidity of the truss w.r.t. the load: the smaller compliance, the better the construction withstands the load.

♠ **Mathematical Model** of the TTD problem is as follows. **Given are:**

• The set of *tentative nodes* – a 2D (*planar truss*) or 3D (*spatial truss*) finite grid of points, where the would-be bars can be connected with each. Some of the nodes in the grid are *fixed* by boundary conditions and cannot move, other *(free nodes)* can move in 2D, resp., 3D.

♢ The nodal grid specifies the space $\mathcal{V}$ of *virtual displacements* – the linear space of block-vectors. The blocks are indexed by free nodes and are virtual displacements of the nodes in the embedding "physical" space $\mathbf{R}^d$ ($d = 2$ for planar, and $d = 3$ for spatial trusses). We set $M = \dim \mathcal{V}$.

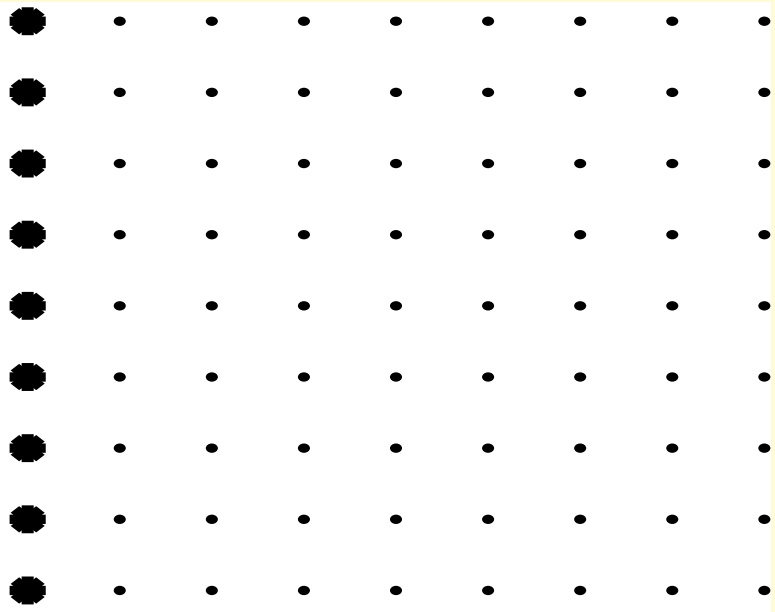
♢ The set of N *tentative bars* – pairs of distinct nodes allowed to be linked by bars

♢ The set of K *loading scenarios* x^k - vectors from $\mathcal{V}$ with blocks representing physical external forces acting at the nodes

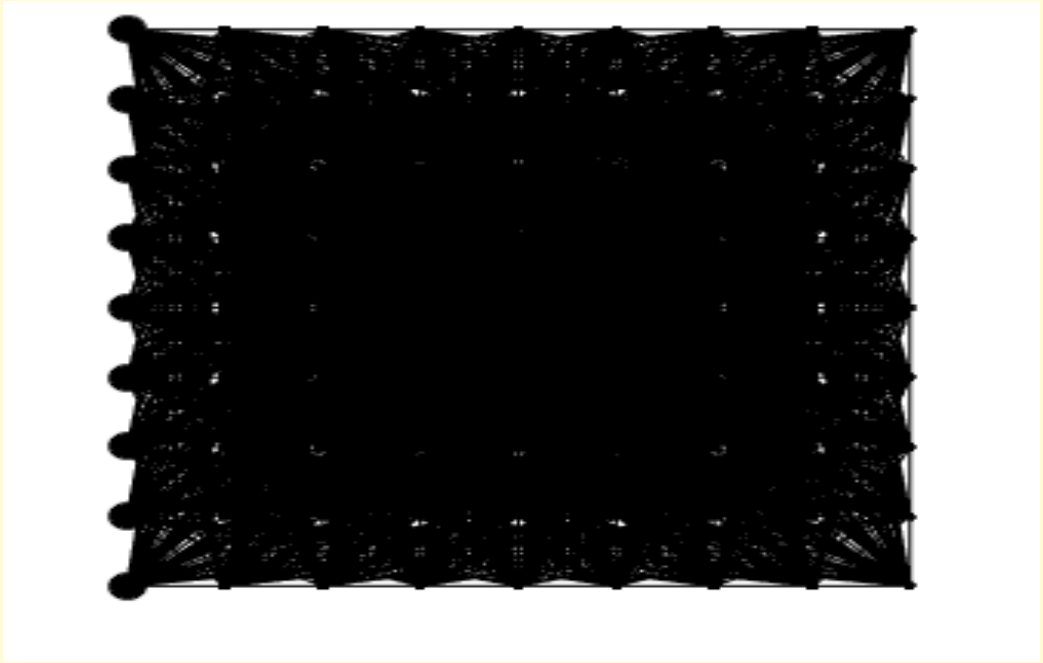
♢ The total volume W of bars.

• What we want is to minimize the worst, over the loading scenarios. compliance of the truss.

♠ Our "guinea pig" will be *design of planar console* with the "given are" as follows:



9×9 nodal grid
•: fixed nodes



3024 tentative bars

• 9×9 grid of tentative nodes with the 9 most left nodes fixed (they are "in the wall") and the remaining 72 nodes free ($m := \dim \mathcal{V} = 2 \times 72 = 144$)
• we allow for every pair of nodes with at least one node free to be linked by a bar (resulting in $N = 3024$ tentative bars)
• there is just one loading scenario, with unit external force applied at the middle node in the most right column of nodes and "looking down"
• total truss volume is $W = 1000$

TTD Model

♠ In *linearly elastic model of a truss* the TTD problem is as follows:

• "Ground structure" (the grid of tentative nodes plus boundary conditions plus the list of tentative bars) specifies vectors $\mathfrak{b}_i \in \mathcal{V} = \mathbf{R}^M$ indexed by tentative bars $i = 1, 2, ..., N$; augmented with *a truss* – a collection $t = \{t_i \geq 0, i \leq N\}$ of volumes of N tentative bars. Vectors $\mathfrak{b}_i$ specify the *stiffness matrix*

$$A(t) = \sum_{i=1}^{N} t_i \mathfrak{b}_i \mathfrak{b}_i^T$$

of truss t. When the "physical displacements" of the nodes form a vector $v \in \mathcal{V}$, the collection of reaction forces caused by deformation of the truss is $-A(t)v$, and the potential energy capacitated in the truss is $\frac{1}{2} v^T A(t) v$

⇒ *The equilibrium displacement* v_f *of truss* t *under external load* $f \in \mathcal{V}$ *satisfies*

$$A(t)v_f = f, \qquad (*)$$

and the compliance is

$$\text{Compl}(t, f) = \frac{1}{2} v_f^T A(t) v_f = \frac{1}{2} f^T v_f$$

Note: When $(*)$ has no solutions, no equilibrium exists – the truss is crushed by load f. $(*)$ may have multiple solutions, meaning that the equilibrium displacement if not uniquely defined; however, wee shall see that the compliance is well defined – all solutions to $(*)$ result in the same value of the compliance.

♠ Here is the justification of the last "Note:"

Fact II.30 *Let $f \in \mathcal{V}$ and $t \in \mathbf{R}^N_+$ be given, and let*

$$F(v) = f^T v - \frac{1}{2} v^T A v : \mathcal{V} \to \mathbf{R} \qquad [A = A(t) = \textstyle\sum_i t_i \mathfrak{b}_i \mathfrak{b}_i^T]$$

(i) Maximizers of F over $v \in \mathcal{V}$ are exactly the solutions to the equation

$$Av = f; \qquad (*)$$

and when this equation is unsolvable $\sup_v F(v) = +\infty$;
(ii) When $(*)$ is solvable and v_f is a solution, it holds

$$\max_v F = F(v_f) = \frac{1}{2} v_f^T A v_f = \frac{1}{2} f^T v_f;$$

As a result, the compliance $\mathrm{Compl}(t, f)$ of truss t under the load f is well defined iff f is bounded from above, in which case $\mathrm{Compl}(t, f) = \max_v F(x)$
(ii) A real τ satisfies $\tau \geq \mathrm{Compl}(t, f)$ if and only if the matrix

$$\left[\begin{array}{c|c} A(t) & f \\ \hline f^T & 2\tau \end{array}\right]$$

is positive semidefinite, As a result, the TTD problem with loading scenarios $f_1, ..., f_K$ can be posed as

$$\min_{\tau,t} \left\{ \tau : \left[\begin{array}{c|c} A(t) & f_k \\ \hline f_k^T & 2\tau \end{array}\right] \succeq 0, \leq l, t \geq 0, \sum_i t_i = W \right\}$$

Proof. Given $f \in \mathcal{V}$ and $t \in \mathbf{R}^N_+$ and setting $A = A(T) = \sum -it_i \mathfrak{b}_i \mathfrak{b}_i^T$, let $A = U\mathrm{Diag}\{\lambda_1, .., \lambda_M\}U^T$ be the eigenvalue decomposition of A. Note that $A \succeq 0$ due to $t \geq 0 \Rightarrow \lambda_j \geq 0 \forall j$.
Substituting $v = Uu$ and setting $\phi = Uf$, equilibrium equation $Av = f$ becomes $U\mathrm{Diag}\{\lambda\}U^TUu = U\phi$, that is, it is nothing but the system

$$\lambda_j u_j = \phi_j,\ 1 \leq j \leq M. \tag{!}$$

We also have

$$\Phi(u) := F(Uu) = \sum_j [\phi_j u_j - \frac{1}{2}\lambda_j u_j^2] \qquad [F(v) = f^Tu - \tfrac{1}{2}v^TAv]$$

We see that F (or, which is the same, Φ is bounded from above iff $\phi_j = 0$ for all j with $\lambda_j = 0$. In this case, by high school algebra, the maximizers of Φ are exactly the solutions to (!), and

$$\max_u \Phi(u) = \frac{1}{2}\sum_j \lambda_j \phi_j^2.$$

$\Rightarrow$ F is bounded from above iff the equation $Av = f$ is solvable, and when it is the case, the maximizers of F are exactly the solutions v_f to this equation, and

$$\max_v F(v) = \frac{1}{2}v_f^TAv_f = \frac{1}{2}f^Tv_f = \mathrm{Compl}(t, f),$$

as claimed in (i) and (ii).
To verify (iii) note that by (i-ii), for a real τ the relation $\tau \geq \mathrm{Compl}(t, f)$ is the same as $\tau \geq F(v)$ for all v, that is, as

$$2\tau - 2f^Tv + v^TAv \geq 0\ \forall v. \tag{!!}$$

Substituting $v = z/s$ with $s \neq 0$, (!!) is the same as

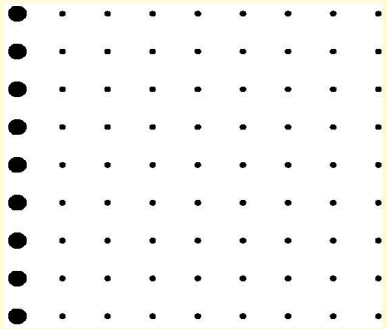
$$2\tau s^2 - 2sf^Tz + z^TAz \geq 0\ \forall (s \neq 0, z),$$

which by continuity and replacing z with $-z$ is equivalent to

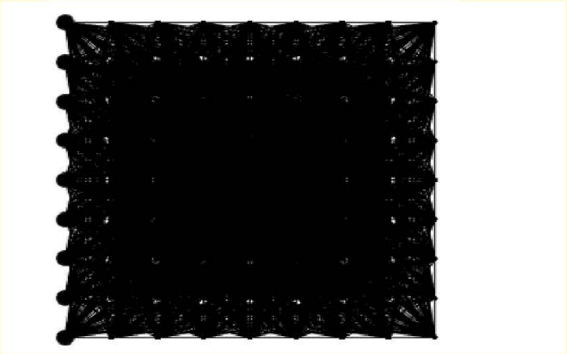
$$2\tau s^2 + 2sf^Tz + z^TAz\ \forall s, z,$$

that is, nothing but $\left[\begin{array}{c|c} A & f \\ \hline f^T & 2\tau \end{array}\right] \succeq 0$, Q.E.D.

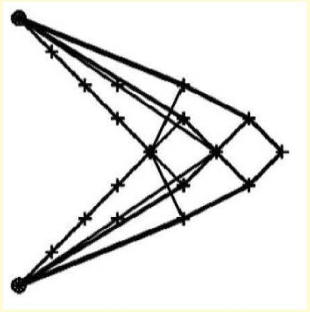
♠ How it works:



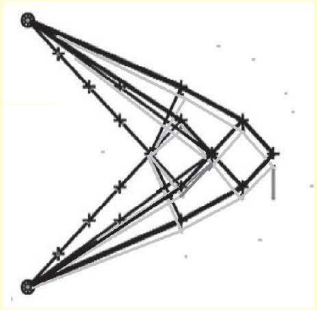
a) 9×9 nodal grid
•: fixed nodes



b) 3024 tentative bars



c) optimal truss, 38 bars
compliance 0.1914



d) displacement under
load of interest

Optimal console. d), gray: bars after deformation and the external force (vertical segment starting at the most right node)

♣ Typically, in TTD with rich list of tentative bars, just a small fraction of them get positive volumes in the optimal truss $\Rightarrow$ TTD is not merely about optimal bar sizing, it indeed is about Topology Design!

Main Theorems on Convex Sets, I: Caratheodory Theorem

♣ **Definition** [dimension of a nonempty set]
• Dimension $\dim L$ of a linear subspace L of $\mathbf{R}^m$ is the linear dimension of L the common cardinality of linear bases of L.
• Dimension $\dim M$ of an affine subspace M in $\mathbf{R}^n$ is the just defined dimension of the parallel to M linear subspace – the common cardinality of all affine bases in M *minus 1*
• Dimension $\dim X$ of a nonempty subset X of $\mathbf{R}^n$ is the just defined dimension of the affine span $\mathrm{Aff}(X)$ of X — the maximum number of affinely independent vectors from X *minus 1*.

Note: Some subsets of $\mathbf{R}^n$ are in the scopes of several "branches" of this definition; for these sets, all applicable "branches" yield the same value of the dimension.

Examples: • The dimension of a singleton is 0.
• The dimension of $\mathbf{R}^n$ is n.
• The affine dimension of an affine subspace $M = \{x : Ax = b\}$ is $n - \mathrm{Rank}(A)$.
• The dimension of the triangle $\mathrm{Conv}\{x^1, x^2, x^3\}$ with affinely independent (or, which is the same, not belonging to a common line) x^1, x^2, x^3 is 2.

♠ **Caratheodory Theorem** *Let $\emptyset \neq X \subset \mathbf{R}^n$. Then every point $x \in \mathrm{Conv}(X)$ is a convex combination of at most $\dim(X)+1$ points of X.*

Proof

1^o. We should prove that if x is a convex combination of finitely many points $x^1, ..., x^k$ of X, then x is a convex combination of at most $m+1$ of these points, where $m = \dim(X)$. The claim is trivial in the case of $m = 0$ where X is a singleton. Assuming $m > 0$ and replacing, if necessary, $\mathbf{R}^n$ with $\mathrm{Aff}(X)$, it suffices to consider the case of $m = n$.

2^o. Consider a representation of x as a convex combination of $x^1, ..., x^k$ *with minimum possible number of nonzero coefficients*; it suffices to prove that this number is $\leq n+1$. Let, on the contrary, the "minimum representation" of x have $p > n+1$ positive coefficients. W.l.o.g.sume that the first p coefficients are positive, so that the representation is

$$x = \sum_{i=1}^{p} \lambda_i x^i \qquad [\lambda_i > 0, \sum_i \lambda_i = 1]$$

.

3^o. Consider the homogeneous system of linear equations in p variables δ_i

$$\begin{cases} (a) & \sum_{i=1}^{p} \delta_i x^i = 0 & [n \text{ linear equations}] \\ (b) & \sum_i \delta_i = 0 & [\text{single linear equation}] \end{cases}$$

Since $p > n+1$, this system has a nontrivial solution δ. Observe that for every $t \geq 0$ one has

$$x = \sum_{i=1}^{p} \underbrace{[\lambda_i + t\delta_i]}_{\lambda_i(t)} x^i \ \& \ \sum_i \lambda_i(t) = 1.$$

♢ When $t = 0$, all coefficients $\lambda_i(t)$ are nonnegative.

♢ $\sum_i \delta_0 = 0$ and $\delta \neq 0$ $\Rightarrow$ some of δ_i are negative $\Rightarrow$ for some i, $\lambda_i(t) \to -\infty$ as $t \to \infty$

$\Rightarrow$ there exists the largest $t = t^* \geq 0$ for which $\lambda_i(t) \geq 0$ for all i; by maximality, some of $\lambda_i(t^*)$ are zeros

$\Rightarrow$ In the representation $x = \sum_{i=1}^{p} \lambda_i(t^*) x^i$ as a convex combination of x^i the number of nonzero terms in the right hand side is $< p$, contradicting to minimality of the representation $\sum_{i=1}^{p} \lambda_i x^i$ of x as a convex combination of x^i.

Note: Given m, n with $1 \le m \le n$, we can point out $m+1$ affinely independent points $x^1, ..., x^{m+1}$ in $\mathbf{R}^n$. We have $\dim\{x^1, ..., x^{m+1}\} = m$ and the vector $\bar{x} = \frac{1}{m+1}\sum_{i=1}^{m+1} x^i$ has exactly one representation as a convex combination of x^i, and this representation involves all $m+1$ of the x^i's

$\Rightarrow$ *Caratheodory Theorem is sharp*

♠ **Caratheodory Theorem, Conic version** *Let $\emptyset \neq X \subset \mathbf{R}^n$. Then every vector from the conic hull $\mathrm{Cone}(X)$ of X is a conic combination of at most $m = \dim X$ vectors from X.*

Proof. The claim is trivially true when X is a singleton, or, which is the same, when $m = 0$. Assuming $m \ge 1$, $\mathrm{Aff}(\mathrm{Cone}(X))$ is m-dimensional linear subs[pace of $\mathbf{R}^n$, and, similarly to the case of the plain Caratheodory Theorem, we lose nothing when assuming that this linear space is the entire $\mathbf{R}^n$, i.e. that $m = n$.

Given $x \in \mathrm{Cone}(X)$, consider the minimal in the number of terms representation $x = \sum_{i=1}^{p} \lambda_i x^i$ of x as a conic combination of vectors from X. We should prove that $p \le n$. Assuming the opposite, consider the homogeneous system of linear equations

$$\sum_{i=1}^{p} \delta_i x^i = 0$$

in variables δ. Since $p > n$, this system has a nontrivial solution δ; replacing, if necessary, δ with $-\delta$, we may further assume that some of δ_i are negative.

For every $t \ge 0$, we have $x = \sum_i \lambda_i(t) x_i$ with $\lambda_i(t) = \lambda_i + t\delta_i$. Same as in the proof of Caratheodory Theorem, $\lambda_i(0) \ge 0$, and for large t some of $\lambda_i(t)$ are negative, implying that there exists the largest $t = t^*$ for which $\lambda_i(t) \ge 0$ for all i. By maximality of t^*, some of $\lambda_i(t^*)$ are zero, implying that the number of nonzero terms in the representation $x = \sum_{i=1}^{p} \lambda_i * t^*) x^i$ of x as a conic combination of x^i is less than p, contradicting the origin of p.

Note: Similarly to the plain Caratheodory Theorem, Caratheodory Theorem in Conic form is sharp.

Illustration I: Blending teas. The story goes as follows:
Supermarkets sell 99 different herbal teas; every one of them is certain blend of 26 herbs A,...,Z. In spite of such a variety of marketed blends, John is not satisfied with any one of them; the only herbal tea he likes is their mixture, in the proportion

$$1 : 2 : 3 : \ldots : 98 : 99$$

Once it occurred to John that in order to prepare his favorite tea, there is no necessity to buy all 99 marketed blends; a smaller number of them will do. With some arithmetics, John found a combination of 66 marketed blends which still allows to prepare his tea. Do you believe John's result can be improved?

Answer: *In fact, just 26 properly selected market bends are enough.*
Indeed, let us represent a blend by its unit weight portion, say, 1g. Such a portion can be identified with 26-dimensional vector $x = [x_1; ...; x_{26}]$ with nonnegative entries summing up to 1, where x_i is the weight, in grams, of herb #i in the portion. Clearly, we have

$$x \in \mathbf{R}^{26}_+ \ \& \ \sum_i x_i = 1.$$

When mixing market blends $x^1, x^2, ..., x^{99}$ to get unit weight portion x of mixture, we take $\lambda_i \geq 0$ grams of market blend x^i, $i = 1, ..., 99$, and mix them together, that is,

$$x = \sum_i \lambda_i x_i.$$

Looking at the weights of both sides, we get $\sum_i \lambda_i = 1$.
The bottom line: blend x can be obtained by mixing market blends $x^1, ..., x^{99}$ if and only if $x \in \text{Conv}\{x^1, ..., x^{99}\}$.
By Caratheodory Theorem, every blend which can be obtained my mixing market blends can be obtained by mixing $m + 1$ of them, where m is the affine dimension of the affine span of $x^1, ..., x^{99}$. In our case, this span belongs to the 25-dimensional affine plane

$$\{x \in \mathbf{R}^{26} : \sum_i x_i = 1\}$$

that is, $m \leq 25$.

♣ **Illustration II: Matrix game, or Caratheodory Theorem in casino.** The story goes as follows. There are two players, A and B. Player A selects her move from $\{1,2,...,m\}$, player B selects her move from $\{1,...,n\}$. With selected moves i,j, A pays to B the sum M_{ij}. The rues of the game – m,n and the matrix $M=[M_{ij}]_{\substack{i\leq m\\ j\leq n}}$ are known to both players, and they select their moves simultaneously, not knowing the selection of the adversary.
A is interested to minimize the money she pays, B is interested to maximize the money she gets..
Question: What should the selections of rational players be?
Answer is easy when M has a *saddle point* – there is a cell i_*,j_* such that the entry M_{i_*,j_*} is minimal in its column and maximal in its row. Such a point is an *equilibrium*: A has no incentive to deviate from i_* when B sticks to j_*, and B has no incentive to deviate from j_* when A sticks to i_* – these deviations cannot improve their outcomes. Thus, saddle point is a pair of player's moves such that no player can improve her position by her unilateral action.

However: Matrices with saddle points are "rare commodity." What to do when the matrix M has no saddle points?

An instructive answer was given by John von Neumann and Oskar Morgenstern in their groundbreaking text *Theory of Games and Economic Behavior* (1944) – they proposed to look at the situation where the game is played repeatedly, round by round, and the players use *mixed strategies* – select their moves in consecutive rounds at random, in an iid fashion, from respective distributions $x \in \Delta_m = \{x \in \mathbf{R}^m_+ : \sum_i x_i = 1\}$, and $y \in \Delta_n = \{y \in \mathbf{R}^n_+ : \sum_j y_j = 1\}$. What matters for the players in the long run, are their *expected payments/rewards per round, that is, the quantity*

$$x^T M y;$$

player A is interested to minimize this quantity by selecting $x \in \Delta_m$, and player B is interested to maximize the quantity in $y \in \Delta_n$. It turns out that *In mixed strategies, there always is an equilibrium: there exist $x_* \in \Delta_m$ and $y_* \in \Delta_n$ such that*

$$\forall (x \in \Delta_m, y \in \Delta_n) : x^T M y_* \geq x_*^T M y_* \geq x_*^T M y.$$

Solving Matrix game in mixed strategies is quite meaningful in many applications, including military ones; however, other things being equal, we would prefer the mixed strategies to be as sparse as possible – the less nonzero entries, the better.

♠ Caratheodory Theorem implies the following nice result: *There always exists a pair of optimal mixed strategies x_*, y_* with no more than* Rank $(M)+1$ *nonzero entries each.* Indeed, Linear Algebra teaches us that an $m \times n$ matrix M of rank r can be represented as

$$M = L^T R,\ L \in \mathbf{R}^{r\times m}, R \in \mathbf{R}^{r\times n}.$$

It follows that the expected payment/reward of the players, their mixed strategies being x, y, is

$$x^T M y = [Lx]^T [Ry],$$

meaning that what matters for the players, are not their mixed strategies x and y *per se*, but the linear images Lx, Ry, living in $\mathbf{R}^r$, of these strategies. In particular, whenever (x_*, y_*) is a solution to the game in mixed strategies, so is every other pair $(\widehat{x}, \widehat{y})$ of mixed strategies, provided that $L\widehat{x} = Lx_*$ and $R\widehat{y} = Ry_*$. Now, Lx_* is a vector from $\mathbf{R}^r$ which is a convex combination of columns of L; by Caratheodory, we can get the same vector Lx_* by taking convex combination of at most $r+1$ columns of L, that is, can find a mixed strategy $\widehat{x}$ with at most $r+1$ nonzero entries such that $L\widehat{x} = Lx_*$. Similarly, we can find a mixed strategy $\widehat{y}$ with at most $r+1$ nonzero entries such that $R\widehat{y} = Ry_*$, thus obtaining a sparse solution $(\widehat{x}.\widehat{y})$ to our game.

♣ **Caratheodory Theorem in War on terror.** Consider the following *Colonel's Blotto game:*

• There are L locations, G good guys and B bad guys. Every day good guys distribute themselves between the locations, in teams of at most g guys each, to prevent attacks of bad guys, and bad guys distribute themselves between the locations, in teams of at most b guys each, to carry out terror attacks.

• The loss from an attack of a team of $\beta \leq b$ bad guys on a location ℓ defended by a team of $\gamma \leq g$ good guys is $M^\ell_{\gamma\beta}$, where M^ℓ, $\ell \leq L$, are given $(g+1)\times(b+1)$ matrices. The good guys want to reduce the total loss, the bad guys want to increase it.

What are rational policies of good and bad guys ?

♠ A pure strategy of good guys can be identified with an ordered collection $\xi = (\xi_1, ..., \xi_L)$ of integers from the range $\{0, 1, ..., g\}$ summing up to G (ξ_ℓ is the size of the good team in location ℓ). Similarly, a pure strategy of bad guys can be identified with an ordered collection $\eta = (\eta_1, .., \eta_L)$ of integers from the range $\{0, 1, ..., b\}$ summing up to B.
Denoting by $\mathcal{X}, \mathcal{Y}$ the sets of pure strategies of good and of bad guys and introducing matrix $\mathcal{M}$ with rows indexed by $\xi \in \mathcal{X}$, columns indexed by $\eta \in \mathcal{Y}$ and entries

$$\mathcal{M}_{\xi,\eta} = \sum_{\ell=1}^{L} M^\ell_{\xi_\ell,\eta_\ell},$$

the total loss of good guys when they are using pure strategy ξ, and their adversaries – pure strategy η, is $\mathcal{M}_{\xi,\eta}$
⇒ *Rational behaviour of players is to use mixed strategies forming a saddle point of the matrix game with matrix* $\mathcal{M}$

However: What are the sizes of $\mathcal{M}$, that is, the cardinalities of $\mathcal{X}$ and $\mathcal{Y}$? Here are some answers

L	$G=B$	$g=b$	Card($\mathcal{X}$)	Card($\mathcal{Y}$)
20	40	2	1	1
20	40	3	12,049,586,631	12,049,586,631
20	40	4	5,966,636,799,745	5,966,636,799,745
20	40	5	81,987,009,993,775	81,987,009,993,775
20	40	6	293,752,173,960,574	293,752,173,960,574
20	40	7	575,564,255,892,036	575,564,255,892,036
20	40	8	835,252,578,607,640	835,252,578,607,640
20	40	9	1,033,320,390,014,830	1,033,320,390,014,830

However: M^ℓ are $(g+1)\times(b+1)$ matrices; setting $r=\min[g,b]+1$, we can represent M^ℓ as $M^\ell = A_\ell^T B_\ell$ with $r\times(g+1)$ matrix A_ℓ and $r\times(b+1)$ matrix B_ℓ $\Rightarrow$

$$M^\ell_{pq} = \mathrm{Col}_p^T[A_\ell]\mathrm{Col}_q[B_\ell],\ 0\le p\le g, 0\le q\le b$$
$$\Rightarrow \mathcal{M}_{\xi,\eta} = \sum_{\ell=1}^L M^\ell_{\xi_\ell,\eta_\ell} = \underbrace{\left[\mathrm{Col}_{\xi_1}[A_1];...;\mathrm{Col}_{\xi_L}[A_L]\right]^T}_{L_\xi^T}\underbrace{[\mathrm{Col}_{\eta_1}[B_1];...;\mathrm{Col}_{\eta_L}[B_L]]}_{R_\eta}$$
$$\Rightarrow \mathcal{M} = L^T R,\ \ L\in\mathbf{R}^{[rL]\times\mathrm{Card}(\mathcal{X})},\ R\in\mathbf{R}^{[rL]\times\mathrm{Card}(\mathcal{Y})}$$

Thus, *while the row and column sizes of $\mathcal{M}$ can be astronomically large, the rank of $\mathcal{M}$ is quite moderate – at most* $\min[g,b]L+L$ $\Rightarrow$ *By Caratheodory, good and bad guys have quite sparse equilibrium mixed strategies – those which are mixtures of at most* $\min[g,b]L+L+1$ *pure strategies.*

Note: On a closest inspection, we can find *efficiently* both the pure strategies participating in the equilibrium, and the probabilities at which these strategies should be used.

Application: Whether the convex hull of a closed set is closed?

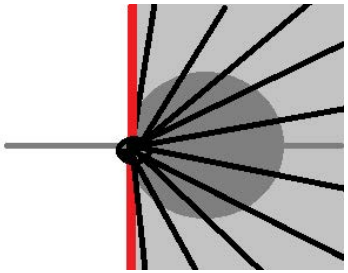
♠ The convex hull of a closed set not necessarily is closed – look what happens when the set is a line augmented by a single point not on the line.

However: *When $X \subset \mathbf{R}^n$ is closed and bounded,* Conv(X) *is closed* (and of course is bounded) $\Rightarrow$

Fact II.31 *The convex hull of a compact set is compact.*

Indeed, let X be compact and $\{x^i \in \mathrm{Conv}(X)\}_i$ be a converging sequence; we should prove that the limit $\bar{x}$ of the sequence belongs to Conv(X). By Caratheodory Theorem, we have $x^i = \sum_{j=1}^{n+1} \lambda_i^j x_i^j$ with $\lambda_j^i \geq 0, \sum_j \lambda_j^i = 1$,and $x_i^j \in X$. By compactness of $[0,1]$ and X, passing to a subsequence, we may assume that the $2(n+1)$ sequences $\{\lambda_j^i\}_i$, $\{x_i^i\}_i$ converge: as $i \to \infty$, we have $\lambda_j^i \to \lambda_j$, $x_i^j \to x^j$, $j \leq n+1$. Note that $x^j \in X$ since X is compact and therefore closed. We clearly have $\lambda_j \geq 0$, $\sum_j \lambda_j = 1$, and $x^i = \sum_j \lambda_j^i x_i^j \to \sum_j \lambda_j x^j$ as $i \to \infty$ $\Rightarrow$ $\bar{x} = \sum_j \lambda_j x^j \in \mathrm{Conv}(X)$.

Note: In contrast to convex hull, the conic hull of a compact set X not necessarily is closed, even when X is convex – look what happens if X is the circle of radius 1 centered at the point $[1;0]$ in $\mathbf{R}^2$.:



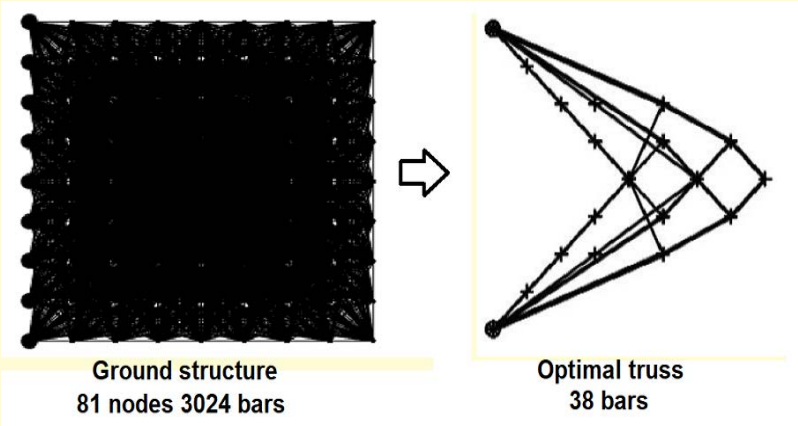
The "conic hull" analogy of the above observation is as follows:

Fact II.32 *Let $X \subset \mathbf{R}^n$ be a nonempty compact set which can be "separated from the origin," meaning that there exists a linear form $f^T x$ on $\mathbf{R}^n$ with positive minimum α over $x \in X$. Then* Cone (X) *is a closed cone.*

Cone (X) always is a cone; all we need is to prove its closedness. Let points $x^i \in$ Cone (X) converge to $\bar{x}$ as $i \to \infty$; we should prove that $\bar{x} \in$ Cone (X). By Caratheodory Theorem in conic form, we have $x^i = \sum_{j=1}^{n} \lambda_i^j x_i^j$ with $\lambda_i^j \geq 0$ and $x_i^j \in X$. We have $f^T \bar{x} = \lim_{i\to\infty} f^T x^i \Rightarrow$ the sequence $\{\sum_j \lambda_i^j f^T x_i^j\}_i$ is bounded. The terms in this sequence are $\geq \alpha \sum_j \lambda_i^j \geq 0$, implying, due to $\alpha > 0$ and $\lambda_i^j \geq 0$, that the sequences $\{\lambda_i^j\}_i$, $j \leq n$, are bounded. With this in mind and since X is compact, we can pass to a subsequence of values of i to ensure that the sequences $\{\lambda_i^j\}_i$ and $\{x_j^j\}_i$, $1 \leq j \leq n$, converge. The latter, by the same argument as in the convex hull case, implies that $\bar{x} \in$ Cone (X).

Embedded story: Truss Topology Design, II

♣ **Question:** In TTD with totally M degrees of freedom of the nodes and K loading scenarios, how many actual bars there should be in optimal truss?



$$M = 144, N = 3024, K = 1$$

Answer: $KM + 1$ bars are enough.

Question: In TTD with totally M degrees of freedom of the nodes and K loading scenarios, how many actual bars there should be in optimal truss?

Answer: $KM+1$ bars are enough.

Indeed, let t_i be the bar volumes in an optimal truss, and v^k be the equilibrium displacement under k-th loading scenario f^k, so that

$$\sum_i \underbrace{\frac{t_i}{W}}_{\lambda_i\geq 0} \underbrace{\begin{bmatrix} [\mathfrak{b}_i^T v^1]\mathfrak{b}_i \\ [\mathfrak{b}_i^T v^2]\mathfrak{b}_i \\ \vdots \\ [\mathfrak{b}_i^T v^K]\mathfrak{b}_i \end{bmatrix}}_{x^k} = \underbrace{\frac{1}{W}\begin{bmatrix} f^1 \\ f^2 \\ \vdots \\ f^K \end{bmatrix}}_{f}$$

and $\sum_i \lambda_i = 1 \Rightarrow KM$-dimensional vector f is convex combination of N vectors x^i with weights λ_i.

By Caratheodory Theorem, f is a convex combination of x^i with coefficients λ_i^* and at most $M+1$ nonzeros among λ_i^*. Setting $t_i^* = \lambda_i^* W$, we get a truss t^* of total volume W, and $\sum_i \lambda_i^* x^i = f$ says that v^k are equilibrium displacements of truss t^* under load f^k, $k = 1, ..., K$, so that the compliance of t^* w.r.t. every one of loading scenarios f^k is the same as the compliance of truss t. Thus, t^* is another optimal truss, and the number of bars of positive volume in this truss is at most $KM+1$.

Note: When a load every free node of truss is affected by a nonzero external force, the number of positive volume bars in a truss capable to withstand the load should be of order of M, since every free node should be incident to aa bar of nonzero volume. Thus,our upper bound on the number of bars of positive volumes in a (properly selected) optimal truss is, in general, tight as far as the dependence of the total number M of degrees of freedom of the nodal set.

Shapley-Folkman Theorem

♣ **Preliminaries.** Let us make the following simple and useful observations:

Fact II.33 *Taking convex hull commutes with taking direct product: when $X_\ell \subset \mathbf{R}^{n_\ell}$, $1 \leq \ell \leq L$, are nonempty sets, one has*

$$\mathrm{Conv}(X_1 \times ... \times X_L) = \mathrm{Conv}(X_1) \times ... \times \mathrm{Conv}(X_L).$$

Proof. Applying induction in L, all we need to prove is that *If $X \subset \mathbf{R}^n$, $Y \subset \mathbf{R}^m$ are nonempty sets, then* $\mathrm{Conv}(X \times Y) = \mathrm{Conv}(X) \times \mathrm{Conv}(Y)$.
Indeed, let $x^i \in X$, $y^i \in Y$, and let $\lambda_i \geq 0$ be such that $\sum_i \lambda_i = 1$. We have $\sum_i \lambda_i [x^i; y^i] = [\sum_i \lambda_i x^i; \sum_i \lambda_i y^i] \in \mathrm{Conv}(X) \times \mathrm{Conv}(Y) \Rightarrow \mathrm{Conv}(X \times Y) \subset \mathrm{Conv}(X) \times \mathrm{Conv}(Y)$. Vice versa, when $x = \sum_i \lambda_i x^i$ with $\lambda_i \geq 0, \sum_i \lambda_i = 1$ and $y = \sum_j \mu_j y^j$ with $\mu_j \geq 0, \sum_j \mu_j = 1$, we have $[\sum_i \lambda_i x^i; \sum_j \mu_j y^j] = [\sum_{i,j} \lambda_i \mu_j x^i; \sum_{i,j} \lambda_i \mu_j y^j] = \sum_{i,j} \lambda_i \mu_j [x^i; y^j]$ and $\lambda_i \mu_j \geq 0, \sum_{i,j} \lambda_i \mu_j = 1 \Rightarrow \mathrm{Conv}(X) \times \mathrm{Conv}(Y) \subset \mathrm{Conv}(X \times Y)$.

Fact II.34 *Taking convex hull commutes with taking affine image: if $X \subset \mathbf{R}^n$ is a nonempty set, $x \mapsto \mathcal{A}(x) := Ax + b : \mathbf{R}^n \to \mathbf{R}^m$ is an affine mapping, and $\mathcal{A}(Y) = \{\mathcal{A}(x), x \in Y\} \subset \mathbf{R}^m$ is the image of $U \subset \mathbf{R}^n$ under the mapping, then*

$$\mathrm{Conv}(\mathcal{A}(X)) = \mathcal{A}(\mathrm{Conv}(X)).$$

Proof: evident.

As a corollary, we have

Fact II.35 *Taking convex hull commutes with taking weighted sum of sets: if $X_\ell \subset \mathbf{R}^n$, $1 \leq \ell \leq L$, are nonempty sets and $\lambda_1, ..., \lambda_L$ are reals, then*

$$\mathrm{Conv}(\lambda_1 X_1 + ... + \lambda_L X_L) = \lambda_1 \mathrm{Conv}(X_1) + ... + \lambda_L \mathrm{Conv}(X_L).$$

Indeed, setting $\mathcal{A}([x^1; ...x^L]) = \sum_\ell \lambda_\ell x^\ell : \underbrace{\mathbf{R}^n \times .. \times \mathbf{R}^n}_{L} \to \mathbf{R}^n$, we get a linear mapping such that $\lambda_1 X_1 + ... + \lambda_L X_L = \mathcal{A}(X_1 \times ... \times X_L)$, and it remains to use Facts II.14-15.

Note: Facts II.14-16 remain true when replacing taking convex hull with taking affine span (but *not* with taking linear span or conic hull – look what happens with Fact II.33 when $L = 2$ and $X_1 = X_2 = \{1\} \subset \mathbf{R}$).

♠ **Shapley-Folkman Theorem** *Let $X_1, ..., X_N$ be nonempty sets in $\mathbf{R}^d$, and let*

$$x \in \mathrm{Conv}(X_1) + ... + \mathrm{Conv}(X_L) \qquad [= \mathrm{Conv}(X_1 + ... + X_L) \text{ by Fact II.35}]$$

Then x can be represented as the sum $x = x^1 + ... + x^L$ where at most d of the terms x^ℓ belong to the convex hulls of the respective X_ℓ, and all remaining terms belong to the respective X_ℓ.

Proof. We have $x = \sum_{\ell=1}^{L} \left[\sum_{i \le I} \lambda_i^\ell x_i^\ell\right]$ with $x_i^\ell \in X_i$, $\lambda_i^\ell \ge 0$, $\sum_i \lambda_i^\ell = 1$. Consider the $(L+d)$-dimensional vectors $y^{\ell i} = [\overbrace{0; ...; 0}^{\ell-1}; 1; \underbrace{0; ...; 0}_{L}; x_i^\ell]$. We have

$$\sum_{\ell \le L, i \le I} \lambda_i^\ell = y := [\underbrace{1; ...; 1}_{L}; x]$$

$\Rightarrow$ y is a conic combination of $y^{\ell i}$ $\Rightarrow$ y is a conic combination of y_i^ℓ with at most $L+d$ nonzero coefficients (Caratheodory Theorem in Conic form)

$$\Rightarrow \quad x = \sum_\ell \sum_i \mu_i^\ell x_i^\ell \ \&\ \mu_i^\ell \ge 0 \, \forall \ell, i \ \&\ \sum_i \mu_i^\ell = 1 \, \forall \ell$$

Besides this, denoting by d_ℓ the number of nonzeros among $\mu_1^\ell, \mu_2^\ell, ..., \mu_i^\ell$, we have $\sum_\ell d_\ell \le L + d$ and $d_\ell \ge 1$ (due to $\sum_i \mu_i^\ell = 1$).

$\Rightarrow$ Setting $\mathcal{L} = \{\ell : d_\ell > 1\}$ and $m = \mathrm{Card}(\mathcal{L})$, we have

$$L + d \ge \sum_\ell d_\ell = \sum_{\ell \in \mathcal{L}} d_\ell + \underbrace{\sum_{\ell \notin \mathcal{L}} d_\ell}_{L-m} \ge 2m + (L - m) = L + m$$

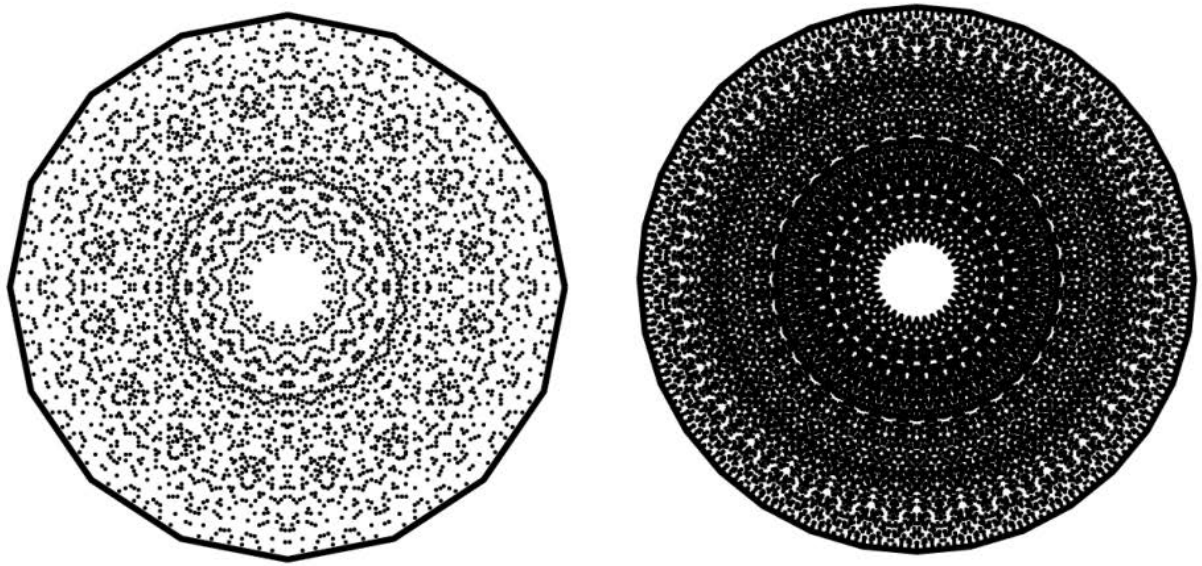
$\Rightarrow$ $m \le d$.

We see that

$$X = \sum_{\ell=1}^{L} \underbrace{\sum_i \mu_i^\ell x_i^\ell}_{x^\ell}$$

with $x^\ell \in \mathrm{Conv}(X_\ell)$ and $x^\ell \notin X_\ell$ for at most d values of ℓ, Q.E.D.

♠ **Comment.** Shapley-Folkman Theorem demonstrates certain "convexification property" of arithmetic summation of sets. Specifically, consider L nonempty sets X_ℓ in $\mathbf{R}^d$, and assume that every one of them is contained in the unit Euclidean ball. Now consider the set $X = X_1 + ... + X_L$ and its convex hull $\widehat{X} = \mathrm{Conv}(X) = \mathrm{Conv}(X_1) + ... + \mathrm{Conv}(X_L)$ (the equality is due to Fact II.35). Of course, $X \subset \widehat{X}$. Shapley-Folkman Theorem implies that *every point from $\widehat{X}$ is at the $\|\cdot\|_2$-distance $\leq d$ from some point of X*. Taking into account that the linear sizes of X can (and typically will) be of order of L, it is prudent to say that *when d is fixed and L grows, the nonconvex set X becomes relatively more and more dense part of its convex hull.*



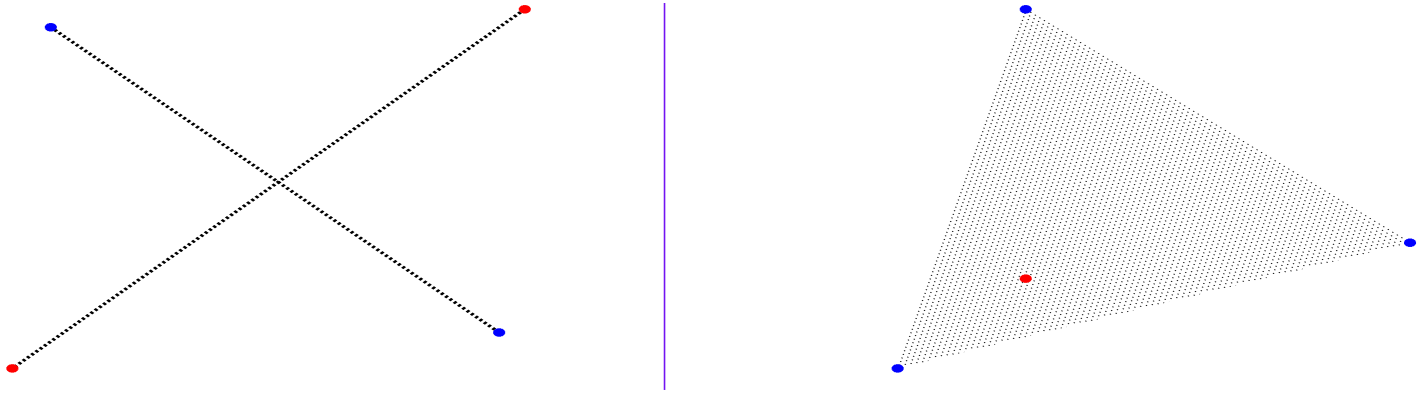
Black dots: the finite set $V = V_1 + V_2 + V_3$, where V_i, $i = 1, 2, 3$, are the vertices of concentric perfect m-side polygons with ratio of linear sizes 4 : 2 : 1. Bold broken line: boundary of the perfect m-side polygon $\mathrm{Conv}(V_1) + \mathrm{Conv}(V_2) + \mathrm{Conv}(V_3)$. Left: $m = 16$; right: $m = 32$.

Illustration: Consider a company producing d types of products on L factories. Given some time period, technologically achievable vectors of product outputs at factory ℓ form a nonempty set $X_\ell \subset \mathbf{R}^n$, while the total product output should be at least a given demand $p \in \mathbf{R}^d$.

- The ideal management goal is to select $x^\ell \in X_\ell$ in such a way that $\sum_\ell x^\ell \geq p$. When X_ℓ are nonconvex, achieving this goal could be a computationally difficult problem, while the convexified version of this problem *Find* $x^\ell \in \mathrm{Conv}(X_\ell)$ *such that* $\sum_\ell x^\ell \geq p$ typically is easy.
- In principle, we can enforce factory ℓ to produce as an output a convex combination $\sum_i \lambda_i x^i$ of outputs $x^i \in X_\ell$; to this end, it suffices to use technological process resulting in output x^i for fraction λ_i of the time period in question. However, this "mixed policy" from the practical viewpoint is much more involving than producing a single output from X_ℓ.
- Shapley-Folkman Theorem says that *if the convexified problem is solvable, then it admits a solution with at most* d *difficult to implement mixed policies.*

Radon & Helly Theorems

Radon Theorem *Let $x^1, ..., x^m$ be $m \geq n+2$ vectors in $\mathbf{R}^n$. One can split the set $\{1, ..., m\}$ of indexes of the vectors into two nonempty and non-overlapping groups A, B such that* $\mathrm{Conv}(\{x^i : i \in A\}) \cap \mathrm{Conv}(\{x^i : i \in B\}) \neq \emptyset$.



Coloring 4 points from $\mathbf{R}^2$ to make convex hulls of red and of blue points intersecting

Proof. Consider the homogeneous system of linear equations in m variables δ_i:

$$\begin{cases} \sum_{i=1}^m \delta_i x_i &= 0 \quad [n \text{ linear equations}] \\ \sum_{i=1}^m \delta_i &= 0 \quad [\text{single linear equation}] \end{cases}$$

Since $m \geq n+2$, the system has a nontrivial solution δ. Setting

$$I = \{i : \delta_i > 0\},\ J = \{i : \delta_i \leq 0\},$$

we split indices $\{1, ..., m\}$ into two *nonempty* (due to $\delta \neq 0, \sum_i \delta_i = 0$) groups such that

$$\sum_{i \in I} \delta_i x_i = \sum_{j \in J} [-\delta_j] x_j,\ \ \gamma = \sum_{i \in I} \delta_i = \sum_{j \in J} -\delta_j > 0$$

whence

$$\underbrace{\sum_{i \in I} \frac{\delta_i}{\gamma} x_i}_{\in \mathrm{Conv}(\{x_i : i \in I\})} = \underbrace{\sum_{j \in J} \frac{-\delta_j}{\gamma} x_j}_{\in \mathrm{Conv}(\{x_j : j \in J\})}.$$

Helly Theorem *Let $A_1, ..., A_M$ be convex sets in $\mathbf{R}^n$. Assume that every $n+1$ sets from the family have a point in common. Then all M sets have point in common.*

• In particular, if every 2 of a finite collection of segments on the real axis have a point in common, all the segments have a common point (evident). If every 3 of a finite collection of triangles on the plane have a point in common, then all the triangles have a common point (why???)

Proof: induction in M.

Base $M \leq n+1$ is trivially true.

Step: Assume that for certain $M \geq n+1$ our statement hods true for every M-member family of convex sets, and let us prove that it holds true for $M+1$-member family of convex sets $A_1, ..., A_{M+1}$.

♢ By inductive hypotheses, every one of the $M+1$ sets

$$B_\ell = A_1 \cap A_2 \cap ... \cap A_{\ell-1} \cap A_{\ell+1} \cap ... \cap A_{M+1}$$

is nonempty. Let us choose $x_\ell \in B_\ell$, $\ell = 1, ..., M+1$.

♢ By Radon Theorem, the collection $x_1, ..., x_{M+1}$ can be split in two sub-collections with intersecting convex hulls. W.l.o.g., let the split be $\{x_1, ..., x_{J-1}\} \cup \{x_J, ..., x_{M+1}\}$, and let

$$z \in \mathrm{Conv}(\{x_1, ..., x_{J-1}\}) \bigcap \mathrm{Conv}(\{x_J, ..., x_{M+1}\}).$$

Claim: $z \in A_\ell$ for all $\ell \leq M+1$, so that $A_1 \cap ... \cap A_{M+1} \neq \emptyset$

Indeed, for $\ell \leq J-1$, the points $x_J, x_{J+1}, ..., x_{M+1}$ belong to the convex set A_ℓ, whence

$$z \in \mathrm{Conv}(\{x_J, ..., x_{M+1}\}) \subset A_\ell.$$

For $\ell \geq J$, the points $x_1, ..., x_{J-1}$ belong to the convex set A_ℓ, whence

$$z \in \mathrm{Conv}(\{x_1, ..., x_{J-1}\}) \subset A_\ell.$$

The inductive step is over and thus the proof of Helly Theorem is complete.

Refinement: *Assume that $A_1, ..., A_M$ are convex sets in $\mathbf{R}^n$ and that*
♢ *the union $A_1 \cup A_2 \cup ... \cup A_M$ of the sets belongs to an affine subspace P of affine dimension m*
♢ *every $m+1$ sets from the family have a point in common*
Then all the sets have a point in common.

Proof We can think of A_j as of sets in P, or, which is the same, as sets in $\mathbf{R}^m$ and apply the Helly Theorem!

What about infinite collections $\{A_\alpha\}_{\alpha\in\mathcal{A}}$?

• When trying to extend Helly Theorem from finite to infinite collections of convex sets, we meet two immediate obstacles:

• *Things can go wrong when the sets A_α are not closed.* E.g. for the collection $\{A_i = (0, 1/i)\}_{i\geq 1}$ of convex subsets of $\mathbf{R}$, intersection of sets from every finite subcollection is nonempty, but the intersection of all A_i is empty

• *Things can go wrong when the intersections of sets from finite subcollections can "run to infinity,"* as is the case for collection $\{A_i = [i, \infty)\}_{i\geq 1}$ of convex subsets of $\mathbf{R}$. Here again intersection of sets from every finite subcollection is nonempty, but the intersection of all A_i is empty.

♠ It turns out that these are the only two obstacles for Helly Theorem to be applicable to infinite collections of convex sets.

Helly Theorem II: *Let A_α, $\alpha \in \mathcal{A}$, be a family of convex sets in $\mathbf{R}^n$ such that every $n+1$ sets from the family have a point in common.*
Assume, in addition, that
♢ *the sets A_α are closed*
♢ *one can find finitely many sets $A_{\alpha_1}, ..., A_{\alpha_M}$ with a bounded intersection.*
Then all sets A_α, $\alpha \in \mathcal{A}$, have a point in common.

Proof. By Helly Theorem, every finite collection of the sets A_α has a point in common, and it remains to apply the following standard fact from Analysis:

Let B_α be a family of closed sets in $\mathbf{R}^n$ such that
♢ every finite collection of the sets has a nonempty intersection;
♢ in the family, there exists finite collection, say, $B_{\alpha_1}, ..., B_{\alpha_N}$, with bounded intersection $B = \cap_{i \le N} B_{\alpha_i}$.
Then all sets from the family have a point in common.

Indeed, B is a closed and bounded subset of $\mathbf{R}^n$, and as such is compact; the sets $\overline{B}_\alpha = B \cap B_\alpha$ are closed subsets of the compact set B, and every finite intersection of these sets is nonempty. By Fact I.13, $\cap_\alpha \overline{B}_\alpha \neq \emptyset$, so that $\cap_\alpha B_\alpha = \cap_\alpha \overline{B}_\alpha \neq \emptyset$.

Exercise: We are given a function $f(x)$ on a 7,000,000-point set $X \subset \mathbf{R}$. At every 7-point subset of X, this function can be approximated, within accuracy 0.001 at every point, by appropriate polynomial of degree 5. To approximate the function on the entire X, we want to use a spline of degree 5 (a piecewise polynomial function with pieces of degree 5). How many pieces do we need to get accuracy 0.001 at every point?

Answer: Just one. Indeed, let A_x, $x \in X$, be the set of coefficients of all polynomials of degree 5 which reproduce $f(x)$ within accuracy 0.001:

$$A_x = \{p = (p_0, ..., p_5) \in \mathbf{R}^6 : |f(x) - \sum_{i=0}^{5} p_i x^i| \leq 0.001\}.$$

The set A_x is polyhedral and therefore convex, and we know that every $6 + 1 = 7$ sets from the family $\{A_x\}_{x \in X}$ have a point in common. By Helly Theorem, all sets A_x, $x \in X$, have a point in common, that is, there exists a *single* polynomial of degree 5 which approximates f within accuracy 0.001 at *every* point of X.

Exercise: We should design a factory which, mathematically, is described by the following Linear Programming model:

$$\begin{array}{rcll} Dx & \geq & d & [d_1, ..., d_{1000}\text{: demands}] \\ Rx & \leq & r & [r_1 \geq 0, ..., r_{10} \geq 0\text{: amounts of resources of various types}] \\ Cx & \leq & c & [\text{other constraints}] \end{array} \qquad (F)$$

The data D, R, C, c are given in advance. We should buy in advance resources $r_i \geq 0$, $i = 1, ..., 10$, in such a way that the factory will be capable to satisfy all demand scenarios d from a given finite set $\mathcal{D}$, that is, (F) should be feasible for every $d \in \mathcal{D}$. Amount r_i of resource i costs us $a_i r_i$ with $0 \neq a \geq 0$.

It is known that in order to be able to satisfy every single demand from $\mathcal{D}$, it suffices to invest \$1 in the resources.

How large should be investment in resources in the cases when $\mathcal{D}$ contains

- ♢ just one scenario?
- ♢ 3 scenarios?
- ♢ 10 scenarios?
- ♢ 2024 scenarios?

Answer: $\mathcal{D} = \{d_1\} \Rightarrow$ \$1 is enough
$\mathcal{D} = \{d_1, d_2, d_3\} \Rightarrow$ \$3 is enough
$\mathcal{D} = \{d_1, ..., d_{10}\} \Rightarrow$ \$10 is enough
$\mathcal{D} = \{d_1, ..., d_{2024}\} \Rightarrow$ \$10 is enough!
Indeed, for $d \in \mathcal{D}$ let $\mathcal{R}_d$ be the set of all nonnegative $r \in \mathbf{R}^{10}$ which cost exactly \$10 and result in solvable system

$$\begin{array}{rcl} Dx & \geq & d \\ Rx & \leq & r \\ Cx & \leq & c \end{array} \qquad (F[d])$$

in variables x. The set $\mathcal{R}_d$ is convex.
Indeed, setting $X = \{x : Cx \leq c\}$ and invoking Illustration to Fact II.10, we conclude that the set

$$\mathcal{RD} = \{(d', r) : \exists x : x \in X, Dx \geq d', Rx \leq r\}$$

is convex. Intersecting $\mathcal{RD}$ with the convex set $\{(d', r) : d' = d, r \geq 0, \sum_j a_j r_j = 10\}$, we get a convex set, and $\mathcal{R}_d$ is just the projection of this convex set onto the space of r-variables.
We claim that *every 10 sets of the family* $\{\mathcal{R}_d : d \in \mathcal{D}\}$ *have a common point.*
Indeed, given 10 scenarios $d^1, ..., d^{10}$ from $\mathcal{D}$, we can meet demand scenario d^i (i.e., make $(F[d^i])$ feasible) investing in resources at most \$1, that is, for $i \leq 10$ there exists $\widetilde{r^i} \geq 0$ such that $\sum_j a_j \widetilde{r^i_j} \leq 1$ and with $r = \widetilde{r^i}$, the system $(F[d^i])$ is feasible. The latter property remains intact when replacing $r = \widetilde{r^i}$ with $r \geq \widetilde{r^i}$, and since $0 \neq a \geq 0$ and the cost of $\widetilde{r^i}$ is at most \$1, we can $\geq$-increase $\widetilde{r^i}$ to make the cost of the resulting vector r^i to be exactly \$1. Thus, for every $i \leq 10$ we can meet demand scenario d^i with vector of resources $r^i \geq 0$ of cost \$1. It remains to note that the vector of resources $r^1 + ... + r^{10}$ is nonnegative, meets every one the demand scenarios d^i, $i \leq 10$, and costs \$ 10, that is, it belongs to every one of the sets $\mathcal{R}_{d^i}$. $i \leq 10$.
Convex sets $\mathcal{R}_d \subset \mathbf{R}^{10}$, $d \in \mathcal{D}$, belong to 9-dimensional affine plane $\{r \in \mathbf{R}^{10} : \sum_j a_j r_j = 10\}$. Since every 10 of these sets have a point in common, all these sets have a point r in common. r costs \$10, and with this r, every one of the systems $(F[d])$, $d \in \mathcal{D}$, is solvable.

Exercise: Consider an optimization program

$$c_* = \min\left\{c^T x : g_i(x) \le 0,\, i = 1, ..., 2024\right\}$$

with 11 variables $x_1, ..., x_{11}$. Assume that the constraints are convex, that is, every one of the sets

$$X_i = \{x : g_i(x) \le 0\},\, i = 1, ..., 2024$$

is convex. Assume also that the problem is solvable with optimal value 0.
Clearly, when dropping one or more constraints, the optimal value can only decrease or remain the same.
♢ Is it possible to find a constraint such that dropping it, we preserve the optimal value? Two constraints which can be dropped simultaneously with no effect on the optimal value? Three of them?

Answer: You can drop as many as $2024 - 11 = 2013$ appropriately chosen constraints without varying the optimal value!
Assume, on the contrary, that every 11-constraint relaxation of the original problem has negative optimal value. Since there are finitely many such relaxations, there exists $\epsilon < 0$ such that every problem of the form

$$\min_x \{c^T x : g_{i_1}(x) \leq 0, ..., g_{i_{11}}(x) \leq 0\}$$

has a feasible solution with the value of the objective $< -\epsilon$. Since this problem has a feasible solution with the value of the objective equal to 0 (namely, the optimal solution of the original problem) and its feasible set is convex, the problem has a feasible solution x with $c^T x = -\epsilon$. In other words, every 11 of the 2024 sets

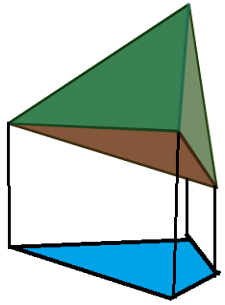
$$Y_i = \{x : c^T x = -\epsilon, g_i(x) \leq 0\},\ i = 1, ..., 2024$$

have a point in common.
The sets Y_i are convex (as intersections of convex sets X_i and an affine subspace). If $c \neq 0$, then these sets belong to affine subspace of affine dimension 10, and since every 11 of them intersect, all 2024 intersect; a point x from their intersection is a feasible solution of the original problem with $c^T x < 0$, which is impossible.
When $c = 0$, the claim is evident: we can drop all 2024 constraints without varying the optimal value!

Lecture I.3

Polyhedral Sets and Theory of Systems of Linear Inequalities

Polyhedrality and Fourier-Motzkin Elimination
Calculus of Polyhedrality
General Theorem of Alternative
Linear Programming Duality



Polyhedrality & Fourier-Motzkin Elimination

♣ **Definition:** A *polyhedral* set $X \subset \mathbf{R}^n$ is a set which can be represented as

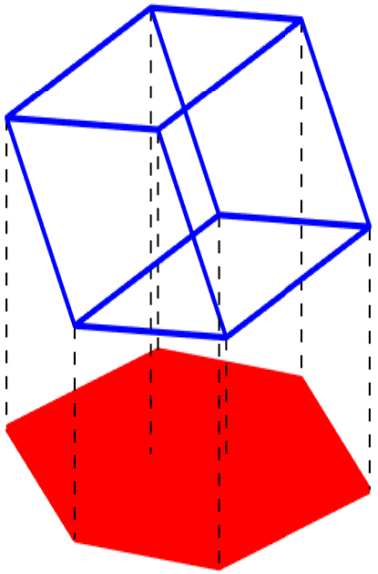
$$X = \{x : Ax \leq b\},$$

that is, as the solution set of a finite system of nonstrict linear inequalities.

♠ **Definition:** A polyhedral representation of a set $X \subset \mathbf{R}^n$ is a representation of X of the form:

$$X = \{x : \exists w : Px + Qw \leq r\},$$

that is, a representation of X as the *a projection* onto the space of x-variables of a polyhedral set $X^+ = \{[x; w] : Px + Qw \leq r\}$ in the space of x, w-variables.



Rotated 3D cube and its 2D projection (hexagon)

♠ **Examples of polyhedral representations:**

- The set $X = \{x \in \mathbf{R}^n : \sum_i |x_i| \leq 1\}$ admits the p.r.

$$X = \left\{ x \in \mathbf{R}^n : \exists w \in \mathbf{R}^n : \begin{array}{r} -w_i \leq x_i \leq w_i, \\ 1 \leq i \leq n, \\ \sum_i w_i \leq 1 \end{array} \right\}.$$

- The set

$$X = \begin{array}{l} \{x \in \mathbf{R}^6 : \max[x_1, x_2, x_3] + 2\max[x_4, x_5, x_6] \\ \quad \leq x_1 - x_6 + 5\} \end{array}$$

admits the p.r.

$$X = \left\{ x \in \mathbf{R}^6 : \exists w \in \mathbf{R}^2 : \begin{array}{l} x_1 \leq w_1, x_2 \leq w_1, x_3 \leq w_1 \\ x_4 \leq w_2, x_5 \leq w_2, x_6 \leq w_2 \\ w_1 + 2w_2 \leq x_1 - x_6 + 5 \end{array} \right\}.$$

Whether a Polyhedrally Representable Set is Polyhedral?

♣ **Question:** Let X be given by a polyhedral representation:

$$X = \{x \in \mathbf{R}^n : \exists w : Px + Qw \leq r\},$$

that is, as the *projection* of the solution set

$$Y = \{[x; w] : Px + Qw \leq r\} \qquad (*)$$

of a finite system of linear inequalities in variables x, w onto the space of x-variables.
Is it true that X is polyhedral, i.e., X is a solution set of finite system of linear inequalities *in variables x only*?

Theorem III.1 *Every polyhedrally representable set is polyhedral.*

Proof is given by the *Fourier — Motzkin elimination scheme* which demonstrates that the projection of the set $(*)$ onto the space of x-variables is a polyhedral set. **Elimination step:** eliminating a *single* slack variable. Given set $(*)$, assume that $w = [w_1; ...; w_m]$ is nonempty, and let Y^+ be the projection of Y on the space of variables $x, w_1, ..., w_{m-1}$:

$$Y^+ = \{[x; w_1; ...; w_{m-1}] : \exists w_m : Px + Qw \leq r\} \qquad (!)$$

Let us prove that Y^+ is polyhedral. Indeed, let us split the linear inequalities

$$p_i^T x + q_i^T w \leq r_i,\ 1 \leq i \leq I$$

defining Y into three groups:

- **black** – the coefficient at w_m is 0
- **red** – the coefficient at w_m is > 0
- **blue** – the coefficient at w_m is < 0

Then

$$Y = \left\{ [x; w] : \begin{array}{ll} a_i^T x + b_i^T [w_1; ...; w_{m-1}] \leq c_i & , i \text{ is black} \\ w_m \leq \overline{a}_i(x, w_1, ..., w_{m-1}) := a_i^T x + b_i^T [w_1; ...; w_{m-1}] + c_i & , i \text{ is red} \\ w_m \geq a_i(x, w_1, ..., w_{m-1}) := a_i^T x + b_i^T [w_1; ...; w_{m-1}] + c_i & , i \text{ is blue} \end{array} \right\}$$

$$Y = \left\{ [x; w] : \begin{array}{ll} a_i^T x + b_i^T [w_1; ...; w_{m-1}] \leq c_i & , i \text{ is black} \\ w_m \leq \overline{a}_i(x, w_1, ..., w_{m-1}) := a_i^T x + b_i^T [w_1; ...; w_{m-1}] + c_i & , i \text{ is red} \\ w_m \geq a_i(x, w_1, ..., w_{m-1}) := a_i^T x + b_i^T [w_1; ...; w_{m-1}] + c_i & , i \text{ is blue} \end{array} \right\}$$

Clearly, a collection $x, w_1, ..., w_{m-1}$ cam be augmented by w_{m+1} to yield a point from y iff the collection satisfies all black inequalities and every blue lower bound $a_i(x, w_!, ..., w_{m-1})$ on w_m is $\leq$ every red upper bound $\overline{a}_j(x, w_1, ..., w_{m-1})$ on w_m:

$$Y^+ := \{x, w_1, ..., w_{m-1} : \exists w_m : [x; w_1; ...; w_m] \in Y\} = \left\{ x, w_1, ..., w_{m-1} : \begin{array}{ll} a_i^T x + b_i^T [w_1; ...; w_{m-1}] \leq c_i & , \text{for all black } i \\ a_i(x, w_1, ..., w_{m-1}) \leq w_m \leq & \text{for all blue } i \\ \overline{a}_j(x, w_1, ..., w_{m-1}) & \text{and all red } j \end{array} \right\},$$

implying that Y^+ is polyhedral. Iterating the process, we conclude that $X = \{x : \exists w_1, ..., w_m : [x; w_1; ...; w_m] \in Y$ is polyhedral, Q.E.D.

♣ As an immediate consequence, we get

Corollary III,1 *Convex hull* $\mathrm{Conv}(X)$ *of a finite set* $X = \{a^1, ..., a^N\} \subset \mathbf{R}^n$ *is a polyhedral set.*

Indeed, there is nothing to prove when $X = \emptyset$. When X is nonempty, $\mathrm{Conv}(X)$ is polyhedral representable:

$$\mathrm{Conv}(X) = \{x \in \mathbf{R}^n : \exists \lambda \in \mathbf{R}^N : \lambda \geq 0, \sum_i \lambda_i = 1, x = \sum_i \lambda_i a^i\}$$

and therefore is polyhedral.

♣ As another immediate consequence of Theorem III.1, let us build a finite algorithms for solving LO problems.
Given an LO program

$$\text{Opt} = \max_x \{c^T x : Ax \le b\}, \qquad (!)$$

observe that the set of values of the objective at feasible solutions can be represented as

$$\begin{aligned} T &= \{\tau \in \mathbf{R} : \exists x : Ax \le b, c^T x - \tau = 0\} \\ &= \{\tau \in \mathbf{R} : \exists x : Ax \le b, c^T x \le \tau, c^T x \ge \tau\} \end{aligned}$$

that is, T *is polyhedral representable*. By Theorem III.1, T is polyhedral, that is, T can be represented by a finite system of linear inequalities *in variable* τ *only.* It immediately follows that *if* T *is nonempty and is bounded from above,* T *has the largest element*. Thus, we have proved

Corollary III.2 *A feasible and bounded LO program admits an optimal solution and thus is solvable.*

Moreover, Fourier-Motzkin elimination scheme suggests a finite algorithm for solving an LO program, where we

- first, apply the scheme to get a representation of T by a finite system S of linear inequalities in variable τ,
- second, analyze S to find out whether the solution set is nonempty and bounded from above, and when it is the case, find out the optimal value $\text{Opt} \in T$ of the program,
- third, use the Fourier-Motzkin elimination scheme in the backward fashion to find x such that $Ax \le b$ and $c^T x = \text{Opt}$, thus recovering an optimal solution to the problem of interest.

Bad news: The resulting algorithm is completely impractical, since the number of inequalities we should handle at a step usually rapidly grows with the step number and can become astronomically large when eliminating just tens of variables.

Calculus of Polyhedrality

♣ Polyhedral sets are convex. All basic convexity-preserving operations as applied to polyhedral sets preserve polyhedrality. Moreover, as applied to polyhedral representations, the corresponding "calculus" is extremely simple and fully algorithmic. Then basics of this calculus is as follows:

A. Taking finite intersections: *Let $X_1, ..., X_K$ be polyhedral sets in $\mathbf{R}^n$ given by polyhedral representations:*

$$X_k = \{x \in \mathbf{R}^n : \exists w_k : P_k x + Q_k w_k \leq r_k\},\ k \leq K.$$

Then the intersection $X = \bigcap_k X_k$ of X_k is polyhedral set with polyhedral representation

$$X = \{x : \exists w = [w_1; ...; w_K] : P_k x + Q_k w^k \leq r_k\},\ k \leq K.$$

B. Taking direct products: *Let $X_k \subset \mathbf{R}^{n_k}$, $k \leq K$, be polyhedral sets given by polyhedral representations:*

$$X_k = \{x_k \in \mathbf{R}^{n_k} : \exists w_k : P_k x_k + Q_k w_k \leq r_k\},\ k \leq K.$$

Then the direct product $X = X_1 \times ... \times X_K$ of X_k is polyhedral set with polyhedral representation

$$X = \{x = [x_1; ...; x_K] : \exists w = [w_1; ...; w_K] : P_k x_k + Q_k w_k \leq r_k,\ k \leq K\}.$$

C. Taking affine image: *Let $X \subset \mathbf{R}^n$, be a polyhedral set given by polyhedral representation:*

$$X = \{x \in \mathbf{R}^n : \exists w : Px + Qw \leq r\}.$$

and $\mathcal{A}(X) = Ax + b : \mathbf{R}^n \to \mathbf{R}^m$ be affine mapping. Then the image $\mathcal{A}(X) = \{\mathcal{A}(x) : x \in X\| \subset \mathbf{R}^m\}$ of X under the mapping is polyhedral set with polyhedral representation

$$\mathcal{A}(X) = \{y \in \mathbf{R}^m : \exists [x; w] : Px + Qw \leq r \ \&\ \underbrace{y - Ax \leq b,\ -y + Ax \leq -b}_{y = Ax + b}\}.$$

D. Taking inverse affine image: *Let $X \subset \mathbf{R}^n$, be a polyhedral set given by polyhedral representation:*

$$X = \{x \in \mathbf{R}^n : \exists w : Px + Qw \leq r\}.$$

and $\mathcal{A}(y) = Ay + b : \mathbf{R}^m \to \mathbf{R}^n$ be affine mapping. Then the inverse image $\mathcal{A}^{-1}(X) = \{y : \mathcal{A}(y) \in X\| \subset \mathbf{R}^m\}$ of X under the mapping is polyhedral set with polyhedral representation

$$\mathcal{A}^{-1}(X) = \{y \in \mathbf{R}^m : \exists w : PAy + Qw \leq r - Pb\}.$$

E. Arithmetic summation: *Let $X_k \subset \mathbf{R}^n$, $k \leq K$, be polyhedral sets given by polyhedral representations:*

$$X_k = \{x \in \mathbf{R}^n : \exists w_k : P_k x + Q_k w_k \leq r_k\},\ k \leq K.$$

Then the arithmetic sum $X = X_1 + ... + X_K$ of X_k is polyhedral set with polyhedral representation

$$X = \{x : \exists [x_1; ..., x_K; w_1; ...; w_K] : P_k x_k + Q_k w_k \leq r_k,\ x = x_1 + ... + x_K\}.$$

F. Taking closed conic transform. *Let X be a nonempty polyhedral set given by polyhedral representation:*

$$X = \{x : \exists u : Px + Qu \leq r\}$$

The closed conic transform $\overline{\mathsf{ConeT}}(X) = \mathsf{cl}\,\{[x; t] : t > 0, x/t \in X\}$ is polyhedral with polyhedral representation

$$\overline{\mathsf{ConeT}}(X) = \{[x; t] : \exists u : Px + Qu \leq tr\ \&\ t \geq 0\}$$

Indeed, denoting $\overline{X}$ the right hand side set in the latter equality, observe that $\overline{X}$ is a polyhedral cone and as such is closed. Besides this, the perspective transform Persp$(X) = \{[x : t] : t > 0.x/t \in X\}$ clearly is nothing but the intersection of $\overline{X}$ with the half-space $t > 0$, and thus Persp$(X) \subset \overline{X}$, and as X is closed, we get $\overline{\text{ConeT}}(X) = \text{cl Persp}(X) \subset$
$overline X$. To prove that this inclusion is equality, let $[x; t] \in \overline{X}$, so that

$$Px + Qu \leq tr \qquad (a)$$

for some u, and $t \geq 0$; When $t > 0$, (a) says that $P(x/t)+Q(u/t) \leq r$, that is, $x/t \in X$,, whence $\{[x; t] \in \text{Persp}(X)$. Now let $t = 0$. As X is nonempty, there exists $\bar{a}$ and $\bar{u}$ such that $P\bar{x}+Q\bar{u} \leq r.$, which combines with $Px+Qu \leq 0$ to imply that

$$[x + \lambda\bar{x}[+Q[u + \lambda\bar{u}] \leq \lambda r, \forall\lambda > 0$$

which, as have already seen, implies that $[x + \lambda\bar{x}; \lambda] \in \text{Persp}(X)$. Passing to limit as $\lambda]to + 0$, we get $[x; 0] \in$ cl Persp(X). Thus, whenever $[x; t] \in \overline{X}$, we have $[x; t] \in \text{cl Persp}(X) = \overline{\text{ConeT}}(X)$, that is, $\overline{X} \subset \overline{\text{ConeT}}(X)$. The opposite inclusion has been already proved, and we end up with $\overline{X} \subset \overline{\text{ConeT}}(X)$, Q.E.D.]small

Theory of Systems of Linear Inequalities, I: Homogeneous Farkas Lemma

♣ Consider a homogeneous linear inequality

$$a^T x \geq 0 \qquad (*)$$

along with a finite system of similar inequalities:

$$a_i^T x \geq 0,\ 1 \leq i \leq m \qquad (!)$$

♣ **Question:** *When* $(*)$ *is a consequence of* $(!)$*, that is, every* x *satisfying* $(!)$ *satisfies* $(*)$ *as well?*
Observation: If a is a conic combination of $a_1, ..., a_m$:

$$\exists \lambda_i \geq 0 : a = \sum_i \lambda_i a_i, \qquad (+)$$

then $(*)$ is a consequence of $(!)$.
Indeed, $(+)$ implies that

$$a^T x = \sum_i \lambda_i a_i^T x \ \forall x,$$

and thus for every x with $a_i^T x \geq 0 \, \forall i$ one has $a^T x \geq 0$.

$$a^T x \geq 0 \qquad (*)$$
$$a_i^T x \geq 0,\ 1 \leq i \leq m \qquad (!)$$

Homogeneous Farkas Lemma: *$(*)$ is a consequence of $(!)$ if and only if a is a conic combination of $a_1, ..., a_m$.*

♣ **Equivalently:** *Given vectors $a_1, ..., a_m \in \mathbf{R}^n$, let $K = \text{Cone}\{a_1, ..., a_m\} = \{\sum_i \lambda_i a_i : \lambda \geq 0\}$ be the conic hull of the vectors. Given a vector a,*

- *it is easy to certify that $a \in \text{Cone}\{a_1, ..., a_m\}$: a certificate is a collection of weights $\lambda_i \geq 0$ such that $\sum_i \lambda_i a_i = a$;*
- *it is easy to certify that $a \notin \text{Cone}\{a_1, ..., a_m\}$: a certificate is a vector d such that $a_i^T d \geq 0\ \forall i$ and $a^T d < 0$.*

Proof of HFL: All we need to prove is that *If a is not a conic combination of $a_1, ..., a_m$, then there exists d such that $a^T d < 0$ and $a_i^T d \geq 0$, $i = 1, ..., m$.*

Fact: The set $K = \text{Cone}\{a_1, ..., a_m\}$ is polyhedral representable:

$$\text{Cone}\{a_1, ..., a_m\} = \left\{ x : \exists \lambda \in \mathbf{R}^m : \begin{array}{c} x = \sum_i \lambda_i a_i \\ \lambda \geq 0 \end{array} \right\}.$$

$\Rightarrow$ By Fourier-Motzkin, K is polyhedral:

$$K = \{x : d_\ell^T x \geq c_\ell, 1 \leq \ell \leq L\}.$$

Observation I: $0 \in K \Rightarrow c_\ell \leq 0\ \forall \ell$

Observation II: $\lambda a_i \in \text{Cone}\{a_1, ..., a_m\}\ \forall \lambda > 0 \Rightarrow \lambda d_\ell^T a_i \geq c_\ell\ \forall \lambda \geq 0 \Rightarrow d_\ell^T a_i \geq 0\ \forall i, \ell$.

Now, $a \notin \text{Cone}\{a_1, ..., a_m\} \Rightarrow \exists \ell = \ell_* : d_{\ell_*}^T a < c_{\ell_*} \leq 0 \Rightarrow d_{\ell_*}^T a < 0$.

$\Rightarrow$ *$d = d_{\ell_*}$ satisfies $a^T d < 0$, $a_i^T d \geq 0$, $i = 1, ..., m$,* Q.E.D.

Theory of Systems of Linear Inequalities, II
Theorem of Alternative

♣ A general (finite!) system of linear inequalities with unknowns $x \in \mathbf{R}^n$ can be written down as

$$\begin{array}{rcl} a_i^T x & > & b_i,\ i = 1, ..., m_{\mathsf{s}} \\ a_i^T x & \geq & b_i,\ i = m_{\mathsf{s}} + 1, ..., m \end{array} \qquad (S)$$

Question: *How to certify that* (S) *is solvable?*
Answer: A solution is a certificate of solvability!
Example: To certify that the system

$$\begin{array}{rrrcr} -4u & -9v & +5w & > & 1.99 \\ -2u & +6v & & \geq & -2 \\ 7u & & -5w & \geq & 1 \end{array}$$

is solvable, it suffices to note that $u = \frac{1}{7}$, $v = -\frac{2}{7}$, $w = 0$ is a solution.

Question: *How to certify that* S *is* **not** *solvable?*
Answer: ???

$$\begin{array}{rcl} a_i^T x & > & b_i,\ i = 1, ..., m_{\mathsf{s}} \\ a_i^T x & \geq & b_i,\ i = m_{\mathsf{s}} + 1, ..., m \end{array} \qquad (S)$$

Question: *How to certify that S is* **not** *solvable?*
Conceptual **sufficient insolvability condition:**
If we can lead the assumption that x solves (S) to a contradiction, then (S) has no solutions.
Example: To certify that the system

$$\begin{array}{rrrcr} -4u & -9v & +5w & > & 2 \\ -2u & +6v & & \geq & -2 \\ 7u & & -5w & \geq & 1 \end{array}$$

has no solutions, it suffices to point out that *aggregating the inequalities of the system with weights $2, 3, 2$, we get a contradictory inequality:*

$$\begin{array}{rr|rrrcr} & 2\times & -4u & -9v & +5w & > & 2 \\ + & & & & & & \\ & 3\times & -2u & +6v & & \geq & -2 \\ + & & & & & & \\ & 2\times & 7u & & -5w & \geq & 1 \\ \hline & & 0\cdot u & +0\cdot v & +0\cdot w & > & 0 \end{array}$$

By how we aggregate, every solution to the system *must* solve the aggregated inequality. The latter has no solutions $\Rightarrow$ so is the system.

$$\begin{array}{rcl} a_i^T x & > & b_i,\ i = 1, ..., m_{\mathsf{s}} \\ a_i^T x & \geq & b_i,\ i = m_{\mathsf{s}} + 1, ..., m \end{array} \qquad (S)$$

"Contradiction by linear aggregation:" Let us associate with inequalities of (S) *nonnegative* weights λ_i and sum up the inequalities with these weights. The resulting inequality

$$\left[\sum_{i=1}^{m} \lambda_i a_i\right]^T x \begin{cases} > \sum_i \lambda_i b_i, & \sum_{i=1}^{m_{\mathsf{s}}} \lambda_i > 0 \\ \geq \sum_i \lambda_i b_i, & \sum_{i=1}^{m_{\mathsf{s}}} \lambda_i = 0 \end{cases} \qquad (C)$$

by its origin is a *consequence* of (S), that is, it is satisfied at every solution to (S). Consequently, *if there exist $\lambda \geq 0$ such that (C) has no solutions at all, then (S) has no solutions!*

Question: When a linear inequality

$$d^T x \begin{cases} > \\ \geq \end{cases} e$$

has no solutions at all?
Answer: This is the case if and only if $d = 0$ and

— either the sign is "$>$", and $e \geq 0$,

— or the sign is "$\geq$", and $e > 0$.

Conclusion: *Consider a system of linear inequalities*

$$\begin{array}{rcl} a_i^T x & > & b_i,\ i = 1, ..., m_{\mathsf{s}} \\ a_i^T x & \geq & b_i,\ i = m_{\mathsf{s}}+1, ..., m \end{array} \qquad (S)$$

in variables x, and let us associate with it two systems of linear inequalities in variables λ:

$$\mathcal{T}_{\mathrm{I}}: \left\{ \begin{array}{rcl} \lambda & \geq & 0 \\ \sum_{i=1}^{m} \lambda_i a_i & = & 0 \\ \sum_{i=1}^{m_s} \lambda_i & > & 0 \\ \sum_{i=1}^{m} \lambda_i b_i & \geq & 0 \end{array} \right. \qquad \mathcal{T}_{\mathrm{II}}: \left\{ \begin{array}{rcl} \lambda & \geq & 0 \\ \sum_{i=1}^{m} \lambda_i a_i & = & 0 \\ \sum_{i=1}^{m_s} \lambda_i & = & 0 \\ \sum_{i=1}^{m} \lambda_i b_i & > & 0 \end{array} \right.$$

If *one of the systems $\mathcal{T}_{\mathrm{I}}$, $\mathcal{T}_{\mathrm{II}}$ is solvable,* **then** *(S) is unsolvable.*
Note: If *$\mathcal{T}_{\mathrm{II}}$ is solvable,* **then** *already the system*

$$a_i^T x \geq b_i,\ i = m_{\mathsf{s}}+1, ..., m$$

is unsolvable!

General Theorem of Alternative: *A system of linear inequalities*

$$\begin{array}{rcl} a_i^T x & > & b_i,\ i = 1, ..., m_{\mathsf{s}} \\ a_i^T x & \geq & b_i,\ i = m_{\mathsf{s}} + 1, ..., m \end{array} \qquad (S)$$

is **un***solvable* **iff** *one of the systems*

$$\mathcal{T}_{\mathrm{I}}: \begin{cases} \lambda & \geq & 0 \\ \sum_{i=1}^{m} \lambda_i a_i & = & 0 \\ \sum_{i=1}^{m_s} \lambda_i & > & 0 \\ \sum_{i=1}^{m} \lambda_i b_i & \geq & 0 \end{cases} \qquad \mathcal{T}_{\mathrm{II}}: \begin{cases} \lambda & \geq & 0 \\ \sum_{i=1}^{m} \lambda_i a_i & = & 0 \\ \sum_{i=1}^{m_s} \lambda_i & = & 0 \\ \sum_{i=1}^{m} \lambda_i b_i & > & 0 \end{cases}$$

is solvable.

Note: *The subsystem*

$$a_i^T x \geq b_i,\ i = m_{\mathsf{s}} + 1, ..., m$$

of (S) *is* **un***solvable* **iff** $\mathcal{T}_{\mathrm{II}}$ *is solvable!*

$$\mathcal{T}_{\rm I}: \begin{cases} \lambda & \geq & 0 \\ \sum_{i=1}^{m} \lambda_i a_i & = & 0 \\ \sum_{i=1}^{m_s} \lambda_i & > & 0 \\ \sum_{i=1}^{m} \lambda_i b_i & \geq & 0 \end{cases} \qquad \mathcal{T}_{\rm II}: \begin{cases} \lambda & \geq & 0 \\ \sum_{i=1}^{m} \lambda_i a_i & = & 0 \\ \sum_{i=1}^{m_s} \lambda_i & = & 0 \\ \sum_{i=1}^{m} \lambda_i b_i & > & 0 \end{cases}$$

Proof. We already know that solvability of one of the systems $\mathcal{T}_{\rm I}$, $\mathcal{T}_{\rm II}$ is a sufficient condition for **un**solvability of (S). All we need to prove is that if (S) is unsolvable, then one of the systems $\mathcal{T}_{\rm I}$, $\mathcal{T}_{\rm II}$ is solvable.
Assume that the system

$$\begin{array}{rcl} a_i^T x & > & b_i,\ i = 1, ..., m_{\rm s} \\ a_i^T x & \geq & b_i,\ i = m_{\rm s}+1, ..., m \end{array} \qquad (S)$$

in variables x has no solutions. Then *every solution x, τ, ϵ to the homogeneous system of inequalities*

$$\begin{array}{rrcl} \tau & -\epsilon & \geq & 0 \\ a_i^T x & -b_i\tau & \geq & 0,\ i = 1, ..., m_{\rm s} \\ a_i^T x & -b_i\tau & \geq & 0,\ i = m_{\rm s}+1, ..., m \end{array}$$

has $\epsilon \leq 0$.
Indeed, in a solution with $\epsilon > 0$ one would also have $\tau > 0$, and the vector $\tau^{-1}x$ would solve (S).

$$\mathcal{T}_{\rm I}: \begin{cases} \lambda & \geq & 0 \\ \sum_{i=1}^{m} \lambda_i a_i & = & 0 \\ \sum_{i=1}^{m_s} \lambda_i & > & 0 \\ \sum_{i=1}^{m} \lambda_i b_i & \geq & 0 \end{cases} \qquad \mathcal{T}_{\rm II}: \begin{cases} \lambda & \geq & 0 \\ \sum_{i=1}^{m} \lambda_i a_i & = & 0 \\ \sum_{i=1}^{m_s} \lambda_i & = & 0 \\ \sum_{i=1}^{m} \lambda_i b_i & > & 0 \end{cases}$$

Situation: Every solution to the system of homogeneous inequalities

$$\begin{array}{rrcl} \tau & -\epsilon & \geq & 0 \\ a_i^T x & -b_i\tau & \geq & 0,\ i = 1, ..., m_{\rm s} \\ a_i^T x & -b_i\tau & \geq & 0,\ i = m_{\rm s}+1, ..., m \end{array} \qquad (U)$$

has $\epsilon \leq 0$, i.e., the homogeneous inequality

$$-\epsilon \geq 0 \qquad (I)$$

is a consequence of system (U) of homogeneous inequalities. By Homogeneous Farkas Lemma, *the vector of coefficients in the left hand side of (I) is a conic combination of the left hand side vectors of coefficients of (U):*

$$\exists \lambda \geq 0, \nu \geq 0: \quad \begin{array}{rcll} \sum_{i=1}^{m} \lambda_i a_i & = & 0 & \text{[coefficients at } x\text{]} \\ -\sum_{i=1}^{m} \lambda_i b_i + \nu & = & 0 & \text{[coefficient at } \tau\text{]} \\ -\sum_{i=1}^{m_{\rm s}} \lambda_i - \nu & = & -1 & \text{[coefficient at } \epsilon\text{]} \end{array}$$

Assuming that $\lambda_1 = ... = \lambda_{m_{\rm s}} = 0$, we get $\nu = 1$, and therefore λ solves $\mathcal{T}_{\rm II}$. In the case of $\sum_{i=1}^{m_{\rm s}} \lambda_i > 0$, λ clearly solves $\mathcal{T}_{\rm I}$.

Corollaries of GTA

♣ **Principle A:** *A finite system of linear inequalities has no solutions iff one can lead it to a contradiction by linear aggregation, i.e., an appropriate weighted sum of the inequalities with "legitimate" weights is either a contradictory inequality*

$$0^T x > a \qquad [a \geq 0]$$

or a contradictory inequality

$$0^T x \geq a \qquad [a > 0]$$

♣ **Principle B:** [Inhomogeneous Farkas Lemma] *A linear inequality*

$$a^T x \leq b$$

is a consequence of **solvable** *system of linear inequalities*

$$a_i^T x \leq b_i,\ i = 1, ..., m$$

iff the target inequality can be obtained from the inequalities of the system **and** *the identically true inequality*

$$0^T x \leq 1$$

by linear aggregation, that is, iff there exist nonnegative $\lambda_0, \lambda_1, ..., \lambda_m$ *such that*

$$\begin{array}{rcl} a & = & \sum_{i=1}^m \lambda_i a_i \\ b & = & \lambda_0 + \sum_{i=1}^m \lambda_i b_i \end{array}$$

$$\left[\Leftrightarrow \left\{ \begin{array}{rcl} a & = & \sum_{i=1}^m \lambda_i a_i \\ b & \geq & \sum_{i=1}^m \lambda_i b_i \end{array} \right. \right]$$

Linear Programming Duality Theorem

♣ The origin of the LP dual of a Linear Programming program

$$\mathrm{Opt}(P) = \min_x \left\{ c^T x : Ax \geq b \right\} \qquad (P)$$

is the desire to get a systematic way to bound from below the optimal value in (P).
The conceptually simplest bounding scheme is *linear aggregation of the constraints*:
Observation: *For every vector λ of nonnegative weights, the constraint*

$$[A^T\lambda]^T x \equiv \lambda^T Ax \geq \lambda^T b$$

is a consequence of the constraints of (P) and as such is satisfied at every feasible solution of (P).
Corollary III.3 *For every vector $\lambda \geq 0$ such that $A^T\lambda = c$, the quantity $\lambda^T b$ is a lower bound on $\mathrm{Opt}(P)$.*
♣ *The problem dual to (P) is nothing but the problem*

$$\mathrm{Opt}(D) = \max_\lambda \left\{ b^T\lambda : \lambda \geq 0, A^T\lambda = c \right\} \qquad (D)$$

of maximizing the lower bound on $\mathrm{Opt}(P)$ given by Corollary III.3.

♣ The origin of (D) implies the following

Weak Duality Theorem: *The value of the primal objective at every feasible solution of the primal problem*

$$\mathrm{Opt}(P) = \min_x \left\{ c^T x : Ax \geq b \right\} \qquad (P)$$

is $\geq$ *the value of the dual objective at every feasible solution to the dual problem*

$$\mathrm{Opt}(D) = \max_\lambda \left\{ b^T \lambda : \lambda \geq 0, A^T \lambda = c \right\} \qquad (D)$$

that is,

$$\left.\begin{array}{l} x \text{ is feasible for } (P) \\ \lambda \text{ is feasible for } (D) \end{array}\right\} \Rightarrow c^T x \geq b^T \lambda$$

In particular,

$$\mathrm{Opt}(P) \geq \mathrm{Opt}(D).$$

♣ **LP Duality Theorem:** *Consider an LP program along with its dual:*

$$\begin{array}{rcll} \mathrm{Opt}(P) & = & \min\limits_{x} \left\{c^T x : Ax \geq b\right\} & (P) \\ \mathrm{Opt}(D) & = & \max\limits_{\lambda} \left\{b^T \lambda : A^T \lambda = c, \lambda \geq 0\right\} & (D) \end{array}$$

Then

♢ *Duality is symmetric: the problem dual to dual is (equivalent to) the primal*

♢ *The value of the dual objective at every dual feasible solution is $\leq$ the value of the primal objective at every primal feasible solution*

♢ *The following 5 properties are equivalent to each other:*

- *(i)* (P) *is feasible and bounded (below)*
- *(ii)* (D) *is feasible and bounded (above)*
- *(iii)* (P) *is solvable*
- *(iv)* (D) *is solvable*
- *(v)* *both* (P) *and* (D) *are feasible*

and whenever they take place, one has $\mathrm{Opt}(P) = \mathrm{Opt}(D)$.

$$\begin{aligned} \mathrm{Opt}(P) &= \min_x \left\{ c^T x : Ax \geq b \right\} \qquad (P) \\ \mathrm{Opt}(D) &= \max_\lambda \left\{ b^T \lambda : A^T \lambda = c, \lambda \geq 0 \right\} \qquad (D) \end{aligned}$$

♢ *Duality is symmetric*

Proof: Rewriting (D) in the form of (P), we arrive at the problem

$$\min_\lambda \left\{ -b^T \lambda : \begin{bmatrix} A^T \\ -A^T \\ I \end{bmatrix} \lambda \geq \begin{bmatrix} c \\ -c \\ 0 \end{bmatrix} \right\},$$

with the dual being

$$\begin{gathered} \max_{u,v,w} \left\{ c^T u - c^T v + 0^T w : \begin{array}{c} u \geq 0, v \geq 0, w \geq 0, \\ Au - Av + w = -b \end{array} \right\} \\ \Updownarrow \\ \max_{x=v-u,w} \left\{ -c^T x : w \geq 0, Ax = b + w \right\} \\ \Updownarrow \\ \min_x \left\{ c^T x : Ax \geq b \right\} \end{gathered}$$

♢ *The value of the dual objective at every dual feasible solution is* $\leq$ *the value of the primal objective at every primal feasible solution*

This is Weak Duality

♢ *The following 5 properties are equivalent to each other:*

(P) is feasible and bounded below (i)

$\Downarrow$

(D) is solvable (iv)

Indeed, by origin of Opt(P), the inequality

$$c^T x \geq \text{Opt}(P)$$

is a consequence of the (solvable!) system of inequalities

$$Ax \geq b.$$

By Principle B, the inequality is a linear consequence of the system:

$$\exists \lambda \geq 0 : A^T \lambda = c \ \& \ b^T \lambda \geq \text{Opt}(P).$$

Thus, the dual problem has a feasible solution with the value of the dual objective $\geq$ Opt(P). By Weak Duality, this solution is dual optimal, and Opt(D) = Opt(P).

♢ *The following properties are equivalent to each other:*

(D) is solvable (iv)

$\Downarrow$

(D) is feasible and bounded above (ii)

Evident

♢ *The following 5 properties are equivalent to each other:*

(D) is feasible and bounded above (ii)

$\Downarrow$

(P) is solvable (iii)

Implied by already proved relation

(P) is feasible and bounded below (i)

$\Downarrow$

(D) is solvable (iv)

in view of primal-dual symmetry

♢ *The following 5 properties are equivalent to each other:*

(P) is solvable (iii)

$\Downarrow$

(P) is feasible and bounded below (i)

Evident

We proved that

$$\text{(i)} \Leftrightarrow \text{(ii)} \Leftrightarrow \text{(iii)} \Leftrightarrow \text{(iv)}$$

and that when these 4 equivalent properties take place, one has

$$\text{Opt}(P) = \text{Opt}(D)$$

It remains to prove that properties (i) – (iv) are equivalent to

$$\text{both } (P) \text{ and } (D) \text{ are feasible} \qquad \text{(v)}$$

♢ In the case of (v), (P) is feasible and below bounded (Weak Duality), so that (v)⇒(i)
♢ in the case of (i)≡(ii), both (P) and (D) are feasible, so that (i)⇒(v)

Optimality Conditions in LP

Theorem III.2 *Consider a primal-dual pair of feasible LP programs*

$$\begin{aligned} \mathrm{Opt}(P) &= \min_x \left\{ c^T x : Ax \geq b \right\} && (P) \\ \mathrm{Opt}(D) &= \max_\lambda \left\{ b^T \lambda : A^T \lambda = c, \lambda \geq 0 \right\} && (D) \end{aligned}$$

and let x, λ *be* **feasible** *solutions to the respective programs. These solutions are optimal for the respective problems*
♢ *iff* $c^T x - b^T \lambda = 0$ ["zero duality gap"]
as well as
♢ *iff* $[Ax - b]_i \cdot \lambda_i = 0$ *for all* i ["complementary slackness"]

Proof: Under Theorem's premise, $\mathrm{Opt}(P) = \mathrm{Opt}(D)$, so that

$$c^T x - b^T \lambda = \underbrace{c^T x - \mathrm{Opt}(P)}_{\geq 0} + \underbrace{\mathrm{Opt}(D) - b^T \lambda}_{\geq 0}$$

Thus, duality gap $c^T x - b^T \lambda$ is always nonnegative and is zero iff x, λ are optimal for the respective problems.

- The complementary slackness condition is given by the identity

$$c^T x - b^T \lambda = (A^T \lambda)^T x - b^T \lambda = [Ax - b]^T \lambda$$

Since both $[Ax - b]$ and λ are nonnegative, duality gap is zero iff the complementary slackness

$$[Ax - b]_i \lambda_i = 0 \ \forall i$$

holds true.

Shadow Prices

♣ Optimality conditions in LP tell us an instructive and nontrivial story. For the sake of this story, let us replace $[A, b]$ with $-[A, b]$, resulting in

$$\begin{array}{rcll} \mathrm{Opt}(P) & = & \min\limits_x \left\{c^T x : Ax \le b\right\} & (P) \\ \mathrm{Opt}(D) & = & \max\limits_\lambda \left\{-b^T\lambda : c + A^T\lambda = 0, \lambda \ge 0\right\} & (D) \end{array} \qquad [A = [a_1^T; ...; a_m^T]]$$

and optimality conditions for a feasible solution x_* to (P) becoming

$$\exists \lambda^* \ge 0 : \left\{ \begin{array}{rcll} c + \sum_i \lambda_i^* a_i & = & 0 & \text{[Karush-Kuhn-Tucker equation]} \\ \lambda_i^*[b_i - a_i^T x_*] & = & 0, 1 \le i \le m & \text{[complementary slackness]} \end{array} \right.$$

♠ To tell the story, let us interpret
— $c^T x$ as the loss (minus profit) incurred when implementing a decision $x \in \mathbf{R}^n$,
— $a_i^T x$ as the amount of resources of type i (manpower, raw materials of different types, etc.) required to implement a decision x,
— b_i - as the amount of resources of type i in your possession.

With this interpretation. (P) becomes the problem of selecting a decision obeying given upper bounds on the resources and minimizing under these restrictions your loss.

♠ Consider now another decision-making environment, where the resources can be bought and solved at per-unit prices $\lambda_i \ge 0$ (*"shadow prices"*). You, as before, have at your possession b_i units of resource i, $i \le m$, and are allowed to make a *whatever* decision $x \in \mathbf{R}^n$, but should, on the top of your "actual loss" $c^T x$, buy $a_i^T x - b_i$ units of resource i which is in shortage (i.e., $a_i^T x > b_i$) and sell $b_i - a_i^T x$ units of resource i which is in abundance (i.e., $a_i^T x \le b_i$). In this model, your total loss incurred by a decision $x \in \mathbf{R}^n$ is

$$[c^T x + \sum_i \lambda_i [a_i^T x - b_i] = \left[c + \sum_i \lambda_i a_i\right]^T x - \sum_i \lambda_i b_i.$$

As before, you want to minimize your total loss.

$$\begin{array}{rcll} \mathrm{Opt}(P) & = & \min\limits_x \left\{c^T x : Ax \le b\right\} & (P) \\ \mathrm{Opt}(D) & = & \max\limits_\lambda \left\{-b^T\lambda : c + A^T\lambda = 0, \lambda \ge 0\right\} & (D) \end{array} \qquad [A = [a_1^T; ...; a_m^T]]$$

• Original loss: $v^T x$ • New loss: $c^T x + \sum_i \lambda_i [a_i^T x - b_i]$, $\lambda_i \ge 0$.

Let us make two observations:

A. The new environment is better for you than the initial one: whenever x is feasible for (P), your new loss is $\le$ the "actual loss" $c^T x$, with the new loss equal to the initial one iff the complementary slackness holds, that is, iff $\lambda_i[a_i^T x - b_i] = 0$ for all i

B. Your new loss is just linear in x, and you can make it as negative as you wish, *unless* $[c + \sum_i \lambda_i a_i = 0$; in this latter case, your new loss is equal to $-\sum_i \lambda_i b_i$ identically in x, and every x minimizes it.

As an immediate corollary, we see that

If x_* is a feasible solution to (P) such that for some nonnegative shadow prices λ_i^* satisfying the complementary slackness condition $\lambda_i^*[b_i - a_i^T x_* = 0$ for all i, x_* is an unconstrained minimizer of your new loss (that is, $c + \sum_i \lambda_i^* = 0$), *then x^* is an optimal solution to (P).*

Indeed, were there a feasible solution $\bar{x}$ to (P) with $c^T\bar{x} < c^T x_*$, the new loss $c^T x + \sum_i \lambda_i^*[a_i^T x - b_i]$ as evaluated at $x = \bar{x}$ would be $\le c^T\bar{x} < c^T x_*$ (as the new loss underestimates the actual one at every feasible solution to (P)). This is impossible, since by complementary slackness the new loss as evaluated at x_* is the same as the actual one, and x_* minimizes this new loss over *all* decisions $x \in \mathbf{R}^n$.

♠ We see that *the existence of nonnegative shadow prices λ_i^* which, taken together with a feasible solution x_* to (P), satisfy complementary slackness and the condition $c + \sum_i \lambda_i^* a_i = 0$, is sufficient for x_* to be an optimal solution to (P).*

We immediately see that what the above sufficient condition wants from the shadow prices is exactly to be feasible for (D) and to be linked to x_* via complementary slackness, so that LP optimality conditions say to us that the above *sufficient* condition for primal optimality of a primal-feasible solution x_* is in fact *necessary and sufficient.*

Helly Theorem: Alternative proof

♣ Here we present an alternative proof of Helly Theorem; this proof does not use Radon Theorem.

Polyhedral version of Helly Theorem: *Let $A_i = \{x : P_i x \leq p_i\}$, $i \leq N$, be a family of polyhedral sets in $\mathbf{R}^n$ with $N \geq n+1$. If every $n+1$ sets from the family have a points in common, then all sets from the family have a common point.*

Proof. All we need to prove is that *If $\cap_{i \leq N} A_i = \emptyset$, then the intersection of properly selected $k \leq n+1$ sets from the family is empty as well.*
Thus, assume that $\cap_i A_i = \emptyset$, that is, the system of linear inequalities

$$P_i x \leq p_i, i \leq N \qquad (*)$$

in variables $x \in \mathbf{R}^n$ has no solutions. By General Theorem of Alternative this means that the weighted sum, with nonnegative weights, of inequalities from the system is a contradictory inequality $0^T x \leq \beta$ with $\beta < 0$. This is the same as to say that the vector $f = [0; ...; 0, \beta]$ is a conic combination of the $n+1$-dimensional vectors of coefficients of our inequalities (n coefficients at variables in the left hand side of inequality augmented by its right hand side). By Caratheodory Theorem in conic form, we can select from our scalar inequalities at most $n+1$ in such a way that f is a conic combination of the selected vectors, implying, by the same General Theorem of Alternative, that the system composed of selected inequalities has no solutions. This, in turn, implies that system of vector inequalities $(*)$ has at most $n+1$-element infeasible subsystem, that is, properly selecting at most $(n+1)$ of the sets A_i, we get a collection with empty intersection. Q.E.D.

From polyhedral to general case. To extract from the just proved Helly Theorem for polyhedral sets Helly Theorem *per se*, consider the family of $N \geq n+1$ convex sets $A_i \subset \mathbf{R}^n$, and assume that every $n+1$ of these sets have a point in common; we want to prove that all N sets have a common point.

Let $\mathcal{I}$ be the set of all collections $\iota = \{i_1, ..., i_{n+1}\}$ of $n+1$ distinct from each other indexes from $I = \{1, ..., N\}$. By our premise, for $\iota \in \mathcal{I}$ there exists $x_\iota \in \cap_{i \in \iota} A_i$. For $i \in I$, let us set $B_i = \mathrm{Conv}\{x_\iota : \iota \in \mathcal{I} \ \& \ i \in \iota\}$. Then

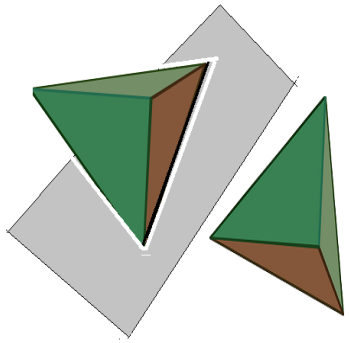
- B_i is polyhedral (Corollary III.1),
- $B_i \subset A_i$ (since $x_\iota \in A_i$ when $i \in \iota$ and A_i is convex), and
- every $n+1$ of the sets $B_i, i \in I$, have a point in common

[indeed, given $i_1, ..., i_{n+1} \in I$, there exists $\iota \in \mathcal{I}$ with $i_s \in \iota$, $s \leq n+1$, whence by construction $x_\iota \in \cap_s B_{i_s}$]

Applying Polyhedral Helly Theorem, we conclude that $\cap_{i \in I} B_i \neq \emptyset$, and since $\cap_i A_i \supset \cap_i B_i$, $\cap_i A_i$ is nonempty as well, Q.E.D.

Lecture I.4
Separation and Extreme Points

Separation Theorem
Supporting planes and Extreme points
Krein-Milman Theorem
Dual cone
Extreme rays and Krein-Milman Theorem in conic form
Dubovitski-Milutin Lemma
Polar of a convex set
Geometry of polyhedral sets



Separation Theorem

♣ Every linear form $f(x)$ on $\mathbf{R}^n$ is representable via inner product:

$$f(x) = f^T x$$

for appropriate vector $f \in \mathbf{R}^n$ uniquely defined by the form. Nontrivial (not identically zero) forms correspond to nonzero vectors f.

♣ A *level set*

$$M = \{x : f^T x = a\} \qquad (*)$$

of a *nontrivial* linear form on $\mathbf{R}^n$ is affine subspace of affine dimension $n-1$; vice versa, every affine subspace M of affine dimension $n-1$ in $\mathbf{R}^n$ can be represented by $(*)$ with appropriately chosen $f \neq 0$ and a; f and a are defined by M up to multiplication by a common nonzero factor.

$(n-1)$-dimensional affine subspaces in $\mathbf{R}^n$ are called *hyperplane*.

$$M = \{x : f^T x = a\} \qquad (*)$$

♣ Level set $(*)$ of nontrivial linear form splits $\mathbf{R}^n$ into two parts:

$$\begin{array}{rcl} M_+ & = & \{x : f^T x \geq a\} \\ M_- & = & \{x : f^T x \leq a\} \end{array}$$

called *closed half-spaces* given by (f, a); the hyperplane M is the common boundary of these half-spaces. The interiors M_{++} of M_+ and M_{--} of M_- are given by

$$\begin{array}{rcl} M_{++} & = & \{x : f^T x > a\} \\ M_{--} & = & \{x : f^T x < a\} \end{array}$$

and are called *open half-spaces* given by (f, a). We have

$$\mathbf{R}^n = M_- \bigcup M_+ \quad [M_- \bigcap M_+ = M]$$

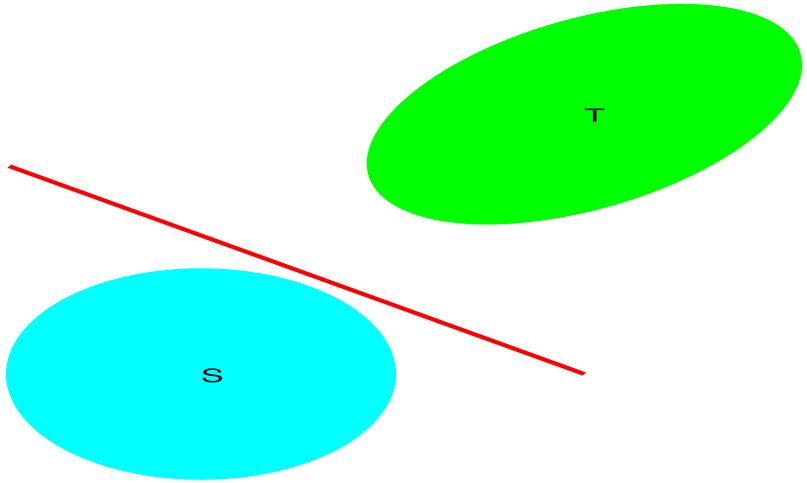
and

$$\mathbf{R}^n = M_{--} \bigcup M \bigcup M_{++}$$

♣ **Definition.** Let T, S be two nonempty sets in $\mathbf{R}^n$.
(i) We say that a hyperplane

$$M = \{x : f^T x = a\} \qquad (*)$$

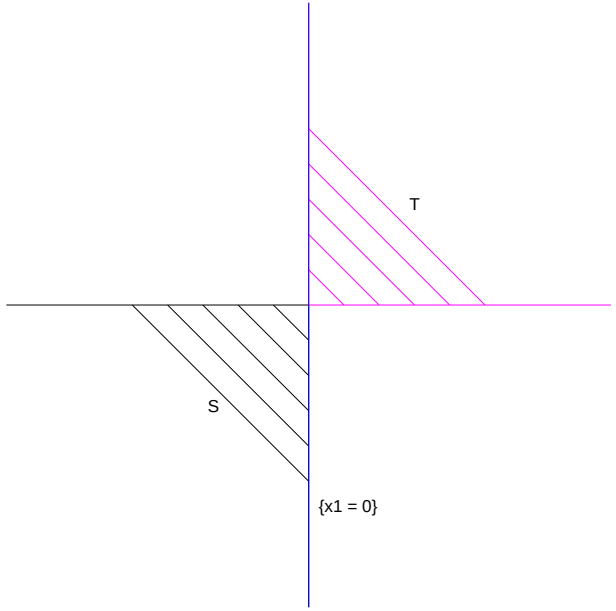
separates S and T, if
♢ $S \subset M_-$, $T \subset M_+$ ("S does not go above M, and T does not go below M")
and
♢ $S \cup T \not\subset M$.
(ii) We say that a nontrivial linear form $f^T x$ separates S and T if, for properly chosen a, the hyperplane $(*)$ separates S and T.



Red hyperplane $2x_1 + 3x_2 = 6$ separates cyan set S and green set T

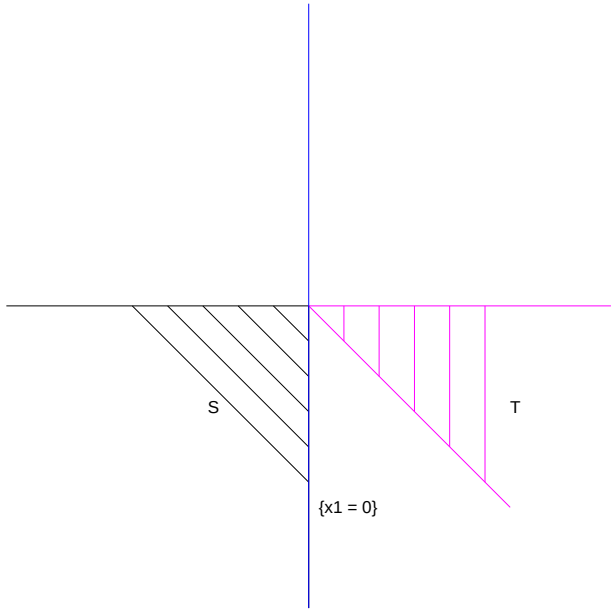
Examples: The linear form x_1 on $\mathbf{R}^2$
1) separates the sets

$$\begin{aligned} S &= \{x \in \mathbf{R}^2 : x_1 \leq 0, x_2 \leq 0\}, \\ T &= \{x \in \mathbf{R}^2 : x_1 \geq 0, x_2 \geq 0\} : \end{aligned}$$



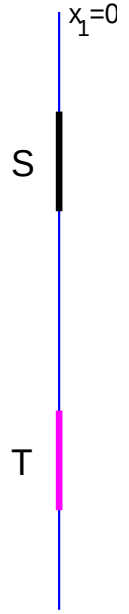
The linear form x_1 on $\mathbf{R}^2$...
2) separates the sets

$$\begin{aligned} S &= \{x \in \mathbf{R}^2 : x_1 \leq 0, x_2 \leq 0\}, \\ T &= \{x \in \mathbf{R}^2 : x_1 + x_2 \geq 0, x_2 \leq 0\} : \end{aligned}$$



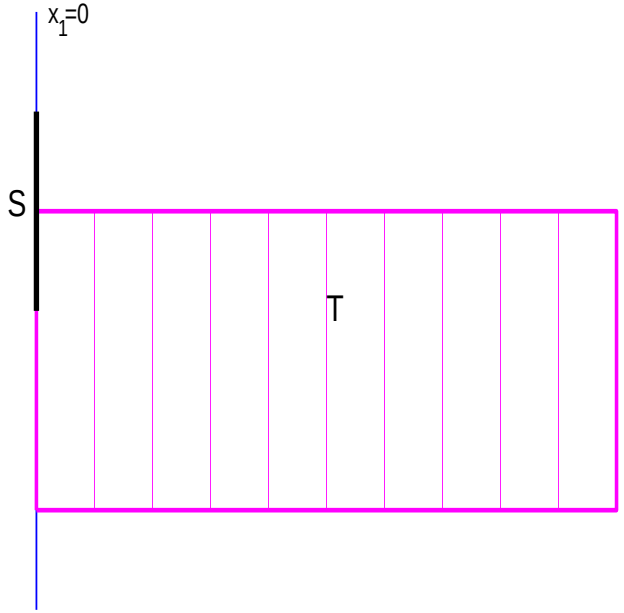
The linear form x_1 on $\mathbf{R}^2$...
3) does **not** separate the sets

$$\begin{aligned} S &= \{x \in \mathbf{R}^2 : x_1 = 0, 1 \leq x_2 \leq 2\}, \\ T &= \{x \in \mathbf{R}^2 : x_1 = 0, -2 \leq x_2 \leq -1\} : \end{aligned}$$



The linear form x_1 on $\mathbf{R}^2$...
4) separates the sets

$$\begin{aligned} S &= \{x \in \mathbf{R}^2 : x_1 = 0, 0 \leq x_2 \leq 2\}, \\ T &= \{x \in \mathbf{R}^2 : 0 \leq x_1 \leq 1, -2 \leq x_2 \leq 1\} : \end{aligned}$$



Observation: *A linear form $f^T x$ separates nonempty sets S, T iff*

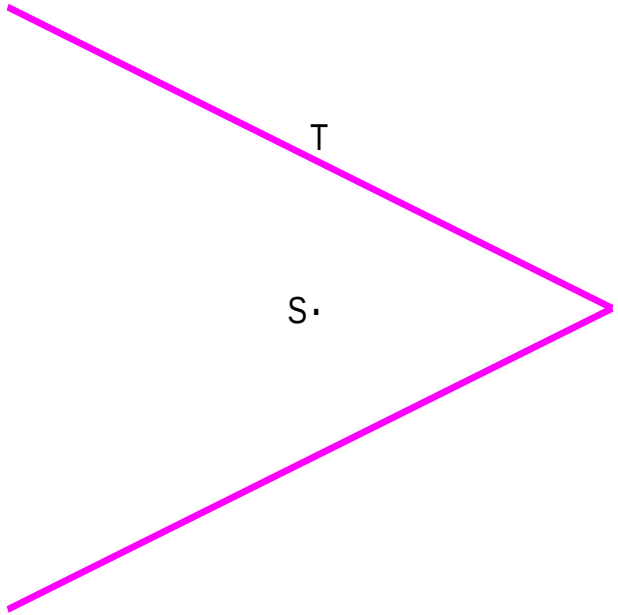
$$\begin{array}{rcl} \sup\limits_{x\in S} f^T x & \leq & \inf\limits_{y\in T} f^T y \\ \inf\limits_{x\in S} f^T x & < & \sup\limits_{y\in T} f^T y \end{array} \qquad (*)$$

In the case of $(*)$, the associated with f hyperplane separating S and T are exactly the hyperplane

$$\{x : f^T x = a\} \text{ with } \sup_{x\in S} f^T x \leq a \leq \inf_{y\in T} f^T y.$$

Fact IV.1 [Separation Theorem] *Two nonempty* **convex** *sets S, T can be separated iff their relative interiors do not intersect.*

Note: In this statement, convexity of both S and T is crucial!



Proof, $\Rightarrow$: (!) If *nonempty convex sets S, T can be separated,* **then** $\operatorname{rint} S \bigcap \operatorname{rint} T = \emptyset$

Lemma. *Let X be a convex set, $f(x) = f^T x$ be a linear form and $a \in \operatorname{rint} X$. Then*

$$f^T a = \max_{x \in X} f^T x \Leftrightarrow f(\cdot)\Big|_X = \text{const}.$$

♣ **Lemma $\Rightarrow$ (!):** Let $a \in \operatorname{rint} S \cap \operatorname{rint} T$. Assume, on contrary to what should be proved, that $f^T x$ separates S, T, so that

$$\sup_{x \in S} f^T x \leq \inf_{y \in T} f^T y.$$

♢ Since $a \in T$, we get $f^T a \geq \sup_{x \in S} f^T x$, that is, $f^T a = \max_{x \in S} f^T x$. By Lemma, $f^T x = f^T a$ for all $x \in S$.

♢ Since $a \in S$, we get $f^T a \leq \inf_{y \in T} f^T y$, that is, $f^T a = \min_{y \in T} f^T y$. By Lemma, $f^T y = f^T a$ for all $y \in T$.

Thus,

$$z \in S \cup T \Rightarrow f^T z \equiv f^T a,$$

so that f does **not** separate S and T, which is a contradiction.

Lemma. *Let X be a convex set, $f(x) = f^T x$ be a linear form and $a \in \operatorname{rint} X$. Then*

$$f^T a = \max_{x \in X} f^T x \Leftrightarrow f(\cdot)\Big|_X = \text{const.}$$

Proof. Shifting X, we may assume $a = 0$. Let, on the contrary to what should be proved, $f^T x$ be non-constant on X, so that there exists $y \in X$ with $f^T y \neq f^T a = 0$. The case of $f^T y > 0$ is impossible, since $f^T a = 0$ is the maximum of $f^T x$ on X. Thus, $f^T y < 0$. The line $\{ty : t \in \mathbf{R}\}$ passing through 0 and through y belongs to $\operatorname{Aff}(X)$; since $0 \in \operatorname{rint} X$, all points $z = -\epsilon y$ on this line belong to X, provided that $\epsilon > 0$ is small enough. At every point of this type, $f^T z > 0$, which contradicts the fact that $\max\limits_{x \in X} f^T x = f^T a = 0$.

Proof, $\Leftarrow$: Assume that S, T are nonempty convex sets such that $\mathrm{rint}\, S \cap \mathrm{rint}\, T = \emptyset$, and let us prove that S, T can be separated.

Step 1: Separating a point and a convex hull of a finite set. Let $S = \mathrm{Conv}(\{b_1, ..., b_m\})$ and $T = \{b\}$ with $b \notin S$, and let us prove that S and T can be separated.

Indeed,

$$S = \mathrm{Conv}(b_1, ..., b_m) = \left\{ x : \exists \lambda : \begin{array}{l} \lambda \geq 0, \sum_i \lambda_i = 1 \\ x = \sum_i \lambda_i b_i \end{array} \right\}$$

is polyhedral representable and thus is polyhedral:

$$S = \{x : a_\ell^T x \leq \alpha_\ell, \ell \leq L\}.$$

Since $b \notin S$, for some $\bar{\ell}$ we have

$$a_{\bar{\ell}}^T b > \alpha_{\bar{\ell}} \geq \sup_{x \in S} a_{\bar{\ell}}^T x$$

which is the desired separation.

Step 2: Separating a point and a convex set which does not contain the point. Let S be a nonempty convex set and $T = \{b\}$ with $b \notin S$, and let us prove that S and T can be separated.

1^0. Shifting S and T by $-b$ (which clearly does not affect the possibility of separating the sets), we can assume that $T = \{0\} \not\subset S$.

2^0. Replacing, if necessary, $\mathbf{R}^n$ with $\mathrm{Lin}(S)$, we may further assume that $\mathbf{R}^n = \mathrm{Lin}(S)$.

Recall that ever nonempty subset X of $\mathbf{R}^n$ is separable (Lecture I.1), that is, there exists a countable subset $\{x_i, i \geq 1\}$ of X *dense in X* – such that every point of X is the limit of a sequence of points from the subset. Let $\{x_i \in S\}_i$ be a countable set which is dense in S. Since S is convex and does not contain 0, we have

$$0 \notin \mathrm{Conv}(\{x_1, ..., x_i\}) \ \forall i$$

whence

$$\exists f_i : 0 = f_i^T 0 > \max_{1 \leq j \leq i} f_i^T x_j. \qquad (*)$$

By scaling, we may assume that $\|f_i\|_2 = 1$.

The sequence $\{f_i\}$ of unit vectors possesses a converging subsequence $\{f_{i_s}\}_{s=1}^{\infty}$; the limit f of this subsequence is, of course, a unit vector. By $(*)$, for every fixed j and all large enough s we have $f_{i_s}^T x_j < 0$, whence

$$f^T x_j \leq 0 \ \forall j. \qquad (**)$$

Since $\{x_j\}$ is dense in S, $(**)$ implies that $f^T x \leq 0$ for all $x \in S$, whence

$$\sup_{x \in S} f^T x \leq 0 = f^T 0.$$

Situation: (a) $\mathrm{Lin}(S) = \mathbf{R}^n$
(b) $T = \{0\}$
(c) We have built a unit vector f such that

$$\sup_{x \in S} f^T x \leq 0 = f^T 0. \qquad (!)$$

By (!), all we need to prove that f separates $T = \{0\}$ and S is to verify that

$$\inf_{x \in S} f^T x < f^T 0 = 0.$$

Assuming the opposite, (!) would say that $f^T x = 0$ for all $x \in S$, which is impossible, since $\mathrm{Lin}(S) = \mathbf{R}^n$ and f is nonzero.

Step 3: Separating two non-intersecting nonempty convex sets. Let S, T be nonempty convex sets which do not intersect; let us prove that S, T can be separated.

Let $\widehat{S} = S - T$ and $\widehat{T} = \{0\}$. The set $\widehat{S}$ clearly is convex and does not contain 0 (since $S \cap T = \emptyset$). By Step 2, $\widehat{S}$ and $\{0\} = \widehat{T}$ can be separated: there exists f such that

$$\begin{cases} \overbrace{\sup_{x\in S,y\in T}[f^Tx - f^Ty]}^{\sup_{x\in S} f^Ts - \inf_{y\in T} f^Ty} & \leq & 0 = \inf_{z\in\{0\}} f^Tz \\ \underbrace{\inf_{x\in S,y\in T}[f^Tx - f^Ty]}_{\inf_{x\in S} f^Tx - \sup_{y\in T} f^Ty} & < & 0 = \sup_{z\in\{0\}} f^Tz \end{cases}$$

whence

$$\begin{array}{rcl} \sup_{x\in S} f^Tx & \leq & \inf_{y\in T} f^Ty \\ \inf_{x\in S} f^Tx & < & \sup_{y\in T} f^Ty \end{array}$$

Step 4: Completing the proof of Separation Theorem. Finally, let S, T be nonempty convex sets with non-intersecting relative interiors, and let us prove that S, T can be separated.
As we know, the sets $S' = \operatorname{rint} S$ and $T' = \operatorname{rint} T$ are convex and nonempty; we are in the situation when these sets do not intersect. By Step 3, S' and T' can be separated: for properly chosen f, one has

$$\begin{array}{rcl} \sup\limits_{x\in S'} f^T x & \leq & \inf\limits_{y\in T'} f^T y \\ \inf\limits_{x\in S'} f^T x & < & \sup\limits_{y\in T'} f^T y \end{array} \qquad (*)$$

Since S' is dense in S and T' is dense in T, inf's and sup's in $(*)$ remain the same when replacing S' with S and T' with T. Thus, f separates S and T.

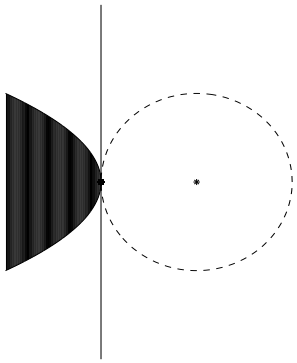
♣ Alternative proof of Separation Theorem starts with separating a point $T = \{a\}$ and a *closed* convex set S, $a \notin S$, and is based on the following fact:

Let S be a nonempty closed convex set and let $a \notin S$. There exists a unique closest to a point in S:

$$\mathrm{Proj}_S(a) = \operatorname*{argmin}_{x \in S} \|a - x\|_2$$

and the vector $e = a - \mathrm{Proj}_S(a)$ separates a and S:

$$\max_{x \in S} e^T x = e^T \mathrm{Proj}_S(a) = e^T a - \|e\|_2^2 < e^T a.$$



Proof: **1^0.** The closest to a point in S does exist. Indeed, let $x_i \in S$ be a sequence such that

$$\|a - x_i\|_2 \to \inf_{x \in S} \|a - x\|_2, \ , \ i \to \infty$$

The sequence $\{x_i\}$ clearly is bounded; passing to a subsequence, we may assume that $x_i \to \bar{x}$ as $i \to \infty$. *Since S is closed*, we have $\bar{x} \in S$, and

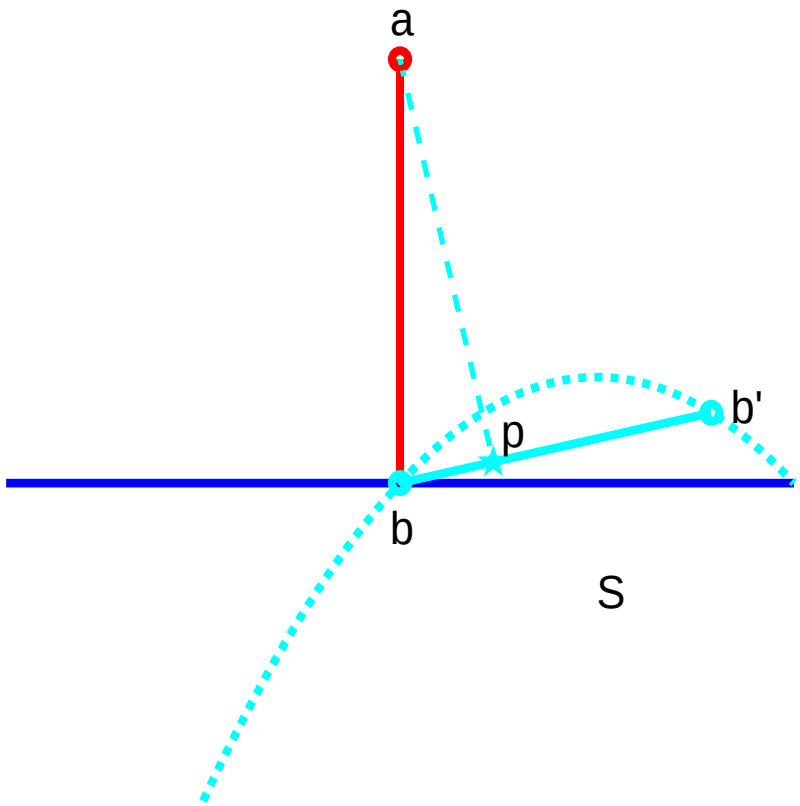
$$\|a - \bar{x}\|_2 = \lim_{i \to \infty} \|a - x_i\|_2 = \inf_{x \in S} \|a - x\|_2.$$

2^0. The closest to a point in S is unique. Indeed, let x, y be two closest to a points in S, so that $\|a - x\|_2 = \|a - y\|_2 = d$. *Since S is convex*, the point $z = \frac{1}{2}(x + y)$ belongs to S; therefore $\|a - z\|_2 \geq d$. We now have

$$\begin{array}{l} \overbrace{\|[a-x]+[a-y]\|_2^2}^{=\|2(a-z)\|_2^2 \geq 4d^2} + \overbrace{\|[a-x]-[a-y]\|_2^2}^{=\|x-y\|^2} \\ = \underbrace{2\|a-x\|_2^2 + 2\|a-y\|_2^2}_{4d^2} \end{array}$$

whence $\|x - y\|_2 = 0$.

3^0. Thus, the closest to a point $b = \mathrm{Proj}_S(a)$ in S exists, is unique and differs from a (since $a \notin S$). The hyperplane passing through b and orthogonal to $a - b$ separates a and S:



Indeed, if there were a point $b' \in S$ "above" the hyperplane, the entire segment $[b, b']$ would be contained in S by convexity of S. Since the angle $\angle abb'$ is $< \pi/2$, performing a small step from b towards b' we stay in S and become closer to a, which is impossible!
With $e = a - \mathrm{Proj}_S(a)$, we have

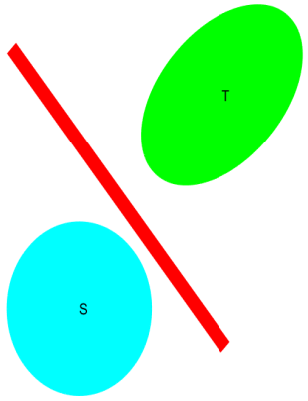
$$\begin{array}{rcl} & & x \in S, f = x - \mathrm{Proj}_S(a) \\ & & \Downarrow \\ \phi(t) & \equiv & \|e - tf\|_2^2 = \|a - [\mathrm{Proj}_S(a) + t(x - \mathrm{Proj}_S(a))]\|_2^2 \\ & \geq & \|a - \mathrm{Proj}_S(a)\|_2^2 = \phi(0), 0 \leq t \leq 1 \\ & \Rightarrow & 0 \leq \phi'(0) = -2e^T(x - \mathrm{Proj}_S(a)) \\ & & \Downarrow \\ & & \forall x \in S : e^T x \leq e^T \mathrm{Proj}_S(a) = e^T a - \|e\|_2^2. \end{array}$$

♣ Separation of sets S, T by linear form $f^T x$ is called *strict*, if

$$\sup_{x \in S} f^T x < \inf_{y \in T} f^T y$$

Geometrically: For properly selected $\delta > 0$ and a, S and T are separated by the stripe $\{x : a - \delta \leq f^T x \leq a + \delta\}$:

$$\sup_{x \in S} f^T x \leq a - \delta < a + \delta \leq \inf_{y \in T} f^T y$$



Fact IV.2 *Let S, T be nonempty convex sets. These sets can be strictly separated iff they are at positive distance:*

$$\mathrm{dist}(S, T) = \inf_{x \in S, y \in T} \|x - y\|_2 > 0.$$

Proof, $\Rightarrow$: Let f strictly separate S, T; let us prove that S, T are at positive distance. Otherwise we could find sequences $x_i \in S$, $y_i \in T$ with $\|x_i - y_i\|_2 \to 0$ as $i \to \infty$, whence $f^T(y_i - x_i) \to 0$ as $i \to \infty$. It follows that the sets on the axis

$$\widehat{S} = \{a = f^T x : x \in S\}, \widehat{T} = \{b = f^T y : y \in T\}$$

are at zero distance, which is a contradiction with

$$\sup_{a \in \widehat{S}} a < \inf_{b \in \widehat{T}} b.$$

Proof, $\Leftarrow$: Let T, S be nonempty convex sets which are at positive distance 2δ:

$$2\delta = \inf_{x \in S, y \in T} \|x - y\|_2 > 0.$$

Let

$$S^+ = S + \{z : \|z\|_2 \leq \delta\}$$

The sets S^+ and T are convex and do not intersect, and thus can be separated:

$$\sup_{x_+ \in S^+} f^T x_+ \leq \inf_{y \in T} f^T y \qquad [f \neq 0]$$

Since

$$\begin{aligned} \sup_{x_+ \in S^+} f^T x_+ &= \sup_{x \in S, \|z\|_2 \leq \delta} [f^T x + f^T z] \\ &= [\sup_{x \in S} f^T x] + \delta \|f\|_2, \end{aligned}$$

we arrive at

$$\sup_{x \in S} f^T x < \inf_{y \in T} f^T y$$

Quiz Below S is a nonempty convex set and $T = \{a\}$.

Statement	True?
If T and S can be separated then $a \notin S$	
If $a \notin S$, then T and S can be separated	
If T and S can be strictly separated, then $a \notin S$	
If $a \notin S$, then T and S can be strictly separated	
If S is closed and $a \notin S$, then T and S can be strictly separated	

Supporting Planes and Extreme Points

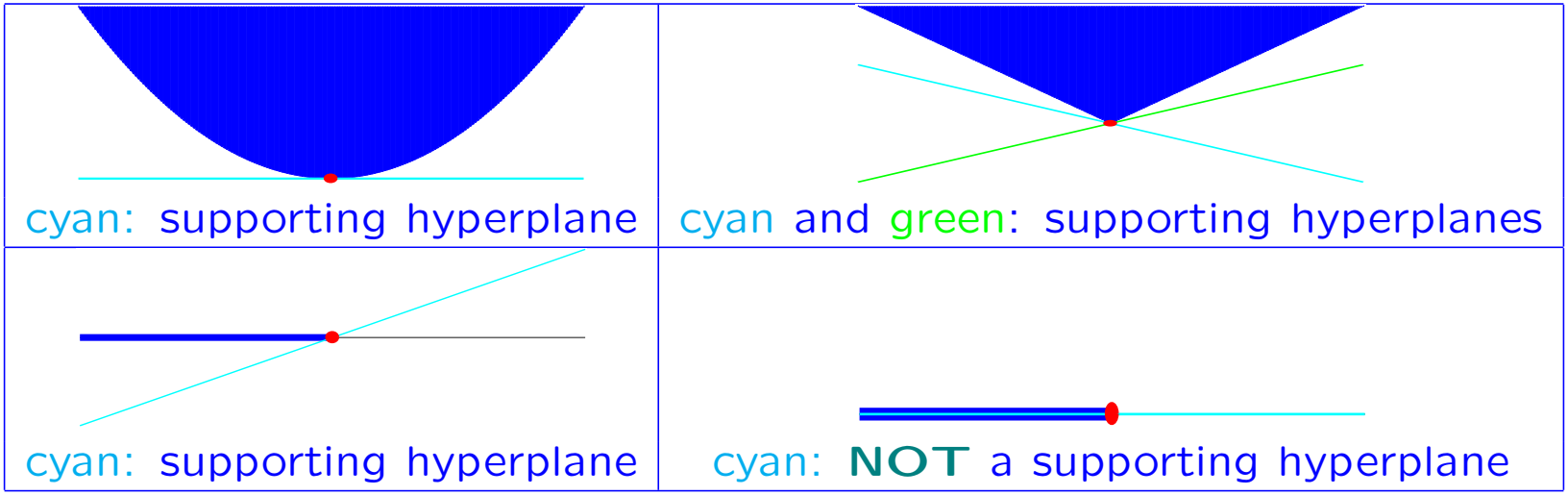
♣ **Definition.** Let Q be a *closed* convex set in $\mathbf{R}^n$ and $\bar{x}$ be a point from the relative boundary of Q. A hyperplane

$$\Pi = \{x : f^T x = a\} \qquad [f \neq 0]$$

is called *supporting to Q at the point $\bar{x}$*, if the hyperplane separates Q and $\{\bar{x}\}$:

$$\begin{array}{lll} \sup\limits_{x\in Q} f^T x & \leq a & \leq f^T\bar{x} \; [\Leftrightarrow \sup_{x\in Q} f^T x = a = f^T\bar{x} \text{ due to } \bar{x}\in Q] \\ \inf\limits_{x\in Q} f^T x & < & f^T\bar{x} \end{array}$$

Equivalently: Hyperplane $\Pi = \{x : f^T x = a\}$ supports Q at $\bar{x}$ iff the linear form $f^T x$ attains its maximum on Q, equal to a, at the point $\bar{x}$ and the form is non-constant on Q.



Q: blue set in 2D; a: red point

Fact IV.3 *Let Q be a convex closed set in $\mathbf{R}^n$ and $\bar{x}$ be a point from the relative boundary of Q. Then*

♢ *There exist at least one hyperplane Π which supports Q at $\bar{x}$;*

♢ *For every such hyperplane Π, the set $Q \cap \Pi$ has dimension less than the one of Q.*

Proof: Existence of supporting plane is given by Separation Theorem. This theorem is applicable since

$$\bar{x} \notin \mathrm{rint}\, Q \Rightarrow \{\bar{x}\} \equiv \mathrm{rint}\, \{\bar{x}\} \cap \mathrm{rint}\, Q = \emptyset.$$

Further,

$$\{\underbrace{Q \cap \Pi}_{\bar{x} \in} \neq \emptyset \ \&\ Q \not\subset \Pi\} \Rightarrow \{\mathrm{Aff}(Q) \not\subset \Pi\} \Rightarrow \{\mathrm{Aff}(\Pi \cap Q) \subset [\mathrm{Aff}(Q) \cap \Pi] \subsetneqq \mathrm{Aff}(Q)\},$$

and if two *distinct* affine subspaces (in our case, $\mathrm{Aff}(\Pi \cap Q)$ and $\mathrm{Aff}(Q)$) are embedded one into another, then the dimension of the embedded subspace is *strictly less* than the dimension of the embedding one.

Extreme Points

♣ **Definition.** Let Q be a convex set in $\mathbf{R}^n$ and $\bar{x}$ be a point of Q. The point is called *extreme*, if it is not a convex combination, with positive weights, of two points of X distinct from $\bar{x}$:

$$\bar{x} \in \mathrm{Ext}(Q)$$
$$\Updownarrow$$
$$\{\bar{x} \in Q\} \ \& \ \left\{ \begin{array}{l} u, v \in Q, \lambda \in (0,1) \\ \bar{x} = \lambda u + (1-\lambda) v \end{array} \right\} \Rightarrow u = v = \bar{x} \Bigg\}$$

Equivalently: A point $\bar{x} \in Q$ is extreme iff it is **not** the midpoint of a nontrivial segment in Q:

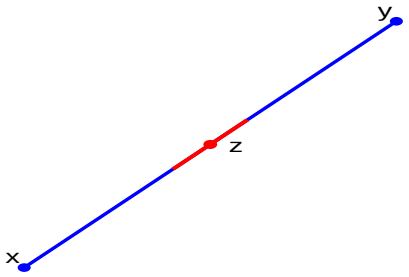
$$\bar{x} \pm h \in Q \Rightarrow h = 0.$$

Equivalently: A point $\bar{x} \in Q$ is extreme iff the set $Q \backslash \{\bar{x}\}$ is convex.

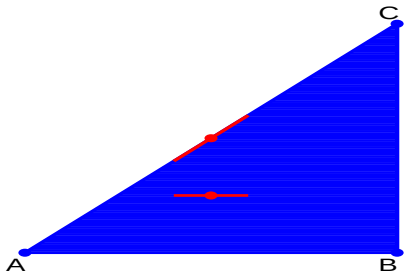
Equivalently: A point $\bar{x} \in Q$ is extreme, if in every representation $\bar{x} = \sum_i \lambda_i x^i$ of the point as a convex combination of points $x^i \in Q$, for all terms with $\lambda_i > 0$ it holds $x^i = \bar{x}$.

Examples:

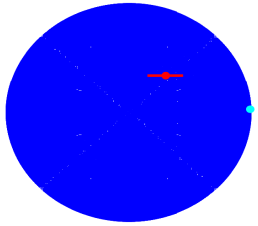
1. Extreme points of $[x, y]$ are ...



2. Extreme points of $\triangle ABC$ are ...



3. Extreme points of the ball $\{x : \|x\|_2 \leq 1\}$ are ...



Quiz, answers

1. Extreme points of $[x, y]$ are the endpoints x and y

2. Extreme points of $\triangle ABC$ are the vertices A, B, C

3. Extreme points of the ball $\{x : \|x\|_2 \leq 1\}$ are the points $\{x : \|x\|_2 = 1\}$ on the boundary of the ball.

Theorem [Krein-Milman] *Let Q be a closed convex and nonempty set in $\mathbf{R}^n$. Then*
♢ *Q possesses extreme points iff Q does not contain lines;*
♢ *If Q is bounded, then Q is the convex hull of its extreme points:*

$$Q = \mathrm{Conv}(\mathrm{Ext}(Q))$$

so that every point of Q is convex combination of extreme points of Q.

Note: If $Q = \mathrm{Conv}(A)$, **then** $\mathrm{Ext}(Q) \subset A$. Thus, extreme points of a *closed convex bounded* set Q give the *minimal* representation of Q as Conv(...).

Proof of KM, A. Let us prove that If closed convex set Q does not contain lines, then Q has extreme points: $\mathrm{Ext}(Q) \neq \emptyset$

Fact IV.4 *Let S be a closed convex set and $\Pi = \{x : f^T x = a\}$ be a hyperplane which supports S at certain point. Then*

$$\mathrm{Ext}(\Pi \cap S) \subset \mathrm{Ext}(S).$$

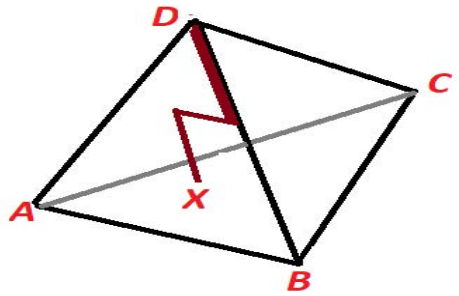
Proof of Fact IV.4. Let $\bar{x} \in \mathrm{Ext}(\Pi \cap S)$; we should prove that $\bar{x} \in \mathrm{Ext}(S)$. Assume, on the contrary, that $\bar{x}$ is a midpoint of a nontrivial segment $[u, v] \subset S$. Then $f^T \bar{x} = a = \max\limits_{x \in S} f^T x$, whence $f^T \bar{x} = \max\limits_{x \in [u,v]} f^T x$. A linear form can attain its maximum on a segment at the midpoint of the segment iff the form is constant on the segment; thus, $a = f^T \bar{x} = f^T u = f^T v$, that is, $[u, v] \subset \Pi \cap S$. But $\bar{x}$ is an extreme point of $\Pi \cap S$ – contradiction!

♠ Let Q be a nonempty closed convex set which does not contain lines. In order to build an extreme point of Q, apply the *Purification algorithm*.
It generates a sequence $Q = S_0 \supset S_1 \supset S_2 \supset ...$ of shrinking closed convex nonempty sets which starts from $S_0 = Q$, along with points $x_t \in S_t$, and is such that
A: *all extreme points of S_t, if any, are extreme points of S_{t-1}* (and therefore are extreme points of $S_0 = Q$), and
B: *whenever S_t is not a singleton, S_{t+1} is well defined and is of dimension strictly less than the dimension of S_t.*
Taking for granted that there is an algorithm capable to produce sequence with these properties, observe that the sequence $S_0 \supset S_t \supset ...$ is finite by **B** (dimension of S_t strictly decreases when passing from S_t to S_{t+1}, and this cannot last indefinitely) and the concluding set S_K in this sequence is a singleton (again by **B**). In particular, S_K has extreme point:

$$S_K = \{\bar{x}\} \Rightarrow \mathrm{Ext}(S_K) = \{\bar{x}\}$$

and by **A** this extreme point is an extreme point of $Q \Rightarrow \mathrm{Ext}(Q) \neq \emptyset$, Q.E.D.

♠ This is how Purification works:
• We start with $S_0 = Q$ and select as x_0 an arbitrary point of S_0
• Given S_t, and $x_t \in S_t$ we check whether S_t is a singleton; if yes, we terminate, otherwise we
— find a point x_{t+1} on the relative boundary of S_t
— build a hyperplane Π_t supporting S_t at x_{t+1}, and set $S_{t+1} = S_t \cap \Pi_t$ **Note:** By construction, S_{t+1}, when defined, is a nonempty closed convex subset of S_t, with $\dim(S_{t+1}) < \dim(S_t)$ (by Fact IV.3) and $\mathrm{Ext}(S_{t+1}) \subset \mathrm{Ext}(S_t)$ (by Fact IV.4), so that we do ensure **A** and **B**.



from point *X* in pyramid *ABCD* to extreme point *D* of the pyramid

Purification Algorithm

♠ Paying debts: How to find a point on the relative boundary of a non-singleton nonempty closed convex set not containing lines?
• To find a point x_{t+1} on the relative boundary of a *non-singleton* closed convex set $S_t \ni x_t$, we take a direction $h \neq 0$ parallel to $\mathrm{Aff}(S_t)$. Since $S_t \subset Q$, S_t does not contain lines
$\Rightarrow$ replacing if necessary h with $-h$, we can assume that the ray

$$\{x_t + sh : s \geq 0\}$$

is not contained in S_t, which combines with closedness of S_t to imply that the largest $s = \bar{s}$ such that $x_t + sh \in S_t$ is well defined
$\Rightarrow$ $x_{t+1} = x_t + \bar{s}h$ is a point from the relative boundary of S_t

Note: Assume you are given a linear form $g^T x$ which is bounded from above on Q. Then in the Purification algorithm one can easily ensure that $g^T x_{t+1} \geq g^T x_t$. Thus,

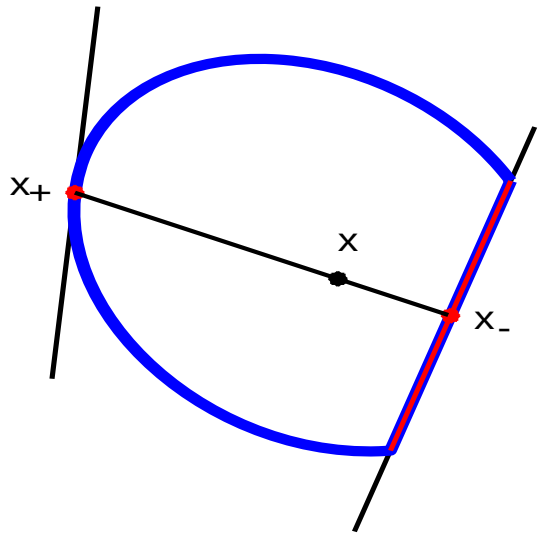
Fact IV.5 *If Q is a nonempty convex closed set in $\mathbf{R}^n$ which does not contain lines and $g^T x$ is a linear form which is bounded above on Q, then for every point $x_0 \in Q$ there exists (and can be found by Purification) a point $\bar{x} \in \mathrm{Ext}(Q)$ such that $g^T \bar{x} \geq g^T x_0$. In particular, if $g^T x$ attains its maximum on Q, then a maximizer can be found among extreme points of Q.*

Proof of KM, B. Let us prove that if a closed convex set Q contains lines, it has no extreme points.
Indeed, when Q contains a line and $v \in Q$, then Q contains the parallel line passing through v (Fact II.15) $\Rightarrow$ v is *not* extreme point of $Q \Rightarrow \mathrm{Ext}(Q) = \emptyset$.

Proof of KM, C. It remains to verify that if a nonempty closed convex set Q is bounded, then $Q = \mathrm{Conv}(\mathrm{Ext}(Q))$.
The inclusion $\mathrm{Conv}(\mathrm{Ext}(Q)) \subset Q$ is evident. Let us prove the opposite inclusion, i.e., prove that every point of Q is a convex combination of extreme points of Q.
Induction in $k = \dim Q$. Base $k = 0$ (Q is a singleton) is evident.
Step $k \mapsto k+1$: Given $(k+1)$-dimensional closed and bounded convex set Q and a point $x \in Q$, we can use the construction for finding a relative boundary point from the Purification algorithm to represent x as a convex combination of two points x_+ and x_- *from the relative boundary of* Q. Let Π_+ be a hyperplane which supports Q at x_+, and let $Q_+ = \Pi_+ \cap Q$. As we know, Q_+ is a closed convex set such that

$$\dim Q_+ < \dim Q,\ \mathrm{Ext}(Q_+) \subset \mathrm{Ext}(Q),\ x_+ \in Q_+.$$

Invoking inductive hypothesis, $x_+ \in \mathrm{Conv}(\mathrm{Ext}(Q_+)) \subset \mathrm{Conv}(\mathrm{Ext}(Q))$. Similarly, $x_- \in \mathrm{Conv}(\mathrm{Ext}(Q))$. Since $x \in [x_-, x_+]$, we get $x \in \mathrm{Conv}(\mathrm{Ext}(Q))$.



♠ Slightly modifying the argument used in item C in the proof of KM, we can prove the following nice statement: the Purification Algorithm combines with induction in dimension to yield

Fact IV.6 *Let $Q \subset$ be a nonempty closed convex set not containing lines. Then the set* Ext(X) *of extreme points of X is nonempty and*

$$Q = \mathrm{Conv}(\mathrm{Ext}(Q)) + \mathrm{Rec}(Q). \qquad (*)$$

The fact that for Q in question, Ext(Q) is nonempty, is part of the KM. To prove ($*$), we use induction in $\dim Q$. Base is evident, and the reasoning at the inductive step is modified as follows. Given that Q is nonempty, closed, convex, and does not contain lines and that ($*$) holds true when $\dim Q = k \geq 0$, we want to prove that ($*$) holds true when $\dim Q = k + 1$. The fact that the left hand side in ($*$) contains the right hand one is due to Fact II.14 combined with the evident inclusion Conv(Ext(Q)). All we need to complete the inductive step is to prove the opposite inclusion holds true. Thus, given $x \in Q$, we want to prove that $x \in \mathrm{Conv}(\mathrm{Ext}(Q) + \mathrm{Rec}(Q)$. To this end we select a nonzero $h \in \mathrm{Aff}(Q)$ (such an h exists since $dim Q > 1$)). As Q does not contain lines, we can, replacing, if necessary, h with $-h$, further assume that $h \notin \mathrm{Rec}(Q)$. With this in mind, we can apply one step of Purification algorithm to get a relative interior point x_+ of Q such that $x_+ = x + t_+ h$ for some $t_+ \geq 0$. Specifying Q_+ as in item C of the proof of KM, we get a nonempty closed subset Q_+ of Q with $\dim Q+ \leq K$ and Ext(Q_+) $\subset$ Ext(Q). In addition, Q_+ does not contain lines due to $Q_+ \subset Q$. Applying inductive hypothesis, we conclude that there exists $v_+ \in \mathrm{Conv}(\mathrm{Ext}(Q_+))$ and $r_+ \in \mathrm{Rec}(Q_+) \subset \mathrm{Rec}(Q)$ (the latter "$\subset$" is due to $Q_+ \subset Q$) such that $x_+ = v_+ + r_+$. Now, two cases are possible: (a) $-h \in \mathrm{Rec}(Q)$, and (b) $-h \notin \mathrm{Rec}(Q)$. In the case of (a), we have

$$x = x_+ + t_+[-h] - \underbrace{v_+}_{\in \mathrm{Ext}(Q)} + \underbrace{[r_+ + t_+[-h]]}_{\in \mathrm{Rec}(Q)}$$

(note that $t_+ \geq 0$ and we are in the case of $h \in \mathrm{Rec}(Q)$), that is, x is in the right hand side set of ($*$).
In the case of (b), $-h \notin \mathrm{Rec}(R)$, and we can apply the construction just described to $-h$ in the role of h to get a point $x_- = x - t_- h$ with $t_- \geq 0$ such that $x_- = v_- + r_- \in \mathrm{Conv}(\mathrm{Ext}(Q)) + \mathrm{Rec}(Q))$. Since x is a convex combination of x_+ and x_-, and both these points belong to the (convex!) right hand side of ($*$), we get $x \in \mathrm{Conv}(\mathrm{Ext}(Q)) + \mathrm{Rec}(Q)$, Q.E.D.

Extreme points and maximizers of linear forms

♣ Let Q be a nonempty closed convex set not containing lines and of positive dimension.

♠ An extreme point of Q is on the relative boundary of Q (since $\dim Q > 0$), and *every point of $\partial_r Q$ is a maximizer of non-constant on Q linear form* (the one coming from a hyperplane supporting Q at the point) $\Rightarrow$ *An extreme point of Q is among maximizers of a properly selected non-constant linear form.*

♠ The reverse also is true:

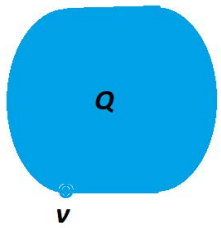
• *If linear form $f^T x$ attains its maximum over $x \in Q$, then all extreme points of* $\mathrm{Argmax}_{x\in Q} f^T x$ (the latter set, when nonempty, is a nonempty closed convex set not containing lines and as such does have extreme points) *are extreme points of Q.*

Indeed, there is nothing to prove when the form is constant on Q. When $f^T x$ is non-constant on Q and $v \in \mathrm{Argmax}_{x\in Q} f^T x$, the hyperplane $\Pi = \{x : f^T x = f^T v\}$ supports Q at v and $\mathrm{Argmax}_{x\in Q} f^T x = \Pi \cap Q$; by Fact IV.4, the extreme points of $\Pi \cap Q$ are extreme points of Q.

As a corollary,

Fact IV.7 *If a linear form attains its maximum over Q at a unique point, the maximizer is an extreme point of Q.*

Question: To which extent Fact IV.7 characterizes extreme point of Q – whether it is true that *every* extreme point of Q is the unique maximizer of properly selected linear form? The answer in general is "no:"



v is *not* the *unique* maximizer of a linear form on Q

The answer becomes "yes," when Q is polyhedral:

Fact IV.8 *For a nonempty polyhedral set Q, $v \in \mathrm{Ext}(Q)$ iff v is the unique maximizer, on Q, of properly selected linear form.*

Indeed, there is nothing to prove when $\dim Q = 0$. Now let $\dim Q > 0$. We already know that the unique maximizer, on Q, of a linear form. To prove the opposite, let $Q = \{x : a_i^T x \leq b_i, i \leq m\}$ and $v \in \mathrm{Ext}(Q)$. Let $I = \{i : a_i^T v = b_i\}$. Then $I \neq \emptyset$ (were I empty, we would have $v \in \mathrm{int}\, Q$, while all extreme points of closed convex set Q of positive dimension are in $\partial_{\mathrm{r}} Q$). Note that v is the unique solution to the system $a_i^T x = b_i, i \in I$; otherwise there would exist $h \neq 0$ such that $a_i^T h = 0, i \in I$, implying that $v \pm th \in Q$ for all small positive v, which is impossible. Setting $f = \sum_{i \in I} a_i$, the linear form $f^T x$ everywhere on Q is $\leq \beta = \sum_i b_i$ and equals to β at v. Moreover, for every $x \in Q$ with $f^T x = \beta$ it holds $a_i^T x = b_i$. $i \in I$, whence by the above $x = v \Rightarrow v$ is the unique maximizer of $f^T x$ over $x \in Q$.

Calculus of extreme points

A. When taking intersections or inverse affine images of closed convex sets, there are no simple rules expressing the extreme point of the result in terms of the extreme points of the operands.

B. Everything is fine with taking direct product: *Whenever $Q_k \subset \mathbf{R}^{n_k}$ are nonempty closed convex sets, one has*

$$\mathrm{Ext}(Q_1 \times ... \times Q_K) = \mathrm{Ext}(Q_1) \times ... \times \mathrm{Ext}(Q_K).$$

C. Situation with taking arithmetic sums is good: *In every representation of an extreme point v of the sum Q of nonempty convex sets $Q_1, ..., Q_K$ as $v = \sum_{k=1}^{K} v_k$ with $v_k \in Q_q$, every v_k is an extreme point of Q_k.*

Indeed, were v_k not an extreme point of Q_k for some k, there would exist $h \neq 0$ such that $v_k \pm h \in Q_k$, implying $v \pm h \in Q$, which is impossible.

Of course, when summing up extreme points of Q_k, the result not necessary is an extreme point of Q (look what happens when $K = 2$ and $Q_1 = Q_2 = [0, 1]$).

D. When taking image $\mathcal{A}(X) = \{ax + b : x \in Q\}$ of a closed nonempty set Q under affine mapping $\mathcal{A}(x) = Ax + b$, simple examples show that the image $\mathcal{A}(v)$ of $v \in \mathrm{Ext}(Q)$ not necessarily is an extreme point of $\mathcal{A}(A)$. Similarly, it may happen that some or all extreme points of $\mathcal{A}(Q)$ are not images of extreme points of Q (look what happens when projecting the stripe $0 \leq x_1 \leq 1$ in $\mathbf{R}^2$ onto the x_1-axis). However,

Fact IV.9 *If $Q \subset \mathbf{R}^n$ is a nonempty closed convex set not containing lines and $x \mapsto Ax + b : \mathbf{R}^n \to \mathbf{R}^m$ is an affine mapping, then every point $v \in \mathrm{Ext}(\mathcal{A}(Q))$ is $\mathcal{A}(u)$ for certain $u \in \mathrm{Ext}(Q)$.*

Indeed, let $v \in \mathrm{Ext}(\mathcal{A}(Q))$, and let $V = \{u \in Q : \mathcal{A}(u) = v\}$. Then V is a nonempty closed convex set not containing lines (since $Q \supset V$ does not contain lines), and therefore has an extreme point u. We claim that $u \in \mathrm{Ext}(Q)$ (this is all we need, as $v = \mathcal{A}(u)$ due $u \in V$). Indeed, assuming $u \pm h \in Q$ and setting $g = Ah$, we get $v \pm Ah \in \mathcal{A}(Q)$, whence $Ah = 0$ since v is an extreme point of $\mathcal{A}(Q)$. As $Ah = 0$ we have $\mathcal{A}(u \pm h) = \mathcal{A}(u) = v$, which combines with $u \pm h \in Q$ to imply $u \pm h \in V$. As $u \in \mathrm{Ext}(V)$, we get $h = 0$. Thus, $u \in \mathrm{Ext}(Q)$.

Cones revisited: Dual cone

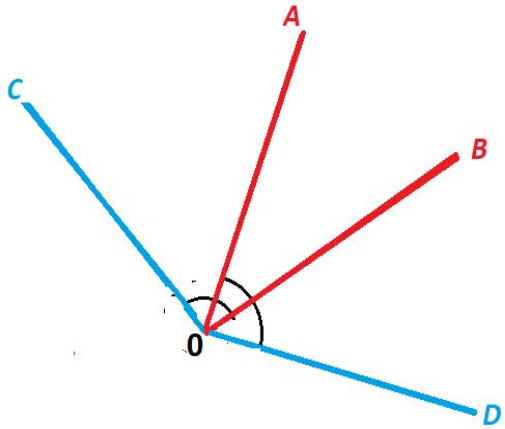
♣ Given a cone $K \subset \mathbf{R}^n$, its *dual cone* K_* *is the set of all vectors making nonnegative inner products with all vectors from* K

$$K_* = \{y : y^T x \geq 0 \, \forall x \in K\}$$

K_* *clearly is a closed cone, and* $[\mathrm{cl}\, K]_* = K_*$

Examples:

- $[\mathbf{R}^n]_* = \{0\}$, $[\{0\}]_* = \mathbf{R}^n$
- $[\mathbf{R}^n_+]_* = \mathbf{R}^n_+$
- A linear subspace L in $\mathbf{R}^n$ is a cone, and $L_* = L^\perp$
- The cone dual to $\angle AOB$ is $\angle COD$, and vice versa



$$\overrightarrow{OC}^T \overrightarrow{OB} = 0 = \overrightarrow{OA}^T \overrightarrow{OD}$$

- By Homogeneous Farkas Lemma, *For $A \in \mathbf{R}^{m\times n}$, the dual of the polyhedral cone $K = \{x : Ax \geq 0\} = \{x : Ax \in \mathbf{R}^m_+\}$, the dual is the polyhedral cone $K_* = A^T\mathbf{R}^m_+ = \{y : \exists \lambda \geq 0 : y = A^T\lambda\}$*
- For a nonempty convex set $X \subset \mathbf{R}^n$, *the dual of the conic transform* $\mathrm{ConeT}(X) = \mathrm{Cone}\,(\{[x;1] : x \in X\})$ *is the cone* $\mathrm{ConeT}_*(X) = \{[-y;s] \in \mathbf{R}^n_y \times \mathbf{R}^1_s : \sup_{x\in X} y^Tx \leq s\}$.

Indeed, $[-y;s]$ makes nonnegative inner products with all vectors from $\mathrm{Cone}\,(\{[x;1] : x \in X\})$ iff it makes nonnegative inner products with all vectors $[x;1]$ with $x \in X$, i.e., iff $s - y^Tx \geq 0$ for all $x \in X$.

$$K_* = \{y : y^T x \geq 0 \, \forall x \in K\}$$

Fact IV.10 *For a cone $K \subset \mathbf{R}^n$ it holds*

$$[K_*]_* = \mathsf{cl}\, K.$$

In particular, twice taken dual of a closed cone is this cone itself.

✓ By definition of K_*, every vector from K makes a nonnegative inner product with every vector from K $\Rightarrow$ $K \subset [K_*]_*$ $\Rightarrow$ $\mathsf{cl}\, K \subset [K_*]_*$.
✓ To prove the inverse inclusion, let $x \in [K_*]_*$; assuming that $x \notin \mathsf{cl}\, K$, we arrive at a contradiction as follows: $x \notin \mathsf{cl}\, K$ $\Rightarrow$ $\{x\}$ and $\mathsf{cl}\, K$ are nonempty convex sets at positive distance $\Rightarrow$ they can be strictly separated: for some y, we have

$$y^T x < \inf_{v \in \mathsf{cl}\, K} y^T v.$$

Since $\mathsf{cl}\, K$ is a cone and $\inf_{v \in \mathsf{cl}\, K} y^T v > -\infty$, we have $0 = \inf_{v \in \mathsf{cl}\, K} y^T v > y^T x \Rightarrow y \in K_*$ and $y^T x < 0$, contradicting $x \in [K_*]_*$.

Fact IV.11 *Let $K \subset \mathbf{R}^n$ be a closed cone. Then*
(i) $\operatorname{int} K_* \neq \emptyset$ *iff* K *is pointed*
(ii) Whenever $y \in \operatorname{int} K_*$, *there exists* $c_y < \infty$ *such that*

$$x \in K \Rightarrow \|x\|_2 \leq c_y[y^T x] \qquad (*)$$

(iii) When K *is nontrivial, one has*

$$\operatorname{int} K_* = \{y : y^T x > 0 \ \forall x \in K \backslash \{0\}\}$$

(i): assuming that K is not pointed: $\pm d \in K$ for some $d \neq 0$, we get $\pm y^T d = 0$ for all $y \in K_*$, implying due to $d \neq 0$ that $\operatorname{int} K_* = \emptyset$. Vice versa, let $\operatorname{int} K_* = \emptyset$, and let us prove that K is *not* pointed. As $\operatorname{int} K_* = \emptyset$, $\operatorname{Lin}(K_*)$ is a proper linear subspace.
Indeed, assuming that $\operatorname{Lin}(K_*) = \mathbf{R}^n$, *there exists a basis* $f_1, ..., f_n$ *of* $\mathbf{R}^n$ *with* $f_i \in K_* \Rightarrow \emptyset \neq \operatorname{int} \operatorname{Conv}(0, f_1, f_2, ..., f_n) \subset K_*$ *– contradiction!*
Since $\operatorname{Lin}(K_*)$ is a proper linear subspace, there exists a nonzero $d \in [\operatorname{Lin}(K_*)]^\perp \Rightarrow d$ is orthogonal to every $y \in K_* \Rightarrow \pm d \in [K_*]_* = K$; since $d \neq 0$, K is not pointed, as claimed.
(ii) When $y \in \operatorname{int} K_*$, for some $r_y > 0$ we have $y + h \in K_*$ for all h with $\|h\|_2 \leq r_y \Rightarrow y^T x + h^T x \geq 0$ for every $x \in K \Rightarrow y^T x - r_y \|x\|_2 = \min_{h:\|h\|_2 \leq r_y} [y + h]^T x \geq 0$ for every $x \in K \Rightarrow (*)$ holds true with $c_y = r_y^{-1}$.
(iii) ✓ Let y be such that $y^T x > 0$ for all $x \in K \backslash \{0\}$, and let us prove that $y \in \operatorname{int} K_*$. There is nothing to prove when $K = \{0\}$, that is, $K_* = \mathbf{R}^n$. Assuming $K \neq \{0\}$ and setting $\alpha = \min_{x \in K : \|x\|_2 = 1} y^T x$, we get $\alpha > 0$ (as the minimum, taken over compact nonempty set, of a continuous function positive on the set). By homogeneity, $y^T x \geq \alpha \|x\|_2$ for all $x \in K$, whence $y + h)^T x \geq 0$ for all $x \in K$ and all h with $\|h\|_2 \leq \alpha \Rightarrow$ centered at y $\|\cdot\|_2$-ball of radius $\alpha > 0$ is contained in $K_* \Rightarrow y \in \operatorname{int} K_*$.
✓ Vice versa, if $y \in \operatorname{int} K$, then $\|x\|_2 \leq c_y[y^T x]$ for all $x \in K$ and some c_y by (ii) $\Rightarrow y^T x > 0$ whenever $x \in K \backslash \{0\}$.

♣ The "conic analogy" of extreme points are *extreme rays* defined as follows:

Let $K \subset \mathbf{R}^n$ be a *closed* cone. A nonzero vector $d \in K$ is called *extreme direction* of K, if in *every* representation $d = d^1 + d^2$ of d as the sum of two vectors *from* K, both d^1 and d^2 are nonnegative multiples of d.

The ray $\mathbf{R} \cdot d$ spanned by an extreme direction d is called *extreme ray* of K.

Examples: • Extreme directions of $\mathbf{R}^n_*$ are positive multiples of the standard basic orth $\Rightarrow$ the extreme rays are the nonnegative rays of the coordinate axes.

Indeed, an extreme direction d is a nonzero nonnegative vector such that when $d = d^1 + d^2$ with $d^i \geq 0$, then d^1 and d^2 are nonnegative multiples of d. This indeed is the case when $d \geq 0$ has exactly one positive entry $d_\ell > 0$; in this case $d^i_j + d^2_j = 0$ for all $j \neq \ell$, and since both summands are ≥ 0, we get $d^1_j = d^2_j = 0$, $j \neq \ell \Rightarrow d^i$ are nonnegative multiples of d. On the other hand, when d has at least 2 nonzero entries, say, $d_! > 0$, $d_2 > 0$, then $d = [d_1; 0; ...; 0] + [0; d_2; d_3; ...; d_n]$, both terms in the right hand side are *not* nonnegative multiples of d.

• Extreme rays of the Lorentz cone $\{[x; t] : t \geq \|x\|_2\}$ of dimension > 1 are the rays of the form $\{[te; t] : t \geq 0\}$ with $\|e\|_2 = 1$

• $\mathbf{R}^n$ and $\{0\}$ have no extreme rays.

Extreme rays and bases

♣ Let $K \subset \mathbf{R}^n$ be a closed cone. Set of the form

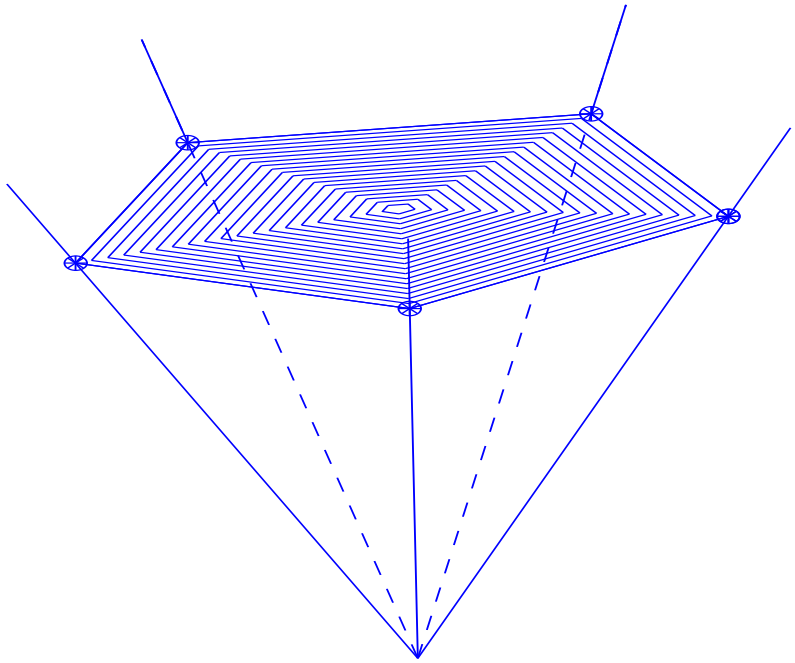
$$B_y = \{x \in K : y^T x = 1\}$$

– intersection of K with a hyperplane *not* passing through the origin – is called a base of K, when B_y is nonempty and contains a positive multiple of every nonzero vector from K.

Example: Bases of $\mathbf{R}^n_+$ are exactly the sets $\{x \geq 0 : y^T x = 1\}$ stemming from $y > 0$.

Note: Trivial cone K has no bases!

Fact IV.12 *Let $K \subset \mathbf{R}^n$ be a nontrivial closed cone. Then*
(i) $B_y = \{x \in K : y^T x = 1\}$ is a base of K iff $y \in \operatorname{int} K_*$. When K is pointed, B_y is a base iff B_y is nonempty and bounded
(ii) K possesses bases iff K is pointed and nontrivial, i.e., iff K is nontrivial and $\operatorname{int} K_* \neq \emptyset$
(iii) A base B_y of K, if any, is a closed and bounded convex set, and there exists one-to-one correspondence between extreme rays of K and extreme points of B_y: nontrivial emanating from the origin rays in K are exactly nonnegative multiples of points from B_y, and the ray $\mathbf{R}_+ \cdot v$ with $v \in B_y$ is extreme ray of K iff v is an extreme point of B_y.



Cone and its base

• K is a nontrivial closed cone. Then
(i) $B_y = \{x \in K : y^T x = 1\}$ is a base of K iff $y \in \operatorname{int} K_*$. When K is pointed, B_y is a base iff B_y is nonempty and bounded
Proof of (i): When $y \in \operatorname{int} K_*$, $y^T x > 0$ for every nonzero $x \in K$ (Fact IV.11.iii) $\Rightarrow$ B_y intersects every nontrivial ray in k emanating from the origin $\Rightarrow$ B_y is a base of K. Vice versa, if B_y is a base of K and $y^T x > 0$ for all nonzero $x \in K$, implying that $y \in \operatorname{int} K_*$ by the same Fact IV.11.iii (recall that K is assumed to be nontrivial); the first "iff" is proved.
Next, let K be pointed. ✓ If B_y is a base, then $y \in \operatorname{int} K_*$ (by the already proved part of (i)); nonbusiness of B_y is part of the definition of a base, and roundedness stems from Fact IV.11.ii. ✓ Vice versa, let B_y be nonempty and bounded, and let us prove that B_y is a base. We already know that to this end it suffices to verify that $y \in \operatorname{int} K_*$; invoking Fact IV.11.iii, all we need to verify that $y^T v > 0$ for every $v \in K \backslash \{0\}$. Assuming the opposite, let $v \in K$ be such that $0 \neq v$ and $y^T v \leq 0$, that $y^T v \leq 0$ for some $v \in \mathbf{K} \backslash K$, and let us lead this assumption to a contradiction. Since $B_y \neq \emptyset$. there exists $u \in K$ with $y^T u = 1$. Note that v is not proportional to u.
Indeed, assuming $v = \lambda u$, λ cannot be neither positive (since $y^T u = 1$, $y^T v \leq 0$), nor 0 (since $v \neq 0$), nor negative (since K is pointed and $u, v \in K$).
Now, $y^T u = 1$, $y^T v \leq 0$ $\Rightarrow$ $y^T d = 0$ for some $d \in [u, v]$, and $d \neq 0$ (since $u \neq 0$ and v is not proportional to u) $\Rightarrow$ $0 \neq d \in \operatorname{Rec}(B_y)$ $\Rightarrow$ the nonempty closed convex set B_y is unbounded, which is a desired contradiction.

• K is a nontrivial closed cone. Then
(ii) K possesses bases iff K is pointed and nontrivial, i.e., iff $\operatorname{int} K_* \neq \emptyset$

Proof of (ii): By (i), K possesses bases iff $\operatorname{int} K_* \neq \emptyset$, which is the same as to say that K is pointed and nontrivial (Fact IV.11.i).

• K is a nontrivial closed cone. Then
(iii) A base B_y of K, if any, is a closed and bounded convex set, and there exists one-to-one correspondence between extreme rays of K and extreme points of B_y: nontrivial emanating from the origin rays in K are exactly nonnegative multiples of points from B_y, and the ray $\mathbf{R}_+ \cdot v$ with $v \in B_y$ is extreme ray of K iff v is an extreme point of B_y.

Proof of (iii): Let B_y be a base of K. ✓ Let $\mathbf{R}_+ \cdot d$ be an extreme ray. It intersects $B_y \Rightarrow$ we can assume that $d \in B_y$. To prove that $d \in \mathrm{Ext}(B_y)$, let $d \pm h \in B_y \Rightarrow y^T h = 0$. We have $d = \underbrace{\frac{1}{2}[d+h]}_{d^1} + \underbrace{\frac{1}{2}[d-h}_{d^2}$ and $d^1, d^2 \in K$ as $B_y \subset K \Rightarrow d^i = \lambda_i d$, $i = 1, 2$. As $1 = y^T d = y^T[d+h] = 2y^T d^1 = 2\lambda_1 y^T d$, we get $\lambda_1 = \frac{1}{2} \Rightarrow d^1 := \frac{1}{2}[d+h] = \frac{1}{2}d \Rightarrow h = 0$. Thus, $d \in B_y$ and $d \pm h \in B_y \Rightarrow h = 0 \Rightarrow d \in \mathrm{Ext}(B_y)$.

✓ Vice versa, let $d \in \mathrm{Ext}(B_y)$, and let us prove that d is an extreme direction of K. Assuming that $d = d^1 + d^2$ with $d^1, d^2 \in K$, to see that d^1, d^2 are nonnegative multiplies of d, it suffices to consider the case when both d^2 and d^2 are nonzero, so that $\lambda_i = y^T d_i > 0$. We have $1 = y^T d = y^T(d^1 + d^2 \Rightarrow \lambda_1 + \lambda_2 = 1 \Rightarrow$ with $f^i = \lambda_i^{-1} d^i$ we get $y^T f^i = 1$ and $f^i \in K$, $i = 1, 2 \Rightarrow d = d^1 + d^2 = \lambda_1 f^1 + \lambda_2 f^2 \Rightarrow d \in \mathrm{Ext}(B_y)$ is a convex combination, with positive coefficients, of $f^1 \in B_y$ and $f^2 \in B_y \Rightarrow f^1 = f^2 = d \Rightarrow d^i = \lambda_i d$, $i = 1, 2$, Q.E.D.

♠

Fact IV.13 [Krein-Milman Theorem in Conic form] *Let $K \subset \mathbf{R}^n$ be a closed cone. K possesses extreme rays iff K is nontrivial and pointed, and in this case K is the conic hull of the set $\mathcal{R}$ of its extreme rays.*

✓ When $\mathcal{K}$ is nontrivial and pointed, if has a base B which is a nonempty compact convex set (Fact IV.12) $\Rightarrow$ Ext$(B) \neq \emptyset$ & $B =$ Conv(Ext(B)) (TM) $\Rightarrow$ $K =$ Conv$(B) \subset$ Cone $(\mathcal{R})$ (Fact IV.12.iii).

✓ When K is trivial, it clearly has no extreme rays. When K is nontrivial and is *not* pointed, it has no extreme rays as well. Indeed, let $e \neq 0$ be a direction of a line contained in K and d be a nonzero direction in K. Then $\{d\} + \mathbf{R} \cdot e \subset K \Rightarrow$ of s, and when e and t is large, at least one of $d^{\pm}$ is not a *nonnegative* multiple of $d \Rightarrow d$ is *not* and extreme direction of K.

Dubovitski–Milutin Lemma

♠ Let $K^1, ..., K^L$ be closed cones in $\mathbf{R}^n$. We have

Fact IV.14 *The cone dual to the intersection $K = \cap_\ell K^\ell$ of the cones K^ℓ is the closure of the sum of the duals K^ℓ_* of the cones:*

$$K_* := \left[\cap_\ell K^\ell\right]_* = \mathrm{cl}\left(K^1_* + ... + K^*_L\right) \qquad (*)$$

In words: *Vectors f such that the linear form $f^T x$ is nonnegative for x running through the intersection of closed cones K_ℓ, $\ell \le L$, are exactly the vectors which can be approximated, to whatever high accuracy, by sums of the form $\sum_\ell f^\ell$ with f^ℓ producing nonnegative on the respective cones K_ℓ linear forms $[f^\ell]^T x$.*

Proof. ✓ When $f = \sum_\ell f^\ell$ with $f^\ell \in K^\ell_*$, then clearly $f \in K_*$, implying that the right hand side set in $(*)$ is contained in the left hand side one.
✓ To prove the inverse inclusion, assume that there is $f \in K_*$ which does not belong to the right hand side set M and let us lead this assumption to a contradiction. Since M is a closed cone, Separation Theorem says that there exists x such that $f^T x \le \inf_{f^\ell \in K^*_\ell, \ell \le L} [\sum_\ell f^\ell]^T x$, that is,

$$f^T x < \sum_\ell \inf_{f^\ell \in K^*_\ell} [f^\ell]^T x. \qquad (*)$$

In particular, every one of $\inf_{f^\ell}$ is finite, implying that it is 0 (K^ℓ_* is a cone!). Thus, for every ℓ it holds $x \in [K^\ell_*]_* = K^\ell$, where the equality is due to Fact IV.10. Thus, $x \in K = \cap_\ell K^\ell$ and the right hand side in $(*)$ is zero, that is, $f^T x < 0$ - the desired contradiction with $f \in K_*$. □

When K^ℓ, $\ell \le L$, are closed cones, one has

$$[\cap_\ell K_\ell]_* = \mathrm{cl}\left(K_*^1 + ... + K_*^L\right). \qquad (*)$$

♠ It can be shown by examples that in general, one cannot get rid of taking the closure in $(*)$, which would be crucial in many applications. Dubovitski-Milutin Lemma presents a simply looking sufficient condition allowing to get rid of taking closure.

Fact IV.15 [Dubovitski-Milutin Lemma] *Let $L \ge 2$, and let $M^1, ..., M^L, M = \cap_\ell M^\ell$ be cones in $\mathbf{R}^n$ with the duals M_*^1,..., M_*^L, M_*, and let*

$$M^1 \cap \mathrm{int}\, M^2 \cap ... \cap \mathrm{int}\, M^L \neq \emptyset \qquad (!)$$

Then, setting $K^\ell = \mathrm{cl}\, M^\ell$ (whence $K_*^\ell = M_*^\ell$), we have
(i) $\mathrm{cl}\, M = K := [\cap_\ell K_\ell]$ $\left[\Rightarrow M_* = K_*\right]$, *and*
(ii) the cone $M_*^1 + ... + M_*^L$ is closed, implying by Fact IV.14 that

$$M_* = M_*^1 + ... + M_*^L.$$

In words: *In the case of (!) a linear form $f^T x$ is nonnegative on the intersection of cones M^ℓ iff the form can be represented as the sum of linear forms nonnegative on the respective cones M^ℓ.*

Let $M^1, ..., M^L, M = \cap_\ell M^\ell$ be cones in $\mathbf{R}^n$ with the duals M^1_*,..., M^L_*, M_*, and let

$$M^1 \cap \operatorname{int} M^2 \cap ... \cap \operatorname{int} M^L \neq \emptyset \qquad (!)$$

Then, setting $K^\ell = \operatorname{cl} M^\ell$ (whence $K^\ell_* = M^\ell_*$), we have
(i) $\operatorname{cl} M = K := [\cap_\ell K_\ell]$ $\left[\Rightarrow M_* = K_*\right]$, and
(ii) the cone $M^1_* + ... + M^L_* = [K^1_* + ... + K^L_*]$ is closed.

Proof. (i): A point in the (nonempty!) set in $(*)$ can be approximated to whatever high accuracy by point from $\operatorname{rint} M^1$, implying due to the structure of the set that there exists a point $\bar{x} \in \operatorname{rint} M^1 \cap \operatorname{int} M^2 \cap ... \cap M^L$. Now, clearly $\operatorname{cl} M \subset \cap_\ell \operatorname{cl} M^\ell$. To prove the inverse inclusion, we need to prove that if $x \in \cap_\ell \operatorname{cl} M^\ell$, then $x \in \operatorname{cl} M$. Indeed, by Fact II.29 the points $x^i = (1/i)\bar{x} + (1 - 1/i)x$ $i = 1, 2...$, belong to M^ℓ, $\ell \leq L$, and therefore belong to $M \Rightarrow x = \lim_{i\to\infty} x^i \in \operatorname{cl} M$, as claimed. (i) is proved.
(ii): There is nothing to prove when $L = 1$. When $L > 1$, for every $\ell \in \{2, ..., L\}$, applying Fact IV.11.ii to $\bar{x} \in \operatorname{int} M^\ell = \operatorname{int} K^\ell$ in the role of y and K^ℓ_* in the role of K, we conclude that there exists $c_\ell < \infty$ such that $\|f^\ell\|_2 \leq c_\ell [f^\ell]^T \bar{x}$ for all $f^\ell \in K^\ell_* = M^\ell_*$. Now we are ready to prove that $Q := M^1_* + ... + M^L_*$ is closed. Indeed, let $f_i = \sum_{\ell=1}^L f^\ell_i$ with $f^\ell_i \in M^\ell_*$ converge as $i \to \infty$ to some $f \in Q$, observe that the sequences $\{[f^\ell_i]^T \bar{x}\}_i$ are nonnegative and their sum converges as $i \to \infty$ to $f^T \bar{x}$. implying that the sequences are bounded: for some real B and all ℓ, i it holds $0 \leq [f^\ell_i]^T \bar{x} \leq B$, implying for $\ell \geq 2$ that the sequences $\{f^\ell_i\}_i$ are bounded: $\|f^\ell_i\|_2 \leq c_\ell B$ for all i and all $\ell \geq 2$. Passing to a subsequence, we may assume w.l.o.g. that for $\ell \geq 2$ the limits $f^\ell = \lim_{i\to\infty} f^\ell_i$ exist. Since $\sum_{\ell=1}^L f^\ell_i \to f$ as $i \to \infty$, the limit $f^1 = \lim_{i\to\infty} f^1_i$ exists as well, $f = \sum_{\ell=1}^L f^\ell$ and $f^\ell \in M^\ell_*$ (cones M^ℓ_* are closed!) $\Rightarrow f \in Q$, Q.E.D.

Note: In the polyhedral case, DML holds "unconditionally:"

Fact IV.16 *When $M^1,,,,,M^L$ are polyhedral cones in $\mathbf{R}^n$, then the cone $M^1_* + ... + M^L_*$ is closed, so that*

$$[\cap_\ell M^\ell]_* = M^1_* + .. + M^L_*.$$

Indeed, the duals of polyhedral cones are polyhedral, and the sum of polyhedral cones is polyhedral and thus is closed; it remains to apply Fact IV.14.

♠ As a corollary of Dubovitski-Milutin Lemma, we get the following useful

Fact IV.17 *Let $L \geq 2$, and let K^ℓ be closed cones in $\mathbf{R}^n$ with duals K^ℓ_*, $\ell \leq L$, and let $K^1_* \cap \operatorname{int} K^2_* \cap ... \cap \operatorname{int} K^L_* \neq \emptyset$. Then the cone $K^1 + ... + K^L$ is closed.*

Indeed, it suffices to apply item (ii) of DML to the cones $M^\ell = K^\ell_*$.

♠ Another useful corollary of DML is as follows. Let $M^+ \subset \mathbf{R}^n$ be a closed cone, $y \mapsto Ay : \mathbf{R}^n \to \mathbf{R}^n$ be a linear map, and let $M = \{y : Ay \in M^+\}$ be the inverse image of M^+ under the mapping, so that M is a closed cone in $\mathbf{R}^m$. Observe that

$$A^T M_*^+ \subset M_*; \qquad (*)$$

indeed, if $f \in M_*^+$, then $[A^T f]^T y = f^T [Ay] \geq 0$ whenever $y \in M$. For numerous applications, it is important to know when $(*)$ is an equality Observe that this is the case iff

$$[M^+ \cap E]_* = M_*^+ + E_*, \quad E = \mathrm{Im}\, A \qquad (**)$$

(as usual $K*$ is the dual of a cone K).
Indeed, $f \in M_*$ must be orthogonal to $\mathrm{Ker} A \subset M$, which by Linear Algebra means that $f = A^T g$ for some g, so that

$$M_* = \{A^T g : g^T Ay \geq 0 \, \forall y \in M\},$$

that is,

$$M_* = \{A^T g : g^T Ay \geq 0 \, \forall y \in M\}, = \{A^T g : g^T x \geq 0 \, \forall x \in [M^+ \cap E]\} = A^T \left([M^+ \cap E]_*\right) \qquad (a)$$

(the first equality in (a) has been just explained, the third is evident, and the second is due to $AM = M^+ \cap E$ by definitions of E and M). By (a), equality $M_* = A^T M_*^+$ means that whenever $g \in [M_*^+ \cap E]_*$, so that $f = A^T g \in M_*$, we have also $A^T g = A^T h$ with $h \in M_*^+$, or, which is the same, for every $g \in [M^+ \cap E]_*$ there exists $h \in M_*^+$ such that $A^T g = A^T h$, or, which again is the same, exists $h \in M_*^+$ such that $h - g \in \mathrm{Ker} A^T = [\mathrm{Im}\, A]^\perp = E^\perp = E_*$. The bottom line is that $M_* = A^T M_*^+$ *iff* $[M^+ \cap E]_* = M_*^+ + E_*$, as claimed in $(**)$.
Applying DML and its polyhedral version, we arrive at

Fact IV.18 *When $M^+ \subset \mathbf{R}^n$ is a closed cone, $A \in \mathbf{R}^{n \times m}$, and $M = \{y : Ay \in M^+\}$, we always have $A^T M_+^* \subset M_*$, with equality taking place iff the cone $M_*^+ + [\mathrm{Im}\, A]^\perp$ is closed. The latter definitely is the case when M^+ is polyhedral, same as when*

$$\mathrm{Im}\, A \cap \mathrm{int}\, M^+ \neq \emptyset.$$

"Inhomogeneous case"

Fact IV.19 *Let $L \geq 2$ closed convex sets $Q^1, ..., Q^L$ in $\mathbf{R}^n$ be given, and let $Q = \cap_\ell Q^\ell \neq \emptyset$.*

- *If a vector $[f; \alpha] \in \mathbf{R}^n_x \times \mathbf{R}^1_\alpha$ can be decomposed as*

$$[f; \alpha] = [f^1; \alpha^1] + ... + [f^L; \alpha^L]$$

with $\sup_{x \in Q^\ell} [f^\ell]^T x \leq \alpha^\ell$, then

$$\sup_{x \in Q} f^T x \leq \alpha,$$

Equivalently: If $f = \sum_\ell f^\ell$, then $\sup_{x \in Q} f^T x \leq \sum_\ell \sup_{x \in Q^\ell} [f^\ell]^T x$

- *When the sum over $\ell \leq L$ of the cones $\{[-y; s] : \sup_{x \in Q^\ell} y^T x \leq s\}$ is closed, which definitely is the case when*

$$Q^1 \cap \operatorname{int} Q^2 \cap ... \cap \operatorname{int} Q^L \neq \emptyset, \qquad (!)$$

same as when all Q^ℓ are polyhedral, the above "If" can be strengthened to "Iff"

In other words: *When the sum over $\ell \leq L$ of the cones $\{[-y; s] : \sup_{x \in Q^\ell} y^T x \leq s\}$ is closed, which definitely is the case when (!) takes place, same as when all Q^ℓ are polyhedral, the relation*

$$\sup_{x \in Q} f^T x \leq \alpha$$

takes place if and only if f can be decomposed as $f = f^1 + ... + f^L$ in such a way that

$$\sum_\ell \sup_{x \in Q^\ell} [f^\ell]^T x \leq \alpha.$$

Proof. Let M^ℓ, M be closed conic transforms of Q^ℓ and $Q = \cap_\ell Q^\ell$:

$$M^\ell = \mathrm{cl}\,\mathrm{Cone}\,(Q^\ell \times \{1\}) = \mathrm{cl}\,\{t[x;1] : x \in Q^\ell, t \geq 0\},\ ,\ M^\ell = \mathrm{cl}\,\mathrm{Cone}\,(Q^\ell \times \{1\}) = \mathrm{cl}\,\{t[x;1] : x \in Q\}$$

so that M^ℓ are closed cones. Observe that $M = \cap_\ell M^\ell$. Indeed, as Q^ℓ and therefore Q are closed, and Q is nonempty.We know from the story of closed conic transforms of closed convex sets that setting $Q_t := \{x : [x;t] \in M\}$,, $Q_t^\ell := \{x : [x:t] \in M^\ell\}$, we have

$$t > 0 \Rightarrow Q_t = \{x : x/t \in Q\},\ Q_t^\ell := \{x : [x:t] \in M^\ell\} = \{x : x/t \in M^\ell\}$$
$$\& \ Q_0 = \mathrm{Rec}(Q),\ Q_0^\ell = \mathrm{Rec}(Q^\ell)$$
$$\& \ t < 0 \Rightarrow Q_t = Q_t^\ell = \emptyset,$$

that is, $Q_t = \cap_\ell Q_t^\ell$ for all t, whence $M = \cap_\ell M_\ell$.
Next, we have

$$\begin{array}{rcl} M_*^\ell & = & \{[-y;s] : ts - y^T x \geq 0\,\forall [x,;t] \in M^\ell\} = \{[-y;s] : ts - y^T x \geq 0 \forall (t > 0/x = tz : z \in Q^\ell\} \\ & = & \{[-y;s] : s - y^T z \geq 0\,\forall z \in Q^\ell\} \\ & = & \{[-y;s] : \sup_{z \in Q^\ell} y^T z \leq s\}. \end{array}$$

and similarly

$$M_* = \{[-y;s] : \sup_{y \in Q} y^T z\} \leq s\}$$

By Fact IV.14, we conclude that
(!) *Whenever the cone $M_*^1 + ... + M_*^L$ is closed, we have*

$$M_* = M_*^1 + ... + M_*^L,$$

which combines with the description of M_* and M_ℓ^* to imply that *when the cone $M_*^1 + ... + M_*^L$ is closed, relation $\sup_{x \in Q} f^T x \leq \alpha$ holds true iff there exist f_ℓ, α_ℓ such that*

$$f = f_1 + ... + f_L \ \& \ \alpha = \alpha_1 + ... + \alpha_L.$$

✓Assume, first, that the intersection $Q \cap \mathrm{int}\, Q^1 \cap ... \cap \mathrm{int}\, Q^L$ is nonempty, and let $\bar{x}$ be a point from this intersection. Then clearly $[\bar{x};1] \in M^1 \cap \mathrm{int}\, M^2 \cap ... \cap M^L$, and therefore by DML we have

$$M_* = M_*^1 + ... + M_*^L,$$

so that the right hand side cone is closed; applying (!), we arrive at the "nonpolyhedral" part of Fact.
✓ Now let Q^ℓ be polyhedral. Then the cones M^ℓ, M are polyhedral as well (Fact **F** in Calculus of polyhedrality), and therefore the cone $M_*^1 + ... + M_*^L$ is polyhedral and thus is closed. Applying (!), we arrive at the "polyhedral" part of Fact. □

Illustration. A state $x \in \mathbf{R}^n$ of certain system is a point of the nonempty polyhedral set $Q = \{x : A_\ell x \le b^\ell, \ell \le L\}$. When in state x, the "tax" $f^T x$ should be paid. It is known that budget α is sufficient to pay the tax, whatever be the state $x \in Q$: $f^T x \le \alpha$ for all $x \in Q$.
Question: Can we "decentralize taxation," that is, instead of paying the tax by central authority which, provided with budget α, observes the state x of the system and pays the tax $f^T x$, use L agents, ℓ-th of them observing $A_\ell x$ and paying tax $\lambda_\ell^T A_\ell x$? Can we specify vectors λ_ℓ and distribute the budget α as $\alpha = \sum_\ell \alpha_\ell$ in such a way that $\sum_\ell \lambda_\ell^T A_\ell x \equiv f^T x$ and $\lambda_\ell^T A_\ell x \le \alpha_\ell$ for all $x \in Q$?

Answer: Yes, we can.
Let $Q^\ell = \{x : A_\ell x \le b^\ell\}$, so that $Q = \cap_\ell Q^\ell$. Observe that

$$[\mathrm{ConeT}(Q^\ell)]_* = \{[-y; s] : \sup_{x \in Q^\ell} y^T x \le s\} = \{[-y; s] : A_\ell x \le b^\ell \Rightarrow y^T x \le s\} = \{[-y; s] : \exists \lambda \ge 0 : y = A_\ell^T \lambda, \lambda^T b^\ell \le s\}$$

where the last equality is by Inhomogeneous Farkas Lemma (applicable since $Q^\ell \ne \emptyset$ due to $Q \ne \emptyset$). We see that the cones $\{[-y; s] : \sup_{x \in Q^\ell} y^T x \le s\}$ are polyhedral, so that their sum over ℓ is polyhedral as well, and thus is closed. Applying Fact IV.19, we conclude that as $\sup_{x \in Q} f^T x \le \alpha$, we can represent f as $\sum_\ell f^\ell$ with $\sum_\ell \underbrace{\sup_{x \in Q^\ell} [f^\ell]^T x}_{\alpha_\ell^*} \le \alpha$. Under the circumstances, $f^\ell = A_\ell^T \lambda_\ell$ with properly selected $\lambda_\ell \ge 0$, and since $\sum_\ell \alpha_\ell^* \le \alpha$, we can find $\alpha_\ell \ge \alpha_\ell^*$ such that $\sum_\ell \alpha_\ell = \alpha$. Thus, $\sum_\ell \lambda_\ell^T A_\ell x \equiv f^T x$ and $\sup_{x \in Q^\ell} \lambda_\ell^T A_\ell x \le \alpha_\ell$ with $\sum_\ell \alpha_\ell = \alpha$ – decentralized taxation indeed is possible.
Note: "Decentralization of taxation" is readily given by LP Duality: our story says that the optimal value in the feasible LP program $\mathrm{Opt} = \max_x \{f^T x : x \in Q_\ell := \{u : A_\ell u \le b^\ell\}, \ell \le L\}$ does not exceed $\alpha \Rightarrow$ the dual problem $\min_\lambda \{\sum_\ell \lambda_\ell^T b^\ell : \sum_\ell A_\ell^T \lambda_\ell = f, \lambda_\ell \ge 0, \ell \le L\}$ is solvable with optimal solution $\lambda^* \ge 0$ satisfying $\sum_\ell \underbrace{A_\ell^T \lambda_\ell^*}_{f^\ell} = f$ and $\sum_\ell \underbrace{[b^\ell]^T \lambda_\ell^*}_{\beta_\ell} = \mathrm{Opt}$. As $\lambda_\ell^* \ge 0$, $\max_{x \in Q_\ell} [f^\ell]^T x \le \beta_\ell$, which is all we need due to $\sum_\ell \beta_\ell = \mathrm{Opt} \le \alpha$. The advantage of our initial argument is that it works when Q^ℓ are convex rather than polyhedral, at the price of assuming $Q^1 \cap \mathrm{int}\, Q^2 \cap ... \cap \mathrm{int}\, Q^L \ne \emptyset$ instead of $\cap_\ell Q^\ell \ne \emptyset$.

Calculus of dual cones

Fact IV.20 *The following claims are true:*

A. Involutive property: *The dual $[K_*]_*$ of the dual K_* of a cone K is* cl K.

B. Taking intersection *When $K^\ell, \ell \le L$, are closed cones in $\mathbf{R}^n$, one has*

$$\left[\bigcap\nolimits_\ell K^\ell\right]_* = \mathrm{cl}\,\left(K_*^1 + ... + K_*^L\right).$$

When $K^1 \cap \mathrm{int}\, K^2 \cap ... \cap \mathrm{int}\, K^L \neq \emptyset$, taking closure in the right hand side can be omitted.

C. Summation: *The dual of the sum of finitely many cones in $\mathbf{R}^n$ is the intersection of their duals. More generally, let $K^\alpha, \alpha \in \mathcal{A}$, be a family of cones in $\mathbf{R}^n$. Then*

$$\left[\mathrm{Cone}\left(\bigcup\nolimits_{\alpha\in\mathcal{A}} K^\alpha\right)\right] = \bigcap\nolimits_{\alpha\in\mathcal{A}} K_*^\alpha$$

Indeed, y belongs to the left hand side set iff $y^T x \ge 0$ for all $x \in \cup_\alpha K^\alpha$, i.e., iff $y \in \cap_\alpha K_*^\alpha$.

D. Taking direct products *When $K^\ell \subset \mathbf{R}^{n_\ell}$, $\ell \le L$, are cones, one has*

$$\left[K^1 \times ... \times K^L\right]_* = K_*^1 \times ... \times K_*^L$$

E. Taking linear image: *Let $K \subset \mathbf{R}^n$ be a cone and $AK = \{Ax, x \in K\}$ be the linear image of K under linear mapping $x \mapsto Ax : \mathbf{R}^n \to \mathbf{R}^m$. Then*

$$[AK]_* = [A^T]^{-1} K_* := \{z : A^T z \in K_*\}$$

Indeed, $\{z^T y \ge \forall y \in AK\} \Leftrightarrow \{z^T A x \ge 0\, \forall x \in K\} \Leftrightarrow A^T z \in K_*$

F. Taking inverse linear image: *Let $K \subset \mathbf{R}^n$ be a closed cone and $A^{-1}K = \{y : Ay \in K\}$ be the inverse image of K under linear mapping $y \mapsto Ay : \mathbf{R}^m \to \mathbf{R}^n$. Then*

$$\big[\underbrace{A^{-1}K}_{Q}\big]_* = \mathrm{cl}\, A^T K_*.$$

When $[\mathrm{Im}\, A]^\perp \cap K = \{0\}$, taking the closure in the right hand side can be omitted.

Indeed, when $z = A^T u$, $u \in K_*$, and $v := Ay \in K$, one has $z^T y = u^T Ay = u^T v \geq 0 \Rightarrow [A^T K* \subset Q_* \Rightarrow \mathrm{cl}\, A^T K_* \subset Q_*$
Vice versa, when $z \notin \mathrm{cl}\, A^T K_*$, there exists linear from $w^T u$ of $u \in \mathbf{R}^m$ strictly separating z from $A^T K_*$;

$$z^T w < \inf_{u \in K_*} w^T A^T u \Rightarrow z^T w < 0 \ \&\ Aw \in [K_*]_* = K \Rightarrow w \in A^{-1}K \text{ and } z^T w < 0 z \notin [A^{-1}K]_* = Q_*,$$

hence $\mathrm{cl}\, A^T K_* \supset Q_*$. Finally, when $[\mathrm{Im}\, A]^\perp \cap K = \{0\}$, the cone $A^T K_*$ is closed (Fact II.23).

Remark: *if $K, L \subset \mathbf{R}^n$ are closed cones and $L \cap K_* \neq \emptyset$, then $[-L_*] \cap K = \{0\}$,. As a result, in the situation if item F, $[\mathrm{Im}\, A] \cap \mathrm{int}\, K_* \neq \emptyset$ implies $[\mathrm{Im}\, A]^\perp \cap K = \{0\}$, whence $A^T K_* = [A^{-1}K]_*$.*

Indeed, let $f \in L \cap \mathrm{int}\, K_*$ and $h \in [-L_*] \cap K$, As $f \in \mathrm{int}\, K_*$ and $h \in K$, we have $f^T h \geq 0$ and $f^T h = 0$ only when $h = 0$; on the other hand, as $f \in L$ and $h \in [-L_*]$, we have $f^T h \leq 0$, which combines with $f^T h \geq 0$ to imply that $f^T h = 0$, whence $h = 0$, Q.E.D.

Calculus of extreme rays

A. When taking intersections or inverse linear images of nontrivial closed pointed cones, there are no simple rules expressing the extreme rays of the result in terms of the extreme rays of the operands.

B. Everything is fine with taking direct product: *Whenever $K^\ell \in \mathbf{R}^{n_\ell}$, $\ell \le L$, are nontrivial closed pointed cones, extreme rays of their direct product $K = K^1 \times ... \times K^L$ are exactly the rays generated by block-vectors $d^1, ..., d^L]$ with exactly one nonzero block that is an extreme direction of the respective factor K^ℓ.*

C. Situation with taking arithmetic sums is good: *When the sum $K = K^1 + ... + K^L$ of nontrivial closed pointed cones is closed and pointed, in every representation of an extreme direction $d = d^1 + ... + d^K$ of K as the sum of directions $d^\ell \in K^\ell$, all nonzero terms are positive multiples of d and are extreme directions of respective cones K^ℓ.*

Indeed, when d is an extreme direction of K, all d^ℓ should be nonnegative multiples of d (as $d^\ell \in K^\ell \subset K$. Besides it, if, say d^1 is nonzero, then d^1 is an extreme direction of K^1; otherwise, we would have $d^1 = d^1_1 + +d^1_2$ with $d^1_1, d^1_2 \in K^1$ with d^1_1 and d^1_2 that are not nonnegative multiples of d^1, or, which is the same, of d, implying representation $d = d^1 + [d^1_2 + d^2 + ... + d^L$ of d as the sum of two vectors, d^1_1 and f, from K, which is impossible, as d is an extreme direction of K, and $d^1_!$ is not a positive multiple of d.

However: not every extreme direction of a factor K^ℓ is an extreme direction of K (look what happens when $K^1 = \{x \in \mathbf{R}^2_+ : x_1 \le x_2\}$ and $K^2 = \{x \in \mathbf{R}^2_+ : x_1 \ge x_2\}$).

D. When taking linear image $K^+ = AK$ of a nontrivial closed pointed cone K, simple examples show that the image AR of an extreme ray R of K is *not* an extreme ray of cl K^+; this may happen even when K^+ is closed, pointed, and nontrivial. However,

Fact IV.21 *If $K \subset \mathbf{R}^n$ is a nontrivial closed pointed cone and its linear image $K^+ = AK$ is nontrivial, closed, and pointed as well, then every extreme ray R of K^+ is the image of an extreme ray of K under the same linear mapping.*

Indeed, let R be an extreme ray of K^+, and $\mathcal{R} = \{x \in K : Ax \in R\}$. Then $\mathcal{R}$ is nontrivial (as $A\mathcal{R} = R \neq \{0\}$) closed pointed cone and therefore it possesses extreme rays and is the conic hull of their union; not all these rays are in Ker$A$, as $A\mathcal{R} \neq \{0\}$. Let $\bar{R}$ be an extreme ray of $\mathcal{R}$ not belonging to Ker$A$; then $A\bar{R} = R$, and all we need to verify is that $\bar{R}$ is an extreme ray of K. Assuming the opposite, there exist $d^1, d^2 \in K$ such that $d = d^1 + d^2$ is a generator of $\bar{R}$ and d^1, d^2 are not nonnegative multiples of d. We have $Ad = Ad^1 + Ad^2$ with $Ad^1, Ad^2 \in K^+$. Now, R is an extreme ray of K^+ and Ad is a generator of $R \Rightarrow Ad^1, Ad^2$ are nonnegative multiples of $Ad \Rightarrow Ad^1, Ad^2 \in R \Rightarrow d^1, d^2 \in \mathcal{R}$, contradicting the fact $\bar{R}$ is an extreme ray of $\mathcal{R}$.

Polar of a convex set

♣ **Definition.** The *polar* of a nonempty convex set $M \subset \mathbf{R}^n$ is the set

$$\text{Polar}(M) := \{y \in \mathbf{R}^n : \ y^T x \leq 1, \ \forall x \in M\}.$$

Examples:

- $\text{Polar}(\mathbf{R}^n) = \{0\}$
- $\text{Polar}(\{0\}) = \mathbf{R}^n$
- Given a linear subspace in $L \subseteq \mathbf{R}^n$, we have $\text{Polar}(L) = L^\perp$ (why?)
- When $K \subset \mathbf{R}^n$ is a cone, $\text{Polar}(K) = -K_*$ (why?)
- The polar of the unit Euclidean ball $B = \{x \in \mathbf{R}^n : \|x\|_2 \leq 1\}$ is B itself (why?)
- Let $M \subset \mathbf{R}^n$ be nonempty convex set and D be a nonsingular $n \times n$ matrix. Then, $\text{Polar}(DM) = D^{-T}\text{Polar}(M)$.
- Let $E = \{x : x^T C x \leq 1\}$, $C \succ 0$, be an ellipsoid centered at the origin. . Then $\text{Polar}(E) = \{x : x^T C^{-1} x \leq 1\}$, i.e., the polar of E is another n-dimensional ellipsoid centered at the origin.

Elementary properties of polars: Let $M \subset \mathbf{R}^n$ be a nonempty convex set.

- $\text{Polar}(M)$ *is a closed convex set containing the origin*
- *Polar of a set remains intact when passing from the set to the closure of the convex full of the union of the set and the origin:*

$$\text{Polar}(M) = \text{Polar}(\text{cl Conv}(M \cup \{0\}))$$

- *Passing to polars reverts inclusions: when* $\emptyset \neq M \subset M' \subset \mathbf{R}^n$, *one has* $\text{Polar}(M') \subset \text{Polar}(M)$.

♠ The polar of a nonempty convex set is a closed convex set containing the origin. In fact, *every set of the latter type is a polar:*

Fact IV.22 *Let $Q \subset \mathbf{R}^n$ be a closed convex set containing the origin. Then* Polar(A) *is a closed convex set containing the origin, and*

$$Q = \text{Polar}(\text{Polar}(Q)) \qquad (!)$$

Thus, *polars of nonempty convex sets are exactly closed convex sets containing the origin, and twice taken polar of a convex set M is the closure of* Conv$(M \cup \{0\}$*. In particular, every closed convex set containing the origin is a polar, specifically, it is the polar of its polar.*

Let $M \subset \mathbf{R}^n$ be a nonempty convex set. We have $Q := \text{Polar}(M) = \{y : y^T x \leq 1 \, \forall x \in M\}$, implying that $\overline{M} := \text{Polar}(Q) \supset M$. Besides this, $\overline{M}$ clearly contains the origin and is closed and convex, whence $\overline{M} \supset \widehat{M} := \text{cl Conv}(M \cup \{0\})$. Let us prove that in fact $\overline{M} = \widehat{M}$. Indeed, assume the opposite and let us lead this assumption to a contradiction. Let $a \in \overline{M} \backslash \widehat{M}$. As $\widehat{M}$ is nonempty, closed, and convex and $a \notin \widehat{M}$, $\{a\}$ and $\widehat{M}$ can be strictly separated: there exists y such that

$$\sup_{x \in \widehat{M}} y^T x < y^T a.$$

As $0 \in \widehat{M}$, we conclude that $y^T a > 0$; passing from y to $\overline{y} = y/(y^T a)$, we arrive at

$$\alpha := \sup_{x \in \widehat{M}} \overline{y}^T x < \overline{y}^T a = 1.$$

As $M \subset \widehat{M}$, we get $\overline{y}^T a = 1 > \alpha \geq \sup_{x \in M} \overline{y}^T x$. Selecting $\theta > 1$ such that $\theta\alpha \leq 1$ and setting $\widehat{y} = \theta \overline{y}$, we get

$$\widehat{y}^T a > 1 \geq \sup_{x \in M} \widehat{y}^T x \qquad (*)$$

The second inequality in $(*)$ says that $\widehat{y} \in \text{Polar}(M)$, and as $a \in \text{Polar}(\text{Polar}(M))$ and $\widehat{y} \in \text{Polar}(M)$, we have $\widehat{y}^T a \leq 1$, contradicting the first inequality in $(*)$; this is the desired contradiction.
The bottom line of our consideration is that *for a nonempty convex set M,* Polar(Polar(M)) = cl Conv$(M \cup \{0\})$. As a result, when M is a closed convex set containing the origin, it is a polar – namely, the polar of its polar.

More facts about polars

♣ Let $M \subset \mathbf{R}^n$ be a closed convex set containing the origin.

A. Question: When Polar(M) is bounded?

Answer: Polar(M) *is bounded iff* M *contains a neighborhood of the origin.*

Recalling that M is the polar of its polar, this is the same as M *is bounded iff* Polar(M) *contains a neighborhood of the origin.*

Proof of the first claim: ✓ When $r > 0$ and $B_r := \{x : \|x\|_2 \leq r\} \subset M$, we have Polar($M$) $\subset$ Polar(B_r) $= \{x \in \mathbf{R}^n : \|x\|_2 \leq 1/r\} \Rightarrow$ Polar(M) is bounded. ✓ Now let $0 \notin \operatorname{int} M$, and let us prove that Polar(M) is unbounded. If Lin(M) $\neq \mathbf{R}^n$, there is a nonzero vector v orthogonal to Lin(M), implying that $\mathbf{R} \cdot v \subset$ Polar(M), and thus Polar(M) is unbounded, as claimed. Now let Lin(M) $= \mathbf{R}^n$. As $0 \in M$, we have Aff(M) $=$ Lin(M) $= \mathbf{R}^n$, and as $0 \notin \operatorname{int} M$, 0 is a point on the boundary of the full-dimensional closed convex set M. As such, 0 is a maximizer, over $x \in M$, of a nonconstant linear form $v^T x$, that is, $0 \neq v$ and $v^T x \leq 0 \, \forall x \in M \Rightarrow$ Polar(M) contains the nontrivial ray $\mathbf{R}_+ \cdot v$ and thus is unbounded. □

B. Question: Let M be polyhedral. Is it true that the polar of M is polyhedral?

Answer: Yes, *The polar of a nonempty polyhedral set* M *is polyhedral, and a polyhedral representation*

$$M = \{x : \exists u : Px + Qu \leq r\}$$

of the set induces explicit polyhedral representation of its polar:

$$\text{Polar}(M) = \left\{y : \exists \lambda : P^T\lambda = y, Q^T\lambda = 0, r^T\lambda \leq 1, \lambda \geq 0\right\}.$$

Indeed, we have

$$\begin{array}{rcl} \text{Polar}(M) & = & \{y : \sup_{x \in M} y^T x \leq 1\} = \left\{y : \max_{x,u}\{y^T x : Px + Qu \leq r\} \leq 1\right\} \\ & = & \left\{y : \min_{\lambda}\{r^T\lambda : P^T\lambda = y, Q^T\lambda = 0, \lambda \geq 0\} \leq 1\right\} \text{ [LP Duality Theorem]} \\ & = & \left\{y : \exists \lambda : P^T\lambda = y, Q^T\lambda = 0, r^T\lambda \leq 1, \lambda \geq 0\right\}. \end{array}$$

Geometry of Polyhedral Sets

♣ **Definition:** A *polyhedral* set Q in $\mathbf{R}^n$ is a subset in $\mathbf{R}^n$ which is a solution set of a finite system of nonstrict linear inequalities:

$$Q \text{ is polyhedral} \Leftrightarrow Q = \{x : Ax \geq b\}.$$

♠ *Every polyhedral set is convex and closed.*
♣ In the sequel, the polyhedral sets in question are assumed to be nonempty.

Extreme points of polyhedral sets

♣ Recall that every nonempty closed convex set not containing lines has extreme points.

Question: *When a nonempty polyhedral set $Q = \{x : Ax \leq b\}$ contains lines? What are these lines, if any?*

Answer: *Q contains lines iff A has a nontrivial kernel:*

$$\mathrm{Ker}A \equiv \{h : Ah = 0\} \neq \{0\}.$$

Directions of lines contained in Q are exactly the nonzero vectors from Ker$A$.

Indeed, a line $\ell = \{x = \bar{x} + th : t \in \mathbf{R}\}$, $h \neq 0$, belongs to Q iff

$$\{\forall t : A(\bar{x} + th) \geq b\} \Leftrightarrow \{\forall t : tAh \geq b - A\bar{x}\} \Leftrightarrow \{Ah = 0 \ \& \ \bar{x} \in Q\}$$

Fact IV.23 *A polyhedral set $Q = \{x : Ax \leq b\}$ always can be represented as*

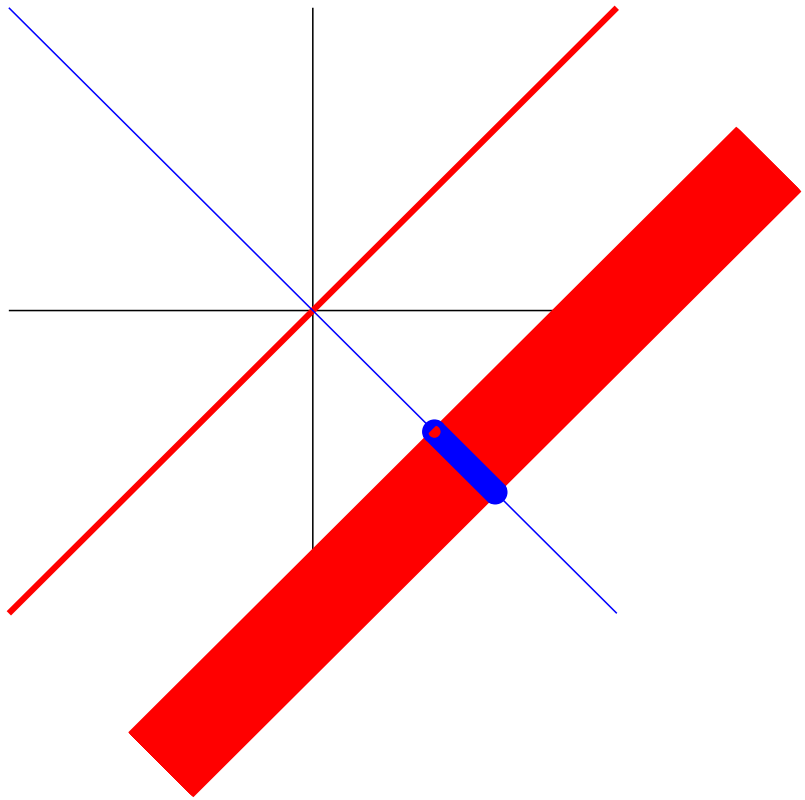
$$Q = Q_* + L,$$

where Q_* is a polyhedral set which does not contain lines and L is a linear subspace. In this representation,

♢ *L is uniquely defined by Q and coincides with* Ker(A),

♢ *Q_* can be chosen, e.g., as*

$$Q_* = Q \cap L^\perp$$



- Red stripe Q: polyhedral set containing lines
- red line L: the recessive subspace of Q
- Blue segment: $Q_* = Q \cap L^\perp$

♠ *Red stripe Q = blue segment Q_* + red subspace L*

♠ *Blue segment Q_*: polyhedral set not containing lines*

Algebraic characterization of extreme point of polyhedral sets

Fact IV.24 *Let $Q = \{x : a_i^T x \leq b_i, i \leq m\}$ be a polyhedral set in $\mathbf{R}^n$, A point $v \in Q$ is an extreme point of Q iff $v \in Q$ and among the constraints $a_i^T x \leq b_i$ which are active at v – are satisfied at v as equalities – there are n constraints with linearly independent vectors of coefficients.*

indeed, let $v \in \text{Ext}(Q)$ and $I = \{i : a_i^T x = b_i\}$ be the set of indices of the constraints active at v. Let us prove that among the vectors a_i, $i \in I$, there are n linearly independent. Assuming the opposite, there exists $h \neq 0$ such that $a_i^T h = 0$, $i \in i$. Setting $d_t^{\pm} = v \pm th$, we get $a_i^T d_t^{\pm} = b_i$ for all $i \in I$; as $a_i 6 < b_i$ for $i \notin I$, we have $a_i^T d_t^{\pm} \leq b_i$ for all small enough $t > 0 \Rightarrow a_i^T]v \pm th \leq b_i$ for all i and all small enough $t > 0 \Rightarrow$ for these t, $v \pm th \in Q$, contradicting $v \in \text{Ext}(Q)$ due to $h \neq 0$.

Vice versa, let $\text{Lin}(a_i : iiI\} = \mathbf{R}^n$, and let us prove that $v \in \text{Ext}(Q)$. To this end we need to verify that when $v \pm h \in Q$, then $h = 0$. Indeed, when $v \pm h \in Q$, we have $a_i^T[v \pm h] \leq b_i$, which for $i \in$ reads $b_i \pm a_i^T h \leq b_i$ $\Rightarrow a_i^T h = 0\, i \in I \Rightarrow h = 0$ (as $\text{Rank}\{a_i : i \in I\} = n$).

Corollary *The number of extreme points of a polyhedral set is finite.*

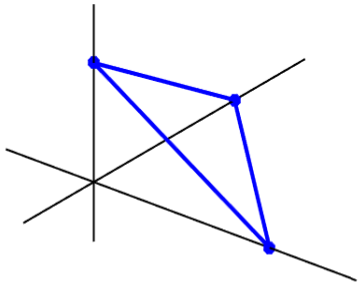
Indeed, among the constraints active at a given extreme point there are n constraints with linearly independent vectors of coefficient $\Rightarrow$ the set I of indices of constraints active at an extreme point v uniquely specifies the point $\Rightarrow$ the number of extreme points of $Q = \{x : a_i^T x \leq b_i, i \leq m\}$ does not exceed $\binom{m}{n}$.

Polyhedral sets with MUST TO KNOW extreme points

A. Let $k \leq n$ be positive integers.
A.1. *The extreme points of the set*

$$\left\{x \in \mathbf{R}^n : 0 \leq x_i \leq 1 \, \forall i, \sum_i x_i = k\right\}$$

are exactly Boolean vectors from the set, that is, 0/1 vectors with exactly k entries equal to 1.
In particular, the extreme points of the "flat (a.k.a. *probabilistic*) simplex"



$$\{x \in \mathbf{R}^n : x \geq 0, \sum_i x_i = 1\}$$

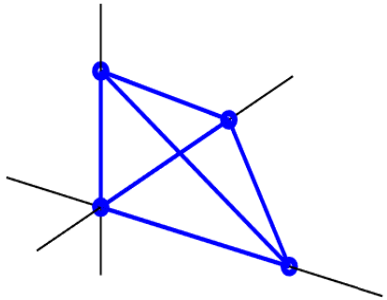
are the standard basic orth (set $k = 1$).

A.2. *The extreme points of the set*

$$\left\{x \in \mathbf{R}^n : 0 \leq x_i \leq 1 \, \forall i, \sum_i x_i \leq k\right\}$$

are exactly Boolean vectors from the set, that is, 0/1 vectors with at most k entries equal to 1.

In particular, the extreme points of the "full-dimensional simplex"



$$\{x \in \mathbf{R}^n : x \geq 0, \textstyle\sum_i x_i \leq 1\}$$

are the standard basic orth *and* the origin (set $k = 1$).

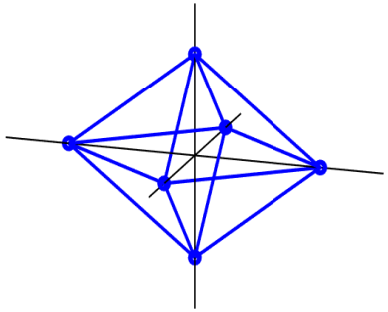
A.3. *The extreme points of the set*

$$\left\{x \in \mathbf{R}^n : |x_i| \leq 1 \,\forall i, \sum_i |x_i| \leq k\right\}$$

are exactly the vectors with k nonzero entries equal to ± 1 each.

In particular,

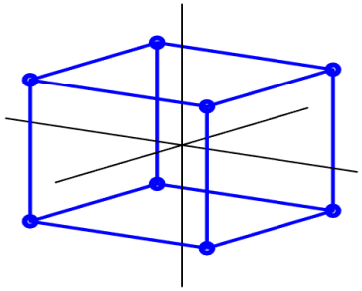
- the extreme points of the unit ℓ_1-ball



$$\{x \in \mathbf{R}^n : \|x\|_1 \leq 1\} = \{x \in \mathbf{R}^n : \textstyle\sum_i |x_i| \leq 1\}$$

are the plus-minus standard basic orth (set $k = 1$).

- the extreme points of the unit ℓ_∞-ball



$$\{x \in \mathbf{R}^n : \|x\|_\infty \leq 1\} = \{x \in \mathbf{R}^n : -1 \leq x_i \leq 1 \,\forall i\}$$

are ± 1 vectors (set $k = n$).

Proof of A.3;

$$Q = \left\{ x \in \mathbf{R}^n : |x_i| \leq 1 \, \forall i, \sum_i |x_i| \leq k \right\}$$

• The only nontrivial part of the claim is that every extreme point of Q is vector with entries 0, ± 1 and exactly k entries equal to ± 1. If you do not see that the inverse is evident, look at the end of this insert.

Proof by bare hands: Let $\bar{x}$ be an extreme point of Q. Then

1) $\bar{x}$ has at most one "fractional entry" - entry of positive magnitude less than 1. Indeed, assuming that there are at least two fractional entries, say, $\bar{x}_1$ and $\bar{x}_2$, let us set $h = [\epsilon; -\epsilon; 0; ...; 0]$ when these entries are of the same sign, and $h = [\epsilon; \epsilon; 0; ...; 0]$, when these entries are of different signs. When $\epsilon > 0$ is small enough, all entries in $\bar{x} \pm h$ are of magnitude ≤ 1, and the sum of their magnitudes is the same as the sum of magnitudes of entries in $\bar{x}$, that is, $\bar{x} \pm h \in Q$ for these ϵ, which is impossible, since $h \neq 0$.

2) $\bar{x}$ has no fractional entries at all. Indeed, by 1) if there is a fractional entry, say, x_1, all other entries are of magnitude 0 or 1, and the sum of magnitudes of all entries is not integer. Consequently, the constraint $\sum_i |x_i| \leq k$ at $\bar{x}$ is satisfied strictly, and therefore the vectors $\bar{x} \pm h$ with $h = [\epsilon; 0; ...; 0]$ belong to Q for small positive ϵ, which again is impossible.

3) The bottom line is that all entries in $\bar{x}$ are 0,± 1, and it remains to see that the number of ± 1 entries, which we know to be $\leq k$ due to $\bar{x} \in Q$, is exactly k. In the opposite case, $\bar{x}$ has a zero entry (since $k \leq n$), say, x_1, and $\bar{x} \pm [\epsilon; 0; ...; 0]$ belongs to Q for all small positive ϵ, which again is impossible $\square$

More intelligent proof: Let $\bar{x}$ be an extreme point of Q. Multiplications by diagonal matrices with ± 1 diagonal entries are symmetries of Q – they map Q onto itself and therefore map extreme points onto extreme points. As a result, we can assume w.l.o.g. that $\bar{x} \geq 0$, and all we need to prove is that $\bar{x}$ has k entries equal to 1 and all remaining entries equal to 0. The set $Q_+ = \{x \in Q : x \geq 0\} = \{x : 0 \leq x_i \leq 1, \sum_i x_i \leq k\}$ is contained in Q and contains $\bar{x}$, so that $\bar{x}$ is an extreme point of Q_+

> I have used the following evident fact: *if $P \subset Q$ are convex sets and $\bar{x} \in P$ is extreme point of Q, then it is extreme point of P* (otherwise $\bar{x}$ would be the midpoint of a nontrivial segment contained in P and therefore contained in Q).

By A.2, $\bar{x}$ has only 0 and 1 entries with at most k entries equal to k. In fact the number of nonzero entries is equal to k, since otherwise $\bar{x}$ would not be an extreme point of Q (last item in the previous proof). □

Finally *every vector $\bar{x}$ with k entries of magnitude 1 and zero remaining entries is an extreme point of Q.* By symmetry, it suffices to verify that the vector $\bar{x}$ with the first k entries of magnitude 1 and zero remaining entries is an extreme point of Q. Indeed, $\bar{x} \in Q$, and assuming that $\bar{x} \pm h \in Q$ for some h, we conclude that $h_1 = ... = h_k = 0$, since otherwise some of the first k entries either in $\bar{x} + h$, or in $\bar{x} - h$ would be of magnitude > 1. We see that the total of magnitudes of entries in $\bar{x} + h$ is $\sum_{i=1}^{k} |\bar{x}|_i + \sum_{i=k+1}^{n} |h_i| = k + \sum_{i=k+1}^{n} |h_i|$, and since this total should be $\leq k$, we conclude that $\sum_{i=k+1}^{n} |h_i| = 0$, the bottom line being that $h = 0$. □

B. A doubly stochastic matrix is a *square matrix with nonnegative entries and all row and column sums equal to 1.* $n \times n$ doubly stochastic matrices form a polytope $\mathcal{P}_n$ in the space $\mathbf{R}^{n \times n}$ of $n \times n$ matrices:

$$\mathcal{P}_n = \{[x_{ij}] \in \mathbf{R}^{n \times n} : x_{ij} \geq 0 \, \forall (i,j), \sum_j x_{ij} = 1 \, \forall i, \sum_i x_{ij} = 1 \, \forall j\}$$

Fact IV.25 [Birkhoff's Theorem] *The extreme points of $\mathcal{P}_n$ are exactly the Boolean matrices from the set, that is, permutation matrices – those with exactly one nonzero entry, equal to 1, in every row and in every column.*

Note: Permutation matrices P are exactly the matrices of linear transformations $x \mapsto Px$ which permute the entries in the argument. Such a matrix is specified by the corresponding permutation, and there are $n!$ of them.

Sketch of the proof: Claim: *If x is an extreme point of $\mathcal{P}$, then the matrix x has an entry equal to 1*
$\Rightarrow$ all other entries in the row and the column of the unit entry are zeros
$\Rightarrow$ eliminating from x the row and the column of the unit entry, we get an $(n-1) \times (n-1)$ doubly stochastic matrix.

Claim: *If x is an extreme point of the polytope $\mathcal{P}$ of doubly stochastic matrices, then matrix x has an entry equal to 1*

Proof: "As is", $\mathcal{P}$ is given by $2n$ linear equalities stating that all row and all column sums in matrix x are equal to 1 plus n^2 inequalities $x_{ij} \geq 0$.

• In fact, we can drop one of the equalities without changing $\mathcal{P}$: if *all column* sums and *all but one row* sums are equal to 1, then *all row and column* sums are equal to 1.

Indeed, the total of all n row sums is equal to the total of all n column sums – both these totals are the sums of all entries in the matrix, and "In fact" follows.

$\Rightarrow$ We lose nothing when assuming that $\mathcal{P}$ is given by n^2 inequalities $x_{ij} \geq 0$ and $2n-1$ linear equalities.

• By algebraic characterization of extreme points, *at an extreme point $\bar{x}$ of $\mathcal{P}$ n^2 of the above constraints should become active*

$\Rightarrow$ *at least $n^2 - 2n + 1$ entries in $\bar{x}$ are zeros*

$\Rightarrow$ *there is a column in $\bar{x}$ with at least $n-1$ zero entries*, since otherwise the total # of zero entries would be at most $n(n-2) < n^2 - 2n + 1$

In the column with at least $n-1$ zero entries the sum of entries is 1, implying that in this column there is exactly one nonzero entry, and this entry is equal to 1.

Application to Assignment problem. *There are n jobs and n workers. Every job takes one man-hour. The profit of assigning worker i with job j is c_{ij}. How to assign workers with jobs in such a way that every worker gets exactly one job, every job is carried out by exactly one worker, and the total profit of the assignment is as large as possible?*
Solution: Assuming for a moment that a worker can distribute his time between several jobs and denoting x_{ij} the fraction of activity of worker i spent on job j, we get a *relaxed* problem

$$\max_x \left\{ \sum_{i,j} c_{ij} x_{ij} : x_{ij} \geq 0, \sum_i x_{ij} = 1 \, \forall j, \sum_j x_{ij} = 1 \, \forall i \right\}$$

The feasible set is polyhedral, nonempty and bounded
$\Rightarrow$ The feasible set is the convex hull of permutation matrices $\Rightarrow$ Program is solvable, and among the optimal solutions there are permutation matrices (since when maximizing a linear function over the convex hull of a finite set, among the maximizers there clearly are points from the set)
$\Rightarrow$ *Relaxation is exact!*

Extreme directions of polyhedral cones: algebraic characterization

♣ Recall that every nontrivial pointed and closed cone has extreme rays, enough for the cone to be the conic hull of the union of these ray. Extreme directions of nontrivial and pointed *polyhedral* cones admit algebraic characterization resembling the algebraic characterization of extreme points of polyhedral sets.

Fact IV.26 *Let $K = \{x : a_i^T x \leq 0, i \leq m\}$ be a nontrivial pointed polyhedral cone. A vector d is an extreme direction of K iff*
— $0 \neq d \in K$, and
— among the constraints $a_i^T x \leq 0\, i \leq m$ which are active at d – are satisfied at d as equalities – there are $n-1$ constraints with linearly independent vectors of coefficient.

Indeed, given $d \in K \backslash \{0\}$, let $I = \{i : a_i^T d = 0\}$.

• Assume, first, that among the vectors $a_i, i \in OI$, there are $n-1$ linearly independent, and let us prove that then d is an extreme direction. Thus, assuming that $d = d^1 + d^2$ with $d^1, d^2 \in K$, let us verify that d^1, d^2 are nonnegative multiples of d. Indeed, for $i \in I$, we have $0 = a_i^T d = a_i^T d^1 + a_i^T d^2$, and both terms in the latter sum are nonpositive due to $d^1, d^2 \in K \Rightarrow a_i^T d^1 = a_i^T d^2 = 0$ for all $i \in I$. Since Rank$\{a_i : i \in I\| \geq n-1$, the linear subspace $L = \{h : a_i^T h = 0, i \in I\}$ is of dimension at most 1; this dimension is exactly 1, since $0 \neq d \in L$. It follows that d^1, d^2, as all vectors from L, are multiples of d; these multiples should be nonnegative, since $0 \neq d \in K$ and K is pointed.

• Now assume that Rank$\{a_i : i \in I\} < n-1$, and let us prove that d is *not* an extreme direction. Indeed, under the circumstances, the linear subspace $L = \{h : a_i^T h = 0, i \in I\}$ is od dimension at least 2 $\Rightarrow$ there exists $h \in L$ not proportional to d. Setting $d_t^{\pm} = \frac{1}{2}[d \pm th]$, we have $d = d_t^+ + d_t^-$ and $a_i^T d_t^{\pm} = 0$, $i \in I$. when $i \notin I$, we have $a_i^T d < 0$ and thus $a_i^T d_t^{\pm} \leq 0$ for all small positive $t \Rightarrow$ for properly selected $t > 0$ $d^{\pm} t \in K$. As $d = d_t^+ + d_t^-$ and $d_t^{\pm}$ for $t \neq 0$ is *not* a multiple of d along with h, d indeed is not an extreme direction, Q.E.D.

Corollary *The number of extreme rays of a nontrivial pointed polyhedral cone is finite.*

Structure of polyhedral set

♣ We are ready to present the fundamental descriptive result on polyhedral sets:

Fact IV.27 [Structure of polyhedral set]
A. *Nonempty polyhedral sets in $\mathbf{R}^n$ are exactly the sets X which can be obtained from finite nonempty set $\{v_1, ..., v_M\} \subset \mathbf{R}^n$ of v-generators and finite (perhaps, empty) set $\{r_1, ..., r_M\} \subset \mathbf{R}^n$ of r-generators representing X according to*

$$X = \mathrm{Conv}(\{v_1, ..., v_M\}) + \mathrm{Cone}\,(\{r_1, ..., r_N\}) = \left\{ x = \sum_i \lambda_i v_i + \sum_j \mu_j r_j : \left\{ \begin{array}{l} \lambda \geq 0, \mu \geq 0 \\ \sum_i \lambda_i = 1 \end{array} \right\} \right. \quad (*)$$

In every representation $(*)$, $\mathrm{Cone}\,(\{r_1, ..., r_N\}) = \mathrm{Rec}(X)$.

• In one direction: $(*)$ is a polyhedral representation of a (clearly, nonempty) set, and polyhedrally representable sets are polyhedral.

• In the opposite direction: Let X be nonempty and polyhedral (and thus closed). Assume, first, that X *does not contain lines*. By Fact IV.6, we have $X = \mathrm{Conv}(\mathrm{Ext}(X)) + \mathrm{Rec}(X)$, and, as we already know, $\mathrm{Ext}(X($ is a nonempty finite set $\{v_1, ..., v_M\}$. Next, $\mathrm{Rec}(X)$ is a pointed (as X has no lines) polyhedral cone. When trivial, we have $\mathrm{Rec}(X)\mathrm{Cone}\,(\emptyset)$, otherwise $\mathrm{Rec}(R)$ is the conic hull of the union of extreme rays, and as we already know, the number of extreme rays is finite. Specifying $\{r_1, ..., r_N\}$ as the set of generators of extreme rays of $\mathrm{Rec}(X)$, we get $\mathrm{Rec}(X) = \mathrm{Cone}\,(\{r_1, ...r_N\}) \Rightarrow X = \mathrm{Conv}(\{v_1, ..., v_M\}) + \mathrm{Cone}\,(\{r_1, ..., r_N\})$ ($N = 0$ when $\mathrm{Rec}(X) = \{0\}$).

When the polyhedral set X contains lines, we have $X = \widehat{X} + \mathrm{Lin}\{f_1, ..., f_K\}$ with not containing lines polyhedral $\widehat{X}$ (Fact IV.23). As we just have seen, $\widehat{X} = \mathrm{Conv}(\{v_1, ..., v_M\}) + \mathrm{Cone}\,(\{r_1, ..., r_N\})$ for properly selected v's and r's$\Rightarrow X = \mathrm{Conv}(\{v_1, ..., v_M\}) + \mathrm{Cone}\,(\{r_1, ..., r_N, \pm f_1, ..., \pm f_K\})$.

Finally, Fact II.26 states that for every representation $X = V + R$ of a closed convex set X as the sum of a bounded set V and closed cone R, one has $R = \mathrm{Rec}(X) \Rightarrow$ *in representation* $(*)$, $\mathrm{Cone}\,(\{r_1, ..., r_N\}) = \mathrm{Rec}(X)$.

B. *Let X be a nonempty polyhedral set not containing lines. Then X has a representation*

$$X = \mathrm{Conv}\{v_1, ..., v_M\} + \underbrace{\mathrm{Cone}\,(\{r_1, ..., r_N\})}_{=\mathrm{Rec}(X)} \qquad (!)$$

where one can select $\{v_1, ..., v_M\} = \mathrm{Ext}(X)$, and this selection if "minimal" – whenever (!) takes place, one has $\mathrm{Ext}(X) \subset \{v_1, ..., v_M\}$.

Application: Extending calculus of polyhedrality

Fact IV.28 *A polyhedral representation*

$$X = \left\{x : \exists u : [x;u] \in X^+ := \{[x;u] : Px + Qu \leq r\}\right\}$$

of a nonempty set X naturally induces polyhedral representation of the recessive cone of the set:

$$\mathrm{Rec}(X) = \{h : \exists v : [h;v] \in \mathrm{Rec}(X^+)\} = \{h : \exists v : Px + Qv \leq 0\}$$

Indeed, by Fact IV.27 we have $X^+ = V + \mathrm{Rec}(X^+)$ with compact V. As $X = \Pi X^+$, where $\Pi[x'u] = x$ is the projection of the space (x,u) where X^+ lives onto the space of x-variables where X lives, we have

$$X = \Pi V + \Pi\mathrm{Rec}(X^+) \qquad (*)$$

ΠV is bounded along with V, and, by polyhedrality, X and the cone $\Pi\mathrm{Rec}(X^+)$ are closed $\Rightarrow$ $(*)$ implies that $\mathrm{Rec}(X) = \Pi\mathrm{Rec}(X^+)$ (Fact II.26), Q.E.D.

Fact IV.29 *A polyhedral representation*

$$X = \{x : \exists u : Px + Qv \leq r\}$$

of nonempty polyhedral set $X \subset \mathbf{R}^n$ naturally induces polyhedral representation of the closed perspective transform

$$\overline{\mathsf{ConeT}}(X) = \mathsf{cl}\,\{[x;t] : t > 0, x/t \in X\}$$

of X:

$$\overline{\mathsf{ConeT}}(X) = \{[x;t] : \exists u : Px + Qu \leq tr, t \geq 0\} \qquad (!)$$

Indeed, it suffices to verify that the cross-sections of both sides in (!) by a hyperplane $\Pi_t = \{[x;t] : x \in \mathbf{R}^n\}$, $t \geq 0$, are the same. When $t > 0$, $\Pi_t \cap \overline{\mathsf{ConeT}}(X) = \{[x;t] : x/t \in X\}$ (this is so for every nonempty closed convex X), that is,

$$\Pi_t \cap \overline{\mathsf{ConeT}}(X) = \{[x;t] : x/t \in X\} = \{[x;t] : \exists u : Px/t + Qu \leq r\} = \{[x;t] : \exists v : Px + Qv \leq tr\},$$

and the concluding set here is the intersection of the right hand side in (!) with Π_t.
As we know from the story about visualization of the recessive cone, $\Pi_0 \cap \overline{\mathsf{ConeT}}(X) = \mathsf{Rec}(X) \times \{0\}$, and by Fact IV.28, one has

$$\mathsf{Rec}(X) \times \{0\} = \{[x;0] : \exists u : Px + Qu \leq 0\},$$

which again is the intersection of the right hand side set in (!) with Π_0.

More on conic hulls

A. Recall that conic hull $\mathrm{Cone}\,(Y)$ of a nonempty set $Y \subset \mathbf{R}^n$ is composed of all conic combinations of vectors from Y, When Y s convex, taking single-term conic combinations already is enough:

Fact IV.30 *Let $Y \subset \mathbf{R}^n$ be nonempty convex set. Then $\mathrm{Cone}\,(Y)$ is composed of nonnegative multiples of vectors from Y.*

Indeed, nonnegative multiples of vectors from Y belong to $\mathrm{Cone}\,(Y)$. To prove that every conic combination $y = \sum_i \lambda_i y^i$ of vectors from Y is a nonnegative multiple of a vector from Y, note that this definitely is so when all λ_i are zero. Otherwise $\sum_i \lambda_i > 0$, and

$$y = [\sum_i \lambda_i]\bar{y},\ \bar{y} = \sum_i \bar{\lambda}_i y^i,\ \bar{\lambda}_i = \lambda_i / \sum_j \lambda_j.$$

$\bar{y}$ is a convex combination of points from Y and thus is a vector from Y (Y is convex!), and y is a positive multiple of $\bar{y}$. □

B. Let $X \subset \mathbf{R}^n$ be a nonempty closed convex set.

$\mathrm{Cone}\,(X)$ necessarily is closed even when X is polyhedral (look at the conic hull f the line $\{x_1 = 1\}$ in $\mathbf{R}^2$ – this is the open right half-plane augmented with the origin). When X is compact and does *not* contain the origin, $\mathrm{Cone}\,(X)$ s closed (Fact II.32 plus Separation Theorem); for compact X containing the origin, all bets are off (the conic hull of the circle in $\mathbf{R}^2$ of radius 1 centered t $[1; 0]$ is the same as the conic hull of the line $\{x_1 = 1\}$). When X is unbounded, $\mathrm{Cone}\,(X)$ can be non-close even when $0 \notin \mathrm{int}\, X$, look at $x = \{[x_1; x_2] \ni \mathbf{R}^2_+ : x_1 x_2 \geq 1\}$).

♠ The situation improves when X is polyhedral:

Fact IV.31 *A polyhedral representation*

$$X = \{x : \exists u : Px + Qu \leq r\}$$

of a nonempty polyhedral set $X \subset \mathbf{R}^n$ naturally induces polyhedral representation of the closure cl Cone (X) *of the conic hull of X:*

$$\text{cl Cone}\,(X) = \widetilde{X} := \{x : \exists \lambda \geq 0, v : Px + Qv \leq \lambda r\}.$$

In one direction: when $x = \sum_i \lambda_i x^i$ is a conic combination of points from X, we have $Px^i + Qu^i \leq r$ for properly selected u^i, whence

$$P[\sum\nolimits_i \lambda_i x^i] + Q[\sum\nolimits_i \lambda_i u^i] \leq [\sum\nolimits_i \lambda_i] r$$

$\Rightarrow x \in \widetilde{X} \Rightarrow$ Cone $(X) \subset \widetilde{X} \Rightarrow$ cl Cone $(X) \subset \widetilde{X}$ (note that $\widetilde{X}$ is polyhedral and thus is closed). Vice versa, let $x \in \widetilde{X}$, and let us prove that $X \in$ cl Cone (X). As $x \in \widetilde{x}$, there exists $\lambda \geq 0$ and u such that

$$Px + Qu \leq r\lambda.$$

Selecting $\bar{x} \in X$ (X is nonempty!), for certain $\bar{u}$ it holds

$$P\bar{x} + Q\bar{u} \leq r.$$

$\Rightarrow$ For every $\epsilon > 0$ it holds

$$P[x + \epsilon\bar{x}] + Q[u + \epsilon\bar{u} \leq [\lambda + \epsilon] r$$

$\Rightarrow x_\epsilon := [x + \epsilon\bar{x}]/[\lambda + \epsilon] \in X \Rightarrow x_\epsilon \in$ Cone (xO). Since Cone $(X) \ni x_\epsilon \to x$ as$\epsilon \to +0$, we have $x \in$ cl Cone X.

The "in addition" part is immediate: representing

$$X = \text{Conv}\{v_1, ..., v_M\} + \text{Cone}\,\{r_1, ..., r_N\}, \qquad (!)$$

we see that when X is bounded (i.e., all r_i are zero), then

$$\text{Cone}\,(X) = \text{Cone}\,(\{v_1, , , v_M\}),$$

which is polyhedral (and thus closed). Similarly, when $0 \in X$, that is,

$$\sum\nolimits_i \bar{\lambda}_i v_i + \sum\nolimits_j \bar{\mu}_j r_j = 0 \qquad [\bar{\lambda}_i \geq 0, \sum\nolimits_i \bar{\lambda}_i = 1, \bar{\mu}_j \geq 0]$$

we have

$$\text{Cone}\,(X) = \text{Cone}\,(\{v_1, ..., v_M, r_1, ..., r_N\}) \qquad (*)$$

(indeed, the left hand side in $(*)$ is a subset of the right hand side by 1). On the other hand, a point from the right hand side is of the generic form

$$y = \sum\nolimits_{i \leq M} \lambda_i v_i + \sum\nolimits_{j \leq N} \mu_j r_j \qquad [\lambda_i \geq 0, \mu_j \geq 0]$$

whence

$$y = \sum\nolimits_i [\lambda_i + \bar{\lambda}_j] v_i + \sum\nolimits_j [\mu_j + \bar{\mu}_j] r_j = (\underbrace{1 + \sum\nolimits_i \lambda_i}_{\kappa}) \bar{y},$$

$$.\bar{y} = \sum\nolimits_i \tfrac{\lambda_i + \bar{\lambda}_i}{\kappa} v_i + \sum\nolimits_j \tfrac{\mu_j + \bar{\mu}_j}{\kappa} r_j$$

By (!), $\bar{y} \in \text{Cone}\,(X).,.$ implying that $y \in \text{Cone}\,(X)..$ We have verified $(*)$, and thus know that $\text{Cone}\,(X)$ is a polyhedral, and thus closed, cone, Q.E.D.

More on convex hulls

♣ The convex hull of the *union* of several polyhedral sets not necessarily is polyhedral (look what happens with the union of the origin and the line $x_1 = 1$ in $\mathbf{R}^2$. However,

Fact IV.32 *Let X^ℓ, $1 \le \ell \le L$, be nonempty polyhedral sets in $\mathbf{R}^n$ given by polyhedral representations*

$$X^\ell = \{x : \exists u^\ell : P_\ell x + Q_\ell u^\ell \le r^\ell\}$$

Then

$$\mathrm{cl}\,\mathrm{Conv}\left(\bigcup_\ell X^\ell\right) = X := \left\{x : \exists y^\ell, u^\ell, \lambda_\ell, \ell \le L : \left\{\begin{array}{l} \lambda_\ell \ge 0, \sum_\ell \lambda_\ell = 1 \\ P_\ell y^\ell + Q\ell u^\ell \le \lambda_\ell r^\ell,\ \ell \le L \\ x = \sum_\ell y_\ell \end{array}\right.\right\}$$

In one direction: since X^ℓ are convex and nonempty, we clearly have

$$\mathrm{cl}\,\mathrm{Conv}\{\cup_\ell X^\ell\} = \mathrm{cl}\left\{\sum_\ell \lambda_\ell x^\ell : \lambda_\ell \ge 0, \sum_\ell \lambda_\ell = 1, x^\ell \in X^\ell\right\} = \mathrm{cl}\,\underbrace{\left\{\sum_\ell \lambda_\ell x^\ell : \lambda_\ell > 0, \sum_\ell \lambda_\ell = 1, x^\ell \in X^\ell\right\}}_{\mathcal{X}}$$

We claim that $\mathcal{X} \subset X$. Indeed, let $\mathcal{X} \ni x = \sum_\ell \lambda_\ell x^\ell$ with $\lambda_\ell > 0$, $\sum_\ell \lambda_\ell = 1$ and $x^\ell \in X^\ell$. As $x^\ell \in X^\ell$, there exists v^ℓ such that $P_\ell x^\ell + Q_\ell v^\ell \le r^\ell \Rightarrow$ setting $y^\ell = \lambda_\ell x^\ell$, $u^\ell = \lambda^\ell v^\ell$, we get $P_\ell y^\ell + Q_\ell u^\ell \le \lambda_\ell r^\ell$ and $x = \sum_\ell y^\ell \Rightarrow x \in X$.

Now let us prove that $\mathcal{X}$ is dense in X. Let $\bar{x}^\ell \in X^\ell$ and $\bar{v}^\ell$ be such that $P_\ell \bar{x}^\ell + Q_\ell \bar{v}^\ell \le r^\ell$, and let $\bar{y}^\ell = L^{-1}\bar{x}^\ell$, $\bar{u}^\ell = L^{-1}\bar{v}^\ell$, $\bar{\lambda}_\ell = L^{-1} \Rightarrow$

$$(a) : P_\ell \bar{y}^\ell + Q_\ell \bar{u}^\ell \le \bar{\lambda}_\ell r^\ell,\ \bar{\lambda}_\ell > 0,\ \sum_\ell \bar{\lambda}_\ell = 1 \quad \forall^n$$

Given $x \in X$, there exist y^ℓ, u^ℓ, λ_ℓ such that

$$(b) : P_\ell y^\ell + Q_\ell u^\ell \le \lambda_\ell r^\ell,\ \lambda_\ell \ge 0,\ \sum_\ell \lambda_\ell = 1 \quad \text{and} \quad (c) : x = \sum_\ell y^\ell$$

By (a) and (b) we have

$$\forall \epsilon \in (0,1) : \left\{\begin{array}{l} \lambda_\ell^\epsilon := (1-\epsilon)\lambda_\ell + \epsilon\bar{\lambda}_\ell > 0, \sum_\ell \lambda_\ell^\epsilon = 1 \\ P_\ell[(1-\epsilon)y^\ell + \epsilon\bar{y}^\ell] + Q_\ell[(1-\epsilon)u^\ell + \epsilon\bar{u}^\ell] \le \bar{\lambda}_\ell r^\ell \Rightarrow x_\epsilon^\ell := [(1-\epsilon)y^\ell + \epsilon\bar{y}^\ell]/\bar{\lambda}_\ell^\epsilon \in X^\ell,\ \ell \le L \\ x_\epsilon := [(1-\epsilon)x + \epsilon\bar{x}] = \sum_\ell \bar{\lambda}_\ell^\epsilon x_\epsilon^\ell \end{array}\right.$$

$\Rightarrow x_\epsilon \in \mathcal{X}$ for all $\epsilon > 0$ and $\lim_{\epsilon \to +0} x_\epsilon = x \Rightarrow \mathcal{X}$ is dense in the *closed* set X (X is polyhedral!) $\Rightarrow \mathrm{cl}\,\mathcal{X} = X$

On the other hand, we knew from the start that $\mathrm{cl}\,\mathcal{X} = \mathrm{cl}\,\mathrm{Conv}(\cup_\ell X^\ell) \Rightarrow \mathrm{cl}\,\{\cup_\ell X^\ell\} = X$, Q.E.D.

Applications to Linear Programming

♣ Consider a Linear Programming program

$$\mathrm{Opt}(c) = \max_{x \in X} c^T x,\ X = \{x \in \mathbf{R}^n : Ax \le b\} \qquad (P)$$

♠ *Assume that* (P) *is feasible* (this assumption imposes no restrictions on c). By Theorem on the structure of polyhedral sets, X admits representation

$$X = \mathrm{Conv}\{\{v_1, ..., v_M\}) + \mathrm{Cone}\,(\{r_1, ..., r_N\}) \qquad (*)$$

where $M > 0$ and $N \ge 0$. This representation says that (P) reduces to a *decoupled* pair of trivial problems

$$\max_\lambda \left\{\sum_i \lambda_i c^T v_i : \lambda \ge 0, \sum_i \lambda_i = 1\right\} \text{ and } \max_\mu \left\{\sum_j \mu_j c^T r_j : \mu \ge 0\right\}$$

making evident the following conclusions:

A. *The domain of the optimal value* $\mathrm{Opt}(c)$ *– the set of* c*'s for which the optimal value is finite – is exactly the same as the domain composed of* c*'s for which the problem is solvable, and is the polyhedral cone*

$$\mathrm{Dom}\,\mathrm{Opt}(\cdot) := \{c : \mathrm{Opt}(c) < \infty\} = \{c : c^T r_j \le 0, j \le N\} = -[\mathrm{Cone}\,(\{r_1, ..., .r_N\})]_*$$

B. *In its domain, the* $\mathrm{Opt}(\cdot)$ *is a piecewise linear convex function*

$$c \in \mathrm{Dom}\,\mathrm{Opt}(\cdot) \Rightarrow \mathrm{Opt}(c) = \max_{i \le M} c^T v_i.$$

positively homogeneous of degree 1

C. *For* $c \in \mathrm{Dom}\,\mathrm{Opt}(\cdot)$*, the set of optimal solutions to* (P) *is*

$$\mathop{\mathrm{Argmax}}_{x \in X} c^T x = \mathrm{Conv}(\{v_i, i \in I_c\}) + \mathrm{Cone}\,(\{r_j : j \in J_c\})$$
$$\left[I_c = \{i : c^T v_i = \max_\iota c^T v_\iota\}, J_c = \{j : c^T r_j = 0\}\right]$$

If X *does not contain lines, which happens iff* $\mathrm{Ker} A = \{0\}$*, then some of the optimal solutions, if any, are extreme points of* X*.*

♣ Additional information on the dependence of the optimal value on the right hand side can be obtained when looking at the dual to (P) problem

$$\mathrm{Opt}_*(b) = \min_{y \in Y} b^T y, \ Y = \{y : A^T y = c, y \geq 0\} \qquad (D)$$

♠ *Assume that (D) is feasible* (this assumption imposes no restrictions on b). By Theorem on the structure of polyhedral sets, Y admits representation

$$Y = \mathrm{Conv}\{\{v_1^*, ..., v_{M_*}^*\}) + \mathrm{Cone}\,(\{r_1^*, ..., r_{N_*}^*\}) \qquad (*)$$

where $M_* > 0$ and $N_* \geq 0$. This representation says that (D) reduces to a *decoupled* pair of trivial problems

$$\min_\lambda \left\{\sum_i \lambda_i b^T v_i^* : \lambda \geq 0, \sum_i \lambda_i = 1\right\} \text{ and } \min_\mu \left\{\sum_j \mu_j b^T r_j^* : \mu \geq 0\right\}$$

making evident the following conclusions:

A*. *The domain of the optimal value $\mathrm{Opt}_*(b)$ – the set of b's for which the optimal value is finite – is exactly the same as the domain composed of b's for which the problem is solvable, and is the polyhedral cone*

$$\mathrm{Dom}\,\mathrm{Opt}_*(\cdot) := \{b : \mathrm{Opt}_*(b) > -\infty\} = \{b : b^T r_j^* \geq 0, j \leq N_*\} = [\mathrm{Cone}\,(\{r_1^*, ..., .r_{N_*}^*\}]_*$$

B*. *In its domain, $\mathrm{Opt}_*(\cdot)$ is a piecewise linear concave function*

$$b \in \mathrm{Dom}\,\mathrm{Opt}_*(\cdot) \Rightarrow \mathrm{Opt}_*(b) = \min_{i \leq M_*} b^T v_i^*.$$

positively homogeneous of degree 1

C*. *For $b \in \mathrm{Dom}\,\mathrm{Opt}_*(\cdot)$, the set of optimal solutions to (D) is*

$$\mathop{\mathrm{Argmin}}_{y \in Y} b^T y = \mathrm{Conv}(\{v_i^*, i \in I_b^*\}) + \mathrm{Cone}\,(\{r_j^* : j \in J_b^*\})$$
$$\left[I_b^* = \{i : b^T v_i^* = \min_\iota b^T v_\iota^*\}, J_b^* = \{j : b^T r_j^* = 0\}\right]$$

Some of the optimal solutions, if any, are extreme point of Y.

Note: As we have seen, the results on the structure of polyhedral sets (Fact IV.27) provide detailed, basically complete, knowledge of the *descriptive* component of LP. Unfortunately, the *operational* (i.e., computational) value of these results is nearly nonexisting – the lists of extreme points and directions appearing in these results usually are astronomically large, and producing these "generators" is incomparably more difficult than solving LP's by existing LP solvers, bad or good alike. "Astronomical" above should be understood literally – the number of extreme points of the feasible domain of Transportation LP with 63 unit capacity suppliers and 63 unit demand customers if $63! \approx 2 \times 10^{87}$ – by 5 orders of magnitude larger than the upper bound $\approx 10^{82}$ on the number of atoms in the universe.
FYI: Solving 63×63 Transportation problem takes less than 1.5"

Law of Diminishing Marginal Returns

♣ Consider a Linear Programming program

$$\mathrm{Opt}_*(b) = \max_x \left\{c^T x : Ax \le b\right\} \qquad (P)$$

and assume that A, c are such that the dual problem is feasible. Then (P) *is solvable for all* b*'s for which it is feasible, and the optimal value considered as the function* $\mathrm{Opt}_*(b)$ *takes values in* $\mathbf{R} \cup \{-\infty\}$ *and is concave on its domain* which is a polyhedral cone.
In addition, by evident reasons, *the recessive cone of* $\mathrm{Dom}\,\mathrm{Opt}_*$ *contains* $\mathbf{R}_+^{\dim b}$, *and* $\mathrm{Opt}_*(b)$ *on its domain is nondecreasing in* b.

♠ It is natural to treat b as the vector of *resources* allocated to certain activity, and the value of the objective – as the profit associated with "production plan" x. Assume that $\Delta \ge 0$ is a given "investment direction;" by investing $t \ge 0$ dollars in the resources, you can increase their "basic level" $\bar{b}$ to $b + t\Delta$. As a result, your optimal profit

$$\Phi(t) = \mathrm{Opt}_*(\bar{b} + t\Delta)$$

will become the larger, the larger t. As you gradually increase your investment t, starting with $t = 0$, your profit can stay $-\infty$ for some (or even all the) time, meaning that investment t is not large enough to make the problem

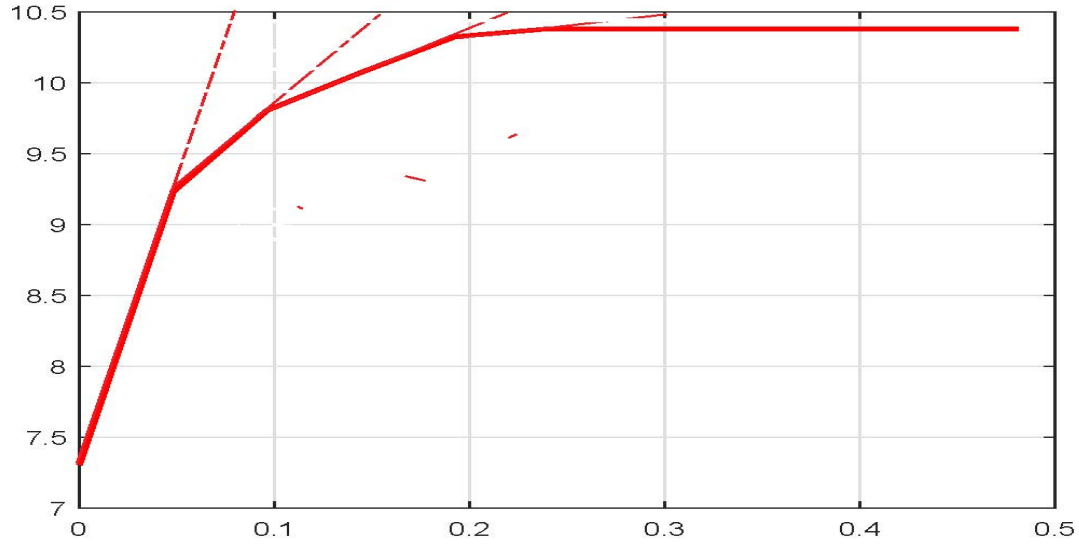
$$\max_x \left\{c^T x : Ac \le \bar{b} + t\Delta\right\}$$

feasible. Assuming that this does not happen for large enough investments t, there will be the smallest $t = \underline{t} \ge 0$ resulting in $\Phi(t) > -\infty$; as $\mathrm{Opt}_*(\cdot)$ is concave piecewise linear function in its domain, your profit $\Phi(t)$ will be real-valued piecewise linear nondecreasing *concave* function on the ray $[\underline{t}, +\infty)$.

♠ That the profit is nondecreasing, is a not so bad news (it hardly is a news – this is what is expected from the very beginning). Concavity of the profit is a not so good news– it means that your marginal return

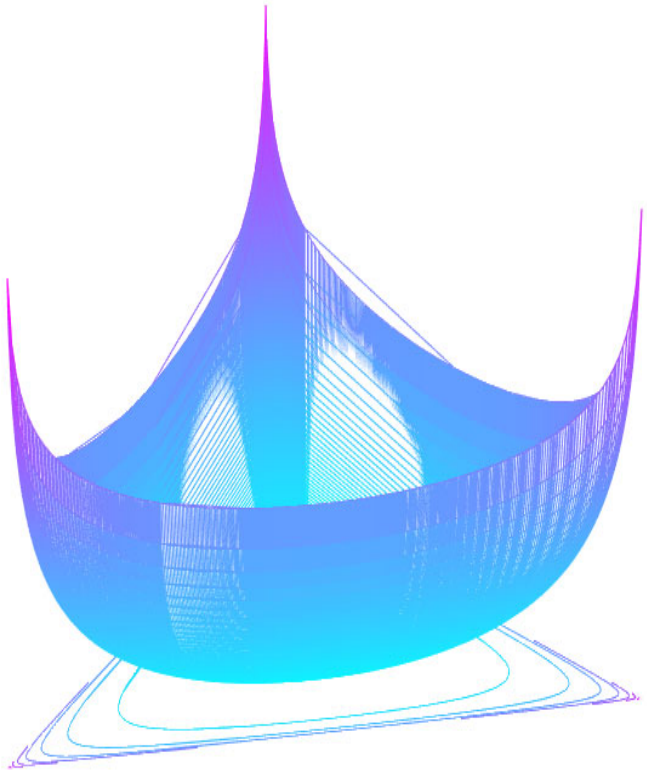
$$\Phi'(d) = \lim_{dt \to +0} \frac{\Phi(t+dt) - \Phi(t)}{dt}$$

is a *nonincreasing* function of $t \in [\underline{t}, \infty)$ – return on investing \$i in resources extra is the smaller the more you have already invested. In Economics, this phenomenon is called *the Law of Diminishing Marginal Returns.*



Profit vs invsetment

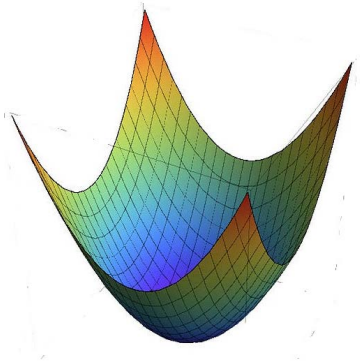
PART II.
Convex Functions



Lecture II.1

Convex Functions – First Acquaintance

Definition
Basic examples
Calculus
How to detect convexity
Local lower boundedness and Lipschitz continuity
Gradient Inequality



Convex Functions: Definition

Definition: *Let f be a real-valued function defined on a nonempty subset* Dom f *in* $\mathbf{R}^n$. *f is called convex, if*

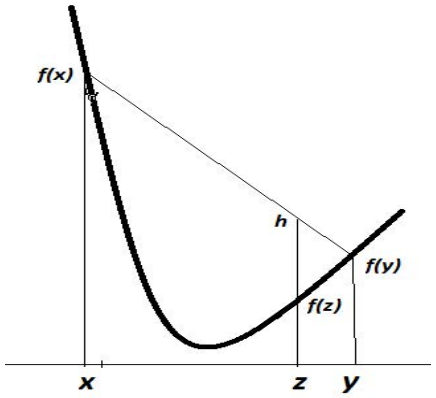
- Dom f *is a convex set*
- *for all* $x, y \in$ Dom f *and* $\lambda \in [0, 1]$ *one has*

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y)$$

Equivalently: Let f be a real-valued function defined on a nonempty subset Dom f in $\mathbf{R}^n$. The function is called convex, if its *epigraph* – the set

$$\text{Epi}\{f\} = \{(x, t) \in \mathbf{R}^{n+1} : f(x) \leq t\}$$

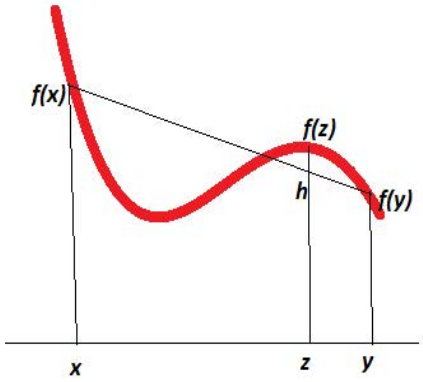
is a convex set in $\mathbf{R}^{n+1}$.



convex function: graph on $[x; y]$ is below secant

$z = \lambda x + (1 - \lambda)y$

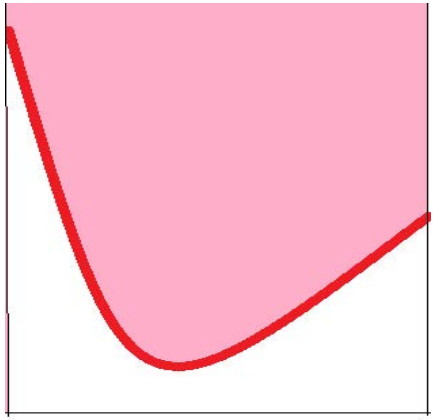
$f(z) \leq h = \lambda f(x) + (1 - \lambda)f(y)$



nonconvex function: graph on $[x; y]$ is *not* entirely below secant

$z = \lambda x + (1 - \lambda)y$

$f(z) > h = \lambda f(x) + (1 - \lambda)f(y)$



epigraph of convex function

What does the definition of convexity actually mean?

♣ The inequality

$$f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y) \qquad (*)$$

where $x, y \in \mathrm{Dom}\, f$ and $\lambda \in [0,1]$ is automatically satisfied when $x = y$ or when $\lambda = 0/1$. Thus, it says something only when the points x, y are distinct from each other and the point $z = \lambda x + (1-\lambda)y$ is a (relative) interior point of the segment $[x, y]$. What does $(*)$ say in this case?

• Observe that $z = \lambda x + (1-\lambda)y = x + (1-\lambda)(y-x)$, whence

$$\|y - x\| : \|y - z\| : \|z - x\| = 1 : \lambda : (1-\lambda)$$

Therefore

$$\begin{array}{c} f(z) \le \lambda f(x) + (1-\lambda) f(y) \qquad (*) \\ \Updownarrow \\ f(z) - f(x) \le \underbrace{(1-\lambda)}_{\frac{\|z-x\|}{\|y-x\|}} (f(y) - f(x)) \\ \Updownarrow \\ \dfrac{f(z) - f(x)}{\|z - x\|} \le \dfrac{f(y) - f(x)}{\|y - x\|} \end{array}$$

Similarly,

$$f(z) \leq \lambda f(x) + (1-\lambda) f(y) \qquad (*)$$

$$\Updownarrow$$

$$\underbrace{\lambda}_{\frac{\|y-z\|}{\|y-x\|}}(f(y) - f(x)) \leq f(y) - f(z)$$

$$\Updownarrow$$

$$\frac{f(y) - f(x)}{\|y-x\|} \leq \frac{f(y) - f(z)}{\|y-z\|}$$

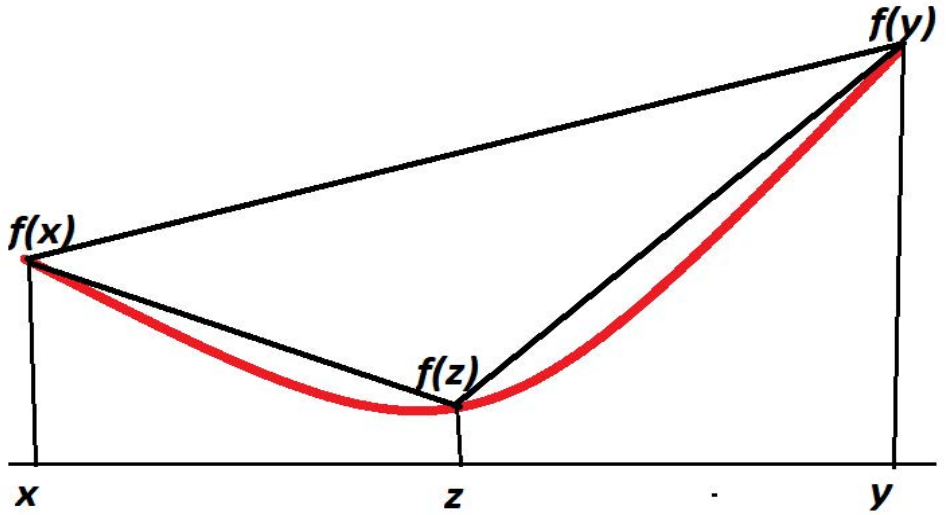
Conclusion: *f is convex iff for every three distinct points x, y, z such that $x, y \in \mathrm{Dom}\, f$ and $z \in [x, y]$, we have $z \in \mathrm{Dom}\, f$ and*

$$\frac{f(z) - f(x)}{\|z - x\|} \leq \frac{f(y) - f(x)}{\|y - x\|} \leq \frac{f(y) - f(z)}{\|y - z\|} \qquad (*)$$

Note: *From 3 inequalities in $(*)$:*

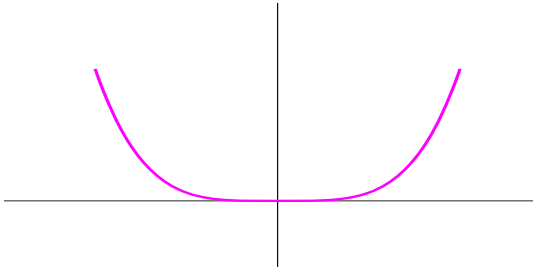
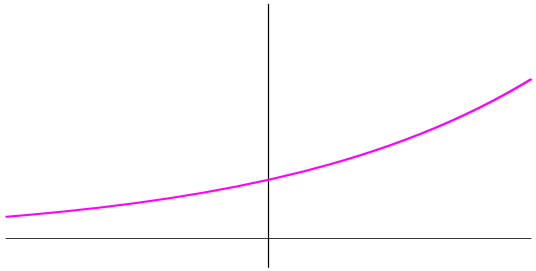
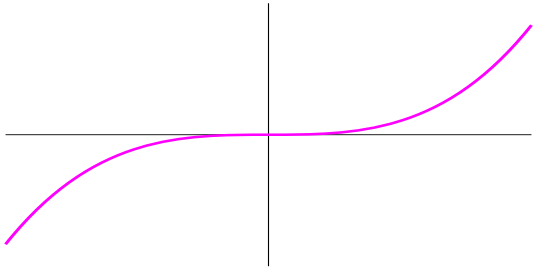
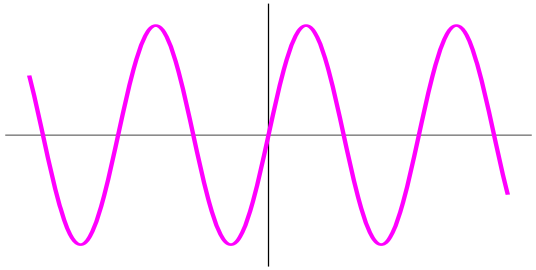
$$\frac{f(z)-f(x)}{\|z-x\|} \leq \frac{f(y)-f(x)}{\|y-x\|}, \quad \frac{f(y)-f(x)}{\|y-x\|} \leq \frac{f(y)-f(z)}{\|y-z\|}, \quad \frac{f(z)-f(x)}{\|z-x\|} \leq \frac{f(y)-f(z)}{\|y-z\|}$$

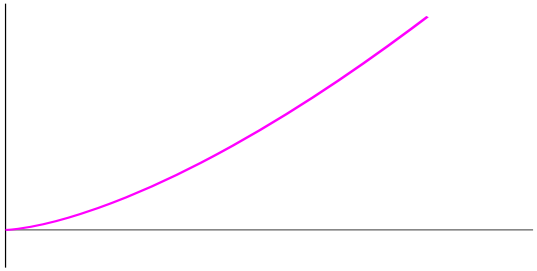
every single one implies the other two.

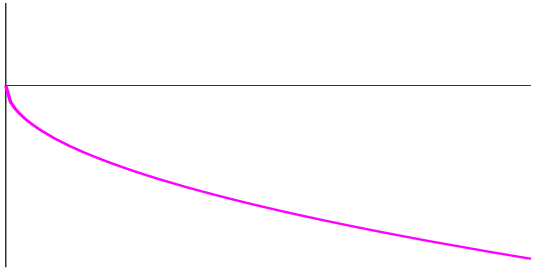


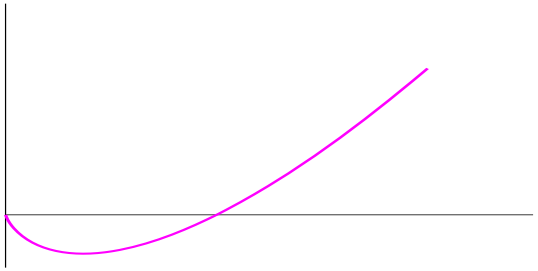
- *When traveling from x to y along $[x, y]$, the overall average rate $\frac{f(y)-f(x)}{\|y-x\|}$ of change in f is in-between the average rate $\frac{f(z)-f(x)}{\|z-x\|}$ of change "in the beginning" and the average rate $\frac{f(y)-f(z)}{\|y-z\|}$ of change "in the end."*

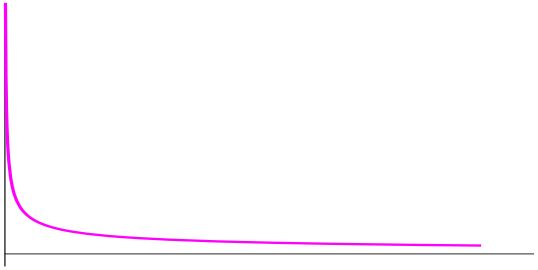
♠ **Examples:**

- **Functions convex on R:**
  - 
    $x^2, x^4, x^6, ...$
  - 
    $\exp\{x\}$
- **Nonconvex functions on R:**
  - 
    x^3
  - 
    $\sin(x)$

- **Functions convex on $\mathbf{R}_+$:**
  - 

    x^p, $p \geq 1$
  - 

    $-x^p$, $0 \leq p \leq 1$
  - 

    $x \ln x$
- **Functions convex on $\mathbf{R}_{++} = \operatorname{int} \mathbf{R}_+ = \{x > 0\}$:**
  - 

    $1/x^p$, $p > 0$

- **Functions convex on $\mathbf{R}^n$:**
- affine function $f(x) = f^T x$
- A norm $\|\cdot\|$ on $\mathbf{R}^n$ is a convex function:

$$\begin{array}{rcll} \|\lambda x + (1-\lambda)y\| & \leq & \|\lambda x\| + \|(1-\lambda)y\| & \text{[Triangle inequality]} \\ & = & \lambda\|x\| + (1-\lambda)\|y\| & \text{[homogeneity]} \end{array}$$

Jensen's Inequality

Fact V.1 [Jensen's Inequality:] *Let $f(x)$ be a convex function. Then*

$$x_i \in \mathrm{Dom}\, f, \lambda_i \geq 0, \sum_i \lambda_i = 1 \Rightarrow$$
$$f(\sum_i \lambda_i x_i) \leq \sum_i \lambda_i f(x_i)$$

Proof:

	$x_i \in \mathrm{Dom}\, f$	
$\Rightarrow$	$(x_i, f(x_i)) \in \mathrm{Epi}\{f\}$	[definition of epigraph]
$\Rightarrow$	$(\sum_i \lambda_i x_i, \sum_i \lambda_i f(x_i)) \in \mathrm{Epi}\{f\}$	[convexity of Epi$\{f\}$]
$\Rightarrow f(\sum_i \lambda_i x_i) \leq \sum_i \lambda_i f(x_i)$	[definition of epigraph]	

Interpretation: *For a convex f and a finitely-valued random vector ξ taking values $x_i \in \mathrm{Dom}\, f$ with probabilities λ_i, $i \leq I$, the value $f(\sum_i \lambda_i x_i)$ of f at the expectation $\sum_i \lambda_i x_i$ of ξ is $\leq$ the expected value $\sum_i \lambda_i f(xi)$ of $f(\xi)$.*

Extension: *Let f be convex, $\mathrm{Dom}\, f$ be closed and f be continuous on $\mathrm{Dom}\, f$. Consider a probability distribution Π supported on $\mathrm{Dom}\, f$. Then*

$$f(\mathbf{E}_\Pi\{x\}) \leq \mathbf{E}_\Pi\{f(x)\}.$$

Illustration: Nonnegativity of Kullback-Leibler distance

Fact V.2 *Let $p = \{p_i > 0\}_{i=1}^N$, $q = \{q_i > 0\}_{i=1}^N$ be two discrete probability distributions. Then the Kullback-Libeler distance*

$$\sum_i p_i \ln \frac{p_i}{q_i}$$

between the distributions is nonnegative.

Indeed, the function $f(x) = -\ln x$, $\mathrm{Dom}\, f = \{x > 0\}$, is convex. Setting $x_i = q_i/p_i$, $\lambda_i = p_i$ we have

$$\begin{aligned} 0 &= -\ln\left(\sum_i q_i\right) = f(\sum_i p_i x_i) \\ &\le \sum_i p_i f(x_i) = \sum_i p_i(-\ln q_i/p_i) \quad \text{[Jensen's Inequality]} \\ &= \sum_i p_i \ln(p_i/q_i) \end{aligned}$$

What is the value of a convex function outside its domain?

♣ **Convention.** To save words, it is convenient to think that a convex function f of n variables is defined *everywhere* on $\mathbf{R}^n$ and takes real values *and value* $+\infty$. With this interpretation, f "remembers" its domain:

$$\text{Dom } f = \{x : f(x) \in \mathbf{R}\} \qquad x \notin \text{Dom } f \Rightarrow f(x) = +\infty$$

and the definition of convexity becomes

$$f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y) \quad \forall (x, y \in \mathbf{R}^n, \lambda \in [0,1])$$

where the arithmetics on the *Extended Real Line* $\overline{\mathbf{R}} = \mathbf{R} \cup \{+\infty\} \cup \{-\infty\}$ is given by the following rules

- **Summation:** $a + b = \begin{cases} a+b & \text{, [the usual sum], both summands are reals} \\ +\infty & \text{, one of the summands is real, another is } +\infty \\ +\infty & \text{, both summands are } +\infty \\ -\infty & \text{, one of the summands is real, another is } -\infty \\ -\infty & \text{, both summands are } -\infty \\ \text{undefined} & \text{, one of the summands is } +\infty, \text{ another is } -\infty \end{cases}$
- **Multiplication:** The magnitude $|a \cdot b|$ of product $a \cdot b$ is

$$|a \cdot b| = \begin{cases} |a||b| & \text{, [the usual product] when } a, b \text{ are reals} \\ +\infty & \text{, one of the factors is } \pm\infty, \text{ another is nonzero} \\ 0 & \text{, one of the factors is zero} \end{cases}$$

The sign of $a \cdot b$ is given by the usual rule (and, of course, the sign of $+\infty$ is "plus", the sign of $-\infty$ is "minus"). For example $2 \cdot 2 = 4$, $2 \cdot [+\infty] = +\infty$, $0 \cdot \pm\infty = 0$, $-1 \cdot \pm\infty = \mp\infty$, $[+\infty] \cdot [\pm\infty] = \pm\infty$.

- **Comparison** between reals is understood in the usual sense, any real is $< +\infty$ and $> -\infty$, and $-\infty < +\infty$.

Convexity of sublevel sets of convex functions

Fact V.3 *Let $f : \mathrm{Dom}\, f \to \mathbf{R}$ be a convex function on convex domain $\mathrm{Dom}\, f$. Then for every real a the sublevel set*

$$\mathrm{Lev}_a(f) = \{x \in \mathrm{Dom}\, f : f(x) \leq a\}$$

of f is convex.

Indeed, if $x, y \in \mathrm{Lev}_a(f)$ and $\lambda \in (0,1)$, then

$$f(z := \lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) \leq \lambda a + (1-\lambda)a = a,$$

that is, $\lambda x + (1-\lambda)y \in \mathrm{Lev}_a(f)$.

Calculus of Convex Functions

Fact V.4 *The following operations with functions taking values in $\mathbf{R}\cup\{+\infty\}$ preserve convexity:*

A. Taking conic combinations: *When $f_i(x):\mathbf{R}^n\to\mathbf{R}\cup\{+\infty\}$, $i\le I$, are convex functions, and λ_i are nonnegative reals, the function*

$$f(x)=\sum_i \lambda_i f_i(x):\mathbf{R}^n\to\mathbf{R}\cup\{+\infty\}$$

is convex.

B. Taking supremum: *The pointwise supremum $f(x)=\sup_{\alpha\in\mathcal{A}} f_\alpha(x)$ of convex functions $f_\alpha(\cdot):\mathbf{R}^n\to\mathbf{R}\cup\{+\infty\}$, $\alpha\in\mathcal{A}$ is convex.*

Indeed, Epi$\{f\}=\cap_\alpha$Epi$\{f_\alpha\}$

C. Affine substitution of argument: *If $f(x):\mathbf{R}^n\to\mathbf{R}\cup\{+\infty\}$ is convex function on $\mathbf{R}^n$ and $x=Ay+b$ is an affine mapping from $\mathbf{R}^m$ to $\mathbf{R}^n$, then the function $g(y)=f(Ay+b)$ is convex on $\mathbf{R}^K$*

D. Partial minimization: *Let $Q\subset\mathbf{R}^n$ be a convex set and $f(x,w):\mathbf{R}^n_x\times\mathbf{R}^k_w\to\mathbf{R}\cup\{+\infty\}$ be a convex function. Assume that the function $g(x)=\inf_w f(x,w)$ does not take value $-\infty$ on Q. Then g is convex on Q.*

Under our assumption, $g:Q\to\mathbf{R}\cup\{+\infty\}$. All we need is to verify is that when $x,y\in Q$ with $g(x)<\infty$, $g(y)<\infty$, and $\lambda\in(0,1)$, one has $f(z:=\lambda x+(1-\lambda)y)\le\lambda f(x)+*1-\lambda)f(y)$. For every $\epsilon>0$ there exist w_x, w_y such that $f(x,w_x)\le g(x)+\epsilon$. $f(y,w_y)\le g(y)+\epsilon\Rightarrow f(z,\lambda w_x+(1-\lambda w_y)\le\lambda f(x,w_x)+(1-\lambda)f(y,w_y)\le\lambda g(x)+(1-\lambda)g(y)=\epsilon$ $\Rightarrow g(z)\le\lambda g(x)+(1-\lambda)g(y)+\epsilon\Rightarrow g(z)\le\lambda g(x)+(1-\lambda)g(y)$.

E. Taking perspective transform: *Let $f(x) : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ be convex. Then so is the perspective transform*

$$F(x,\tau) = \begin{cases} \tau f(x/\tau) & ,\tau > 0 \\ +\infty & ,\tau \leq 0 \end{cases}$$

of f.
Indeed,

$$\mathrm{Epi}\{F\} = \{[[x;\tau];t] : \tau > 0, \tau f(x/\tau) \leq t\} = \{[[x;\tau];t] : \tau > 0, f(x/\tau) \leq t/\tau\} = \{[[x;\tau];t] : \tau > 0, [x;t]/\tau \in \mathrm{Epi}\{f\}\}$$
$$= \{[x;\tau;t] : [[x;t];\tau] \in \mathrm{Persp}\{\mathrm{Epi}\{f\}\}$$

and the perspective transform of a convex set is convex.

Illustration: *The function $\alpha \ln(\alpha/\beta)$ is convex in the quadrant $\{\alpha > 0, \beta > 0\}$.*
Indeed, the function is projective transformation of the convex function

$$f(\beta) = \begin{cases} -\ln(\beta) & ,\beta > 0 \\ +\infty & ,\beta \leq 0 \end{cases}$$

F. Monotone superposition:

Fact V.5 [Monotone superposition] *Let $f_i(x) : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$, $i \leq K$, and $F : \mathbf{R}^K \to \mathbf{R} \cup \{+\infty\}$ be convex functions, and let*

$$g(x) = \begin{cases} F(f_1(x), ..., f_K(x)) & , x \in \mathrm{Dom}\, f_i, \forall i \\ +\infty & , \textit{otherwise} \end{cases}$$

Assume that $F(y_1, ..., y_K)$ is nondecreasing in every one of its arguments y_k for which f_k is not affine. Then g is convex.

Indeed, let $x, x' \in \mathrm{Dom}\, g$ and $\lambda \in (0,1)$, and let us prove that $g(z := \lambda(1-\lambda)x') \leq \lambda g(x) + (1-\lambda)g(x')$. Indeed, setting $f = [f_1; ...; f_K]$, $y = f(x)$, $y' = f(x')$, $z_k = \lambda f_k(x) + (1-\lambda) f_k(x')$, we have by convexity/affinity of f_k

$$z_k \begin{cases} = f_k(z) & , f_k \text{ is affine} \\ \leq f_k(z) & , \text{otherwise} \end{cases},$$

which combines with (partial) monotonicity of F to imply that $g(z) \leq F(\lambda f(x) + (1-\lambda) f(x'))$; taken together with the convexity of F this implies that $g(\lambda x + (1-\lambda)x') \leq \lambda F(f(x)) + (1-\lambda)F(f(x')) = \lambda g(x) + (11-\lambda)g(x')$, Q.E.D.

Refinement: *Let $f_k : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$, $k \leq K$ and $F : \mathbf{R}^K \to \mathbf{R} \cup \{+\infty\}$ be convex functions, and $Y \subset \mathbf{R}^k$ be a convex set such that $f(x) \in Y$ whenever all $f_k(x)$ are finite. Let also $\mathcal{K}$ be the set of indices k for which f_k is not affine. Assume that F is nonincreasing in everyone of y_k, $k \in \mathcal{K}$ on Y only:*

$$\forall \left(y, y' \in Y : y_k = y'_k,\, k \notin \mathcal{K},\, y_k \leq y'_k,\, k \in K\right) : F(y) \leq F(y').$$

Then g is convex.

Illustration. Theorem on Superposition is not applicable to the composition $g = F(f(x))$ of $f(x) = x^2 - 1$ and $F(y) = y^2$, and the composition is in fact nonconvex. Theorem remains unapplicable when $f(x) = x^2 - 1$ is replaced with $f(x) = x^2 + 1$, but its Refinement works (set $Y = \{y \geq 0\}$), certifying the convexity of $(x^2 + 1)^2$

How to detect convexity?

♣ *Convexity is one-dimensional property:*

Fact V.6 • *A set $X \subset \mathbf{R}^n$ is convex iff the set*

$$\{t : a + th \in X\}$$

is, for every (a, h), a convex set on the axis

• *A function $f : \mathbf{R}^n \to \mathbf{R} \cup \{\infty\}$ is convex iff the univariate function*

$$\phi(t) = f(a + th)$$

is, for every (a, h), a convex function on the axis.

♣ **When a function ϕ on the axis is convex?**

Let ϕ be convex and finite on (a, b). This is exactly the same as

$$\frac{\phi(z)-\phi(x)}{z-x} \leq \frac{\phi(y)-\phi(x)}{y-x} \leq \frac{\phi(y)-\phi(z)}{y-z}$$

when $a < x < z < y < b$. Assuming that $\phi'(x)$ and $\phi'(y)$ exist and passing to limits as $z \to x+0$ and $z \to y-0$, we get

$$\phi'(x) \leq \frac{\phi(y)-\phi(x)}{y-x} \leq \phi'(y)$$

that is, *$\phi'(x)$ is nondecreasing on the set of points from (a, b) where it exists*.

Differential Criteria of Convexity

Fact V.7 *A differentiable function $f : (a, b) \to \mathbf{R}$ is convex on (a, b) iff its derivative f' is nondecreasing on (a, b). A twice differentiable function $f : (a, b) \to \mathbf{R}$ is convex on (a, b) iff f'' is nonnegative everywhere on (a, b).*

Indeed, we have just seen that when f is convex and real-valued on (a, b), the derivative is nondecreasing on the set where it exists $\Rightarrow$ when f is convex *and differentiable* on (a, b), f' is nondecreasing on (a, b). Vice versa, let f be differentiable and f' be nondecreasing on (a, b). To prove that f is convex, we should verify that when $a < x < z < y < b$, then

$$\frac{f(z) - f(x)}{z - x} \leq \frac{f(y) - f(z)}{y - z} \qquad (*)$$

By the Mean Value Theorem the left hand side ratio if $f'(\xi)$ for some $\xi \in (x, z)$, and the right hand side ratio is $f'(\eta)$ for some $\eta \in (z, y)$. Since f' is nondecreasing and $\eta \geq \xi$, $(*)$ follows.

When f' is differentiable on (a, b), f' is nondecreasing on (a, b) iff $f'' \geq 0$ on (a, b), which combines with the already proved part of Fact to complete the proof.

♠ Recalling that a multivariate function is convex iff its restrictions on all lines are so, we arrive at

Corollary *Let $f : A \to \mathbf{R}$ be a function defined on an open convex domain Q and let f be twice continuously differentiable on Q. F is convex iff the second order directional derivative $\frac{d^2}{dt^2}\big|_{t=0} f(x + th)$ of f taken at any point $x \in Q$ along any direction $h \in \mathbf{R}^n$ is nonnegative, or, which is the same, iff*

$$f''(x) \succeq 0 \ \forall x \in Q.$$

♠ The above results allow to establish convexity of multivariate functions with convex, not necessarily open, domains due to the following

Observation: *Let $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ be a function with convex domain* Dom f. *When f is continuous on* Dom f *and is convex on* rint Dom f, *f is convex.*

Immediate consequences

♠ The differential criteria of convexity we have just established are sufficient to justify convexity of all univariate functions, like x^{2k}, $k = 0, 1, ...$, $\exp\{x\}$, $x^p : [0, \infty) \to \mathbf{R}$, $p \geq 1$, $-x^p : [0, \infty) \to \mathbf{R}$, $0 \leq p \leq 1$, $x \ln x : [0, \infty \to \mathbf{R}$ claimed so far to be convex.
As a matter of fact, these univariate convex "row materials" plus Calculus of convexity allow to establish convexity of typical convex functions arising in applications. There are few "exceptions" – important convex functions for which convexity is established "by bare hands" – via multivariate differential criteria.

Examples:

- *The function $f(x) = \ln(\sum_{i=1}^N \exp\{x_i\})$ is convex.*

Indeed, given $x \in \mathbf{R}^n$ and $n \in \mathbf{R}^n$ and setting $p_i = \frac{\exp\{x_i\}}{\sum_j \exp\{x_j\}}$, so that $\sum_i p_i = 1$, direct computation shows

$$\frac{d^2}{dt^2}\Big|_{t=0} f(x + th) = \sum_i p_i h_i^2 - (\sum_i p_i h_i)^2$$

We see that $\frac{d^2}{dt^2}\Big|_{t=0} f(x + th)$ is the variance (expected square minus squared expectation) of random variable taking values $h_1, ..., h_N$ with probabilities $p_1, ..., p_N$, and variance is nonnegative.
Verification:

$$\left(\sum_i p_i h_i\right)^2 = \left(\sum_i \sqrt{p_i}[\sqrt{p_i} h_i]\right)^2 \leq \left(\sum_i p_i\right)\left(\sum_i p_i h_i^2\right) \leq \sum_i p_i h_i^2.$$

However: We can extract the convexity of f from Calculus of convexity:

$$\mathrm{Epi}\{f\} = \{[x; t] : \ln(\sum_i \exp\{x_i\}) \leq t\} = \{[x; t] : \sum_i \exp\{x_i\} \leq \exp\{t\}\} = \{[x; t] : \sum_i \exp\{x_i - t\} \leq 1\},$$

and the concluding set is convex as the sublevel set of convex function.

Corollary: *When $c_i > 0$, the function $g(y) = \ln\left(\sum_i c_i \exp\{a_i^T y\}\right)$ is convex.*

Indeed, $g(y) = \ln\left(\sum_i \exp\{\ln c_i + a_i^T y\}\right)$ is obtained from the convex function $\ln\left(\sum_i \exp\{x_i\}\right)$ by affine substitution of argument.

• *Let $\pi_i > 0$, $\sum_i \pi_i \le 1$. Then the function $f(x) = \prod_i x_i^{\pi_i} : \mathbf{R}^n_+ \to \mathbf{R}$ is concave*
(that is, the function $-f(x)$ with the domain $\mathbf{R}^n_+$ is convex).

]Indeed, f is continuous on $\mathbf{R}^n_+$ $\Rightarrow$ to establish convexity $g(x) = -f(x)$, is suffices to prove that if $x > 0$, then $g''(x) \succeq 0$. Indeed, given $x \in \operatorname{int}\mathbf{R}^n_+$ and $h \in \mathbf{R}^n$, direct computation results in

$$\begin{aligned} h^T\nabla g(x) &= \left[\sum_i \pi_i (h_i/x_i)\right] g(x) \\ h^T\nabla^2 g(x) h &= \left[\left[\sum_i \pi_i \tfrac{h_i}{x_i}\right]^2 - \sum_i \pi_i \tfrac{h_i^2}{x_i^2}\right] g(x) \end{aligned}$$

We have $g(x) \le 0$ and

$$\left[\sum_i \pi_i \frac{h_i}{x_i}\right]^2 \le \Big[\underbrace{\sum_i \pi_i}_{\in[0,1]}\Big]^2 \left[\sum_i \pi_i \frac{h_i^2}{x_i^2}\right] \le \left[\sum_i \pi_i \frac{h_i^2}{x_i^2}\right]$$

$\Rightarrow h^T\nabla^2 g(x) h \ge 0$.

• *When $\pi_i > 0$, the function $f(x) = \prod_i x_i^{-\pi_i} : \operatorname{int}\mathbf{R}^n_+ \to \mathbf{R}$ is convex,*

Indeed, the function $\ln(g(x)) = \sum_i \pi_i \ln(1/x_i) : \operatorname{int}\mathbf{R}^n_+ \to \mathbf{R}$ is convex $\Rightarrow$ $g(x) = \exp\{\text{convex function}\}$ is convex.

Below boundedness and Lipschitz continuity of a convex function

Fact V.8 [EM, Proposition II.10.11] *A convex function $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ is below bounded on every bounded subset of $\mathbf{R}^n$*

Fact V.9 [EM Theorem II.8.1] *Let f be a convex function, and let K be a closed and bounded set belonging to* rint Dom f. *Then f is Lipschitz continuous on K, that is, there exists a constant $L < \infty$ such that*

$$|f(x) - f(y)| \leq L\|x - y\|_2 \quad \forall x, y \in K.$$

Note: All three assumptions on K are essential, as is shown by the following examples:
- $f(x) = -\sqrt{x}$, Dom $f = \{x \geq 0\}$, $K = [0, 1]$. Here $K \subset$ Dom f is closed and bounded, but is not contained in rint Dom f, and f is *not* Lipschitz continuous on K (as $\lim_{t\to+0}(f(0) - f(t))/t = \infty$)
- $f(x) = x^2$, Dom $f = K = \mathbf{R}$. Here K is closed and belongs to rint Dom f, but is unbounded, and f is *not* Lipschitz continuous on K (as $\lim_{t\to\infty}(f(t) - f(0))/t = +\infty$)
- $f(x) = \frac{1}{x}$, Dom $f = \{x > 0\}$, $K = (0, 1]$. Here K is bounded and belongs to rint $Dom f$, but *is not closed*, and f is *not* Lipschitz continuous on K (as $\lim_{t\to+0} \lim_{\tau\to t+0}(f(t) - f(\tau))/(\tau - t) = +\infty$)

Gradient Inequality

Fact V.10 [Gradient Inequality] *Let $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ be a convex function and $x \in \mathrm{Dom}\, f$ be such that f is differentiable at x, meaning that there exists a vector $\nabla f(x)$ such that*

$$\forall \epsilon > 0 \exists \delta > 0 : y \in \mathrm{Dom}\, f \ \&\ \|y - x\| \leq \delta \Rightarrow |f(y) - f(x) - (y-x)^T \nabla f(x)| \leq \epsilon \|y - x\|.$$

Then

$$\forall y : f(y) \geq f(x) + (y-x)^T \nabla f(x). \qquad (*)$$

Proof. Let $y \in \mathbf{R}^n$, and let us prove that $(*)$ takes place. There is nothing to prove when $y = x$ or $f(y) = +\infty$, thus, assume that $f(y) < \infty$ and $y \neq x$. Let is set $z_\epsilon = x + \epsilon(y - x)$, $0 < \epsilon < 1$. Then z_ϵ is an interior point of the segment $[x, y]$. Since f is convex, we have

$$\frac{f(y) - f(x)}{\|y - x\|} \geq \frac{f(z_\epsilon) - f(x)}{\|z_\epsilon - x\|} = \underbrace{\frac{f(x + \epsilon(y-x)) - f(x)}{\epsilon}}_{\to (y-x)^T f'(x) \text{ as } \epsilon \to +0} \cdot \frac{1}{\|y - x\|}$$

Passing to limit as $\epsilon \to +0$, we arrive at

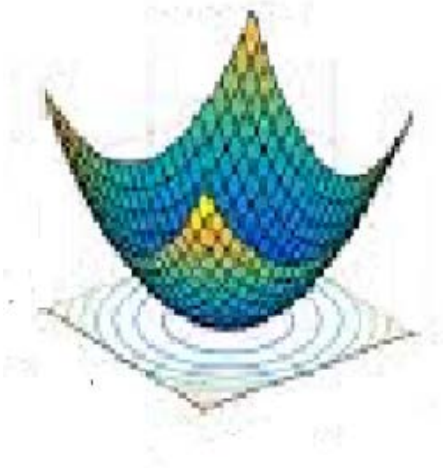
$$\frac{f(y) - f(x)}{\|y - x\|} \geq \frac{(y-x)^T f'(x)}{\|y - x\|},$$

as required by $(*)$.

Lecture II.2

Maxima and Minima of Convex Functions

Minimizing convex functions: Unimodality
Minimizing convex functions: Optimality conditions
Maxima of convex functions
Subgradients



Minimizing Convex Functions: Unimodality

Fact VI.1 [Unimodality] *Let $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ be a convex function and x_* be a local minimizer of f:*

$$x_* \in \mathrm{Dom}\, f \ \&\ \exists r > 0 : f(x) \geq f(x_*)\ \forall (x : \|x - x_*\|_2 \leq r).$$

Then x_* is a global minimizer of f:

$$f(x) \geq f(x_*)\ \forall x.$$

Proof. All we need to prove is that if $x \neq x_*$ and $x \in \mathrm{Dom}\, f$, then $f(x) \geq f(x_*)$. To this end let $z \in (x_*, x)$. By convexity we have

$$\frac{f(z) - f(x_*)}{\|z - x_*\|} \leq \frac{f(x) - f(x_*)}{\|x - x_*\|}.$$

When $z \in (x_*, x)$ is close enough to x_*, we have $\frac{f(z)-f(x_*)}{\|z-x_*\|} \geq 0$, whence $\frac{f(x)-f(x_*)}{\|x-x_*\|} \geq 0$, that is, $f(x) \geq f(x_*)$, Q.E.D.

Fact VI.2 *Let $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ function. Then the set X_* of global minimizers of f is convex.*

Indeed, when $X_* \neq \emptyset$, X_* is the sublevel set $\mathrm{Lev}_a = \{x : f(x) \leq a := \min_x f(x) \in \mathbf{R}\}$, and a sublevel set of a convex function is convex (Fact V.3).

When the minimizer of a convex function is unique?

Definition: A convex function $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ is called *strictly convex*, if

$$f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda) f(y)$$

whenever $x, y \in \operatorname{Dom} f$, $x \neq y$ and $\lambda \in (0,1)$.

Note: *If a convex function f has open domain and is twice continuously differentiable on this domain with*

$$h^T f''(x) h > 0 \quad \forall (x \in \operatorname{Dom} f, h \neq 0),$$

then f is strictly convex.

Fact VI.3 *For a strictly convex function f a minimizer, if it exists, is unique.*

Proof. Assume that $X_* = \operatorname{Argmin} f$ contains two distinct points x', x''. By strong convexity,

$$f(\frac{1}{2}x' + \frac{1}{2}x'') < \frac{1}{2}\left[f(x') + f(x'')\right] = \inf_x f,$$

which is impossible.

Minimizing convex functions: Optimality conditions

Fact VI.4 [Optimality conditions in convex minimization] *Let $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ be a convex function and $x_* \in \mathrm{Dom}\, f$ be a point at which f is differentiable. x_* is a minimizer of f iff*

$$\forall x \in \mathrm{Dom}\, f : [x - x_*]^T \nabla f(x_*) \geq 0. \qquad (*)$$

In one direction: let $(*)$ hold true. By Gradient inequality we have $f(x) \geqq [s - s_*]^T \nabla f(x_*)$, which combines with $(*)$ to imply that $f(x) \geq f(x_*)$ for all $x \in \mathrm{Dom}\, f$. The same inequality holds true when $x \notin \mathrm{Dom}\, f$, that is when $f(x) = +\infty$.

In the opposite direction: Let x_* be a minimizer of f, and let us prove that $(*)$ holds. When $x = x^*$, $(*)$ is trivially true. Now let $x \in \mathrm{Dom}\, f$ be different form x_*, and let $x_t = x_* + t(x - x_*)$, As $t \in (0, 1)$, we have $x_t \in \mathrm{Dom}\, f$ and $f(x_t) \geq f(x_*) \Rightarrow [x - x_*]^T \nabla f(x_*) = \lim_{t \to +0} (f(x_t) - f(x_*))/t \geq 0$, as claimed in $(*)$. Note that this reasoning does *not* use the convexity of f and utilizes solely the convexity of $\mathrm{Dom}\, f$, the differentiability of f at x_*, and the fact that x_* is a local minimizer of f.

♣ **Equivalent reformulation:**

• **Radial and Normal cones,** Let $Q \subset \mathbf{R}^n$ and $\bar{x} \in Q$.

– The *radial cone* $T_Q(x)$ of Q at $\bar{x}$ is the cone $\mathrm{Cone}\,(Q - \{\bar{x}\})$ spanned by directions $x - \bar{x}$ with $x \in Q$

– The *normal cone* $N_Q(\bar{x})$ of Q at $\bar{x}$ is the negation of the cone dual to the radial cone $T_Q(\bar{x})$:

$$N_Q(\bar{x}) = \{y \in \mathbf{R}^n : y^T(\bar{x} - x) \geq 0 \, \forall x \in Q\}$$

Note: the radial cone not necessarily is closed; the normal cone *is* closed.

• Fact VI.4 can be reformulated as follows:

Fact VI.5 *Let $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ be a convex function and $x_* \in \mathrm{Dom}\, f$ be a point at which f is differentiable. x_* is a minimizer of f iff*

$$\nabla f(x_*) \in -N_{\mathrm{Dom} f}(x_*). \qquad (!)$$

Let us look what (!) says in several standard situations.

Example I: $x_* \in \mathrm{int}\,\mathrm{Dom}\, f$**.** Here $T_{\mathrm{Dom} f}(x_*) = \mathbf{R}^n \Rightarrow N_{\mathrm{Dom} x}(x_*) = \{0\}$, and $(*)$ becomes the *Fermat equation*

$$\nabla f(x_*) = 0$$

– the *necessary and sufficient* condition for an interior point of the domain of a convex function f to be a minimizer of the function.

Example II: $x_* \in \mathrm{rint}\, Q$**.** Let L be the linear subspace parallel to $\mathrm{Aff}(\mathrm{Dom}\, f)$, so that $L = \mathrm{Lin}(\mathrm{Dom}\, f - \{x_*\}) = T_{\mathrm{Dom}\, f}(x_*)$, whence $N_{\mathrm{Dom} f}(x_*) = L^\perp$. Thus, Fact VI.5 states that *the necessary and sufficient condition for a point* $x_* \in \mathrm{rint}\, \mathrm{Dom}\, f$ *where a convex function* f *is differentiable to me a minimizer of the function is*

$$\nabla f(x) \in [\mathrm{Aff}(\mathrm{Dom}\, f) - \{x_*\}]^\perp.$$

Equivalently: Let $\mathrm{Aff}(Q) = \{x : Ax = b\}$. Then $L = \{x : Ax = 0\}$, $L^\perp = \{y = A^T\lambda\}$, and the optimality condition becomes

$$\exists \lambda^* : \quad \begin{array}{c} \nabla\big|_{x=x_*}[f(x) + (\lambda^*)^T(Ax - b)] = 0 \\ \Updownarrow \\ f'(x_*) + \sum_i \lambda_i^* \nabla(a_i^T x - b_i) = 0 \end{array} \qquad [A = [a_1^T; ...; a_m^T]]$$

Example III: $\mathrm{Dom}\, f = \{x : Ax - b \leq 0\}$ **is polyhedral.** Here

$$T_{\mathrm{Dom} f}(x_*) = \left\{h : a_i^T h \leq 0 \ \forall i \in I(x_*) = \{i : a_i^T x_* - b_i = 0\}\right\}. \qquad [A = [a_1^T; ...; a_m^T]]$$

By Homogeneous Farkas Lemma,

$$\begin{array}{rcl} N_{\mathrm{Dom} f}(x_*) & \equiv & \{y : a_i^T h \leq 0,\, i \in I(x_*) \Rightarrow y^T h \leq 0\} \\ & = & \{y = \sum\limits_{i \in I(x_*)} \lambda_i a_i : \lambda_i \geq 0\} \end{array}$$

and the optimality condition becomes

$$\exists (\lambda_i^* \geq 0, i \in I(x_*)) : \nabla f(x_*) + \sum_{i \in I(x_*)} \lambda_i^* a_i = 0$$

or, which is the same:

$$\exists \lambda^* \geq 0 : \left\{ \begin{array}{rcll} \nabla f(x_*) + \sum\limits_{i=1}^m \lambda_i^* a_i & = & 0 & \text{[Karush-Kuhn-Tucker equation]} \\ \lambda_i^* (a_i^T x_* - b_i) & = & 0,\, i = 1, ..., m & \text{[complementary slackness]} \end{array} \right.$$

The point is that in the *convex* case these conditions are necessary *and sufficient* for x_* to be a minimizer of f.

Note: The "common denominator" of Examples II – III is as follows: *A point x_* where a convex function f with polyhedral domain is differentiable minimizes the function over the domain if and only if it minimizes over this domain the linearization $\overline{f}(x) = f(x_*) + [x - x_*]^T \nabla f(x_*)$, taken at x_*, of the function.*

Example: Let us solve the problem

$$\min_x \left\{ f(x) := c^T x + \sum_{i=1}^m x_i \ln x_i : x \geq 0, \sum_i x_i = 1 \right\}.$$

The objective is convex, the domain $\mathrm{Dom}\, f = \{x \geq 0, \sum_i x_i = 1\}$ is convex (and even polyhedral). *Assuming* that the minimum is achieved at a point $x_* \in \mathrm{rint}\, Q$, the optimality condition becomes

$$\begin{gathered}
\nabla \left[c^T x + \sum_i x_i \ln x_i + \lambda [\sum_i x_i - 1] \right] = 0 \\
\Updownarrow \\
\ln x_i = -c_i - \lambda - 1 \ \forall i \\
\Updownarrow \\
x_i = \exp\{-1-\lambda\} \exp\{-c_i\}
\end{gathered}$$

Since $\sum_i x_i$ should be 1, we arrive at

$$x_i = \frac{\exp\{-c_i\}}{\sum_j \exp\{-c_j\}}.$$

At this point, the optimality condition is satisfied, so that the point indeed is a minimizer.

Karush-Kuhn-Tucker conditions – an interpretation

♣ Informal interpretation of Karush-Kuhn-Tucker optimality conditions goes back to Joseph-Louis Lagrange (1736 – 1813) and is as follows.
Optimization problem

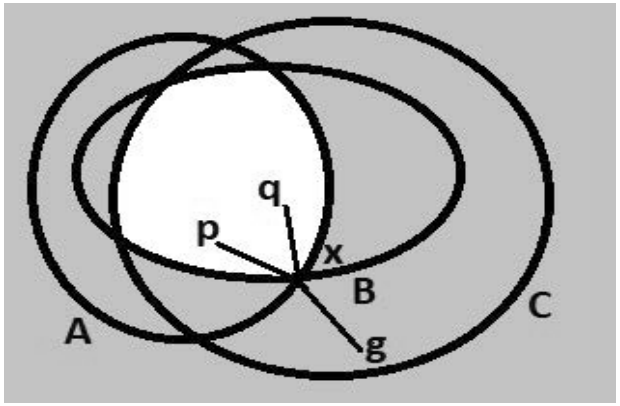
$$\min_{x\in\mathbf{R}^n}\{f(x):\ a_i(x)\le 0,\quad i=1,\dots,m\}$$

can be interpreted as locating the equilibrium position of a particle that is moving through $\mathbf{R}^n$ while being affected by an external force (like gravity) with potential f, meaning that

- When the position of the particle is $x\in\mathbf{R}^n$, the force acting at the particle is $-\nabla f(x)$.
- The domain in which the particle can actually travel is $Q:=\{x\in\mathbf{R}^n:\ a_i(x)\le 0,\ i\le m\}$; think about areas $a_i(x)>0$ as rigid obstacles that the particle cannot penetrate into.
- When the particle touches i-th obstacle (i.e., is in position x with $a_i(x)=0$), the obstacle produces a reaction force directed along the inward normal $-\nabla a_i(x)$ to the boundary of the obstacle, so that the reaction force is $-\lambda_i\nabla a_i(x)$; here $\lambda_i\ge 0$ depends on the pressure on the obstacle exerted by the particle.
- At an equilibrium x^* (which, by Physics, should minimize, at least locally, the potential f over Q), the total of the forces acting at the particle should be zero, that is, for properly selected $\lambda_i\ge 0$ one should have

$$-\nabla f(x^*)-\sum_{i:a_i(x^*)=0}\lambda_i\nabla a_i(x^*)=0,$$

which is exactly what is said by our Karush-Kuhn-Tucker (KKT) optimality condition as applied to the problem where the functions $a_i(x)=a_i^Tx-b_i$ are affine.



Physical illustration of KKT optimality onditions for optimization problem
$\min_{x\in\mathbf{R}^2}\{f(x) : a_i(x) \le 0,\, i = 1,2,3\}$.
White area represents the feasible domain Q, while ellipses **A**, **B**, **C** represent the sets $a_1(x) \le 0$, $a_2(x) \le 0$, $a_3(x) \le 0$. The point x is a candidate feasible solution located at the intersection $\{u \in \mathbf{R}^2 : a_1(u) = a_2(u) = 0\}$ of boundaries of **A** and **B**. $g = -\nabla f(x)$ is external force acting at particle located at x, p and q are reaction forces created by obstacles **A** and **B**. The condition for x to be an equilibrium reduces to $g + p + q = 0$, as on the picture. Equilibrium condition $g + p + q = 0$ translates to the KKT equation
$$\nabla f(x) + \lambda_1 a_1(x) + \lambda_2 \nabla a_2(x) = 0$$
holding for some nonnegative λ_1, λ_2.

Maxima of convex functions

Fact VI.6 *Let f be a convex function. Then*

A. *If f attains its maximum over $\mathrm{Dom}\, f$ at a point $x^* \in \mathrm{rint}\,\mathrm{Dom}\, f$, then f is constant on $\mathrm{Dom}\, f$*

Indeed, assuming that $f(x) < f(x^*)$ for some $x \in \mathrm{Dom}\, f$, $y = x^* + \alpha[x^* - x] \in \mathrm{Dom}\, f$ for small $\alpha > 0 \Rightarrow x^*$ is in the relative interior of segment $[x, y] \subset \mathrm{Dom}\, f \Rightarrow f(x^*) \leq \lambda \underbrace{f(x)}_{<f(x^*)} + (1-\lambda) f(y)$ for some $\lambda \in (0, 1) \Rightarrow f(y) > f(x^*)$

– contradiction!

B. *If $\mathrm{Dom}\, f$ is closed and does not contain lines and f attains its maximum on $\mathrm{Dom}\, f$, then among the maximizers there is an extreme point of $\mathrm{Dom}\, f$*

C. *If $\mathrm{Dom}\, f$ is polyhedral and f is bounded from above on $\mathrm{Dom}\, f$, then f attains its maximum on $\mathrm{Dom}\, f$.*

- **Good news:** Maximizing convex function f over a bounded polyhedral set $X \neq \emptyset$ reduces to computing the function at finitely many extreme points of the set. For example, problem $\max_x \{f(x) : \|x\|_1 \leq 1\}$ is easy
- **Bad news:** For a bounded polyhedral X, the number of extreme points usually is astronomically large, as is the case for the box $X = \{x : \|x\|_\infty \leq 1\}$, making maximizing over extreme points by looking at them one by one intractable. *In general, maximizing convex function is a computationally intractable task.*

Subgradients of convex functions

♣ Let $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ be a convex function and $\bar{x} \in \operatorname{int} \operatorname{Dom} f$. If f differentiable at $\bar{x}$, then, by Gradient Inequality, there exists an affine function, specifically,

$$h(x) = f(\bar{x}) + [\nabla f(\bar{x})]^T (x - \bar{x}),$$

which underestimates f everywhere and coincides with f at $\bar{x}$:

$$f(x) \geq h(x) \forall x \ \& \ f(\bar{x}) = h(\bar{x}) \qquad (*)$$

Affine function with property $(*)$ may exist also in the case when f is *not* differentiable at $\bar{x} \in \operatorname{Dom} f$. $(*)$ implies that

$$h(x) = f(\bar{x}) + g^T (x - \bar{x}) \qquad (**)$$

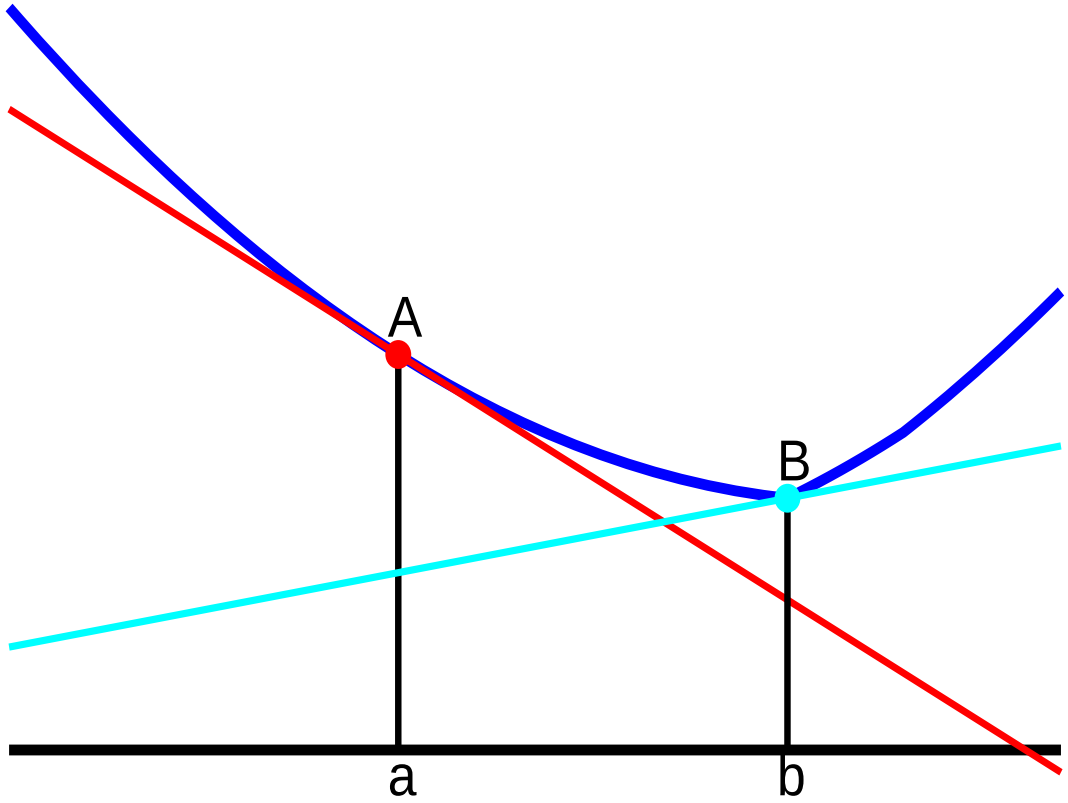
for certain g. Function $(**)$ indeed satisfies $(*)$ if and only if g is such that

$$f(x) \geq f(\bar{x}) + g^T (x - \bar{x}) \quad \forall x \qquad (!)$$

Definition. Let f be a convex function and $\bar{x} \in \operatorname{Dom} f$. Every vector g satisfying

$$f(x) \geq f(\bar{x}) + g^T (x - \bar{x}) \quad \forall x \qquad (!)$$

is called a *subgradient* of f at $\bar{x}$. The set of all subgradients, if any, of f at $\bar{x}$ is called *subdifferential* $\partial f(\bar{x})$ of f at $\bar{x}$.



Geometrically: A hyperplane supporting the epigraph Epi$\{f\}$ of f at a point $(\bar{x}, f(\bar{x}))$ is, at least for $\bar{x} \in \operatorname{int} \operatorname{Dom} f$, the graph of an affine function $h(x) = f(\bar{x}) + g^T(x - \bar{x})$ which underestimates f everywhere and is equal to f at the point $x = \bar{x}$.
The slope g of this affine function is a subgradient of f at x.

Definition. Let f be a convex function and $\bar{x} \in \mathrm{Dom}\, f$. Every vector g satisfying

$$f(x) \geq f(\bar{x}) + (x - \bar{x})^T g \quad \forall x \qquad (!)$$

is called a *subgradient* of f at $\bar{x}$. The set of all subgradients, if any, of f at $\bar{x}$ is called *subdifferential* $\partial f(\bar{x})$ of f at $\bar{x}$.

Example I: By Gradient Inequality, *if convex function f is differentiable at $\bar{x}$, then $\nabla f(\bar{x}) \in \partial f(\bar{x})$. Moreover, if $x \in \mathrm{int}\,\mathrm{Dom}\, f$, then $\nabla f(x)$ is the only element of $\partial f(x)$.*
To verify that $\partial f(x) = \{\nabla f(x)\}$ when $x \in \mathrm{int}\,\mathrm{Dom}\, f$, let $g \in \partial f(x)$. Then for every $h \in \mathbf{R}^n$ and for all small $t > 0$ we have $x + th \in \mathrm{Dom}\, f$, whence

$$\frac{f(x+th) - f(x)}{t} \geq \frac{[f(x) + tg^T h] - f(x)}{t} = g^T h.$$

Passing to limit as $t \to +0$, we get $[\nabla f(x) - g]^T h \geq 0$ for all $h \in \mathbf{R}^n$, that is, $g = \nabla f(x)$, Q.E.D.

Example II: Let $f(x) = |x| : \mathbf{R} \to \mathbf{R}$. When $\bar{x} \neq 0$, f is differentiable at $\bar{x}$, whence $\partial f(\bar{x}) = f'(\bar{x})$. When $\bar{x} = 0$, subgradients g are given by

$$|x| \geq 0 + gx = gx \ \forall x,$$

that is, $\partial f(0) = [-1, 1]$.
Note: In the case in question, f has directional derivative

$$Df(x)[h] = \lim_{t \to +0} \frac{f(x+th) - f(x)}{t}$$

at every point $x \in \mathbf{R}$ along every direction $h \in \mathbf{R}$, and this derivative is nothing but

$$Df(x)[h] = \max_{g \in \partial f(x)} g^T h$$

Example III: Consider feasible LP problem

$$\mathrm{Opt}(c) = \max_x \left\{ c^T x : Ax \leq b \right\} \qquad (P[c])$$

and assume that $\mathrm{Opt}(\bar{c}) < \infty$, so that $(P[\bar{c}])$ is solvable; let $\bar{x}_*$ is an optimal solution to $P(\bar{x})$.
As we have seen, $\mathrm{Opt}(c)$ is a convex function. We have

$$\mathrm{Opt}(c) \geq c^T \bar{x} = \bar{c}^T \bar{x} + \bar{x}^T [c - \bar{c}] = \mathrm{Opt}(\bar{c}) + \bar{x}^T [c - \bar{c}]$$

$\Rightarrow \bar{x} \in \partial \mathrm{Opt}(\bar{c})$.

♠ The most important fact about subgradients of convex functions f is that *the subgradient does exist at least at any point* $x \in \mathrm{rint}\,\mathrm{Dom}\, f$.

Fact VI.7 *Let* $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ *be convex,* $\mathrm{Dom}\, f$ *be nonempty, and let* $\mathcal{L} = \mathrm{Aff}(\mathrm{Dom}\, f) - \mathrm{Aff}(\mathrm{Dom}\, f)$ *be the linear subspace parallel to* $\mathrm{Aff}(\mathrm{Dom}\, f)$. *Then*

A. *For every* $x \in \mathrm{Dom}\, f$, *the subdifferential* $\partial f(x)$ *is a closed convex set*

B. *If* $x \in \mathrm{rint}\,\mathrm{Dom}\, f$, *then* $\partial f(x)$ *is nonempty*

C. *The multivalued mapping* $f \mapsto \partial f(x)$ *is locally bounded and closed on* $\mathrm{int}\,\mathrm{Dom}\, f$, *meaning that if* $Q \subset \mathrm{int}\,\mathrm{Dom}\, f$ *is a compact set, then for some* $L < \infty$ *it holds*

$$\forall (x \in Q, g \in \partial f(x)) : \|g\|_2 \le L,$$

and if a sequence $\{x^i \in Q, g^i \in \partial f(x^i)\}_i$ *converges, as* $i \to \infty$, *to* (x, g), *then* $g \in \partial f(x)$.

D. *Assume that* $\bar{x} \in \mathrm{Dom}\, f$ *is represented as* $\lim\limits_{i\to\infty} x^i$ *with* $x^i \in \mathrm{Dom}\, f$ *and that* $f(\bar{x}) \le \liminf\limits_{i\to\infty} f(x^i)$. *If a sequence* $g^i \in \partial f(x^i)$ *converges to certain vector* g, *then* $g \in \partial f(\bar{x})$.

E. *Being a subgradient is a local property: if* $x \in \mathrm{rint}\,\mathrm{Dom}\, f$ *and* g *is such that for certain* $r > 0$ *one has*

$$f(x) + g^T(y - x) \le f(y)\ \forall (y \in \mathrm{Dom}\, f, \|y - x\| \le r),$$

then $g \in \partial f(x)$.

F. *If* $x \in \mathrm{rint}\,\mathrm{Dom}\, f$, *then, for every* $h \in \mathcal{L}$, *there exists the directional derivative taken at* x *along direction* h *– the quantity*

$$Df(x)[h] \equiv \lim_{t\to+0} \frac{f(x+th) - f(x)}{t}$$

$Df(x)[h]$ *is a convex positively homogeneous, of degree 1, function of* $h \in \mathcal{L}$ *such that*

$$\begin{array}{rcll} Df(x)[h] & = & \max\limits_{g\in\partial f(x)} g^T h & (a) \\ f(x) + Df(x)[h] & \le & f(x+h) & (b) \\ \partial Df(x)[0] & = & \partial f(x) & (c) \end{array}$$

Proofs

A. For every $x \in \mathrm{Dom}\, f$, the subdifferential $\partial f(x)$ is a closed convex set
Indeed, $\partial f(x) = \{g : f(y) \geq f(x) + g^T(y-x)\, \forall y \in \mathrm{Dom}\, f\}$ is the solution set of a (infinite) system of nonstrict linear inequalities.

B. If $\bar{x} \in \mathrm{rint}\, \mathrm{Dom}\, f$, then $\partial f(\bar{x})$ is nonempty
W.l.o.g. let $\mathrm{Dom}\, f$ be full-dimensional, so that $\bar{x} \in \mathrm{int}\, \mathrm{Dom}\, f$. Consider the convex set

$$T = \mathrm{Epi}\{f\} = \{[x;t] : t \geq f(x)\}.$$

Since f is convex, it is continuous on $\mathrm{int}\, \mathrm{Dom}\, f$, whence T has a nonempty interior. The point $\bar{y} := [\bar{x}; f(\bar{x}]$ clearly does not belong to this interior (as $\bar{y} - [0;\epsilon] \notin \mathrm{Epi}\{f\}$ for all $\epsilon > 0$) whence $S = \{(\bar{x}, f(\bar{x}))\}$ can be separated from T: there exists $(\alpha, \beta) \neq 0$ such that

$$\alpha^T \bar{x} + \beta f(\bar{x}) \leq \alpha^T x + \beta t \quad \forall (x, t \geq f(x)) \qquad (*)$$

Clearly $\beta \geq 0$ (otherwise $(*)$ will be impossible when $x = \bar{x}$ and $t > f(\bar{x})$ is large).
Claim: $\beta > 0$. Indeed, with $\beta = 0$, $(*)$ implies that

$$\alpha^T \bar{x} \leq \alpha^T x \; \forall x \in \mathrm{Dom}\, f \qquad (**)$$

Since $(\alpha, \beta) \neq 0$ and $\beta = 0$, we have $\alpha \neq 0$; but then $(**)$ contradicts $\bar{x} \in \mathrm{int}\, \mathrm{Dom}\, f$.
- Since $\beta > 0$, $(*)$ implies that if $g = -\beta^{-1}\alpha$, then

$$-g^T \bar{x} + f(\bar{x}) \leq -g^T x + f(x) \; \forall x \in \mathrm{Dom}\, f,$$

that is,

$$f(x) \geq f(\bar{x}) + (x - \bar{x})^T g \; \forall x.$$

C. The multivalued mapping $f \mapsto \partial f(x)$ is locally bounded and closed on $\operatorname{int}\operatorname{Dom} f$, meaning that if $Q \subset \operatorname{int}\operatorname{Dom} f$ is a compact set, then for some $L < \infty$ it holds

$$\forall (x \in Q, g \in \partial f(x)) : \|g\|_2 \leq L,$$

and if a sequence $\{x^i \in Q, g^i \in \partial f(x^i)\}_i$ converges, as $i \to \infty$, to (x, g), then $g \in \partial f(x)$.
Given a compact set $Q \subset \operatorname{int}\operatorname{Dom} f$, we can find $\epsilon > 0$ such that the sets

$$Q_\epsilon = \{x \in \mathbf{R}^n : \operatorname{dist}(x, Q) := \min_{y \in Q} \|x - y\|_2 \leq \epsilon\}$$

(which is compact along with Q) is contained in $\operatorname{int}\operatorname{Dom} f$ (why?). As we know, f is Lipschitz continuous, with some constant L, on Q_ϵ. It follows that

$$\forall (x \in Q, g \in \partial f(x), h \in \mathbf{R}^n, \|h\|_2 \leq r) : g^T h \leq f(x + h) - f(x) \leq L\|h\|,$$

whence $\|g\|_2 \leq L$. Thus, $\partial g(x) \in \{g : \|g\|_2 \leq L\|$ for all $x \in Q$. It remains to verify that if $x^i \in Q$, $g_i \in \partial f(x^i)$ and $x^i \to x$ and $g^i \to g$ as $i \to \infty$, then $g \in \partial f(x)P$, indeed, for every y and every i we have

$$f(y) \geq f(x^i) + [g^i]^T (y - x^i). \qquad (!)$$

Since $x = \lim_i x^i \in Q$ (as Q is compact and thus is closed) and $Q \subset \operatorname{int}\operatorname{Dom} f$, f is continuous (and even Lipschitz continuous) on Q, we have $f(x^i) \to f(x)$ as $i \to \infty$, and passing to limits as $i \to \infty$ in (!), we get

$$f(y) \geq f(x) + +g^R (y - x).$$

This inequality holds true for every y, implying that $g \in \partial f(x)$, Q.E.D.
D. Assume that $\bar{x} \in \operatorname{Dom} f$ is represented as $\lim\limits_{i \to \infty} x^i$ with $x^i \in \operatorname{Dom} f$ and that $f(\bar{x}) \leq \liminf\limits_{i \to \infty} f(x^i)$ If a sequence $g^i \in \partial f(x^i)$ converges to certain vector g, then $g \in \partial f(\bar{x})$.
For every $y \in \mathbf{R}^n$, we still have at our disposal (!) and thus the relation

$$f(y) \geq \liminf_{i \to \infty} f(x^i) + g^T (y - x)$$

which combines with the premise of **D** to imply that $f(y) \geq f(x) + g^T (y - x)$. Since y is arbitrary, we get $g \in \partial f(x)$, Q.E.D.

E. Being a subgradient is a local property: if $x \in \mathrm{rint}\,\mathrm{Dom}\, f$ and g is such that for certain $r > 0$ one has

$$f(x) + g^T(y - x) \leq f(y)\ \forall (y \in \mathrm{Dom}\, f, \|y - x\| \leq r),$$

then $g \in \partial f(x)$.

We want to prove that $f(y) \geq f(x) + g^T(y - x)$ for any y. There is nothing to prove when $f(y) = \infty$, same as when $y = x$, Assuming that $x \neq y$ and $y \in \mathrm{Dom}\, f$, we can find a point $z \in (x, y)$ such that $\|z - x\|_2 \leq r$. Setting $e(w) = f(w) - g^T(w - x)$, we get a convex function such that $e(z) - e(x) \geq 0$ due to the origin of g, while by convexity of e it holds

$$\frac{e(y) - e(x)}{\|y - x\|_2} \geq \frac{e(z) - e(x)}{\|z - x\|_2} \geq 0$$

that is, $e(y) \geq e(x)$, which is nothing but the desired inequality $f(y) - g^T(y - x) \geq f(x)$.

F. If $x \in \mathrm{rint}\,\mathrm{Dom}\, f$, then, for every $h \in \mathcal{L}$, there exists the directional derivative taken at x along direction h – the quantity

$$Df(x)[h] \equiv \lim_{t \to +0} \frac{f(x+th) - f(x)}{t}$$

$Df(x)[h]$ is a convex positively homogeneous, of degree 1, function of $h \in \mathcal{L}$ such that

$$\begin{array}{rcll} Df(x)[h] & = & \max\limits_{g \in \partial f(x)} g^T h & (a) \\ f(x) + Df(x)[h] & \leq & f(x+h) & (b) \\ \partial Df(x)[0] & = & \partial f(x) & (c) \end{array}$$

It is immediately seen that it suffices to prove **F** in the case when $\mathrm{Dom}\, f$ is full-dimensional, that is, when $\mathcal{L} = \mathbf{R}^n$, which we assume from now on.
Given $x \in \mathrm{int}\,\mathrm{Dom}\, f$ and $h \in \mathbf{R}^n$, we can find $r > 0$ such that $x \pm rh \in \mathrm{Dom}\, f$, implying by convexity that

$$\frac{f(x) - f(x-rh)}{r} \leq \frac{f(x+th) - f(x)}{t},\ 0 < t < r, \qquad (!!)$$

and that the right hand side ratio is a nonincreasing function of $t \in (0, r]$. As this ratio is below bounded by the left hand side of (!!), we conclude that the limit

$$D(x)[h] = \lim_{t \to +0} \frac{f(x+th) - f(x)}{t}$$

does exist. Next, setting $f_t(h) = \frac{f(x+th)-f(x)}{t}$, observe that for every $R > 0$ and all small enough $t > 0$, the function $f_t(h)$ is convex on the ball $\{h : \|h\|_2 \leq R\}$, implying that the pointwise limit, as $t \to +0$, $Df(x)[\cdot]$ of the functions $f_t(\cdot)$ is convex. positive homogeneity, of degree 1, of $Df(x)[\cdot]$ is evident.
To prove (a), note that if $g \in \partial f(x)$, then for all $t > 0$ and all h one has $\frac{f(x+th)-f(x)}{t} \geq g^T h$; passing to limits as $t \to +0$, we get $Df(x)[h] \geq g^T h$ for all h and all $g \in \partial f(x)$, whence

$$Df(x)[h] \geq \max_{g \in \partial f(x)} g^T h. \qquad (a.1)$$

If $x \in \operatorname{rint}\operatorname{Dom} f$, then, for every $h \in \mathcal{L}$, there exists the directional derivative taken at x along direction h – the quantity

$$Df(x)[h] \equiv \lim_{t \to +0} \frac{f(x+th) - f(x)}{t}$$

$Df(x)[h]$ is a convex positively homogeneous, of degree 1, function of $h \in \mathcal{L}$ such that

$$\begin{array}{rcll} Df(x)[h] & = & \max\limits_{g \in \partial f(x)} g^T h & (a) \\ f(x) + Df(x)[h] & \leq & f(x+h) & (b) \\ \partial Df(x)[0] & = & \partial f(x) & (c) \end{array}$$

♠ We have proved that

$$Df(x)[h] \geq \max_{g \in \partial f(x)} g^T h. \qquad (a.1)$$

To prove the opposite inequality, let us fix h and select a small $\tau > 0$ such that $x + \tau h \in \operatorname{int}\operatorname{Dom} f$, and select $g_\tau \in \partial f(x + \tau h)$ (this is possible by **B**), By convexity of f, for $0 < t < \tau$ we have

$$\frac{f(x+th) - f(x)}{t} \leq \frac{f(x+\tau h) - f(x)}{\tau} \leq f(x+\tau h) - \frac{f(x+\tau h) + g_\tau^T(x - [x+\tau h])}{\tau} = g_\tau^T h.$$

whence, passing to limit as $t \to +0$, we get

$$Df(x)[h] \leq g_\tau^T h. \qquad (a.2)$$

as $\tau \to +0$, $\|g_\tau\|_2$ remains bounded by **C**, thus,, we can find a sequence $\tau_1 > \tau_2 > ...$ of positive reals converging to 0 as $i \to \infty$ and such that $g^i = g_{\tau_i}$ converge, as $i \to \infty$, to some g. Noting that $x^i := x + \tau_i h \to x$ as $i \to \infty$ and invoking the same **C**, we conclude that $g \in \partial f(x)$. At the same time, from $(a.2)$ it follows that $g^T h \geq Df(x)[h]$, which combines with $(a.1)$ to imply (a).

If $x \in \operatorname{rint} \operatorname{Dom} f$, then, for every $h \in \mathcal{L}$, there exists the directional derivative taken at x along direction h – the quantity

$$Df(x)[h] \equiv \lim_{t \to +0} \frac{f(x+th) - f(x)}{t}$$

$Df(x)[h]$ is a convex positively homogeneous, of degree 1, function of $h \in \mathcal{L}$ such that

$$\begin{array}{rcll} Df(x)[h] & = & \max\limits_{g \in \partial f(x)} g^T h & (a) \\ f(x) + Df(x)[h] & \leq & f(x+h) & (b) \\ \partial Df(x)[0] & = & \partial f(x) & (c) \end{array}$$

To prove (b), it suffices to consider the case when $x + h \in \operatorname{Dom} f$, By convexity of f, for $0 < t < 1$ it holds $\frac{f(x+th)-f(x)}{t} \leq \frac{f(x+h)-f(x)}{1}$; passing to limit as $t \to +0$, we arrive at (b).

To prove (c), note that when $\operatorname{int} \operatorname{Dom} f$ contains the centered at the origin ball of radius $r > 0$ and $g \in \partial Df(x)[0]$, for h with $\|h\|_2 =\leq r$ we have $f(x) + g^T h \leq f(x) + Df(x)[h] \leq f(x+h)$, with the last inequality given by (b); the resulting inequality, in view of **E**, implies that $g \in \partial f(x)$, that is $\partial Df(x)[0] \subset \partial f(x)$. To verify the opposite inclusion, note that by (a) for $g \in \partial f(x)$ and every $h \in \mathbf{R}^n$ it holds $Df(x)[h] \geq g^T h$, that is, $g \in \partial Df(x)[0]$. Q.E.D.

Elementary Calculus of Subgradients

♣ **Disclaimer:** All functions participating in the rules below are assumed to be convex.

- If $g_i \in \partial f_i(x)$ and $\lambda_i \geq 0$, then

$$\sum_i \lambda_i g_i \in \partial(\sum_i \lambda_i f_i)(x)$$

- If $g_\alpha \in \partial f_\alpha(x)$, $\alpha \in \mathcal{A}$,

$$f(\cdot) = \sup_{\alpha \in \mathcal{A}} f_\alpha(\cdot)$$

and

$$f(x) = f_\alpha(x), \ \alpha \in \mathcal{A}_*(x) \neq \emptyset,$$

then every convex combination of vectors g_α, $\alpha \in \mathcal{A}_*(x)$, is a subgradient of f at x

- [Chain rule] If $g_i \in \partial f_i(x)$, $i = 1, ..., m$, $F(y_1, ..., y_m)$ is monotone w.r.t. every y_i such that f_i is not affine, and $d \in \partial F(f_1(x), ..., f_m(x))$, then the vector

$$g := \sum_i d_i g_i$$

is a subgradient of the composition

$$G(\xi) = \begin{cases} F(f_1(\xi), ..., f_m(\xi)) & , \xi \in \cap_i \text{Dom } f_i \\ +\infty & , \text{otherwise} \end{cases}$$

at $\xi = x$.

Chain rule: If $g_i \in \partial f_i(x)$, $i = 1, ..., m$, $F(y_1, ..., y_m)$ is monotone w.r.t. every y_i such that f_i is not affine, and $d \in \partial F(f_1(x), ..., f_m(x))$, then the vector

$$g := \sum_i d_i g_i$$

is a subgradient of the composition

$$G(\xi) = \begin{cases} F(f_1(\xi), ..., f_m(\xi)) & , \xi \in \cap_i \mathrm{Dom}\, f_i \\ +\infty & , \text{otherwise} \end{cases}$$

at $\xi = x$.

Under the premise of Chain rule, $x \in \mathrm{Dom}\, f_i$, $i \le m$, and $y := [f_1(x), ..., f_m(x)] \in \mathrm{Dom}\, F$. Let $h \in \mathbf{R}^n$; we need to prove that

$$G(x + h) \ge G(x) + g^T h \qquad (*)$$

There is noting to prove when $x + h \notin \mathrm{Dom}\, G$. Assuming that $x + h \in \mathrm{Dom}\, G$, let I be the set of those $i \le m$ for which f_i is *not* affine, and let us set

$$y_i = f_i(x),\ z_i = f_i(x + h),\ \bar{y}_i = y_i + g_i^T h,\ i \le m,\ \bar{h} = [g_1^T h; , , , ; g_m^T h].$$

As $g_i \in \partial f_i(x)$, we have

$$\bar{y}_i \left\{ \begin{array}{ll} \le z_i & , i \in I \\ = z_i & , i \notin I \end{array} \right\}$$

which combines with partial monotonicity of F to imply the first inequality in the following chain:

$$\begin{array}{l} G(x + h) = F(z_1, ..., z_m) \ge F(\bar{y}_1, ..., \bar{y}_m) = F([y_1; ...; y_m] + \bar{h}) \\ \ge F(y_1, ..., y_m) + d^T \bar{h} \ [\text{as } d \in \partial F(y_1, ..., y_m)] \\ = G(x) + \sum_i d_i \bar{h}_i = G(x) + \sum_i d_i g_i^T h \\ = G(x) + g^T h, \end{array}$$

whence $g \in \partial G(x)$, Q.E.D.

Hahn-Banach Theorem

♣ Here is the [finite-dimension version of] one of the cornerstones of Functional Analysis

Fact VI.8 [Hahn-Banach Theorem] *Let $\mathfrak{n}(x) : \mathbf{R}^n \to \mathbf{R}$ be a convex positively homogeneous, of degree 1, function and $E \subset \mathbf{R}^n$ be a linear subspace. Let also $e^T x$ be a linear function which is dominated by $\mathfrak{n}(x)$ on E:*

$$e^T x \leq \mathfrak{n}(x) \; \forall x \in E.$$

Then there exists a linear function $\overline{e}^T x : \mathbf{R}^n \to \mathbf{R}$ which coincides with $e^T x$ on E and is dominated by $\mathfrak{n}(x)$ everywhere.

Canonical wording: *A linear form dominated by $\mathfrak{n}(\cdot)$ on a linear subspace can be extended from this subspace onto the entire $\mathbf{R}^n$, the domination being preserved.*

Note: We clearly have $\mathfrak{n}[\cdot] = D\mathfrak{n}(0)[\cdot]$; with this in mind, the relation

$$D\mathfrak{n}(0)[h] = \max_{g \in \partial \mathfrak{n}(0)} g^T h$$

becomes a special case of HBT – the one when $E = \mathbf{R} \cdot h$. Taking into account that for convex $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ and $x \in \operatorname{int} \operatorname{Dom}$ one has $\partial D f(x)[\cdot] = \partial f(x)$, HBT implies that whenever $x \in \operatorname{int} \operatorname{Dom} F$ and E is a linear subspace, the linear form $e^T h$ satisfying

$$f(x+h) \geq f(x) + e^T h \; \forall h \in E$$

– the subgradient of the restriction of f onto $x + E$ taken at x – can be obtained from a "full-dimensional" subgradient of f: there exists $g \in \partial f(x)$ such that

$$f^T h = e^T h \; \forall h \in E.$$

Proof of HBT: It suffices to prove the HBT for the case when E is of dimension $n-1$ – given this fact, we can extend linear form from E onto the entire $\mathbf{R}^n$ step by step, adding one dimension at a time. Thus let $\dim E = n-1$, and let $a \in \mathbf{R}^n \backslash E$. An extension $f^T x$ of the linear form $e^T x$ from E onto the entire $\mathbf{R}^n$ is fully determined by the quantity $\alpha = f^T a$: every $x \in \mathbf{R}^n$ can be uniquely decomposed as $x = ta + d$ with $t \in \mathbf{R}$ and $d \in E \Rightarrow f^T x = \alpha t + e^T d$. To get an extension dominated by $\mathfrak{n}(\cdot)$, we need to select α in such a way that

$$\forall (d \in E, t \in \mathbf{R}) : \alpha t + e^T d \leq \mathfrak{n}(ta + d) \qquad (*)$$

When $t = 0$, the conclusion in $(*)$ holds true automatically, as $e^T x$ is dominated by $\mathfrak{n}(x)$ on E. By homogeneity, all we need to ensure a is to select α in such a way that

$$\alpha + e^T d \leq \mathfrak{n}(d + a) \ \forall d \in E \ \& \ \underbrace{-\alpha + e^T r \leq \mathfrak{n}(r - a)}_{\Leftrightarrow \alpha \geq e^T r - \mathfrak{n}(r-a)} \ \forall r \in E.$$

An evident necessary and sufficient condition for the existence of the required α is

$$\mathfrak{n}(d + a) - e^T d \geq -\mathfrak{n}(r - a) + e^T r \ \forall r, d \in E,$$

that is,

$$n(d + a) + n(r - a) \geq e^T [d + r] \ \forall d, r \in E.$$

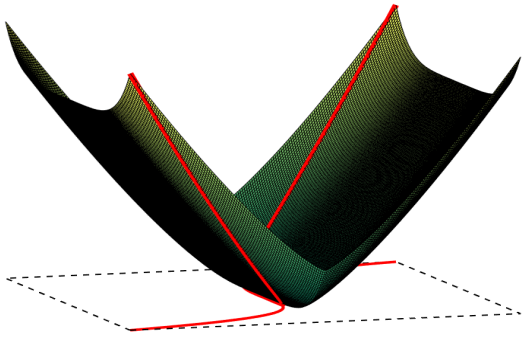
The latter inequality indeed takes place, since by convexity and positive homogeneity of $\mathfrak{n}$ one has $n(d+a) + \mathfrak{n}(r-a) \geq \mathfrak{n}(d+r)$, and $\mathfrak{n}(d+r) \geq e^T[d+r]$ for all $d, r \in E$, Q.E.D.

Note: This proof straightforwardly combines with *transfinite induction* to imply the "true" infinite-dimensional, HBT.

Lecture II.3

Legendre Transform and Fenchel Duality

Back to metric spaces: lower semicontinity
Legendre transform
Hölder and Moment Inequalities
Support and Minkowski functions of convex sets



$z = \frac{1}{5}|x|^5 + \frac{4}{5}|y|^{5/4}$

Preliminaries: Extended real line as a metric space

♣ Convex functions take values in the extended real line $\mathbf{R} \cup \{+\infty\}$, concave ones – in $\mathbf{R} \cup \{-\infty\}$, and for our further developments it makes sense to equip the extended real line $\overline{R} = \mathbf{R} \cup \{+\infty\} \cup \{-\infty\}$ with metric (and thus with convergence).
A good way to do it is
• to select a *strictly increasing continuous encoding function* $\theta : \mathbf{R} \to \mathbf{R}$ such that $\lim_{t\to+\infty} \theta(t) = 1$, $\lim_{t\to-\infty} \theta(t) = -1$ e.g., $\theta(t) = \frac{2}{\pi}\mathrm{atan}(t)$,
• to extend $\theta(\cdot)$ from $\mathbf{R}$ onto $\overline{\mathbf{R}}$ by setting $\theta(\pm\infty) = \pm 1$, and
• to use the resulting one-to-one correspondence $t \mapsto \theta(t) : \overline{\mathbf{R}} \to [-1, 1]$ to "translate" the standard metric on $[-1, 1]$ into the metric

$$d(t, t') = |\theta(t) - \theta(t')|$$

on $\overline{\mathbf{R}}$.
Note: Equipped with this metric, $\overline{\mathbf{R}}$ becomes a *compact* metric space, as $[-1, 1]$ is so.
♠ As is immediately seen, the resulting convergence of sequences of points from $\overline{\mathbf{R}}$ is as follows:
A sequence $\{t_i \in \overline{\mathbf{R}}\}_i$ converges to $\bar{t} \in \overline{\mathbf{R}}$ iff
— $\bar{t} \in \mathbf{R}$ and for every reals $\underline{a} < \bar{t}$, $\overline{a} > \bar{t}$, all but finitely many terms in the sequence satisfy $\underline{a} < t_i < \overline{a}$,
— $\bar{t} = +\infty$ and for every real $\underline{a}$, all but finitely many terms in the sequence satisfy $\underline{a} < t_i$,
— $\bar{t} = -\infty$ and for every real $\overline{a}$, all but finitely many terms in the sequence satisfy $t_i < \overline{a}$
Note: The resulting notion of convergence (this is the only notion we will be interested in) does not depend on the choice of the encoding function and is in full accordance with how the sentence "$t_i \to \bar{t} \in \overline{\mathbf{R}}$ as $i \to \infty$" is understood in Calculus.
Note: From now on, we treat $\overline{\mathbf{R}}$ as a metric space, and can therefore speak about continuity of functions defined on metric spaces and taking values in $\overline{\mathbf{R}}$.

lim inf and lim sup

♠ *Limiting points* of a sequence $\{x^i \in X\}_i$ of points in metric space (X, d) are, by definition, the limits of converging subsequences of the sequence. It is immediately seen that
The set of limiting points of a given sequence is closed (and nonempty when the space is compact).

As a result,
Among the limiting points of a sequence $\{x^i \in [-1, 1]\}$ *there are the minimal and the maximal ones.*

♠ The one-to-one correspondence, given by an encoding function, between $\overline{\mathbf{R}}$ and $[-1, 1]$ preserves convergence and order
⇒ *Among the (nonempty) set of limiting points of a sequence* $\{x^i \in \overline{\mathbf{R}}\}$ *there is the smallest one (the lower limit of the sequence* $\liminf_{i\to\infty} x^i$*), and the largest one (the upper limit of the sequence* $\limsup_{i\to\infty} x^i$*). The sequence converges iff*

$$\liminf_{i\to\infty} x^i = \limsup_{i\to\infty} x^i,$$

in which case the common value of lim inf *and* lim sup *is* $\lim_{i\to\infty} x^i$.

For example,
- $\liminf_{i\to\infty}(-1)^i/i = \limsup_{i\to\infty}(-1)^i/i = \lim_{i\to\infty}(-1)^i/i = 0$
- $\liminf_{i\to\infty}(-1)^i = -1$, $\limsup_{i\to\infty}(-1)^i = 1$,
- $\liminf_{i\to\infty}(-1)^i i = -\infty$, $\limsup_{i\to\infty}(-1)^i i = +\infty$.

Lower and upper semicontinuity

♣ Recall that a mapping $f : X \to Z$ (X, Z are metric spaces) is continuous iff the inverse image of every open subset of Z is an open subset of X, or, which is the same, the inverse image of every closed subset of Z is a closed subset in X.

♠ It is immediately seen that in the cases of $Z = \mathbf{R}$ and $Z = \overline{\mathbf{R}}$ these continuity criteria can be simplified as follows:

• *A function $f : X \to \mathbf{R}$ ($f : X \to \overline{\mathbf{R}}$) is continuous iff for every real a, the set $\{x \in X : f(x) > a\}$ is open (or, equivalently, the set $\{x : f(x) \le a\}$ is closed) and for every real a, the set $\{x \in X : f(x) < a\}$ is open (or, equivalently, the set $\{x : f(x) \ge a\}$ is closed).*

♠ Functions $f : X \to \mathbf{R}$ ($f : X \to \overline{\mathbf{R}}$) satisfying "halves" of the latter characterisation have names:

• A function $f : X \to \mathbf{R}$ ($f : X \to \overline{\mathbf{R}}$) is called *lower semicontinuous* (lsc), if for every real a the set $\{x \in X : f(x) > a\}$ is open (or, equivalently, the set $\{x : f(x) \le a\}$ is closed)

• A function $f : X \to \mathbf{R}$ ($f : X \to \overline{\mathbf{R}}$) is called *upper semicontinuous* (usc), if for every real a the set $\{x \in X : f(x) < a\}$ is open (or, equivalently, the set $\{x : f(x) \ge a\}$ is closed)

Example: The function $f_\alpha(x) = \begin{cases} 0 & , x \neq 0 \\ \alpha & , x = 0 \end{cases} : \mathbf{R} \to \mathbf{R}$ is

– lower semicontinuous, iff $\alpha \le 0$
– upper semicontinuous, iff $\alpha \ge 0$
– continuous, iff $\alpha = 0$

Lower semicontinuous functions with values in $\mathbf{R} \cup \{+\infty\}$

Fact VII.1 *Let $f : X \to \mathbf{R} \cup \{+\infty\}$ be a function. The following three properties of f are equivalent to each other:*
A. *f is lower semicontinuous*
B. *whenever $x^i \in X$ and $\overline{x} = \lim_i x^i$, one has $f(\overline{x}) \leq \liminf_i f(x_i)$*
Equivalently: *whenever $x^i \in X$ converge to $\overline{x}$ and $f(x^i) \to a \in \overline{\mathbf{R}}$ as $i \to \infty$, one has $f(\overline{x}) \leq a$*
C. *the epigraph $\mathrm{Epi}\{f\} = \{[x; t] : t \geq f(x)\} \subset X \times \mathbf{R}$ is closed.*

Proof:
A $\Rightarrow$ **B:** Assume that f is lsc, $x = \lim_i x^i$, and $a = \liminf_i f(x_i)$. Assume that $f(x) > a$, and let us lead this assumption to contradiction. As $f(x) > a$, we have $a \in \overline{\mathbf{R}} \backslash \{+\infty\}$, and there exists $b \in \mathbf{R} : a < b < f(\overline{x})$. There exists a subsequence $\{x^{i_j}\}_j$ of $\{x^i\}$ such that $f(x^{i_j}) \leq a$ as $j \to \infty \Rightarrow$ all but finitely many of the points x^{i_j} belong to $X_b = \{x : f(x) \leq b\}$. As f is lsc, X_b is closed, and as $x^{i_j} \in X_b$ for large j and $x^{i_j} \to \overline{X}$ as $j \to \infty$, we have $\overline{x} \in X_b$, that is, $f(\overline{x}) \leq b$, which is a contradiction.
B $\Rightarrow$ **C:** Assume that B is the case, and let $[x^i; t_i] \in \mathrm{Epi}\{f\}$ converge to $[\overline{x}; \overline{t}]$ as $i \to \infty$; we should prove that $[\overline{x}; \overline{t}] \in \mathrm{Epi}\{f\}$. Indeed, by **B** we have $f(\overline{x}) \leq \liminf_i f(x^i) \leq \liminf_i t_i = \overline{t}$.
C $\Rightarrow$ **A:** Assume that **C** takes place, and let $a \in \mathbf{R}$; we want to prove that the set $X_a = \{x : f(x) \leq a\}$ is closed. Indeed, assuming the opposite, there exists a converging sequence $\{x^i \in X_a\}$ such that $x^i \to \overline{x}$ as $i \to \infty$ and $f(\overline{x}) > a$. As $f(x^i) \leq a$, we have $[x^i; a] \in \mathrm{Epi}\{f\}$, and since $\mathrm{Epi}\{f\}$ is closed and $[x^i; a] \to \overline{x}; a]$ as $i \to \infty$, we have $[\overline{x}; a] \in \mathrm{Epi}\{f\}$, i.e., $f(\overline{x}) \leq a$, which is a contradiction.

Basic properties of lsc functions

Fact VII.2 *Let X be a metric space.*

I. *If $f_\alpha : X \to \mathbf{R} \cup \{+\infty\}$, $i = 1, ..., I$, are lsc and $\lambda_i \geq 0$, then*

$$f(x) = \sum_i \lambda_i f_i(x) : X \to \mathbf{R} \cup \{+\infty\}$$

is lsc.

This is given by characterization **B** of lsc functions due to the evident relations (check them!) $\liminf_i \lambda a_i = \lambda \liminf_i a_i$ whenever $\lambda \geq 0$, $a_i \in \mathbf{R} \cup \{+\infty\}$, and $\liminf_i [a_i + b_i] \leq \liminf_i a_i + \liminf b_i$ whenever $a_i, b_i \in \mathbf{R} \cup \{+\infty\}$ and $\liminf_i a_i > -\infty$, $\liminf_i b_i > -\infty$

2. *If function $f_\alpha : X \to \mathbf{R} \cup \{+\infty\}$, $\alpha \in \mathcal{A}$, are lsc, so is $f(x) = \sup_\alpha f_\alpha(x)$*

This is given by characterization **C** of lsc functions, as $\mathrm{Epi}\{f\} = \cap_\alpha \mathrm{Epi}\{f\}$

3. *If X, Y are metric spaces, $f : X \to Y$ is continuous, and $F : Y \to \mathbf{R} \cup \{+\infty\}$ is lsc, the composition $F(f(\cdot))$ is lsc*

This is given by characterization **B** of lsc functions

4. *Let X be a compact metric space and $f : X \to \mathbf{R} \cup \{+\infty\}$ be lsc. Then $\inf_{x \in X} f(x) = f(x_*)$ for some $x_* \in X$; in particular, $\inf_X f > -\infty$.*

Indeed, let $\underline{a} = \inf_X f \in \overline{\mathbf{R}}$. When $\underline{a} = +\infty$, the claim is clearly true. Now let $\underline{a} < \infty$. For every real $a > \underline{a}$, the set $X_a = \{x : f(x) \leq a\}$ is closed (as f is lsc) and nonempty (as $a > \inf_X$. Every finite collection of sets from the family $\{X_a : a > \underline{a}\}$ clearly has a nonempty intersection, implying, by compactness of X, that $X_* := \cap_{a > \underline{a}} X_a \neq \emptyset$ (Fact I.13), and clearly $f(x) = \underline{a}$ on X_*.

Closed convex functions

♣ Convex lsc functions are nothing but functions $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ with convex *and closed* epigraph Epi$\{f\}$. "You can get much farther with a kind word and a gun than you can with a kind word alone" – the convex lsc (a.k.a. *closed convex*) functions are much nicer than merely convex ones.

♠ As we know, for a convex $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ and $\bar{x} \in \mathrm{rint}\,\mathrm{Dom}\, f$ one has $\partial f(\bar{x}) \neq \emptyset$, implying that *there exists an affine function dominated by f everywhere and equal to f at $\bar{x}$*,, whence *the pointwise supremum $\overline{f}$ of all affine functions dominated by convex function f is equal to f on* rint Dom f.

In fact $\overline{f} = f = +\infty$ *outside of* cl Dom f *as well.*

Indeed, there is nothing to prove when $\mathrm{Dom}\, f = \emptyset$. When $\mathrm{Dom}\, f \neq \emptyset$ and $\bar{x} \notin \mathrm{cl}\,\mathrm{Dom}\, f$, there exists a linear form which strongly separates Dom f and $\bar{x}$, or, which is the same, there exists an affine function $a(\cdot)$ such that $a(\bar{x}) > 0$ and $a(x) \leq 0$ for $x \in \mathrm{Dom}\, f$. Besides this, there exists affine function $b(\cdot)$ dominated by f $\Rightarrow$ when $\lambda > 0$, the affine function $b(\cdot) + \lambda a(\cdot)$ is dominated by f $\Rightarrow$ $\overline{f}(\bar{x}) \geq b(\bar{x}) + \lambda a(\bar{x}) \to +\infty$, $\lambda \to \infty$ $\Rightarrow$ $f(\bar{x}) = +\infty$, Q.E.D.

We see that a convex function f may differ from the supremum $\overline{f}$ of its affine minorants on a "subtle" set – the relative boundary of Dom f only. The convex lsc functions are exactly those for which f and $\overline{f}$ are the same.

Fact VII.3 *A function $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ is convex lsc if and only if it is the pointwise supremum of all affine functions dominated by f.*

Proof. ✓ In one direction: the pointwise supremum of any family of affine functions is convex and lsc.
✓ In the opposite direction: Let f be convex and lsc, and let us prove that f is exactly the supremum $\overline{f}$ of all affine functions dominated by f. There is nothing to prove when $f \equiv +\infty$. Assuming f to be proper (i.e., $\mathrm{Dom}\, f \neq \emptyset$), we already know that $f = \overline{f}$ on $\mathrm{rint}\,\mathrm{Dom}\, f$ and outside of $\mathrm{cl}\,\mathrm{Dom}\, f$, and all we need to verify is that $f(\bar{x}) = \overline{f}(\bar{x})$ when $\bar{x} \in \mathrm{cl}\,\mathrm{Dom}\, f$. As $\overline{f} \leq f$, the relation $\overline{f}(\bar{x}) = f(\bar{x})$ folds true when $\overline{f}(\bar{x}) = +\infty$. Now let $\overline{f}(\bar{x}) \in \mathbf{R}$. Selecting $x' \in \mathrm{rint}\,\mathrm{Dom}\, f$, let $x^i = x' + \lambda_i[\bar{x} - x']$ with $0 \leq \lambda_i \to 1 - 0$ as $i \to \infty$. Then $x^i \in \mathrm{rint}\,\mathrm{Dom}\, f$ (Fact II.29), whence

$$f(x^i) = \overline{f}(x^i) \underbrace{\leq \overline{f}(x') + \lambda_i[\overline{f}(\bar{x}) - \overline{f}(x')]}_{\overline{f} \text{ is convex!}} \to \overline{f},\ i \to \infty \ \&\ \ x^i \to \bar{x},\ i \to \infty.$$

As f is lsc, we get $f(\bar{x}) \leq \overline{f}(\bar{x})$, implying that $f(\bar{x}) = \overline{f}(\bar{x})$ due to $\overline{f} \leq f$. Q.E.D.

Closure of a convex function

♣ Let $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ be a convex function.
♠ f not necessarily is closed; for example, the epigraph of the *characteristic function*

$$\Upsilon_G(x) = \begin{cases} 0 & , x \in G \\ +\infty & , x \notin G \end{cases}$$

of a convex set $G \subset \mathbf{R}^n$ is $G \times \mathbf{R}_+$ and is closed iff G is so.
However: *There exist convex lsc functions* (even affine ones) *dominated by* f $\Rightarrow$ *There exists the largest convex lsc function which is dominated by* f*, specifically, the pointwise supremum* $\mathrm{cl}\, f$ *(called closure of* f*) of all convex lsc functions dominated by* f (recall that the pointwise supremum of a whatever (nonempty) family of convex functions is convex, and similarly for lsc functions). The geometry of $\mathrm{cl}\, f$ is as it should be:

Fact VII.4 *For a convex function* F*, the epigraph of its closure* $\mathrm{cl}\, f$ *is exactly the closure* $\mathrm{cl}\,\mathrm{Epi}\{x\}$ *of the epigraph of* f*:*

$$\mathrm{Epi}\{\mathrm{cl}\, f\} = \mathrm{cl}\,\mathrm{Epi}\{f\}. \qquad (*)$$

Besides this, $\mathrm{cl}\, f$ *is the pointwise supremum of affine functions dominated by* f*.*

For a convex function F, the epigraph of its closure $\mathrm{cl}\, f$ is exactly the closure $\mathrm{cl}\,\mathrm{Epi}\{x\}$ of the epigraph of f:

$$\mathrm{Epi}\{\mathrm{cl}\, f\} = \mathrm{cl}\,\mathrm{Epi}\{f\}. \qquad (*)$$

Besides this, $\mathrm{cl}\, f$ is the pointwise supremum of affine functions dominated by f.

✓ To prove $(*)$, it suffices to prove that the set $\overline{E} = \mathrm{cl}\,\mathrm{Epi}\{f\}$ is an epigraph of a function talking values in $\mathbf{R} \cup \{+\infty\}$. Indeed, assuming that $\overline{E} = \mathrm{Epi}\{\widetilde{f}\}$, function $\widetilde{f}$ is convex and closed along with $\overline{E}$, and dominates every closed convex function dominated by f (since the epigraph of such a function should be closed convex set containing $\mathrm{Epi}\{f\}$ and thus containing $\overline{E}$). As $\mathrm{cl}\, f$ is the largest closed convex function dominated by f, it follows that $\widetilde{\{f\}} = \mathrm{cl}\, f$, and $(*)$ follows due to $\mathrm{Epi}\{\widetilde{f}\} = \overline{E} = \mathrm{cl}\,\mathrm{Epi}\{f\}$.

It remains to verify that $\overline{E}$ is an epigraph. The necessary and sufficient condition for a set $E \subset \mathbf{R}^n_x \times \mathbf{R}_t$ to be an epigraph is to have as the intersection with any vertical line $L_x = [x; t] : t \in R\}$ either an empty set or a closed ray $\{x; t : t \geq a_x\}$ with $a_x \in \mathbf{R}$, As applied to $\overline{E}$, this condition definitely holds true. Assuming $\overline{E}$ nonempty, this closed convex set contains a ray directed by $[0_{n\times 1}; 1]$, and thus its intersection with a vertical line L_x is either empty, or is a closed ray $[\{: t \geq a_x\}$ with $a_x \in \mathbf{R}$, or is the entire line L_x; the latter option is impossible, since convex function f is below bounded on every bounded set in $\mathbf{R}^n$, that is, the intersection of $\mathrm{Epi}\{f\}$ with the inverse image, under the projection $[x; t] \mapsto t$, of a bounded set U in $\mathbf{R}^n$ is contained in a half-space of the form $\{x, t] : t \geq a_U \in \mathbf{R}\}$, and this property clearly remains intact when passing from $\mathrm{Epi}\{f\}$ to $\mathrm{cl}\,\mathrm{Epi}\{f\}$. Thus, $\overline{E}$ indeed is an epigraph, and $(*)$ is proved.

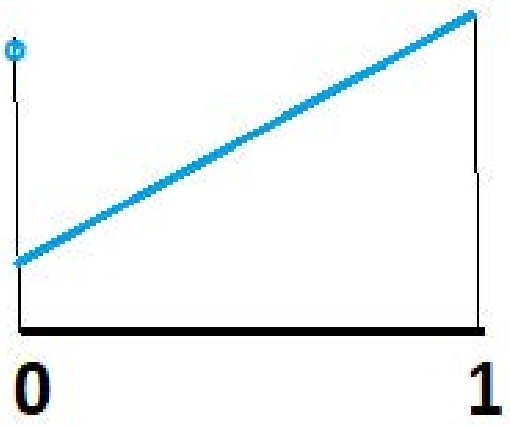
✓ To prove the "Besides this" claim, note that as $\mathrm{cl}\, f \leq f$ and $\mathrm{cl}\, f$ is the largest closed convex function dominated by f, affine function dominated by $\mathrm{cl}\, f$ are exactly the same as affine functions dominated by f, and $cl f$, as any closed convex function, is the pointwise supremum of its affine minorants (Fact VII.3).

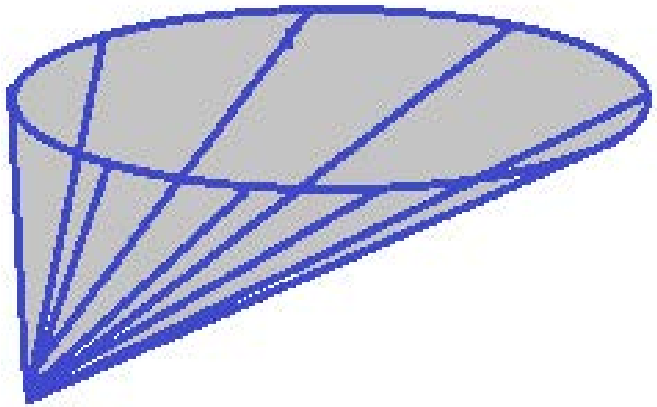
Note: On a straightforward inspection, we have proved the following

Fact VII.5 *Let $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ be a function, not necessarily convex, which possesses affine minorants. The pointwise supremum $\mathrm{cl}\, f$ of these minorants is the largest closed convex function dominated by f, and*

$$\mathrm{Epi}\{\mathrm{cl}\, f\} = \mathrm{cl}\,\mathrm{Conv}(\mathrm{Epi}\{f\})$$

Illustration: Discontinuous convex functions.





Left: Discontinuous convex usc function f with polyhedral domain $\mathrm{Dom}\, f = [0, 1]$.
Note: f is convex and discontinuous: it "jumps up" at a boundary point of $\mathrm{Dom}\, f$. "Jumps up" at some points from the domain's boundary is the only type of discontinuity allowed for convex functions *with polyhedral domains.* In particular, *every lsc convex function with polyhedral domain is continuous on this domain.*

Right: Discontinuous convex lsc function with non-polyhedral domain.
Note: The function is the supremum of all affine functions on the 2D plane which are ≤ 0 at the origin and are ≤ 1 on the circumference

$$C = \{(x-1)^2 + y^2 \leq 1\}.$$

.The function is convex lsc and discontinuous: it is 0 at the origin and is 1 at all distinct from the origin points of C.

Legendre Transform of convex function

♣ Let $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ be a *proper* function, i.e., Dom $f \neq \emptyset$.
- **Question:** When an affine function $y^T x - a$ is dominated by f?
- **Answer:** It is the case iff $a \geq y^T x - f(x)$, i.e. iff

$$a \geq f_*(y) := \sup_x [y^T x - f(x)] = \sup_{x \in \mathrm{Dom} f} [y^T z - f(x)] \qquad (!)$$

The function $f_*(\cdot)$ given by (!) takes values in $\mathbf{R} \cup \{+\infty\}$ (as f is proper) and is called the *Legendre* (a.k.a. *Fenchel*) *transform of* f.

♠ Immediate observations:

Fact VII.6 *Let function $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ be proper and possess affine minorants (e.g. f is convex). Then*

A. *The Legendre transform f_* of f is a proper lsc convex function, and*

$$(f_*)_*(x) := \sup_x [x^T y - f_*(y)] = \sup_{y \in \mathrm{Dom} f} [y^T x - f_*(y)]$$

is the closure cl $f(x)$ *of f – the largest convex lsc function dominated by f.*
in particular, when f is proper convex lsc, so is f_*, and f is the Legendre transform of f_*.

Indeed, function f_* is proper (as f possesses affine minorants) convex lsc (as the supremum of nonempty family of convex lsc functions). Setting $\{a \in \mathbf{R} : y^T \xi - a \leq f(\xi)\, \forall \xi\} = \{a : a \geq f_*(y)\}$, we have

$$\sup_y [y^T x - f_*(y)] = \sup_{y,a} \{y^T x - a : a \geq f_*(y)\} = \sup_{y,a} \{y^T x - a : y^T \xi - a \leq f(\xi)\, \forall \xi\},$$

i.e., $(f_*)_*(x)$ is the supremum, as evaluated at x, of all affine minorants of f, that is, cl $f(x)$, Q.E.D.

When f is proper convex lsc, $f =$ cl f, that is, $(f_*)_* = f$.

B. *Let f (or, which is the same, f_*) be proper convex lsc. Then*

$$\underbrace{\operatorname*{Argmax}_y[x^Ty - f_*(y)] = \partial f(x)}_{(a)} \quad \& \quad \underbrace{\operatorname*{Argmax}_x[x^Ty - f(x)] = \partial f_*(y)}_{(b)}$$

Indeed, if $f(x) = +\infty$, then both $\partial f(x) = \emptyset$ and $\sup_y[x^Ty - f_*(y)] = f(x) = +\infty$, whence $\operatorname{Argmax}_y[x^Ty - f_*(y)] = \emptyset = \partial f(x)$. When $x \in \operatorname{Dom} f$, we have

$$\begin{array}{l}\bar{y} \in \operatorname{Argmax}_y[x^Ty - f_*(y)] \Leftrightarrow x^T\bar{y} - f_*(\bar{y}) = f(x) \Leftrightarrow f(x) = x^T\bar{y} - \sup_\xi[\xi^T\bar{y} - f(\xi)] \\ \Leftrightarrow f(x) = x^T\bar{y} + \inf_\xi[f(\xi) - \xi^Tx] \Leftrightarrow \inf_\xi\{f(\xi) - [f(x) + \bar{y}^T[\xi - x]]\} = 0 \Leftrightarrow \bar{y} \in \partial f(x)\end{array}$$

and we arrive at (a). Swapping f, f_*, the same reasoning results in (b).

C. *For all x, y one has*

$$x^Ty \le f(x) + f_*(y) \qquad (!)$$

If one of the functions f, f_* is proper convex lsc (whence, by **A**, *both f, f_* are proper convex lsc), inequality (!) is tight:*

$$\underbrace{\forall(x \in \operatorname{Dom} f) : \inf_y(f(x) + f_*(y) - x^Ty) = 0}_{(c)} \quad \& \quad \underbrace{\forall(y \in \operatorname{Dom} f_*) : \inf_x(f(x) + f_*(y) - x^Ty) = 0}_{(d)}$$

Moreover, when one of the functions f, f_* (and then the other one as well) is proper convex lsc, then a pair x, y makes (!) equality iff $y \in \partial f(x)$, same as iff $x \in \partial f_*(y)$.

Indeed, by the definition of Legendre transform, $\forall(x, y) : f_*(y) \ge x^Ty - f(x)$, implying (!). Now let one of f, f_* be proper convex lsc; then both of these functions are so, and $\infty > f_*(y) = \sup_x[y^Tx - f(x)] \Leftrightarrow \infty > f_*(y) = \sup_{x \in \operatorname{Dom} f}[y^Tx - f(x)]$ is the same as $y \in \operatorname{Dom} f_* \Rightarrow \inf_{x \in \operatorname{Dom} f}[f_*(y) + f(x) - x^Ty] = 0$, which is (d). Swapping f and F_*, we arrive at (c).

The "Moreover" part of **C** follows from **B**, since (!) combines with the definition of the Legendre transform to imply that (!) is equality iff $x \in \operatorname{Argmax}_\xi[y^T\xi - f(\xi)]$, same as iff $y \in \operatorname{Argmax}_\eta[x^T\eta - f_*(\eta)]$.

♠ As a corollary of Fact VII.6.B,

Fact VII.7 *A proper convex lsc function* $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ *is below bounded iff* $0 \in \mathrm{Dom}\, f_*$, *in which case*

$$\inf f = -f_*(0),$$

and the set of minimizers of f, *if any, is*

$$\operatorname*{Argmin}_x f(x) = \partial f_*(0),$$

Indeed, $f_*(0) = \sup_x\{-f(x)\} = -\inf_x f(x)$.

Legendre transform of a polyhedral function

♠ **Polyhedral functions.** We call a function $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ *polyhedral*, if it is proper and its epigraph is a polyhedral set. A polyhedral representation

$$\mathrm{Epi}\{f\} = \{[x;t] : \exists u : Px + tp + Qu \leq r\}$$

of the epigraph of f is called a *polyhedral representation of f.*

• A polyhedral function is proper, convex and lsc (as its epigraph is a polyhedral set and as such is convex and closed).

♠ Calculus of convex sets immediately implies that *the results of basic convexity-preserving operations with functions – taking linear combinations, finite maxima, affine substitution of arguments, as applied to polyhedral functions given by polyhedral representations are polyhedral as well, with a polyhedral representation of the results readily given by those of the operands.* The same takes place for monotone superposition

$$f_1, ..., f_m, F \mapsto \{g(x) = \begin{cases} F(f_1(x), ..., f_m(x))) & , f_i(x) < \infty \, \forall i \\ +\infty & , \text{otherwise} \end{cases}$$

when all f_i and F are polyhedral, and F satisfies the standard convexity-preserving condition "$F(y_1, , , .y_m)$ is nondecreasing in every i for which f_i is not affine."

♠ Polyhedrality is respected by Legendre transformation: given a polyhedral representation

$$t \geq f(x) \Leftrightarrow \exists u : Px + tp + Qu \leq r,$$

we have

$$\begin{array}{rcl} f_*(y) \leq t & \Leftrightarrow & \max_x[y^T x - f(x)] \leq t \\ & \Leftrightarrow & \max_{x,\tau}\{y^T x - \tau : \tau \geq f(x)\} \leq t \\ & \Leftrightarrow & \max_{x,\tau,u}\{y^T x - \tau : Px + \tau p + Qu \leq r\} \leq t \\ & \Leftrightarrow & \min_\lambda\{r^T\lambda : P^T\lambda = y, p^T\lambda = -1, Q^T\lambda = 0, \lambda \geq 0\} \leq t \text{ [LP duality]} \\ & \Leftrightarrow & \exists \lambda : r^T\lambda \leq t, P^T\lambda = y, p^T\lambda = -1, Q^T\lambda = 0, \lambda \geq 0 \end{array}$$

and we end up with a polyhedral representation of f_*.

"Classical case"

♠ Let $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ be convex lsc function with domain $\operatorname{int} G$, where G is a closed convex set with a nonempty interior, and let f be twice continuously differentiable on $\operatorname{int} G$ and strongly convex, meaning that

$$\forall (x \in \operatorname{int} G, h \neq 0) : \frac{d^2}{dt^2} f(x + th) > 0.$$

Note: Since $\operatorname{Dom} f = \operatorname{int} G$, so that $f = +\infty$ on the boundary of G, and f is lsc, f is a *barrier* for G:

$$x^i \in \operatorname{int} G, \lim_{i \to \infty} x^i \in \partial G \Rightarrow f(x^i) \to \infty \text{ as } i \to \infty$$

Assume also that

(!) *Whenever the function $y^T x - f(x)$ is bounded from above, it achieves its maximum*

Examples:

1. $G = \mathbf{R}^n$ is the entire space, $f(x) = \frac{1}{2} x^T Q x - 2 q^T x$ with $Q \succ 0$

2. $G = \{x : \|x\|_2 \leq 1\}$ is the unit Euclidean ball, $f(x) = -\ln(1 - x^T x)$

3. $G = \{x : \|x\|_\infty \leq 1\}$ is the unit box, $f(x) = -\sum_i \ln(1 - x_i^2)$

4. $G = \mathbf{R}^n_+$ is the nonnegative orthant, $f(x) = -\sum_i \ln x_i$

5. $G = \mathbf{L}^n = \{x : x_n \geq \sqrt{x_1^2 + ... + x_{n-1}^2}\}$ is the Lorentz cone, $f(x) = \ln(x_n^2 - s_1^2 - ... - x_{n-1}^2)$

6. $G = \mathbf{S}^n_+ := \{x \in \mathbf{S}^n : x \succeq 0\}$ is the semidefinite cone, $f(x) = -\ln \operatorname{Det} x$

Note: In some of these examples, strong convexity (and just convexity itself) of f and (!) are not evident and require verification (for the time being, postponed).

Fact VII.8 *In the case in question, the domain of f_* is open, and thus is $\mathrm{int}\, G_*$ for a closed convex set G_*, f_* is twice continuously differentiable and strongly convex on $\mathrm{int}\, G_*$, the function $x^T y - f_*(y)$ attains its maximum in y whenever it is bounded from above, and the mappings $x \mapsto \nabla f(x) : x \in \mathrm{int}\, G \to \mathbf{R}^n$ and $y \mapsto \nabla f_*(y) : \mathrm{int}\, G_* \to \mathbf{R}^n$ establish inverse to each other continuously differentiable correspondences between $\mathrm{int}\, G$ and $\mathrm{int}\, G_*$.*

Proof. ✓ Observe, first, that the mapping $x \mapsto F * x) := f'(x)$ is a one-to-one mapping of $\mathrm{int}\, G$ onto $\mathrm{int}\, G_*$. Indeed, $\mathrm{int}\, G_*$ is exactly the set of those y for which the function $y^T x - f(x)$ is above bounded, or, which is the same by (!) and the Fermat rule, of those y for which $y = \nabla f(x_y)$ for some x_y. As f is strongly convex, x_y is uniquely defined by y. Vice versa, if $x \in \mathrm{int}\, G$, we have $\mathrm{Argmax}_\xi[[\nabla f(x)]^T \xi - f(\xi)] = \{x\}$, implying that $\nabla f(x) \in \mathrm{Dom}\, f_*$. The bottom line is that the mapping $x \mapsto y(x) := \nabla f(x) : \mathrm{int}\, G \to \mathbf{R}^n$ maps $\mathrm{int}\, G$ onto $\mathrm{Dom}\, f_*$, and this mapping is one-to-one (as when $\nabla f(x) = \nabla f(x') = y$, the strongly convex function $f(\xi) - y^T \xi$ attains its minimum on $\mathrm{int}\, G$ both at x and at x', implying that $x = x'$).
✓ The Jacobian of the continuously differentiable on its domain $\mathrm{int}\, G$ mapping f is the Hessian of f, and therefore it is nondegenerate at every point from this domain. Applying the Inverse Function Theorem of Calculus, nondegeneracy of the Jacobian of the one-to-one mapping F implies that the image $F(\mathrm{int}\, G)$ of the domain of F (this image, as we know, is $\mathrm{Dom}\, f_*$) is open, and the inverse F^{-1} of F is continuously differentiable on $\mathrm{int}\, G_*$. Now, when $y \in \mathrm{int}\, G_*$ and $x = F^{-1}(y)$, that is, $x \in \mathrm{int}\, G$ and $y = \nabla f(x)$, we have $f_*(y) = \max_\xi [y^T \xi - f(\xi)] = y^T x - f(x)$ (by Fermat rule), that is, $f_*(y) = y^T F^{-1}(y) - f(F^{-1}(y))$, implying that f_* is continuously differentiable on $\mathrm{int}\, G_*$. Finally, for our y and x we have $f_*(y) = y^T x - f(x)$, implying by Fact VII.6.C that $x = F^{-1}(y) \in \partial f_*(y)$. As f_* is continuously differentiable, $\partial f_*(y) = \{\nabla f_*(y)\} \Rightarrow \nabla f_*(y) = F^{-1}(y)$. Thus, $\nabla f_*(y)$ is continuously differentiable ($\Rightarrow$ f_* is twice continuously differentiable), and the mappings $x \mapsto F(x) = \nabla f(x)$ and $y \mapsto F^{-1}(y) = \nabla f_*(y)$ establish one-to-one correspondence between $\mathrm{int}\, G$ and $\mathrm{int}\, G_*$. By Chain rule the Jacobians of F and of $F * -1$ (i.e., the Hessians of f and of f_*) taken at the corresponding to each other points $x \in \mathrm{int}\, G$ and $y \in \mathrm{int}\, G_*$ are inverses of each other $\Rightarrow$ f_* is strongly convex on $\mathrm{int}\, G_*$. Finally, from our analysis, when the function $x^T y - f_*(y)$ is bounded from above (that is, $x \in \mathrm{Dom}\, (f_*)_* = \mathrm{Dom}\, f$, the function achieves its maximum (namely, at the point $\nabla f(x)$). The proof is complete.

How it works

1. $f(x) = \frac{1}{2}x^TQx - q^Tx$ $[Q \succ 0]$, $\mathrm{Dom}\, f = \mathbf{R}^n \Rightarrow f_*(y) = \frac{1}{2}(y+q)^TQ^{-1}(y+q)$, $\mathrm{Dom}\, f_* = \mathbf{R}^n$

$$\nabla f(X) = Qx - q, \quad \nabla f_*(y) = Q^{-1}(y+q)$$

2. $f(x) = -\ln(1-x^Tx)$, $\mathrm{Dom}\, f = \{x : x^Tx < 1\} \Rightarrow f_*(y) = \frac{y^Ty}{\sqrt{y^Ty+1}+1} - \ln(\sqrt{y^Ty+1}+1) + \ln 2$, $\mathrm{Dom}\, f_* = \mathbf{R}^n$,

$$\nabla f(x) = \frac{x}{1-x^Tx}, \quad \nabla f_*(y) = \frac{y}{\sqrt{y^Ty+1}+1}$$

3. $f(x) = -\sum_i \log(1-x_i^2)$, $\mathrm{Dom}\, f = \{x : \|x\|_\infty < 1\} \Rightarrow f_*(y) = \sum_i \left[\frac{y_i^2}{\sqrt{y_i^2+1}+1} - \ln(\sqrt{y_i^2+1}+1) + \ln 2\right]$, $\mathrm{Dom}\, f_* = \mathbf{R}^n$,

$$[\nabla f(x)]_i = \frac{2x_i}{1-x_i^2}, \quad [\nabla f_*(y)]_i = \frac{2y_i}{\sqrt{y_i^2+1}+1}$$

4. $f(x) = -\sum_i \ln x_i$, $\mathrm{Dom}\, f = \mathrm{int}\, \mathbf{R}^n_+ = \{x : x > 0\} \Rightarrow f_*(y) = -\sum_i \ln(-y_i) - n$, $\mathrm{Dom}\, f_* = -\mathrm{int}\, \mathbf{R}^n_+$

$$[\nabla f(x)]_i = -1/x_i, \quad [\nabla f_*(y)]_i = -1/y_i$$

5. $f(x) = -\ln(x_n^2 - x_1^2 - \ldots - x_{n-1}^2) = -\ln(x^TJx)$, $J = \mathrm{Diag}\{-1, \ldots, -1, 1\}$. $\mathrm{Dom}\, f = \mathrm{int}\, \mathbf{L}^n = \{x : x_n > 0, x^TJx > 0\}$ $\Rightarrow f_*(y) = -\ln(y^TJy) + 2\ln(2) - 2$, $\mathrm{Dom}\, f_* = -\mathrm{int}\, \mathbf{L}^n$,

$$Df(x)[h] = 2h^TJx/(x^TJx), \quad Df_*(y)[h] = 2h^TJy/(y^TJy)$$

6. $f(x) = -\ln \mathrm{Det}(x)$, $\mathrm{Dom}\, f = \mathrm{int}\, \mathbf{S}^n_+ = \{x \in \mathbf{S}^n : x \succ 0\} \Rightarrow f_*(y) = -\ln \mathrm{Det}(-y) + n$. $\mathrm{Dom}\, f_* = -\mathrm{int}\, \mathbf{S}^n$

$$Df(x)[h] = -\mathrm{Tr}(x^{-1}h), \quad Df_*(y)[h] = -\mathrm{Tr}(y^{-1}h)$$

Legendre transforms of smooth and of strongly convex functions

♣ A. Conjugate norms

♠ Let $\|\cdot\|$ be a norm on $\mathbf{R}^n$. Its *conjugate norm* is defined as

$$\|y\|_* = \max_x \{y^T x : \|x\| \leq 1\}$$

As is immediately seen, $\|\cdot\|_*$ indeed is a norm, and its unit ball is the polar of the unit ball of $\|\cdot\|$:

$$\{y : \|y\|_* \leq 1\} = \text{Polar}\left(\{x : \|x\| \leq 1\}\right).$$

Since every closed convex set containing the origin is the polar of its polar, *the conjugate of the conjugate norm $\|\cdot\|_*$ is the original norm $\|\cdot\|$*

Note: $\|\cdot\|_*$ is the smallest norm for which the inequality

$$y^T x \leq \|x\|\|y\|_* \ \forall x, y$$

holds true. This inequality is tight, meaning that *for every x there exists nonzero y making this inequality an equality*, and similarly, *for every y there exists a nonzero x making the inequality an equality*.

Examples:

- The conjugate of $\|\cdot\|_2$ is $\|\cdot\|_2$ itself;
- The conjugate of $\|\cdot\|_1$ is $\|\cdot\|_\infty$, and the conjugate of $\|\cdot\|_\infty$ is $\|\cdot\|_1$.

More generally, we shall see in a while that

The conjugate of the norm $\|\cdot\|_p$, $1 \leq p \leq \infty$, is the norm $\|\cdot\|_q$ with q given by $\frac{1}{p} + \frac{1}{q} = 1$.

Note: *For $L > 0$. the functions $\frac{L}{2}\|x\|^2$ and $\frac{1}{2L}\|y\|_*^2$ are Legendre transforms of each other.*

♣ B. Smooth and strongly convex functions

♠ Let f be a convex function, $\|\cdot\|$ be a norm on $\mathbf{R}^n$, and L be a nonnegative real. We say that f *is* $(L, \|\cdot\|)$*-smooth*, if $\mathrm{Dom}\, f = \mathbf{R}^n$ and

$$\forall (x, z \in \mathbf{R}^n, e \in \partial f(x)) : f(z) \le f(x) + e^T[z - x] + \frac{L}{2}\|z - x\|^2.$$

It is easily seen that *a convex function $f : \mathbf{R}^n \to \mathbf{R}$ is $(L, \|\cdot\|)$-smooth if and only if it is continuously differentiable,and the mapping $x \mapsto \nabla f(x)$ is Lipschitz continuous, with constant L, from the norm $\|\cdot\|$ to the norm $\|\cdot\|_*$:*

$$\|\nabla f(x) - \nabla f(y)\|_* \le L\|x - y\| \ \forall x, y$$

same as *if and only if f is continuously differentiable and*

$$[x - y]^T[\nabla f(x) - \nabla f(y)] \le L\|x - y\|^2 \ \forall x, y.$$

A twice continuously differentiable function $f : \mathbf{R}^n \to \mathbf{R}$ is $(L, \|\cdot\|)$-smooth if and only if $0 \le \frac{d^2}{dt^2}\big|_{t=0} f(x + th) \le L\|h\|^2$ for all x, h.

For example: *Convex quadratic function $f = \frac{1}{2}x^TQx - q^Tx + c$ $(Q \succeq 0)$ is $(L, \|\cdot\|_2)$-smooth whenever the eigenvalues of Q are upper-bounded by L.*

♠ Let $g : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ be a proper convex lsc function, $\|\cdot\|_*$ be the conjugate of a norm $\|\cdot\|$, and L be a positive real. g is called $(L, \|\cdot\|_*)$*-strongly convex*, if for every $\bar{y} \in \mathrm{Dom}\, g$ it holds

$$\forall (y \in \mathbf{R}^n, e \in \partial g(\bar{y})) : g(y) \ge g(\bar{y}) + [y - \bar{y}]^T e + \frac{1}{2L}\|y - \bar{y}\|_*^2.$$

It can be proved that *a proper convex lsc function g is $(L, \|\cdot\|_*)$-strongly convex if and only if*

$$[e' - e]^T[y' - y] \ge \frac{1}{L}\|y' - y\|_*^2 \ \forall (y, y' \in \mathrm{Dom}\, g, e \in \partial g(y), e' \in \partial g(y'))$$

A twice continuously differentiable on $\mathrm{rint}\,\mathrm{Dom}\, g$ *convex lsc function g is $(L, \|\cdot\|_*)$-strongly convex if and only if $\frac{d^2}{dt^2}\big|_{t=0} g(y + th) \ge L^{-1}\|h\|_*^2$ for all $y \in \mathrm{rint}\,\mathrm{Dom}\, g$ and all h from the linear subspace parallel to* $\mathrm{Aff}(\mathrm{Dom}\, g)$.

For example: *Convex quadratic form $f(x) = \frac{1}{2}x^TQx - q^Tx + c : \mathbf{R}^n \to \mathbf{R}$ is $(L, \|\cdot\|_2)$-strongly convex if and only if $Q \succ 0$ and all eigenvalues of Q are lower-bounded by L^{-1}.*

Note: When f is convex quadratic form with the matrix of the quadratic part equal to $Q \succ 0$, f_* is convex quadratic form with the matrix of the quadratic part equal to Q^{-1}. From the examples above, *A quadratic form with positive definite matrix of the quadratic part, the form is $(L, \|\cdot\|_2)$-smooth if and only if the Legendre transform f_* of f is $(L, \|\cdot\|_2)$-strongly convex.*

♠ The just outlined relation between the smoothness and strong convexity of convex quadratic form is a particular case of the following general result:

Fact VII.9 *The Legendre transform f_* of an $(L, \|\cdot\|)$-smooth convex function $f : \mathbf{R}^n \to \mathbf{R}$ is $(L, \|\cdot\|_*)$-strongly convex. Vice versa, if the Legendre transform of a convex function $f : \mathbf{R}^n \to \mathbf{R}$ is $(L, \|\cdot\|_*)$-strongly convex, then f is $(L, \|\cdot\|)$-smooth.*

Proof.
✓ Let $f : \mathbf{R}^n \to \mathbf{R}$ be $(L, \|\cdot\|)$-smooth convex function, let $\bar{y} \in \mathrm{Dom}\, f_*$ be such that $\partial f_*(\bar{y}) \neq \emptyset$, and let $\bar{x} \in \partial f_*(\bar{y})$. By Fact VII.6.B.$(b)$, $\bar{x}$ is a maximizer over $x \in \mathbf{R}^n$ of the function $\bar{y}^T x - f(x)$, whence $f_*(\bar{y}) = \bar{y}^T\bar{x} - f(\bar{x})$. Besides this, as $\bar{x} \in \partial f(\bar{y})$, $\bar{y}$ is a maximizer of $y^T\bar{x} - f_*(y)$ over y, whence $\bar{y} \in \partial f(\bar{x})$ by Fact VII.6.B.(a) $\Rightarrow$ $f(x) \leq f(\bar{x}) + \bar{y}^T[x - \bar{x}] + \frac{L}{2}\|x - \bar{x}\|^2$ (f is $(L, \|\cdot\|)$-smooth !)$\Rightarrow$

$$\begin{array}{l} f_*(y) = \sup_x[y^T x - f(x)] \geq \sup_x\{y^T x - f(\bar{x}) - \bar{y}^T[x - \bar{x}] - \frac{L}{2}\|x - \bar{x}\|^2\} \\ = \sup_x\{[y - \bar{y}]^T[x - \bar{x}] - \frac{L}{2}\|x - \bar{x}\|^2\} + y^T\bar{x} - f(\bar{x}) = \frac{1}{2L}\|y - \bar{y}\|_*^2 + y^T\bar{x} - f(\bar{x}) \\ = \frac{1}{2L}\|y - y_*\|_*^2 + [y - \bar{y}]^T\bar{x} + [\bar{y}^T\bar{x} - f(\bar{x})] = f_*(\bar{y}) + [y - \bar{y}]^T\bar{x} + \frac{1}{2L}\|y - \bar{y}\|_*^2 \end{array}$$

(recall that the Legendre transform of $\frac{L}{2}\|\cdot\|^2$ is $\frac{1}{2L}\|\cdot\|_*^2$). Thus,

$$f_*(y) \geq \ f_*(\bar{y}) + \bar{x}^T[y - \bar{y}] + \frac{1}{2L}\|y - \bar{y}\|^2$$

whenever $\bar{x} \in \mathrm{Argmax}_x[\bar{y}^T x - f(x)] = \partial f_*(\bar{y})$ (the latter relation is given by Fact VII.6.B.(b)). Thus, f_* is $(L, \|\cdot\|_*)$-strongly convex.
✓ Vice versa, let f_* be $(L, \|\cdot\|_*)$-strongly convex, and let us prove that f is $(L, \|\cdot\|)$-smooth. To verify that $\mathrm{Dom}\, f = \mathbf{R}^n$, take $y \in \mathrm{rint}\,\mathrm{Dom}\, f_*$ and $e \in \partial f_*(y)$; then $f_*(z) \geq \overline{f}_*(z) := f_*(y) + e^T[z - y] + \frac{1}{3L}\|z - y\|_*^2$, implying that the function $x^T z - f_*(z) \leq x^T z - \overline{f}_*(z)$ of z is bounded from above for every x, whence $\mathrm{Dom}\, f = \mathrm{Dom}\,(f_*)_* = \mathbf{R}^n$. Now let $\bar{x} \in \mathbf{R}^n$, $\bar{y} \in \partial f(\bar{x})$. Then $\bar{x}$ is a maximizer of $\bar{y}^T x - f(x)$ over x, whence $\bar{y} \in \mathrm{Dom}\, f_*$, $f(\bar{x}) = \bar{y}^T\bar{x} - f_*(\bar{y})$, and $\bar{x} \in \partial f_*(\bar{y})$ by Fact VII.6.B.(b). As f_* is strongly convex and $\bar{x} \in \partial f_*(\bar{y})$, we have $f_*(y) \geq f_*(\bar{y}) + \bar{x}^T[y - \bar{y}]^T + \frac{1}{2L}\|y - \bar{y}\|_*^2$ for every y. Therefore, as f is the Legendre transform of f_*, we have

$$\begin{array}{l} f(x) = \sup_y[y^T x - f_*(y)] \leq \sup_y\left\{y^T x - f_*(\bar{y}) - \bar{x}^T[y - \bar{y}] - \frac{1}{2L}\|y - \bar{y}\|_*^2\right\} \\ = \sup_y\left\{[y - \bar{y}]^T[x - \bar{x}] - \frac{1}{2L}\|y - \bar{y}\|_*^2\right\} + \bar{y}^T[x - \bar{x}] + [\bar{y}^T\bar{x} - f_*(\bar{y})] = \frac{L}{2}\|x - \bar{x}\|^2 + [x - \bar{x}]^T\bar{y} + f(\bar{x}) \end{array}$$

The resulting inequality holds true for every $x \in \mathbf{R}^n$ and every $\bar{y} \in \partial f(\bar{x})$, implying that f is $(L, \|\cdot\|)$-smooth. □

Example: Let $f(x) = \ln(\sum_{i=1}^n e^{x_i})$. As we know, f is convex. Moreover, f is $(1, \|\cdot\|_\infty)$-smooth, since setting $p_i = e^{x_i}/(\sum_j e^{x_j})$, we have $p_i > 0$, $\sum_i p_i = 1$, whence

$$\frac{d^2}{dt^2}\Big|_{t=0} f(x+th) = \sum_i p_i h_i^2 - (\sum_i p_i h_i)^2 \le \sum_i p_i h_i^2 \le \|h\|_\infty^2.$$

Direct computation shows that

$$f_*(y) = \begin{cases} \sum_i y_i \ln(y_i) & , y \in \Delta = \{y \ge 0 : \sum_i y_i = 1\} \\ +\infty & , y \notin \Delta \end{cases},$$

and we conclude that f_* is $(1, \|\cdot\|_1)$-strongly convex – the fact playing important role in the design of *proximal First Order algorithms* for minimizing convex functions over the probabilistic simplex.

Characteristic, Support, and Minkowski functions of convex sets

♣ Let a set $X \subset \mathbf{R}^n$ be convex and nonempty.
♠ The **Characteristic function** Υ_X of X is

$$\Upsilon_X(x) = \begin{cases} 0 & , x \in X \\ +\infty & , x \notin X \end{cases}$$

As is immediately seen *The characteristic function Υ_X of a nonempty convex set X is proper convex function with the domain X. Υ_X is lsc iff X is closed, and one always has*

$$\mathrm{cl}\,\Upsilon_X = \Upsilon_{\mathrm{cl}\,X}$$

♠ The **Support function** $\mathrm{Supp}_X(y)$ of X is the Legendre transform of the characteristic function:

$$\mathrm{Supp}_X(y) = \sup_{x \in \mathbf{R}^n} [y^T x - \Upsilon_X(x)] = \sup_{x \in X} y^T x.$$

As is immediately seen,
S.0. *The support function of X is the same as the support function of $\mathrm{cl}\,X$, and the Legendre transform of $\mathrm{Supp}_X \equiv \mathrm{Supp}_{\mathrm{cl}\,X}$ is the characteristic function $\Upsilon_{\mathrm{cl}\,X}$ of $\mathrm{cl}\,X$* (as the latter function is proper convex lsc with Legendre transform $\mathrm{Supp}_{\mathrm{cl}\,X}$),
S.1. $\mathrm{Supp}_X(\cdot)$ *is proper convex lsc function which is positively homogeneous of degree 1:*

$$\forall (y \in \mathrm{Dom}\,\mathrm{Supp}_X, t \geq 0) : \mathrm{Supp}_X(ty) = t\mathrm{Supp}_X(y).$$

S.2 $\mathrm{Supp}_X(\cdot)$ *"remembers" the closure* $\mathrm{cl}\,X$ *of X, specifically,*

$$\mathrm{cl}\,X = \partial \mathrm{Supp}_X(0)$$

To verify **S.2**, note that the function $\Upsilon_{\mathrm{cl}\,X}(\cdot)$ is proper convex lsc, and the set of its minimizers is $\mathrm{cl}\,X$; it remains to refer to Fact VII.7.

♠ **S.1** can be inverted:

Fact VII.10 *Every proper convex lsc function $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ which is positively homogeneous of degree 1 is the support function of a nonempty closed convex set, namely, the set $\partial f(0)$.*

Indeed, with f as stated we have

$$f_*(x) = \sup_{y \in \mathrm{Dom} f} [x^T y - f(y)] = \sup_{y \in \mathrm{Dom} f, t \geq 0} [tx^T y - f(ty)] = \sup_{y \in \mathrm{Dom} f, t \geq 0} t[x^T y - f(y)]$$

and the latter sup is either zero (when x is such that $x^T y \leq f(y)$ for all y,) or $+\infty$ (when $x^T y > f(y)$ for some y), that is, f_* is the characteristic function of $\mathrm{Dom}\, f_*$, and since f_* is proper, convex, and lsc, $\mathrm{Dom}\, f_*$ is a nonempty closed convex set, and $f = (f_*)_*$ is the support function of this set. This combines with **S.2** to imply that the set in question is $\partial f(0)$.

Example I: *Kullback-Leibler divergence.* The function

$$x \ln(x/y) : \{x \geq 0\} \times \{y > 0\} \to \mathbf{R}$$

is convex (as the projective transform of the convex function $x \ln x$ with the domain $\{x \geq 0\}$). This function is *not* lsc; the closure of the function (which we still denote $x \ln(x/y)$) is obtained by adding to the original domain the origin $x = 0, y = 0$ in $\mathbf{R}^2$ (call this extended domain $\mathcal{D}$), preserving the values of the function at its original domain and setting $0 \ln(0/0) = 0$. Note that $x \ln(y/x)$ is positively homogeneous of degree 1.

♠ The *Kullback-Leibler divergence* is the function

$$f(x, y) = \sum_i x_i \ln(x_i/y_i) : \mathbf{R}^n \times \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}.$$

This is positively homogeneous of degree 1 proper convex lsc function (by Fact VII.2.I, the sum of lsc functions is lsc). Direct computation shows that

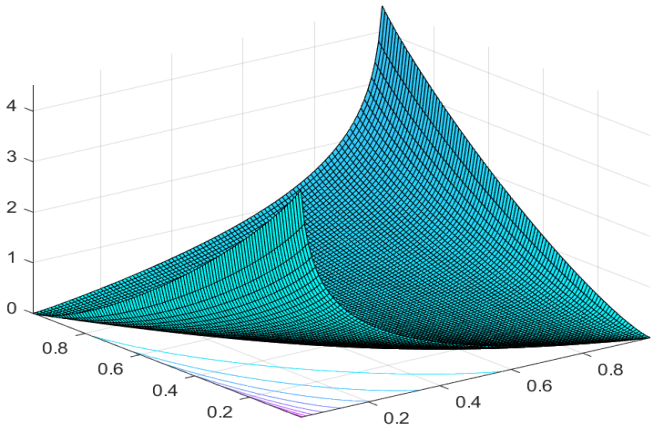
$$\partial f(0) = \mathcal{X} := \{(p, q) \in \mathbf{R}^n \times \mathbf{R}^n : q_i < 0, 1 + \ln(-q_i) \geq p_i, 1 \leq i \leq n\},$$

and therefore $f_* = \Upsilon_{\mathcal{X}}$.

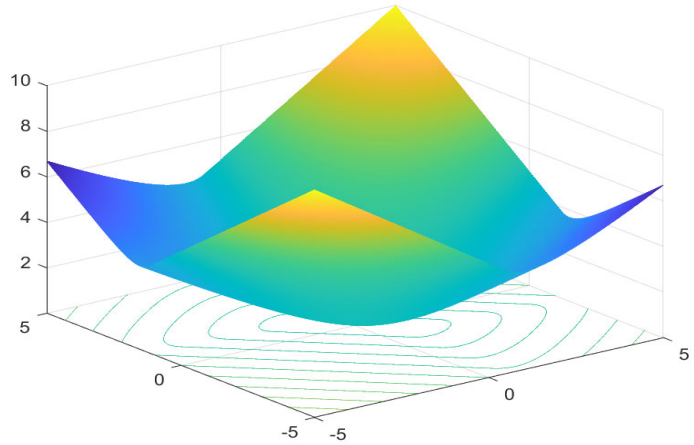
Example I continued: Typical applications of the KL divergence are in quantifying proximity between probability distributions. As a result, the "actual" KL divergence is the restriction of the KL divergence just defined onto the set of *probabilistic* vectors (x, y) (nonnegative entries summing up to 1). This function is

$$\mathcal{KL}(x,y) = \begin{cases} \sum_i x_i \ln(x_i/y_i) & , x, y \in \Delta_n = \{z \in \mathbf{R}^n : z \geq 0, \sum_i z_i = 1\} \\ +\infty & , \text{otherwise} \end{cases}$$

This function is *not* homogeneous, and its Legendre transform has nothing to do with the above f_* – its domain is the entire $\mathbf{R}^n$ (as is the case for the Legendre transform of any proper convex function with bounded domain)



$$KL(r,s) := x \ln\left(\tfrac{x}{1-y}\right) + (1-x)\ln\left(\tfrac{1-x}{y}\right)$$



$$KL_*(\cdot,\cdot)$$

Comment. When $n = 2$, the KL divergence $f(x, y)$ restricted onto Δ_2 becomes function of two real variables $r = x_1$, $s = y_2$ with the unit square $[0 \leq r, s \leq 1]$ (with vertices $[1; 0]$ and $[0; 1]$ excluded) as the domain. This function is presented on the left picture, and its Legendre transform is on the right picture.

Example II. The support function of the ball $B_R(\bar{x}) = \{x : \|x - \bar{x}\| \leq R\}$ $(R > 0)$ is the scaled linear perturbation of $\|\cdot\|_*$:

$$\mathrm{Supp}_{x \in B_R(\bar{x})}(y) = R\|y\|_* + \bar{x}^T y.$$

This is a special case of the general fact as follows:

Fact VII.11 *If f is a proper convex lsc function with Legendre transform f_*, A is nonsingular $n \times n$ matrix, $\alpha > 0$, and $s, e \in \mathbf{R}^n$, then the function*

$$g(x) = \alpha f(A[x - s]) + e^T x$$

is proper convex lsc with the Legendre transform

$$g_*(y) = \alpha f_*(A^{-T}[y - e]/\alpha) + s^T[y - e]$$

Indeed,

$$\begin{array}{rcl} g_*(y) & = & \sup_x\{y^T - \alpha f(Ax - s) - e^T y\} = \alpha \sup_x\{[\alpha^{-1}[y - e]^T x - f(\overbrace{A[x - s]}^{z})\} \\ & = & \alpha \sup_z\{\alpha^{-1}[y - e]^T[A^{-1}z + s] - f(z)\} = \alpha\left[\sup_z\{[\alpha^{-1}A^{-T}(y - e)]^T z - f(z) + \alpha^{-1}[y - e]^T s\right] \\ & = & \alpha f_*(A^{-T}[y - e]/\alpha) + s^T[y - e] \end{array}$$

♠ Domain of support function.

Fact VII.12 *The domain* $\mathrm{Dom}\,\mathrm{Supp}_X$ *of the support function of a closed nonempty convex set* $X \subset \mathbf{R}^n$ *is a cone (not necessarily closed) which is in-between the interior of the negation of the cone dual to the recessive cone of* X *and this negation itself:*

$$\mathrm{int}\,(-[\mathrm{Rec}(X)]_*) \subset \mathrm{Dom}\,\mathrm{Supp}_X \subset -[\mathrm{Rec}(X)]_*. \qquad (*)$$

First, Supp_X is positively homogeneous of degree 1 proper convex function, whence $D := \mathrm{Dom}\,\mathrm{Supp}_X$ is a cone. Let $K = -[\mathrm{Rec}(X)]_* = \{y : y^T h \le 0, \forall h \in \mathrm{Rec}(X)\}$, and let $\bar{x} \in X$. ✓ When $y \notin K$, there exists $h \in \mathrm{Rec}(X)$ with $y^T h > 0$, implying that $y^T[\bar{x} + th] \to \infty$ as $t \to \infty$. Since $\bar{x} + \mathbf{R}_+ h \in X$ due to $\bar{x} \in X$,, $h \in \mathrm{Rec}(X)$, we get $\mathrm{Supp}_X(y) = +\infty$, that is,, $y \notin D$. The right inclusion in $(*)$ is proved.
✓ To prove the left inclusion in $(*)$, let $y \in \mathrm{int}\,K$, and let us prove that $\mathrm{Supp}_X(y) < \infty$. Indeed, as $y \in \mathrm{int}\,K$, for certain C and all $h \in \mathrm{Rec}(X)$ it holds

$$\|h\|_2 \le C[-y^T h] \qquad (!)$$

(Fact IV.11). Assume now, on the contrary to what should be proved, that $y^T x^i \to +\infty$ for some sequence $\{x_i \in X\}$. Then the sequence clearly diverges; let h be an asymptotic direction of the sequence, that is, for some $i_1 < i_2 < ...$ we have

$$\|x^{i_j}\|_2 \to \infty \ \& \ x^{i_j}/\|x^{i_j}\|_2 \to h \text{ as } j \to \infty.$$

Passing to a subsequence, we can assume that $x^i/\|x^i\|_2 \to h$ as $i \to \infty$. Thus

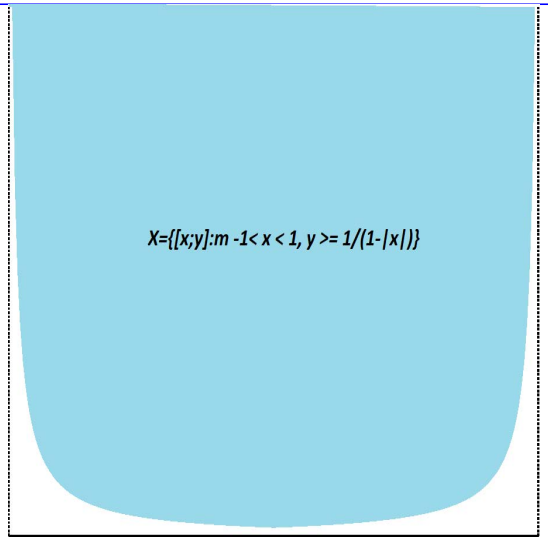
$$x^i = r_i h + r_i d^i \text{ with } r_i = \|x^i\|_2 \to \infty,\ i \to \infty, \text{ and } d_i \to 0,\ i \to \infty..$$

Besides this, h is the $\|\cdot\|_2$-unit vector from $\mathrm{Rec}(X)$, whence $y^T h \le -\alpha < 0$ by (!). We have

$$y^T x_i = r_i y^T h + r_i y^T d_i \le -\alpha r_i + r_i \|y\|_2 \|d_i\|_2;$$

as $i \to \infty$, the right hand side in this inequality goes to $-\infty$ since $d_i \to 0$, $i \to \infty$, and the left hand side goes to $+\infty$, which is the desired contradiction. .

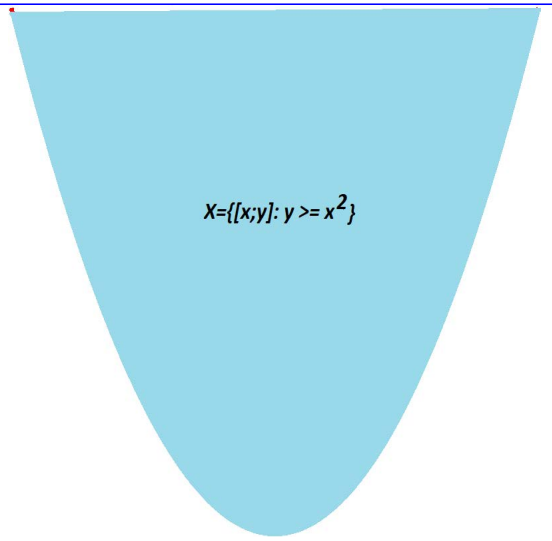
Quiz: Point out an unbounded closed nonempty convex set X such that
— $\operatorname{Dom}\operatorname{Supp}_X$ is nonclosed
— $\operatorname{Dom}\operatorname{Supp}_X$ is closed



X is closed, $\mathrm{Rec}(X) = \{[0; y] :\geq 0\}$

$\mathrm{Dom}\,\mathrm{Supp}_X = \{[p; q] : q \leq 0\} = -[\mathrm{Rec}(X)]_*$ is closed

$$\mathrm{Supp}_X(p,q) = \begin{cases} -q & , q \leq 0, q + |p| \geq 0 \\ |p| - 2\sqrt{-|p|q} & , q \leq 0, q + |p| \leq 0 \\ +\infty & , q > 0 \end{cases}$$



X is closed,vRec$(X) = \{[0, y] : y \geq 0\}$

$\mathrm{Dom}\,\mathrm{Supp}_X = \{[p; q] : q < 0\} \cup \{[0, 0]\}$ is not closed

$$\mathrm{Supp}_X(p,q) = \begin{cases} -\frac{p^2}{4q} & , q < 0 \\ 0 & , p = q = 0 \\ +\infty & , \text{otherwise} \end{cases}$$

Note:

Fact VII.13 *The domain of the support function* Supp_X *of a nonempty polyhedral set* X *is a polyhedral cone, and the epigraph of the support function is polyhedral..*

Indeed, when a nonempty X is given by polyhedral representation

$$X = \{x \in \mathbf{R}^n : \exists u : Px + Qy \leq r\},$$

one has

$$\begin{array}{rcl} \mathrm{Supp}_X(y) \leq t & \Leftrightarrow & \max_{x\in X} y^T x \leq t \\ & \Leftrightarrow & \max_{x,u} \left\{ y^T x : Px + Qu \leq r \right\} \leq t \\ & \Leftrightarrow & \min_\lambda \{r^T\lambda : P^T\lambda = y, Q^T\lambda = 0, \lambda \geq 0\} \leq t \\ & & \text{[by LP duality]} \\ & \Leftrightarrow & \exists \lambda : r^T\lambda \leq t, P^T\lambda = y, Q^T\lambda = 0, \lambda \geq 0 \end{array}$$

and we end up with explicit polyhedral representation of $\mathrm{Epi}\{\mathrm{Supp}_X\}$ given by homogeneous linear inequalities (and thus being a polyhedral cone).

♠ **Subdifferential of the support function.** Let $X \subset \mathbf{R}^n$ be a nonempty closed convex set. As Supp_X is the Legendre transform of proper convex semicontinuous function Υ_X, Fact VII.6.B states that

$$y \in \mathrm{Dom}\,\mathrm{Supp}_X \Rightarrow \partial \mathrm{Supp}_X(y) = \mathop{\mathrm{Argmax}}_{x \in X} x^T y$$

In particular,

$$\partial \mathrm{Supp}_X(0) = X$$

and for $y \in \mathrm{Dom}\,\mathrm{Supp}_X \backslash \{0\}$, $\partial \mathrm{Supp}(ty) = \partial \mathrm{Supp}_X(y)$ for all $t > 0$ and, in addition,

$$\partial \mathrm{Supp}_X(y) = \{g \in \partial \mathrm{Supp}_X(0) : g^T y = \mathrm{Supp}_X(y).\}$$

Examples:

- The support function of the unit $\|\cdot\|_2$-ball is $\|\cdot\|_2$-norm; $\|\cdot\|_2$, and

$$\partial \|x\|_2 = \begin{cases} \{\xi : \|\xi\|_2 \le 1\} & , x = 0, \\ \{x\} = \{\nabla \|x\|_2\} & , x \neq 0 \end{cases}$$

- The support function of the unit $\|\cdot\|_1$-ball is $\|\cdot\|_\infty$-norm;,$\|\cdot\|_\infty$, and

$$\partial \|x\|_\infty = \left\{ \xi : \xi_i \begin{cases} = \mathrm{sign}(x_i) & , x_i \neq 0 \\ \in [-1, 1] & , x_i = 0 \end{cases} \right\}.$$

- The support function of the unit $\|\cdot\|_\infty$-ball is $\|\cdot\|_1$-norm; $\|\cdot\|_1$, and

$$\partial \|x\|_1 = \begin{cases} \{\xi : \|\xi\|_\infty \le 1\} & , x = 0, \\ \left\{ \xi : \xi_i = \begin{cases} \mathrm{sign}(x_i) & , |x_i| = \|x\|_\infty \\ 0 & , |x_i| < \|x\|_\infty \end{cases} \right\} & , x \neq 0 \end{cases}$$

Minkowski function

♣ Let $X \subset \mathbf{R}^n$ be a *closed* convex set *containing the origin*.

♠ The **Minkowski function** $\mathfrak{M}_X(\cdot)$ is defined as

$$\mathfrak{M}_X(x) = \inf_{t>0} \left\{ t : t^{-1}x \in X \right\} : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\} \qquad [\inf_{t\in\emptyset} t = +\infty]$$

Example: *The Minkowski function of the unit ball $\{x : \|x\| \le 1\}$ of a norm $\|\cdot\|$ is this norm.*

• Clearly, *The domain of Minkowski function is the radial cone of X taken at the origin:* $\mathrm{Dom}\,\mathfrak{M}_X = \mathrm{Cone}\,(X)$, *and in this domain the function is nonnegative and positively homogeneous of degree 1:*

$$\alpha \ge 0, x \in \mathrm{Dom}\,\mathfrak{M}_X \Rightarrow \mathfrak{M}_X(\alpha x) = \alpha \mathfrak{M}(x) \ge 0.$$

Besides this, as $0 \in X$ and X is convex, for $\tau > 0$ the point $\tau^{-1}x$ belongs to X iff $t^{-1}x \in X$ for all $t \in (0, \tau]$. As a result,

• $\mathfrak{M}_X(x) = 0$ iff $tx \in X$ for all $t \ge 0$, that is, *iff* $x \in \mathrm{Rec}(X)$.

♠ To understand the geometry of the Minkowski function, note that a point $[x; t]$ belongs to $\mathrm{Epi}\{\mathfrak{M}_X\}$ iff $x/\tau \in X$ for all $\tau > t$. As X is closed, we conclude that when $t > 0$, $[x; t] \in \mathrm{Epi}\{\mathfrak{M}_X\}$ iff $x/t \in X$. Besides this, as we have seen, $[x; 0] \in \mathrm{Epi}\{\mathfrak{M}_X\}$ iff $x \in \mathrm{Rec}(X)$, and $\mathrm{Epi}\{\mathfrak{M}_X\}$ lives in the half-space $\{[x; t] : t \ge 0\}$. The bottom line is that

• $\mathrm{Epi}\{\mathfrak{M}_X\}$ *is exactly the closed conic transform* $\overline{\mathrm{ConeT}}(X) = \mathrm{cl}\,\mathrm{Cone}\,(X \times \{1\})$ *of the closed convex set* X.

As a result,

Fact VII.14 *The Minkowski function $\mathfrak{M}_X$ of a closed convex set $X \ni 0$ is a nonnegative proper convex lsc function, positively homogeneous of degree 1, and its sublevel set $\{x : \mathfrak{M}_X(x) \le 1\}$ is exactly X.*

The last claim is due to the fact that

$$\{x : \mathfrak{M}_X(x) \le 1\} = \{x : [x; 1] \in \mathrm{Epi}\{\mathfrak{M}_X\}\} = \{x : [x; 1] \in \overline{\mathrm{ConeT}}(X)\} = \mathrm{cl}\,X = X.$$

♠ **Question:** What is the Legendre transform of $\mathfrak{M}_X$?
Answer: For a positively homogeneous, of degree 1, nonnegative proper convex lsc function f, setting $X = \{x : f(x) \leq 1)$, we have

$$\begin{array}{rl} f_*(y) & = \sup_x[y^T x - f(x)] = \sup_{t\geq 0,x}[y^T[tx] - f(tx)] \\ & = \sup_{t\geq 0,x} t[y^T x - f(x)] \\ & = \begin{cases} +\infty & , \sup_x[y^T x - f(x)] > 0 \\ 0 & , y^T x \leq f(x)\, \forall x \end{cases} \\ & = \begin{cases} +\infty & , \exists (x, f(x) \leq 1) : y^T x > 1 \\ 0 & , y^T x \leq 1 \forall (x : f(x) \leq 1) \end{cases} \\ & = \Upsilon_{\text{Polar}\,(X)}(y). \end{array}$$

In words:

Fact VII.15 *The Legendre transform of a nonnegative proper convex homogeneous function f is the characteristic function of the polar of the sublevel set $\{x : f(x) \leq 1\}$ of f.*

• As a result,

Fact VII.16 *The Legendre transform of the Minkowski function $\mathfrak{M}_X$ of a closed convex set $X \ni 0$ is the characteristic function of the polar* Polar (X) *of X.*

♠ Combining the last two facts, we arrive at the following conclusion:

Fact VII.17 *Every nonnegative proper lsc positively homogeneous of degree 1 function f is a Minkowski function, specifically, the Minkowski function of the closed convex and containing the origin set $X = \{x : f(x) \leq 1\}$ (X is indeed convex and closed, as a sublevel set of a convex lsc function).*

♣ Let $X \subset \mathbf{R}^n$ be a closed convex set containing the origin, and let

$$X_* := \text{Polar}\,(X) = \{u : u^T x \leq 1 \, \forall u \in X\}$$

♠ **Domain and subdifferential of Minkowski function $\mathfrak{M}_X$.** We already know that the domain of $\mathfrak{M}_X$ is the radial cone $\text{Cone}\,(X)$ taken at the origin,

Subdifferential of $\mathfrak{M}_X$ at a point $x \in \text{Dom}\,\mathfrak{M}_X$ is readily given by Fact VII.6 in view of the fact that $\mathfrak{M}_X$ is the Legendre transform of the characteristic function of X_*, and we know the structure of subdifferentials of support functions. Specifically,

- *The subdifferential of $\mathfrak{M}_X$ taken at the origin is the polar X_* of X:*

$$\partial \mathfrak{M}_X(0) = X_*$$

- *At a point $x \neq 0$ from $\text{Dom}\,\mathfrak{M}_X$ the subdifferential is*

$$\partial \mathfrak{M}_X(x) = \underset{y}{\text{Argmax}} \left\{y^T x : y \in X_*\right\}$$

When $x \neq 0$ and $t > 0$, one has

$$\partial \mathfrak{M}_X(tx) = \partial \mathfrak{M}_X(x) = \{g \in X_* : g^T x = \mathfrak{M}_X(x)\}$$

♠ **Minkowski function of a polyhedral set.** Let X be a polyhedral set containing the origin. As we remember, the polar X_* of X is polyhedral as well, and the support function of X_* – which is nothing but the Minkowski function of X is polyhedral. Assuming that X is given by polyhedral representation

$$X = \{x : \exists u : Px + Qu \leq r\},$$

let us compute "from scratch" a polyhedral representation of $\mathfrak{M}_X$.

♣ First, let us convert representation of X into representation of X_*:

$$\begin{array}{l} X_* = \{y : y^T x \leq 1 \, \forall x \in X\} = \{y : \max_{x \in X} y^T x \leq 1\} = \{y : \max_{x,u} \left\{y^T x : Px + Qu \leq r\right\} \leq 1\} \\ = \{y : \min_\lambda \{r^T \lambda : P^T \lambda = y, Q^\lambda = 0, \lambda \geq 0\} \leq 1\} \text{ [LP duality]} \\ = \{y : \exists \lambda : r^T \lambda \leq 1, P^T \lambda = y, Q^\lambda = 0, \lambda \geq 0\} \end{array}$$

• Now we are ready to compute a polyhedral representation of $\mathfrak{M}_X$ (recall that it is the Legendre transform of Υ_{X_*}):

$$\begin{array}{rcl} \mathfrak{M}_X(x) \leq t & \Leftrightarrow & \max_{y \in X_*} x^T y :\leq t \\ & \Leftrightarrow & \max_y \{y^T x : \exists \lambda : P^t \lambda = y, Q^T \lambda = 0, r^T \lambda \leq 1\} \leq t \\ & \Leftrightarrow & \max_{y,\lambda} \{y^T x : P^T \lambda = y, Q^T \lambda = 0, r^T \lambda \leq 1\} \leq t \\ & \Leftrightarrow & \min_{\mu,\nu,\sigma,\gamma} \{\sigma : P\mu + Q\nu + \sigma r - \gamma = 0, -\mu = x\} \leq t \\ & & \text{[LP duality: } \mu, \nu, \gamma(\gamma \geq 0), \sigma(\sigma \geq 0), \text{ are Lagrange multipliers} \\ & & \text{for the respective constraints } P^T \lambda = 0, Q^T \lambda = 0, \lambda \geq 0, r^T \lambda \leq 1] \\ & \Leftrightarrow & \min_{\mu,\nu,\sigma,\gamma} \{: \sigma, : P\mu + Q\nu + \sigma r - \gamma = 0, -\mu = x\} \leq t \\ & \Leftrightarrow & \min\limits_{\nu,\sigma} \{\sigma : Q\nu + \sigma r - Px \geq 0, \sigma \geq 0r\} \leq t \\ & \Leftrightarrow & \exists \sigma, \upsilon : 0 \leq \sigma \leq t, Px + Q\upsilon \leq \sigma r \end{array}$$

and we end up with polyhedral representation of $\mathfrak{M}_X$.

$$\mathfrak{M}_X(x) \leq t \Leftrightarrow \exists \sigma, v : 0 \leq \sigma \leq t, Px + Qv \leq \sigma r$$

Note: We could arrive at the above result without computations, just after some brief thought .The point of the above derivation is to demonstrate that our Calculus of convexity allows to get meaningful results *without any thought at all.*
Here are somehow relevant words of Gottfried Wilhelm Leibniz (1646-1716), great philosopher and one of the founders of Calculus:

> *The only way to rectify our reasonings is to make them as tangible as those of the Mathematicians, so that we can find our error at a glance, and when there are disputes among persons, we can simply say: Let us calculate, without further ado, to see who is right.*

Question: When the Minkowski function of a closed convex set $X \ni 0$ is real-valued?
Answer: This is the case iff $\mathrm{Dom}\,\mathfrak{M}_X = \mathrm{Cone}\,(X) = \mathbf{R}^n$, that is, *iff* $0 \in \mathrm{int}\,X$. *In this case,* $\mathfrak{M}_X$ *is a nonnegative convex positively homogeneous of degree 1 real-valued function on* $\mathbf{R}^n$. By Fact VII.17 the inverse is also true: *Every nonnegative convex real-valued function on* $\mathbf{R}^n$ *is the Minkowski function of a closed convex set* X *such that* $0 \in \mathrm{int}\,X$.

• A positively homogeneous, of degree 1, real-valued function on $\mathbf{R}^n$ is convex iff it is *subadditive*, i.e., iff

$$f(x+y) \le f(x) + f(y)\ \forall x, y$$

Indeed, for positively homogeneous of degree 1 real-valued f and $\lambda \in (0,1)$ the inequality $f(\lambda x + (1-\lambda y) \le \lambda f(x) + (1-\lambda) f(y)$ is exactly the same as $f(u+v) \le f(u) + f(v)$, $u = \lambda x, v = (1-\lambda)y$; as x, y run through $\mathbf{R}^n$, so do u, v. We see that *the Minkowski function* $\mathfrak{M}_X$ *of a closed convex set* X, $0 \in \mathrm{int}\,X$, *inherits some properties of a norm, specifically it is continuous* (as every real-valued convex function on $\mathbf{R}^n$) *and*

- *it is nonnegative and positively homogeneous, of degree 1,*
- *satisfies the Triangle Inequality* $\mathfrak{M}_X(x+y) \le \mathfrak{M}_X(x) + \mathfrak{M}_X(y)$, *and*
- *has* X *as its unit ball:* $X = \{x : \mathfrak{M}_X(x) \le 1\}$

What is missing, is *symmetry:* $\mathfrak{M}_X(x) = \mathfrak{M}_X(-x)$ (it takes place iff $X = -X$) and *positivity:* $\mathfrak{M}_X(x) > 0$ whenever $x \ne 0$. To ensure positivity, we need $\mathrm{Rec}(X) = \{0\}$, which for closed convex X is the same as the boundedness of X.

♠ As a byproduct of our considerations, we can pay out our long-standing debt:

Fact VII.18 *When* $1 \le p \le \infty$, *the function* $\|\cdot\|_p$ *on* $\mathbf{R}^n$ *is a norm.*

Indeed, when $p < \infty$, $\|\cdot\|_p$ is the Minkowski function of the set $V = \{x \in \mathbf{R}^n : \sum_i |x_i|^p \le 1\}$, which is the sublevel set of convex continuous function, so that V is closed and convex; the facts that $V = -V$ and $0 \in \mathrm{int}\,V$ are evident. The same reasoning works for $p = \infty$, with the unit box $\{x : |x_i| \le 1, i \le n\}$ in the role of V.

Useful Inequalities – Young, Hölder, Moment

♣ **Young's Inequality.** *When $p \in (1,\infty)$ and $q = \frac{p}{p-1}$ (that is, $\frac{1}{p} + \frac{1}{q} = 1$), one has*

$$xy \le \frac{|x|^p}{p} + \frac{|y|^q}{q},$$

with inequality being equality iff $y = |x|^{p-1}\mathrm{sign}(x)$, whence also $x = |y|^{q-1}\mathrm{sign}(y)$.

Indeed, direct computation shows that if $f(x) = |x|^p/p$, then $f_*(y) = |y|^q/q$.

Extension: *Let $p_1, ..., p_k$ be positive reals such that*

$$\frac{1}{p_1} + ... + \frac{1}{p_k} = 1.$$

Then

$$|x_1 x_2 ... x_k| \le \frac{|x_1|^{p_1}}{p_1} + ... + \frac{|x_k|^{p_k}}{p_k}. \qquad (*)$$

Indeed, one can extract $(*)$ from Young's Inequality by induction or, much easier, note that $(*)$ is trivially true when some of x_i are zeros; when all x_i are nonzero, setting $\xi_i = \ln(|x_i|)$, we have

$$|x_1 x_2 ... x_k| = \exp\{\xi_1 + ... + \xi_k\} = \exp\{\frac{1}{p_1}[p_1\xi_i] + ... + \frac{1}{p_k}[p_k\xi_k]\} \le \frac{1}{p_1}\exp\{p_1\xi_1\} + ... + \frac{1}{p_k}\exp\{p_k\xi_k\} = \frac{|x_1|^{p_1}}{p_1} + ... + \frac{|x_k|^{p_k}}{p_k}$$

where $\le$ is by convexity of the exponent and due to $\frac{1}{p_i} > 0$ combined with $\frac{1}{p_1} + ... + \frac{1}{p_k} = 1$.

$$p_\ell > 0,\ \frac{1}{p_1} + ... + \frac{1}{p_k} = 1 \Rightarrow |x_1 x_2 ... x_k| \le \frac{|x_1|^{p_1}}{p_1} + ... + \frac{|x_k|^{p_k}}{p_k}. \qquad (*)$$

♣ **Weighted ℓ_p norms.** Let $w_1, ..., w_n$ be positive weights, and let $p \in [1, \infty]$. *Weighted ℓ_p-norm on $\mathbf{R}^n$* is defined as

$$\|x\|_{w,p} = \begin{cases} \left(\sum_i w_i |x_i|^p\right)^{1/p} & , 1 \le p < \infty \\ \max_i |x_i| = \lim_{p' \to \infty} \|x\| & , p = \infty \end{cases}$$

The usual $\|\cdot\|_p$ is weighted ℓ_p-norm with unit weights.

Notation: In the sequel, for vectors $x, x^1, .., x^k \in \mathbf{R}^n$ and $\alpha > 0$ we denote by

- $|x|$ – the vector $[|x_1|; ...; |x_n|]$ of magnitudes of entries in x,
- $|x|^\alpha$ – the vector $[|x_1|^\alpha; ...; |x_n|^\alpha]$, the entrywise α-power of $|x|$,
- $x^1 \cdot ... \cdot x^k$ – the entrywise product of x^ℓ: $[x^1 \cdot ... \cdot x^k]_i = x_i^1 ... x_i^k$, $1 \le i \le n$.

♠ Let $p_1, ..., p_k$ be positive reals such that $\frac{1}{p_1} + ... + \frac{1}{p_k} = 1$. By $(*)$ for every i we have

$$|w_i [x^1 \cdot ... \cdot x^k]_i| = [w_i^{\frac{1}{p_1}} |x_i^1|][w_i^{\frac{1}{p_2}} |x_i^2|]...[w_i^{\frac{1}{p_k}} |x_i^k|] \le \sum_{\ell=1}^k \frac{w_i |x_i^\ell|^{p_\ell}}{p_\ell};$$

summing over i, we get

$$\|x^1 \cdot ... \cdot x_i^k\|_{w,1} \le \sum_{\ell=1}^k \frac{\|x_i^\ell\|_{w,p_\ell}^{p_\ell}}{p_\ell}. \qquad (**)$$

$$\|x^1 \cdot ... \cdot x^k\|_{w,1} \leq \sum_{\ell=1}^{k} \frac{\|x^\ell\|_{w,p_\ell}^{p_\ell}}{p_\ell}. \qquad (**)$$

As an immediate consequence, we arrive at

♣ **Hölder Inequality.** *Let p_ℓ be positive reals such that $\frac{1}{p_1} + ... + \frac{1}{p_k} = 1$. Then for any k vectors x^ℓ, $\ell \leq k$, it holds*

$$\|x^1 \cdot ... \cdot x^k\|_{w,1} \leq \|x^1\|_{w,p_1}\|x^2\|_{w,p_2}...\|x^k\|_{w,p_k} \qquad (!)$$

Indeed, (!) is trivially true when some of x^ℓ are zero. Assuming that it is not the case, note that both sides in (!) are of homogeneity degree 1 w.r.t. every one of x^ℓ: when multiplying x^ℓ by θ, both sides in (!) are multiplied by $|\theta|$. As a result, to prove (!) when all x^ℓ are nonzero is the same as to prove the relation when $\|x^\ell\|_{w,p_\ell}$=1 for all ℓ, that is, to verify that

$$\|x^\ell\|_{w,p_\ell} = 1 \, \forall \ell \Rightarrow \|x^1 \cdot ... \cdot x^k\|_{w,1} \leq 1, \qquad (\#)$$

which is immediate: by $(**)$, we have

$$\|x^1 \cdot ... \cdot x^k\|_{w,1} \leq \sum_{\ell} \frac{\|x^\ell\|_{w,p_\ell}^{p_\ell}}{p_\ell}$$

and under the premise of (#) the right hand side in the latter inequality is 1.

Remark. It is immediately seen that (!) *holds true when $p_\ell = \infty$ for some ℓ, the corresponding terms $\frac{1}{p_\ell}$ in the condition $\frac{1}{p_1} + ... + \frac{1}{p_\ell}$ being set to 0.*

Illustration: *Let the weights w_i sum up to 1. Then*

$$1 \le p \le r \le \infty \Rightarrow \|x\|_p \le \|x\|_r. \qquad (*)$$

Indeed, assuming $1 \le p < r < \infty$ and setting $\alpha = \frac{r}{p}$, , $\beta = \frac{\alpha}{\alpha-1}$, $|x|^\pi = [|x_1|^\pi; ...; |x_n|^\pi]$ and $\mathbf{1} = [1; ...; 1]$, we have by Hölder Inequality

$$\|x\|_{w,p}^p = \|[|x|^p] \cdot \mathbf{1}\|_{w,1} \le \||x|^p\|_{w,\alpha} \underbrace{\|\mathbf{1}\|_{w,\beta}^{1/\beta}}_{=1} = \||x|^{p\alpha}\|_{w,1}^{1/\alpha}, = \|x\|_{w,r}^{p/r}$$

$\Rightarrow \|x\|_{w,p} \le \|x\|_{w,r}$, as claimed in $(*)$. Thus, the conclusion in $(*)$ holds true whenever $1 \le p < r < \infty$, and by continuity – whenever $1 \le p \le r \le \infty$.

Note: With unit weights, the dependence of $\|\cdot\|_p$ on p is completely opposite:

$$1 \le p \le r \le \infty, x \in \mathbf{R}^n \Rightarrow \|x\|_q \le \|x\|_p \le n^{\frac{1}{p}-\frac{1}{r}}\|x\|_r.$$

Indeed, by continuity in r, p, it suffices to verify the conclusion when $1 \le p < r < \infty$, and by homogeneity in x – when $\|x\|_p = 1$. In this case, $|x_i| \le 1$ for all i, which combines with $r \ge p$ to imply that $|x_i|^r \le |x_i|^p$, whence $\|x\|_r^r \le \|x\|_p^p = 1$, that is, $\|x\|_r \le \|x\|_p = 1$. Thus, $\|x\|_r \le \|x\|_p$. To upper-bound $\|x\|_p$ via $\|x\|_r$, set $\alpha = \frac{r}{p} \in [1, \infty)$ and $\beta = \frac{\alpha}{\alpha-1} = \frac{r}{r-p}$. By Hölder Inequality,

$$\begin{array}{l} \|x\|_p^p = \||x|^p\|_1 = \|[|x|^p] \cdot \mathbf{1}\|_1 \le \||x|^p\|_\alpha \|\mathbf{1}\|_\beta = \||x|^{p\alpha}\|_1^{1/\alpha} n^{1/\beta} = \|x\|_{p\alpha}^p n^{\frac{r-p}{r}} = \|x\|_r^p n^{1-\frac{p}{r}} \\ \Rightarrow \|x\|_p \le \|x\|_r n^{\frac{r-p}{pr}} \\ \Rightarrow \|x\|_p \le n^{\frac{1}{p}-\frac{1}{r}}\|x\|_q \end{array} .$$

♣ **Moment Inequality.** *Let $a \in \mathbf{R}^n$ be a nonzero vector, and $w = (w_1, ..., w_n)$ be a collection of positive weights. The function*

$$f(\rho) = \ln(\|a\|_{w,1/\rho}) : [0,1] \to \mathbf{R}$$

is convex. In other words, when $0 \le \rho \le \sigma \le 1$ and $\lambda \in [0,1]$, one has

$$\tau = \lambda\rho + (1-\lambda)\sigma \Rightarrow \|a\|_{w,1/\tau} \le \|a\|^{\lambda}_{w,1/\rho}\|a\|^{1-\lambda.}_{w,1/\sigma} \qquad (*)$$

Indeed, by continuity, it suffices to verify $(*)$ when $0 < \rho < \sigma < 1$ and $\lambda \in (0,1)$. In this case, setting

$$x^1 = |a|^{\lambda/\tau},\ x^2 = |a|^{(1-\lambda)/\tau},\ \alpha = \lambda\rho/\tau,\ \beta = (1-\lambda)\sigma/\tau, \qquad (1)$$

so that

$$|a|^{1/\tau} = x^1 \cdot x^2 \ \&\ \alpha, \beta > 0. \alpha + \beta = 1,$$

we have

$$\begin{aligned}
&\|a\|^{1/\tau}_{w,1/\tau} = \||a|^{1/\tau}\|_{w,1} = \|x^1 \cdot x^2\|_{w,1} \\
&\le \|x^1\|^{\alpha}_{w,1/\alpha}\|x^2\|^{\beta}_{w,1/\beta} \text{ [by Hölder Inequality]} \\
&= \||a|^{\lambda/(\tau/\alpha)}\|^{\alpha}_{w,1}\||a|^{(1-\lambda)/(\tau\beta)}\|^{\beta}_{w,1} \text{ [see (1)]} \\
&= \||a|^{1/\rho}\|^{\lambda\rho/\tau}_{w,1}\||a|^{1/\sigma}\|^{(1-\lambda)\sigma/\tau}_{w,1} \text{ [see (1)]} \\
&= \|a\|^{\lambda/\tau}_{w,1/\rho}\|a\|^{(1-\lambda)/\tau}_{w,1/\sigma} \\
&\Rightarrow \|a\|^{1/\tau}_{w,1/\tau} \le \|a\|^{\lambda/\tau}_{w,1/\rho}\|a\|^{(1-\lambda)/\tau}_{w,1/\overline{\tau}} \\
&\Rightarrow \|a\|_{w,1/\tau} \le \|a\|^{\lambda}_{w,1/\rho}\|a\|^{1-\lambda}_{w,1/\sigma}, \quad \text{as claimed.}
\end{aligned}$$

Application: Conjugates of ℓ_p norms, $1 \leq p$

♣ We have seen that the standard norms $\|\cdot\|_p$, $1 \leq p \leq \infty$, are norms. The weighted ℓ_p-norms are obtained from the standard ones by invertible linear transformation

$$\|x\|_{w,p} = \|Wx\|_p,\ W = \mathrm{Diag}\{w_1^{1/p}, ..., w_n^{1/p}\},\ p < \infty \ \&\ \|\cdot\|_{w,\infty} = \|\cdot\|_\infty.$$

and therefore are norms themselves.

Question: What is the norm conjugate to $\|\cdot\|_p$?

Answer: *The conjugate of $\|\cdot\|_p$ is the norm $\|\cdot\|_q$ with $q = \frac{p}{p-1}$, that is, with $\frac{1}{p} + \frac{1}{q} = 1$.*

Indeed, We know that the result is true when $p = 1$ or $p = \infty$. Now assume that $1 < p < \infty$. By Hölder Inequality, the weights being unit, for $x, y \in \mathbf{R}^n$ we have

$$|x^T y| = \|x \cdot y\|_1 \leq \|x\|_p \|y\|_q,$$

$\Rightarrow$ *the conjugate to $\|\cdot\|_p$ norm $\|y\|_p^* := \max_x\{y^T x : \|x\|_p \leq 1\}$ is $\leq \|\cdot\|_q$.* On the other hand, given y with $\|y\|_q = 1$, setting $x_i = |y_i|^{q-1}\mathrm{sign}(y_i)$, and taking into account that $p = \frac{q}{q-1}$, we have

$$\|x\|_p = (\sum_i |y_i|^{(q-1)p})^{1/p} = (\sum_i |y_i|^{[q-1]\frac{q}{q-1}})^{1/p} = 1 \ \&\ y^T x = \sum_i y_i[|y_i|^{q-1}\mathrm{sign}(y_i)] = \sum_i |y_i|^q = 1$$

$\Rightarrow \|y\|_p^* \geq y^T x = \|y\|_q$ whenever $\|y\|_q = 1$; by homogeneity, it follows that $\|\cdot\|_q \leq \|\cdot\|_p^*$, the bottom line being that $\|\cdot\|_p^* \equiv \|\cdot\|_q$, Q.E.D.

♠ Let $\|\cdot\|$ and $\|\cdot\|_*$ be a norm on $\mathbf{R}^n$, and its conjugate, and B, B_* be the unit balls of these norms, so that

$$B_* = \{y : y^T x \leq 1 \, \forall x \in B\} \Leftrightarrow B_* = \text{Polar}\,(B) \Leftrightarrow B = \text{Polar}\,(B_*) \Leftrightarrow \|x\| = \max_y \{y^T x : y \in B_*\} \Rightarrow (\|\cdot\|_*)_* = \|\cdot\|$$

Besides this,

$$\begin{array}{ll} \|\cdot\| = \mathfrak{M}_B(\cdot) & \Rightarrow \partial\Big|_{x=0}\|x\| = \text{Polar}\,(B) = B_* \\ \|\cdot\|_* = \mathfrak{M}_{B_*}(\cdot) & \Rightarrow \partial\Big|_{y=0}\|y\|_* = \text{Polar}\,(B_*) = B \end{array}$$

For example,

$$1 \leq p \leq \infty \Rightarrow \text{Polar}\,(\{x \in \mathbf{R}^n : \|x\|_p \leq 1\}) = \{y \in \mathbf{R}^n : \|y\|_q \leq 1\} \qquad [\tfrac{1}{p} + \tfrac{1}{q} = 1]$$

TTD illustration

♠ Let us look at the compliance $\mathrm{Compl}(t,f)$ of truss $t=[t_1;...;t_N]\in\mathbf{R}^N_+$ w.r.t. load $f\in\mathbf{R}^m$ as a function of t,f. By Fact II.30.ii, the epigraph of this function is the set

$$\mathrm{Epi}\{\mathrm{Compl}(t,f)\}:=\{[t;f;\tau]:\tau\geq\mathrm{Compl}(t,f),t\geq 0\}=\left\{[t;f;\tau]:\left[\begin{array}{c|c}B\mathrm{Diag}\{t\}B^T & f\\ \hline f^T & 2\tau\end{array}\right]\succeq 0,t\geq 0\right\}$$

where $M\times N$ matrix B is given by the geometry of nodal grid.

♠ From now on, we make the following assumption:

$$BB^T\succ 0$$

This assumption means that a *positive truss* $t>0$ *can withstand whatever load:* $\mathrm{Compl}(t,f)<\infty$ *for all* f.

Question: What can we say about the compliance?

A. The epigraph of $\mathrm{Compl}(t,f)$ is a closed convex cone $\Rightarrow$ $\mathrm{Compl}(t,f)$ *is a proper convex lsc function, positively homogeneous of degree 1 and even in* f. *It is the Minkowski function of the closed convex set*

$$X=\{[t,f]:t\geq 0,\left[\begin{array}{c|c}B\mathrm{Diag}\{t\}B^T & f\\ \hline f^T & 2\end{array}\right]\succeq 0\}$$

containing the origin.

B. Taking into account that for positive μ, ν the matrices

$$\left[\begin{array}{c|c} B\mathrm{Diag}\{t\}B^T & f \\ \hline f^T & 2\tau \end{array}\right] \quad \left[\begin{array}{c|c} B\mathrm{Diag}\{\mu t\}B^T & [\nu f] \\ \hline [\nu f]^T & 2\nu^2\tau/\mu \end{array}\right]$$

are/are not positive semidefinite simultaneously (since the second is obtained from the first by multiplication from the left and from the right by the nonsingular diagonal matrix $\mathrm{Diag}\{\sqrt{\mu}, ..., \sqrt{\mu}, \nu/\sqrt{\mu}\}$) we see that $\mathrm{Compl}(t, f)$ *is homogeneous, of degree 2, in* f*-variable and homogeneous, of degree -1, in* t*-variable:*

$$\mu > 0, \nu \in \mathbf{R} \Rightarrow \mathrm{Compl}(\mu t, \nu f) = \frac{\nu^2}{\mu}\mathrm{Compl}(t, f).$$

C. The domain of the compliance is a (not necessarily closed) cone – the radial cone of X taken at the origin. This cone definitely contains the open set

$$\{[0_{N\times 1}; 0_{M\times 1}]\} \cup \left\{\{t > 0\} \times \mathbf{R}_f^M\right\},$$

since a truss $t > 0$ can withstand any load. Which points $[t; f]$ with a non-strictly positive $t \geq 0$ belong, and which do not belong to $\mathrm{Dom}\,\mathrm{Compl}$, it depends on B.

$$\mathrm{Compl}(t,f) = \mathfrak{M}_X(t,f),\ X = \{[t,f] : t \geq 0,\ \left[\begin{array}{c|c} B\mathrm{Diag}\{t\}B^T & f \\ \hline f^T & 2 \end{array}\right] \succeq 0\}$$

D. What are the subdifferentials of Compl ? By our general theory, $\partial\mathrm{Compl}(0,0)$ is the polar of the set X, and subdifferential of Compl at a nonzero points are cut off the subdifferential $\mathrm{Polar}\,(X)$ at 0 by equality constraint

$$\partial\mathrm{Compl}(t,f) = \left\{[\alpha;\beta] \in \mathrm{Polar}\,(X) : \alpha^T t + \beta^T f = \mathrm{Compl}(t,f)\right\}.$$

Let us find $\mathrm{Polar}\,(X)$.

✓ First of all, we have

$$K := \mathrm{Epi}\{\mathrm{Compl}\} = \{[t;f;\tau] : t \geq 0,\ \left[\begin{array}{c|c} B\mathrm{Diag}\{t\}B^T & f \\ \hline f^T & 2\tau \end{array}\right] \succeq 0\}$$

Observe that K is closed convex cone which lives in the half-space $\mathbf{R}^{N+M}_{t,f} \times \mathbf{R}_+$. Setting $X = \{[t;f] : [t;f;1] \in K\}$, we get

$$K = \overline{\mathrm{ConeT}}(X)$$

by Fact II.18. Now, $[\alpha;\beta] \in \mathrm{Polar}\,(X)$ iff the linear form $[-\alpha;-\beta;1]^T z$ of z is nonnegative on the set $X \times \{1\} = \{z \in \mathrm{ConeT}(X) : z_{M+N+1} = 1\}$, or, which is the same, is nonnegative on $\mathrm{ConeT}(X)$, or, which again is the same, on $\overline{\mathrm{ConeT}}(X) = K$. The bottom line is that *A vector* $[\alpha;\beta]$ *is a subgradient of* Compl *at the origin iff the vector* $[-\alpha;-\beta;1]$ *belongs to the dual to* K *cone* K_*.

✓ To understand what K_* is, note that

$$K = \{[t; f; \tau] : \mathcal{A}[t; f; \tau] \in K^+\},$$
$$\mathcal{A}[t; f; \tau] = \left(\left[\begin{array}{c|c} B\mathrm{Diag}\{t\}B^T & f \\ \hline f^T & 2\tau \end{array}\right], t\right), \quad K^+ = \mathbf{S}^{M+1}_+ \times \mathbf{R}^N_+$$

and the image space of $\mathcal{A}$ intersects the interior of the closed convex cone K^+ (indeed, when $t = [1; ...; 1]$, $f = 0$, and $\tau = 1$, $\mathcal{A}[t; f; \tau] = \big(\mathrm{Diag}\{BB^T, 2\}, [1; ...; 1]\big)$ is an interior point of K^+). Next, equipping $\mathbf{S}^{M+1}$ with the *Frobenius inner product*

$$\langle A, B\rangle = \mathrm{Tr}(AB) = \sum_{i,j} A_{ij}B_{ij},$$

and equipping $\mathbf{S}^{M+1}$ with orthonormal w.r.t this inner product basis, we can identify matrices from $\mathbf{S}^{M+1}$ with vectors of their coefficients in this basis, thus identifying

— $\mathbf{S}^{M+1}$ with $\mathbf{R}^K$ ($K = \frac{M(M+1)}{2}$),

— the cone $\mathbf{S}^{M+1}_+$ (which is a regular cone in $\mathbf{S}^{M+1}$) – with certain regular cone S in $\mathbf{R}^K$,

— the space $\mathbf{S}^{M+1} \times \mathbf{R}^N$ where the cone K^+ lives with the space $\mathbf{R}^{K+N}$, the cone K^+ itself – with the cone $\widetilde{K}^+ = S \times \mathbf{R}^N_+ \subset \mathbf{R}^{K+N}$, and

— the linear mapping $[t; f; \tau] \mapsto \mathcal{A}[t; f; \tau]$ – with linear mapping $[t; f; \tau] \to A[t; f; \tau] \in \mathbf{R}^{K+N}$, where A is an appropriately selected matrix.

• It is known that *the semidefinite cone $\mathbf{S}^p_+$ is self-dual w.r.t. the Frobenius inner product:*

$$P \in \mathbf{S}^p, \mathrm{Tr}(PQ) \geq 0 \, \forall Q \in \mathbf{S}^p_+ \Leftrightarrow P \in \mathbf{S}^p_+.$$

Consequently, the cone $\widetilde{K}^+$ is self-dual:

$$\widetilde{K}^+_* = \widetilde{K}^+$$

(as the product of two self-dual cones), and $\mathrm{Im}\,\mathcal{A} \cap \mathrm{int}\, K^+ \neq \emptyset$ implies that $\mathrm{Im}\, A \cap \mathrm{int}\, \widetilde{K}^+_* \neq \emptyset$. The latter, by Remark in Fact IV.20.F, implies the first of the equalities to follow:

$$K_* \equiv [A^{-1}\widetilde{K}^+]_* = A^T\widetilde{K}^+_* = A^T\widetilde{K}^+,$$

This translates to

$$[\alpha; \beta; \sigma] \in K_* \Leftrightarrow \exists \left(\gamma \in \mathbf{R}^N_+, \left[\begin{array}{c|c} P & q \\ \hline q^T & r \end{array}\right] \in \mathbf{S}^{M+1}_+\right) : \left\{\begin{array}{r} \alpha_i = \gamma_i + [B^TPB]_{ii} \\ 1 \leq i \leq N \\ \beta = 2q, \ \sigma = 2r \end{array}\right.$$

✓ Combining the above observations, we arrive at the following result:

$$\begin{array}{rcl}
\partial\mathrm{Compl}(0,0) & = & \{[\alpha;\beta] \in \mathbf{R}^N_\alpha \times \mathbf{R}^M_\beta : [-\alpha;-\beta;1] \in K_*\} \\
& = & \left\{[\alpha;\beta] \in \mathbf{R}^N \times \mathbf{R}^M : \exists P \in \mathbf{S}^M : \left[\begin{array}{c|c} P & -\beta/2 \\ \hline -\beta^T/2 & 1/2 \end{array}\right] \succeq 0, \alpha_i + [B^T P B]_{ii} \le 0,\ i = 1,...,N\right\}. \\
& = & \left\{[\alpha;\beta] : \exists P \in \mathbf{S}^M : \left[\begin{array}{c|c} 2P & \beta \\ \hline \beta^T & 1 \end{array}\right] \succeq 0, \alpha_i + [B^T P B]_{ii} \le 0,\ i = 1,...,N\right\}
\end{array}$$

(note that symmetric block matrices $\left[\begin{array}{c|c} P & Q \\ \hline Q^T & R \end{array}\right]$ and $\left[\begin{array}{c|c} P & -Q \\ \hline -Q^T & R \end{array}\right]$ are/are not positive semidefinite simultaneously (why?))

E. Now let us look at the optimal value in the TTD problem

$$\mathrm{Opt}(f,W)=\min_t\left\{\mathrm{Compl}(t,f):t\geq 0,\sum_i t_i=W\right\}$$

as a function of f,W. We intend to consider this function in the domain

$$\mathcal{FW}=\{[f;W]:W>0\}$$

– the largest domain where the function could be of interest. Setting $\mathcal{W}=\{[t;W]\in\mathbf{R}^N_+\times\mathbf{R}_+:W=\sum_i t_i>0\}$, we get a convex set, so that the function

$$\mathcal{C}(t;f;W)=\mathrm{Compl}(t,f)+\Upsilon_{\mathcal{W}}(t,W),$$

is convex (and of course, nonnegative) function such that

$$\mathrm{Opt}(f,W)=\inf_t\mathcal{C}(t,f,W)\qquad(*)$$

Since $\mathcal{C}(t,f,W)$ is convex and nonegative, Calculus of Convexity says that the function

$$\mathrm{Opt}(f;W):\mathbf{R}^M_f\times\mathbf{R}_W\to\mathbf{R}\cup\{+\infty\}$$

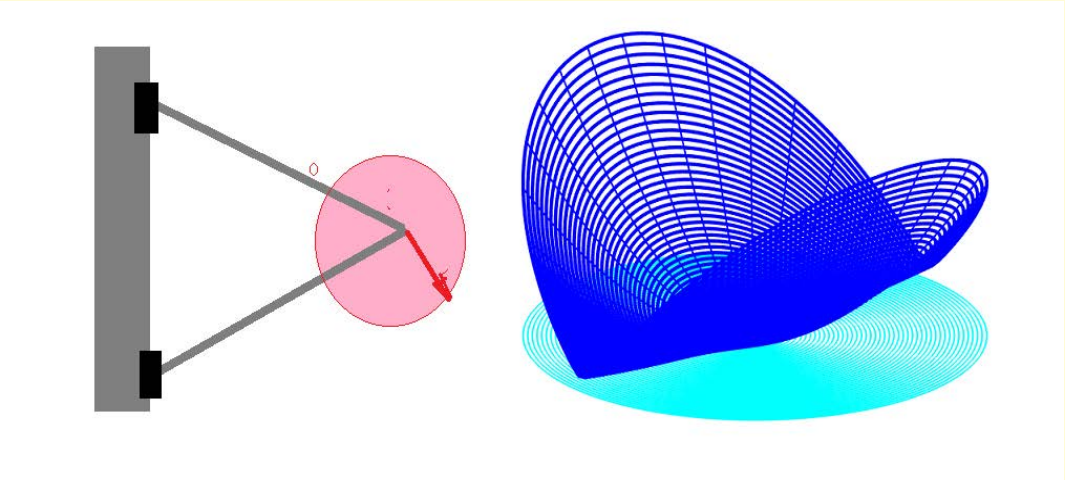
is convex; the domain of this function clearly is contained in $\mathcal{FW}$. In fact, the domain is *exactly* $\mathcal{FW}$. Indeed, due to $BB^T\succ 0$, the set

$$\mathcal{Z}_{f,W}=\{[t;\tau]:t\geq 0,\sum_i t_i=W,\tau\geq\mathrm{Compl}(t,f)\}$$

is nonempty, and since the function $\mathrm{Compl}(t,f)-\tau$ as a function of $t;\tau$ is lsc along with $\mathrm{Compl}(t,f)$, this set is also closed. The function $\mathcal{T}(t,\tau)\equiv\tau$ is continuous and *coercive* on $\mathcal{Z}_{f,W}$ – its sublevel sets $\{[t;\tau]\in\mathcal{Z}_{t,W}:\mathcal{T}(t,\tau)\leq q\}$ are compact for every real a, the continuous function $\mathcal{T}(t,\tau)$ attains its minimum on $\mathcal{Z}_{f,W}$ by the Weierstrass Theorem, implying that *the TTD problem $(*)$ is solvable whenever $W>0$*, so that $\mathrm{Opt}(t,W)$ is a real-valued convex function with the *open* domain $\mathcal{FW}$; as a result, *the function $\mathrm{Opt}(f,W)$ is real-valued nonnegative convex and continuous on $\mathcal{FT}$.*

♠ Additional useful information can be extracted from the homogeneity properties of Compl(t, f); these properties immediately imply that *the function* Opt(f, W) *is homogeneous, of homogeneity degree* -1*, in* W*-variable and homogeneous, of degree 2, in* f*-variable.*

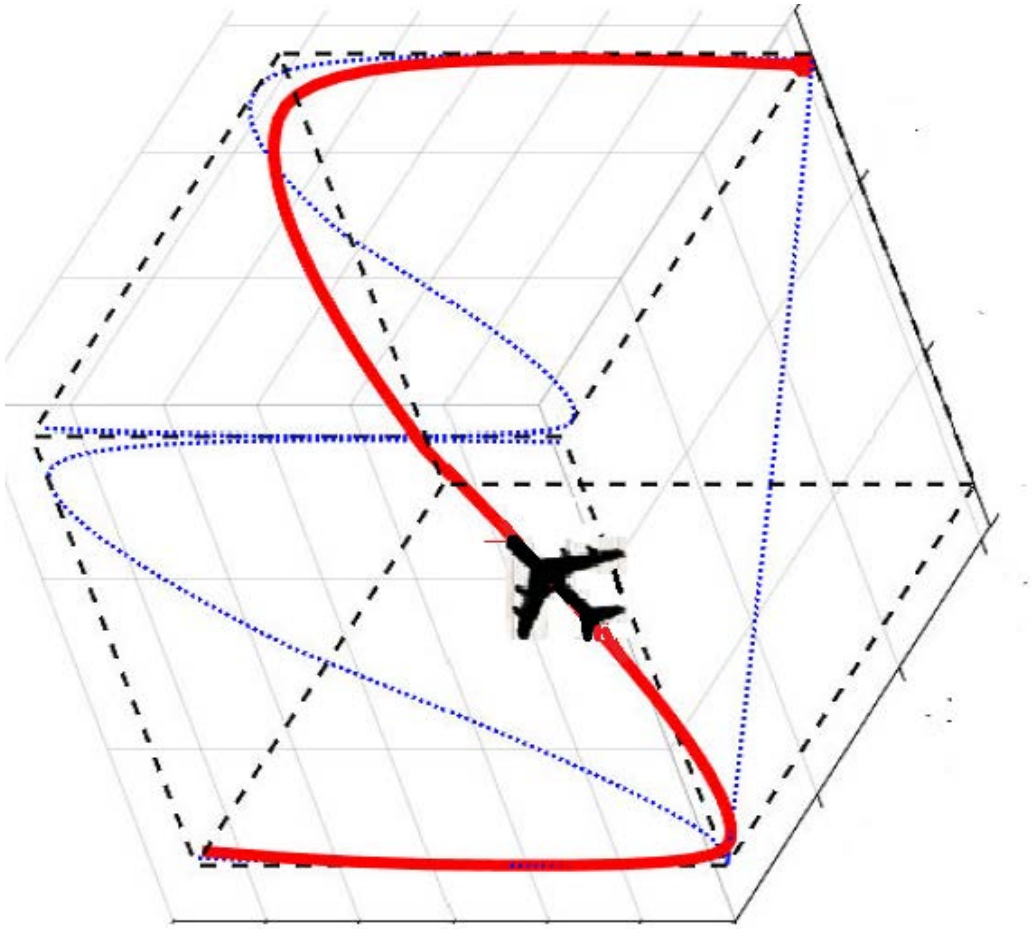
• We could further investigate the closure of Opt$(\cdot, \cdot)$, which is a homgeneous, of homogeneity degree 1, function of f, W, and thus is the Minkowski function of certain closed convex set containing the origin, identify this set and its polar, etc., etc., but enough is enough...



What you see is a toy planar ground structure (left) and the optimal compliance Opt$(f, 50)$ as a function of external force f running through the unit circle (right),

PART III.

Convex Programming



minimum acceleration maneuver
in restricted space

Lecture III.1
Convex Programming

Convex Programs in Mathematical Programming,
Cone-constrained and Conic Forms
Convex Theorems on Alternative
Lagrange Duality
Optimality conditions in Saddle Point and
Karush-Kuhn-Tucker forms

Mathematical Programming Problem

♣ Mathematical Programming problem (a.k.a *Mathematical Programming program*) reads:

$$\min_x \left\{ f(x) : \begin{array}{c} (g_1(x), g_2(x), ..., g_m(x)) \leq 0 \\ (h_1(x), ..., h_k(x)) = 0 \\ x \in X \end{array} \right\}. \qquad (P)$$

In this problem, f, g_i, h_j are real-valued functions on the problem's *domain* X which is a nonempty subset in some $\mathbf{R}^n$.

- $x \in \mathbf{R}^n$ is called *decision vector*, and its entries are called *decision variables*
- f is called the *objective*,and g_i, h_j are called *constraints* – inequality and equality constraints, respectively.

Note: The *relations* $g_i(x) \leq 0$, $h_j(x) = 0$ also are called *constraints*; It always will be clear from the context what "constraint" means under the circumstances – the relation or its left hand side.

- A *solution* to (P) is a whatever value of the decision vector. A solution is called *feasible*, if it satisfies all the inequality and equality constraints, same as the *domain constraint* $x \in X$. A solution which is not feasible is called *infeasible*.
- The set Feas(P) of all feasible solutions to (P) is called *the feasible set* of (P); if this set is nonempty, (P) is called *feasible*, otherwise the problem is called *infeasible*.

$$\mathrm{Opt}(P) = \min_x \left\{ f(x) : \begin{array}{c} (g_1(x), g_2(x), ..., g_m(x)) \leq 0 \\ (h_1(x), ..., h_k(x)) = 0 \\ x \in X \end{array} \right\}. \qquad (P)$$

• (P) is called *bounded*, if the objective is below bounded on the feasible set (e.g., due to the fact that the latter set is empty) and *unbounded* otherwise. The infimum of values on the objective on the feasible set is called *the optimal value* $\mathrm{Opt}(P)$ of problem (P):

$$\mathrm{Opt}(P) = \begin{cases} +\infty & , \mathrm{Feas}(P) = \emptyset \\ \inf\{f(x) : x \in \mathrm{Feas}(P)\} & , \text{otherwise} \end{cases}$$

Thus, $\mathrm{Opt}(P) = \infty$ when (P) is infeasible, $\mathrm{Opt}(P) = -\infty$ when the problem is unbounded;. When (P) is feasible and bounded, $\mathrm{Opt}(P)$ is a real a such that
— for every $\epsilon > 0$, there exists a feasible solution with $f(x) \leq a + \epsilon$, and
— there are no feasible solutions x with $f(x) < a$.

• A feasible solution x with $f(x) = \mathrm{Opt}(P)$ is called *an optimal solution* to the problem. (P) is called *solvable*, it the problem has optimal solutions, and is called *insolvable* otherwise.

Note: Solvable problem definitely is feasible and bounded, but not vice versa (look at the problem $\min_{x \in \mathbf{R}} \mathrm{e}^x$).

Convention: *Whenever $X = \mathbf{R}^n$, we take the liberty to omit writing the constraint $x \in \mathbf{R}^n$ explicitly. Thus, "by default" – whenever the domain X is not explicitly specified – it is the entire $\mathbf{R}^n$.*

♠ Our terminology is adjusted to minimization problems and needs modification when speaking about maximization problem

$$\mathrm{Opt}(P) = \max_x \left\{ f(x) : \begin{array}{c} (g_1(x), g_2(x), ..., g_m(x)) \leq 0 \\ (h_1(x), ..., h_k(x)) = 0 \\ x \in X \end{array} \right\}. \qquad (P)$$

Specifically, a maximization problem is called *bounded*, if its objective is bounded from *above* on the feasible set, and the optimal value in a maximization problem is the *supremum* of the values of the objective on the feasible set.

Convex problems in MP form

♣ A Mathematical Programming problem

$$\mathrm{Opt}(P) = \min_x \left\{ f(x) : \begin{array}{c} (g_1(x), g_2(x), ..., g_m(x)) \leq 0 \\ (h_1(x), ..., h_k(x)) = 0 \\ x \in X \end{array} \right\}. \qquad (P)$$

is called *convex*, if

- the domain X is a convex set
- the functions f, g_i are convex (and, as always, real-valued) on X, and
- all equality constraints are linear.

Clearly, the feasible set of a convex problem is convex.

Convex problems in Cone-constrained form

♣ The equality constraints in a *convex* MP problem

$$\mathrm{Opt}(P) = \min_x \left\{ f(x) : \begin{array}{c} (g_1(x), g_2(x), ..., g_m(x)) \leq 0 \\ (h_1(x), ..., h_k(x)) = 0 \\ x \in X \end{array} \right\}. \qquad (P)$$

are linear and just say that a feasible solution should satisfy a system of linear inequalities

$$Ax - b \leq 0 \qquad (*)$$

representing equivalently the system of linear equations $(h_1(x), ..., h_k(x)) = 0$.
The inequality constraints say that at a feasible solution, the *vector-valued function* $g(x) := [g_1(x); ...; g_m(x)]$ should take value in - the negation $-\mathbf{R}^m_+$ of the specific cone, the nonnegative orthant $\mathbf{R}^m_+$. Moreover, $g(x)$ is "adjusted" to this specific cone real-valuedness and convexity of g_i on X mean that $g(x)$ is well-defined on X and satisfies the relation

$$\forall (x, y \in X, \lambda \in [0,1]) : \lambda g(x) + (1-\lambda) g(y) - g(\lambda x + (1-\lambda) y) \in \mathbf{R}^m_+.$$

Thus, convex program (P) can be rewritten equivalently as

$$\mathrm{Opt}(P) = \min_x \{ f(x) : Ax - b \leq 0, g(x) \in -\mathbf{K}, x \in X \}$$

where the system of linear constraints $Ax - b \leq 0$ represents equivalently a system of linear equations, and $\mathbf{K}$ is a specific *regular* (i.e., closed, pointed, convex and with a nonempty interior) cone (namely, $\mathbf{R}^m_+$) adjusted to the *convex* domain X and vector-valued function g via the relation

$$\forall (x, y \in X, \lambda \in [0,1]) : \lambda g(x) + (1-\lambda) g(y) - g(\lambda x + (1-\lambda) y) \in \mathbf{K}.$$

It turns out that *As far as Convex Programming is concerned, it is highly rewarding to extend the MP formulation of a convex problem by allowing for* $(*)$ *to be a whatever system of linear inequalities, and for* $\mathbf{K}$ *– to be a regular cone*, not necessarily nonnegative orthant, *thus arriving at convex problems in cone-constrained form.*

Preliminaries, I: Vector inequalities

♣ In LP and MP, we all the time meet the relations like $a \geq b$ with *vectors* a, b. Although denoted exactly as the "arithmetic" $\geq$ – relation between reals, *coordinate-wise vector inequality* $a \geq b$ is a quite different beast: it means that a, b are vectors from some $\mathbf{R}^m$ such that $a - b$ belongs to a specific regular cone $\mathbf{R}^m_+$.

To see the difference between the "arithmetic" and the coordinate-wise vector $\geq$, note that the arithmetic $\geq$ is a *complete* order on $\mathbf{R}$ – for every two reals a, b, we either have $a \geq b$, or $b \geq a$, or both, while the order induced by the vector $\geq$ on $\mathbf{R}^m$, $m > 1$, is just partial: we cannot compare vectors $[0; 1]$ and $[1; 0]$ in $\mathbf{R}^2$!

♠ Given a regular (i.e., closed, pointed, and with a nonempty interior) cone $\mathbf{K} \subset \mathbf{R}^m$, we associate with it vector inequality $\geq_{\mathbf{K}}$ – relation between vectors from $\mathbf{R}^m$ given by

$$[b \leq_K a \Leftrightarrow] a \geq_{\mathbf{K}} b \;\; \Leftrightarrow a - b \in \mathbf{K}$$

Note: $\mathbf{K} = \{a : a \geq_{\mathbf{K}} 0\}$ is just the set of all $\mathbf{K}$-nonnegative vectors.

$$[b \le_K a \Leftrightarrow] a \ge_{\mathbf{K}} b \ \Leftrightarrow a - b \in \mathbf{K}$$

♠ Relation $\ge_{\mathbf{K}}$ shares all basic properties of the coordinate-wise $\ge$, specifically

- $\ge_{\mathbf{K}}$ is al partial order:

$$\begin{array}{ll} a \ge_{\mathbf{K}} a \, \forall a & \text{reflexivity} \\ a \ge_{\mathbf{K}} b \ \& \ b \ge_{\mathbf{K}} a \Rightarrow a = b & \text{antisymmetry} \\ a \ge_{\mathbf{K}} b, b \ge_{\mathbf{K}} c \Rightarrow a \ge_{\mathbf{K}} c & \text{transitivity} \end{array}$$

- $\ge_K$ is compatible with linear operations:

$$a \ge_{\mathbf{K}} b, c \ge_K d \Rightarrow a + c \ge_{\mathbf{K}} b + d$$
$$a \ge_{\mathbf{K}} b, \mathbf{R} \ni \lambda \ge 0 \Rightarrow \lambda a \ge_{\mathbf{K}} \lambda b$$

- One can pass to sidewise limits in $\ge_{\mathbf{K}}$:

$$a_i \ge_{\mathbf{K}} b_i,\ a_i \to a,\ b_i \to b \text{ as } i \to \infty \Rightarrow a \ge_{\mathbf{K}} b.$$

- We can define a strict version $a >_{\mathbf{K}} b$ of $\ge_{\mathbf{K}}$:

$$[b <_{\mathbf{K}} a \Leftrightarrow] a >_{\mathbf{K}} b \Leftrightarrow a - b \in \operatorname{int} \mathbf{K}$$

Elementary arithmetics of $\ge_{\mathbf{K}}$ and $>_{\mathbf{K}}$ is the same as for usual $\ge$: the inequalities of the same, up to strictness, type can be added, the result being strict if one of the operands is so, inequalities are preserved when multiplying both sides by a nonnegative real (up to the fact that $>_{\mathbf{K}}$ becomes $\ge_{\mathbf{K}}$ when the real is zero), strict inequality $a >_{\mathbf{K}} b$ is stable - it remains valid when a and b are subject to small enough perturbations, etc.

♠ **Note:** *Taking inner products of both sides in a valid vector inequality $a \ge_{\mathbf{K}} b$ with $\lambda \in \mathbf{K}_*$, we get a valid scalar inequality:*

$$a \ge_{\mathbf{K}} b,\ \lambda \in \mathbf{K}_* \Rightarrow \lambda^T a \ge \lambda^T b.$$

Preliminaries, II: Cone-convex functions

♣ **Cone-convex functions.** Let $Q \subset \mathbf{R}^n$ be a nonempty convex set, and $\mathbf{K} \subset \mathbf{R}^m$ be a regular cone. *A function $f : Q \to \mathbf{R}^m$ is called $\mathbf{K}$-convex on Q, if*

$$\forall (x, y \in Q, \lambda \in [0,1]) : f(\lambda x + (-\lambda)y) \leq_{\mathbf{K}} \lambda f(x) + (1-\lambda) f(y).$$

Examples: • When $m = 1$ and $\mathbf{K} = \mathbf{R}_+$, $\mathbf{K}$-convex functions are exactly the functions convex and real-valued on Q

• When $\mathbf{K} = \mathbf{R}^m_+$, $\mathbf{K}$-convex functions are exactly the vector-valued functions

$$f(x) = [f_1(x); ...; f_m(x)]$$

with convex real-valued on Q components $f_1, ..., f_m$

• The function $f(X) = X^T X : \mathbf{R}^{m \times n} \to \mathbf{S}^n$ is $\mathbf{S}^n_+$-convex

Indeed, when $X, Y \in \mathbf{R}^{m \times n}$ and $\lambda \in (0,1)$, we have

$$\begin{aligned} &\lambda X^T X + (1-\lambda) Y^T Y - [\lambda X + (1-\lambda) Y]^T [[\lambda X + (1-\lambda) Y] \\ &= \lambda(1-\lambda) X^T X + \lambda(1-\lambda) Y^T Y - \lambda(1-\lambda)[X^T Y + Y^T X] = \lambda(1-\lambda)[X-Y]^T [X-Y] \succeq 0. \end{aligned}$$

♠ **Immediate and crucial observation:**

Fact VIII.1 *A vector-valued function $f : Q \to \mathbf{R}^m$ is $\mathbf{K}$-convex on a convex set Q iff for every $\phi \in \mathbf{K}_*$ the real-valued function*

$$f_\phi(x) = \phi^T f(x)$$

is convex on Q. Thus, $\mathbf{K}$-convexity of a vector-valued function f is equivalent to plain convexity of certain family of real-valued functions associated with f

Indeed, as $\mathbf{K} = [\mathbf{K}_*]_*$, we have for $x, y \in Q$ and $\lambda \in [0,1]$:

$$\lambda f(x) + (1-\lambda) f(y) - f(\lambda x + *1 - \lambda) y) \in \mathbf{K} \Leftrightarrow \phi^T [\lambda f(x) + (1-\lambda) f(y) - f(\lambda x + *1 - \lambda) y)] \geq 0] \, \forall \phi \in \mathbf{K}_*$$

Due to the above observation, significant part of known to us facts about convex functions can be immediately extended to cone-convex ones:

♠ **Epigraph characterization of $\mathbf{K}$-convexity:** *A function $f : Q \to \mathbf{R}^m$ defined on a convex set $Q \subset \mathbf{R}^n$ is $\mathbf{K}$-convex iff its "$\mathbf{K}$-epigraph"*

$$\{(x,t) \in \mathbf{R}^n_x \times \mathbf{R}^m_y) : x \in Q, f(x) \leq_K t\}$$

is a convex set.

♠ **Calculus:**

• *An affine mapping $x \mapsto Ax + b : \mathbf{R}^n \to \mathbf{R}^m$ is $\mathbf{K}$-convex, for every regular cone $\mathbf{K} \subset \mathbf{R}^m$, on every nonempty convex set $Q \subset \mathbf{R}^n$.*

• *If $\lambda_i \geq 0$ and $f_i : Q \to \mathbf{R}^m$ are $\mathbf{K}$-convex, so is $\sum_i \lambda_i f_i$. Moreover, if $\mathbf{K}$ is a regular cone in $\mathbf{R}^k$, $\mathbf{K}_i$ are regular cones in $\mathbf{R}^{m_i}$, the functions $f_i : Q \to \mathbf{R}^{m_i}$ are $\mathbf{K}_i$-convex, and $m_i \times k$ matrices Λ_i are $(\mathbf{K}_i, \mathbf{K})$-nonnegative, meaning that $\Lambda_i \mathbf{K}_i \subset \mathbf{K}$, then the function $\sum_i \Lambda_i f_i : Q \to \mathbf{R}^k$ is $\mathbf{K}$-convex,*

• *If $\mathbf{K}_i \subset \mathbf{R}^{m_i}$, $i \leq I$, are regular cones, $Q_i \subset \mathbf{R}^{n_i}$ are convex sets, and functions $f_i : Q_i \to \mathbf{R}^{m_i}$ are $\mathbf{K}_i$-convex, the function*

$$f(x^1, ..., x^I) = [f_1(x^1); ...; f_I(x^I)] : Q_1 \times ... \times Q_I \to \mathbf{R}^{m_1 + ... + m_I}$$

is $(\mathbf{K} = \mathbf{K}_1 \times ... \times \mathbf{K}_I)$-convex on $Q = Q_1 \times ... \times Q_I$

• *If Q is a convex set in $\mathbf{R}^n$, $\mathbf{K}$ is a regular cone in $\mathbf{R}^m$, $f : Q \to \mathbf{R}^m$ is $\mathbf{K}$-convex and $y \to \mathcal{A}(y) = Ay + b$ is an affine mapping from $\mathbf{R}^p$ to $\mathbf{R}^n$ such that the set $\mathcal{A}^{-1}(Q) = \{y : \mathcal{A}(y) \in Q\}$ is nonempty, the function $f(\mathcal{A}(y))$ is $\mathbf{K}$-convex on $\mathcal{A}^{-1}(Q)$.*

But: Forget about taking pointwise maximum: since the order given by $\geq_{\mathbf{K}}$ is incomplete (unless $\mathbf{K}$ is one-dimensional), there is no such thing as the $\geq_{\mathbf{K}}$-maximum of two vectors.

♠ **Local regularity:** *A $\mathbf{K}$-convex function $f : Q \to \mathbf{R}^m$ is Lipschitz continuous on every compact set $X \subset \operatorname{rint} Q$.*

Indeed, as $\mathbf{K}$ is regular, so is $\mathbf{K}_*$; in particular, $\operatorname{int}\mathbf{K}_* \neq \emptyset$, thus we can find m linearly independent $\phi_\ell \in \mathbf{K}_*$, so that $p(z) = \max_\ell |\phi_\ell^T z|$ is a norm on $\mathbf{R}^m$. As $\phi_\ell^T f$ are convex real-valued functions on Q, their restrictions on X are Lipschitz continuous, with properly selected constant L_X, with respect to the norm $\|\cdot\|_2$ on the argument space $\mathbf{R}^n$, implying that $p(f(x) - f(y)) \leq L_X \|x - y\|_2$ for all $x, y \in X$.

♠ **"Gradient Inequality:"** *Let $Q \subset \mathbf{R}^n$ be a convex set, $f : Q \to \mathbf{R}^m$ be $\mathbf{K}$-convex and differentiable at a point $x \in Q$ function, and let $\mathcal{J}(x) \in \mathbf{R}^{m\times n}$ be the Jacobian of f at x, (i.e., the matrix of the linear mapping $h \mapsto Df(x)[h] : \mathbf{R}^n \to \mathbf{R}^m$). Then*

$$\forall y \in Q : f(y) \geq_{\mathbf{K}} f(x) + \mathcal{J}(x)[y - x] \qquad (*)$$

Indeed, for $\phi \in \mathbf{K}_*$ the real-valued on Q function $f_\phi(y) = \phi^T f(y)$ is convex and differentiable at x, with $\nabla f_\phi(x) = [\mathcal{J}(x)]^T \phi$, whence by the usual Gradient inequality

$$\forall y \in Q : \phi^T f(y) = f_\phi(y) \geq f_\phi(x) + [\nabla f_\phi(x)]^T [y - x] = \phi^T [f(x) + \mathcal{J}(x)[y - x]],$$

and since this relation holds true for all $\phi \in \mathbf{K}_*$, $(*)$ follows.

Example: We shall see eventually that the fractional-quadratic function

$$f(X, Y) = X^T Y^{-1} X : \mathbf{R}^{m\times n}_X \times \{Y \in \mathbf{S}^m : Y \succ 0\} \to \mathbf{S}^n$$

– the matrix analogy of the convex fractional-quadratic function $t^{-1} x^T x : \mathbf{R}^n_x \times \{t > 0\}$ (this function is convex – it is the perspective transform of the convex function $x^T x$) is $\mathbf{S}^m_+$-convex. Here is how the Gradient Inequality looks for it:

$$\begin{array}{l}\forall (X, \overline{X} \in \mathbf{R}^{m\times n}, Y \succ 0, \overline{Y} \succ 0) : \\ \begin{array}{rcl} X^T Y^{-1} X & \succeq & \overline{X}^T \overline{Y}^{-1} \overline{X} + [X - \overline{X}]^T \overline{Y}^{-1} \overline{X} + \overline{X}^T \overline{Y}^{-1} [X - \overline{X}] - \overline{X}^T \overline{Y}^{-1} [Y - \overline{Y}] \overline{Y}^{-1} \overline{X} \\ & = & \overline{X}^T \overline{Y}^{-1} X + X^T \overline{Y}^{-1} \overline{X} - \overline{X}^T \overline{Y}^{-1} Y \overline{Y}^{-1} \overline{X}. \end{array}\end{array}$$

Convex problem in cone-constrained form

♣ By definition, a *cone-constrained convex problem reads*

$$\mathrm{Opt}(C) = \min_{x in X} \left\{ f(x) : \overline{g}(x) := \overline{A}x - \overline{b} \leq 0, \widehat{g}(x) \leq_{\mathbf{K}} 0 \right\}, \qquad (C)$$

where:

- *the domain X is a nonempty convex set in some $\mathbf{R}^n$*
- *the objective $f : X \to \mathbf{R}$ is convex*
- *$\mathbf{K}$ is a regular cone in some $\mathbf{R}^m$*
- *$\widehat{g} : X \to \mathbf{R}^m$ is a $\mathbf{K}$-convex function*

Note that

- A convex in the standard sense MP problem is a problem of the form (C) with the nonnegative orthant of appropriate dimension in the role of $\mathbf{K}$
- Vice versa, convex cone-constrained problem (C) can be reformulated as a convex MP problem

  Indeed, $\mathbf{K}_*$ is a regular cone, and the function

  $$\Phi(y) = \max_z \{z^T y : z \in \mathbf{K}_*, \|z\|_2 \leq 1\}$$

  is a real-valued convex function which clearly is $\mathbf{K}$-monotone: $y \leq_{\mathbf{K}} y' \Rightarrow \Phi(y) \leq \Phi(y')$, which immediately implies that the function $g(x) := \Phi(\widehat{g}(x)) : X \to \mathbf{R}$ is convex. Besides this, $\Phi(y) \leq 0$ iff $y^T z \leq 0$ for all $z \in \mathbf{K}_*$, that is, iff $-y \in [\mathbf{K}_*]_* = \mathbf{K}$. $\Rightarrow$ for $x \in X$ it holds

  $$\widehat{g}(x) \leq_{\mathbf{K}} 0 \Leftrightarrow \widehat{g}(x) \in -\mathbf{K} \Leftrightarrow -\widehat{g}(x) \in \mathbf{K} \Leftrightarrow \Phi(\widehat{g}(x) \leq 0 \Leftrightarrow g(x) \leq 0.$$

  The bottom line is that (C) is equivalent to the convex MP problem

  $$\min_{x \in X} \left\{ f(x) : \overline{A}x - \overline{b} \leq 0, g(x) \leq 0 \right\}.$$

- If $\widehat{g}_i : X \to \mathbf{R}^{m_i}$ are $\mathbf{K}_i$-convex problem, optimization problem

  $$\min_{x \in X} \left\{ f(x) : \overline{g}(x) := \overline{A}x - \overline{b} \leq 0, \widehat{g}_i(x) \leq_{\mathbf{K}_i} 0, \, i \leq I \right\}$$

  is equivalent to the convex cone-constrained problem

  $$\min_{x \in X} \left\{ f(x) : \overline{A}x - \overline{b} \leq 0, \widehat{g}(x) := [\widehat{g}_1(x); ...; \widehat{g}_I(x)] \leq_{\mathbf{K}} 0, \; \mathbf{K} = \mathbf{K}_1 \times ... \times \mathbf{K}_I. \right\}$$

Conic problems

♣ **Conic problem** reads

$$\mathrm{Opt}(C) = \min_{x \in \mathbf{R}^n} \left\{ c^T x : \overline{g}(x) := \overline{A}x - \overline{b} \leq 0, \widehat{A}x - \widehat{b} \leq_{\mathbf{K}} 0 \right\}, \qquad (C)$$

where $\mathbf{K}$ is a regular cone in some $\mathbf{R}^n$.
Thus, a conic problem is a cone-constrained problem where

- the domain X is the entire $\mathbf{R}^n$
- the objective is linear, and
- the left hand side $\widehat{g}(x) = \widehat{A}x - \widehat{b}$ of the constraint $\widehat{g}(x) \leq_{\mathbf{K}} 0$ is affine

Note: Conic problem automatically is a *convex* cone-constrained problem, as every affine mapping is $\mathbf{K}$-convex, whatever be a regular cone $\mathbf{K}$. It can be easily verified that every convex MP problem can be reformulated as a conic one (and vice versa, since conic problem is a special case of a convex cone-constrained one, and problems of the latter type can be rewritten as convex MPs).

Motivation

♣ We have presented three "universal" ways to pose convex optimization problems – convex MP, convex cone-constrained, and conic formulations.

Question: What for ??? Convex MP problem seems to be a clear enough entity; what for all these regular cones, cone convexity, etc., etc.?

Answer is not so trivial. Of course, the proof of the pudding is in the eating, and we shall taste the pudding we just have started to cook till the end of our course...

However, it makes sense to outline a short answer right now:

Conic formulation of a convex optimization problem is "structure-revealing," allowing
— in many cases – to get deep understanding of the problem at hand "on paper," prior to any number-crunching
—- in all cases – to utilize the revealed structure by solution algorithms, enabling unified and efficient numerical treatment of seemingly quite different from each other problems.

As for cone-constrained form of a convex problem, investigating it will allow us to kill two birds with one stone – to arrive at basic duality results and optimality conditions for both classical Convex MP and for Conic Programming simultaneously.

♠ Convex problems have a lot of structure – otherwise, how could you know that your problem is convex?

• MP form "summarizes" all the structure in convexity. It is enough to be able to solve, under minimal computability and boundedness assumptions, convex MP's in a theoretically efficient manner.

• However: the corresponding "universal" algorithms of Convex Programming are "black-box-oriented:" they learn the instance to be solved by computing the values and the derivatives of the objective and the constraints at subsequently generated search points – what else, computationally speaking, can you do with general-type functions, even convex ones? In other words, *all your detailed knowledge of problem's structure and data* (you definitely possess this knowledge: how else could you be sure that the problem is convex?) *is used to compute local information – value and derivatives of the objective and constraints – at various points*. which is a very poor way to utilize your a priori knowledge of the problem.

In contrast: the LP Simplex method has no idea how to solve convex MP's, just linear ones. However, as applied to an LP,, it never computes values and gradients; it works directly on problem's data and converts it into the optimal solution.

This is how a high performance algorithm should work – it should be adjusted to problem's structure and utilize it to accelerate computations.

However - what is problem's structure? "Structure" has no formal definition, this is something we recognise in hindsight only. E.g., it is clear what is the structure of LP – it "sits" in the very simple and transparent cone $\mathbf{R}^m_+$ responsible for the coordinate-wise "$\leq$", all the rest are just linear functions.
In contrast, in convex MP "structure" sits in the word "convex;" of course, convex functions are much better suited for optimization than general ones, but still – "convexity": is an abstract notion; solution algorithm has no access to the specific reasons making your problem convex and thus cannot utilize these reasons, just the outcome – convexity...
Conic problem is not much better. Of course, we understand where the structure sits - in the regular cone $\mathbf{K}$ responsible for the vector inequality $\leq_{\mathbf{K}}$; all the rest, as in LP, are just linear functions. However, a general regular cone has no more "visible structure" than a general-type convex function, so what is the point?
The answer is highly unexpected and is, as it should be, "experimental," and not academic: *As a matter of fact, for all practical purposes,* whatever it means, *the entire Convex Programming is in the scope of just three "magic" families of conic problems: those where the cones are*
— *nonnegative orthants – finite direct products of nonnegative rays* $\mathbf{R}_+$ *on real line* (LP), or
— *finite direct products of Lorentz cones* (Conic Quadratic Programming, CQP, a.k.a. Second Order Conic Programming), or
— *finite direct products of semidefinite cones* (Semidefinite Programming, SDP)
♠ These three magic families of cones are well understood, and
• form a hierarchy – LP's are special cases of CQP's, which in turn are special cases of SDP's
• possess deep intrinsic mathematical similarity allowing for unified design of efficient solvers.

♠ Practice demonstrates that *as far as convex optimization models are concerned, users usually are capable to utilize their private knowledge of their models to convert them into an LP/CQP/SDP.* After it is done, the structure is revealed, and you can start number-crunching (or try to get understanding "on paper," utilizing powerful existing tools, primarily, *Conic Duality*)

♣ In addition, *there exists simple fully algorithmic "calculus" of representations of convex sets and functions via magic cones, which in practice simplifies dramatically converting your private knowledge into LP/CQP/SDP reformulation of your model*. This calculus can – and is — implemented in dedicated compilers ("Disciplined Convex Programming" — `CVX` software designed by Michael Grant and Stephen Boyd, second–to-none in user-friendliness and scope).

♣ **Why all this?** Why magic families of conic problems are enough for basically all applications of Convex Optimization?

> A cow's tail grows downward. I do not attempt to explain why the cow's tail grows downward. I merely cite the fact.
> – Jack London, *The End of the Story* (1891)

Convex Duality, I: Convex Theorem on Alternative

♣ **Duality: Motivation.** Consider a convex cone-constrained problem

$$\mathrm{Opt}(P) = \min_{x\in X}\left\{f(x) : \overline{g}(x) := \overline{A}x - \overline{b} \le 0, \widehat{g}(x) \le_{\mathbf{K}} 0\right\}, \qquad (P)$$

("primal problem"). As in LP, Convex duality is motivated by the desire to build a mechanism for lower-bounding the optimal value of (P). The mechanism is completely similar to the one used in LP. We start with answering the following

Question: *Given a real c, how to certify the relation $c \le \mathrm{Opt}(P)$?* In other word, *how to certify insolvability of the system of constraints*

$$\begin{array}{rcll} f(x) & < & c & (a)\\ \overline{g}(x) & \le & 0 & (b)\\ \widehat{g}(x) & \le_{\mathbf{K}} & 0 & (c)\\ x & \in & X & (d)\end{array} \qquad (S)$$

in variables x?

Answer: *A certificate is a collection of weights – weight 1 for (a), weight $\overline{\lambda} \in \mathbf{R}^{\mu}_{+}$ for (b), weight $\widehat{\lambda} \in \mathbf{K}_*$ for (c) – such that the aggregated system*

$$f(x) + \overline{\lambda}^T\overline{g}(x) + \widehat{\lambda}^T\widehat{g}(x) < c \;\&\; x \in X \qquad (\Sigma)$$

has no solutions, or, which is the same, such that

$$\inf_{x\in X}\left[f(x) + \overline{\lambda}^T\overline{g}(x) + \widehat{\lambda}^T\widehat{g}(x)\right] \ge c \qquad (!)$$

$$\left.\begin{array}{rcll} f(x) & < & c & (a)\\ \overline{g}(x) & \leq & 0 & (b)\\ \widehat{g}(x) & \leq_{\mathbf{K}} & 0 & (c)\\ x & \in & X & (d)\end{array}\right\} \quad (S)$$

$$\inf_{x\in X}\left[f(x)+\overline{\lambda}^T\overline{g}(x)+\widehat{\lambda}^T\widehat{g}(x)\right]\geq c \quad (!)$$

♠ The resulting condition *there exists aggregation weight* $\lambda=[\overline{\lambda};\widehat{\lambda}]$ *which is legitimate:*

$$[\overline{\lambda};\widehat{\lambda}]\in\Lambda:=\mathbf{R}^\mu_+\times\mathbf{K}_*$$

satisfying (!)
is sufficient for infeasibility of (S). In the linear case (f is linear, $\mathbf{K}_*=\mathbf{R}^m_+$, $X=\mathbf{R}^n$) and when the subsystem $(b),(c)$ composed by nonstrict inequalities of (S) is feasible, GTA says that this condition is also necessary. To get necessity in the general convex case, we need (S) to satisfy the Relaxed Slater condition

♠ **Slater/Relaxed Slater condition:** *We say that* (S) *satisfies*
– *Slater condition, if there exist* $\overline{x}\in\mathrm{rint}\,X$ *such that* $\overline{g}(\overline{x})<0$ *and* $\widehat{g}(\overline{x})<_{\mathbf{K}}0$,
– *Relaxed Slater condition, if there exist* $\overline{x}\in\mathrm{rint}\,X$ *such that* $\overline{g}(\overline{x})\leq 0$ *and* $\widehat{g}(\overline{x})<_{\mathbf{K}}0$.

Fact VIII.2 [Cone-constrained Convex Theorem on Alternative] *Let* (S) *be convex (i.e.,* $X\subset\mathbf{R}^n$ *is nonempty and convex,* $f:X\to\mathbf{R}$ *is convex,* $\mathbf{K}\subset\mathbf{R}^m$ *is a regular cone, and* $\widehat{g}:X\to\mathbf{R}^m$ *is* $\mathbf{K}$*-convex) and satisfy Relaxed Slater condition. Then* (S) *is infeasible iff* (!) *is feasible.*

This is Theorem IV.16.13, the most technically involving fact in our textbook. Its proof combines Separation Theorem and Dubovitski-Milutin Lemma.

Convex Duality, II: Lagrange Duality Theorem

Consider convex cone-constrained problem

$$\mathrm{Opt}(P) = \min_{x\in X}\left\{f(x) : \overline{g}(x) := \overline{A}x - \overline{b} \le 0, \widehat{g}(x) \le_{\mathbf{K}} 0\right\}, \quad \left[\overline{A} : \mu \times n,\ \widehat{A} : m \times n\right] \tag{P}$$

CTA attracts our attention to the *Lagrange function*

$$L(x.\lambda) = f(x) + \overline{\lambda}^T\overline{g}(x) + \widehat{\lambda}^T\widehat{g}(x) : X \times \Lambda \to \mathbf{R} \quad \left[\lambda = [\overline{\lambda}; \widehat{\lambda}].\ \Lambda = \mathbf{R}^\mu_+ \times \mathbf{K}_*\right]$$

of (P), associated *dual objective*

$$\underline{L}(\lambda) = \inf_{x\in X} L(x,\lambda) : \Lambda \to \mathbf{R} \cup \{-\infty\}$$

and the *Lagrange Dual*

$$\mathrm{Opt}(D) = \max_{\lambda\in\Lambda} \underline{L}(\lambda) \tag{D}$$

of primal *problem* (P), (in exception of our convention, the objective in (D) may take value $-\infty$).
When $\lambda \in \Lambda$, the Lagrange function clearly underestimates the objective f of (P) everywhere on the feasible set $\mathrm{Feas}(P)$ of (P), implying that $\underline{L}(\lambda) \le \mathrm{Opt}(P)$ for all $\lambda \in \Lambda$, whence

- $\mathrm{Opt}(D) \le \mathrm{Opt}(P)$ [weak duality]

Moreover, *assuming that* (P) *satisfies the Relaxed Slater condition, for every real* $c < \mathrm{Opt}(P)$, *that is. such that the system of constraints*

$$f(x) < c,\ \overline{g}(x) \le 0,\ \widehat{g}(x) \le_{\mathbf{K}} 0,\ x \in X$$

in variables x *is infeasible, the system of constraints* $\underline{L}(\lambda) \ge c, \lambda \in \Lambda$ *in variables* λ *is feasible.*

$$\begin{array}{rcll}\mathrm{Opt}(P) & = & \min\limits_{x\in X}\left\{f(x):\overline{g}(x):=\overline{A}x-\overline{b}\le 0,\widehat{g}(x)\le_{\mathbf{K}}0\right\} & (P)\\ \mathrm{Opt}(D) & = & \max\limits_{[\overline{\lambda};\widehat{\lambda}]\in\mathbf{R}^{\mu}_{+}\times\mathbf{K}_*}\left\{\underline{L}(\lambda):=\inf\limits_{x\in X}\left[f(x)+\overline{\lambda}^T\overline{g}(x)+\widehat{\lambda}^T\widehat{g}(x)\right]\right\} & (D)\end{array}$$

♠ We have arrived at the following

Fact VIII.3 [Lagrange Duality Theorem, cone-constrained form] *Assume that (P) is a feasible and bounded convex cone-constrained problem satisfying Relaxed Slater condition. Then (D) is solvable, and*

$$\mathrm{Opt}(P)=\mathrm{Opt}(D)$$

Indeed, we always have $\mathrm{Opt}(P)\ge\mathrm{Opt}(D)$ by weak duality. Now, under the premise of Theorem $\mathrm{Opt}(P)$ is a real, and of course the system of constraints

$$f(x)<\mathrm{Opt}(P),\overline{g}(x)\le 0,\widehat{g}(x)\le_{\mathbf{K}}0,x\in X$$

in variables x is infeasible. Applying CTA with $c=\mathrm{Opt}(P)$, we conclude that the system of constraints

$$\underline{L}(\lambda)\ge\mathrm{Opt}(P),\lambda\in\Lambda$$

in variables λ has a solution λ_*. Note that λ_* is a feasible solution to (D) with the value of the dual objective $\ge\mathrm{Opt}(P)$, while by weak duality $\mathrm{Opt}(D)\le\mathrm{Opt}(P)$. We see that $\mathrm{Opt}(P)\le\underline{L}(\lambda^*)\le\mathrm{Opt}(D)\le\mathrm{Opt}(P)$, where the second inequality is due to (D) being a maximization problem, and we conclude that $\mathrm{Opt}(P)=\mathrm{Opt}(D)$ and that λ_* is an optimal solution to (D). □

.

Convex Programming Optimality Conditions, Saddle Point form

♣ Consider a convex cone-constrained problem

$$\mathrm{Opt}(P) = \min_{x\in X}\left\{f(x) : \overline{g}(x) := \overline{A}x - \overline{b} \le 0, \widehat{g}(x) \le_{\mathbf{K}} 0\right\}, \qquad (P)$$

along with its Lagrange function

$$L(x.\lambda) = f(x) + \overline{\lambda}^T\overline{g}(x) + \widehat{\lambda}^T\widehat{g}(x) : X \times \Lambda \to \mathbf{R} \qquad \left[\lambda = [\overline{\lambda};\widehat{\lambda}].\Lambda = \mathbf{R}^\mu_+ \times \mathbf{K}_*\right]$$

A point (x_*, λ_*) is called a *saddle point* of the Lagrange function, if
— $x_* \in X$, $\lambda_* \in \Lambda$, and
— at point x_*, the function $L(x, \lambda_*)$ of $x \in X$ attains its minimum, and the function $L(x_*, \lambda)$ of $\lambda \in \Lambda$ attains its maximum:

$$\forall(x, \lambda) \in X \times \Lambda : L(x, \lambda_*) \ge L(x_*, \lambda_* \ge L(x_*, \lambda).$$

Immediate observation:

Fact VIII.4 *When $x \in X$, the function $L(x, \lambda)$ of $\lambda \in \Lambda$ is bounded from above iff $x \in$ Feas(P), and for $x \in$ Feas(P), a point $\lambda_* = [\overline{\lambda}_*; \widehat{\lambda}_*] \in \Lambda$ maximizes $L(x, \lambda)$ in $\lambda \in \Lambda$ iff x and λ_* are linked by complementary slackness:*

$$\overline{\lambda}_*^T\overline{g}(x) = 0 \ \&\ \widehat{\lambda}_*^T\widehat{g}(x) = 0,$$

or, which for $x \in$ Feas(P) and $\lambda_* \in \Lambda$ is the same, the relation $\lambda_*^T[\overline{g}(x); \widehat{g}(x)] = 0$,

Indeed, for a closed cone K, $\sup_{\phi\in K_*} \phi^T z = +\infty$ iff $z \notin -K$. When $z \in -K$, the supremum is 0 and is achieved exactly at those $\phi \in K_*$ for which $\phi^T z = 0$. As

$$\sup_{[\overline{\lambda};\widehat{\lambda}]\in\mathbf{R}^\mu_+\times\mathbf{K}_*} [\overline{\lambda}^T\overline{g}(x) + \widehat{\lambda}^T\widehat{g}(x)] = \sup_{\overline{\lambda}\in\mathbf{R}^\mu_+} \overline{\lambda}^T\overline{g}(x) + \sup_{\widehat{\lambda}\in\mathbf{K}_*} \widehat{\lambda}^T\widehat{g}(x),$$

we conclude that the left hand side sup is $< +\infty$ iff $\overline{g}(x) \le 0$ and $\widehat{g}(x) \le_{\mathbf{K}} 0$ (i.e., iff $x \in$ Feas(P)), and the maximizers in this case are exactly the pairs $\overline{\lambda} \in \mathbf{R}^\mu_+$, $\widehat{\lambda} \in \mathbf{K}_*$ linked to x by complementary slackness, Q.E.D.

$$\begin{array}{rcll}\mathrm{Opt}(P) & = & \min\limits_{x\in X}\left\{f(x):\overline{g}(x):=\overline{A}x-\overline{b}\le 0,\widehat{g}(x)\le_{\mathbf{K}}0\right\} & (P)\\ \mathrm{Opt}(D) & = & \max\limits_{[\overline{\lambda};\widehat{\lambda}]\in\mathbf{R}^{\mu}_{+}\times\mathbf{K}_*}\left\{\underline{L}(\lambda):=\inf\limits_{x\in X}\left[L(x,\lambda):=f(x)+\overline{\lambda}^T\overline{g}(x)+\widehat{\lambda}^T\widehat{g}(x)\right]\right\} & (D)\end{array}$$

Fact VIII.5 [Optimality conditions for cone-constrained convex problem, saddle point form] *Consider convex cone-constrained problem (P).*

(i) *Assume that $x_*\in X$ can be augmented by $\lambda_*\in\Lambda$ to yield a saddle point of the Lagrange function. Then x_* is an optimal solution to (P)*

Indeed, under the premise of the claim, λ_* is a maximizer of the function $L(x_*,\lambda)$ in $\lambda\in\Lambda=\mathbf{R}^{\mu}_{+}\times\mathbf{K}_*$, implying by Fact VIII.4 that x_* is a feasible solution to (P) linked to λ_* by complementary slackness: $\lambda_*^T[\overline{g}(x_*);\widehat{g}(x_*)]=0$. Now let x be a feasible solution to (P). We have

$$f(x)\ge f(x)+\lambda_*^T[\overline{g}(x);\widehat{g}(x)]=L(x.\lambda_*)\ge L(x_*,\lambda_*)=f(x_*)+\underbrace{\lambda_*^T[\overline{g}(x_*);\widehat{g}(x_*)]}_{=0}=f(x_*),$$

where the first inequality is due to $[\overline{g}(x);\widehat{g}(x)]\in-[\mathbf{R}^{\mu}_{+}\times\mathbf{K}]$ (x is feasible!) and $\lambda_*\in\Lambda=[\mathbf{R}^{\mu}_{+}\times\mathbf{K}]_*$, and the second inequality holds true since (x_*,λ_*) is a saddle point of L. Thus, x_* is an optimal solution to (P), Q.E.D.

(ii) *Assume that x_* is an optimal solution and (P) satisfies the Relaxed Slater condition. Then x_* can be augmented by $\lambda_*\in\Lambda$ to yield a saddle point of the Lagrange function.*

Indeed, by the Lagrange Duality Theorem, under the premise of the claim (D) has an optimal solution $\lambda_*\in\Lambda$, and $\mathrm{Opt}(D)=\mathrm{Opt}(P)=f(x_*)$. Let us prove that (x_*,λ_*) is a saddle point of L. We have

$$f(x_*)=\mathrm{Opt}(P)=\mathrm{Opt}(D)=\underline{L}(\lambda_*)=\inf_{x\in X}\{L(x,\lambda_*)\equiv f(x)+\lambda_*^T[\overline{g}(x);\widehat{g}(x)]\}\le f(x_*)+\lambda_*^T[\overline{g}(x_*);\widehat{g}(x_*)]\le f(x_*)$$

where the last inequality is due to $[\overline{g}(x_*);\widehat{g}(x_*)]\in\mathbf{R}^{\mu}_{+}\times\mathbf{K}$ and $\lambda_*\in\Lambda=[\mathbf{R}^{\mu}_{+}\times\mathbf{K}]_*$. We see that the inequalities in the above chain are equalities, implying that

$$\lambda_*^T[\overline{g}(x_*);\widehat{g}(x_*)]=0\ \&\ \inf_{x\in X}L(x,\lambda_*)=f(x_*).\qquad(*)$$

Due to $x_*\in\mathrm{Feas}(P)$ and $\lambda_*\in\Lambda$, the first relation in $(*)$ implies that x_* and λ_* are linked by complementary slackness, so that λ_* maximizes $L(x_*,\lambda)$ in $\lambda\in\Lambda$ by Fact VIII.4.Taken together, relations $(*)$ say that x_* minimizes $L(x,\lambda*)$ in $x\in X$. The bottom line is that (x_*,λ_*) is a saddle point of L, Q.E.D.

Convex Programming Optimality Conditions, Karush-Kuhn-Tucker form

♣ We are about to translate Saddle point optimality conditions into something "more verifiable." Consider a convex cone-constrained problem

$$\mathrm{Opt}(P) = \min_{x\in X}\left\{f(x) : \overline{g}(x) := \overline{A}x - \overline{b} \le 0, \widehat{g}(x) \le_{\mathbf{K}} 0\right\}, \qquad [\overline{A} : \mu \times n] \tag{P}$$

along with its Lagrange function

$$L(x.\lambda) = f(x) + \overline{\lambda}^T\overline{g}(x) + \widehat{\lambda}^T\widehat{g}(x) : X \times \Lambda \to \mathbf{R} \qquad \left[\lambda = [\overline{\lambda};\widehat{\lambda}],\ \Lambda = \mathbf{R}^\mu_+ \times \mathbf{K}_*\right]$$

and let $x_* \in X$ be a point where f and $\widehat{g}$ are differentiable.

Question: When $x_* \in X$ is the x-component of a saddle point of the Lagrange function?

Answer: *x_* should be a Karush-Kuhn-Tucker (KKT) point of (P)*, meaning that

x_* is feasible for (P), and there exists $\lambda_* \in \Lambda$ such that:

$$\begin{array}{rcll} \lambda_*^T[\overline{g}(x_*);\widehat{g}(x_*)] & = & 0 & \text{[complementary slackness]} \\ \nabla_x L(x_*,\lambda_*) & \in & -N_X(x_*) & \text{[KKT equation]} \end{array}$$

where

$$N_X(x_*) = \{h : h^T[x_* - x] \ge 0\ \forall x \in X\}$$

is the taken at x_* normal cone of X.

Indeed, $(x_*,\lambda_*) \in X \times \Lambda$ is a saddle point of L iff the following two conditions hold:

- λ_* is a maximizer of $L(x_*,\lambda)$ in $\lambda \in \Lambda$, which by Fact VIII.4 is the same as x_* being feasible and x_*, λ_* being linked by complementary slackness, and
- x_* is a minimizer of the function $L(x,\lambda_*)$ in $x \in X$. This function is convex in $x \in X$ (as (P) is convex and $\lambda_* \in \Lambda = \mathbf{R}^\mu_+ \times \mathbf{K}_*$), and we are in the situation when the function is differentiable at x_*; consequently, $x_($ minimizes the function in $x \in X$ iff the KKT equation holds true.

$$\mathrm{Opt}(P) = \min_{x\in X}\left\{f(x) : \overline{g}(x) := \overline{A}x - \overline{b} \le 0, \widehat{g}(x) \le_{\mathbf{K}} 0\right\}, \qquad (P)$$

♠ With the observation just made in mind, Fact VIII.5 translates into

Fact VIII.6 [Optimality conditions for cone-constrained convex problem, KKT form] *Consider convex cone-constrained problem (P), and let $x_* \in X$ be such that f, $\widehat{g}$ are differentiable at x_*.*

(i) *If $x_* \in X$ is a KKT point of (P), then x_* is an optimal solution to (P).*

(ii) *If x_* is an optimal solution to (P) and the .Relaxed Slater condition holds, x_* is a KKT point of (P).*

Illustration I: Given $a_i > 0$, $1 \leq i \leq n$, let us solve the optimization problem

$$\mathrm{Opt} = \min_x \left\{ \sum_i a_i/x_i : x > 0, \sum_i x_i \leq 1 \right\}$$

this is convex cone-constrained problem with $X = \{x \in \mathbf{R}^n : x > 0\}$ (more exactly, it becomes such a problem when setting $\overline{g}(x) = \sum_i x_i - 1$ and, say, $\mathbf{R} = \mathbf{R}_+$ and $\widehat{g}(x) \equiv -1$ to respect our cone-constrained format). Let us make an educated guess that there exists a KKT point where the constraint $\overline{g}(x) \leq 0$ is active, and let us find this point. As the constraint $\widehat{g}(x) \leq 0$ never is active, its Lagrange multiplier is 0, complementary slackness does not impose additional to nonnegativity restrictions on the Lagrange multiplier for the constraint $\overline{g}(x) \leq 0$ – we are looking at the point where the constraint is active!, the normal cone of X at every point $x \in X$ is the entire $\mathbf{R}^n$, and the KKT equation reads

$$\nabla_x (\sum_i a_i/x_i + \lambda[\sum_i x_i - 1]) = 0 \Leftrightarrow \{a_i/x_i^2 = \lambda, i \leq n\}$$

Augmenting the KKT equation with our guessed $\sum_i x_i = 1$, we immediately find λ and x_i, arriving at

$$\lambda = [\sum_i \sqrt{a_i}]^2,\ x_i = \frac{\sqrt{a_i}}{\sum_j \sqrt{a_j}},\ \mathrm{Opt} = \left[\sum_i \sqrt{a_i}\right]^2,$$

and our computation shows that what we have found indeed is a KKT point, and thus, by Fact VIII.6, the x we have found is an optimal solution to the problem.

Illustration II: Given reals a_i, $i \le n$, let us solve the problem

$$\mathrm{Opt} = \min_x \left\{ \sum_i [x_i \ln(x_i) - a_i x_i] : x \ge 0, \sum_i x_i = 1 \right\}.$$

(we have already solved this problem when illustrating optimality conditions in minimization of a convex function over a convex set). What we have is a convex cone-constrained problem with $X = \{x \in \mathbf{R}^n : x \ge 0, \sum_i x_i = 1\}$ and "dummy" $\overline{g}(x)$, $\widehat{g}(x)$, $\mathbf{K}$, say, $\overline{g}(x) \equiv \widehat{g}(x) = -1$, $\mathbf{K} = \mathbf{R}_+$. As a result, complementary slackness releases us from looking for Lagrange multipliers – they are zeros, With the educated guess that there is a KKT point in rint X, where the radial cone is $\{h : \sum_i h_i = 0$, and the normal cone is $\mathbf{R} \times [1; ...; 1]$, the KKT equation is

$$\exists \mu \in \mathbf{R} : \ln(x_1) + 1 - a_i = \mu,\ 1 \le i \le n,$$

which combines with the feasibility requirement $\sum_i x_i = 1$ to yield a solution:

$$x_i = \exp\{a_i\} / \sum_j \exp\{a_j\},\ i \le n,\ \mathrm{Opt} = -\ln(\sum_j \exp\{a_j\}).$$

Note that we have computed the Legendre transform of the function $\sum_i x_i \ln(x_i)$ restricted onto the probabilistic simplex: it is $\ln(\sum_i \exp\{y_i\})$.

Pay attention to how the existence of a *verifiable* sufficient optimality condition simplifies our life: whatever guess we make, upon success – after a KKT point is found – we are done: we have found *an* optimal solution. All's well that ends well...

Application: Optimal value in parametric convex cone-constrained problem

♣ When speaking about the optimal value in LP as a function of the right hand side vector, we have seen that the subgradient of this function is given by an optimal solution to the dual problem. We are about to establish a "nonlinear analogy" of this fact.

Situation: Consider a parametric family of convex cone-constrained problems defined by a parameter $p \in P$

$$\mathrm{Opt}(p) := \min_{x \in X} \{ f(x,p) : \ g(x,p) \leq_{\mathbf{M}} 0 \}, \qquad (P[p])$$

where

- $X \subseteq \mathbf{R}^n$, $P \subseteq \mathbf{R}^\mu$ are nonempty and convex,
- $\mathbf{M} \subset \mathbf{R}^\nu$ is a regular cone,
- $f : X \times P \to \mathbf{R}$ is convex, and $g : X \times P \to \mathbf{R}^\nu$ is $\mathbf{M}$-convex

[to save notation, we stick to "single-constraint" formulation]

♠ **Question:** What is the status of the function $\mathrm{Opt}(\cdot) : P \to \mathbf{R} \cup \{\pm\infty\}$?

$$\mathrm{Opt}(p) := \min_{x\in X}\{f(x,p) :\ g(x,p) \leq_{\mathbf{M}} 0\}, \qquad (P[p])$$

♠ We make the following assumption:
$\overline{x} \in X$, $\overline{p} \in P$ are such that

- $\overline{x}$ is a KKT point of $(P[\overline{p}])$ ($\Rightarrow$ $\overline{x}$ is an optimal solution to the problem)
- $f(x,p)$ and $g(x,p)$ are differentiable at the point $[\overline{x};\overline{p}]$, the derivatives being

$$Df([\overline{x};\overline{p}])[[dx;dp]] = F_x^T dx + F_p^T dp,\ \ Dg([\overline{x};\overline{p}])[[dx;dp]] = G_x dx + G_p dp.$$

- Let $\overline{\mu} \in \mathbf{M}_*$ be the Lagrange multiplier associated with $\overline{x}$ and (P[$\overline{p}$]):

$$\overline{\mu}^T g(\overline{x},\overline{p}) = 0 \ \&\ [x-\overline{x}]^T[F_x + G_x^T\overline{\mu}] \geq 0,\ \forall x \in X. \qquad (*)$$

Fact VIII.7 *Under the circumstances,* Opt(·) *is a convex function on* P *taking values in* $\mathbf{R}\cup\{+\infty\}$ *and finite at* $\overline{p}$*, and the vector*

$$F_p + G_p^T\overline{\mu}$$

is a subgradient of Opt(·) *at* $\overline{p}$*:*

$$\mathrm{Opt}(p) \geq \mathrm{Opt}(\overline{p}) + [p-\overline{p}]^T[F_p + G_p^T\overline{\mu}],\ \forall p \in P.$$

Proof. ✓ f is convex $\Rightarrow$

$$f(x,p) \geq f(\overline{x},\overline{p}) + F_x^T[x-\overline{x}] + F_p^T[p-\overline{p}], \quad \forall (x\in X, p\in P).$$

$\overline{\mu} \in \mathbf{M}_*$, g is $\mathbf{M}$-convex $\Rightarrow$ $\overline{\mu}^T g(x,p) : X\times P \to \mathbf{R}$ is convex $\Rightarrow$

$$\overline{\mu}^T g(x,p) \geq \underbrace{\overline{\mu}^T g(\overline{x},\overline{p})}_{=0 \text{ by } (*)} + \overline{\mu}^T G_x[x-\overline{x}] + \overline{\mu}^T G_p[p-\overline{p}],\ \forall (x\in X, p\in P).$$

$Df([\overline{x};\overline{p}])[[dx;dp]] = F_x^T dx + F_p^T dp,\ Dg([\overline{x};\overline{p}])[[dx;dp]] = G_x dx + G_p dp$	$(a.1)$
$f(x,p) \geq f(\overline{x},\overline{p}) + F_x^T[x-\overline{x}] + F_p^T[p-\overline{p}],\ \forall (x \in X, p \in P)$	$(a.2)$
$\overline{\mu} \in \mathbf{M}_*\ \&\ : \overline{\mu}^T g(\overline{x},\overline{p}) = 0\ \&\ [x-\overline{x}]^T[F_x + G_x^T\overline{\mu}] \geq 0,\ \forall x \in X$	(b)
$\overline{\mu}^T g(x,p) \geq \overline{\mu}^T G_x[x-\overline{x}] + \overline{\mu}^T G_p[p-\overline{p}],\ \forall (x \in X, p \in P)$	(c)
$\mathrm{Opt}(p) := \min_{x \in X}\{f(x,p) :\ g(x,p) \leq_{\mathbf{M}} 0\}$	
$\mathrm{Opt}(p)?\geq?\mathrm{Opt}(\overline{p}) + [p-\overline{p}]^T[F_p + G_p^T\overline{\mu}],\ \forall p \in P$	(d)

Now let $p \in P$ and x be feasible for $(P[p])$. Then,

$$\begin{array}{ll} f(x,p) & \geq f(x,p) + \overline{\mu}^T g(x,p) \\ & [\text{as } \overline{\mu} \in \mathbf{M}_*\ \&\ g(x,p) \leq_{\mathbf{M}} 0] \\ & \geq f(\overline{x},\overline{p}) + F_x^T[x-\overline{x}] + F_p^T[p-\overline{p}] + \overline{\mu}^T G_x[x-\overline{x}] + \overline{\mu}^T G_p[p-\overline{p}] \\ & [\text{by } (a.2) \text{ and } (c)] \\ & = \mathrm{Opt}(\overline{p}) + (F_x + G_x^T\overline{\mu})^T[x-\overline{x}] + (F_p + G_p^T\overline{\mu})^T[p-\overline{p}] \\ & \geq \mathrm{Opt}(\overline{p}) + (F_p + G_p^T\overline{\mu})^T[p-\overline{p}], \\ & [\text{by } (b)] \end{array}$$

The resulting inequality holds true for all x feasible for $(P[p])$ $\Rightarrow$? $\geq$? in (d) is $\geq$
✓ It remains to verify that $\mathrm{Opt}(\cdot)$ is convex on P. As ? $\geq$? in (d) is $\geq$, Opt on P does not take value $-\infty$. Let $p', p'' \in P \cap \mathrm{Dom}\,(\mathrm{Opt}(\cdot))$ and $\lambda \in [0,1]$, and let

$$p = \lambda p' + (1-\lambda)p''$$

Given $\epsilon > 0$, there exist $x', x'' \in X$:

$$g(x',p') \leq_{\mathbf{M}} 0,\ g(x'',p'') \leq_{\mathbf{M}} 0,\ f(x',p') \leq \mathrm{Opt}(p') + \epsilon,\ f(x'',p'') \leq \mathrm{Opt}(p'') + \epsilon.$$

Setting $x = \lambda x' + (1-\lambda)x''$, by convexity of f and $\mathbf{M}$-convexity of g, we have

$$g(x,p) \leq_{\mathbf{M}} 0,\ f(x,p) \leq [\lambda\mathrm{Opt}(p') + (1-\lambda)\mathrm{Opt}(p'')] + \epsilon.$$

We conclude that

$$\mathrm{Opt}(\lambda p' + (1-\lambda)p'') \leq \lambda\mathrm{Opt}(p') + (1-\lambda)\mathrm{Opt}(p'') + \epsilon\ \forall \epsilon > 0,$$

and the convexity of $\mathrm{Opt}(\cdot)$ follows, Q.E.D.

♠ **Standard Example:**

$$\mathrm{Opt}(p) := \min_{x \in X} \{ f(x) : \ g(x) - p \leq_{\mathbf{M}} 0 \}, \qquad (P[p])$$

In this case $P = \mathbf{R}^\nu$, and Fact VIII.7 says that *If X is convex, $f : X \to \mathbf{R}$ is convex, $g : X \to \mathbf{R}^\nu \supset X$ is $\mathbf{M}$-convex, and $\overline{x}$ is a KKT point of $(P[\overline{p}])$, $\overline{\mu}$ being the associated Lagrange multiplier, then* $\mathrm{Opt}(p)$ *is convex on X and*

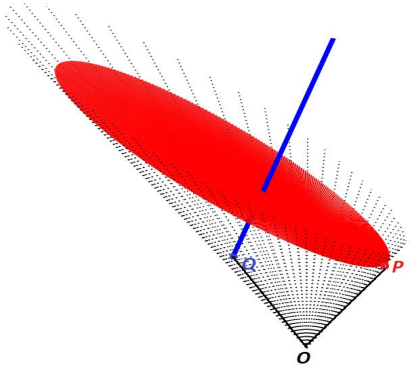
$$\forall (p \in P) : \mathrm{Opt}(p) \geq \underbrace{\mathrm{Opt}(\overline{x})}_{f(\overline{x})} - \overline{\mu}^T [p - \overline{p}]$$

– minus the Lagrange multiplier $\overline{\mu}$ certifying optimality of $\overline{x}$ for $(P[\overline{p}])$ is a subgradient of $\mathrm{Opt}(p)$ *at $p = \overline{p}$.*

Lecture III.2

Conic Programming
Saddle Points

Conic Programming programs
Conic Duality & Optimality conditions
Geometry of Primal-Dual pair of conic programs
Saddle Points



Conic Programming and Conic Duality

♣ **Conic problem** is a cone-constrained problem where the objective is linear, the left hand side $\widehat{g}(\cdot)$ of the nonlinear constraint is affine, and the domain X is the entire space, that is, a conic problem reads

$$\begin{array}{c}\mathrm{Opt}(P) = \min\limits_{x\in\mathbf{R}^n}\left\{c^Tx : Ax \le b, \widehat{A}x \le_{\mathbf{K}} \widehat{b}\right\}\\ \left[A : \mu\times n,\ \widehat{A} : m\times n\right]\end{array} \qquad (P)$$

where $\mathbf{K}\subset\mathbf{R}^m$ is a regular cone. As an affine mapping is $\mathbf{K}$-convex, whatever be a regular cone $\mathbf{K}$, the problem is a *convex* cone-constrained one.

♣ **The Lagrange function** of (P) is

$$L(x,\lambda=[\overline{\lambda};\widehat{\lambda}]) = c^Tx + \overline{\lambda}^T[Ax-b] + \widehat{\lambda}^T[\widehat{A}x-\widehat{b}] = [c+A^T\overline{\lambda}+\widehat{A}^T\widehat{\lambda}]^Tx - b^T\overline{\lambda} - \widehat{b}^T\widehat{\lambda}.$$

Consequently, the objective of the Lagrange dual problem is

$$\underline{L}(\lambda=[\overline{\lambda};\widehat{\lambda}]) = \begin{cases} -b^T\overline{\lambda}-\widehat{b}^T\widehat{\lambda} & , A^T\overline{\lambda}+\widehat{A}^T\widehat{\lambda} = -c\\ -\infty & ,\ \text{otherwiise}\end{cases}$$

and the Lagrange dual problem itself becomes the *conic dual problem*

$$\mathrm{Opt}(D) = \max_{\lambda=[\overline{\lambda};\widehat{\lambda}]}\left\{-[b^T\overline{\lambda}+\widehat{b}^T\widehat{\lambda}] : \overline{\lambda}\ge 0, \widehat{\lambda}\in\mathbf{K}_*, A^T\overline{\lambda}+\widehat{A}^T\widehat{\lambda} = -c\right\} \qquad (D)$$

of (P). We see that *The Lagrange dual of a conic problem is a conic problem itself*.

$$\begin{array}{rcll}\mathrm{Opt}(P) & = & \min\limits_{x\in\mathbf{R}^n}\left\{c^Tx : Ax\le b, \widehat{A}x\le_{\mathbf{K}}\widehat{b}\right\}] & (P)\\ \mathrm{Opt}(D) & = & \max\limits_{\overline{\lambda};\widehat{\lambda}}\left\{-[b^T\overline{\lambda}+\widehat{b}^T\widehat{\lambda}] : \overline{\lambda}\ge 0, \widehat{\lambda}\in\mathbf{K}_*, A^T\overline{\lambda}+\widehat{A}^T\widehat{\lambda}=-c\right\} & (D)\end{array}$$

(D) becomes problem of the form (P) when replacing maximization of $-[b^T\overline{\lambda}+\widehat{b}^T\widehat{\lambda}]$ with minimization of $b^T\overline{\lambda}+\widehat{b}^T\widehat{\lambda}$, representing the equality constraints $A\overline{\lambda}+\widehat{A}\widehat{\lambda}=-c$ by inequalities

$$[A^T;-A^T]\overline{\lambda}+[\widehat{A}^T;-\widehat{A}^T]\widehat{\lambda}\le[-c;c] \qquad (1)$$

and rewriting the constraints $\overline{\lambda}\ge 0, \widehat{\lambda}\in\mathbf{K}_*$ as

$$\begin{array}{rcll}-\overline{\lambda} & \le & 0 & (2)\\ -\widehat{\lambda} & \le_{\mathbf{K}_*} & 0 & (3)\end{array}$$

♠ Let us build the conic dual of the conic representation of (D). Denoting by $u=[u_+;u_-]\ge 0$, $v\ge 0$, $w\in[\mathbf{K}_*]_*=\mathbf{K}$ the Lagrange multipliers for the constraints (1), (2), (3), respectively, and recalling that $[\mathbf{K}_*]_*=\mathbf{K}$, the conic dual of (D) becomes the problem

$$\max_{u_\pm,v,w}\left\{-c^T[u_--u_+] : A[u_--u_+]+v=b,\ \widehat{A}[u_--u_+]+w=\widehat{b}, u_+\ge 0, u_-\ge 0, v\ge 0, w\ge_{\mathbf{K}}\right\}.$$

Eliminating v,w, setting $x=u_--u_+$, and passing from maximization of $-c^T[u_--u_+]$ to minimization of $c^T[u_--u_+]$, the latter problem becomes

$$\min_x\left\{c^Tx : Ax-b\le 0, \widehat{A}x-\widehat{b}\le_{\mathbf{K}}0\right\},$$

which is nothing but (P).

$$\begin{array}{rcll}\mathrm{Opt}(P) & = & \min\limits_{x\in\mathbf{R}^n}\left\{c^Tx : Ax\le b, \widehat{A}x\le_{\mathbf{K}}\widehat{b}\right\} & (P)\\ \mathrm{Opt}(D) & = & \max\limits_{\overline{\lambda};\widehat{\lambda}}\left\{-[b^T\overline{\lambda}+\widehat{b}^T\widehat{\lambda}] : \overline{\lambda}\ge 0, \widehat{\lambda}\in\mathbf{K}_*, A^T\overline{\lambda}+\widehat{A}^T\widehat{\lambda}=-c\right\} & (D)\end{array}$$

♣ We have established

Fact IX.1 [Primal-dual symmetry in Conic Programming] *The conic duality is symmetric: the conic dual to the conic dual (D) of (P) is* (equivalent to) *the primal problem (P).*

♠ Combining primal-dual symmetry with the Lagrange Duality Theorem in cone-constrained form, we arrive at

Fact IX.2 [Conic Duality Theorem] *Consider a primal-dual pair (P), (D) of conic problems. Then*

$\mathring{}$(i) [Primal-dual symmetry] *Conic duality is symmetric: the problem dual to (D) is* (equivalent to) *(P)*

$\mathring{}$(ii) [Weak duality] *One has $\mathrm{Opt}(D)\le\mathrm{Opt}(P)$.*

$\mathring{}$(ii) [Strong duality] *Assume that one of the problems (P), (D) satisfies the Relaxed Slater condition and is bounded. Then the other problem is solvable, and*

$$\mathrm{Opt}(P)=\mathrm{Opt}(D).$$

As a result, when both problems satisfy the Relaxed Slater condition, both are solvable with equal optimal values. Finally, one has

$$\mathrm{Opt}(P)=\mathrm{Opt}(D)$$

whenever one of the problems satisfies the Relaxed Slater condition, whether this problem is or is not bounded.

Conic Programming Optimality conditions

$$\begin{array}{rcll} \mathrm{Opt}(P) & = & \min\limits_{x\in\mathbf{R}^n}\left\{c^Tx : Ax\le b, \widehat{A}x\le_{\mathbf{K}}\widehat{b}\right\}] & (P)\\ \mathrm{Opt}(D) & = & \max\limits_{\overline{\lambda};\widehat{\lambda}}\left\{-[b^T\overline{\lambda}+\widehat{b}^T\widehat{\lambda}] : \overline{\lambda}\ge 0, \widehat{\lambda}\in\mathbf{K}_*, A^T\overline{\lambda}+\widehat{A}^T\widehat{\lambda}=-c\right\} & (D)\end{array}$$

Fact IX.3 [Conic Programming Optimality conditions] *Given a primal-dual pair (P), (D) of conic programs, assume that both satisfy the Relaxed Slater condition, and let x_*, $\lambda_*=[\overline{\lambda}_*;\widehat{\lambda}_*]$ be a pair of primal-dual feasible solutions. The pair is composed of optimal solutions to the respective problems*

- [Zero duality gap] *Iff*

$$\mathrm{DualityGap}(x_*,\lambda_*) := c^Tx_* - \left[-[b^T\overline{\lambda}_*+\widehat{b}^T\widehat{\lambda}_*]\right] \equiv c^Tx_*+b^T\overline{\lambda}_*+\widehat{b}^T\widehat{\lambda}_* = 0$$

same as

- [Complementary slackness] *Iff*

$$\overline{\lambda}_*^T[b-Ax_*]=0 \ \&\ \widehat{\lambda}_*^T[\widehat{b}-\widehat{A}x_*]=0$$

$$\begin{array}{rcll}
\mathrm{Opt}(P) & = & \min_{x\in\mathbf{R}^n}\left\{c^Tx : Ax\le b, \widehat{A}x\le_{\mathbf{K}}\widehat{b}\right\}] & (P)\\
\mathrm{Opt}(D) & = & \max_{\overline{\lambda};\widehat{\lambda}}\left\{-[b^T\overline{\lambda}+\widehat{b}^T\widehat{\lambda}] : \overline{\lambda}\ge 0, \widehat{\lambda}\in\mathbf{K}_*, A^T\overline{\lambda}+\widehat{A}^T\widehat{\lambda}=-c\right\} & (D)\\
\mathrm{DualityGap}(x_*,\lambda_*) & := & c^Tx_*+b^T\overline{\lambda}_*+\widehat{b}^T\widehat{\lambda}_*=0 \text{ [zero duality gap]} & \\
\overline{\lambda}_*^T[b-Ax_*] & = & 0 \;\&\; \widehat{\lambda}_*^T[\widehat{b}-\widehat{A}x_*] = \text{ [complementarry slackness]} &
\end{array}$$

Indeed, as both problems satisfy the Relaxed Slater condition, both are solvable with equal optimal values Therefore

- One has

$$\mathrm{DualityGap}(x_*,\lambda_*) = \underbrace{\left[c^Tx_*-\mathrm{Opt}(P)\right]}_{\ge 0} + \underbrace{\left[\mathrm{Opt}(D)-\left[[-b^T\overline{\lambda}_*-\widehat{b}^T\widehat{\lambda}_*]\right]\right]}_{\ge 0},$$

⇒ *The duality gap as evaluated at a primal-dual feasible pair (x_*,λ_*) is nonnegative and is zero iff x_* is primal-, and λ_* is dual optimal.*

- For primal-dual feasible $(x,\lambda=[\overline{\lambda};\widehat{\lambda}])$ we have

$$\mathrm{DualityGap}(x,\lambda) = c^Tx+b^T\overline{\lambda}+\widehat{b}^T\widehat{\lambda} = -[A^T\overline{\lambda}+\widehat{A}^T\widehat{\lambda}]^Tx+b^T\overline{\lambda}+\widehat{b}^T\widehat{\lambda} = [\underbrace{b-Ax}_{\ge 0}]^T\underbrace{\overline{\lambda}}_{\ge 0}+[\underbrace{\widehat{b}-\widehat{A}x}_{\in\mathbf{K}}]^T\underbrace{\widehat{\lambda}}_{\in\mathbf{K}_*}$$

⇒ *The duality gap as evaluated at a primal-dual feasible pair (x,λ) vanishes iff x and λ are linked by complementary slackness.*

Illustration Steiner sum problem

♣ **Steiner sum problem**:

$$\min_{x\in\mathbf{R}^n}\sum_{i=1}^m\|x-a_i\|_2. \qquad [m>1, a_1,...,a_m \text{ are distinct points in } \mathbf{R}^n]$$

Cover story ($n=2$): There are m oil wells located at points $a_1,...,a_m\in\mathbf{R}^2$. Where should one place an oil collector in order to minimize the total length of pipelines connecting the wells to the collector?

♠ The problem can be reformulated as conic:

$$\min_{t_1,..,t_m,x}\left\{\sum_{i=1}^m t_i : \underbrace{[x-a_i;t_i]\in\mathbf{L}^{n+1}}_{\Leftrightarrow\|x-a_i\|_2\le t_i}, i=1,...,m\right\} \qquad (P)$$

Lorentz cones are self-dual, so that the problem dual to (S) is obtained by
— assigning the constraints $[x-a_i;t_i]\in\mathbf{L}^{n+1}$ with Lagrange multipliers $[y_i;z_i]\in\mathbf{L}^{n+1}$ giving rise to the aggregated constraint

$$\sum_i\left[[x-a_i]^Ty_i]+t_iz_i\right]\ge 0\Leftrightarrow[\sum_i y_i^T]x+\sum_i z_it_i\ge\sum_i y_i^Ta_i$$

— imposing on the multipliers the restriction that the left hand side in the aggregated constraint is, identically in the primal variables x,t_i, equal to the primal objective $\sum_i t_i$, which amounts to

$$\sum_i y_i=0,\ z_1=...=z_m=1$$

and maximizing under this restriction the right hand side of the aggregated constraint. Thus, the dual problem reads

$$\max_{y_1,...,y_m}\left\{\sum_i a_i^Ty_i : \sum_i y_i=0, \|y_i\|_2\le 1, i\le m\right\} \qquad (D)$$

$$\begin{array}{rcll}
\mathrm{Opt}(P) & = & \min\limits_{t_1,..,t_m,x} \left\{\sum_{i=1}^m t_i : [x-a_i; t_i] \in \mathbf{L}^{n+1},\, i=1,...,m\right\} & (P)\\
\mathrm{Opt}(D) & = & \max\limits_{y_1,...,y_m} \left\{\sum_i a_i^T y_i : \sum_i y_i = 0, \|y_i\|_2 \le 1, i \le m\right\} & (D)
\end{array}$$

- (P) clearly is solvable and strictly feasible $\Rightarrow$ (D) is solvable and $\mathrm{Opt}(P) = \mathrm{Opt}(D)$.
- From optimality conditions it is easily seen that

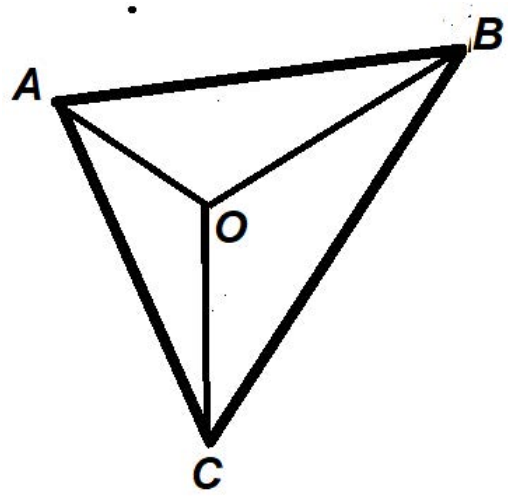
— A point x distinct from $a_1, .., a_m$ is an optimal solution to the Steiner sum problem iff

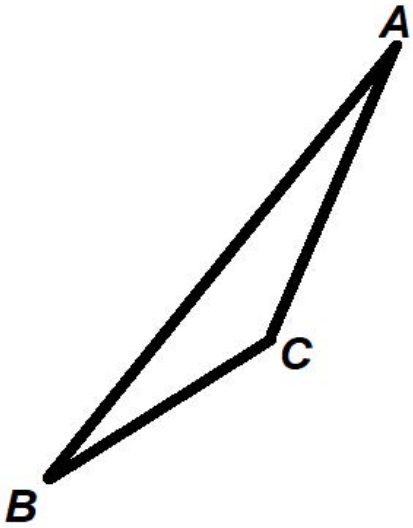
$$\sum_i \frac{a_i - x}{\|a_i - x\|_2} = 0.$$

— point $x = a_\ell$ is an optimal solution iff

$$\left\| \sum_{i \neq \ell} \frac{a_i - x}{\|a_i - x\|_2} \right\|_2 \le 1.$$

♠ In the simplest case of 3 points $a_1 = A, a_2 = B, a_3 = C$ in 2D plane, the optimal solution is
— either the point from which all 3 sides of the triangle ΔABC are seen at the angle 120° (such a point exists if angles of the triangle are $< 120°$)
— or the vertex of the triangle corresponding to the angle $\geq 120°$, if such an angle is present





$\angle CAB < 120^o$, $\angle ABC < 120^o$, $\angle BCA < 120^o$
solution $O, \angle AOB = \angle BOC = \angle COA = 120^o$

$\angle ACB \geq 120^o$
solution C

Note: Quoting "Fermat point" in Wikipedia, "This question [to minimize the sum of distances from a point to the vertices of triangle] was proposed by Fermat, as a challenge to Evangelista Torricelli. He solved the problem in a similar way to Fermat's [...] His pupil, Viviani, published the solution in 1659.

Consequences of Conic Duality Theorem

♣ **Question:** When a linear vector inequality

$$Ax \geq_{\mathbf{K}} b \qquad \text{(I)}$$

with regular cone $\mathbf{K}$ has no solutions?

Tautological answer: (I) *has no solutions iff*

$$b \notin \mathcal{B} := \{b : Ax \geq_{\mathbf{K}} b\} = A\mathbf{R}^n - \mathbf{K}$$

Note: For $\lambda \in \mathbf{K}_*$, the scalar inequality $[A^T\lambda]^T x \geq b^T\lambda$ is a consequence of (I). $\Rightarrow$

♠ **Immediate sufficient condition for infeasibility of (I):** *If by "admissible aggregation" of* (I) *one can obtain a contradictory scalar inequality:*

$$\exists \lambda \geq_{\mathbf{K}_*} 0 : \quad A^T\lambda = 0,\ \lambda^T b > 0. \qquad \text{(II)}$$

then (I) *has no solutions.*

Fact IX.4 *Let*

$$\overline{\mathcal{B}} = \mathrm{cl}\,\mathcal{B} = \mathrm{cl}\,[A\mathbf{R}^n - \mathbf{K}]$$

be the set of b's for which (I) *is "almost solvable," meaning that appropriately chosen arbitrarily small perturbations of b make* (I) *solvable.*
(II) *is solvable iff $b \notin \overline{\mathcal{B}}$. Thus,*

- *if* (II) *is solvable, then* (I) *is unsolvable*
- *if* (I) *is solvable and $\mathcal{B}$ is closed (as is the case when the cone $\mathbf{K}$ is polyhedral), then* (II) *is unsolvable,*

$$Ax \geq_{\mathbf{K}} b \qquad \text{(I)}$$
$$\lambda \geq_{\mathbf{K}_*} 0,\ A^T\lambda = 0,\ \lambda^T b > 0 \qquad \text{(II)}$$

Proof. All we need is to prove that (II) is solvable iff $b \notin \overline{\mathcal{B}}$. ✓ Let (II) be solvable for $b = \bar{b}$. Then (II) remains solvable for all small enough perturbations b of $\bar{b}$, implying by Immediate sufficient condition for infeasibility of (I) that all these perturbations are outside of $\mathcal{B}$, whence $\bar{b} \notin \overline{\mathcal{B}}$. ✓ Now let $b \notin \overline{\mathcal{B}}$, and let us prove that (II) is solvable. For $f \in \operatorname{int}\mathbf{K}$, consider the conic problem

$$\text{Opt} = \min_{x,t}\{t : Ax - b + tf \geq_{\mathbf{K}} 0\}.$$

As $b \notin \overline{\mathcal{B}}$, the t-components of feasible solutions are bounded away from 0, so that Opt > 0, and as $f > 0$, all solutions $(x = 0, t)$ with large positive t satisfy the $>_{\mathbf{K}}$-version of the constraint. By Conic Duality Theorem, the dual . problem

$$\max_{\lambda}\{b^T\lambda : A^T\lambda = 0, f^T\lambda = 1,\ \lambda \geq_{\mathbf{K}_*} 0\}$$

is solvable with optimal value Opt > 0, implying that (II) is solvable, Q.E.D.

$$
\begin{array}{cc}
Ax \geq_{\mathbf{K}} b & \text{(I)} \\
\lambda \geq_{\mathbf{K}_*} 0,\ A^T\lambda = 0,\ \lambda^T b > 0 & \text{(II)} \\
\mathcal{B} = A\mathbf{R}^n - \mathbf{K} &
\end{array}
$$

♣ When $\mathbf{K} = \mathbf{R}^m_+$, Fact IX.4 says exactly the same as the General Theorem on Alternative: *a finite system of nonstrict linear inequalities has no solutions iff the stemming from some* $\lambda \in \mathbf{R}^m_+$ *scalar linear inequality* $\lambda^T[Ax - b]$ *in variables* x *is contradictory*. Moreover, Fact IX.4 states that *if* $\mathcal{B}$ *is closed,* the same holds true, provided $\lambda \in \mathbf{R}^m_+ = [\mathbf{R}^m_+]_*$ is replaced with $\lambda \in \mathbf{K}_*$. In this respect, note that $\mathcal{B}$ is the linear image of the closed cone $\mathbf{M} = \mathbf{R}^n_x \times \mathbf{K}$ under the linear mapping $[x; y] \mapsto \widehat{A}[x; y] \equiv Ax - y$, so that by Fact II.23 a *sufficient* condition for $\mathcal{B}$ to be closed is $\mathrm{Ker}\widehat{A} \cap M = \{0\}$, that is, $A^{-1}\mathbf{K} = \{0\}$.

$$Ax \geq_{\mathbf{K}} b \qquad \text{(I)}$$
$$\lambda \geq_{\mathbf{K}_*} 0,\ A^T\lambda = 0,\ \lambda^T b > 0 \qquad \text{(II)}$$
$$\mathcal{B} = A\mathbf{R}^n - \mathbf{K}$$

♠ In general, $\mathcal{B}$ can be non-closed, meaning that there is a "gap" between insolvability of (I) and solvability of (II) – both the systems can be infeasible.

Example: Let

$$Ax - b := [3; 4; 5] \cdot x - [4; -3; 0] = [3x - 4; 4x + 3; 5x] \geq_{\mathbf{L}^3} 0 \qquad (I)$$

A solution x should satisfy

$$25x^2 \geq \underbrace{(3x-4)^2 + (4x+3)^2}_{25x^2+25} \ \&\ 5x \geq 0$$

and clearly does not exist. However, with $b_\epsilon = [4; -3+\epsilon; 0]$, the inequality $Ax - b_\epsilon \geq_{\mathbf{L}^3} 0$ reads

$$25x^2 \geq \underbrace{(3x-4)^2 + (4x+3-\epsilon)^2}_{25x^2-8\epsilon x+16+(3-\epsilon)^2} \&\ x \geq 0$$

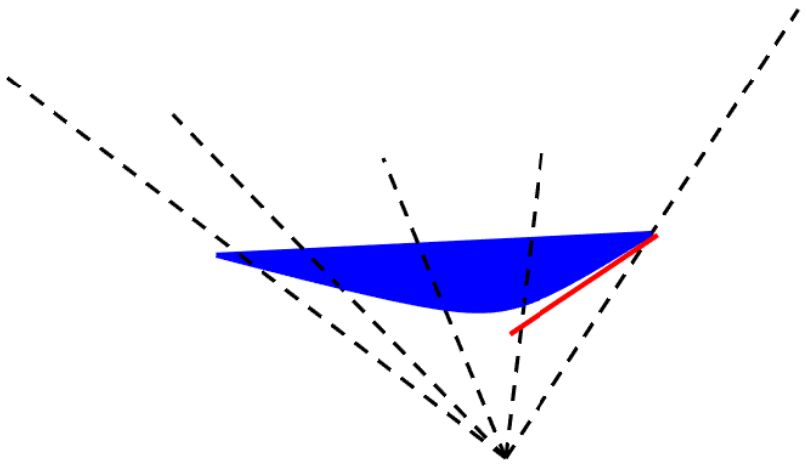
and becomes solvable whenever $\epsilon > 0$.

⇒ *(I) almost solvable, albeit insolvable*, which by Fact IX.4 implies that the alternative (II, which under the circumstances is the system of constraints

$$3\lambda_1 + 4\lambda_2 + 5\lambda_3 = 0,\ 4\lambda_1 - 3\lambda_2 > 0, \lambda_3 \geq \sqrt{\lambda_1^2 + \lambda_2^2}$$

in variables λ has no solutions (check that this indeed is the case!)

The geometry of this example is as follows. (I) wants to find a point in the intersection of a straight line ℓ in 3 (red line on the picture) D which happens to be an asymptote of (a branch of) the hyperbola (blue area on the picture) bounding the intersection of a 2D plane $\Pi \supset \ell$ with the ice-cream cone $\mathbf{L}^3$. This intersection is empty $\Rightarrow$ (I) is unsolvable. However, appropriate, whatever small, shifts of ℓ do intersect $\mathbf{L}^3$, making (I) almost solvable and thus making (II) infeasible.



Question: *When a scalar inequality*

$$c^T x \geq d \qquad (S)$$

is a consequence of a vector inequality

$$Ax \geq_{\mathbf{M}} b \qquad (V)$$

where $\mathbf{M}$ *is a regular cone ?*

Answer:

A. *If* (S) *can be obtained from* (V) *and the trivial inequality* $0 \geq -1$ *by "admissible linear aggregation:"*

$$\exists y \geq_{\mathbf{M}_*} 0 : A^T y = c \;\&\; y^T b \geq d, \qquad (*)$$

then (S) *is a consequence of* (V)*.*

This is evident.

B. *If* (S) *is a consequence of* (V) *and* (V) *satisfies the relaxed Slater condition –* $\mathbf{M}$ *can be decomposed as* $\mathbf{R}^m_+ \times \mathbf{K}$ *with regular cone* $\mathbf{K}$ *and* $A\bar{x} - b \in \mathbf{R}^m_+ \times \mathrm{int}\,\mathbf{K}$ *for some* $\bar{x}$*, then* (S) *can be obtained from* (V) *by admissible linear aggregation.*

Indeed, under the premise of **B** the conic problem

$$\mathrm{Opt}(P) = \min_x \left\{ c^T x : Ax \geq_{\mathbf{M}} b \right\}$$

satisfies the Relaxed Slater condition and $\mathrm{Opt}(P) \geq d$, implying by Conic Duality Theorem that the optimal solution to the dual problem

$$\mathrm{Opt}(D) = \max_y \left(b^T y : A^T y = c, y \in \mathbf{K}_* \right\} \qquad (D)$$

exists and satisfies $b^T y \geq d$, meaning that $(*)$ does take place.

Geometry of primal-dual pair of conic problems

♣ Consider a primal-dual pair of conic problems (we slightly change notation and format)

$$\begin{array}{rcll} \mathrm{Opt}(P) & = & \min_x \left\{c^T x : Ax - b \in \mathbf{K}, Rx - r = 0\right\} & (P) \\ & & \boxed{\begin{array}{c} Ax - b \in \mathbf{K} \ \&\ Rx = r \ \&\ y \in \mathbf{K}_* \ \&\ A^T y + R^T s = c \\ \Rightarrow c^T x \geq b^T y + r^T s \end{array}} & \\ \mathrm{Opt}(D) & = & \max_{y,s} \left\{b^T y + r^T s : y \in \mathbf{K}_*, A^T y + R^T s = c\right\} & (D) \end{array}$$

Assumption: *The systems of linear equality constraints in (P) and (D) are solvable:*

$$\exists \bar{x}, [\bar{y}, \bar{s}] : R\bar{x} = r,\ A^T\bar{y} + R^T\bar{s} = c.$$

♠ Let us pass in (P) from variable x to *primal slack* $\eta = Ax - b$. Whenever x satisfies $Rx = r$, we have

$$c^T x = [A^T\bar{y} + R^T\bar{s}]^T x = \bar{y}^T Ax + \bar{s}^T Rx = \bar{y}^T[Ax - b] + [b^T\bar{y} + r^T\bar{s}]$$

$\Rightarrow$ *(P) is equivalent to the conic problem*

$$\begin{array}{c} \mathrm{Opt}(\mathcal{P}) = \min\limits_{\eta} \left\{\bar{y}^T\eta : \eta \in [\mathcal{L} - \bar{\eta}] \cap \mathbf{K}\right\},\ \mathcal{L} = \{Ax : Rx = 0\},\ \bar{\eta} = b - A\bar{x} \\ \left[\mathrm{Opt}(\mathcal{P}) = \mathrm{Opt}(P) - [b^T\bar{y} + r^T\bar{s}]\right] \end{array} \qquad (\mathcal{P})$$

Explanation: (P) wants of $\eta := Ax - b$ (a) to belong to $\mathbf{K}$, and (b) to be representable as $Ax - b$ for some x satisfying $Rx = r$. (b) says that η should belong to the *primal affine plane* $\{Ax - b : Rx = r\}$, which is the shift of the parallel linear subspace $\mathcal{L} = \{Ax : Rx = 0\}$ by a (whatever) vector from the primal affine plane, e.g., the vector $-\bar{\eta} = A\bar{x} - b$.

$$\begin{array}{rcll}\mathrm{Opt}(P) & = & \min_x\left\{c^Tx : Ax-b\in\mathbf{K}, Rx=r\right\} & (P)\\ \mathrm{Opt}(D) & = & \max_{y,s}\left\{b^Ty+r^Ts : y\in\mathbf{K}_*, A^Ty+R^Ts=c\right\} & (D)\end{array}$$

♠ Let us pass in (D) from variables $[y;s]$ to variable y. Whenever $[y;s]$ satisfies $A^Ty+R^Ts=c$, we have

$$b^Ty+r^Ts=b^Ty+\bar{x}^TR^Ts=b^Ty+\bar{x}^T[c-A^Ty]=[b-A\bar{x}]^Ty+c^T\bar{x}=\bar{\eta}^Ty+c^T\bar{x},$$

$\Rightarrow$ *(D) is equivalent to the conic problem*

$$\begin{array}{c}\mathrm{Opt}(\mathcal{D})=\max_y\left\{\bar{\eta}^Ty : y\in[\mathcal{L}^\perp+\bar{y}]\cap\mathbf{K}_*\right\}\\ \left[\mathrm{Opt}(\mathcal{D})=\mathrm{Opt}(D)-c^T\bar{x}\right]\end{array}\qquad(\mathcal{D})$$

Explanation: (D) wants of y (a) to belong to $\mathbf{K}_*$, and (b) to satisfy $A^Ty=c-R^Ts$ for some s. (b) says that y should belong to the *dual affine plane* $\{y:\exists s: A^Ty+R^Ts=c\}$, which is the shift of the parallel linear subspace $\widetilde{\mathcal{L}}=\{y:\exists s: A^Ty+R^Ts=0\}$ by a (whatever) vector from the dual affine plane, e.g., the vector $\bar{y}$.

Elementary Linear Algebra says that $\widetilde{\mathcal{L}}=\mathcal{L}^\perp$. Indeed,

$$\begin{array}{rcl}[\widetilde{\mathcal{L}}]^\perp & = & \{z : z^Ty=0\,\forall y:\exists s: A^Ty+R^Ts=0\}=\{z : z^Ty+0^Ts=0 \text{ whenever } A^Ty+R^Ts=0\}\\ & = & \{z:\exists x:[z^T,0]=x^T[A^T,R^T]\}=\{z:\exists x: Ax=z, Rx=0\}=\mathcal{L}.\end{array}$$

$$\begin{array}{rcll}\mathrm{Opt}(P) & = & \min_x\left\{c^Tx : Ax-b\in\mathbf{K}, Rx=r\right\} & (P)\\ \mathrm{Opt}(D) & = & \max_{y,s}\left\{b^Ty+r^Ts : y\in\mathbf{K}_*, A^Ty+R^Ts=c\right\} & (D)\end{array}$$

♣ **Bottom line:** *Problems (P), (D) are equivalent, respectively, to*

$$\begin{array}{rcll}\mathrm{Opt}(\mathcal{P}) & = & \min_\eta\left\{\bar{y}^T\eta : \eta\in[\mathcal{L}-\bar{\eta}]\cap\mathbf{K}\right\} & (\mathcal{P})\\ \mathrm{Opt}(\mathcal{D}) & = & \max_y\left\{\bar{\eta}^Ty : y\in[\mathcal{L}^\perp+\bar{y}]\cap\mathbf{K}_*\right\} & (\mathcal{D})\end{array}$$
$$\left[\mathcal{L}=\{Ax : Rx=0\},\ R\bar{x}=r,\ \bar{\eta}=b-A\bar{x},\ A^T\bar{y}+R^T\bar{s}=c\right]$$

Note: *When x is feasible for (P), and $[y;s]$ is feasible for (D), the vectors $\eta=Ax-b$, y are feasible for $(\mathcal{P})$, resp., $(\mathcal{D})$, and*

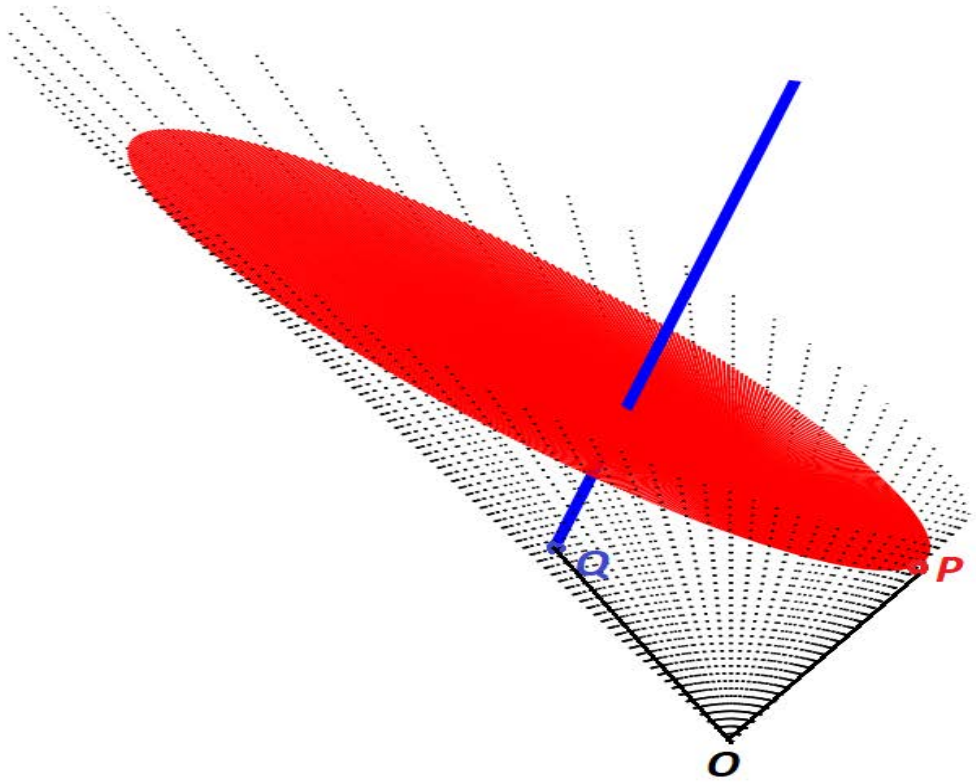
$$\mathrm{DualityGap}(x;[y;s]):=c^Tx-b^Ty-r^Ts=[A^Ty+R^Ts]^Tx-b^Ty-r^Ts=[Ax-b]^Ty=\eta^Ty$$

⇒ *Geometrically, (P), (D) are as follows: "geometric data" of the problems are the pair of linear subspaces $\mathcal{L}$, $\mathcal{L}^\perp$ in the space where $\mathbf{K}$, $\mathbf{K}_*$ live, the subspaces being orthogonal complements to each other, and pair of vectors $\bar{\eta}$, $\bar{y}$ in this space.*

- *(P) is equivalent to minimizing $f(\eta)=\bar{y}^T\eta$ over the intersection of $\mathbf{K}$ and the primal feasible plane $\mathcal{M}_P$ which is the shift of $\mathcal{L}$ by $-\bar{\eta}$*
- *(D) is equivalent to maximizing $g(y)=\bar{\eta}^Ty$ over the intersection of $\mathbf{K}_*$ and the dual feasible plane $\mathcal{M}_D$ which is the shift of $\mathcal{L}^\perp$ by $\bar{y}$*
- *taken together, (P) and (D) form the problem of minimizing the duality gap over feasible solutions to the problems, which is exactly the problem of finding pair of vectors in $\mathcal{M}_P\cap\mathbf{K}$ and $\mathcal{M}_D\cap\mathbf{K}_*$ as close to orthogonality as possible.*

♣ Pay attention to the ideal geometrical primal-dual symmetry we observe.

♣ **Illustration:** Primal-dual pair of conic problems on 3D Lorentz cone



Red: feasible set of $(\mathcal{P})$ Blue: feasible set of $(\mathcal{D})$
P – optimal solution to $(\mathcal{P})$; Q – optimal solution to $(\mathcal{D})$.

- Pay attention to orthogonality of $\overrightarrow{OP}$ to $\overrightarrow{OQ}$.

Conic Duality – Postscriptum

♣ **So far**, our "universes" - the linear spaces where all the actions are taking place – were the spaces $\mathbf{R}^n$ of n-dimensional column vectors *equipped with the standard inner product* $\mathbf{R}^n \ni x, y \mapsto x^T y \in \mathbf{R}$. This inner product was used on many occasions, e.g.

- when representing a linear mapping as multiplication by the corresponding matrix
- when representing a linear form as the inner product with an appropriate vector
- when defining the cone dual to a given cone
- when representing the derivative of a function by its gradient – the vector with inner products with directions being the corresponding directional derivatives of the function,
- etc., etc.

♣ **In fact**, universes we can handle are *Euclidean spaces* – finite-dimensional linear spaces E over reals equipped with inner products $E \ni x, y \Rightarrow \langle x, y\rangle_E \in \mathbf{R}$ with bilinear function $\langle x, y\rangle_E$ (linear in x for y fixed and linear in y for x fixed) which is symmetric: $\langle x, y\rangle \equiv \langle y, x\rangle$ and "positive on the diagonal" $\langle x, x\rangle_E > 0$ whenever $x \neq 0$.

♠ Linear Algebra teaches that *every Euclidean space E is equivalent to appropriate* $\mathbf{R}^n$, meaning that for properly selected n (namely, equal to the linear dimension of E) there exists one-to-one correspondence between E and $\mathbf{R}^n$ which preserves linear operations and converts the inner product in E into the standard inner product on $\mathbf{R}^n$; to establish this correspondence, it suffices to build an orthonormal basis $e_1, .., e_n$ in E:

$$\langle e_i, e_j\rangle_E = \begin{cases} 0 & , i \neq j \\ 1 & , i = j \end{cases}$$

which is always possible, and to pass from a vector from E to the vector of its coefficients in this basis.

♠ We tacitly used the possibility to "standardize" our universes when it made no harm. When speaking about the cone of positive semidefinite matrices and conic programs on this cone, this approach becomes inconvenient. $\mathbf{S}^n_+$ lives in the Euclidean space $\mathbf{S}^n$ of symmetric $n \times n$ matrices equipped with the Frobenius inner product

$$\langle x, y \rangle := \mathrm{Tr}(xy) = \sum_{i,j=1}^n x_{ij} y_{ij}$$

♠ Of course, we could without any difficulty build an orthonormal basis in this space and identify $\mathbf{S}^n$ with $\mathbf{R}^{\frac{n(n+1)}{2}}$, but what for? All we could get is notational havoc... It is much better to express all our constructions in terms of linear operations and inner product, and understand what they become when dealing with $\mathbf{S}^n$. This is what happens:

• Linear mappings from $\mathbf{R}^k$ to $\mathbf{S}^n$ (for our needs, this is enough) are not represented by multiplication by matrices; their natural representation is

$$x \mapsto \mathcal{A}x = \sum_{i=1}^k x_i A_i, \qquad (*)$$

where $A_i \in \mathbf{S}^n$ are "columns" of the mapping (cf. the usual representation $Ax = \sum_k x_i \mathrm{Col}_i[A]$ of a linear mapping from $\mathbf{R}^k$ to $\mathbf{R}^n$);

• Every linear mapping $x \mapsto \mathcal{A}x$ from Euclidean space $E, \langle \cdot, \cdot \rangle_E$ to Euclidean space $F, \langle \cdot, \cdot \rangle_F$ has its *conjugate* – linear mapping $y \mapsto \mathcal{A}^* y : F \to E$ given by the identity

$$\langle y, \mathcal{A}x \rangle_F \equiv \langle \mathcal{A}^* y, x \rangle_E.$$

Note: As is immediately seen, *the mapping conjugate to* $(*)$ *is*

$$y \mapsto \mathcal{A}^* y \equiv [\mathrm{Tr}(A_1 y); ...; \mathrm{Tr}(A_k y)] : \mathbf{S}^n \to \mathbf{R}^k.$$

• The conjugate to a linear combination of linear mappings is the same linear combination of their conjugates, the conjugate to the product (composition) $\mathcal{A}\mathcal{B}$ of two mappings: $\mathcal{A}\mathcal{B}(x) \equiv \mathcal{A}(\mathcal{B}(x))$ is $\mathcal{B}^*\mathcal{A}^*$, and $\mathcal{A}$ is the conjugate to $\mathcal{A}^*$.

• When E, F are $\mathbf{R}^n$, $\mathbf{R}^m$ with the standard inner products, and we represent a linear map $A : E \to F$ as multiplication by $m \times n$ matrix: $\mathcal{A}(x) = Ax$, $\mathcal{A}^*$ is represented by A^T: $\mathcal{A}^*(y) = A^T y$, and this is where the transposes of matrices come to our constructions.

- The dual of a cone $K \subset E$ becomes

$$\mathbf{K}_* = \{y \in E : \langle y, x\rangle_E \geq 0 \, \forall x \in \mathbf{K}\}.$$

It is immediately seen that *the positive semidefinite cone* $\mathbf{S}^n_+ = \{X \in \mathbf{S}^n : X \succeq 0\}$ *is self-dual:*

$$\{Y \in \mathbf{S}^n : \mathsf{Tr}(YX) \geq 0 \; \forall X \in \mathbf{S}^n_+\} = \mathbf{S}^n_+$$

♣ Equipped with broader sight, let us build the dual to the SDP conic problem

$$\mathrm{Opt}(P) = \min_{x\in\mathbf{R}^n}\left\{c^Tx : Ax \geq b,\, \mathcal{A}_ix \equiv \sum_j x_jA_j^i \succeq B^i,\, 1\leq i\leq m\right\} \qquad (P)$$
$$\left[A_j^i, B^i \in \mathbf{S}^{m_i}\right]$$

To build the dual, we

• equip the system $Ax \geq 0$ of scalar inequalities with Lagrange multiplier $\lambda \in \mathbf{R}_+^{\dim b}$, and LMI's – with Lagrange multipliers $\Lambda_i \in \mathbf{S}_+^{m_i}$

• take the inner products of both sides of our constraints with the corresponding Lagrange multipliers and sum the results up, thus arriving at the scalar linear inequality

$$\lambda^TAx + \sum_i \mathrm{Tr}(\Lambda_i\mathcal{A}_ix) \geq b^T\lambda + \sum_i \mathrm{Tr}(B_i\lambda) \qquad (*)$$

which, by its origin, is a consequence of the system of constraints of (P)

• impose on the Lagrange multipliers the restriction for the left hand side in $(*)$ to be, as a function of x, identically equal to c^Tx and maximize under this restriction (and initial "sign restrictions") on the multipliers the right hand side in $(*)$, thus arriving at the dual problem

$$\mathrm{Opt}(D) = \max_{\lambda,\Lambda_i}\left\{b^T\lambda + \sum\nolimits_i \mathrm{Tr}(B_i\Lambda_i) : \lambda \geq 0, \Lambda_i \succeq 0, A^T\lambda + \sum_i \mathcal{A}_i^*\Lambda_i = c\right\} \qquad (D)$$

where for a linear mapping $x \mapsto \sum_j x_jA_j : \mathbf{R}^n \to \mathbf{S}^N$. the conjugate mapping $Y \mapsto \mathcal{A}^*Y$ is given by

$$Y \mapsto \mathcal{A}^*Y = [\mathrm{Tr}(A_1Y); ...; \mathrm{Tr}(A_nY)] : \mathbf{S}^N \to \mathbf{R}^n.$$

Truss Topology Design and Conic Duality

♣ The TTD problem reads

$$\mathrm{Opt}(P) = \min_{\tau,r}\left\{\tau : \left[\begin{array}{c|c} B\mathrm{Diag}\{t\}B^T & f \\ \hline f^T & 2\tau \end{array}\right] \succeq 0, t \geq 0, \sum_i t_i = W\right\} \qquad (P)$$
$$\left[B = [\mathfrak{b}_1, ..., \mathfrak{b}_N] \in \mathbf{R}^{M\times N},\ BB^T \succ 0\right]$$

(P) is a conic problem involving the semidefinite cone $\mathbf{S}_+^{M+1}$ that is self-dual w.r.t. the inner product

$$\langle A, B\rangle = \mathrm{Tr}(AB) = \sum_{i,j} A_{ij}B_{ij}$$

on the space $\mathbf{S}^{M+1}$.

A. It is easily seen that $BB^T \succ 0 \Rightarrow$ *the Relaxed Slater condition holds & (P) is solvable.*

B. Passing from problem (P) to its conic dual (this is a purely mechanical process), we arrive at the problem

$$\max_{V,g,\theta,\lambda,\mu}\left\{-2f^Tg - W\mu : 2\theta = 1, \mathfrak{b}_i^TV\mathfrak{b}_i + \lambda_i - \mu = 0\,\forall i, \lambda \geq 0, \left[\begin{array}{c|c} V & g \\ \hline g^T & \theta \end{array}\right] \succeq 0\right\}$$

Eliminating variable θ (fixed by the constraint $2\theta = 1$), and variables λ_i, the dual becomes

$$\mathrm{Opt}(D) = \max_{V,g,\mu}\left\{-2f^Tg - W\mu : \mathfrak{b}_i^TV\mathfrak{b}_i \leq \mu = 0\,\forall i, \lambda \geq 0, \left[\begin{array}{c|c} V & g \\ \hline g^T & \frac{1}{2} \end{array}\right] \succeq 0\right\} \qquad (D)$$

By Conic Duality Theorem, *(D) is solvable, and* $\mathrm{Opt}(P) = \mathrm{Opt}(D)$.

$$\mathrm{Opt}(D) = \max_{V,g,\mu}\left\{-2f^Tg - W\mu : \mathfrak{b}_i^T V \mathfrak{b}_i \le \mu = 0\,\forall i,\ \left[\begin{array}{c|c} V & g \\ \hline g^T & \frac{1}{2}\end{array}\right] \succeq 0\right\} \qquad (D)$$

To proceed, we need

Fact IX.5 [Schur Complement Lemma] *Consider symmetric block-matrix* $\left[\begin{array}{c|c} Q & Q \\ \hline Q^T & R\end{array}\right]$ *with* $R \succ 0$, *Then*

$$\left[\begin{array}{c|c} P & Q \\ \hline Q^T & R\end{array}\right] \succeq 0 \Leftrightarrow P - OR^{-1}Q^T \succeq 0.$$

Indeed,

$$\left[\begin{array}{c|c} Q & Q \\ \hline Q^T & R\end{array}\right] \succeq 0 \Leftrightarrow [u;v]^T \left[\begin{array}{c|c} Q & Q \\ \hline Q^T & R\end{array}\right] [u;v] \equiv u^TPu + 2u^TQv + v^TRv \ge 0 \forall u,v$$
$$\Leftrightarrow u^TPu + \min_v[2u^TQv + v^TRv] \equiv u^TPu - u^TQR^{-1}Q^Tu \ge 0\,\forall u \Leftrightarrow [P - QR^{-1}Q^T] \succeq 0$$

♠ The SCL allows to eliminate in (D) the matrix variable V – by the SCL, in a feasible solution to (D) one has $V \succeq \overline{V} = 2gg^T$, and replacing V with $\overline{V}$, we preserve feasibility and keep the value of the objective intact. The resulting problem reads

$$\max_{g,\mu}\left\{-2f^Tg - W\mu : \mu \ge 2[\mathfrak{b}_i^Tg]^2\,\forall i\right\}$$

or, again applying the SCL,

$$\mathrm{Opt}(\overline{D}) = \max_{g,\mu}\left\{-2f^Tg - W\mu : \left[\begin{array}{c|c} \mu & \mathfrak{b}_i^Tg \\ \hline \mathfrak{b}_i^Tg & \frac{1}{2}\end{array}\right] \succeq 0\,\forall i\right\} \qquad (\overline{D})$$

which is solvable with $\mathrm{Opt}(\overline{D}) = \mathrm{Opt}(D) = \mathrm{Opt}(P)$. Besides, it is immediately seen that $(\overline{D})$ satisfies the Relaxed Slater condition.

Surprise # 1: *The constraints in* $(\overline{D})$ *involve the cone* $\mathbf{S}^2_+$ *only, and this cone, up to one-to-one linear transformation of* $\mathbf{S}^2 = \mathbf{R}^3$, *is the 3D Lorentz cones! Thus,* $(\overline{D})$ *is a Conic Quadratic problem...*

$$\mathrm{Opt}(\overline{D}) = \max_{g,\mu}\left\{-2f^Tg - W\mu : \left[\begin{array}{c|c}\mu & \mathfrak{b}_i^Tg \\ \hline \mathfrak{b}_i^Tg & \frac{1}{2}\end{array}\right] \succeq 0\,\forall i\right\} \qquad (\overline{D})$$

C. Let us pass from $(\overline{D})$ to *its* conic dual.

Note: Were $(\overline{D})$ the conic dual (D) of our original problem (P), the result would be known in advance: by the symmetry of conic duality, we wold come bask to (P). However, $(\overline{D})$ is not (D), it was obtained from (D) by partial optimization in variables V and λ, and nobody knows in advance what this reduction does with the dual... Well, immediate purely mechanical computation says that the dual to $(\overline{D})$ is the problem

$$\mathrm{Opt}(\overline{P}) = \min_{s,t,q}\left\{\frac{1}{2}\sum_i s_i : \sum_i t_i = W, \sum_i q_i\mathfrak{b}_i = f, \left[\begin{array}{c|c}t_i & q_i \\ \hline q_i & s_i\end{array}\right] \succeq 0\,\forall i\right\} \qquad (\overline{P})$$

By Conic Duality Theorem, $(\overline{P})$ is solvable, and $\mathrm{Opt}(\overline{P}) = \mathrm{Opt}(\overline{D}) = \mathrm{Opt}(D) = \mathrm{Opt}(P)$.

$$\mathrm{Opt}(\overline{P}) = \min_{s,t,q}\left\{\frac{1}{2}\sum_i s_i : \sum_i q_i \mathfrak{b}_i = f, \left[\begin{array}{c|c} t_i & q_i \\ \hline q_i & s_i \end{array}\right] \succeq 0 \,\forall i, \sum_i t_i = W\right\} \qquad (\overline{P})$$

D. Let us compare $(\overline{P})$ with our initial form of the TTD problem

$$\mathrm{Opt}(P) = \min_{\tau,r}\left\{\tau : \left[\begin{array}{c|c} B\mathrm{Diag}\{t\}B^T & f \\ \hline f^T & 2\tau \end{array}\right] \succeq 0, t \geq 0, \textstyle\sum_i t_i = W\right\} \qquad (P)$$

While a candidate truss t in (P) and variable t in $(\overline{P})$ are of the same dimension, and the constraints of $(\overline{P})$ and those of (P) impose the same restrictions $t \geq 0$, $\sum_i t_i = W$ on t, nobody told us that (P) and $(\overline{P})$ equally well model the design of the optimal truss. What we know is that

for a given $t \geq 0$ with $\sum_i t_i = W$, the smallest τ which, taken together with t, yields a feasible solution to (P), is the compliance $\mathrm{Compl}(t, f)$, so that (P) is the problem of minimizing $\mathrm{Compl}(t, f)$ over trusses t of total bar volume W.

Similarly,

denoting by $\overline{\mathrm{Compl}}(t, f)$ the minimal, over s, q resulting in a feasible for $(\overline{P})$ triple (s, t, q), value of the objective $\frac{1}{2}\sum_i s_i$ of $(\overline{P})$, $(\overline{P})$ is the problem of minimizing $\overline{\mathrm{Compl}}(t, f)$ over trusses t of total bar volume W.

All we do know is that *the optimal values in (P) and $(\overline{P})$ are the same*, which in no sense guarantees that $\mathrm{Compl}(t, f) \equiv \overline{\mathrm{Compl}}(t, f)$, and that solving $(\overline{P})$ to optimality, we indeed have minimized $\mathrm{Compl}(t, f)$ over t.

However, an educated *guess* is that $\mathrm{Opt}(P) = \mathrm{Opt}(\overline{P})$ is not a coincidence. And indeed, a not too difficult reasoning heavily utilizing Optimality conditions yields

Surprise # 2: *One has $\overline{\mathrm{Compl}}(t, f) \equiv \mathrm{Compl}(t, f)$.*

$$\mathrm{Opt}(\overline{P}) = \min_{s,t,q}\left\{\frac{1}{2}\sum_i s_i : \sum_i q_i\mathfrak{b}_i = f, \left[\begin{array}{c|c} t_i & q_i \\ \hline q_i & s_i \end{array}\right] \succeq 0\,\forall i, \sum_i t_i = W\right\} \qquad (\overline{P})$$

E. Observe that the nonlinear constraints $\left[\begin{array}{c|c} t_i & q_i \\ \hline q_i & s_i \end{array}\right] \succeq 0$ are nothing but the inequalities $q_i^2/t_i \le s_i$, where the convex function a^2/b is extended from the its natural domain – the half-plane $b > 0$ – onto the entire (a,b)-plane "by lover semicontinuity" - by setting $0^2/0 = 0$,$a^2/0 = +\infty$ for $a \neq 0$, and $a^2/b = +\infty$ for $b < 0$. As a result, $(\overline{P})$ becomes the problem

$$\min_{t,q}\left\{\frac{1}{2}\sum_i q_i^2/t_i : \sum_i q_i\mathfrak{b}_i = f, t \ge 0, \sum_i t_i = W\right\} \qquad (R)$$

(R) admits partial optimization in t, resulting in the problem

$$\min_q\left\{\frac{1}{2W}\left[\sum_i |q_i|\right]^2 : \sum_i q_i\mathfrak{b}_i = f\right\} \qquad (!)$$

Surprise # 3: (!) *is nothing but the LP problem*

$$\min_q\left\{\sum_i |q|_i : \sum_i q_i\mathfrak{b}_i = f\right\}$$

This is a kind of miracle - we reduced highly nonlinear problem to simply-looking LP.
ГЛЯДЯ НА МИР, НЕЛЬЗЯ НЕ УДИВЛЯТЬСЯ!
[Looking at the world, one cannot help but be surprised! – The 110th aphorism from the collection of thoughts and aphorisms "The Fruits of Thought" (1854) by Kozma Prutkov. "Kozma Prutkov was a collective pseudonym created by several Russian writers (Aleksey Konstantinovich Tolstoy and the Zhemchuzhnikov brothers), the character they created was known for his pompous, often absurd, and yet strangely insightful pronouncements. This particular aphorism, while seemingly simple, has become one of their most enduring and recognizable creations." – Gemini]

♣ **The "LP miracle,"** whatever surprising, is of restricted interest – it vanishes when passing from the simplest single-load TTD to multi-load TTD and to other problems (like *Shape Design*) of optimal design of linearly elastic mechanical structures. What *is* of actual interest, is the alternative description of the compliance

$$\mathrm{Compl}(t,f) = \inf_q \left\{ \frac{1}{2}\sum\nolimits_i q_i^2/t_i : \sum\nolimits_i q_i \mathfrak{b}_i = f \right\} \qquad (!)$$

Question: *Can we derive* (!) *from the first principles of Mechanics, circumventing mathematical stuff like taking twice conic duality, partial minimization, etc. ?*
Answer: *Yes and no*

To understand what is going on, let us start with deriving the TTD model.

♣ Consider the design of d-dimensional truss ($d = 2$ or $d = 3$)

♠ Let m be the number of free nodes in the original nodal grid, and let us assign the nodes *serial numbers* α in such a way that the numbers of the m free nodes are $1, 2, ..., m$, and let $p_\alpha \in \mathbf{R}^d$ be the pre-deformation position of node with serial number α ("node α" for short).

• Recall that the vectors v (nodal displacement), f (external load), and g (*reaction*) are $M = md$-dimensional block vectors with m blocks of dimension d indexed by serial numbers of the free nodes; the α-th blocks $v_\alpha, f_\alpha, g_\alpha$ of these vectors are, respectively, physical displacements, external forces, and reaction forces corresponding to node α.

♠ Consider i-th bar; assume it links free nodes $\alpha = \alpha(i)$ and $\beta = \beta(i)$, and let $\ell = \ell(i) = \|p_{\alpha(i)} - p_{\beta(i)}\|_2$ be the length of the bar, $\mathbf{e} = \mathbf{e}(i) = [p_{\alpha(i)} - p_{\beta(i)}]/\ell(i)$ be the pre-deformation direction of the bar, and $S = S(i)$ be the $(d-1)$-dimensional cross-sectional size of the bar $\Rightarrow$ the d-dimensional volume of the bar is

$$t_i = S(i)\ell(i)$$

• As a result of nodal displacement v, *in linear in v approximation of the actual physical phenomena, and with properly selected length and force units,*

— the post-deformation length of the bar is $\ell + \delta$, $\delta = [v_\alpha - v_\beta]^T\mathbf{e}$

— the caused by deformation reaction force at node α is $-\sigma S\mathbf{e}$, where the *stress* σ, by Hooke's Law, is

$$\sigma = \frac{\delta}{\ell} = \frac{[v_\alpha - v_\beta]^T\mathbf{e}}{\ell}$$

The reaction force at node β is, of course, $\sigma S\mathbf{e}$

— the potential energy capacitated in the bar as the result of its deformation is

$$\tfrac{1}{2}[S\sigma]\delta = \tfrac{1}{2}S\delta^2/\ell$$

• Let us define $\mathfrak{b}_i$ as the block-vector with m blocks, the nonzero ones being the α-th, equal to $\mathbf{e}(i)/\ell$, and β-th, equal to $-\mathbf{e}(i)/\ell$ Denoting by g_i the contribution of bar i to the reaction g of the truss caused by deformation, g_i is block-vector with two nonzero blocks, α-th and β-th, given by

$$[g_i]_\alpha = -[g_i]_\beta = -S\sigma\mathbf{e} = S\tfrac{\delta}{\ell}\mathbf{e} = -S\tfrac{[v_\alpha - v_\beta]^T\mathbf{e}}{\ell}\mathbf{e} = -[S\ell]\left[[v_\alpha - v_\beta]^T\mathbf{e}/\ell\right]\mathbf{e}/\ell = -t_i[\mathfrak{b}_i^T v][\mathfrak{b}_i]_\alpha.$$

This expression for the "physical" reaction forces caused by deformation of i-th bar at its end-nodes α, β was derived when both nodes α, β are free. If only one of them, say, α, is free, the expression still works, provided v_β is set to 0 and $\mathfrak{b}_i$ is defined as the block-vector with just α-th block nonzero (and equal to $\mathbf{e}(i)/\ell$). Thus, the contribution of i-th bar to the overall reaction g of the truss is $g_i = -t_i[\mathfrak{b}_i\mathfrak{b}_i^T]v$, whence

$$g = -\left[\sum\nolimits_i t_i\mathfrak{b}_i\mathfrak{b}_i^T\right]v$$

as required in the TTD model.

Let us define *scaled tension in bar* i as $q_i = t_i \mathfrak{b}_i^T v$ While defined in terms of a displacement v of the entire nodal grid, q_i depends solely on the "physical " displacements of the nodes linked by the bar:

$$q_i = t_i[\mathbf{e}(i)]^T[v_{\alpha(i)} - v_{\beta(i)}]/\ell(i).$$

Note that

$$g_i = -q_i \mathfrak{b}_i, \; g = -\sum_i q_i \mathfrak{b}_i$$

and the energy capacitated in bar i is

$$\tfrac{1}{2}S(i)\delta^2(i)/\ell(i) = \tfrac{1}{2}S(i)\left([v_{\alpha(i)} - v_{\beta(i)}]^T\mathbf{e}(i)\right)^2/\ell(i) = \tfrac{1}{2}\underbrace{\ell(i)S(i)}_{t_i}[\mathfrak{b}_i^T v]^2 = \tfrac{1}{2}q_i^2/t_i,$$

Note: q_i has a transparent mechanical sense: it is the $\ell(i)$ times minus tension (minus reaction force acting the end-node $\alpha(i)$ of bar i). In terms of these scaled tensions, the equilibrium deformation under load f and the compliance $\mathrm{Compl}(t, f)$ – the capacitated in the truss under this deformation potential energy – are given, respectively, by $\sum_i q_i \mathfrak{b}_i = f$ and $\frac{1}{2}\sum_i q_i^2/t_i$.

♠ Have we reached our goal to derive the representation

$$\mathrm{Compl}(t, f) = \inf_q \left\{ \frac{1}{2}\sum_i q_i^2/t_i : \sum_i q_i \mathfrak{b}_i = f \right\} \qquad (!)$$

of the compliance from the first principles? — Absolutely not!

♠ Have we reached our goal to derive the representation

$$\mathrm{Compl}(t,f) = \inf_q \left\{ \frac{1}{2}\sum_i q_i^2/t_i : \sum_i q_i \mathfrak{b}_i = f \right\} \qquad (!)$$

of the compliance from the first principles? — Absolutely not!

• Of course, every real q_i can be obtained as the scaled tension of i-th bar under certain displacements of its nodes, same as any 3D vector can be obtained as the velocity of aircraft's passenger John Doe. However, all passengers of an aircraft are flying with the same velocity, same as for t given, the N scaled tensions of bars stem from a displacement of the nodal grid and thus form the linear image of the M-dimensional vector v of nodal displacements. Thus, *mechanically meaningful scaled tensions q_i cannot be treated as independent decision variables as we see them in* (!) (think of Console design where $N = 3024$ and $M = 72$). "Nearly all" feasible solutions to (!) have nothing to do with the displacements of the nodal grid and thus make no actual mechanical sense.

However: *At the optimum, q_i do come from nodal displacements!* These displacements are the Lagrange multipliers for the equality constraints in (!) certifying optimality of the optimal solution.

As a matter of fact, to the best of my knowledge, no one was brave (or crazy) enough to arrive at (!) "from scratch" – from considerations originating in Mechanics, and the outlined circumstantial way to discover the *bar-force* formulation ($\overline{P}$) of the TTD problem (and similar formulations of other Structural Design problems), and ($\overline{P}$) was discovered exactly as explained - by twice taking conic duals and eliminating variables...

Saddle Points

♣ We have spoken about saddle points of the Lagrange function. It is time to focus on saddle points *per se*.

♣ Let $X \subset \mathbf{R}^n$, $\Lambda \subset \mathbf{R}^m$ be nonempty sets, and let $F(x, \lambda)$ be a real-valued function on $X \times \Lambda$. This function gives rise to two optimization problems

$$\mathrm{Opt}(P) = \inf_{x \in X} \overbrace{\sup_{\lambda \in \Lambda} F(x, \lambda)}^{\overline{F}(x)} \quad (P)$$

$$\mathrm{Opt}(D) = \sup_{\lambda \in \Lambda} \underbrace{\inf_{x \in X} F(x, \lambda)}_{\underline{F}(\lambda)} \quad (D)$$

Game interpretation: Player I chooses $x \in X$, player II chooses $\lambda \in \Lambda$. With choices of the players x, λ, player I pays to player II the sum $F(x, \lambda)$. What should the players do to optimize their wealth?

♢If Player I chooses x first, and Player II knows this choice when choosing λ, II will maximize her profit, and the loss of I will be $\overline{F}(x)$. To minimize her loss, Player I should solve (P), thus ensuring herself loss $\mathrm{Opt}(P)$ or less.

♢If Player II chooses λ first, and Player I knows this choice when choosing x, Player I will minimize her loss, and the profit of Player II will be $\underline{F}(\lambda)$. To maximize her profit, Player II should solve (D), thus ensuring herself profit $\mathrm{Opt}(D)$ or more.

$$\mathrm{Opt}(P) = \inf_{x\in X} \overbrace{\sup_{\lambda\in\Lambda} F(x,\lambda)}^{\overline{F}(x)} \quad (P)$$
$$\mathrm{Opt}(D) = \sup_{\lambda\in\Lambda} \underbrace{\inf_{x\in X} F(x,\lambda)}_{\underline{F}(\lambda)} \quad (D)$$

For Player I, the second situation seems better, so that it is natural to guess that her anticipated loss in this situation is $\leq$ her anticipated loss in the first situation:

$$\mathrm{Opt}(D) \equiv \sup_{\lambda\in\Lambda} \inf_{x\in X} F(x,\lambda) \leq \inf_{x\in X} \sup_{\lambda\in\Lambda} F(x,\lambda) \equiv \mathrm{Opt}(P).$$

This indeed is true:

Fact IX.6 *One has* $\mathrm{Opt}(D) \leq \mathrm{Opt}(P)$.

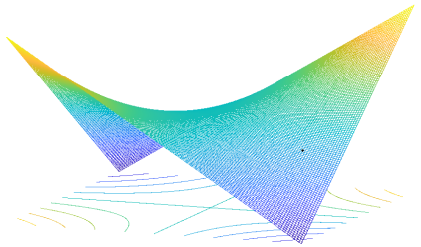
Indeed, the claim is trivially true when $\mathrm{Opt}(P) = \infty$. When $\mathrm{Opt}(P) < \infty$, for every real $a > \mathrm{Opt}(P)$ there exists $x_a \in X$ such that $F(x_a,\lambda) \leq a$ for all $\lambda \Rightarrow \underline{F}(\lambda) \leq \mathrm{Opt}(P)$ for all $\lambda \Rightarrow \mathrm{Opt}(D) \leq a$. Thus, $\mathrm{Opt}(D)$ is $\leq a$ whenever real a is $>$ real $\mathrm{Opt}(P)$, implying $\mathrm{Opt}(D) \leq \mathrm{Opt}(P)$, $\square$

$$\mathrm{Opt}(P) = \inf_{x\in X} \overbrace{\sup_{\lambda\in\Lambda} F(x,\lambda)}^{\overline{F}(x)} \quad (P)$$
$$\mathrm{Opt}(D) = \sup_{\lambda\in\Lambda} \underbrace{\inf_{x\in X} F(x,\lambda)}_{\underline{F}(\lambda)} \quad (D)$$

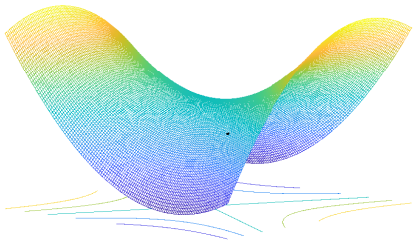
♣ What should the players do when making their choices simultaneously?
A "good case" when we can answer this question is when F has a *saddle point*.
Definition: *We call a point* $(x_*, \lambda_*) \in X \times \Lambda$ *a saddle point of* F*, if*

$$F(x, \lambda_*) \geq F(x_*, \lambda_*) \geq F(x_*, \lambda) \ \forall (x \in X, \lambda \in \Lambda).$$

In game terms, a saddle point is an *equilibrium* – no one of the players can improve her wealth, provided the adversary keeps her choice unchanged.



$F(x,\lambda) = -x\lambda$



$F(x,\lambda) = x^2 - \lambda^2 + x\lambda$

In both cases, $F(x,0) \geq F(0,0) \geq F(0,\lambda) \Rightarrow (0,0)$ is a saddle point

$$\mathrm{Opt}(P) = \inf_{x\in X} \overbrace{\sup_{\lambda\in\Lambda} F(x,\lambda)}^{\overline{F}(x)} \quad (P)$$
$$\mathrm{Opt}(D) = \sup_{\lambda\in\Lambda} \underbrace{\inf_{x\in X} F(x,\lambda)}_{\underline{F}(\lambda)} \quad (D)$$

Fact IX.7 [Existence and structure of saddle points] *F has a saddle point iff both (P) and (D) are solvable with equal to each other real optimal values. In this case, the saddle points of F are exactly the pairs (x_*,λ_*), where x_* is an optimal solution to (P), and λ_* is an optimal solution to (D). In addition, at every saddle point, (x_*,λ_*), $F(x_*,\lambda_*)$ equals to the common value of $\mathrm{Opt}(P)$ and $\mathrm{Opt}(D)$.*

Proof. ✓ Assume that (x_*,λ_*) is a saddle point of F, and let us prove that x_* solves (P), λ_* solves (D), and $\mathrm{Opt}(P) = \mathrm{Opt}(D)$.
Indeed, we have

$$F(x,\lambda_*) \geq F(x_*,\lambda_*) \geq F(x_*,\lambda) \;\forall (x\in X, \lambda\in\Lambda)$$

whence

$$\mathrm{Opt}(P) \leq \overline{F}(x_*) = \sup_{\lambda\in\Lambda} F(x_*,\lambda) = F(x_*,\lambda_*)$$
$$\mathrm{Opt}(D) \geq \underline{F}(\lambda_*) = \inf_{x\in X} F(x,\lambda_*) = F(x_*,\lambda_*)$$

Since $\mathrm{Opt}(P) \geq \mathrm{Opt}(D)$, we see that all inequalities in the chain

$$\mathrm{Opt}(P) \leq \overline{F}(x_*) = F(x_*,\lambda_*) = \underline{F}(\lambda_*) \leq \mathrm{Opt}(D)$$

are equalities. Thus, x_* solves (P), λ_* solves (D) and $\mathrm{Opt}(P) = \mathrm{Opt}(D)$.

$$\mathrm{Opt}(P) = \inf_{x\in X} \overbrace{\sup_{\lambda\in\Lambda} F(x,\lambda)}^{\overline{F}(x)} \quad (P)$$
$$\mathrm{Opt}(D) = \sup_{\lambda\in\Lambda} \underbrace{\inf_{x\in X} F(x,\lambda)}_{\underline{F}(\lambda)} \quad (D)$$

✓ Assume that (P), (D) have optimal solutions x_*, λ_* and $\mathrm{Opt}(P) = \mathrm{Opt}(D)$, and let us prove that (x_*, λ_*) is a saddle point. We have

$$\begin{array}{c} \mathrm{Opt}(P) = \overline{F}(x_*) = \sup_{\lambda\in\Lambda} F(x_*,\lambda) \geq F(x_*,\lambda_*) \\ \mathrm{Opt}(D) = \underline{F}(\lambda_*) = \inf_{x\in X} F(x,\lambda_*) \leq F(x_*,\lambda_*) \end{array} \qquad (*)$$

Since $\mathrm{Opt}(P) = \mathrm{Opt}(D)$, all inequalities in $(*)$ are equalities, so that

$$\sup_{\lambda\in\Lambda} F(x_*,\lambda) = F(x_*,\lambda_*) = \inf_{x\in X} F(x,\lambda_*). \qquad \square$$

Existence of Saddle Points

$$\mathrm{Opt}(P) = \inf_{x\in X} \overbrace{\sup_{\lambda\in\Lambda} F(x,\lambda)}^{\overline{F}(x)} \quad (P)$$
$$\mathrm{Opt}(D) = \sup_{\lambda\in\Lambda} \underbrace{\inf_{x\in X} F(x,\lambda)}_{\underline{F}(\lambda)} \quad (D)$$

Fact IX.8 [Sion-Kakutani Theorem] *Let $X \subset \mathbf{R}^n$, $\Lambda \subset \mathbf{R}^m$ be nonempty convex sets, with X being compact, and let $F(x,\lambda) : X \times \Lambda \to \mathbf{R}$ be a continuous function which is convex in $x \in X$ and concave in $\lambda \in \Lambda$. Then*

(i) *One has*

$$\mathrm{Opt}(P) = \mathrm{Opt}(D)$$

(ii) *Assume that Λ is closed and that for every real a there exists $\bar{x}_a \in X$ such that the set*

$$\Lambda_a : \{\lambda \in \Lambda : F(\bar{x}_a, \lambda) \geq a\}$$

is bounded (e.g., Λ is bounded). Then F possesses a saddle point on $X \times \Lambda$.

Note: Same as maximization of a function f can be reduced to minimization of $-f$, saddle point problem

$$\max_{\lambda\in\Lambda} \min_{x\in X} F(x,\lambda) \Rightarrow \begin{cases} \min_{x\in X} \sup_{\lambda\in\Lambda} F(x,\lambda) & (P) \\ \max_{\lambda\in\Lambda} \inf_{x\in X} F(x,\lambda) & (D) \end{cases}$$

is equivalent to the saddle point problem

$$\max_{x\in X} \min_{\lambda\in\Lambda} [-F(x,\lambda)] \Rightarrow \begin{cases} \min_{\lambda\in\Lambda} \sup_{x\in X} [-F(x,\lambda)] & (P') \equiv (D) \\ \max_{x\in X} \inf_{\lambda\in\Lambda} [-F(x,\lambda)] & (D') \equiv (P) \end{cases}$$

Therefore the conclusion (i) in Sion-Kakutani Theorem holds true when instead of compactness of X, compactness of Λ is assumed.

♠ The key role in the proof of Sion-Kakutani Theorem is played by

Fact IX.9 [MinMax Lemma] *Let $X \subset \mathbf{R}^n$ be a convex compact set and $f_i(x) : X \to \mathbf{R}$, $i = 1, ..., m$, be convex continuous functions. Then there exists $\mu^* \geq 0$ with $\sum_i \mu_i^* = 1$ such that*

$$\min_{x \in X} \max_{1 \leq i \leq m} f_i(x) = \min_{x \in X} \sum_i \mu_i^* f_i(x)$$

Note: Setting $\Delta = \{\mu \in \mathbf{R}^m : \mu \geq 0, \sum_i \mu_i = 1\}$, consider the convex-concave saddle point problem

$$\min_{x \in X} \max_{\mu \in \Delta} \sum_i \mu_i f_i(x) \Rightarrow \begin{cases} \mathrm{Opt}(P) = \min_{x \in X} \overline{F}(x) := \underbrace{\max_{\mu \in \Delta} \sum_i \mu_i f_i(x)}_{\equiv \max_i f_i(x)} & (P) \\ \mathrm{Opt}(D) = \max_{\mu \in \Delta} \underline{F}(\mu) := \min_{x \in X} \sum_i \mu_i f_i(x) & (D) \end{cases}$$

MinMax Lemma states that $\mathrm{Opt}(D) = \mathrm{Opt}(P)$, or (since (P) and (D) under the premise of MinMax lemma clearly are solvable) that *the convex-concave function $\sum_i \mu_i f_i(x)$ has a saddle point on $X \times \Delta$.*

Proof of MinMax Lemma. Consider the optimization program

$$\min_{t,x}\{t : f_i(x) - t \le 0,\, i \le m, (t,x) \in X_+\},$$
$$X_+ = \{(t,x) : x \in X\} \qquad (P)$$

The optimal value in this problem clearly is

$$t_* = \min_{x\in X} \max_i f_i(x).$$

The program clearly is convex, solvable and satisfies the Relaxed Slater condition (which, due to the absence of the "$\le_{\mathbf{K}}$-part, reduces to existence of a feasible solutions in rint X_*, which is evident). whence there exists $\lambda^* \ge 0$ and an optimal solution (x_*, t_*) to (P) such that $(x_*, t_*; \lambda^*)$ is the saddle point of the Lagrange function on $X^+ \times \{\lambda \ge 0\}$:

$$\min_{x\in X,t}\left\{t + \sum_i \lambda_i^*(f_i(x) - t)\right\} = t_* + \sum_i \lambda_i^*(f_i(x_*) - t_*) \quad (a)$$
$$\max_{\lambda\ge 0}\left\{t_* + \sum_i \lambda_i(f_i(x_*) - t_*)\right\} = t_* + \sum_i \lambda_i^*(f_i(x_*) - t_*) \quad (b)$$

(b) implies that $t_* + \sum_i \lambda_i^*(f_i(x_*) - t_*) = t_*$. (a) implies that $\sum_i \lambda_i^* = 1$. Thus, $\lambda^* \ge 0, \sum_i \lambda_i^* = 1$ and

$$\min_{x\in X} \sum_i \lambda_i^* f_i(x) = \min_{x\in X,t}\left\{t + \sum_i \lambda_i^*(f_i(x) - t)\right\} = t_* + \sum_i \lambda_i^*(f_i(x_*) - t_*) = t_* = \min_{x\in X} \max_i f_i(x). \qquad \square$$

Proof of Sion-Kakutani Theorem.
1^o As F is continuous on X and X is compact, we have

$$\underline{F}(\lambda) = \min_{x \in X} F(x, \lambda) : \Lambda \to \mathbf{R}$$

Besides, $\underline{F}(\cdot)$ is concave on Λ as the minimum of a family of concave functions of λ.

2^o Let us start with proving the following

Fact IX.10 *Let X, $\underline{\Lambda}$ be nonempty convex compact sets and $G : X \times \underline{\Lambda} \to \mathbf{R}$ be continuous function which is convex in $x \in X$ and concave in $\lambda \in \underline{\Lambda}$. Then G has a saddle point.*

Proof. Indeed, G is a continuous function on a compact set, and is therefore uniformly continuous $\Rightarrow$ the objectives in the problems

$$\begin{array}{rcll} \mathrm{Opt}(\mathcal{P}) & = & \inf\limits_{x \in X} \overbrace{\sup\limits_{\lambda \in \underline{\Lambda}} G(x, \lambda)}^{\overline{G}(x)} & (\mathcal{P}) \\ \mathrm{Opt}(D) & = & \sup\limits_{\lambda \in \underline{\Lambda}} \underbrace{\inf\limits_{x \in X} G(x, \lambda)}_{\underline{G}(\lambda)} & (\mathcal{D}) \end{array}$$

are continuous on the problem's domains and the right hand side sup and inf are achieved $\Rightarrow$ $(\mathcal{P})$, $\mathcal{D})$ are problems of maximization/minimization of continuous functions over nonempty compact sets and as such are solvable. Thus, all we need to prove that G has a saddle point is the equality $\mathrm{Opt}(\mathcal{P}) = \mathrm{Opt}(\mathcal{D})$. This is what we are about to do. Note that $\mathrm{Opt}(\mathcal{D}) \leq \mathrm{Opt}(\mathcal{P})$ by Fact IX.6.

By shifting G by a constant, assume w.l.o.g. that $\mathrm{Opt}(\mathcal{D}) = 0$, that is, $\mathrm{Opt}(\mathcal{P}) \geq 0$, and let, for a contradiction, $\epsilon := \mathrm{Opt}(\mathcal{P}) > 0$. Then for every $x \in X$ there exists $\lambda_x \in \Lambda$ such that $G(x, \lambda_x) \geq \epsilon$; since $G(\cdot, \lambda_x)$ is continuous on X, there is a neighborhood U_x of x such that $G(x', \lambda_x) \geq \epsilon/2$ for all $x' \in U_x \cap X$. As X is compact, we can extract from the open covering $\cup_{x \in X} U_x$ of X a finite subcovering: for properly selected integer m and points $x^i \in X$, $1 \leq i \leq m$, we have $X \subset \cup_{i \leq m} U_{x^i}$. The latter means that setting $f_i(x) = G(x, \lambda_{x^i})$, $i \leq m$, we get convex continuous functions on X such that $\max_{i \leq m} f_i(x) \geq \epsilon/2$ for every $x \in X$. Applying MinMax Lemma, we conclude that there exist convex combination weights μ_i such that $\sum_i \mu_i f_i(x) \geq \epsilon/2$ for all $x \in X$. Setting $\overline{\lambda} = \sum_i \mu_i \lambda_{x^i}$, we see that $\overline{\lambda} \in \Lambda$ and

$$\forall x \in X : G(x, \overline{\lambda}) \geq \sum_i \mu_i G(x, \lambda_{x^i}) = \sum_i \mu_i f_i(x) \geq \epsilon/2,$$

with the first inequality due to the concavity of $G(x, \lambda)$ in $\lambda \in \underline{\Lambda}$. Thus, $\underline{G}(\overline{\lambda}) \geq \epsilon/2$, contradicting $\mathrm{Opt}(\mathcal{D}) = 0$. □

Note: Looking how the equality $\mathrm{Opt}(\mathcal{D}) = \mathrm{Opt}(\mathcal{P})$ has been proved, we see that in fact we have proved a bit more than was claimed; specifically, we have proved

(!) *When $X, \underline{\Lambda}$ are nonempty and convex, X is compact* (just X!) *and the function $G(x, \lambda) : X \times \underline{\Lambda} \to \mathbf{R}$ is continuous, convex in x and concave in λ, then for every real a such that $\mathrm{Opt}(\mathcal{P}) > a$* (in the proof we used $a = 0$) *one has $\mathrm{Opt}(\mathcal{D}) > a$ as well.*

It follows that under the premise of the Sion-Kakutani Theorem, one has $\mathrm{Opt}(P) = \mathrm{Opt}(D)$. Indeed, specifying $\underline{\Lambda}$ as Λ and G as F, in the case of $\mathrm{Opt}(P) = \infty$ (!) says that $\mathrm{Opt}(D) = \infty$ (as (!) as applicable with every real a), when $\mathrm{Opt}(P) \in \mathbf{R}$, we get $\mathrm{Opt}(D) = \mathrm{Opt}(P)$ by applying (!) to all $a < \mathrm{Opt}(P)$ and taking into account Fact IX.6; the same Fact IX.6 says that $\mathrm{Opt}(D) = \mathrm{Opt}(P)$ when $\mathrm{Opt}(P) = -\infty$.

3^o. It remains to prove item (ii) of the Sion-Kakutani Theorem. We already know that the optimal values in the optimization problems

$$\begin{array}{rcl} \mathrm{Opt}(P) & = & \inf\limits_{x\in X} \overbrace{\sup\limits_{\lambda\in\Lambda} F(x,\lambda)}^{\overline{F}(x)} \quad (P) \\ \mathrm{Opt}(D) & = & \sup\limits_{\lambda\in\Lambda} \underbrace{\inf\limits_{x\in X} F(x,\lambda)}_{\underline{F}(\lambda)} \quad (D) \end{array}$$

are equal to each other; therefore, in view of Fact IX.7 all we need to establish the existence of a saddle point is to prove that under the premise of (ii) (P) and (D) are solvable.

✓ As we know from item 1^o, $\underline{F}(\lambda)$ is a concave real-valued function of $\lambda \in \Lambda$. Besides this, from continuity of F, and compactness of X it follows that $\underline{F}$ is continuous on every compact subset K of Λ (since the continuous function $F(x,\lambda)$ on the *compact* set $X \times K$ is uniformly continuous); as Λ is closed, we conclude that $\underline{F}$ is continuous on Λ. By the premise of (ii), for every $a \in \mathbf{R}$ and some $\overline{x}_a \in X$ the set

$$\Lambda_a = \{\lambda \in \Lambda : F(\overline{x}_a, \lambda) \geq a\}$$

is bounded; as $\underline{F}(\lambda) \leq F(\overline{x}_a, \lambda)$, the superlevel set $\{\lambda : \underline{F}(\lambda) \geq a\}$ of $\underline{F}$ is bounded as well. As Λ is closed and $\underline{F}$ is continuous, these bounded superlevel sets are closed and therefore are compact $\Rightarrow$ (D) is solvable by the Weierstrass Theorem; in particular, $\mathrm{Opt}(D) = \mathrm{Opt}(P)$ is a real.

✓ Function $\overline{F} : X \to \mathbf{R} \cup \{+\infty\}$ is the supremum of a family $F(x,\lambda)$, $\lambda \in \Lambda$ of continuous real-valued functions and as such is lsc; besides this, $\mathrm{Opt}(P) = \inf_{x\in X} \overline{F}(x)$ is finite, implying that $\overline{F}$ is a proper lsc function. As X is nonempty and compact, (P) is solvable. □

Illustration: Matrix Game

♣ Recall that in Matrix game the players operate with the quantity

$$F(x,y) = x^T M y; \qquad [M \in \mathbf{R}^{m\times n}]$$

The first player want to minimize F over $x \in \Delta_m = \{x \in \mathbf{R}^m_+ : \sum_i x_i = 1\}$, and the second player wants to maximize F over $y \in \Delta_n$. The equilibria are the saddle points (min in $x \in \Delta_m$, max in $Y \in \Delta_+$) of F.

We see that Matrix game is a convex-concave Saddle Point problem on the direct product of two convex compact sets and as such is solvable – the saddle points do exist. The corresponding primal and dual problems are just LP's:

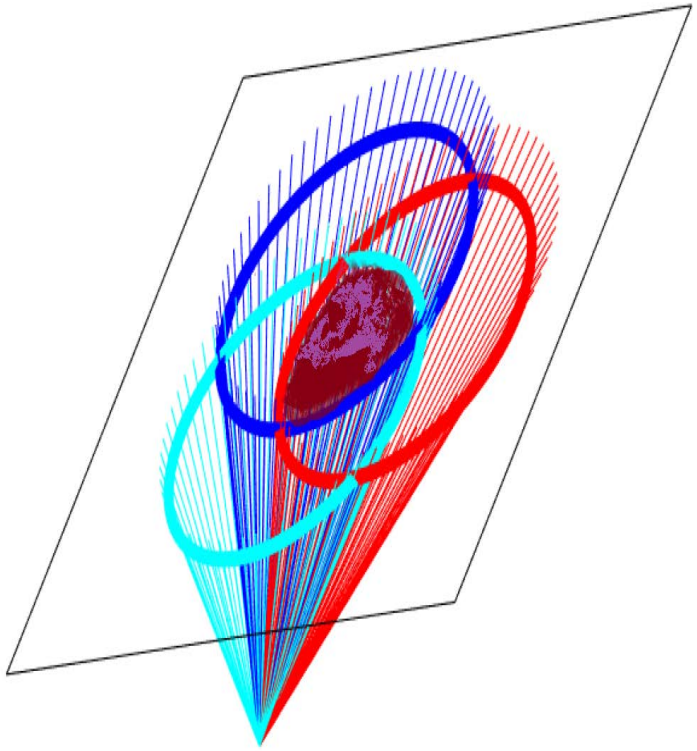
$$\begin{array}{rcll}
\mathrm{Opt}(P) & = & \min_{x\in\Delta_m}\left[\max_{y\in\Delta_n}[A^Tx]^Ty\right] = \min_{x\in\Delta_m}\min_t\left\{t : t\mathbf{1}_n \geq A^Tx\right\} & \\
& & [\mathbf{1}_k = [1;...;1] \in \mathbf{R}^k] & \\
& = & \min_{x,t}\left\{t : t\mathbf{1}_n - A^Tx \geq 0, x \geq 0, \sum_i x_i = 1\right\} & (P) \\
\mathrm{Opt}(D) & = & \max_{y\in\Delta_n}\left[\min_{x\in\Delta_m}[Ay]^Tx\right] = \max_{y\in\Delta_n}\max_s\left\{s : s\mathbf{1}_m \leq Ay\right\} & \\
& = & \max_{y,s}\left\{s : Ay - s\mathbf{1}_m \geq 0, y \geq 0, \sum_j y_j = 1\right\} & (D)
\end{array}$$

Denoting by $y \in \mathbf{R}^n_+, s \in \mathbf{R}, \mu \in \mathbf{R}^m_+$ the Lagrange multipliers for the constraints $t\mathbf{1}_n - A^Tx \geq 0$, $\sum_i x_i = 1$, $x \geq 0$ of (P), the LP dual of the LP problem (P) reads

$$\max_{y,s,\mu}\left\{s : Ay - s\mathbf{1}_m - \mu = 0, \sum_i y_i = 1, y \geq 0, \mu \geq 0\right\},$$

which, after eliminating μ, is nothing but the LP problem (D)

PART IV.
Conic Representations
of
Convex Sets & Functions



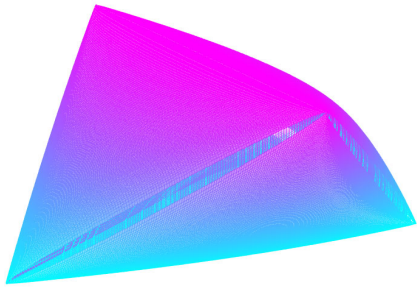
Lecture IV.1

Conic Represntations

of

Convex Sets & Functions

Conic representation: Definition
Calculus of conic representations: Calculus rules
Schur Complement and $\mathcal{S}$-Lemmae
Majorization
Calculus of conic representations: Raw materials



$$\{(x,y,z) : \left[\begin{array}{c|c|c} 1 & x & z \\ \hline x & 1 & y \\ \hline z & y & 1 \end{array}\right] \succeq 0\}$$

Conic representations: Motivation and Definition

♣ Recall that a *polyhedral representation* of a convex set $X \subset \mathbf{R}^n$ is

$$X = \{x \in \mathbf{R}^n : \exists u : Px + Qu \leq r\};$$

while polyhedrally representable set are polyhedral, the notion of a polyhedral representation plays the central role both in *calculus of polyhedrality*, and in *techniques for recognizing the possibility to reduce an optimization problem to LP* and carrying out this reduction when it is possible, thus allowing to enjoy the power of LP solvers.

♠ *Conic representations* of convex sets and functions are aimed at recognizing the possibility to reduce optimization problems to conic problems on "good," computationally friendly, cones (primarily, from the "magic" families $\mathfrak{P}/\mathfrak{C}/\mathfrak{S}$) and to carry out the reduction when it is possible, thus bringing the problems into the scope of powerful solvers.

♣ **Conic representation** of a set $X \subset \mathbf{R}^n$ is a representation of X of the form

$$X = \{x \in \mathbf{R}^n : \exists u : Px + Qu \leq_{\mathbf{K}} r\}$$

where $\mathbf{K}$ is a regular cone.

Conic representation of a function $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ is a conic representation of its epigraph, the right hand side in an equivalence

$$t \geq f(x) \Leftrightarrow \exists u : Px + tp + Qu \leq_{\mathbf{K}} r,$$

where $\mathbf{K}$ is a regular cone.

Note: *Sets/functions admitting conic representations automatically are convex!*

Convention: In the sequel, we refer to a constraint of the form

$$Ax + b \leq_{\mathbf{K}} Cx + d$$

as to a *conic constraint* in variables x on (regular) cone $\mathbf{K}$.

♠ In the sequel we say that a conic constraint

$$Ax + b \leq_{\mathbf{K}} Cx + d$$

involving regular cone $\mathbf{K}$ is *essentially strictly feasible* (synonym: satisfies the Relaxed Slater condition), of one can represent $\mathbf{K}$ as the direct product $\mathbf{P} \times \mathbf{M}$ with polyhedral factor $\mathbf{P}$ and regular factor $\mathbf{M}$ in such a way that

$$[C\bar{x} + d] - [A\bar{x} + b] \in \mathbf{P} \times \operatorname{int} \mathbf{M}$$

for some $\bar{x}$.

♣ **Immediate observations:**

A. Conic representation

$$t \geq f(x) \Leftrightarrow \exists u : Px + tp + Qu \leq_{\mathbf{K}} r,$$

of a function immediately yields conic representations of its sublevel sets:

$$\{x : f(x) \leq a\} = \{x : \exists u : Px + Qu \leq_{\mathbf{K}} r - ap,$$

B. Conic representations

$$\begin{array}{rcl} X & = & \{x \in \mathbf{R}^n : \exists u : P_X x + Q_X u \leq_{\mathbf{K}_X} r_X\} \\ t \geq f(x) & \Leftrightarrow & \exists v : P_f x + tp_f + Q_f v \leq_{\mathbf{K}_f} r_f \end{array}$$

of a set $X \subset \mathbf{R}^n$ and function $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ imply straightforwardly to reformulate the optimization problem

$$\min_{x \in X} f(x) \qquad (O)$$

as the conic problem

$$\min_{x,t,u,v} \left\{ t : \left[\begin{array}{c} P_X x + Q_X u \\ P_f x + tp_f + Q_f v \end{array} \right] \leq_{\mathbf{K}_x \times \mathbf{K}_f} \left[\begin{array}{c} r_X \\ r_f \end{array} \right] \right\} \qquad (C)$$

on the cone $\mathbf{K}_x \times \mathbf{K}_f$: the x-component of a feasible solution (x, t, u, v) to (C) is feasible solution of (O) with the value of the objective $\leq t$, and vice versa: every pair x, t with x feasible for (O) and satisfying $f(x) \leq t$ can be extended to a feasible solution (x, t, u, v) to (C). In particular, both problems have the same optimal values, simultaneously are/are not solvable. and the x-part of an optimal solution, if any, to (C) is an optimal solution to (O).

♣ Let us fix a family $\mathfrak{K}$ of regular cones such that

- $\mathfrak{K}$ is closed w.r.t. taking direct products:

$$\mathbf{K}_\ell \in \mathfrak{K},\ \ell \le L \Rightarrow \mathbf{K}_1 \times ... \times \mathbf{K}_L \in \mathfrak{K}$$

- $\mathfrak{K}$ is closed w.r.t. passing from a cone to its dual:

$$\mathbf{K} \in \mathfrak{K} \Rightarrow \mathbf{K}_* \in \mathfrak{K}$$

Examples:

• $\mathfrak{P}$– the family of nonnegative orthants (i,e,, finite direct products of nonnegative rays $\mathbf{R}_+$); conic problems on the cones from this family are the usual LP's

• $\mathfrak{C}$– the family of finite direct products of Lorentz cones $\mathbf{L}^k = \{x \in \mathbf{R}^k : x_k \ge \sqrt{x_1^2 : ... : x_{k-1}^2}\}$, $k = 1, 2, ...$; conic problems on the cones from this family are called *Conic Quadratic problems (CQPs)*, or *Second Order Conic (SOCP) problems*

• $\mathfrak{S}$– the family of finite direct products of semidefinite cones $\mathbf{S}^k_+ = \{x \in \mathbf{S}^k : x \succeq 0\}$, $k = 1, 2, ...$; conic problems on the cones from this family are called *Semidefinite problems (SDP)*, or *problems with Linear Matrix Inequalities (LMIs)*.

♠ We call sets/functions with conic representations utilizing cones from $\mathfrak{K}$ *$\mathfrak{K}$-representable* ($\mathfrak{K}$-r for short), and refer to the corresponding conic representations as to $\mathfrak{K}$-r.'s (pay attention to the dot after "r").

♠ A *concave* function $f : \mathbf{R}^n \to \mathbf{R} \cup \{-\infty\}$ is called $\mathfrak{K}$-representable, if the convex function $-f$ is so; a $\mathfrak{K}$-representation of a concave function f is the hand side in an equivalence

$$t \le f(x) \Leftrightarrow \exists u : Px + tp + Qu \le_{\mathbf{K}} r,$$

Note: By Fourier-Motzkin, every $\mathfrak{P}$-r set is polyhedral – it admits *polyhedral description*, i.e., it can be described by finitely many nonstrict linear inequalities in the variables from the space where the set lives; in this case, polyhedral representations allow for no more (and no less!) than convenient and fully algorithmic calculus of polyhedrality.
Beyond $\mathfrak{P}$, conic representations play more significant role: there is no FM elimination anymore, and a set $X \subset \mathbf{R}^n$ admitting, say, $\mathfrak{C}$-r

$$X = \{x \in \mathbf{R}^n : \exists u : P_k x + Q_k u \leq_{\mathbf{L}^{m_k}} b_k, k \leq K\}$$

not necessarily admits "Lorentz description"

$$X = \{x \in \mathbf{R}^n : Q_\ell x \leq_{\mathbf{L}^{n_\ell}} q_\ell, \ell \leq L\}$$

– a $\mathfrak{C}$-r. without additional variables.

Calculus of $\mathfrak{K}$-representability: Sets

Fact X.1 *Basic convexity-preserving operations with sets preserve $\mathfrak{K}$-representability:*

A. [Taking finite intersections] *$\mathfrak{K}$-r.'s $X_k = \{x \in \mathbf{R}^n : \exists u_k : P_k x + Q_k u_k \leq_{\mathbf{K}_k} r_k\}$ of sets $X_k \subset \mathbf{R}^n$, $k \leq K$, induce $\mathfrak{K}$-r.*

$$\cap_{k \leq K} X_k = \{x : \exists u = [u_1; ...; u_K] : [P_1; ...; P_K]x + [Q_1 u_1; ...; Q_K u_k] \leq_{\mathbf{K}_1 \times ... \times \mathbf{K}_K} [r_1; ...; r_K]\}$$

of the intersection of the sets.

B. [Taking direct product] *$\mathfrak{K}$-r.'s $X_k = \{x \in \mathbf{R}^{n_k} : \exists u_k : P_k x + Q_k u_k \leq_{\mathbf{K}_k} r_k\}$ of sets $X_k \subset \mathbf{R}^{n_k}$, $k \leq K$, induce $\mathfrak{K}$-r.*

$$X_1 \times ... \times X_K = \{x = [x_1; ...; x_K] \in \mathbf{R}^{n_1 + ... + n-K} : \exists u = [u_1; ...; u_K] : [P_1 x_1; ...; P_K x_k] + [Q_1 u_1; ...; Q_K u_K] \leq$$

of the direct product of the sets,

C. [Taking affine image] *$\mathfrak{K}$-r. $X = \{x : \exists u : Px + Qu \leq_{\mathbf{K}_k} r\}$ of a set $X \subset \mathbf{R}^n$ induces $\mathfrak{K}$-r.*

$$\mathcal{A}(X) = \{y : \exists [x; u] : y = \mathcal{A}(x), Px + Qu \leq_{\mathbf{K}} r\} \qquad (*)$$

of the image $\mathcal{A}(X) = \{y = Ax + b : x \in X\}$ of X under the affine mapping $x \mapsto \mathcal{A}(x) = Ax + b : \mathbf{R}^n \to \mathbf{R}^m$ (note that $y = \mathcal{A}(x)$ is a conic constraint $y \leq \mathcal{A}(x), -y \leq -\mathcal{A}(x)$ on $\mathbf{R}^{2m}_+ \in \mathfrak{K}$, so that the constraints in $(*)$ form a conic constraint on the cone $\mathbf{R}^{2m}_+ \times \mathbf{K}$ belonging to $\mathfrak{K}$ along with $\mathbf{K}$).

D. [Taking inverse affine image] *$\mathfrak{K}$-r. $X = \{x : \exists u_k : Px + Qu \leq_{\mathbf{K}} r\}$ of a set $X \subset \mathbf{R}^n$ induces $\mathfrak{K}$-r.*

$$\mathcal{A}^{-1}(X) = \{y : \exists u : P[Ay + b] + Qu \leq_{\mathbf{K}} r\}$$

of the inverse image $\mathcal{A}^{-1}(X) = \{y : Ay + b \in X\}$ of X under the affine mapping $y \mapsto \mathcal{A}(y) = Ay + b : \mathbf{R}^m \to \mathbf{R}^n$.

E. [Summation] *$\mathfrak{K}$r.'s $X_k = \{x \in \mathbf{R}^n : \exists u_k : P_k x + Q_k u_k \leq_{\mathbf{K}_k} r_k\}$ of sets $X_k \subset \mathbf{R}^n$, $k \leq K$, induce $\mathfrak{K}$-r.*

$$X_1 + ... + X_K = \left\{x : \exists([x^1;,,,;x^K], u = [u_1;...;u_K]) : x = x^1 : ... + x^K, P_k x^k \leq_{\mathbf{K}_k} rk,\, k \leq K\right\}$$

of the sum of the sets.

♣ More advanced operations with convex sets "nearly preserve" $\mathfrak{K}$-representability:

Fact X.2 *One has:*

A. [Closedness] *A convex set X given by $\mathfrak{K}$-r. $X = \{x : \exists u : Px + Qu \leq_{\mathbf{K}} r\}$ not necessarily is closed. It definitely if closed when bK s a polyhedral cone, same as when $\mathbf{K}$ is regular cone and $\mathrm{Im}Q \cap \mathbf{K} = \{0\}$.*

Indeed, by calculus of polyhedrality, when $\mathbf{K}$ is polyhedral, so is X, and polyhedral cones are closed. When $\mathbf{K}$ is regular cone, what affects X is just the linear space $\mathrm{Im}\, Q$, not Q itself $\Rightarrow$ w.l.o.g. we can assume that Q is an embedding: $\mathrm{Ker}Q = \{0\}$. X is the projection of the nonempty closed convex set $X^+ = \{[x; u] : Px + Qu \leq_{\mathbf{K}} r\}$ in the (x, u)-space onto the plane of x-variables. By Fact II.23, a sufficient condition for this projection to be closed is the triviality $-$ $K = \{0\}$ $-$ of the intersection K of the kernel $\{0\} \times \mathbf{R}^{\dim u}$ of the mapping $[x; u] \mapsto x$ and $\mathrm{Rec}(X^+)$. We clearly have $\mathrm{Rec}(X^+) = \{[dx; du] : Pdx + Qdu \leq_{\mathbf{K}} 0\}$, resulting in $K = \{[0; du] : Qdu \leq_{\mathbf{K}} 0\}$, that is, $K = \{0\}$ is the same as the validity of the implication $[0; du] \in K \Rightarrow du = 0$, which for embedding Q is the same as $\mathrm{Im}\, Q \cap \mathbf{K} = \{0\}$.

B. [Passing from a set to its recessive cone] *Let a nonempty set X be given by $\mathfrak{K}$-r. $X = \{x : \exists u : Px + Qu \leq_{\mathbf{K}} r\}$. Then the $\mathfrak{K}$-r cone*

$$R = \{x : \exists u : Px + Qu \leq_{\mathbf{K}} 0\}$$

is contained in $\mathrm{Rec}(\mathrm{cl}\, X)$. When $\mathbf{K}$ is polyhedral, same as when $\mathrm{Im}\, Q \cap \mathbf{K} = \{0\}$, X is closed and $\mathrm{Rec}(X) = R$.

The first claim is evident. The "polyhedral" part of the second claim is given by calculus of polyhedrality. Finally, when $\mathrm{Im}Q \cap \mathbf{K} = \{0\}$, X is closed by item **A**. Same as in the proof of item **A**, we lose nothing when assuming that Q is an embedding. The set $X^+ = \{[x; u] : Px + Qu] \leq_{\mathbf{K}} r\}$ is nonempty along with X, closed, and its recessive cone clearly is $R^+ = \{[xd; du] : Pdx + Qdu \leq_{\mathbf{K}} 0\}$. X is the image of X^+ under the projection $[x; u] \mapsto x$. Same as in the proof of item **A**, assuming that Q is an embedding and $\mathrm{Im}\, Q \cap \mathbf{K} = \{0\}$, the intersection K of the kernel of this projection with the closed cone R^+ is $\{0\}$, so that $\mathrm{Rec}(X)$ is , by Fact II.23, exactly the projection R of R^+ onto the x-space, Q.E.D.

C. [Taking perspective transform] *Let a nonempty set X be given by $\mathfrak{K}$-r. $X = \{x : \exists u : Px + Qu \leq_{\mathbf{K}} r\}$, and assume that $\mathfrak{K}$ contains 3D Lorentz cone $\mathbf{L}^3 = \{[\alpha;\beta;\gamma] : \gamma \geq \sqrt{\alpha^2+\beta^2}\}$. Then the $\mathfrak{K}$-r set*

$$R = \{[x;t] : \exists u, \alpha : Px + Qu - tr \leq_{\mathbf{K}} 0 \ \&\ [2; t-\alpha; t+\alpha] \geq_{\mathbf{L}^3} 0,\ \alpha \geq 0\}$$

is $\mathfrak{K}$-r. of the perspective transform

$$\mathrm{Persp} = \{[x;t] : x/t \in X,\ t > 0\}$$

of X. Indeed, as is immediately seen, the constraints $[2; t-\alpha; t+\alpha] \geq_{\mathbf{L}^3} 0$, $\alpha \geq 0$ are exactly equivalent to $\alpha t \geq 1$, $\alpha, t \geq 0$, that is,

$$\begin{array}{c} R = \{[x;t] : \exists u, \alpha : Px + Qu - tr \leq_{\mathbf{K}} 0,\ t > 0\} = \{[x;t]; t > 0 \\ \& \\ \exists v : P[x/t] + Qv \leq_{\mathbf{K}} r\} = \{[x;t] : t > 0,\ x/t \in X\} = \mathrm{Persp}(X). \end{array}$$

D. [Taking the closed conic transform] *Let a nonempty set X be given by $\mathfrak{R}$-r. $X = \{x : \exists u : Px + Qu \leq_{\mathbf{K}} r\}$. Then the $\mathfrak{R}$-r cone*

$$R = \{[x; t] : \exists u : Px + Qu - tr \leq_{\mathbf{K}} 0 \ \& \ t \geq 0\}$$

is in-between the perspective transform $\mathrm{Persp}(X) = \{[x, t] : t > 0, x/t \in X\}$ *of X and the closed conic transform* $\overline{\mathrm{ConeT}}(X) = \mathrm{cl}\,\mathrm{Persp}(X)$*:*

$$\mathrm{Persp}(X) \subset R \subset \overline{\mathrm{ConeT}}(X).$$

In particular, when R is closed (which, by item **A**, *definitely is the case when $\mathbf{K}$ is polyhedral, same as when $\mathrm{Im}\, Q \cap \mathbf{K} = \{0\}$), R is the closed conic transform of X*

Indeed, if $[x; t] \in \mathrm{Persp}(X)$, so that $t > 0$ and $x/t \in X$, we have $P[x/t] + Qu \leq_{\mathbf{K}} r$ for some u, whence $Px + Q[tu] - tr \leq \mathbf{K}_0$, that is, $[x; t] \in R$. On the other hand, $X \neq \emptyset$,m whence $P\bar{x} + Q\bar{u} \leq_{\mathbf{K}} r$ for some $\bar{x}$, that is $[\bar{x}; 1] \in R$. Now let $[x; t] \in R$, so that $Px + Qu - rt \leq_{\mathbf{K}} 0$ for some u and $t \geq 0$. We have $[x; t] = \lim_{\epsilon \to +0} [x + \epsilon\bar{x}; t + \epsilon]$and $P[x + \epsilon\bar{x}] + Q[u + \epsilon\bar{u} - [t + \epsilon]r \leq_{\mathbf{K}} 0$, implying, due to , $t + \epsilon > 0$ whenever $\epsilon > 0$, that $[x + \epsilon\bar{t}]/[t + \epsilon] \in X$, that is, $[x + \epsilon\bar{x}; t + \epsilon \in \mathrm{Persp}(X)$, whence $[x; t] = \lim_{\epsilon \to +0} [x + \epsilon\bar{x}; t + \epsilon] \in \mathrm{cl}\,\mathrm{Persp}(X) = \overline{\mathrm{ConeT}}(X)$. Thus, $R \subset \overline{\mathrm{ConeT}}(X)$, in addition to the already proved $\mathrm{Persp}(X) \subset R$. The "in addition" part of the claim readily given by the already proved part of it. □

E. [Taking convex hull of finite union] *Let nonempty sets $X_k \subset \mathbf{R}^{n_k}$, $k \le K$, be given by $\mathfrak{K}$-r.'s $X_k = \{x : \exists u_k : P_k x + Q_k u_k \le_{\mathbf{K}_k} r_k\}$. Then the $\mathfrak{K}$-r set*

$$X = \left\{ x \in \mathbf{R}^n : \exists y_1, ..., y_K, u_1, ..., u_K, \lambda_1, ..., \lambda_K : \left\{ \begin{array}{cc} \lambda_k \ge 0, \sum_k \lambda_k = 1 & (a) \\ P_k y_k + Q_k u_k - \lambda_k r_k \le_{\mathbf{K}_k} 0,\ k \le K & (b) \\ \sum_k y_k = x & (c) \end{array} \right\} \right.$$

is in-between the convex hull of $\cup_k X_k$ and the closure of this convex hull:

$$\mathsf{Conv}\,(\cup_k X_k) \subset X \subset \mathsf{cl}\,\mathsf{Conv}\,(\cup_k X_k)$$

$$X=\left\{x\in\mathbf{R}^n:\exists y_1,...,y_K,u_1,...,u_K,\lambda_1,...,\lambda_K:\left\{\begin{array}{ll}\lambda_k\geq 0,\sum_k\lambda_k=1 & (a)\\ P_ky_k+Q_ku_k-\lambda_kr_k\leq_{\mathbf{K}_k}0,\ k\leq K & (b)\\ \sum_k y_k=xq & (c)\end{array}\right\}\right\}$$

✓ Let us prove that $\mathrm{Conv}\,(\cup_k X_k)\subset X$. Let $x=\sum_i\lambda_k x_k\in\mathrm{Conv}\,(\cup_k X_k)\subset(\lambda_k\geq 0,\sum_k\lambda_k=1,\ x_k\in X_k)$. Setting $I=\{k:\lambda_k>0\}$ and $y_k=x_k/\lambda_k$, $k\in I$, for $k\in I$ we have for properly selected v_k: $P_kx_k+Q_kv_k\leq_{\mathbf{K}_k}r_k$ $\Rightarrow$ $P_ky_k+Q_ku_k-\lambda_kr_k\leq_{\mathbf{K}_k}0$ with $u_k:=v_k/\lambda_k$, $k\in I$ Setting $y_k=0$, $u_k=0$ for $k\notin I$, we ensure $(a)-(c)$ $\Rightarrow$ $x\in X$.

✓ Now let us prove that $X\subset\mathrm{cl}\,\mathrm{Conv}\,(\cup_k X_k)$. Let $x\in X$, and let us verify that $x\in\mathrm{cl}\,\mathrm{Conv}\,(\cup_k X_k)$

As $x\in X$, there exist y_k,λ_k,u_k such that (see $(a)-(c)$)

$$\lambda_k\geq 0,\ \sum_k\lambda_k=1,\ P_ky_k+Q_ku_k-\lambda_kr_k\leq_{\mathbf{K}_k}0,\ x=\sum_k y_k$$

As X_k are nonempty, we can find $\overline{x}_k$, $\overline{v}_k$ such that

$$P_k\overline{x}_k+Q_k\overline{v}_k-r_k\leq_{\mathbf{K}_k}0.$$

For $\epsilon\in(0,1)$, let

$$\overline{y}_k=\overline{x}_k/K,\ \overline{\lambda}_k=1/K,\ \overline{u}_k=\overline{v}_k/K\quad\left[\Rightarrow P_k\overline{y}_k+Q_k\overline{u}_k-\overline{\lambda}_kr_k\leq_{\mathbf{K}_k}0,\ k\leq K\right]$$
$$x_\epsilon=(1-\epsilon)x+\epsilon\textstyle\sum_k\overline{y}_k,\ y_k^\epsilon=(1-\epsilon)y_k+\epsilon\overline{y}_k,\ \lambda_k^\epsilon=(1-\epsilon)\lambda_k+\epsilon\overline{\lambda}_k,\ u_k^\epsilon=(1-\epsilon)u_k+\epsilon\overline{u}_k$$
$$\Rightarrow\left\{\begin{array}{l}P_ky_k^\epsilon+Q_ku_k^\epsilon-\lambda_k^\epsilon r_k=(1-\epsilon)\left[P_ky_k+Q_ku_k-\lambda_kr_k\right]+\epsilon\left[P_k\overline{y}_k+Q_k\overline{u}_k-\overline{\lambda}_kr_k\right]\leq_{\mathbf{K}_k}0,\ k\leq K\\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\Rightarrow x_k^\epsilon:=y_k^\epsilon/\lambda_k^\epsilon\in X_k,\ k\leq K\\ 0<\lambda_k^\epsilon,\ \sum_k\lambda_k^\epsilon=(1-\epsilon)\sum_k\lambda_k+\epsilon\sum_k\overline{\lambda}_k=1\\ \sum_k\lambda_k^\epsilon x_k^\epsilon=\sum_k y_k^\epsilon=(1-\epsilon)\sum_k y_k+\epsilon\sum_k\overline{y}_k=(1-\epsilon)x+\epsilon\sum_k\overline{y}_k=x_\epsilon\end{array}\right.$$

We see that $x_\epsilon=\sum_k\lambda_k^\epsilon x_k^\epsilon\in\mathrm{Conv}\,(\cup_k X_k)$ and $x_\epsilon\to x$ as $\epsilon\to+0$ $\Rightarrow$ $x\in\mathrm{cl}\,\mathrm{Conv}\,(\cup_k X_k)$ □

F. [Taking conic hull of finite union] *Let nonempty sets $X_k \subset \mathbf{R}^{n_k}$, $k \le K$, be given by $\mathfrak{R}$-r.'s $X_k = \{x : \exists u_k : P_k x + Q_k u_k \le_{\mathbf{K}_k} r_k\}$. Then the $\mathfrak{R}$-r cone*

$$X = \left\{ x \in \mathbf{R}^n : \exists y_1, ..., y_K, u_1, ..., u_K, \lambda_1, ..., \lambda_K : \left\{ \begin{array}{rl} \lambda_k \ge 0 & (a) \\ P_k y_k + Q_k u_k - \lambda_k r_k \le_{\mathbf{K}_k} 0,\ k \le K & (b) \\ \sum_k y_k = x & (c) \end{array} \right\} \right.$$

is in-between the conic hull of $\cup_k X_k$ and the closure of this conic hull:

$$\text{Cone } (\cup_k X_k) \subset X \subset \text{cl Cone } (\cup_k X_k)$$

✓ Let us prove that Cone $(\cup_k X_k) \subset X$. Let $x = \sum_i \lambda_k x_k \in$ Cone $(\cup_k X_k) \subset (\lambda_k \ge 0, x_k \in X_k)$. Let $I = \{k : \lambda_k > 0\}$. When $I = \emptyset$, $x = 0$, and setting $y_k = 0$, $u_k = 0$, $k \le K$, we fir $(a) - (c)$, that is, $x \in X$. When $I \ne \emptyset$, setting $y_k = x_k/\lambda_k$, $k \in I$, for $k \in I$ we have for properly selected v_k: $P_k x_k + Q_k v_k \le_{\mathbf{K}_k} r_k \Rightarrow P_k y_k + Q_k u_k - \lambda_k r_k \le_{\mathbf{K}_k} 0$ with $u_k := v_k/\lambda_k$, $k \in I$ Setting $y_k = 0$, $u_k = 0$ for $k \notin I$, we ensure $(a) - (c) \Rightarrow x \in X$.
✓ Now let us prove that $X \subset$ cl Cone $(\cup_k X_k)$. Let $x \in X$, and let us verify that $x \in$ cl Cone $(\cup_k X_k)$
As $x \in X$, there exist y_k, λ_k, u_k such that (see $(a) - (c)$)

$$\lambda_k \ge 0,\ P_k y_k + Q_k u_k - \lambda_k r_k \le_{\mathbf{K}_k} 0,\ x = \sum_k y_k$$

As X_k are nonempty, we can find $\overline{x}_k$, $\overline{v}_k$ such that

$$P_k \overline{x}_k + Q_k \overline{v}_k - r_k \le_{\mathbf{K}_k} 0.$$

For $\epsilon \in (0, 1)$, let

$$\begin{array}{c} \overline{y}_k = \overline{x}_k,\ \overline{\lambda}_k = 1,\ \overline{u}_k = \overline{v}_k \quad \left[\Rightarrow P_k \overline{y}_k + Q_k \overline{u}_k - \overline{\lambda}_k r_k \le_{\mathbf{K}_k} 0,\ k \le K\right] \\ x_\epsilon = (1-\epsilon)x + \epsilon \sum_k \overline{y}_k,\ y_k^\epsilon = (1-\epsilon) y_k + \epsilon \overline{y}_k,\ \lambda_k^\epsilon = (1-\epsilon)\lambda_k + \epsilon \overline{\lambda}_k,\ u_k^\epsilon = (1-\epsilon) u_k + \epsilon \overline{u}_k \\ \Rightarrow \left\{ \begin{array}{l} P_k y_k^\epsilon + Q_k u_k^\epsilon - \lambda_k^\epsilon r_k = (1-\epsilon)\left[P_k y_k + Q_k u_k - \lambda_k r_k\right] + \epsilon \left[P_k \overline{y}_k + Q_k \overline{u}_k - \overline{\lambda}_k r_k\right] \le_{\mathbf{K}_k} 0,\ k \le K \\ \qquad\qquad \Rightarrow x_k^\epsilon := y_k^\epsilon/\lambda_k^\epsilon \in X_k,\ k \le K \\ 0 < \lambda_k^\epsilon,\ \sum_k \lambda_k^\epsilon x_k^\epsilon = \sum_k y_k^\epsilon = (1-\epsilon)\sum_k y_k + \epsilon \sum_k \overline{y}_k = (1-\epsilon)x + \epsilon \sum_k \overline{y}_k = x_\epsilon \end{array} \right. \end{array}$$

We see that $x_\epsilon = \sum_k \lambda_k^\epsilon x_k^\epsilon \in$ Cone $(\cup_k X_k)$ and $x_\epsilon \to x$ as $\epsilon \to +0 \Rightarrow x \in$ cl Cone $(\cup_k X_k)$ □

G. [Taking polar] *Let a nonempty set X be given by essentially strictly feasible $\mathfrak{K}$-r. $X = \{x : \exists u : Px + Qu \leq_{\mathbf{K}} r\}$. Then the polar* Polar $(X) = \{y : \sup_{x \in X} yh^T x\}$ *admit $\mathfrak{K}$-r.*

$$\text{Polar}\,(X) = \{y : \exists z : P^T z = y,\ r^T z \leq 1,\ z \leq_{\mathbf{K}_*} 0\} \qquad (*)$$

Indeed, $y \in$ Polar (X) iff the optimal value Opt in the problem $\max_{x \in X} y^T x$, or, which is the same, in the conic problem $\max_{x,u} \left\{y^T x : Px + Qu \leq_{\mathbf{K}} r\right\}$ is ≤ 1. We are in the situation when the latter problem satisfies the Relaxed Slater condition $\Rightarrow$ by Conic Duality Theorem, Opt ≤ 1 iff the conic dual problem

$$\min_{\lambda} \left\{-r^T \lambda : P^T \lambda = -y,\ Q^T \lambda = 0,\ \lambda_{\mathbf{K}_*} \geq 0\right\}$$

has a feasible solution with the value if the objective ≤ 1. As $\mathbf{K}_* \in \mathfrak{K}$ along with $\mathbf{K}$, $(*)$ indeed is a $\mathfrak{K}$-r. of Polar (X).

Calculus of $\mathfrak{K}$-representability: Functions

Fact X.3 *Basic convexity-preserving operations with functions preserve $\mathfrak{K}$-representability:*

A. [Taking restriction onto $\mathfrak{K}$-r set] *$\mathfrak{K}$-r.'s $t \geq f(x) \Leftrightarrow \exists u : P_f x + tp_f + Q_f u \leq_{\mathbf{K}} r_f$ of a function $f : \mathbf{R}^n \to \mathbf{R} \cup \{\infty\}$ and a set*

$$X = \{x : \exists v : P_X x + Q_X u \leq_{\mathbf{K}_X} r_x\} \subset \mathbf{R}^n$$

induce $\mathfrak{K}$-r. of the restriction $f\big|_X(x) = \begin{cases} f(x) & , x \in X \\ +\infty & , \text{otherwise} \end{cases}$ of f onto X:

$$t \geq f\big|_X(x) \leftrightarrow \exists u, v : P_X x + Q_X v \leq_{\mathbf{K}_X} r_X,\ P_f x + tp_f + Q_f u \leq_{\mathbf{K}_f} 0$$

- **B.** [Taking conic combinations] *$\mathfrak{K}$-r.'s $t \geq f_k(x) \Leftrightarrow \exists u_k : P_k x + tp_k + Q_k u_k \leq_{\mathbf{K}_k} r_k,\ k \leq k$, of functions $f_k : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ induce $\mathfrak{K}$-r. of their conic combination with coefficients $\alpha_k > 0$:*

$$t \geq \sum_k \alpha_k f_k(x) \Leftrightarrow \exists t_1, ..., t_K, u_1, ..., u_K : P_k x + t_k p_k + Q_k u_k \leq_{\mathbf{K}_k} r_k,\ k \leq K, \sum_k \alpha_k t_k \leq t$$

B. [Direct summation] *$\mathfrak{K}$-r.'s $t \geq f_k(x_k) \Leftrightarrow \exists u_k : P_k x_k + tp_k + Q_k u_k \leq_{\mathbf{K}_k} r_k,\ k \leq k$, of functions $f_k : \mathbf{R}^{n_k} \to \mathbf{R} \cup \{+\infty\}$ induce $\mathfrak{K}$-r. of the direct sum $\sum_k f_k(x_k)$ of the functions:*

$$t \geq \sum_k f_k(t_k) \Leftrightarrow \exists t_1, ..., t_K, u_1, ..., u_K : P_k x_k + t_k p_k + Q_k u_k \leq_{\mathbf{K}_k} r_k,\ k \leq K, \sum_k t_k \leq t$$

C. [Affine substitution of variables] *$\mathfrak{K}$-r. $t \geq f(x) \Leftrightarrow \exists u : Px + tp + Qu \leq_{\mathbf{K}} r$ of a function $f : \mathbf{R}^n \to \mathbf{R} \cup \{\infty\}$ induces $\mathfrak{K}$-r.*

$$t \geq f(\mathcal{A}(y)) \Leftrightarrow \exists u : P\mathcal{A}(y) + tp + Qu \leq_{\mathbf{K}} r$$

of f with the affine mapping $y \mapsto \mathcal{A}(y) = Ay + b : \mathbf{R}^m \to \mathbf{R}^n$.

D. [Monotone superposition] *Let*
— *functions* $f_k : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ *be affine* $f_k(x) = a_k^T x + b_k$ *for* $k < \underline{K}$ *and given by* $\mathfrak{K}$*-r.'s*

$$t \geq f_k(x) \Leftrightarrow \exists u_k : P_k x + t p_k + Q_k u_k \leq_{\mathbf{K}_k} r_k$$

for $\underline{K} \leq k \leq K$,
— *set* $Y \subset \mathbf{R}^m$ *be given by* $\mathfrak{K}$*-r.* $Y = \{y : \exists w : Ay + Bw \leq_{\mathbf{K}} c$,
— *function* $F : \mathbf{R}^m \to \mathbf{R} \cup \{+\infty\}$ *be given by* $\mathfrak{K}$*-r.* $t \geq F(y) \Leftrightarrow \exists v : Pq + tp + Qv \leq_{\mathbf{M}} r$.
Assume that whenever $x \in \cap_k \mathrm{Dom}\, f_k$, *the function* $f(x) = [f_1(x); ...; f_m(x)]$ *takes values in* Y, *and that* F *is nonincreasing in its arguments* y_k, $\underline{K} \leq k \leq K$ *on* Y*:*

$$y, y' \in Y, y_k = y'_k,\ k < \overline{K},\ y_k \leq y'_k,\ \underline{K} \leq k \leq K \Rightarrow F(y) \leq F(y')$$

Then the composition

$$G(x) = \begin{cases} F(f(x)) & , x \in \cap_k \mathrm{Dom}\, f_k \\ +\infty & , \textit{otherwise} \end{cases}$$

is $\mathfrak{K}$*-r:*

$$t \geq G(x) \Leftrightarrow \exists s_k, k \leq K, u_k, \underline{K} \leq k \leq K, w, v :$$
$$\begin{cases} s_k = a^T x + b_k & , 1 \leq k < \underline{K} \quad \text{[says that } f_k(x) = s_k\text{]} \\ P_k x + s_k p_k + Q_k u_k \leq_{\mathbf{K}_k} r_k & , \underline{K} \leq k \leq K \quad \text{[says that } f_k(x) \leq s_k\text{]} \\ A[s_1; ...; s_K] + Bw \leq_{\mathbf{K}} c & \quad \text{[says that } [s_1; ...; s_K] \in Y\text{]} \\ P[s_1; ...; s_K] + tp + Qv \leq_{\mathbf{M}} r & \quad \text{[says that } F(s_1, ..., s_K) \leq t\text{]} \end{cases}$$

E. [Characteristic function] *The characteristic function* χ_X *of a* $\mathfrak{K}$*-r set*

$$X = \{x \in \mathbf{R}^n : \exists u : Px + Qy] \leq_{\mathbf{K}} r\}$$

is $\mathfrak{K}$*-r:*

$$t \geq \chi_X(x) \Leftrightarrow \exists u : Px + Qu \leq_{\mathbf{K}} r,\ t \geq 0$$

♠ Some calculus rules require minor additional assumptions

Fact X.4 *One has*

A. [Closedness] *Let* $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ *be given by* $\mathfrak{R}r$.

$$\mathrm{Epi}\{f\} = \{[x;t] : \exists u : Px + tp + u \leq_{\mathbf{K}} r\}$$

When $\mathbf{K}$ *is polyhedral, same as when* $\mathrm{Im}\, Q \cap \mathbf{K} = \{0\}$, f *is lsc.*

Indeed, a function with values in $\mathbf{R} \cap \{+\infty\}$ is lsc iff its epigraph is closed, and it remains t refer to Fact X.2.

B. [Partial minimization] *Let function* $f(x,y) : \mathbf{R}^n_x \times \mathbf{R}^m y \to \mathbf{R} \cup \{+\infty\}$ *be given by* $\mathfrak{R}$*-r.*

$$t \geq f(x,y) \Leftrightarrow \exists u : P_x x + P_y y + tp + Qu \leq_{\mathbf{K}} r.$$

Assume that $\inf_y f(x,y)$, *whenever it is* $< +\infty$, *is achieved. Then* $g(x) = \inf_y f(x,y)$ *is* $\mathfrak{R}$*-r:*

$$t \geq f(x) \Leftrightarrow \exists y, u : P_x x + tp + [P_y y + Qu] \leq_{\mathbf{K}} r.$$

C. [Support function] *Support function of a set*

$$X = \{x : \exists u : Px + Qu \leq_{\mathbf{K}} r\}$$

essentially strictly feasible $\mathfrak{R}$*-r. is* $\mathfrak{R}$*-r:*

$$t \geq \mathrm{Supp}_X(y) \Leftrightarrow \exists \lambda : P^T\lambda = y, Q^T\lambda = 0, r^T\lambda \leq t, \lambda \geq_{\mathbf{K}_0} 0 \qquad (*)$$

Indeed, $t \geq \mathrm{Supp}(x)$ *iff the optimal value* Opt *in the problem* $\sup_{x \in X} y^T x$, *or, which is the same, in the conic problem* $\max_{x,u}\{y^T x : Px + Qu \leq_{\mathbf{K}} r\}$ *is* $\leq t$. *Under the premise of item* **C**, *this problem satisfies the Relaxed Slater condition* $\Rightarrow$ *by Conic Duality Theorem,* Opt $\leq t$ *iff the conic dual*

$$\min_\lambda\{r^T\lambda : P^T\lambda = y, Q^T\lambda = 0, \lambda \geq_{\mathbf{K}_*} 0\}$$

of the above conic problem has a feasible solution with the value of the objective $\leq t$, *resulting in* $(*)$,

D. [Legendre transform] *Let a function $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ be given by essentially strictly feasible $\mathfrak{K}$-r.*

$$t \geq f(x) = \{[x;t] : \exists u : Px + tp + Qu \leq_{\mathbf{K}} r\}$$

Then the Legendre transform $f_*(y) = \sup_x[y^T x - f(x)]$ is $\mathfrak{K}$-r:

$$t \geq f_*(y) \Leftrightarrow \exists \lambda : P^T\lambda = y, p^T\lambda = -1, Q^T\lambda = 0, r^T\lambda \leq t, \lambda \geq_{\mathbf{K}_*} 0 \qquad (!)$$

Indeed, $t \geq f_*(y)$ iff the optimal value Opt in the problem $\sup_x[y^T x - f(x)]$, or, which is the same, in the conic problem

$$\max_{x,u,s}\{y^T x - s : \underbrace{Px + st + Qu \leq_{\mathbf{K}} r}_{\text{says that } s \geq f(x)}\}$$

is $\leq t$. Under the premise of item **D**, this problem satisfies the Relaxed Slater condition $\Rightarrow$ by Conic Duality Theorem, Opt $\leq t$ iff the conic dual

$$\min_{\lambda}\{r^T\lambda : P^T\lambda = y, p^T\lambda = -1, Q^T\lambda = 0, \lambda \geq_{\mathbf{K}_*} 0\}$$

has a feasible solution with the value of the objective $\leq t$, resulting in (!).

E. [Perspective transform] *Let a function $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ be given by $\mathfrak{K}$-r.*

$$t \geq f(x) \Leftrightarrow \{[x;t] : \exists u : Px + tp + Qu \leq_{\mathbf{K}} r\}$$

and assume that $\mathfrak{K}$ contains 3D Lorentz cone $\mathbf{L}^3 = \{[\alpha;\beta;\gamma] : \gamma \geq \sqrt{\alpha^2 + \beta^2}\}$. Then the perspective transform

$$F(x,\tau) = \begin{cases} \tau f(x/\tau) & , \tau > 0 \\ +\infty & , \text{otherwise} \end{cases}$$

of f is $\mathfrak{K}$-r:

$$t \geq F(x,\tau) \Leftrightarrow \exists u, \alpha : Px + tp + Qu - \tau r \leq_{\mathbf{K}} 0, [2;\tau - \alpha;\tau + \alpha] \geq_{\mathbf{L}^3} 0, \alpha \geq 0.$$

Indeed, the constraints $[2;\tau - \alpha;\tau + \alpha] \geq_{\mathbf{L}^3} 0$, $\alpha \geq 0$ are equivalent to $t \geq 0, \alpha \geq 0, t\alpha > 0$ $\Rightarrow$

$\exists u, \alpha : Px + tp + Qu - \tau r \leq_{\mathbf{K}} 0, [2;\tau - \alpha;\tau + \alpha] \geq_{\mathbf{L}^3} 0, \alpha \geq 0 \Leftrightarrow \exists u : Px + tp + Qu - \tau r \,\&\, \tau > 0$
$\Leftrightarrow \exists u : P[x/\tau] + [t/\tau]p + Q[u/\tau] \leq_{\mathbf{K}} r \;\&\; \tau > 0 \Leftrightarrow f(x/\tau) \leq t/\tau \;\&\; \tau > 0 \Leftrightarrow F(x,\tau) \leq t.$

$\mathfrak{K}$-representability : Raw materials

♣ Rules of Grammar (or of Calculus) become useful after we get at our disposal "raw materials" these rules can be applied to: dictionary of words (or of elementary functions with "bare hands" computed derivatives) which we keep in our memory. Situation with $\mathfrak{P}/\mathfrak{C}/\mathfrak{S}$-representability is similar: we need a dictionary of "elementary" functions/sets of this type.

In contrast to Calculus of $\mathfrak{K}$-representability completely independent of what is the family $\mathfrak{K}$ of regular cones we are speaking about, the raw materials do depend on the family.

♣ Let us start with several general observations.

A. *Assume we are given two families of regular cones $\mathfrak{K}$ and $\mathfrak{M}$, both closed w.r.t. taking direct products and passing from a cone to its dual, and let every cone $\mathbf{K}$ from $\mathfrak{K}$ admit $\mathfrak{M}$-representation:*

$$\mathbf{K} = \{x : \exists u : Px + Qu \leq_{\mathbf{M}} r\} \qquad [\mathbf{M} \in \mathfrak{M}]$$

Then every $\mathfrak{K}$-representable set/function is $\mathfrak{M}$-representable. Given $\mathfrak{M}$-r.'s of cones from $\mathfrak{K}$, we can immediately convert a $\mathfrak{K}$-r. of a set/function into its $\mathfrak{M}$-r.

B. $\mathfrak{P}/\mathfrak{C}/\mathfrak{S}$ Hierarchy: *Nonnegative orthants are $\mathfrak{C}$-representable; Lorentz cones are $\mathfrak{S}$-representable.*

Indeed, $\mathbf{R}_+$ is the same as $\mathbf{L}^1$, so that the finite direct products of positive rays – nonnegative orthants – are direct products of Lorentz cones as well.

To see that a Lorentz cone $\mathbf{L}^n$ is $\mathfrak{S}$-r, we need the following

Fact X.5 *The Lorentz cone $\mathbf{L}^n$ admits $\mathfrak{S}$-r., specifically,*

$$\mathbf{L}^n = \{x \in \mathbf{R}^n : \mathrm{Arrow}(x) := \left[\begin{array}{c|c|c|c} x_n & x_1 & \ldots & x_{n-1} \\ \hline x_1 & x_n & & \\ \hline \vdots & & \ddots & \\ \hline x_{n-1} & & & x_n \end{array}\right] \succeq 0\}.$$

✓ Let $x \in \mathbf{L}^n$; then either $x = 0$, whence $\mathrm{Arrow}(x) = 0 \succeq 0$, or $x_n > 0$, so that the South-Eastern block $x_n I_{n-1}$ in $\mathrm{Arrow}(x)$ is $\succ 0$. The Schur complement to this block in $\mathrm{Arrow}(x)$ is

$$x_n - \frac{\sum_{i=1}^{n-1} x_i^2}{x_n} \geq 0 \qquad \text{[due to } x_n > 0 \ \&\ x_n^2 \geq x_1^2 + \ldots + x_{n-1}^2\text{]}$$

and $\mathrm{Arrow}(x) \succeq 0$ by Schur Complement Lemma.

✓ Vice versa, Let $\mathrm{Arrow}(x) \succeq 0$, and let us verify that $x \in \mathbf{L}^n$. As x_n is a diagonal element of $\succeq 0$-matrix, we have $x_n \geq 0$. If $x_n = 0$, then $x = 0$ (By Sylvester, 2×2 principal minors in a $\succeq 0$-matrix are nonnegative $\Rightarrow$ *if a diagonal entry in a positive semidefinite matrix is 0, all entries in its row and column are zeros as well.* And if $x = 0$,then, of course, $x \in \mathbf{L}^n$. When $x_n > 0$, Schur Complement Lemma as applied to $\mathrm{Arrow}(x) \succeq 0$ says that $x_n \geq \frac{\sum_{i=1}^{n-1} x_i^2}{x_n}$, which combines with $x_n > 0$ to imply that $x_n \geq \sqrt{x_1^2 + \ldots + x_{n-1}^2} \Rightarrow x \in \mathbf{L}^n$. □

C. The next observation is a bit surprising:

Fact X.6 *As far as Lorentz representability is concerned, we need nothing but finite direct products of 1D and 3D Lorentz cones. Specifically, let $\underline{\mathfrak{C}}$ be the family of finite direct products of nonnegative rays and copies of $\mathbf{L}^3$. Then the cone $\mathbf{L}^n$, for every n, is $\underline{\mathfrak{C}}$-representable.*

Indeed, it suffices to consider the case when $n-1=2^K$ is an integer power of 2 (as for every $K=0,1,...$ and every positive integer $n<N:=2^K+1$, the cone $\mathbf{L}^n$ is the inverse image of $\mathbf{L}^N$ under the linear mapping $\mathbf{R}\ni x\mapsto[0;...;0;x]\in\mathbf{R}^N$). Here is a $\underline{\mathfrak{C}}$-r. of $\mathbf{L}^{2^K+1}$:

$$x_{2^K+1}\geq\sqrt[2^K]{x_1^2+,,,+x_{2^K}^2}$$
$$\Updownarrow$$
$$\exists u_{k\ell},\ 1\leq k\leq K-1, 1\leq\ell\leq 2^{K-k}:$$
$$u_{1,1}\geq\sqrt{x_1^2+x_2^2}, u_{1,2}\geq\sqrt{x_3^2+x_4^2}, u_{1,3}\geq\sqrt{x_5^2+x_6^2}, ..., u_{1,2^{K-1}}\geq\sqrt{x_{2^K-1}^2+x_{2^K}^2}$$
$$u_{2,1}\geq\sqrt{u_{1,1}^2+u_{1,2}^2}, u_{2,2}\geq\sqrt{u_{1,3}^2+u_{1.4}^2}, ..., u_{2,2^{K-2}}\geq\sqrt{u_{1,2^{K-1}-1}^2+u_{1,2^{K-1}}^2}$$
$$.....$$
$$u_{K-1,1}\geq\sqrt{u_{K-2,1}^2+u_{K-2,2}^2}, u_{K-1,2}\geq\sqrt{u_{K-2,3}^2+u_{K-2.4}^2}$$
$$x_{2^K+1}\geq\sqrt{u_{K-1,1}^2+u_{K-1,2}^2}$$

D. The next observation is much more surprising: in a sense, $\mathfrak{C}$ does not exist as an independent entity.

Fact X.7 [Fast polyhedral approximation of Lorentz cone] *For every n and every ϵ, $0 < \epsilon < 1/2$, one can point out a polyhedral set $\mathbf{L}^+$ given by an explicit system of homogeneous linear inequalities in variables $x \in \mathbf{R}^n$, $t \in \mathbf{R}$, $w \in \mathbf{R}^k$:*

$$\mathbf{L}^+ = \{[x;t;w] : Px + tp + Qw \leq 0\} \qquad (!)$$

such that

- *the number of inequalities in the system ($\approx 2n\ln(1/\epsilon)$) and the dimension of the slack vector w ($\approx 0.7n\ln(1/\epsilon)$) do not exceed $O(1)n\ln(1/\epsilon)$*
- *the projection*

$$\mathbf{L} = \{[x;t] : \exists w : Px + tp + Qw \leq 0\}$$

of $\mathbf{L}^+$ on the space of x,t-variables is in-between the Second Order Cone and $(1+\epsilon)$-extension of this cone:

$$\mathbf{L}^{n+1} := \{[x;t] \in \mathbf{R}^{n+1} : \|x\|_2 \leq t\} \subset \mathbf{L} \subset \mathbf{L}_\epsilon^{n+1} := \{[x;t] \in \mathbf{R}^{n+1} : \|x\|_2 \leq (1+\epsilon)t\}.$$

In particular, we have

$$B_n^1 \subset \{x : \exists w : Px + p + Qw \leq 0\} \subset B_n^{1+\epsilon}$$
$$B_n^r = \{x \in \mathbf{R}^n : \|x\|_2 \leq r\}$$

Note: When $\epsilon = 1.e\text{-}17$, your laptop does not distinguish between $r = 1$ and $r = 1 + \epsilon$. Thus, *for all practical purposes*, the n-dimensional Euclidean ball admits explicit polyhedral representation with $\approx 28n$ variables w and $\approx 79n$ linear inequality constraints.

For proof, see section 2.5 in [LMCO] (A. Ben-Tal, A. Nemirovski, *Lectures on Modern Convex Optimization 2020/2021/2022/2023/2024* `https://www2.isye.gatech.edu/~nemirovs/LMCOLN2024Spring.pdf`)

Note: A straightforward representation $X = \{x : Ax \leq b\}$ of a polyhedral set X satisfying

$$B_n^1 \subset X \subset B_n^{1+\epsilon}$$

requires at least $N = O(1)\epsilon^{-\frac{n-1}{2}}$ linear inequalities. With $n = 100$, $\epsilon = 0.01$, we get

$$N \geq 3.0\text{e}85 \approx 300,000 \times [\#\text{ of atoms in universe}]$$

With "fast polyhedral approximation" of B_n^1, a 0.01-approximation of B_{100} requires just 922 linear inequalities on 100 original and 325 additional variables.

♠ **Illustration:** Approximating 2D unit circle by projecting "high-dimensional" polytope:

##	Dumension of polytope	# of linear inequalities	Quality of approximation	Sides in equal quality polygonal approximation
1	10	12	5.e-3	16
2	13	18	3.e-4	127
3	19	30	7.e-8	8,192
4	31	54	4.e-15	34,200,93

♣ With fast polyhedral approximation of the cone $\mathbf{L}^{n+1} = \{[x;t] \in \mathbf{R}^{n+1} : \|x\|_2 \le t\}$, Conic Quadratic Optimization programs "for all practical purposes" become LO programs. For example, the program

$$\begin{gathered}
\text{minimize } c^T x \text{ subject to} \\
Ax = b \\
x \ge 0 \\
\left(\sum_{i=1}^{8} |x_i|^3\right)^{1/3} \le x_2^{1/7} x_3^{2/7} x_4^{3/7} + 2x_1^{1/5} x_5^{2/5} x_6^{1/5} \\
5x_2 \ge \frac{1}{x_1^{1/2} x_2^2} + \frac{2}{x_2^{1/3} x_3^3 x_4^{5/8}} \\
\begin{bmatrix} x_1 & x_2 & & \\ x_2 & x_3 & x_4 & \\ & x_4 & x_5 & x_6 \\ & & x_6 & x_7 \end{bmatrix} \succeq 8I
\end{gathered}$$

can be *in a systematic fashion* converted to Conic Quadratic Programming and thus "for all practical purposes" is just an LP program.

♠ **More surprises:** Exponent e^x which lives in our mind is defined on the real axis and rapidly grows/goes to zero as $x \to \infty / x \to -\infty$. The exponent which lives in your laptop is a different beast: it is a function with bounded domain.; according to your computer, $e^{759} = \infty$, and $e^{-759} = 0$. $\ln(e^{750}) = \infty$, And as it should be with this arithmetic, the log which lives in your laptop is a function with bounded range,
It turns out that *the exponent and the logarithm which live in computer are, for all practical purposes, polyhedral functions: given* $\epsilon \in (0, 1/2)$, *you can build polyhedral representations of functions* Exp *and* Ln *which approximate* exp *with relative accuracy* ϵ, *and* ln *with absolute accuracy* ϵ *in their "computer domains." These polyhedral representations are explicit and use just* $O(\ln(1/\epsilon))$ *variables...*

Historical note: Once upon a time... Fast polyhedral approximation of Lorentz cone was discovered circa 2000 when trying to process numerically about 100 CQP's with ≈1000 decision variables and conic quadratic constraints. This was done in course of a case study aimed at investigating the then-new methodology of *Robust Linear Programming.* At that time, in contrast to today, solvers capable to handle CQP's of these sizes were nonexistent, and fast polyhedral approximation allowed to process the CQP's by commercial LP solvers, quite powerful even in these old times.

Paupertas omnes artes perdocet, ubi quem attingit [Poverty teaches all arts, wherever it touches.– Titus Maccius Plautus (254 BC - 184 BC) , Roman comedian]

$\mathfrak{C}$-representable functions and sets

♣ Let us list several useful $\mathfrak{C}$-r functions and sets (for their explicit $\mathfrak{C}$-r.'s, see Lecture 2 in [LMCO])

♠ **Functions:**

- Euclidean norm $\|\cdot\|_2$: $t \geq \|x\|_2 \Leftrightarrow [x;t] \in \mathbf{L}^{\dim x+1}$
- Squared Euclidean norm x^Tx: $t \geq x^Tx \Leftrightarrow [2x; t-1; t+1] \in \mathbf{L}^{\dim x+2}$
- Univariate power functions:
  $(\max[x;0])^\pi : \mathbf{R} \to \mathbf{R}$ $[\pi \geq 1]$; $-x^\pi : \mathbf{R}_+ \to \mathbf{R}$ $[0 \leq \pi \leq 1]$; $x^{-\pi} : \operatorname{int}\mathbf{R}_+ \to \mathbf{R}$ $[\pi > 0]$
  with rational π
- Concave algebraic monomial $f(x) = x_1^{\pi_1}x_2^{\pi_2}...x_n^{\pi_n} : \mathbf{R}^n_+ \to \mathbf{R}$ with positive rational π_i, $\sum_i \pi_i \leq 1$
- Convex algebraic monomial $f(x) = x_1^{-\pi_1}x_2^{-\pi_2}...x_n^{-\pi_n} : \operatorname{int}\mathbf{R}^n_+ \to \mathbf{R}$ with positive rational π_i
- p-norm $\|x\|_p$ with rational $p \in [1, \infty]$
- Fractional-quadratic function $x^Tx/s : \mathbf{R}^n \times \mathbf{R}_+ \to \mathbf{R} \cup \{+\infty\}$:

$$t \geq x^Tx/s \Leftrightarrow [2x; t-s; t+s] \in \mathbf{L}^{n+2}$$

♠ **Sets:**

- Rotated Lorentz cone

$$\{[x;\alpha;\beta] \in \mathbf{R}^{n+2} : x^T x \leq \alpha\beta, \alpha, \beta \geq 0\} = \{[x;\alpha;\beta] : [2x, \beta-\alpha, \beta+\alpha] \in \mathbf{L}^{n+2}\}$$

- Epigraph of compliance $\mathrm{Compl}(t, f)$ – the set $\{[t; f; \tau] : \left[\begin{array}{c|c} B\mathrm{Diag}\{t\}B^T & f \\ \hline f^T & 2\tau \end{array}\right]\}$

$\mathfrak{S}$-representable sets and functions: Preliminaries

♣ In Semidefinite Programming with its tremendous expressive abilities, two facts are of paramount importance. The first is the already known to us Fact IX.5 :

[Schur Complement Lemma] *Consider symmetric block-matrix* $\left[\begin{array}{c|c} P & Q \\ \hline Q^T & R \end{array}\right]$ *with* $R \succ 0$, *Then*

$$\left[\begin{array}{c|c} P & Q \\ \hline Q^T & R \end{array}\right] \succeq 0 \Leftrightarrow P - QR^{-1}Q^T \succeq 0.$$

The other one is

Fact X.8 [$\mathcal{S}$-Lemma] *Let homogeneous quadratic inequality*

$$x^T Ax \geq 0 \qquad (A)$$

be strictly feasible: $\exists \bar{x} : \bar{x}^T A\bar{x} > 0$. *A homogeneous quadratic inequality*

$$x^T Bx \geq 0 \qquad (B)$$

is a consequence of (A) *iff it can be obtained by summing up a nonnegative multiple of* (A) *and an identically true homogeneous quadratic inequality, or, which is the same.*

$$\exists \lambda \geq 0 : B \succeq \lambda A.$$

$$x^TAx \geq 0 \Rightarrow x^TBx \geq 0$$
$$\Updownarrow$$
$$\exists \lambda \geq 0 : B \succeq \lambda A$$

Some comments are in order.

♠ $\mathcal{S}$-Lemma resembles the Homogeneous Farkas Lemma. HFL says that a homogeneous linear inequality is a consequence of a finite system of homogeneous linear inequalities iff the target inequality can be obtained by taking weighted sum, with nonnegative weights, of the inequalities from the system. One could add also "and adding an identically true homogeneous linear inequality," but it does not change anything – the only identically true homogeneous linear inequality is $0^Tx \geq 0$, and adding it does not help.
$\mathcal{S}$-Lemma is a similar statement but for homogeneous quadratic inequalities. with iwo important caveats:

• "a system" is now restricted to be a *single* homogeneous quadratic inequality and is assumed to be strictly feasible; the first restriction is severe, the second – of no actual importance;

• "adding identically true homogeneous quadratic inequality" now does help – there are plenty of identically true homogeneous quadratic inequalities, namely, those of the form $x^TCx \geq 0$ with $C \succeq 0$.

• Of course, the possibility to get the target inequality as a weighted sum of inequalities from the system and identically true inequality as a *sufficient* condition for the target inequality to be a consequence of the system is absolutely evident and has nothing to do with what are the inequalities in question and how many of them are there in the system. However, the actual power of HFL and $\mathcal{S}$-Lemma stems from the fact that under their premises the simple condition in question is *necessary*. In this respect situation with quadratic inequalities is intrinsically worse than with linear ones: already the "two inequalities in the system" version of $\mathcal{S}$-Lemma fails to be true in general...

• In spite of its seemingly heavily restricted, as compared to the HFL, scope, $\mathcal{S}$-Lemma is an indispensable tool in SDP.

♠ $\mathcal{S}$-Lemma admits inhomogeneous version:

Fact X.9 [Inhomogeneous $\mathcal{S}$-Lemma] *Assume that the quadratic inequality*

$$x^T Bx + 2b^T x + \beta \geq 0 \qquad (B)$$

is a consequence f strictly feasible quadratic inequality

$$x^T Ax + 2b^T x + \alpha \geq 0 \qquad (A)$$

iff the homogeneous version

$$x^T Bx + 2tb^T x + \beta t^2 \geq 0$$

of (B) is a consequence of the homogeneous version

$$x^T Ax + 2ta^T x + \alpha t^2 \geq 0$$

of (A), that is (by the plain $\mathcal{S}$-Lemma) iff

$$\exists \lambda \geq 0 : \left[\begin{array}{c|c} B & b \\ \hline b^T & \beta \end{array}\right] \geq \lambda \left[\begin{array}{c|c} A & a \\ \hline a^T & \alpha \end{array}\right].$$

For the proof of the latter fact, see section 18.4 in our Lecture Notes. An intelligent proof of $\mathcal{S}$-lemma follows.

The sufficiency of the condition

$$\exists \lambda : B \succeq \lambda A \qquad (*)$$

for the validity of the implication $x^T Ax \geq 0 \Rightarrow x^T Bx \geq 0$ is evident: in the case of $(*)$, the quadratic form $x^T Bx$ *everywhere* majorates the quadratic form $x^T Ax$ and therefore is nonnegative at every x where the latter form is so.

To prove the necessity, consider pair of implications

$$\begin{aligned} \forall \left(x \in \mathbf{R}^n, x^T Ax \geq 0\right) : & \quad x^T Bx \geq 0 \quad (!) \\ \forall \left(X \in \mathbf{S}^n_+, \mathrm{Tr}(AX) \geq 0\right) : & \quad \mathrm{Tr}(BX) \geq 0 \quad (!!) \end{aligned}$$

Note: $x^T Qx = \mathrm{Tr}(Axx^T)$ for all $x \in \mathbf{R}^n$ and all $Q \in \mathbf{S}^n$ $\Rightarrow$ $(!)$ *is weaker than* $(!!)$: $(!!)$ says that $\mathrm{Tr}(BX) \geq 0$ whenever $X \succeq 0$ is such that $\mathrm{Tr}(AX) \geq 0$, and $(!)$ says the same, but only for those $X \succeq 0$ which are representable as xx^T, that is, for *rank 1* matrices $X \succeq 0$.

• It is immediately seen that $(*)$ *is a necessary (and sufficient) condition for the validity of* $(!!)$. Indeed, consider the SDP problem

$$\mathrm{Opt} = \min_X \left\{\mathrm{Tr}(BX) : \mathrm{Tr}(AX) \geq 0, X \succeq 0\right\}.$$

$(!!)$ says exactly that $\mathrm{Opt} \geq 0$. When the inequality $x^T Ax \geq 0$ is strictly feasible, the above SDP satisfies the Relaxed Slater condition (why?) $\Rightarrow$ by Conic Duality Theorem, $\mathrm{Opt} \geq 0\}$ iff the optimal value in the dual problem

$$\max_{\lambda, \Lambda} \{\mathrm{Tr}(0_{n\times n}\Lambda) : B = \lambda A + \Lambda, \lambda \geq 0, \Lambda \succeq 0\}$$

is solvable with optimal value ≥ 0, that is, iff $(*)$ holds true.

$$\begin{array}{rll} & \exists \lambda : B \succeq \lambda A & (*) \\ \forall \left(x \in \mathbf{R}^n, x^T A x \geq 0 \right) : & x^T B x \geq 0 & (!) \\ \forall \left(X \in \mathbf{S}^n_+, \mathrm{Tr}(AX) \geq 0 \right) : & \mathrm{Tr}(BX) \geq 0 & (!!) \end{array}$$

• We see that *all we need in order to prove that* $(*)$ *is necessary for the validity of* (!) *is to verify that the implications* (!) *and* (!!) *are equivalent to each other*; given that (!) is weaker than (!!), this is the same as to verify that

(!!!) Whenever (!) holds true, so is (!!), that is, *if* $x^T B x \geq 0$ *whenever* $x^T A x \geq 0$, *then* $\mathrm{Tr}(BX) \geq 0$ *for every* $X \succeq 0$ *such that* $\mathrm{Tr}(AX) \geq 0$.

Here is the demonstration: Let (!)) and hold true, and let $X \succeq 0$ be such that $\mathrm{Tr}(AX) \geq 0$; we should prove that $\mathrm{Tr}(BX) \geq 0$. Let

$$X^{1/2} A X^{1/2} = U \mathrm{Diag}\{\lambda\} U^T \quad [\Leftrightarrow \mathrm{Diag}\{\lambda\} = U^T [X^{1/2} A X^{1/2}] U$$

be the eigenvalue decomposition of the symmetric matrix x, and let ζ be a Rademacher random vector of dimension n, that is, entries in ζ are independent of each other and take values ± 1 with probabilities $1/2$. Let us set $\xi = X^{1/2} U \zeta$. Then

$$\begin{array}{l} \xi^T A \xi = \zeta^T U^T X^{1/2} A X^{1/2} U \zeta = \zeta^T U^T [X^{1/2} A X^{1/2}] U \zeta = \zeta^T U^T [U \mathrm{Diag}\{\lambda\} U^T] U \zeta \\ = \zeta^T \mathrm{Diag}\{\lambda\} \zeta \underbrace{\equiv}_{a} \mathrm{Tr}(\Lambda) = \mathrm{Tr}(U^T X^{1/2} A X^{1/2} U^T) \underbrace{=}_{b} \mathrm{Tr}(A [X^{1/2} U^T U X^{1/2}]) = \mathrm{Tr}(AX) \geq 0 \end{array}$$

where a ia due to all realizations of ζ being vectors with ± 1 entries, and b is due to the following nice property of trace:

Whenever rectangular matrices P, Q *are such that* PQ *makes sense and is a square matrix, it holds* $\mathrm{Tr}(PQ) = \mathrm{Tr}(QP)$. In (B), we use this property for $P = U^T X^{1/2}$ and $Q = A X^{1/2} U^T$.

As $\xi^T A \xi \geq 9$ for all realizations of ξ, we have

$$\begin{array}{l} 0 \leq \mathbf{E}\{\xi^T B \xi\} = \mathbf{E}\{\zeta^T U^T X^{1/2} B X^{1/2} U \zeta\} = \mathbf{E}\{\zeta^T U^T [X^{1/2} B X^{1/2}] U \zeta\} = \mathbf{E}\{\sum_{i,j} \zeta_i \zeta_j [U^T [X^{1/2} B X^{1/2}] U]_{ij}\} \\ = \sum_{i,j} \mathbf{E}\{\zeta_i \zeta_j\} [U^T [X^{1/2} B X^{1/2}] U]_{ij} = \mathrm{Tr}(U^T [X^{1/2} B X^{1/2}] U) = \mathrm{Tr}(B [X^{1/2} U] [U^T X^{1/2}]) = \mathrm{Tr}(BX), \end{array}$$

i.e., $\mathrm{Tr}(BX) \geq 0$, Q.E.D.

For "bare hands" proof of $\mathcal{S}$-Lemma, based on Optimality conditions in Mathematical Programming, see section 21.3.1 of Lecture Notes.

$\mathfrak{S}$-representable sets and functions: Majorization Principle

♣ To proceed we need some basic facts on symmetric functions and Majorization principle.

♠ **Symmetric functions.** A function $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ is called *symmetric*, if its values remain intact under permutation of entries in the argument: that is,

$$\forall (x \in \mathbf{R}^n, \sigma \in \Sigma_n) : f(x^\sigma) = f(x),$$
$$\left[\Sigma_n : \text{ set of all } n! \text{ permutations of } 1, ..., n, [x^\sigma]_i = x_{\sigma(i)},\ i = 1, ..., n\right]$$

Examples include: $\sum_i x_i$, $\prod_i x_i$, the sum $s_k(x)$ of the $k \leq n$ largest entries in x,...

Fact X.10 *Let $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ be a convex symmetric function, π be a doubly stochastic $n \times n$ matrix, and let $x \in \mathbf{R}^n$. Then*

$$f(\pi x) \leq f(x).$$

Indeed, by Birkhoff Theorem, π is a convex combination of permutation matrices $\Rightarrow$ $\pi x \in \text{Conv}(\{x^\sigma, \sigma \in \Sigma_n\})$ $\Rightarrow$ $f(\pi x) \leq \max_\sigma f(x^\sigma)$ (as f is convex) $\Rightarrow$ $f(\pi x) \leq f(x)$ (as $f(x^\sigma) = f(x)$ for all σ due to the symmetry of f). □

♠ **Majorization principle** provides characterization of points you can get by multiplying a given x by doubly stochastic matrices (or, which is the same, describes $\mathrm{Conv}(\{x^\sigma, \sigma \in \Sigma_n\})$:

Fact X.11 *Let $x \in \mathbf{R}^n$, A vector $y \in \mathbf{R}^n$ can be represented as πx with doubly stochastic π iff*

$$s_k(y) \le s_k(x),\ k < n \ \&\ s_n(y) = s_n(x),$$

where $s_k(z)$ is the sum of k largest entries in z

Proof of the "only if" part (for the proof of the "if" part, see the proof of Theorem II.7.15 in Lecture Notes). Observe that $s_k(x)$ is convex – it is the maximum of linear functions $\sum_{i\in I} x_i$ taken over all k-element subsets of the index set $\{1, 2, ..., n\}$; and of course s_k is symmetric. Therefore if $y = \pi x$ with doubly stochastic π, by Fact X.10 for every $k \le n$ it holds $s_k(y) \le s_k(x)$, and of course $s_n(y) = \sum_i y_i = \mathbf{1}^T \pi x = \sum_i x_i = s_n(x)$.

Note: *For $x, y \in \mathbf{R}^m$, the condition $s_k(y) \le s_k(x)$, $k \le m$, is necessary and sufficient for existence of double-stochastic matrix π such that $y \le \pi x$.*

Proof: ✓ If part: The functions $s_k(\cdot)$ clearly are monotone, so that when $y \le \pi x$ with double-stochastic π, we have $s_k(y) \le s_k(\pi x)$, and the latter quantity, as we know, is $\le s_k(x)$.

✓ Only if part: Let $s_k(y) \le s_k(x)$, $k \le m$. Let x_t be obtained from x by decreasing by t the smallest entry in x and keeping the remaining entries intact. We have $s_k(x_t) = s_k(x)$, $k < m$, and $s_m(x_t) = s_m(x) - t$. Setting $t = s_m(x) - s_m(y)$, we get $s_k(x_t) \ge s_k(y)$, $k < m$, and $s_m(x_t) = s_m(y)$. By Majorization principle, $y = \pi x_t$ for some double-stochastic matrix π, and $\pi x_t \le \pi x$ since $x_t \le x \Rightarrow y \le \pi x$.

♣ To prepare ourselves to what follows, let us answer the following question: as it was already explained, s_k is the maximum of $\binom{n}{k}$ linear forms, whence its epigraph admits polyhedral description – polyhedral representation with no additional variables – with $\binom{n}{k}$ linear inequalities. Does it admit a shorter $\mathfrak{P}$-r,? The answer is positive:

Fact X.12 *Whenever $k \leq n$, one has*

$$t \geq s_k(x) \Leftrightarrow \exists z \in \mathbf{R}^n, s \in \mathbf{R} : x \leq z + s\mathbf{1},\ z \geq 0, \sum_i z_i + ks \leq t \qquad [x \in \mathbf{R}^n]$$

where $\mathbf{1}$, as always, is the all-ones vector of the context-specified dimension (in our case, of dimension n).

Indeed, from the original description of s_k as the maximum of linear form it follows that s_k is $\leq$-nondecreasing, and, moreover, $s_k(x) = \max_I x^T e^I$, where $e^i \in \mathbf{R}^n$ is a Boolean vector with $e^I_i = 1$ exactly for $i \in I$, and I runs through the set of k-element subsets of $\{1, ..., n\}$. As we remember, e^I are exactly the extreme points of the bounded polyhedral set $X_k = \{u \in \mathbf{R}^n : 0 \leq u_i \leq 1\, \forall i, \sum_i u_i = k\}$, yielding the first equality in the following computation

$$\begin{array}{rcl} s_k(x) & = & \max_u\{x^T u : 0 \leq u_i \leq 1\, \forall i, \sum_i u_i = k\} \\ & = & \min_{s,z}\{\sum_i z_i + ks : z \geq 0, x \leq z + s\mathbf{1}\} \end{array}$$

where the second equality is due to LP duality (z is the vector of Lagrange multipliers for the bounds $u_i \leq 1$, s – for the equality constraint $\sum_i u_i = k$; the multipliers for the lower bounds $u_i \geq 0$ admit immediate elimination). □

Preliminaries on eigenvalues

♣ **Must to know: eigenvalues.** Aside of their definition and Eigenvalue Decomposition Theorem (see, e..g., Appendix D in Lecture Notes), you should remember that

• for a matrix $A \in \mathbf{S}^n$,

$$\lambda(A) = [\lambda_1(A); ...; \lambda_n(A)]$$

stands for the vector of eigenvalues of A taken with their multiplicities in the non-ascending order:

$$\lambda_1(A) \geq \lambda_2(A) \geq ... \geq \lambda_n(A).$$

Sometimes we shall write $\lambda_{\max}(A)$ instead of $\lambda_1(A)$ and $\lambda_{\min}(A)$ instead of $\lambda_n(A)$.

Note: *Eigenvalues are rotationally invariant: $\Lambda(X) = \lambda(UXU^T)$ for orthogonal U.*

♠ **$\succeq$-monotonicity of eigenvalues.** *The vector-valued function $X \mapsto \lambda(X) : \mathbf{S}_n \to \mathbf{R}^n$ is $\succeq$-monotone:*

$$X' \succeq X \Rightarrow \lambda(X') \geq \lambda(X).$$

This fact is one of the consequences of *Variational Characterization of Eigenvalues* of symmetric matrix – the single most important fact about eigenvalues of symmetric matrices, essentially more informative than their initial algebraic definition.

> ♠ [Variational Characterization of Eigenvalues] *For an $n \times n$ symmetric matrix A one has*
>
> $$\lambda_i(A) = \min_{E \in \mathcal{E}_i} \max_{e \in E, \|e\|_2 = 1} e^T A e, \ 1 \leq i \leq n,$$
>
> *where $\mathcal{E}_i$ is the family of all linear subspaces of $\mathbf{R}^n$ of dimension $n - i + 1$.*

♠ **Eigenvalues: convexity status.** As a matter of fact, $\lambda_{\max}(X)$ is convex function of $X \in \mathbf{S}^n$, $\lambda_{\min}(x) = -\lambda_{\max}(-X)$ is concave; intermediate eigenvalues have no definite convexity status. And of course $S_n(X) = \mathrm{Tr}(X)$ is just linear. What *is* convex, are the sums

$$S_k(X) = \sum_{i \le k} \lambda_i(X)$$

of $k \le n$ of the largest eigenvalues (whence the sum of $k \le n$ smallest eigenvalues, that is, the function $-S_k(-X)$ is concave). Convexity of S_k will play the central role in specifying "$\mathfrak{G}$–raw materials," this is why we start with deriving $\mathfrak{G}$-r. of S_k.

Fact X.13 *The function $S_k(X) : \mathbf{S}^n \to \mathbf{R}$ $(k \le n)$ is $\succeq$-nondecreasing and is $\mathfrak{S}$-r:*

$$S_k(X) \le t \Leftrightarrow \exists s, Z : \left\{ \begin{array}{lrcl} (a) & ks + \mathrm{Tr}(Z) & \le & t \\ (b) & Z & \succeq & 0 \\ (c) & X & \preceq & Z + sI_m \end{array} \right.$$

Pay attention to similarity of this $\mathfrak{S}$-r. of $S_k(X) = s_k(\lambda(X))$ and our $\mathfrak{P}$-r. of $s_k(x)$; the first is inspired by the second via the following idea which sometimes (not always!) works: *think of vectors $x \in \mathbf{R}^n$ as of diagonal entries of diagonal matrices $X \in \mathbf{S}^n$* (and thus - vectors of eigenvalues of X) *do what you need to do, and then erase the word "diagonal" and look what happens...*

Proof. We should prove that

- If a pair X, t can be extended, by properly chosen s, Z, to a solution of $(a) - (c)$, then $S_k(X) \le t$;
- If $S_k(X) \le t$, then the pair X, t can be extended by properly chosen s, Z, to a solution of $(a) - (c)$.

✓ Let us prove that if a pair X, t can be extended, by properly chosen s, Z, to a solution of $(a) - (c)$, then $S_k(X) \le t$. Indeed, let (X, t, s, Z) solve $(a) - (c)$. Then

$$\begin{array}{lcl} & X \preceq Z + sI_m & [\text{by } (c)] \\ \Rightarrow & \lambda(X) \le \lambda(Z + sI_m) = \lambda(Z) + s \left[\begin{array}{c} 1 \\ \vdots \\ 1 \end{array} \right] & [\text{by } \succeq\text{-monotonicity of eigenvalues}] \\ \Rightarrow & S_k(X) \le S_k(Z) + sk & \\ \Rightarrow & S_k(X) \le \mathrm{Tr}(Z) + sk & \left[\begin{array}{l} \text{since } S_k(Z) \le \mathrm{Tr}(Z) \\ \text{due to } (b) \end{array} \right] \\ \Rightarrow & S_k(X) \le t & [\text{by } (a)] \end{array}$$

$$S_k(X) \le t \ ?\Leftrightarrow? \ \exists s, Z : \left\{ \begin{array}{lrcl} (a) & ks + \mathrm{Tr}(Z) & \le & t \\ (b) & Z & \succeq & 0 \\ (c) & X & \preceq & Z + sI_m \end{array} \right.$$

✓ Now let us prove that if $S_k(X) \le t$, then X, t can be augmented by s, Z to satisfy $(a) - (c)$. Indeed, let $S_k(X) \le t$, let $X = U\mathrm{Diag}\{\lambda\}U^T$, $\lambda = \lambda(X)$, be the eigenvalue decomposition of X. Setting

$$s = \lambda_k, \quad Z = U \underbrace{\left[\begin{array}{cccccc} \lambda_1 - \lambda_k & & & & & \\ & \ddots & & & & \\ & & \lambda_{k-1} - \lambda_k & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{array} \right]}_{\mathrm{Diag}_{\{\lambda(Z)\}}} U^T,$$

we have

$$\begin{array}{c} Z \succeq 0, \\ \mathrm{Diag}\{\lambda(X)\} \le \mathrm{Diag}\left\{ \lambda(Z) + s \left[\begin{array}{c} 1 \\ \vdots \\ 1 \end{array} \right] \right\} \Rightarrow X \preceq Z + sI_m, \\ t \ge S_k(X) = ks + \mathrm{Tr}(Z), \end{array}$$

so that (t, X, s, Z) solves the system of LMIs

$$\begin{array}{llcl} (a) & ks + \mathrm{Tr}(Z) & \le & t \\ (b) & Z & \succeq & 0 \\ (c) & X & \preceq & Z + sI_m \end{array}$$

Q.E.D.

$\mathfrak{S}$-representability of functions of eigenvalues

Fact X.14 *Let $f(x)$ be a symmetric convex function on $\mathbf{R}^m$. Then the function*

$$F(X) = f(\lambda(X))$$

is convex on $\mathbf{S}^m$, and, moreover,

$$F(X) = \max_{U:U^TU=I} f(\mathrm{Dg}(UXU^T)). \qquad (*)$$

[$\mathrm{Dg}(A)$ – *vector of diagonal entries of square matrix A*]

Proof: It suffices to verify $(*)$; indeed, given $(*)$, $F(\cdot)$ is convex as the upper bound, w.r.t. orthogonal U, of the family of (clearly convex) functions $f_U(\cdot)$. For properly chosen orthogonal U we have

$$UXU^T = \mathrm{Diag}\{\lambda(X)\} \Rightarrow \max_{U:U^TU=I} f(\mathrm{Dg}(UXU^T)) \geq f(\lambda(X)).$$

To prove the opposite inequality, observe that every matrix of the form UXU^T with orthogonal U is of the form $V\mathrm{Diag}\{\lambda(X)\}V^T$ with orthogonal V as well. Now,

$$[\mathrm{Dg}(UXU^T)]_i = [V\mathrm{Diag}\{\lambda(X)\}V^T]_{ii} = \sum_j V_{ij}^2\lambda_j(X),$$

that is, $\mathrm{Dg}(UXU^T) = \pi\lambda(X)$ for the *double stochastic* matrix $\pi = [V_{ij}^2]_{i,j}$. Therefore, by Fact X.10,

$$f(\mathrm{Dg}(UXU^T)) = f(\pi\lambda(X)) \leq f(\lambda(X)) \qquad \square\ .$$

Examples: For every symmetric matrix X with the vector of eigenvalues λ one has

- The sum of k largest diagonal entries of X does not exceed $S_k(X) = \lambda_1 + ... + \lambda_k$

$[f(x) = \max\limits_{i_1<i_2<...<i_k} [x_{i_1} + ... + x_{i_k}]$ is the sum of k largest entries in $x]$

- The sum of k smallest diagonal entries in X is at least the sum of k smallest of λ_i's
- If $X \succ 0$, then the product of the k smallest diagonal entries in X is at least the product of the k smallest of λ_i's. In particular, the product of all diagonal entries in X is $\geq \mathrm{Det}(X)$.

$[g(x) = \min\limits_{i_1<i_2<...<i_k} [\ln x_{i_1} + ... + \ln x_{i_k}]$ is the sum of logs of k smallest entries in $x > 0$, $f(x) = -g(x)]$

♣ Combining $\mathfrak{S}$-representability of $S_k(\cdot)$ and Majorization Principle, we arrive at the following important result:

Fact X.15 [SDP-r.'s of symmetric $\mathfrak{S}$-r functions of eigenvalues] *Let $f(x)$ be a $\mathfrak{S}$-r symmetric function on $\mathbf{R}^m$. Then the function*

$$F(X) = f(\lambda(X)) : \mathbf{S}^m \to \mathbf{R} \cup \{+\infty\}$$

is $\mathfrak{S}$-r with $\mathfrak{S}$-r. readily given by $\mathfrak{S}$-r. of f. In particular, the following functions are $\mathfrak{S}$-r with explicit $\mathfrak{S}$-r.'s

- $-\mathrm{Det}^\pi(X)$, $X \in \mathbf{S}^m_+$ *($\pi \in (0, \frac{1}{m}]$ is rational);*
- $\mathrm{Det}^{-\pi}(X)$, $X \succ 0$ *($\pi > 0$ is rational);*
- $|X|_\pi = \|\lambda(X)\|_\pi$, $X \in \mathbf{S}^m$ *($\pi \in [1, \infty)$ is rational or $\pi = \infty$).*

Proof. Let

$$t \geq f(x) \Leftrightarrow \exists u : \mathcal{A}(t, x, u) \succeq 0.$$

Then

$$t \geq F(X) \Leftrightarrow t \geq f(\lambda(X)) \Leftrightarrow \exists (y \in \mathbf{R}^m, \pi \in P_m) : \left\{ \begin{array}{l} y_1 \geq y_2 \geq ... \geq y_m \\ f(y) \leq t \\ \lambda(X) = \pi y \\ [\text{since } f(\pi y) \leq f(y)] \end{array} \right.$$

$$\Rightarrow \quad t \geq F(X) \Leftrightarrow \exists y \in \mathbf{R}^m : \left\{ \begin{array}{l} y_1 \geq y_2 \geq ... \geq y_m,\ f(y) \leq t \\ s_k(\lambda(X)) \leq y_1 + ... + y_k,\ k < m \\ s_m(\lambda(X)) = y_1 + ... + y_m \\ [\text{by Majorization Principle}] \end{array} \right.$$

$$\Rightarrow \quad t \geq F(X) \Leftrightarrow \exists (y \in \mathbf{R}^m, u) : \left\{ \begin{array}{l} y_1 \geq y_2 \geq ... \geq y_m,\ \mathcal{A}(y, t, u) \succeq 0 \\ \underbrace{S_k(X) \leq y_1 + ... + y_k}_{\text{SD-representable!}},\ k < m \\ \mathrm{Tr}(X) = y_1 + ... + y_m \end{array} \right.$$

□

$\mathfrak{S}$-representability of functions of singular values

♣ **Must to know: singular values.** Linear Algebra says that an $m \times n$ matrix A admits *singular value decomposition* (svd)

$$A = UDV^T,$$

where U is $m \times m$, and V is $n \times n$ orthogonal matrix:

$$U^TU = I_m,\ V^TV = I_n,$$

and D is $m \times n$ is *diagonal* matrix with nonnegative diagonal entries, diagonality meaning that the only nonzero entries are among D_{ii},$1 \le i \le r = \min[m; n]$. These diagonal entries are called *singular values* of A.
When taken in the non-ascending order, singular values form a vector

$$\sigma(A) = [\sigma_1(A); ...\sigma_r(A)],\ \sigma_1(A) \ge \sigma_2(A) \ge ... \ge \sigma_r(A) \qquad [r = \min[m, n]]$$

which, in contrast to the svd, are uniquely defined by A. In other words, A can be represented as

$$A = \sum_{i=1}^{r} \sigma_i(A) u_i v_i^T,$$

where the vectors u_i, $1 \le i \le r$, form an orthonormal system in $\mathbf{R}^m$, and the vectors v_i form an orthonormal system in $\mathbf{R}^n$.
Note: $\sigma(X) = \sigma(X^T)$, and $\sigma(X)$ is rotationally invariant: $\sigma(X) = \sigma(UXV^T)$ for orthogonal U, V.
• When A is symmetric, the singular values of A are the magnitudes of the eigenvalues.

♠ Singular values of $A \in \mathbf{R}^{m\times n}$ are closely related to eigenvalues of the symmetric $(m+n)\times(m+n)$ matrix

$$\mathsf{S}(A) = \left[\begin{array}{c|c} & A \\ \hline A^T & \end{array}\right],$$

linearly depending on A; specifically, *the eigenvalues of* $\mathsf{S}(A)$ *are the* $r = \min[m,n]$ *singular values of* A, r *negations of these singular values, and* $m+n-2r$ *zeros.*

♠ Combining the latter fact with calculus of $\mathfrak{S}$-r.'s with our results on $\mathfrak{S}$-r.'s of eigenvalues of symmetric matrices, we get "for free" $\mathfrak{S}$-r.'s of functions of singular values:

Fact X.16 *Let m, n be positive integers, $r = \min[m, n]$, and $f : \mathbf{R}^r_+ \to \mathbf{R} \cup \{+\infty\}$ be a symmetric, convex nondecreasing function. Then the function*

$$F(X) = f(\sigma(X)) : \mathbf{R}^{m\times n} \to \mathbf{R} \cup \{+\infty\}$$

is convex, and an $\mathfrak{S}$-r. of f induces straightforwardly an $\mathfrak{S}$-r. of F. In particular, the following functions of $X \in \mathbf{R}^{m\times n}$ are $\mathfrak{S}$-r with explicit $\mathfrak{S}$-r.'s:

- *spectral norm $\sigma_1(X) = \|X\| := \max_x\{\|X\|_2 : \|x\|_2 \le 1\}$:*

$$t \ge \|X\| \Leftrightarrow \left[\begin{array}{c|c} I_m & X \\ \hline X^T & I_n \end{array}\right] \succeq 0$$

- *the sum $\Sigma_k(X) = \sum_{i=1}^k \sigma_i(X)$ of $k \le r$ largest singular values*
- *the Shatten p-norm $\|X\|_{\mathsf{Sh},p} = \|\sigma(X)\|_p$, $p \in [1, \infty]$; this function is convex (and is indeed a norm); it is $\mathfrak{S}$-r with explicit $\mathfrak{S}$-r. when p is rational*

Note: *Shatten 2-norm, a.k.a. Frobenius norm, is just $\sqrt{Tr(XX^T)} = \sqrt{\sum_{i,j} X_{ij}^2}$, and Shatten norms are rotationally invariant: $\|X\|_{\mathsf{Sh},p} = \|UXV^T\|_{\mathsf{Sh},p}$ for orthogonal U, V; besides this, $\|X\|_{\mathsf{Sh},p} = \|X^T\|_{\mathsf{Sh},p}$.*

Note: *Same as for the usual ℓ_p-norm, the conjugate of $\|X\|_{\mathsf{Sh},p}$ is $\|X\|_{\mathsf{Sh},q}$, $\frac{1}{p} + \frac{1}{q} = 1$, and "matrix Hölder inequality" reads*

$$\mathsf{Tr}(XY^T) \le \|X\|_{\mathsf{Sh},p}\|Y\|_{\mathsf{Sh},q}, \frac{1}{p} + \frac{1}{q} = 1 \qquad [X, Y \in \mathbf{R}^{m\times n}, p, q \in [1, \infty]]$$

SDP-representability and polynomials

♣ We associate with univariate algebraic polynomial

$$p(s) = p_0 + p_1 s + ... + p_d s^d \qquad [p_i \in \mathbf{R}]$$

of degree $\leq d$ the vector $p = [p_0; ...; p_d] \in \mathbf{R}^{d+1}$. Similarly, we associate with univariate trigonometric polynomial

$$p(s) = p_0 + \sum_{\ell=1}^{d} [p_{2\ell-1} \cos(\ell s) + p_{2\ell} \sin(\ell s)] \qquad [p_i \in \mathbf{R}]$$

of degree $\leq d$ the vector of its coefficients $p = [p_0; ..., p_{2d}] \in \mathbf{R}^{2d+1}$.

♠ Given a set $\Delta \subset \mathbf{R}$ (Δ can be a segment, a ray, the entire reals axis, or a finite union of segments and rays) and nonnegative integer d the sets $\mathcal{P}_d(\Delta)$, $\mathcal{T}_d(\Delta)$ of vectors of coefficients of algebraic/trigonometric polynomials of degree $\leq d$ which are nonnegative on Δ. Clearly, $\mathcal{P}_d(\Delta)$ and $\mathcal{T}_d(\Delta)$ are closed convex cones. A remarkable fact, due to Yu. Nesterov, is

Fact X.17 *The cones $\mathcal{P}_d(\Delta)$, $\mathcal{T}_d(\Delta)$ are $\mathcal{G}$-r with explicit $\mathcal{G}$-r.'s — they are images of the semidefinite cone $\mathbf{S}^D_+$ of appropriate dimension under explicitly given linear mapping.*

For example

$$\mathcal{P}_{2d}(\mathbf{R}) = \{p \in \mathbf{R}^{2d+1} : \exists [P]_{\substack{0\leq i\leq d \\ 0\leq j\leq d}} \in \mathbf{S}^{d+1}_+ : p_k = \sum_{i,j:i+j=k} P_{ij}\}$$

This is a direct consequence of the remarkable algebraic fact which states that *a univariate* (unfortunately, just univariate!) *algebraic polynomial of degree $\leq 2d$ is everywhere nonnegative iff it is the sum of squares of polynomials of degree $\leq d$* (in fact, just two of them).

♠ As a result of $\mathfrak{S}$-representability of $\mathcal{P}_d$ and $\mathcal{T}_d$,

- finding the minimum of an algebraic polynomial $p(s)$ over a segment Δ reduces to SDP

$$\max_t \{t : p(\cdot) - t \in \mathcal{P}_d(\Delta)\}$$

and similarly for finding the minimum of a trigonometric polynomial

- design specification $\underline{q}(s) \le p(s) \le \overline{q}(s)$, $s \in \Delta$, on the coefficients of variable trigonometric polynomial of degree $\le d$, $\underline{q}, \overline{q}$ being given trigonometric polynomials of degree $\le d$ (these specifications arise when designing controllers and arrays of antennae) – a specific *infinite* system of linear constraints on p – reduces to LMI constraints

$$p - \underline{q} \in \mathcal{T}_d(\Delta),\ \overline{q} - p \in \mathcal{T}_d(\Delta)$$

SDP representability of some matrix-valued functions

♣ Some important functions F taking values in $\mathbf{S}^m$ are $\mathfrak{G}$-representable, meaning that their $\succeq$-epigraphs

$$\{[x, Y] : Y \succeq F(x)\}$$

are $\mathfrak{G}$-representable, so that the constraints $F(X) \preceq A$ are representable by LMIs.
Examples include

- $\succeq$-convex quadratic form $F(X) = AXQQ^TX^TA^T + BXC + C^TX^TB^T + D$ of rectangular matrix $X \in \mathbf{R}^{p\times q}$, with coefficients $D \in \mathbf{S}^m$, Q, A, B, C of appropriate size. Indeed, by Schur Complement Lemma (SCL for short),

$$Y \succeq F(X) \Leftrightarrow \left[\begin{array}{c|c} Y - BXC - C^TX^TB^T - D & AXQ \\ \hline X^TA^TQ^T & I \end{array}\right] \succeq 0$$

- fractional-quadratic function $F(U, V) = UV^{-1}U^T : \mathbf{R}^{p\times q}_U \times \operatorname{int}\mathbf{S}^q_+ \to \mathbf{S}^p$. By SCL,

$$Y \succeq F(U, V) \Leftrightarrow \exists W \in \mathbf{S}^q : \left[\begin{array}{c|c} Y & U \\ \hline U^T & V \end{array}\right] \succeq 0, \underbrace{\left[\begin{array}{c|c} V & I_q \\ \hline I_q & W \end{array}\right] \succeq 0}_{\text{says that } V \succ 0}$$

• matrix square root $F(X) = X^{1/2} : \mathbf{S}^n_+ \to \mathbf{S}^n$ This function is $\succeq$-concave, so that what is $\mathfrak{G}$-r, is the $\succeq$-hypograph of F:

$$X \succeq 0, Y \preceq X^{1/2}, \Leftrightarrow \exists U \succeq 0 : \left[\begin{array}{c|c} X & U \\ \hline U & I \end{array}\right] \ \& \ Y \preceq U$$

Note: $F(X)$ is not only $\succeq$-concave, it is also $\succeq$-monotone on $\mathbf{S}^n_+$: whenever $0 \preceq Z \preceq X$, one has $Z^{1/2} \preceq X^{1/2}$. This (not that trivial!) fact combines with $\mathfrak{G}$-representability of $X^{1/2}$ to imply $\succeq$-monotonicity and $\mathfrak{G}$-representability of the functions $X^{1/2^k} : \mathbf{S}^n_+ \to \mathbf{S}^n$, $k = 1, 2, ...$ Taking into account that $\ln(s) = \lim_{k\to\infty} 2^k[s^{1/2^k} - 1]$, this implies $\succeq$-monotonicity and $\succeq$-concavity of *matrix logarithm* $\ln(X)$ of $X \succ 0$ – symmetric matrix Y such that $\mathrm{e}^Y = X$, where the *matrix exponent* e^X is given by the standard definition

$$\mathrm{e}^X = \lim_{k\to\infty} (I + X/k)^k = \sum_{i=0}^{\infty} \frac{1}{i!} X^i.$$

Note that e^X for any $X \in \mathbf{S}^n$, and $\ln(X)$ for $X \succ 0$, share with X its eigenvectors, and their eigenvalues, as it should be, are

$$\lambda_i(\mathrm{e}^X) = \mathrm{e}^{\lambda_i(X)}, \ \lambda_i(\ln(X)) = \ln(\lambda_i(X)),$$

and, of course, $e^{\ln(X)} \equiv X$ for $X \succ 0$.

Note: e^X is neither $\succeq$-monotone, nor $\succeq$-convex, unless $n = 1$. However, $\mathrm{Tr}(\mathrm{e}^X) = \sum_i e^{\lambda_i(X)}$ *is* $\succeq$-monotone (along with $\lambda(X)$) and convex (as a symmetric convex function of $\lambda(X)$, see Fact X.15).

• I do not know whether the (minus) matrix entropy $X \ln(X) : \mathbf{S}^n_+ \to \mathbf{S}^n$ is or is not $\succeq$-convex; its trace, called the (minus) von Neumann, or quantum, entropy is convex by Fact X.15.

Funny fact: For $X, Y \in \mathbf{S}^n$, the sets

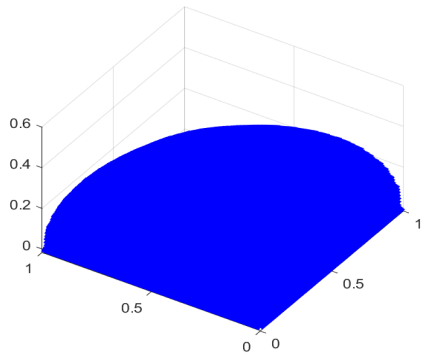
$$\{(X,Y) : X \succeq 0, 0 \preceq Y \leq X^{1/2}\}$$

and

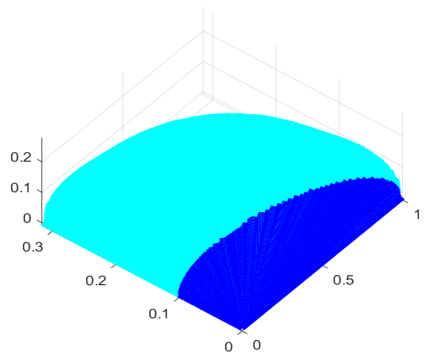
$$\{(X,Y) : X \succeq 0, Y \succeq 0, Y^2 \preceq X\}$$

are both convex and $\mathfrak{S}$-r with explicit $\mathfrak{S}$-r.'s, but differ from each other (unless $n = 1$), and the second set is a "negligible part" of the first one. The reason is that when $n \geq 2$, the function $X \mapsto X^2 : \mathbf{S}^n_+ \to \mathbf{S}^n$ is not $\succeq$-monotone.

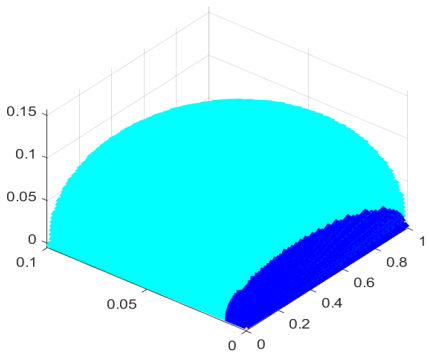
Illustration:



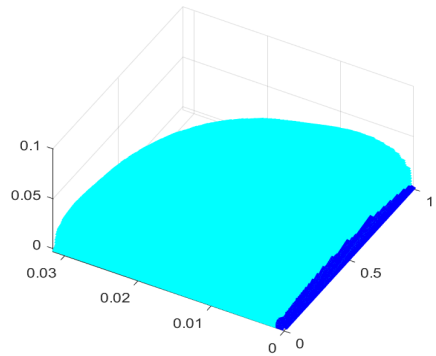
$$X = \left[\begin{array}{c|c} 1 & \\ \hline & 1 \end{array}\right]$$



$$X = \left[\begin{array}{c|c} 1 & \\ \hline & 0.1 \end{array}\right]$$



$$X = \left[\begin{array}{c|c} 1 & \\ \hline & 0.01 \end{array}\right]$$



$$X = \left[\begin{array}{c|c} 1 & \\ \hline & 0.001 \end{array}\right]$$

- What you see: 3D domains of entries x, y, z in matrix $\left[\begin{array}{c|c} x & z \\ \hline z & y \end{array}\right]$ satisfying $0 \preceq Y$, $Y^2 \preceq X$ (dark blue) and $0 \preceq Y \preceq X^{1/2}$ (dark blue and cyan). Note that left-most domain is just the set $\{x, y, z : 0 \preceq \left[\begin{array}{c|c} x & z \\ \hline z & y \end{array}\right] \preceq I\}$ – it is half of rotated ice-cream cone.

Morale: Be careful! Matrices do not commute, repealing many guesses inspired by our experience with reals!

Miscellaneous

♣ Let us look at two useful SDP-representable sets

♠ Let $A, B \in \mathbf{S}^n_+$ and $a, b \in \mathbf{R}^n$ with $b \in \mathrm{Im}\, B$. Consider the ellipsoid

$$E = \mathcal{E}(A, a) := \{x = Au + a, u^T u \leq 1\}$$

(this is called "image representation;" E is an ellipsoid in $\mathrm{Im}\, A$) along with elliptic cylinder

$$C = \mathcal{C}(B, b) = \{x : \|b - Bx\|_2 \leq 1\}$$

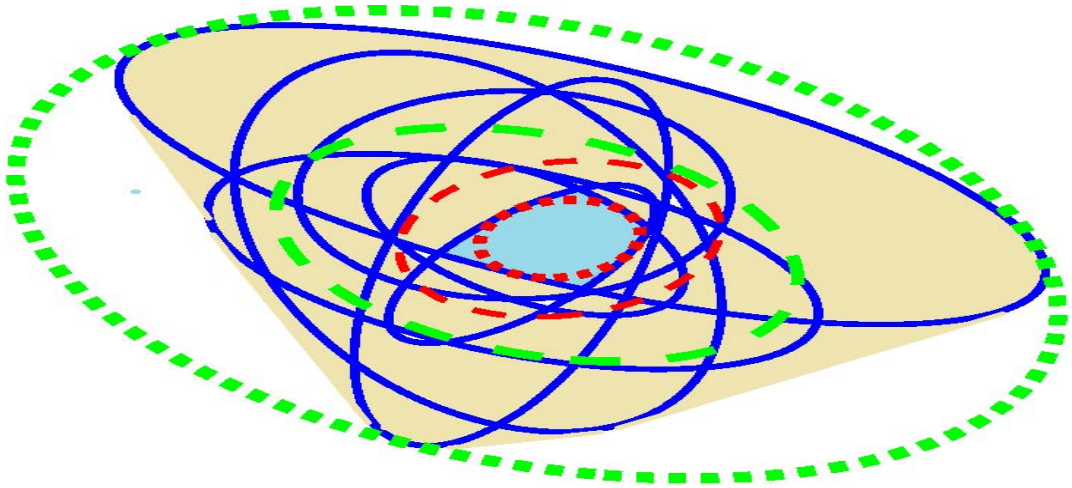
(this is called "inequality representation;" elliptic cylinder is a full-dimensional ellipsoid when $B \succ 0$, otherwise it is either the direct product of $\mathrm{Ker} B$ and an ellipsoid in $\mathrm{Im}\, B$ or is empty.
Question: When $E \subset C$?
Answer [Boyd et al] *This is the case iff there exist $\lambda \in \mathbf{R}$ such that the matrix inequality*

$$\left[\begin{array}{c|c|c} 1-\lambda & & a^T B - b^T \\ \hline & \lambda I & AB \\ \hline Ba - b & BA & I \end{array}\right] \succeq 0 \qquad (*)$$

holds true.
This is an easy consequence of $\mathcal{S}$-Lemma; for proof, see, e.g., section 3.7.3 in [LMCO].

Note: For E fixed, $(*)$ is an LMI in variable λ and in the parameters B, b of C. For C fixed, $(*)$ is an LMI in variable λ and in the parameters A, a of E. Thus, *both the facts that*
— a varying ellipsoid is contained in a fixed elliptic cylinder
— a varying elliptic cylinder contains a fixed ellipsoid
are semidefinite representable!
This fact is instrumental in SDP formulations of various problems related to *extremal ellisoids/elliptic cylinders* containing/contained in given sets, e.g. minimum volume ellipsoid containing the union of a given finite collections of points/ellipsoids ("outer ellipsoidal approximation"), or the maximum volume ellipsoid contained in the intersection of finitely many ellipsoids/elliptic cylinders ("inner ellipsoidal approximation"), see section 3.7 in [LMCO].

Illustration: Inner and outer elliptic approximation



What you see are 7 ellipses (blue) and

- the smallest area ellipse (dotted green) containing the union of the blue ellipses
- the largest area ellipse (dotted red) contained in the intersection of the blue ellipses

Pay attention: twice shrunk outer ellipse (dashed green) is inside the convex hull (yellow) of the union of blue ellipses; twice enlarged inner ellipse (dashed red) contains the intersection (light blue) of the blue ellipses.

This is the 2D case of the following nice result:

Frits John Theorem: *Shrinking by factor n the smallest volume ellipsoid containing n-dimensional solid X* (comapct convex set with nonempty interior), *you get an ellipsoid contained in X. Enlarging by factor n the largest volume ellipsoid contained in X, you get an ellipsoid containing X* (in both cases, we keep the centers of the ellipsoids intact and preserve the directions of the axes).

• The factors are sharp for n-dimensional simplex. When X is centrally symmetric, the factors reduce to $\sqrt{n}$, the new factors being sharp for n-dimensional box.

Note: It is difficult to find the smallest volume ellipsoid containing the intersection of several given ellipsoids, same as it is diffuclt to find the largest volume ellipsoid contained in the convex hull of the union of several given ellipsoids. The diffculties stem from the intractability of the corresponding *analysis* problem – cheking whether a given ellipsoid contains the intersection/is contained in the convex hull of a given family of ellipsoids.

♠ Consider an infinite system of LMI's

$$S(x) - Q^T \Delta^T R(x) - R^T(x) \Delta Q \succeq 0 \ \forall (\Delta, \|\Delta\| \leq 1) \qquad (*)$$

in variables $x \in \mathbf{R}^n$, with the left hand side taking values in some $\mathbf{S}^m$, where $Q \neq 0$, P, Z, $R(x)$, $S(x)$ are matrices of appropriate sizes *with $S(x)$ and $R(x)$ affine in x*, variable matrix Δ is a "perturbation," and $\|\cdot\|$ stands for the spectral norm.

The following fact due to Boyd et al (for proof, see, e.g., section 3.3 in [LMCO]) is a simple consequence of $\mathcal{S}$-Lemma:

The set of feasible solutions to $(*)$ is $\mathfrak{S}$-r: x is feasible for $(*)$ iff it can be extended by a real λ to solve the LMI

$$\left[\begin{array}{c|c} S(x) - \lambda Q^T Q & R^T(x) \\ \hline R(x) & \lambda I \end{array}\right] \succeq 0$$

Note: The fact that the feasible set of $(*)$ admits and explicit $\mathfrak{S}$-r. is useful in various applications, e.g., in synthesis of linear controllers for uncertainty-affected Linear Dynamical Systems.

Concluding remarks

♣ You could get the impression that what pretends to be a "rich list of raw materials for the calculus of well-structured convex sets and functions" is (perhaps aside from polyhedral sets and several $\mathfrak{C}$-r.'s of algebraic functions) rather esoteric; I do not think you frequently hear the word "eigenvalue" in your ISyE classes.

• Note: ISyE is about "soft engineering;" our researchers and students are dealing with quantitative models of extremely complex industrial and societal situations (logistics, supply chain, healthcare, inventory management, multistage decision making,...). These models typically are structurally simple (and where to take knowledge and data allowing to specify a structurally rich model in healthcare?), and difficulties come from potentially huge problems' sizes, and from stochasticity and other forms of data uncertainty (when planning development of electricity generation and distribution, what can you say about future demands for electricity, would-be annual precipitations, climate, wind intensity, etc., etc., on the 10-year time horizon?). As for OR-related research at ISyE, it primarily focuses on design and analysis of algorithms.

• Hard-core engineering is different: here optimization is aimed at systems governed by well understood laws giving rise to structurally rich quantitative models (not always easy for numerical processing). Newton and Coulomb Laws are above Congress and in the Millenia to come will stay the same as they were Millenia ago...

Страна не та уже давно, а скорость звука все та же, что при батюшке-царе
Тимур Шаов, современный русский бард

[The country is not the same anymore, but the speed of sound is still the same as it was under the Tsar-Father – Timur Shaov, contemporary Russian bard]

• As for eigenvalues, you, consciously or unconsciously, meet with them every day when springing on springs, playing your guitar, driving your car, traveling by air, listening to music, speaking by your phone, running your computer, undergoing CT scans, and so on.

• In fact, every one of $\mathfrak{C}/\mathfrak{S}$ representations we have seen gives rise to a wide spectrum of convex (and thus computation-friendly) optimization models in Control, Engineering, Communications, Signal Processing, Medical Imaging, Statistics, etc., etc. List of these models includes, but not reduces to

  • stability analysis of uncertainty-affected dynamical systems (cars, aircrafts,...) and synthesis of controllers for these systems

  • design of mechanical structures (buildings, bridges, aircrafts,...) optimizing their *dynamical stability* (ability to withstand, to the extend possible, earthquakes like the recent one that devastated Turkey) and *static stability* (small deformation under load)

  • design of arrays of antennae capable to send/receive energy along (narrow cones around of) prescribed directions,

  • optimal signal recovery in Communications

  • phase retrieval with applications in X-ray crystallography, transmission electron microscopy, and coherent diffractive imaging

  • optimizing chips aimed at improving the clock speed of computer's CPU

  • design of efficient Signal Processing and Statistical recovery procedures

  • sparsity-oriented image recovery in Medical Imaging and beyond

  • computing extremal ellipsoids, with applications ranging from efficient algorithms for low-dimensional black-box-oriented convex minimization to approximating reachability domains of dynamical systems and stable integration of Ordinary Differential Equations

  • .....

This list has been continuously extending over decades, in significant part due to the pioneering research of Prof. Stephen Boyd from Stanford University; I highly recommend the wonderful book Stephen Boyd and Lieven Vandenberghe, *Convex Optimization* https://stanford.edu/~boyd/cvxbook/bv_cvxbook.pdf. Some of the models can be found in [LMCO] https://www2.isye.gatech.edu/~nemirovs/LMCOLN2024Spring.pdf

МАТЕМАТИКА наука о величинах и количествах; все, что можно выразить цифрою, принадлежит математике. Математика чистая, занимается величинами отвлеченно; Математика прикладная, прилагает первую к делу, к предметам.
Владимир Даль (1801-1872) Толковый словарь живаго великорускаго языка

MATHEMATICS is the science of magnitudes and quantities; everything that can be expressed numerically belongs to mathematics. Pure mathematics, deals with quantities abstractly; applied mathematics, applies the former to business, to objects.
Vladimir Dal (1801-1872) Explanatory dictionary of the living Great Russian language

♠ Illustrations of applying Convex Optimization "to business, to objects" to follow are neither the most advanced nor the most important ones; their "common denominator" is, first, that once upon a time I worked on them, and second, that they allow for visualization.

Illustration: Antenna Design

♣ Consider linear antenna array composed of 24 harmonic oscillators forming an equidistant grid.

• with proper actuation (selecting amplitudes and initial phases of the 24 interfering harmonic oscillations), we can control directional distribution of the energy send by the antenna, e.g., concentrate it in a narrow cone around a prescribed direction:

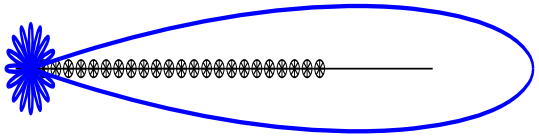
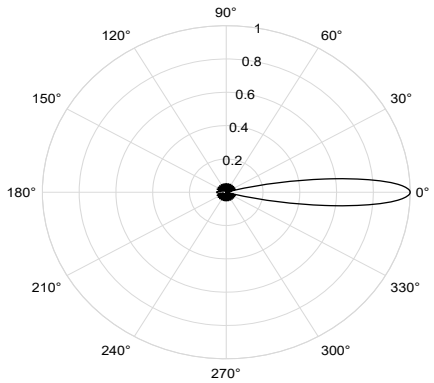


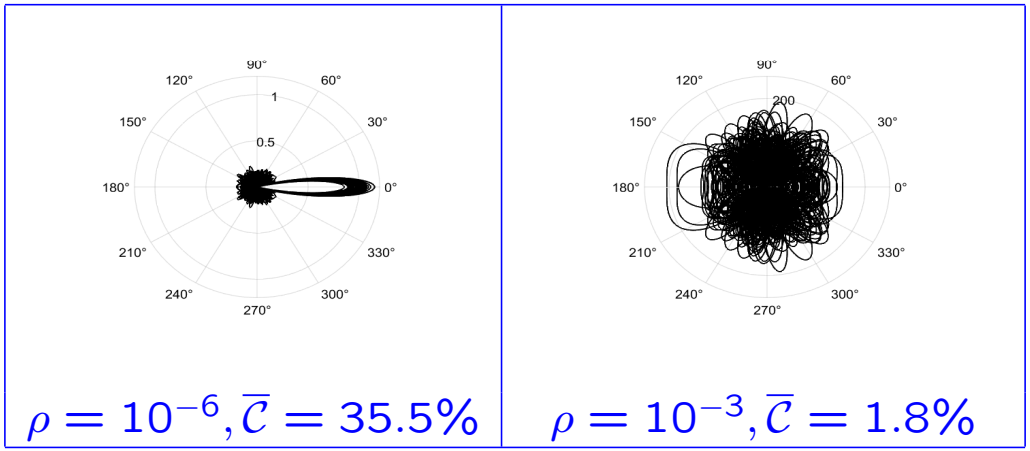
Diagram of 24-element linear antenna array
Directional distribution of energy vs. the angle
between a direction and the direction of the array

♠ Selecting "actuation weights" to maximize energy concentration in a given cone of interest, we end up with the diagram as follows:



Optimal design, energy concentration $\mathcal{C} = 74.8\%$

• **However:** Actuation weights are characteristics of physical devices and therefore cannot be implemented exactly as computed. This is what happens with random implementation errors of small relative magnitude ρ:

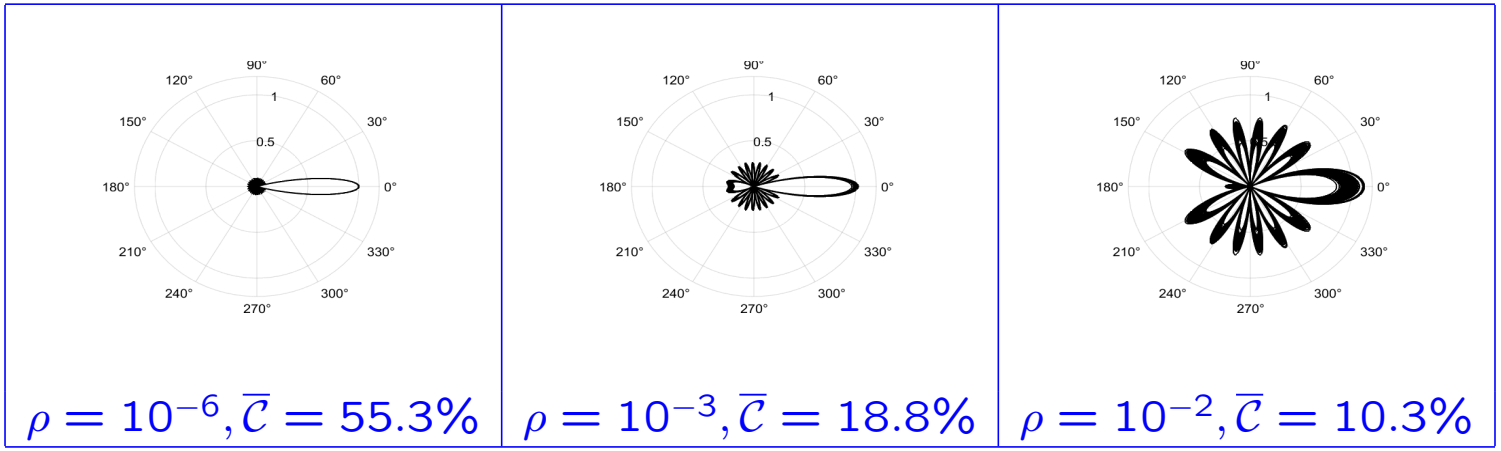


$\rho = 10^{-6}, \overline{\mathcal{C}} = 35.5\%$ | $\rho = 10^{-3}, \overline{\mathcal{C}} = 1.8\%$

Sample of 100 diagrams with implementation errors

$\overline{C}$: average energy concentration

♠ **Robust design** immunizes the solution against implementation errors:



$\rho = 10^{-6}, \overline{\mathcal{C}} = 55.3\%$ | $\rho = 10^{-3}, \overline{\mathcal{C}} = 18.8\%$ | $\rho = 10^{-2}, \overline{\mathcal{C}} = 10.3\%$

Robust design, sample of 100 diagrams with implementation errors

Illustration: Controlling Peak-to-Peak Gain

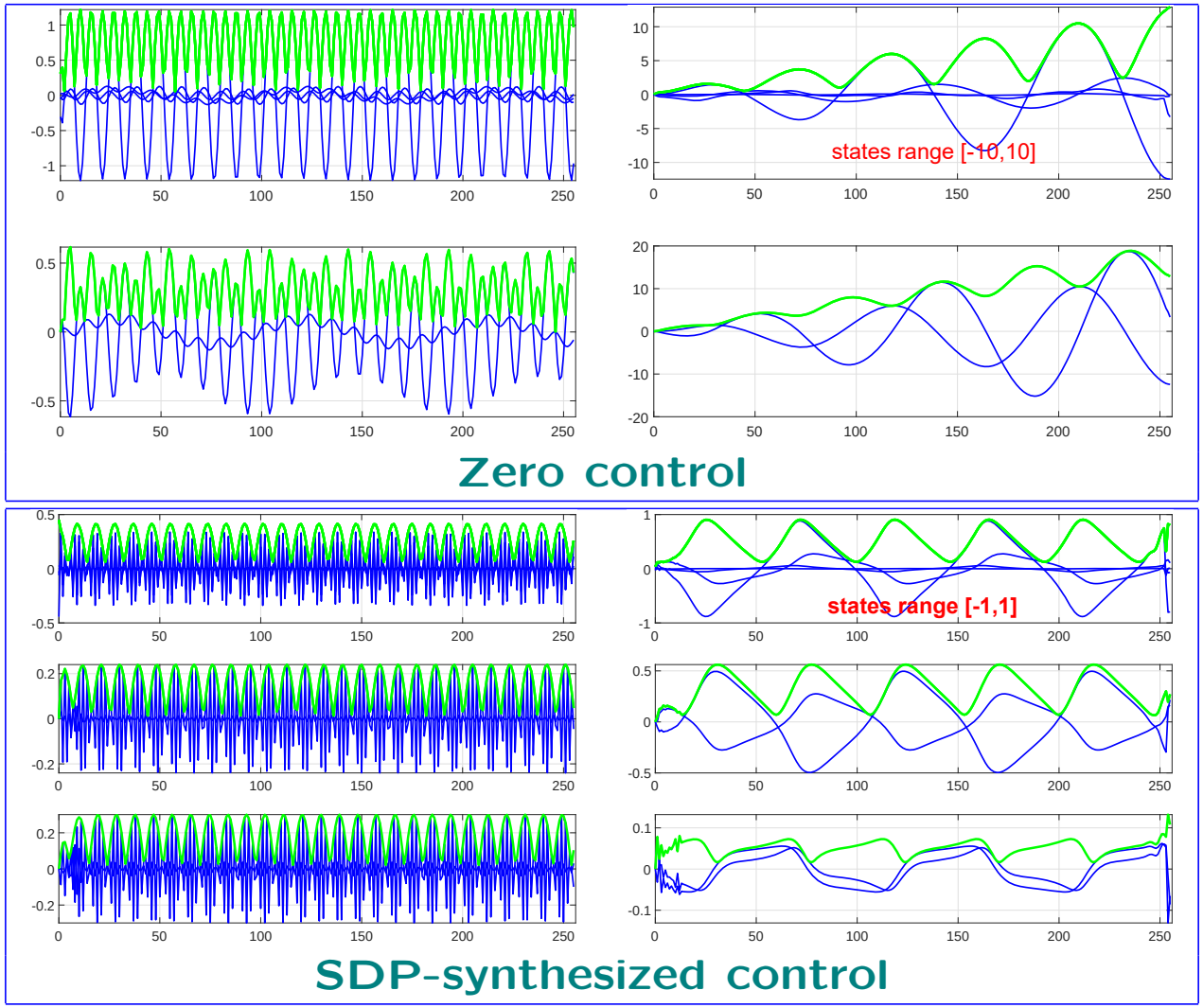
♣ **External disturbances** affecting (discrete time) linear dynamical system on time horizon $1, 2, ..., T$ enforce the consecutive system's states to deviate from their nominal values. This phenomenon is quantified by *peak-to-peak gain* – the worst-case, over the disturbances, ratio of the maximum, over time, of magnitudes of state deviations to the maximum, over time, of magnitudes of the disturbances:

$$\text{Gain} = \sup \frac{\|\text{sequence of states unde zero control}\|_\infty}{\|\text{sequence of distrubances}\|_\infty}$$

♠ Design of linear controllers accounting for peak-to-peak gain is an old and not so easy-to-solve problem in Control.

♠ SDP offers computation-friendly way to handle the peak-to-peak gain.

- Peak-to-peak gain:



In blue, from top to bottom: perturbations in states/outputs/controls (on the synthesized control plots) vs. time. In the left pane: random harmonic oscillation disturbance, in the right pane: "bad disturbance." In green: $\|\cdot\|_2$-norms of states, outputs and controls, respectively

Cruise flight of Boeing 747 (linearized model)
4 states, 2 outputs, 2 controls, 2 disturbances (wind)

Illustration: Stabilizing Linear System with unknown dynamics

♣ Consider a discrete time linear time-invariant dynamical system

$$x_{t+1} = Ax_t + Bu_t + d_t,\ t = 0, 1, 2, ..., \qquad (S)$$

where

- $x_t \in \mathbf{R}^{n_x}$ is system's state,
- $u_t \in \mathbf{R}^{n_u}$ is control,
- $x_t \in \mathbf{R}^{n_x}$ is external disturbance

at time instant t.

♠ In Control, the standard design specification is to ensure *stability:* with identically zero disturbances, the controller must ensure convergence of the states to 0 as $t \to \infty$. A widely used way to ensure stability is to specify a *stabilizing static feedback* $x_t \mapsto u_t = Fx_t$ making the resulting *closed loop system*

$$x_{t+1} = [A + BF]x_t + d_t$$

stable.

$$x_{t+1} = Ax_t + Bu_t + d_t,\ t = 0, 1, 2, ..., \qquad (S)$$

Note: We restrict ourselves (and this is a severe restriction!) to the case when the states are observable "in real time" – we see x_t when generating u_t. In more general settings, what we see when generating u_t is the *output* Cx_t, where C, same as A and B, specify the system. In this case, linear controller to be designed is $u_t = FCx_t$. *In what follows, we consider the case $C = I$ of fully observable states.*

♠ Linear Algebra says a linear dynamical system

$$x_{t+1} = \mathcal{A}x_t \qquad (\mathcal{A})$$

is stable iff the *spectral radius* $\rho(\mathcal{A})$ – the maximum of modulae of the eigenvalues of $\mathcal{A}$ *is* < 1. Equivalent, and more convenient for us, criterion of stability is the existence of *Lyapunov stability certificate* – a positive definite matrix U and $\gamma < 1$ such that

$$\mathcal{A}^T U \mathcal{A} \preceq \gamma U. \qquad (!)$$

Sufficiency of (!) for the stability is clear: as $U \succ 0$, the quantity $\|x\|_U = \sqrt{x^T U x}$ is a norm, and (!) says that in this norm, the mapping $x \mapsto \mathcal{A}x$ is a *contraction:* $\|\mathcal{A}x\|_U \leq \sqrt{\gamma}\|x\|_U$, implying that in the norm $\|\cdot\|_U$, every trajectory of $(\mathcal{A})$ approaches the origin exponentially fast.

$$\{U \succ 0 \ \&\ \mathcal{A}^T U \mathcal{A} \preceq \gamma U\} \Rightarrow [\mathcal{A}]^t x_0 \to 0,\ t \to \infty\ \forall x_0$$

♠ We see that designing a stabilising static feedback F for (S) reduces to finding $U \succ 0$ and $\gamma < 1$ such that

$$[A + BF]^T U [A + BF] \preceq \gamma U.$$

Find $U \succ 0$, $\gamma < 1$, F: $[A+BF]^T U[A+BF] \preceq \gamma U$.

♠ To achieve our goal, observe that we can treat γ as a given real rather than as a variable. Indeed, set $\gamma = 0.9$ and try to find the required U and F; upon success, you are done, otherwise set $\gamma = 0.99$ and repeat your attempt; in the case of failure, try $\gamma = 0.999$, and so on. If you are not in business after 6=10 attempts, forget about stabilizing your system by static linear feedback — for all practical purposes, this is impossible.

♠ With the above remark in mind, our goal becomes to solve the system

$$U \succ 0,\ [A+BF]^T U[A+BF] \preceq \gamma U \qquad (*)$$

in variables $U \in \mathbf{S}^{n_x}$, $F \in \mathbf{R}^{n_u \times n_x}$. We are about to reduce this system to an equivalent system of *Linear* Matrix Inequalities.

• $(*)$ can be rewritten as

$$\left[U^{-1/2}[A+BF]U^{-1/2}\right]^T \left[U^{1/2}[A+BF]U^{-1/2}\right] \leq \gamma I$$

or, which is the same, as

$$\|U^{1/2}[A+BF]U^{-1/2}\| \leq \sqrt{\gamma}$$

where $\|\cdot\|$ is the spectral norm. As the spectral norm remains intact when passing from a matrix to its transpose, the target can be rewritten as

$$\|U^{-1/2}[A+BF]^T U^{1/2}\| \leq \sqrt{\gamma}$$

and therefore as

$$\left[U^{1/2}[A+BF]U^{-1/2}\right]\left[U^{-1/2}[A+BF]^T U^{1/2}\right] \preceq \gamma I$$

that is, as $U^{1/2}[A+BF]U^{-1}[A+BF]^T U^{1/2} \preceq \gamma I$, which is the same as

$$[A+BF]V[A+BF]^T \preceq \gamma V.$$

where $V = U^{-1}$. Note that $V \succ 0$ is exactly the same as $U \succ 0$.

Find $V \succ 0$, F: $[A+BF]V[A+BF]^T \preceq \gamma V$. $(*)$

♠ Our target matrix inequality is homogeneous in $V \Rightarrow$ the restriction $V \succ 0$ can be safely modeled as $V \succeq I$. Passing from F to new variable $G = FV$, $(*)$ becomes the system of matrix inequalities

$$V \succeq I \ \& \ AVA^T + BGA^T + AG^TB^T + [BG]V^{-1}[BG]^T \preceq \gamma V.$$

Introducing "matrix upper bound" W onto $[BG]V^{-1}[BG]^T$:

$$W \succeq [BG]V^{-1}[BG]^T \Leftrightarrow \left[\begin{array}{c|c} W & BG \\ \hline [BG]^T & V \end{array}\right] \succeq 0$$

(equivalence is due to the Schur Complement Lemma), we end up with a system of Linear Matrix Inequalities

$$V \succeq I \ \& \ AVA^T + BGA^T + AG^TB^T + W \preceq \gamma V \ \& \ \left[\begin{array}{c|c} W & BG \\ \hline [BG]^T & V \end{array}\right] \succeq 0;$$

in matrix variables V, G, W; The matrix inequality of interest is solvable iff the resulting system is so, and a solution V, G, W to the latter system induces the solution

$$V,\ F = GV^{-1}$$

to $(*)$. The associated stability certificate is $U = V^{-1}$.

$$x_{t+1} = Ax_t + Bu_t + d_t,\ t = 0, 1, 2, ..., \qquad (S)$$

♣ The above story is just a preamble to the question which we actually want to address: *How to stabilize (S) in the case when we know B, bur do not know A ?*
This is one of the basic questions in Control, and there are various answers to it, depending primarily on what we assume about the disturbances. The characteristic properties of "answer" I am about to present are:

- my answer is absolutely "non-scientific" – I am not going to prove anything (and even did not try to do it - I think that without restrictive assumptions on disturbances, like their iid stochastic nature with positive definite covariance matrix, to justify the approach formally is just impossible)
- the underlying rationale seems to be reasnable
- the resulting control policy is computation-friendly and thus is implementable, and, most importantly, *it seems to work.*

- As an excuse for me giving up science, here is an appropriate citation:
  "Shall I refuse my dinner because I do not fully understand the process of digestion?"
  Oliver Heaviside (1850-1925), British physicist and mathematician, in relation to invented by him *operational calculus* which worked, in spite of years of absence of rigorous justification.

♠ **Main assumption:** *Disturbances are uncertain-but-bounded: for some known in advance $\delta \geq 0$ it holds*

$$\|d_t\|_\infty \leq \delta,\ t = 0, 1, ...$$

Besides this, we assume that we have a priori bounds $\underline{A}$, $\overline{A}$ on the entries of A:

$$\underline{A}_{ij} \leq A_{ij} \leq \overline{A}_{ij} \ \forall i, j$$

These bounds can be loose $\Rightarrow$ the assumption is not too restrictive. On the other hand, it allows to take into account some a priori information on A, e.g., the one that (S) comes from finite-difference equation $z_{t+1} = \sum_{\tau=0}^{n_x-1} \alpha_\tau z_{t-\tau} + u_t + \Delta_t$, so that (S) is the system

$$\underbrace{\begin{bmatrix} z_{t-n_x+2} \\ \vdots \\ \vdots \\ z_{t+1} \end{bmatrix}}_{x_{t+1}} = \left[\begin{array}{c|c|c|c|c} & 1 & & & \\ \hline & & 1 & & \\ \hline & & & \ddots & \\ \hline & & & & 1 \\ \hline \alpha_{n_x-1} & \alpha_{n_x-2} & \alpha_{n_x-3} & \cdots & \alpha_0 \end{array}\right] \begin{bmatrix} z_{t-n_x+1} \\ \vdots \\ \vdots \\ z_t \end{bmatrix} + \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} u_t + \underbrace{\begin{bmatrix} 0 \\ \vdots \\ 0 \\ \Delta_t \end{bmatrix}}_{d_t}$$

$$x_{t+1} = Ax_t + Bu_t + d_t,\ \|d_t\|_\infty \le \delta,\ t = 0, 1, 2, ..., \qquad (S)$$

♠ **Main observation:** Assume that we have somehow generated controls at instants $0, 1, ..., t-1$ and have observed $x_0, x_1, ..., x_t$. At time t, when u_t should be generated, we know something about the matrix A, namely, that it satisfies the system of linear inequalities

$$\underline{A}_{ij} \le A_{ij} \le \overline{A}_{ij}\ \forall i, j\ \&\ \|x_{\tau+1} - Ax_\tau - Bu_\tau\|_\infty \le \delta,\ 0 \le \tau \le t-1 \qquad (P_t)$$

♠ There are at least two ways to utilize this information when generating u_t:

A. Find the best possible upper entrywise bounds U^{t+1} and lower entrywise bounds L^{t+1} on $\bar{x}_{t+1} := Ax_t$:

$$\begin{array}{rcl} U_i^{t+1} & = & \max_A \left\{\sum_j A_{ij}[x_t]_j :\ A \text{ satisfies } (P_t)\right\} \\ L_i^{t+1} & = & \min_A \left\{\sum_j A_{ij}[x_t]_j :\ A \text{ satisfies } (P_t)\right\} \end{array}$$

so that upper and lower entrywise bounds on $x_{t+1} - d_t$ are $U^{t+1} + Bu_t$ and $L^{t+1} + Bu_t$, and select u_t minimizing, say,

$$\max\left[\|U^{t+1} + Bu_t\|_p, \|L^{t+1} + Bu_t\|_p\right]$$

B. Find somehow the "most central" point $\overline{A}_t$ in the solution set $\mathcal{A}_t$ of (P_t), e.g., specify $\bar{A}$ as the *Tschebyshev center* of $\mathcal{A}_t$ – as the A-component of the solution to the LP problem

$$\max_{s,A}\{s : [A_{ij} + s]_{i,j} \in \mathcal{A}_t,\ [A_{ij} - s]_{i,j} \in \mathcal{A}_t\}$$

After $\overline{A}_t$ is found, try to find the feedback F_t stabilizing the system obtained from (S) by replacing unknown matrix A with its "estimate" $\overline{A}_t$. Upon success, set $u_t = F_t x_t$, otherwise specify u_t according to Recipe **A**.

The rationale behind the proposed control policy is extremely simple. The larger are the states, the more information on A is stored in the system of constraints (P_t), and, consequently, the better is the estimate $\overline{A}_t$ of A, and when it is good enough, our control is as it would be, were we knowing A in advance. This phenomenon is a "common denominator" of numerous identification problems in linear models with additive noise/disturbance: were identification of A our only goal, the best way to solve it would be generating at random huge controls ("flying aircraft to Moon"). Of course, in reality you hardly would like to fly to Moon and back to learn the aircraft's dynamics when flying to Moon and using your knowledge to control the aircraft after coming back. The policy we have outlined is more natural: *do not care to identify A, act as if the current estimate of the matrix were precise; if you are wrong, "the nature" hopefully will take care of improving your estimate well before visiting Moon.*

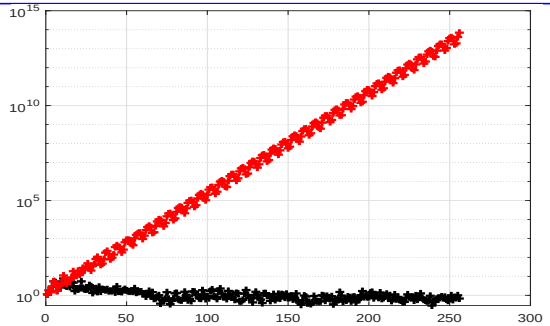
How It Works

♠ I am about to present results of several numerical experiments. In these experiments,

• A was generated as an unstable (with spectral radius slightly greater than 1) matrix stabilizable by static feedback, with the best achievable with such a feedback spectral radius of the "closed loop" matrix $A+BF$ close, but not too close, to 1 (I used the value 0.95, corresponding to $\gamma = 0.95^2$),

• I used $\delta = 0.1$, and there were two modes for generating d_t: the "extreme," where d_t's were generated as random vectors with entries $\pm\delta$, with probability θ for a particular entry to be equal to δ, and "modulated," where the entries in the vectors produced by "extreme" generation were further multiplied by random factors uniformly distributed on $[0,1]$. In both cases, the entries of disturbances were i.i.d. samples drawn from the distribution just described.

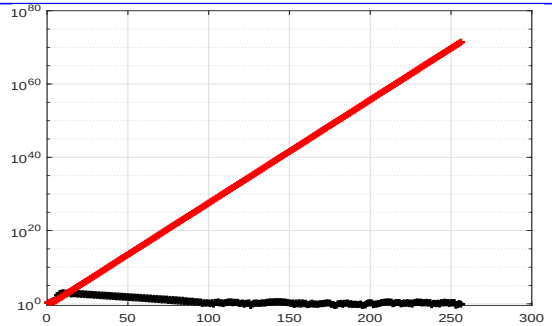
♠ **The results** of my experiments are as follows (T: time horizon; $\overline{A}(t)$: recovery of A at step t):



$n_x = 7, n_u = 1$

$t = 1$	7.745/5.048
$t = 32$	2.955/1.657
$t = 64$	2.295/1.402
$t = 128$	1.684/1.301
$t = 256$	0.677/0.442

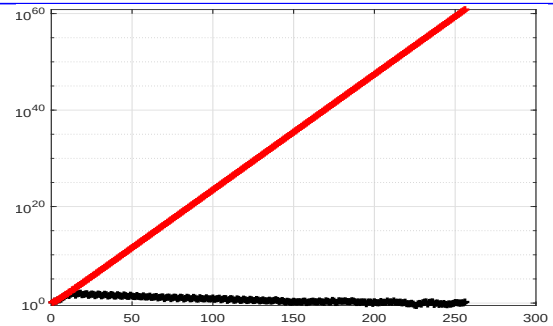
$\max\{\|A - \overline{A}(t)\| : t \geq 32\} = 0.212$
$\max\{\|A - \overline{A}(t)\| : t \geq 128\} = 0.092$
"modulated" generation



$n_x = 5, n_u = 2$

$t = 1$	1136/24.21
$t = 32$	308.7/4.293
$t = 64$	3.865/4.293
$t = 128$	4.351/4.293
$t = 256$	0.824/1.068

$\max\{\|A - \overline{A}(t)\| : t \geq 32\} = 0.0084$
$\max\{\|A - \overline{A}(t)\| : t \geq 128\} = 0.0073$
"modulated" generation



$n_x = 7, n_u = 2$

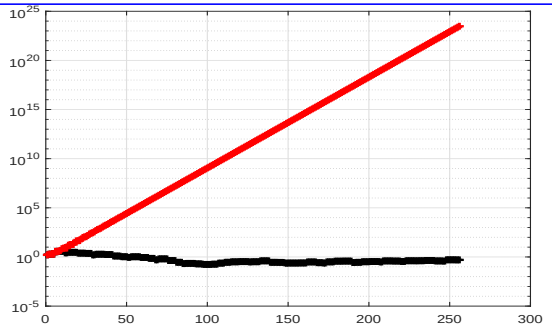
$t = 1$	1.140/5.623
$t = 32$	6.871/3.905
$t = 64$	2.353/3.905
$t = 128$	5.190/2.691
$t = 256$	1.669/1.994

$\max\{\|A - \overline{A}(t)\| : t \geq 32\} = 2.1\text{e-}11$
$\max\{\|A - \overline{A}(t)\| : t \geq 128\} = 4.9\text{e-}13$
"extreme" generation

Magnitudes of states vs time, $T = 256$ **Red:** zero control, **Back:** proposed control policy (pay attention to the log-scale along the y-axis)

In the tables: $\max\{\|x_s\|_\infty : t \leq s \leq 256\}$ for the proposed control policy (blue) and similar quantity (red) for the "ideal" feedback control – one you would use were A known in advance; both controls are used on the same sequence of disturbances.

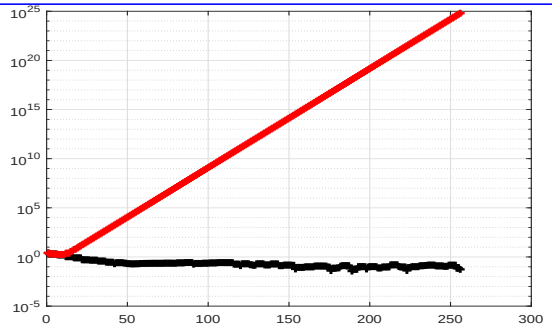
♠ This is how our policy works when the dynamical system stems from a finite-difference equation:



$n_x = 7, n_u = 1$

$t = 1$	3.738/1.527
$t = 32$	1.888/0.330
$t = 64$	0.717/0.330
$t = 128$	0.513/0.330
$t = 256$	0.504/0.330

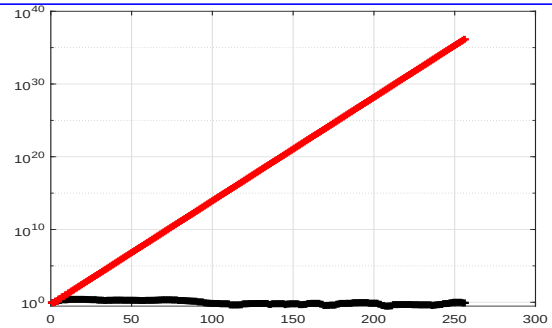
$\max\{\|A - \overline{A}(t)\| : t \geq 32\} = 0.024$
$\max\{\|A - \overline{A}(t)\| : t \geq 128\} = 0.018$
"modulated" generation



$n_x = 5, n_u = 1$

$t = 1$	2.564/1.735
$t = 32$	0.383/0.216
$t = 64$	0.270/0.186
$t = 128$	0.291/0.156
$t = 256$	0.055/0.083

$\max\{\|A - \overline{A}(t)\| : t \geq 32\} = 0.013$
$\max\{\|A - \overline{A}(t)\| : t \geq 128\} = 0.010$
"modulated" generation



$n_x = 7, n_u = 2$

$t = 1$	2.492/1.174
$t = 32$	2.162/1.174
$t = 64$	2.162/1.174
$t = 128$	0.985/0.829
$t = 256$	0.818/0.619

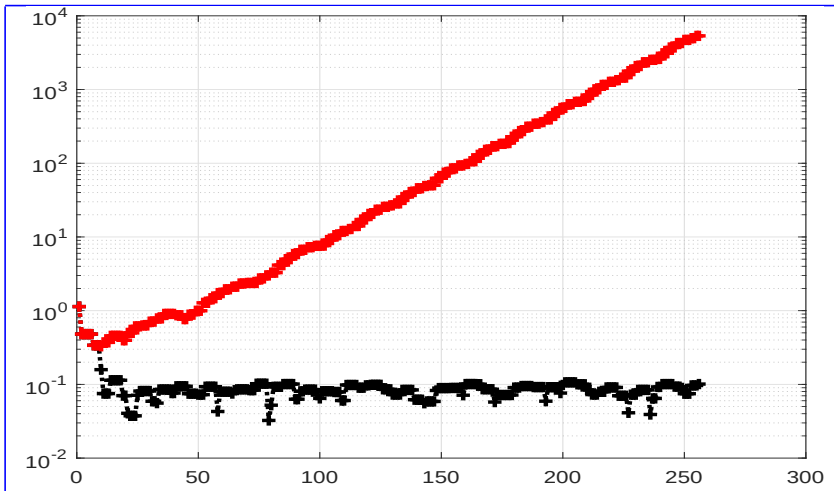
$\max\{\|A - \overline{A}(t)\| : t \geq 32\} =$ 2.65e-11
$\max\{\|A - \overline{A}(t)\| : t \geq 128\} =$ 1.51e-12
"extreme" generation

Magnitudes of states vs time, $T = 256$ **Red:** zero control, **Black:** proposed control policy (pay attention to the log-scale along the y-axis)
In the tables: $\max\{\|x_s\|_\infty : t \leq s \leq 256\}$ for the proposed control policy (blue) and similar quantity (red) for the "ideal" feedback control – one you would use were A known in advance; both controls are used on the same sequence of disturbances.

$$x_{t+1} = Ax_t + Bu_t + d_t,\ \|d_t\|_\infty \le \delta,\ t = 0, 1, 2, ..., \qquad (S)$$

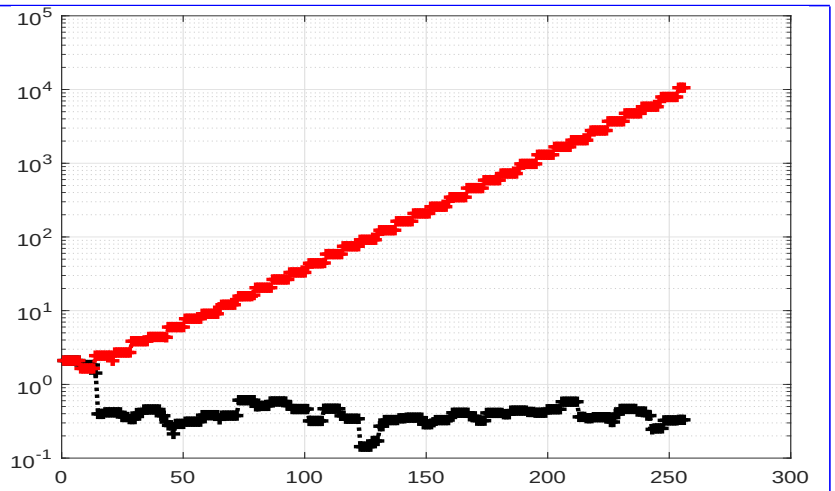
♣ In fact, the outlined control policy can be easily extended to the case where *both* the open loop matrix A and the actuation matrix B are unknown. However, now the policy

- does *not* work when A is a general type matrix
- still works when the dynamical system stems from a finite-difference equation ($\overline{B}(t)$: recovery of B at time t):



$t = 1$	1.136/1.174
$t = 64$	0.107/0.299
$t = 128$	0.107/0.920
$t = 256$	0.101/0.127

$\max\{\|A - \overline{A}(t)\| : t \ge 32\} = 0.351$
$\max\{\|B - \overline{B}(t)\| : t \ge 32\} = 0.154$



$t = 1$	2.103/0.942
$t = 64$	0.611/0.440
$t = 128$	0.414/0.414
$t = 256$	0.227/0.227

$\max\{\|A - \overline{A}(t)\| : t \ge 32\} = 9.7\text{e-}12$
$\max\{\|B - \overline{B}(t)\| : t \ge 32\} = 2.1\text{e-}12$

Magnitudes of states vs time, $T = 256$ **Red:** zero control, **Black:** proposed control policy (pay attention to the log-scale along the y-axis)
In the tables: $\max\{\|x_s\|_\infty : t \le s \le 256\}$ for the proposed control policy (blue) and similar quantity (red) for the "ideal" feedback control – one you would use were A known in advance; both controls are used on the same sequence of disturbances.

$$x_{t+1} = Ax_t + Bu_t + d_t,\ \|d_t\|_\infty \le \delta,\ t = 0, 1, 2, ..., \qquad (S)$$

Strange phenomena: Numerical experimentation raises several questions:

♠ Why the policy in question works when B is known, still works when B is unknown and (S) stems from a finite-difference equation, and does *not* work when B is unknown and A is a general-type matrix?

♠ Recall that with the policy on question, the control at time t is generated according to option **B** ("step B"), provided it is available (i.e., certain SDP program is feasible), otherwise option **A** ("step A") is used. Experiments say that

- when B is known, all steps, except for ≈ 5 at the very beginning, are B-steps
- when B is unknown (and (S) stems from a finite-difference equation – otherwise our policy does not work), all steps, except for ≈ 5 at the very beginning, are A-steps
- when B is known and A is a general type matrix, forbidding to use the option **B** crushes the policy.

Why all this?

♣ If you have nothing better to do, you could think how to explain these phenomena. I have no idea why they take place.

Morale: `Меньше знаешь - крепче спишь [the less you know, the better you sleep]` Detailed information is not a must for happy living. You may succeed not knowing much! Except you are a graduate student.

More seriously: My empirical observation

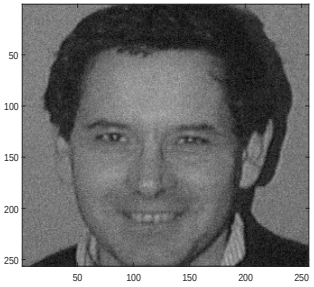
> *The proposed policy works reasonably well when B is known, same as when B is unknown and the system stems from a finite-difference equation.*

is supported, *without a single exception*, by few tens of simulations.

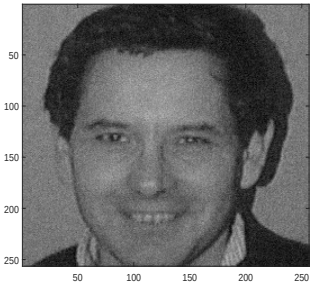
However: The "common denominator" in all these simulations is the way how the disturbances were generated. I have strong doubts that the above empirical observations are applicable to observations of essentially different nature.

Note: "Unreliability" of empirical conclusions based on "passive learning" seems to be unavoidable. I prefer not to live in a house built by a whatever successful practitioner who have never heard about earthquakes...

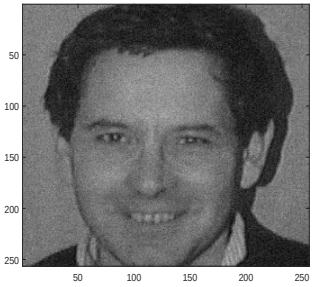
Illustration: Change point detection



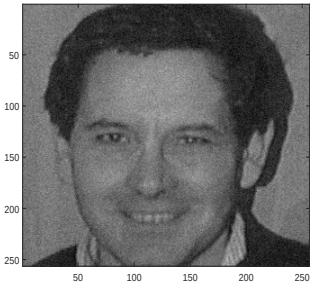
1



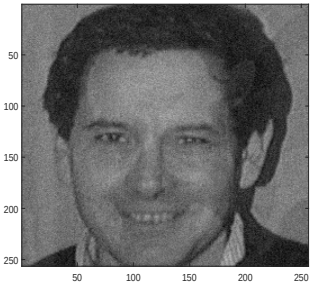
2



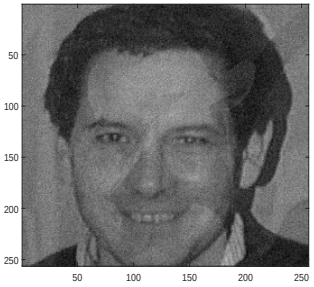
3



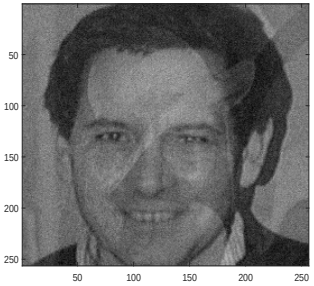
4



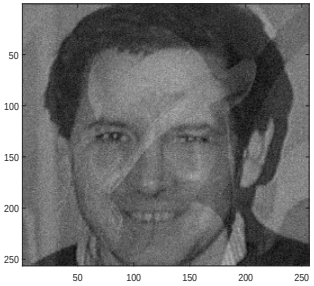
5



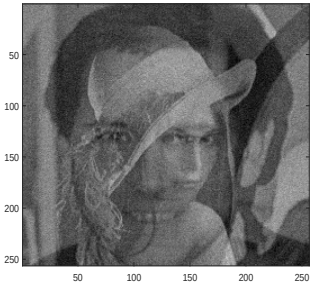
6



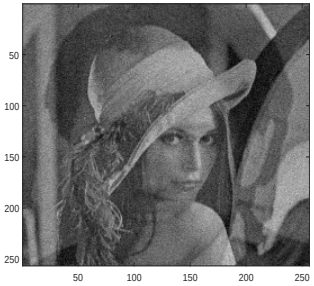
7



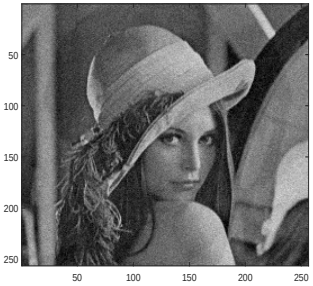
8



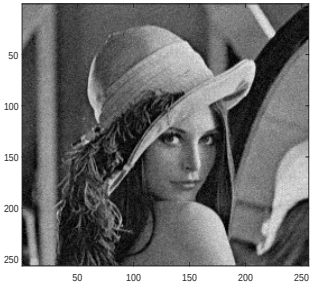
15



20



28



34

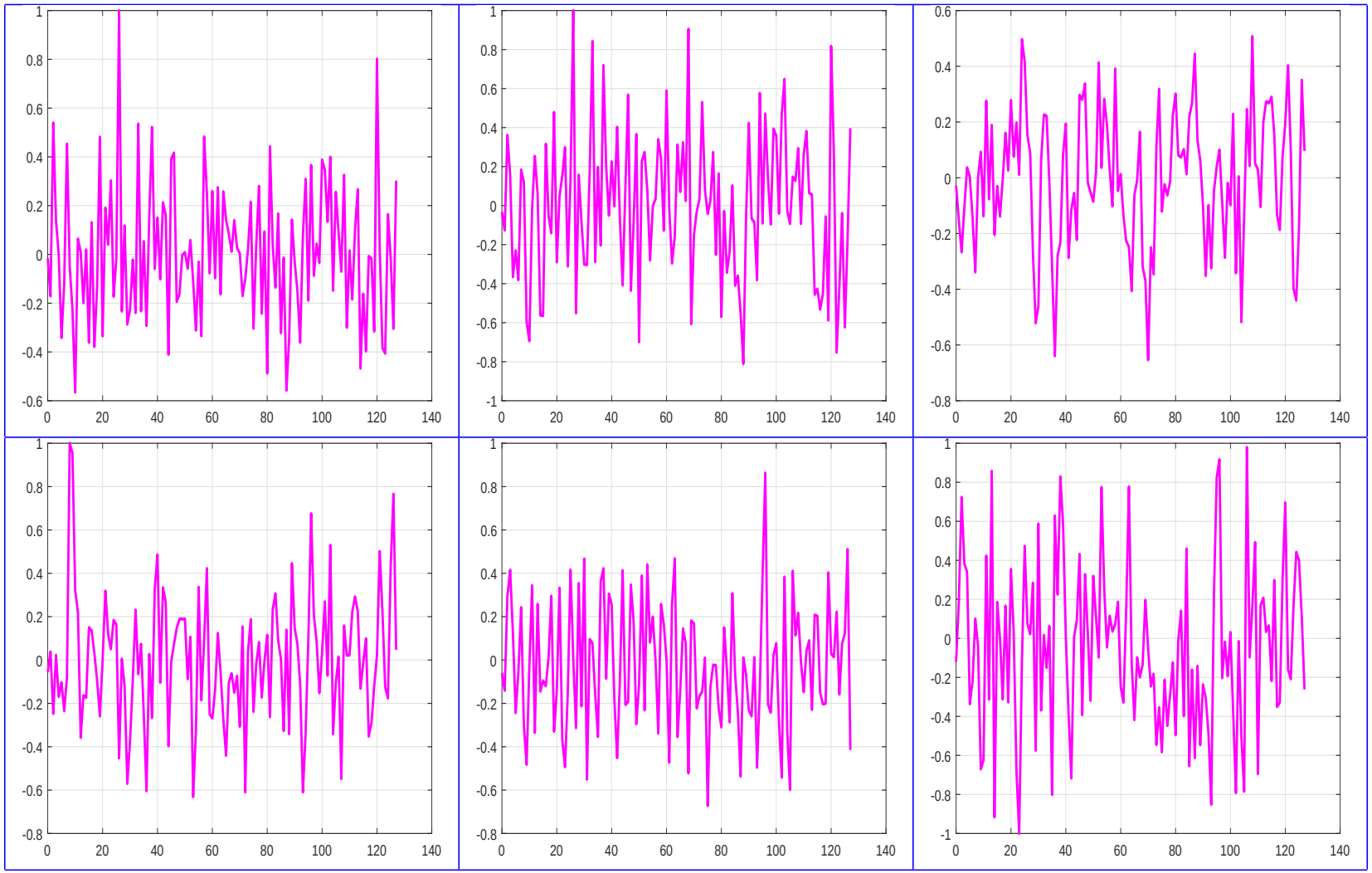
Frames from a noisy "movie"

Question: When the picture starts to change?

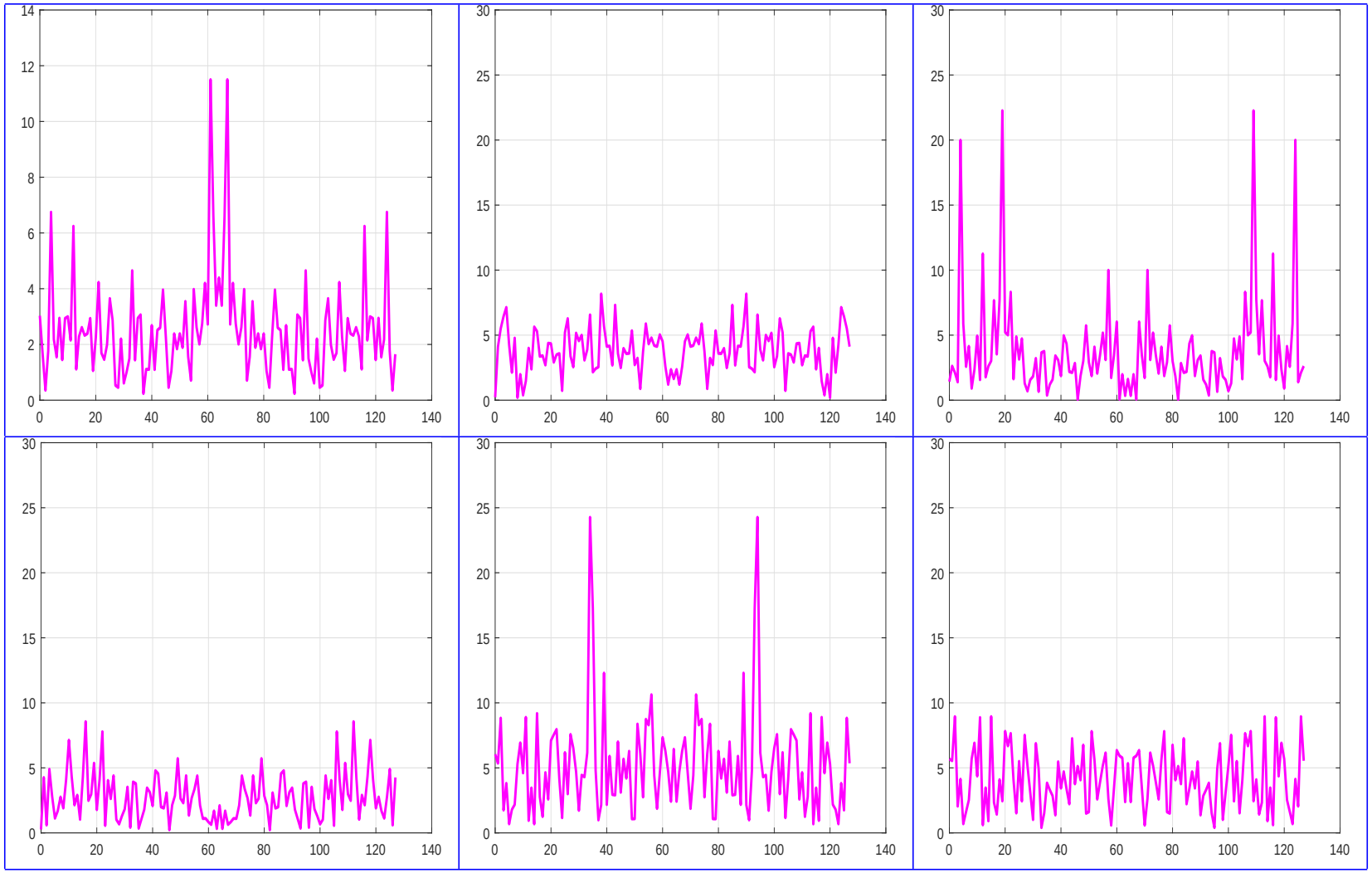
Illustration: Detecting presence of signal in time series

♣ Some of the signals below are pure Gaussian noise, some are the sums of this noise and 5 harmonic oscillations of unknown frequencies and amplitudes.

• Can you decide who is who?



♠ This is how the previous signals look in the frequency domain:



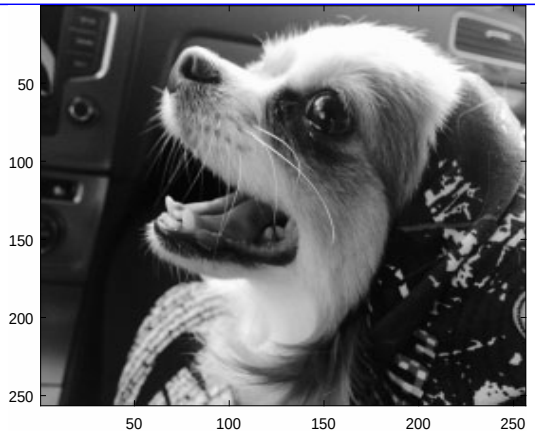
♠ **A provably good test:** *Claim that signal is present if the maximum magnitude of the Fast Fourier Transform of observation is above a properly selected level* (depending solely on the noise's intensity, the length of the time series, and the required false alarm probability).

Illustration: Recovery in Generalized Linear Model

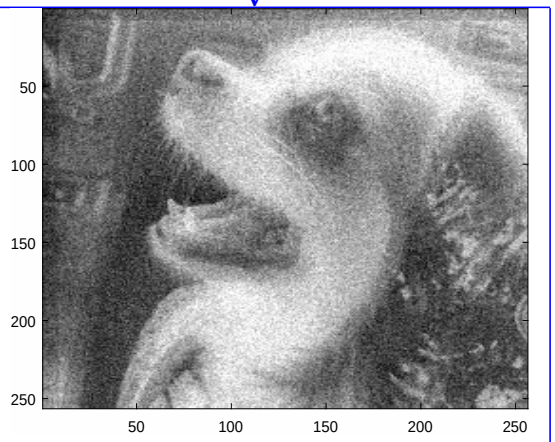
♣ We want to recover 2D image x from noisy observation y of *nonlinear* transformation of x:

$$y = [\varkappa \star x]^{1/2} + \sigma\xi$$

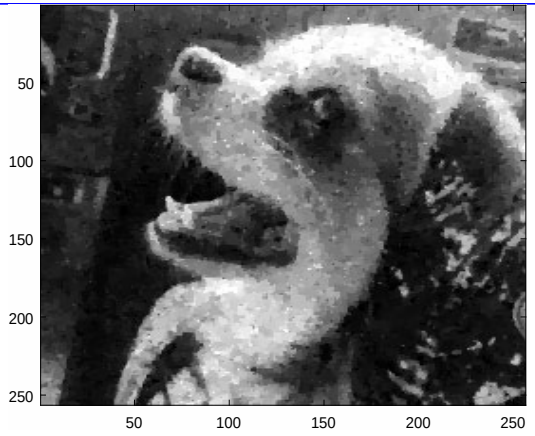
$\varkappa$:	known 2D kernel, $\|\varkappa_1\|_1 = 1$	$\star$:	2D convolution
x:	2D image to be recovered	$[\cdot]^{1/2}$:	entrywise square root
ξ:	white Gaussian noise	σ:	$1.2 \approx 0.075\sqrt{\|x\|_\infty}$



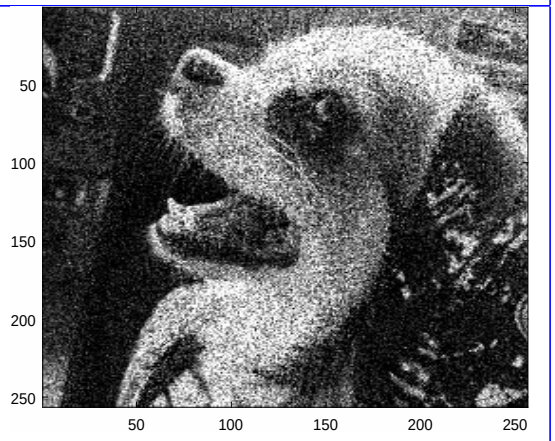
True image



Observation y



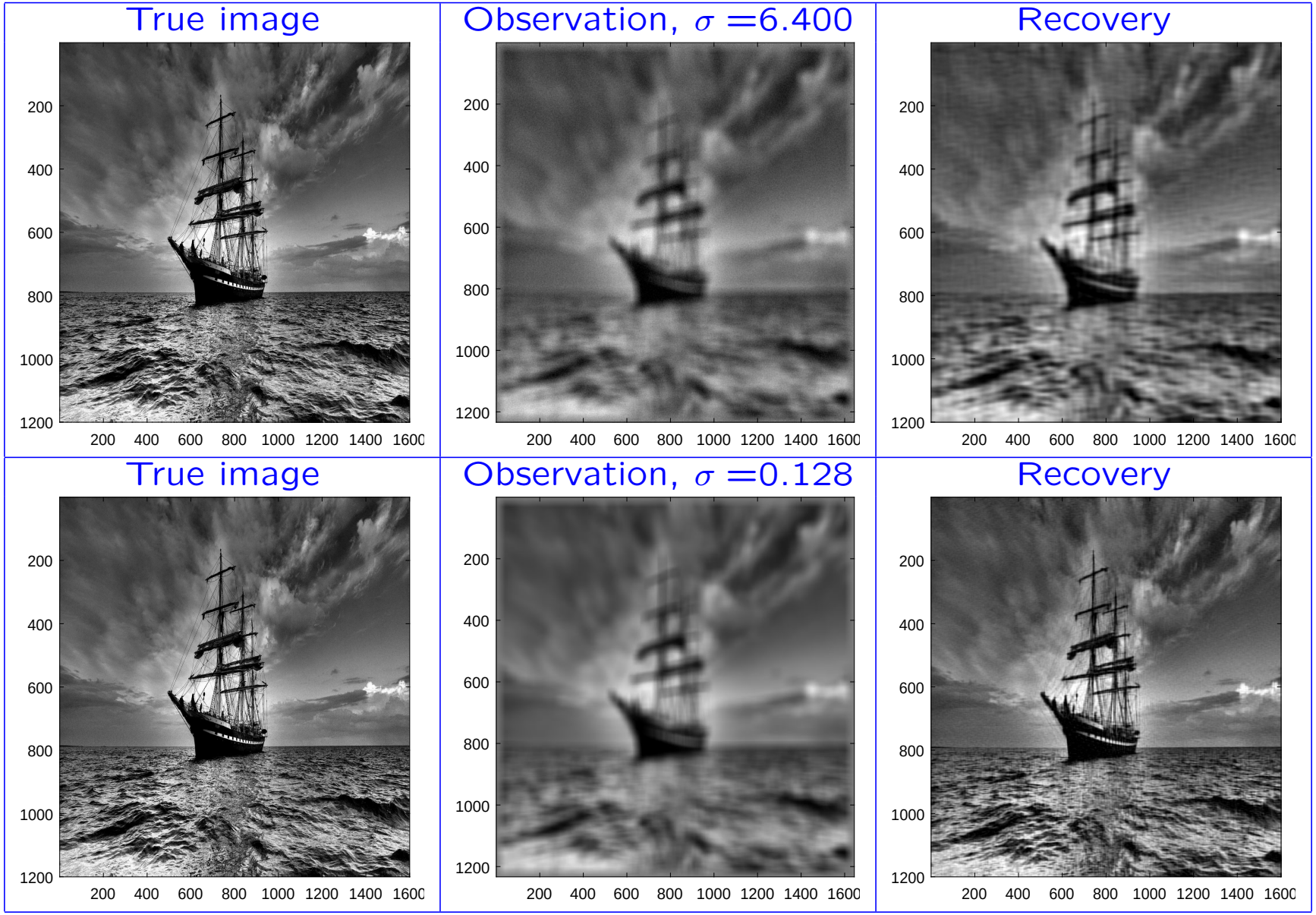
SAA recovery
$\mathcal{X}$: nonnegative part of Total Variation ball



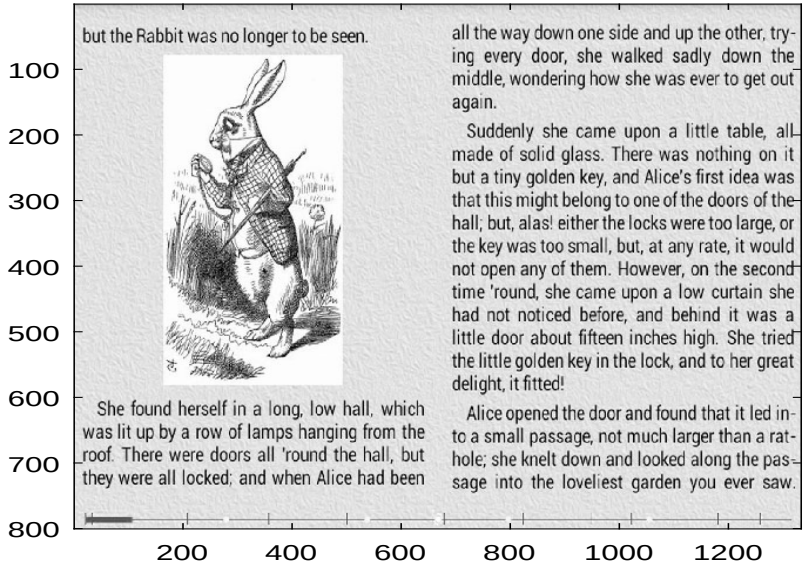
SAA recovery
$\mathcal{X}$: nonnegative orthant

Illustration: Image reconstruction from blurred noisy observations

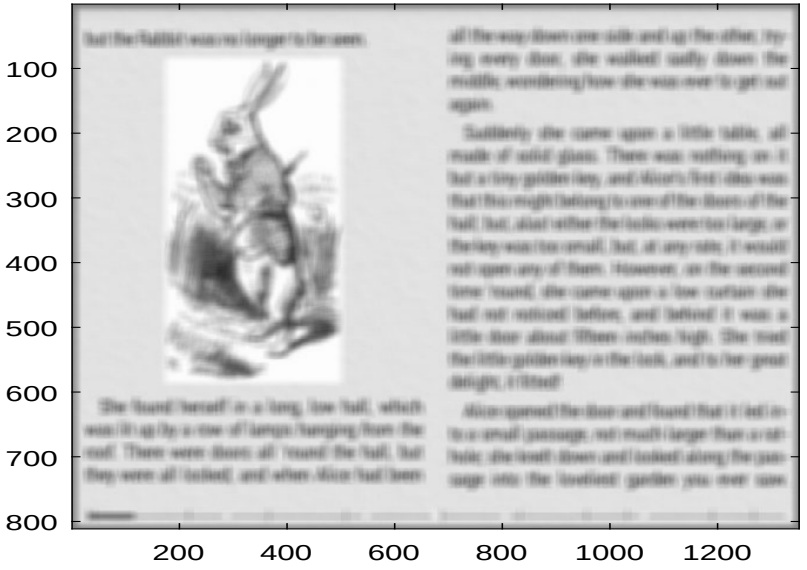
♣ We want to recover 2D image from noisy observation of its 2D convolution with known "blurring kernel." This amounts to recovering a large vector x_* from its noisy observation $y = Ax_* + \sigma\xi$ (A is known sensing matrix, ξ is white Gaussian noise - zero mean, unit covariance).


True image
Observation, σ =6.400
Recovery
True image
Observation, σ =0.128
Recovery
200
400
600
800
1000
1200
1400
1600

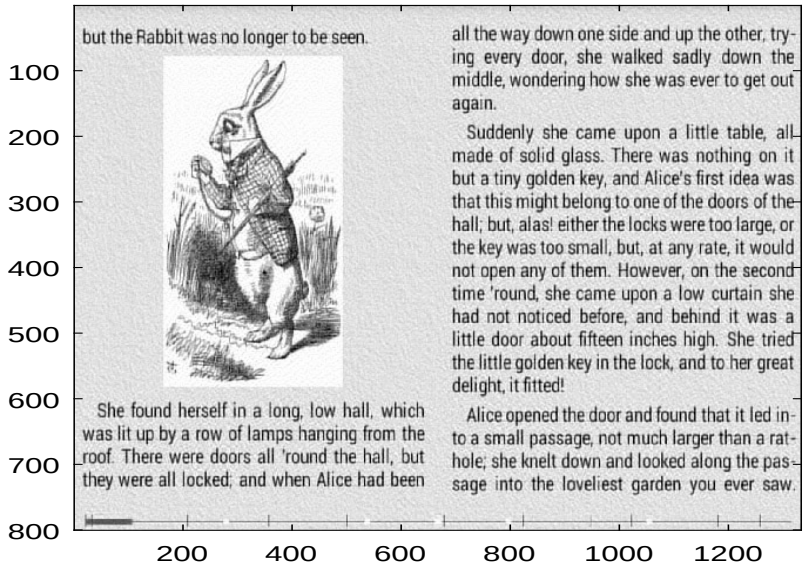
True image



Observation, $\sigma = 0.128$



Recovery



true image

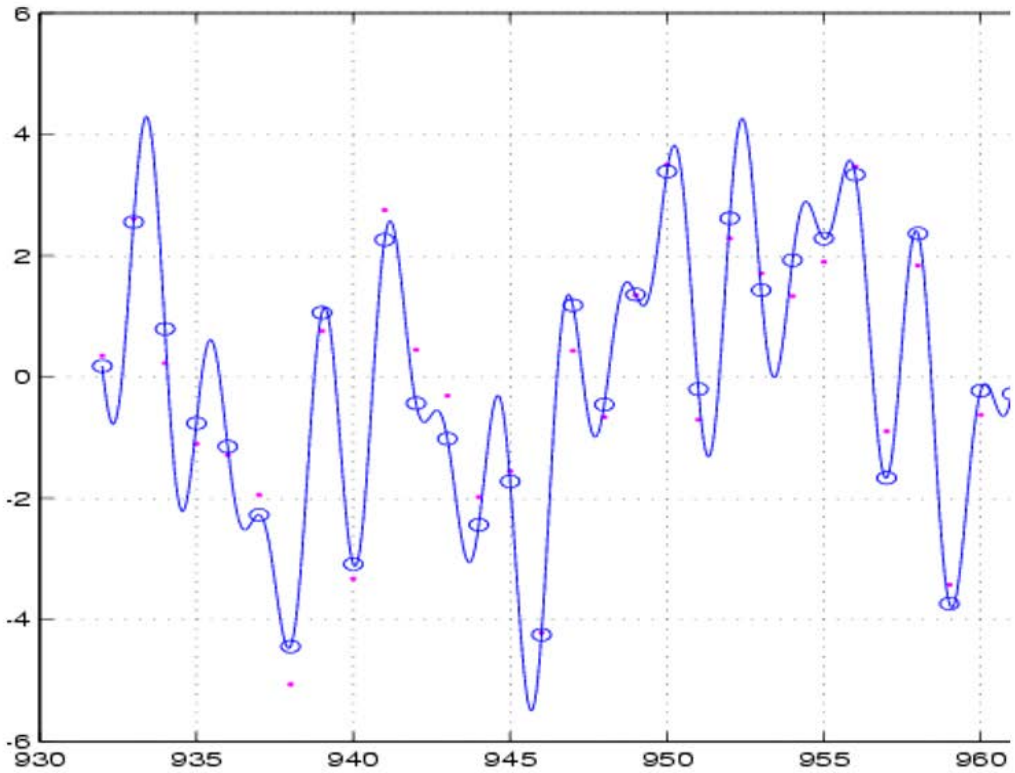
observation

recovery

Illustration: Predicting sparse sum of harmonic oscillations observed in noise

♣ Given noisy observations of a sparse sum of harmonic oscillations with unknown frequencies and amplitudes, we want to predict the future value of the sum.

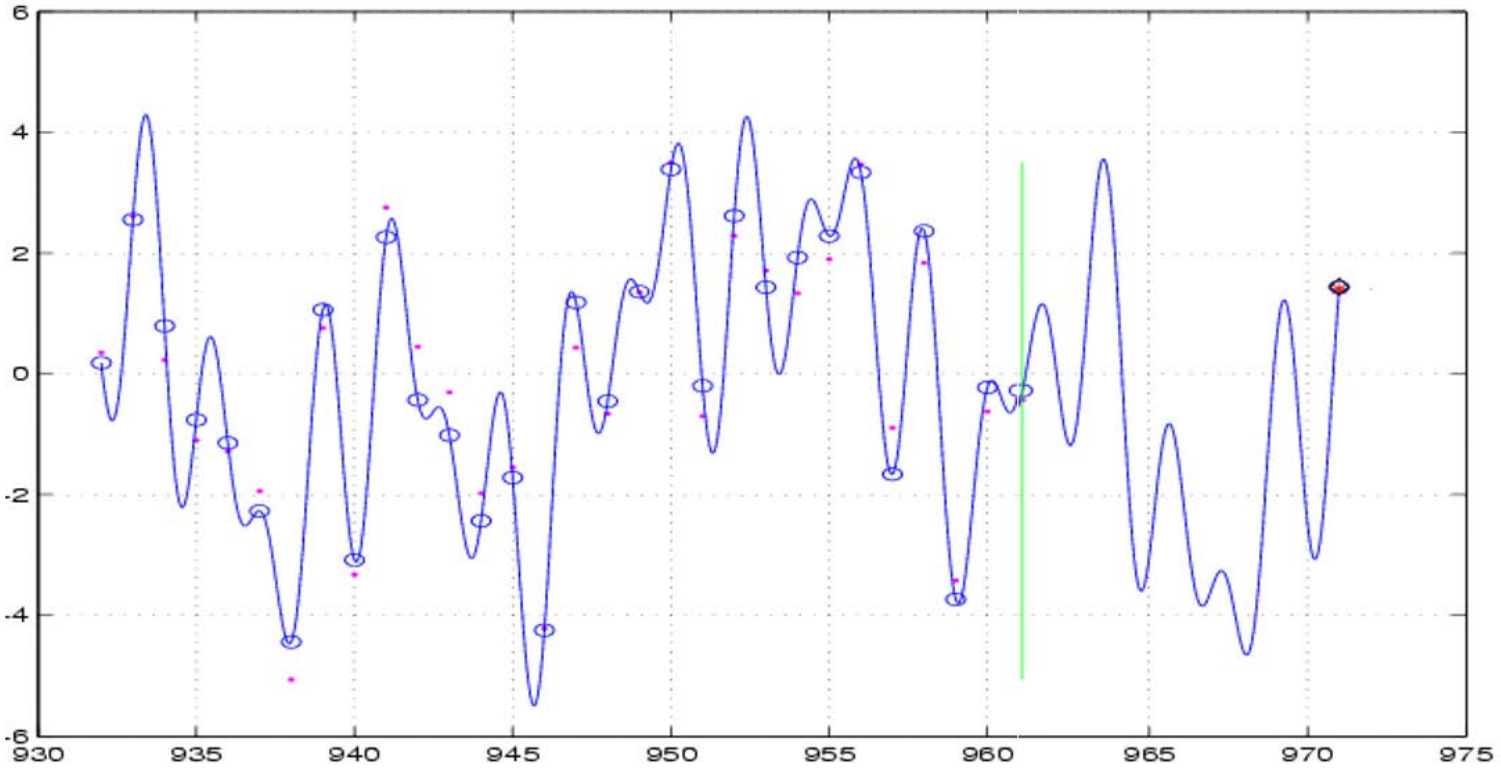
Signal: 5 harmonic components, minimal wavelength 1.9, $\sigma = 0.50$



Sample 10-step forecast based on 512 observations: **up or down?**

o: true signal ▪: observations

Signal: 5 harmonic components, minimal wavelength 1.9, $\sigma = 0.50$

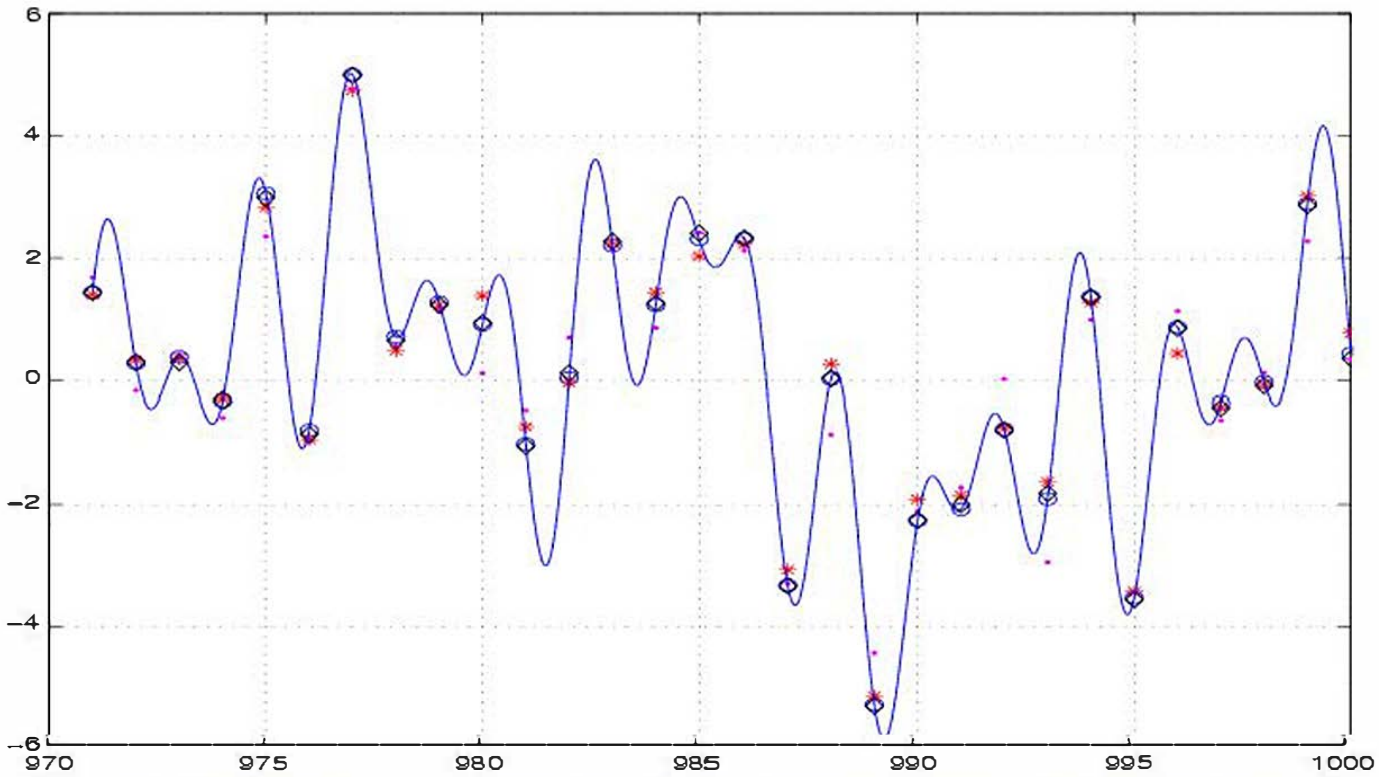


Sample 10-step forecast based on 512 observations: up!

o: true signal ▪: observations $*$: forecast

Note: Recovering routine has no idea what are the frequencies and the amplitudes of oscillations; in fact, *they cannot be recovered at all*, but a provably consistent forecast *is* nevertheless possible!

Signal: 5 harmonic components, minimal wavelength 1.9, $\sigma = 0.50$



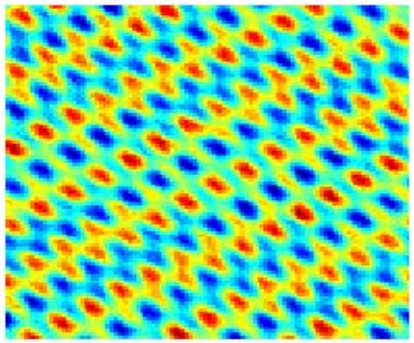
30 subsequent 10-step forecasts based on 512 observations

o: true signal ▪: observations ∗: forecasts

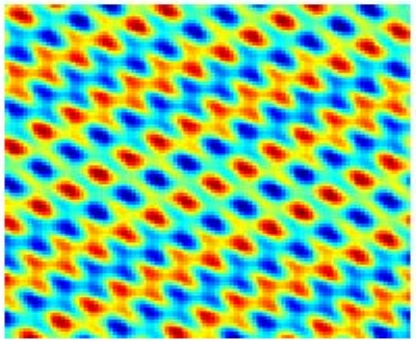
Illustration: Denoising sparse sum of high-frequency harmonics

♠ Given noisy observations of a mixture of 2D harmonic waves with unknown amplitudes, wavelengths and directions of wavefronts, we want to denoise the mixture.

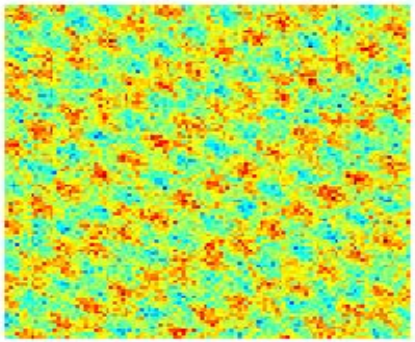
3 harmonic components, 128 × 128 grid, minimal wavelength 8.0



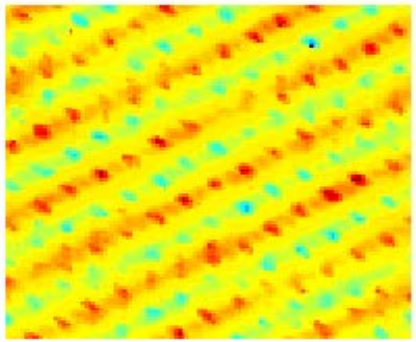
Recovery, $T = 10$
MSE=0.25



True image



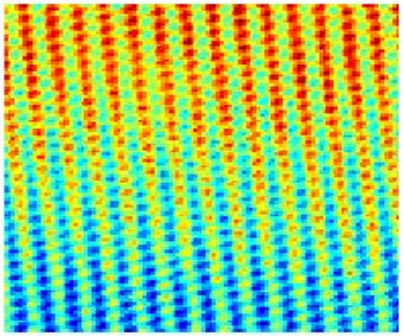
Observations
$\sigma = 1.00$



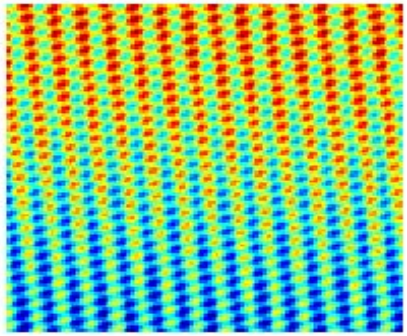
Standard recovery
MSE=0.79

Note: Here again, the recovering routine has no idea about the amplitudes, wavelengths, and directions of wavefronts of interfering waves!

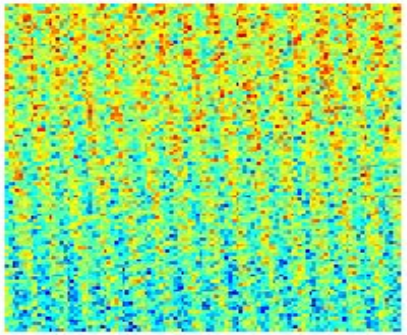
3 harmonic components, 128 × 128 grid, minimal wavelength 4.0



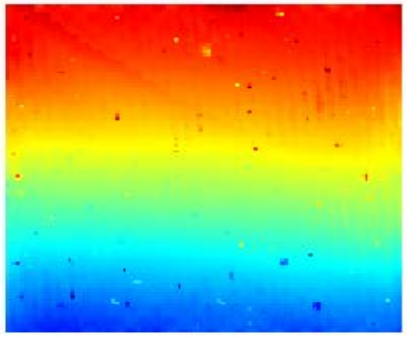
Recovery, $T = 10$
MSE=0.29



True image

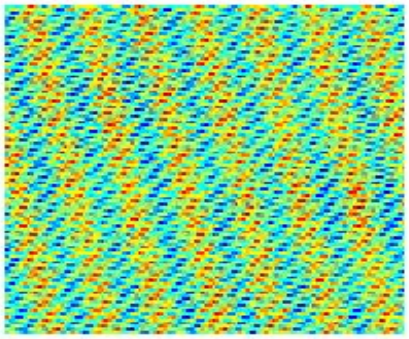


Observations
$\sigma = 1.00$

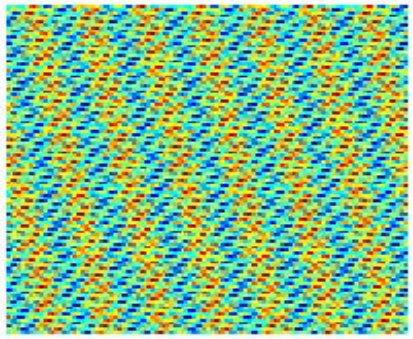


Standard recovery
MSE=0.75

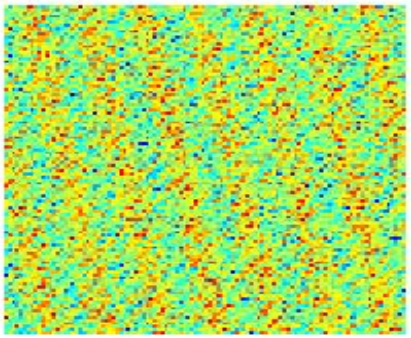
3 harmonic components, 128 × 128 grid, minimal wavelength 2.0



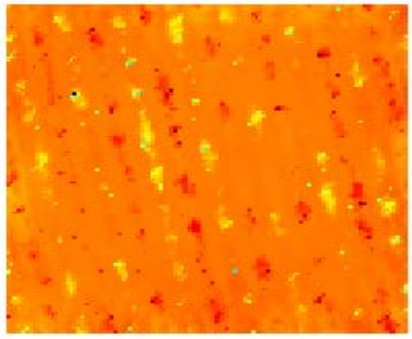
Recovery, $T = 10$
MSE=0.27



True image



Observations
$\sigma = 1.00$



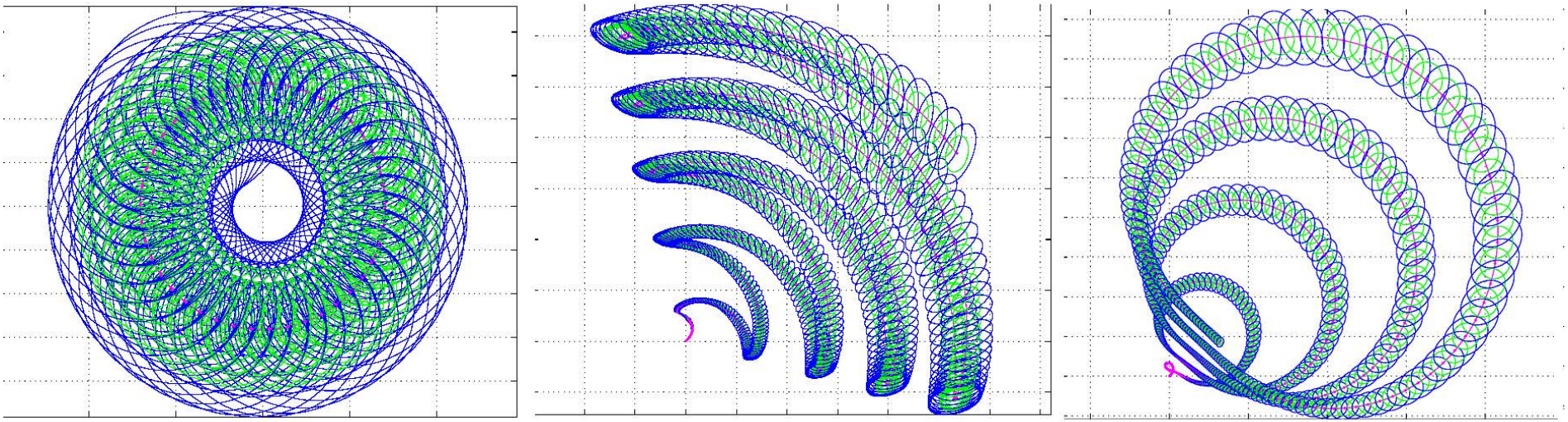
Standard recovery
MSE=0.98

Illustration: Inner & outer ellipsoidal approximations of reachable sets

♠ Given continuous time linear dynamical system

$$\frac{d}{dt}x(t) = Ax(t) + Bu(t) + f(t),\ x(0) = 0$$

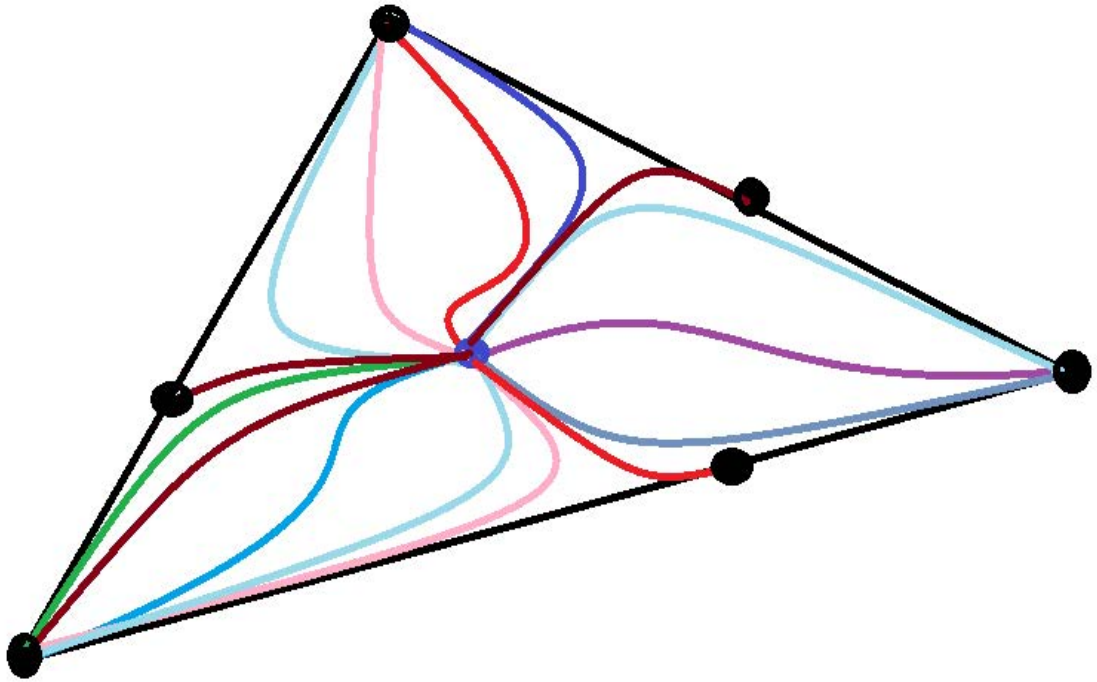
with control $u(\cdot)$ subject to norm bound $\|u(t)\|_2 \leq 1$ for all t, we want to approximate by ellipsoids, from inside and from outside, the set of states which can be reached at time t.



Inner (green) and outer (blue) elliptic approximations of reachable sets of continuous time linear dynamical systems under norm-bounded control vs. time

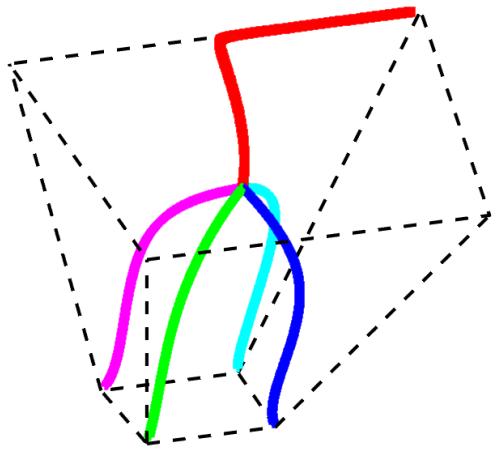
(a)	$\frac{d}{dt}$	$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}$	$=$	$\left[\begin{array}{c\|c} 0 & 1 \\ \hline -1 & 0 \end{array}\right]$	$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}$	$+u(t)$	$\begin{bmatrix} 0 \\ 0.05 \end{bmatrix}$	,					$x(0) = 0, \|u(\cdot)\| \leq 1, 0 \leq t \leq 30$
(b)	$\frac{d}{dt}$	$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}$	$=$	$\left[\begin{array}{c\|c} 0 & -\sin(t) \\ \hline \sin(t) & 0 \end{array}\right]$	$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}$	$+u(t)$	$\begin{bmatrix} \cos(t) \\ \sin(t) \end{bmatrix}$	$+$	$\begin{bmatrix} 10 \\ 10 \end{bmatrix}$	,	$x(0) = 0, \|u(\cdot)\| \leq 1, 0 \leq t \leq 30$		
(c)	$\frac{d}{dt}$	$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}$	$=$	$\left[\begin{array}{c\|c} \cos(t) & -\sin(t) \\ \hline \sin(t) & \cos(t) \end{array}\right]$	$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}$	$+u(t)$	$\begin{bmatrix} \cos(t) \\ \sin(t) \end{bmatrix}$	$+$	$\begin{bmatrix} 10 \\ 10 \end{bmatrix}$	,	$x(0) = 0, \|u(\cdot)\| \leq 1, 0 \leq t \leq 30$		

Part V.
Interior Point Algorithms
for
Linear and Semidefinite Optimization



Lecture V.1
Interior Point Algorithms
for
Linear and Semidefinite Optimization

Executive Summary



Interior Point Methods for LP and SDP

♣ **Interior Point Methods** (IPM's) are state-of-the-art theoretically and practically efficient polynomial time algorithms for solving well-structured convex optimization programs, primarily Linear, Conic Quadratic and Semidefinite ones.
Modern IPMs were first developed for LP, and the words "Interior Point" are aimed at stressing the fact that instead of traveling along the vertices of the feasible set, as in the Simplex algorithm, the new methods work in the interior of the feasible domain.
♠ Basic theory of IPMs remains the same when passing from LP to SDP
$\Rightarrow$ It makes sense to study this theory in the more general SDP case.

Note: For proofs of the facts to follow, see A. Nemirovski, *Introduction to Linear Optimization*. WorldScientific 2024 `https://www2.isye.gatech.edu/~nemirovs/WSLOPostPrint.pdf`

Primal-Dual Pair of SDP Programs

♣ Consider an SDP program in the form

$$\mathrm{Opt}(P) = \min_x \left\{ c^T x : \mathcal{A}x := \sum_{j=1}^n x_j A_j \succeq B \right\} \qquad (P)$$

where A_j, B are $m \times m$ block diagonal symmetric matrices of a given block-diagonal structure $\nu = (\nu_1, ..., \nu_K)$ (i.e., with a given number K and given sizes ν_k, $k \leq K$, of diagonal blocks). (P) can be thought of as a conic problem on the self-dual and regular positive semidefinite cone $\mathbf{S}^\nu_+$ in the space $\mathbf{S}^\nu$ of symmetric block diagonal $m \times m$ matrices with block-diagonal structure ν.
Note: In the diagonal case (with the block-diagonal structure in question, all diagonal blocks are of size 1), (P) becomes a LP program with m linear inequality constraints and n variables.

♠ **Standing Assumption A:** The mapping $x \mapsto \mathcal{A}x$ has trivial kernel, or, equivalently, the matrices $A_1, ..., A_n$ are linearly independent.

$$\mathrm{Opt}(P) = \min_x \left\{ c^T x : \mathcal{A}x := \sum_{j=1}^n x_j A_j \succeq B \right\} \quad [A_j, B \in \mathbf{S}^\nu] \qquad (P)$$

♠ The problem dual to (P) is

$$\mathrm{Opt}(D) = \max_{S \in \mathbf{S}^\nu} \{ \mathrm{Tr}(BS) : S \succeq 0, \mathrm{Tr}(A_j S) = c_j \, \forall j \} \qquad (D)$$

Recall where (D) comes from. The cone $\mathbf{S}^\nu_+$ of positive semidefinite matrices from $\mathbf{S}^\nu$ is self-dual $\Rightarrow$ *when x is feasible for (P) and $0 \preceq S \in \mathbf{S}^\nu$, one has*

$$\sum_j \mathrm{Tr}(SA_j)x_j = \mathrm{Tr}(S[\mathcal{A}x]) \geq \mathrm{Tr}(SB).$$

Imposing on S, in addition to $S \in \mathbf{S}^\nu_+$, the restriction $\mathrm{Tr}(SA_j) = c_j$, $j \leq n$, $\mathrm{Tr}(SB)$ becomes a lower bound on $\mathrm{Opt}(P)$, and (D) is the problem of maximizing this lower bound.

♠ **Standing Assumption B:** Both (P) and (D) are strictly feasible (i.e., have feasible solutions satisfying the "$\succ$" versions of the LMIs).

Notation: In the sequel, $m = \sum_i \nu_i$

$$\mathrm{Opt}(P) = \min_x \left\{ c^T x : \mathcal{A}x := \sum_{j=1}^n x_j A_j \succeq B \right\} \qquad (P)$$

$$\mathrm{Opt}(D) = \max_{S \in \mathbf{S}^\nu} \{ \mathrm{Tr}(BS) : S \succeq 0, \mathrm{Tr}(A_j S) = c_j \, \forall j \} \qquad (D)$$

♠ Let $C \in \mathbf{S}^\nu$ satisfy the equality constraint in (D). Passing in (P) from x to the primal slack $X = \mathcal{A}x - B$, (P) becomes the problem

$$\begin{array}{c} \mathrm{Opt}(\mathcal{P}) = \min\limits_{X \in \mathbf{S}^\nu} \left\{ \mathrm{Tr}(CX) : X \in [\mathcal{L}_P - B] \cap \mathbf{S}^\nu_+ \right\} \\ \mathcal{L}_P = \{X = \mathcal{A}x\} = \mathrm{Lin}\{A_1, ..., A_n\} \\ C : \mathrm{Tr}(A_i C) = c_i, i \le n \end{array} \qquad (\mathcal{P})$$

while (D) is the problem

$$\begin{array}{c} \mathrm{Opt}(D) = \max\limits_{S \in \mathbf{S}^\nu} \left\{ \mathrm{Tr}(BS) : S \in [\mathcal{L}_D + C] \cap \mathbf{S}^\nu_+ \right\} \\ \mathcal{L}_D = \mathcal{L}_P^\perp = \{S \in \mathbf{S}^\nu : \mathrm{Tr}(A_j S) = 0, \, 1 \le j \le n\} \end{array} \qquad (D)$$

• Since (P) and (D) are strictly feasible, both problems are solvable with equal optimal values, and a pair of feasible solutions X to $(\mathcal{P})$ and S to $(\mathcal{D})$ is composed of optimal solutions to the respective problems iff $\mathrm{Tr}(XS) = 0$.

Note: *x is feasible for (P) iff $X = \mathcal{A}x - B$ is feasible for $(\mathcal{P})$. When x is feasible for (P), S is feasible for (D) and $X = \mathcal{A}x - b$,, one has*

$\mathrm{DualityGap}(x, S) := c^T x - \mathrm{Tr}(BS) = \sum_i x_i \mathrm{Tr}(A_i S) - \mathrm{Tr}(BS) = \mathrm{Tr}([\mathcal{A}x]S) - \mathrm{Tr}(BS) = \mathrm{Tr}(XS)$

⇒ **Conclusion:** *For primal-dual feasible pair of solutions x to (P) and S to (D), the sum of their non-optimalities in the respective problems is $\mathrm{Tr}(XS)$, where $X = \mathcal{A}x - B$ is the feasible solution to $(\mathcal{P})$ induced by x. The necessary and sufficient condition for the pair (x, S) to be composed of optimal solutions to the respective problems is $\mathrm{Tr}(XS) = 0$.*

$$\mathrm{Opt}(P) = \min_x \left\{ c^T x : \mathcal{A}x := \sum_{j=1}^n x_j A_j \succeq B \right\} \qquad (P)$$
$$\Leftrightarrow \mathrm{Opt}(\mathcal{P}) = \min_X \left\{ \mathrm{Tr}(CX) : X \in [\mathcal{L}_P - B] \cap \mathbf{S}^\nu_+ \right\} \qquad (\mathcal{P})$$
$$\mathrm{Opt}(D) = \max_S \left\{ \mathrm{Tr}(BS) : S \in [\mathcal{L}_D + C] \cap \mathbf{S}^\nu_+ \right\} \qquad (D)$$
$$\left[\mathcal{L}_P = \mathrm{Im}\mathcal{A},\ \mathcal{L}_D = \mathcal{L}_P^\perp \right]$$

Fact: *For positive semidefinite X, S, $\mathrm{Tr}(XS) = 0$ if and only if $XS = SX = 0$.*

✓ Indeed, $XS = 0$ clearly implies $\mathrm{Tr}(XS) = 0$. ✓ Vice versa, let $X \succeq 0, S \succeq 0$ and $\mathrm{Tr}(XS) = 0$. We have

$$\begin{array}{rl} & 0 = \mathrm{Tr}(XS) = \mathrm{Tr}(X^{1/2}X^{1/2}S) = \mathrm{Tr}(X^{1/2}SX^{1/2}) \\ & \quad [\text{as } \mathrm{Tr}(AB) = \mathrm{Tr}(BA) \text{ whenever } AB \text{ makes sense and is square}] \\ \Rightarrow & X^{1/2}SX^{1/2} = 0 \ [\text{as } X^{1/2}SX^{1/2} \succeq 0] \\ \Rightarrow & 0 = X^{1/2}S^{1/2}S^{1/2}X^{1/2} = [X^{1/2}S^{1/2}][X^{1/2}S^{1.2}]^T \\ \Rightarrow & X^{1/2}S^{1/2} = 0 \ [\text{as } \sum_{i,j}[X^{1/2}S^{1/2}]^2_{ij} = \mathrm{Tr}([X^{1/2}S^{1/2}][X^{1/2}S^{1.2}]^T)] \\ \Rightarrow & XS = X^{1/2}[X^{1/2}S^{1/2}]S^{1/2} = 0 \\ \Rightarrow & SX = [XS]^T = 0 \end{array}$$

♠ We have arrived at

Fact XI.18 *Assuming (P), (D) strictly feasible, feasible solutions X for $(\mathcal{P})$ and S for (D) are optimal for the respective problems if and only if*

$$XS = SX = 0$$

("SDP Complementary Slackness").

Logarithmic Barrier for the Semidefinite Cone $\mathbf{S}^\nu_+$

$$\begin{array}{rl}
\mathrm{Opt}(P) = \min\limits_x \left\{c^T x : \mathcal{A}x := \sum_{j=1}^n x_j A_j \succeq B\right\} & (P)\\
\Leftrightarrow \mathrm{Opt}(\mathcal{P}) = \min\limits_X \left\{\mathrm{Tr}(CX) : X \in [\mathcal{L}_P - B] \cap \mathbf{S}^\nu_+\right\} & (\mathcal{P})\\
\mathrm{Opt}(D) = \max\limits_S \left\{\mathrm{Tr}(BS) : S \in [\mathcal{L}_D + C] \cap \mathbf{S}^\nu_+\right\} & (D)\\
\left[\mathcal{L}_P = \mathrm{Im}\mathcal{A},\ \mathcal{L}_D = \mathcal{L}_P^\perp\right] &
\end{array}$$

♣ A crucial role in building IPMs for (P), (D) is played by the *logarithmic barrier for the positive semidefinite cone:*

$$K(X) = -\ln \mathrm{Det}(X) : \mathrm{int}\,\mathbf{S}^\nu_+ \to \mathbf{R}$$

Back to Basic Analysis: Gradient and Hessian

♣ Consider a smooth (3 times continuously differentiable) function $f(x) : D \to \mathbf{R}$ defined on an open subset D of Euclidean space E with inner product $\langle \cdot, \cdot \rangle$.

♠ The *first order directional derivative of* f taken at a point $x \in D$ along a direction $h \in E$ is the quantity

$$Df(x)[h] := \frac{d}{dt}\Big|_{t=0} f(x + th)$$

Fact: *For a smooth* f, $Df(x)[h]$ *is linear in* h *and thus*

$$Df(x)[h] = \langle \nabla f(x), h \rangle \ \forall h$$

for a uniquely defined vector $\nabla f(x)$ *called the gradient of* f *at* x.

If E *is* $\mathbf{R}^n$ *with the standard Euclidean structure, then*

$$[\nabla f(x)]_i = \frac{\partial}{\partial x_i} f(x),\ 1 \leq i \leq n$$

♠ The *second order directional derivative* of f taken at a point $x \in D$ along a *pair* of directions g, h is defined as

$$D^2 f(x)[g,h] = \frac{d}{dt}\big|_{t=0} [Df(x+tg)[h]]$$

Fact: *For a smooth f, $D^2 f(x)[g,h]$ is bilinear and symmetric in g, h, and therefore*

$$D^2 f(x)[g,h] = \langle g, \nabla^2 f(x) h\rangle = \langle \nabla^2 f(x) g, h\rangle \forall g, h \in E$$

for a uniquely defined linear mapping $h \mapsto \nabla^2 f(x) h : E \to E$, called the Hessian of f at x. If E is $\mathbf{R}^n$ with the standard Euclidean structure, then

$$[\nabla^2 f(x)]_{ij} = \frac{\partial^2}{\partial x_i \partial x_j} f(x)$$

Fact: *Hessian is the derivative of the gradient:*

$$\nabla f(x+h) = \nabla f(x) + [\nabla^2 f(x)]h + R_x(h),\ \|R_x(h)\| \le C_x \|h\|^2 \forall (h : \|h\| \le \rho_x), \rho_x > 0$$

Fact: *Gradient and Hessian define the second order Taylor expansion*

$$\widehat{f}(y) = f(x) + \langle y - x, \nabla f(x)\rangle + \tfrac{1}{2}\langle y - x, \nabla^2 f(x)[y-x]\rangle$$

of f at x which is a quadratic function of y with the same gradient and Hessian at x as those of f. This expansion approximates f around x, specifically,

$$|f(y) - \widehat{f}(y)| \le C_x \|y - x\|^3 \ \forall (y : \|y - x\| \le \rho_x), \rho_x > 0$$

Back to SDP

$$\mathrm{Opt}(P) = \min_x \left\{ c^T x : \mathcal{A}x := \sum_{j=1}^n x_j A_j \succeq B \right\} \qquad (P)$$

$$\Leftrightarrow \mathrm{Opt}(\mathcal{P}) = \min_X \left\{ \mathrm{Tr}(CX) : X \in [\mathcal{L}_P - B] \cap \mathbf{S}^\nu_+ \right\} \qquad (\mathcal{P})$$

$$\mathrm{Opt}(D) = \max_S \left\{ \mathrm{Tr}(BS) : S \in [\mathcal{L}_D + C] \cap \mathbf{S}^\nu_+ \right\} \qquad (D)$$

$$\left[\mathcal{L}_P = \mathrm{Im}\mathcal{A},\ \mathcal{L}_D = \mathcal{L}_P^\perp \right]$$

$$K(X) = -\ln \mathrm{Det} X : \mathbf{S}^\nu_{++} := \{X \in \mathbf{S}^\nu : X \succ 0\} \to \mathbf{R}$$

Fact: *$K(X)$ is a smooth function on its domain $\mathbf{S}^\nu_{++} = \{X \in \mathbf{S}^\nu : X \succ 0\}$. The first- and the second order derivatives of this function taken at a point $X \in \mathrm{Dom}\, K$ along directions $H, G \in \mathbf{S}^\nu$ are given by*

$$\begin{array}{rcl} DK(X)[H] & := & \frac{d}{dt}\big|_{t=0} K(X+tH) \\ & = & -\mathrm{Tr}(X^{-1}H) \quad \left[\Leftrightarrow \nabla K(X) = -X^{-1}\right] \end{array}$$

$$\begin{array}{rcl} D^2K(X)[H,G] & := & \frac{d}{dt}\big|_{t=0} DK(x+tG)[H] \\ & = & \mathrm{Tr}(X^{-1}HX^{-1}G) \quad \left[\Leftrightarrow [\nabla^2 K(X)]H = X^{-1}HX^{-1}\right] \\ \frac{d^2}{dt^2}\big|_{t=0} K(X+tH) & = & D^2K(X)[H,H] \\ & = & \mathrm{Tr}(H[X^{-1}HX^{-1}]) = \mathrm{Tr}([X^{-1/2}HX^{-1/2}]^2) \end{array}$$

In particular, K is strongly convex:

$$X \in \mathrm{Dom}\, K, 0 \neq H \in \mathbf{S}^\nu \Rightarrow \frac{d^2}{dt^2}\big|_{t=0} K(X+tH) > 0$$

♠ **Additional properties of $K(\cdot)$:**

• $\nabla K(tX) = -[tX]^{-1} = -t^{-1}X^{-1} = t^{-1}\nabla K(X)$

• The mapping $X \mapsto -\nabla K(X) = X^{-1}$ maps the domain $\mathbf{S}^\nu_{++}$ of K onto itself and is self-inverse:

$$S = -\nabla K(X) \Leftrightarrow X = -\nabla K(S) \Leftrightarrow XS = SX = I$$

• The function $K(X)$ is an *interior penalty* for the positive semidefinite cone $\mathbf{S}^\nu_+$: whenever points $X_i \in \mathrm{Dom}\, K = \mathbf{S}^\nu_{++}$ converge to a boundary point of $\mathbf{S}^\nu_+$, one has $K(X_i) \to \infty$ as $i \to \infty$.

Primal-Dual Central Path

$$\mathrm{Opt}(P) = \min_x \left\{ c^T x : \mathcal{A}x := \sum_{j=1}^n x_j A_j \succeq B \right\} \qquad (P)$$
$$\Leftrightarrow \mathrm{Opt}(\mathcal{P}) = \min_X \left\{ \mathrm{Tr}(CX) : X \in [\mathcal{L}_P - B] \cap \mathbf{S}^\nu_+ \right\} \qquad (\mathcal{P})$$
$$\mathrm{Opt}(D) = \max_S \left\{ \mathrm{Tr}(BS) : S \in [\mathcal{L}_D + C] \cap \mathbf{S}^\nu_+ \right\} \qquad (D)$$
$$K(X) = -\ln \mathrm{Det}(X)$$

♠ Let

$$\mathcal{X} = \{X \in \mathcal{L}_P - B : X \succ 0\}$$
$$\mathcal{S} = \{S \in \mathcal{L}_D + C : S \succ 0\}.$$

be the (nonempty!) sets of strictly feasible solutions to $(\mathcal{P})$ and (D), respectively. Given *path parameter* $\mu > 0$, consider the functions

$$P_\mu(X) = \mathrm{Tr}(CX) + \mu K(X) : \mathcal{X} \to \mathbf{R}$$
$$D_\mu(S) = -\mathrm{Tr}(BS) + \mu K(S) : \mathcal{S} \to \mathbf{R} \ .$$

Fact: *For every $\mu > 0$, the function $P_\mu(X)$ achieves its minimum at $\mathcal{X}$ at a unique point $X_*(\mu)$, and the function $D_\mu(S)$ achieves its minimum on $\mathcal{S}$ at a unique point $S_*(\mu)$. These points are related to each other:*

$$X_*(\mu) = \mu S_*^{-1}(\mu) \Leftrightarrow S_*(\mu) = \mu X_*^{-1}(\mu)$$
$$\Leftrightarrow X_*(\mu) S_*(\mu) = S_*(\mu) X_*(\mu) = \mu I$$

♠ We associate with $(\mathcal{P})$, (D) the *primal-dual central path – the curve* $\{X_*(\mu), S_*(\mu)\}_{\mu>0}$; for every $\mu > 0$, $X_*(\mu)$ is a strictly feasible solution to $(\mathcal{P})$, and $S_*(\mu)$ is a strictly feasible solution to (D).

Duality Gap on the Central Path

$$\mathrm{Opt}(P) = \min_x \left\{ c^T x : \mathcal{A}x := \sum_{j=1}^n x_j A_j \succeq B \right\} \qquad (P)$$

$$\Leftrightarrow \mathrm{Opt}(\mathcal{P}) = \min_X \left\{ \mathrm{Tr}(CX) : X \in [\mathcal{L}_P - B] \cap \mathbf{S}^\nu_+ \right\} \qquad (\mathcal{P})$$

$$\mathrm{Opt}(D) = \max_S \left\{ \mathrm{Tr}(BS) : S \in [\mathcal{L}_D + C] \cap \mathbf{S}^\nu_+ \right\} \qquad (D)$$

$$\Rightarrow \left\{ \begin{array}{l} X_*(\mu) \in [\mathcal{L}_P - B] \cap \mathbf{S}^\nu_{++} \\ S_*(\mu) \in [\mathcal{L}_D + C] \cap \mathbf{S}^\nu_{++} \end{array} \right\} : X_*(\mu) S_*(\mu) = \mu I$$

Observation: *On the primal-dual central path, the duality gap is*

$$\mathrm{Tr}(X_*(\mu) S_*(\mu)) = \mathrm{Tr}(\mu I) = \mu m.$$

Therefore sum of non-optimalities of the strictly feasible solution $X_*(\mu)$ to $(\mathcal{P})$ and the strictly feasible solution $S_*(\mu)$ to (D) in terms of the respective objectives is equal to μm and goes to 0 as $\mu \to +0$.

$\Rightarrow$ *Our ideal goal would be to move along the primal-dual central path, pushing the path parameter μ to 0 and thus approaching primal-dual optimality, while maintaining primal-dual feasibility.*

♠ Our ideal goal is not achievable – how could we move along a curve? A *realistic* goal could be to move in a neighborhood of the primal-dual central path, staying close to it. A good notion of "closeness to the path" is given by the *proximity measure* of a triple $\mu > 0, X \in \mathcal{X}, S \in \mathcal{S}$ to the point $(X_*(\mu), S_*(\mu))$ on the path:

$$\begin{aligned}
\mathrm{dist}(X,S,\mu) &= \sqrt{\mathrm{Tr}(X[X^{-1}-\mu^{-1}S]X[X^{-1}-\mu^{-1}S])} \\
&= \sqrt{\mathrm{Tr}(X^{1/2}[X^{1/2}[X^{-1}-\mu^{-1}S]X^{1/2}]\left[X^{1/2}[X^{-1}-\mu^{-1}S]\right])} \\
&= \sqrt{\mathrm{Tr}([X^{1/2}[X^{-1}-\mu^{-1}S]X^{1/2}]\left[X^{1/2}[X^{-1}-\mu^{-1}S]X^{1/2}\right])} \\
&= \sqrt{\mathrm{Tr}([X^{1/2}[X^{-1}-\mu^{-1}S]X^{1/2}]^2)} \\
&= \sqrt{\mathrm{Tr}([I-\mu^{-1}X^{1/2}SX^{1/2}]^2)}.
\end{aligned}$$

Note: We see that $\mathrm{dist}(X,S,\mu)$ is well defined and $\mathrm{dist}(X,S,\mu)=0$ iff $X^{1/2}SX^{1/2}=\mu I$, or, which is the same,

$$SX = X^{-1/2}[X^{1/2}SX^{1/2}]X^{1/2} = \mu X^{-1/2}X^{1/2} = \mu I,$$

i.e., iff $X = X_*(\mu)$ and $S = S_*(\mu)$.

Note: In the LP case, $\mathrm{dist}(X,S,\mu) = \sqrt{\sum_i [1 - X_{ii}S_{ii}/\mu]^2}$.

Note: We have

$$\begin{aligned}
\mathrm{dist}(X,S,\mu) &= \sqrt{\mathrm{Tr}(X[X^{-1}-\mu^{-1}S]X[X^{-1}-\mu^{-1}S])} \\
&= \sqrt{\mathrm{Tr}([I-\mu^{-1}XS][I-\mu^{-1}XS])} \\
&= \sqrt{\mathrm{Tr}(\left[[I-\mu^{-1}XS][I-\mu^{-1}XS]\right]^T)} \\
&= \sqrt{\mathrm{Tr}([I-\mu^{-1}SX][I-\mu^{-1}SX])} \\
&= \sqrt{\mathrm{Tr}(S[S^{-1}-\mu^{-1}X]S[S^{-1}-\mu^{-1}X])},
\end{aligned}$$

⇒ The proximity is defined in a symmetric w.r.t. X, S fashion.

Fact: *Whenever $X \in \mathcal{X}$, $S \in \mathcal{S}$ and $\mu > 0$, one has*

$$\mathrm{Tr}(XS) \leq \mu[m + \sqrt{m}\,\mathrm{dist}(X, S, \mu)]$$

Corollary. *Let us say that a triple (X, S, μ) is close to the path, if $X \in \mathcal{X}$, $S \in \mathcal{S}$, $\mu > 0$ and* $\mathrm{dist}(X, S, \mu) \leq 0.1$. *Whenever (X, S, μ) is close to the path, one has*

$$\mathrm{Tr}(XS) \leq 2\mu m,$$

that is, if (X, S, μ) is close to the path, then X is at most $2\mu m$-nonoptimal strictly feasible solution to $(\mathcal{P})$, and S is at most $2\mu m$-nonoptimal strictly feasible solution to (D).

How to Trace the Central Path?

♣ **The goal:** To follow the central path, staying close to it and pushing μ to 0 as fast as possible.

♣ **Question.** *Assume we are given a triple $(\bar{X}, \bar{S}, \bar{\mu})$ close to the path. How to update it into a triple (X_+, S_+, μ_+), also close to the path, with $\mu_+ < \mu$?*

♠ **Conceptual answer:** Let us choose μ_+, $0 < \mu_+ < \bar{\mu}$, and try to update $\bar{X}, \bar{S}$ into $X_+ = \bar{X} + \Delta X$, $S_+ = \bar{S} + \Delta S$ in order to make the triple (X_+, S_+, μ_+) close to the path. Our goal is to ensure that

$$\begin{array}{rcl} X_+ = \bar{X} + \Delta X & \in & \mathcal{L}_P - B \quad \& \quad X_+ \succ 0 \quad (a) \\ S_+ = \bar{S} + \Delta S & \in & \mathcal{L}_D + C \quad \& \quad S_+ \succ 0 \quad (b) \\ & & G_{\mu_+}(X_+, S_+) \approx 0 \qquad\qquad\quad (c) \end{array}$$

where $G_\mu(X, S) = 0$ expresses equivalently the *augmented slackness* condition $XS = \mu I$. For example, we can take

$$\begin{array}{l} G_\mu(X, S) = S - \mu X^{-1}, \text{ or} \\ G_\mu(X, S) = X - \mu S^{-1}, \text{ or} \\ G_\mu(X, S) = XS + SX - 2\mu I, \quad \text{or...} \end{array}$$

$$\begin{array}{rcccccl} X_+ = \bar{X} + \Delta X & \in & \mathcal{L}_P - B & \& & X_+ & \succ & 0 \quad (a) \\ S_+ = \bar{S} + \Delta S & \in & \mathcal{L}_D + C & \& & S_+ & \succ & 0 \quad (b) \\ & & G_{\mu_+}(X_+, S_+) \approx 0 & & & & (c) \end{array}$$

♠ Since $\bar{X} \in \mathcal{L}_P - B$ and $\bar{X} \succ 0$, (a) amounts to $\Delta X \in \mathcal{L}_P$, which is a system of linear equations on ΔX, and to $\bar{X} + \Delta X \succ 0$. Similarly, (b) amounts to the system $\Delta S \in \mathcal{L}_D$ of linear equations on ΔS, and to $\bar{S} + \Delta S \succ 0$. To handle the troublemaking *nonlinear in* $\Delta X, \Delta S$ condition (c), we *linearize* G_{μ_+} *in* ΔX *and* ΔS*:*

$$\begin{array}{l} G_{\mu_+}(X_+, S_+) \approx G_{\mu_+}(\bar{X}, \bar{S}) \\ \quad + \left.\frac{\partial G_{\mu_+}(X,S)}{\partial X}\right|_{(X,S)=(\bar{X},\bar{S})} \Delta X + \left.\frac{\partial G_{\mu_+}(X,S)}{\partial S}\right|_{(X,S)=(\bar{X},\bar{S})} \Delta S \end{array}$$

and enforce the linearization, as evaluated at ΔX*,* ΔS*, to be zero.* We arrive at the *Newton system*

$$\left\{ \begin{array}{l} \Delta X \in \mathcal{L}_P, \ \ \Delta S \in \mathcal{L}_D \\ \frac{\partial G_{\mu_+}}{\partial X} \Delta X + \frac{\partial G_{\mu_+}}{\partial S} \Delta S = -G_{\mu_+} \end{array} \right. \qquad (N)$$

(the value and the partial derivatives of $G_{\mu_+}(X, S)$ are taken at the point $(\bar{X}, \bar{S})$).

♠ We arrive at conceptual *primal-dual path-following method* where one iterates the updates

$$(X_i, S_i, \mu_i) \mapsto (X_{i+1} = X_i + \Delta X_i, S_{i+1} = S_i + \Delta S_i, \mu_{i+1})$$

where $\mu_{i+1} \in (0, \mu_i)$ and $\Delta X_i, \Delta S_i$ are the solution to the Newton system

$$\left\{ \begin{array}{l} \Delta X_i \in \mathcal{L}_P, \;\; \Delta S_i \in \mathcal{L}_D \\ \frac{\partial G^{(i)}_{\mu_{i+1}}}{\partial X} \Delta X_i + \frac{\partial G^{(i)}_{\mu_{i+1}}}{\partial S} \Delta S_i = -G^{(i)}_{\mu_{i+1}} \end{array} \right. \qquad (N_i)$$

and $G^{(i)}_\mu(X, S) = 0$ represents equivalently the augmented complementary slackness condition $XS = \mu I$ and the value and the partial derivatives of $G^{(i)}_{\mu_{i+1}}$ are evaluated at (X_i, S_i).

♠ Initialized by a close to the path triple (X_0, S_0, μ_0), this conceptual algorithm should

- be well-defined: (N_i) should remain solvable, X_i should remain strictly feasible for $(\mathcal{P})$, S_i should remain strictly feasible for (D), and
- maintain closeness to the path: for every i, (X_i, S_i, μ_i) should remain close to the path.

Under these limitations, we want to push μ_i to 0 as fast as possible.

Example: Primal Path-Following Method

$$\mathrm{Opt}(P) = \min_x \left\{ c^T x : \mathcal{A}x := \sum_{j=1}^n x_j A_j \succeq B \right\} \qquad (P)$$

$$\Leftrightarrow \mathrm{Opt}(\mathcal{P}) = \min_X \left\{ \mathrm{Tr}(CX) : X \in [\mathcal{L}_P - B] \cap \mathbf{S}^\nu_+ \right\} \qquad (\mathcal{P})$$

$$\mathrm{Opt}(D) = \max_S \left\{ \mathrm{Tr}(BS) : S \in [\mathcal{L}_D + C] \cap \mathbf{S}^\nu_+ \right\} \qquad (D)$$

$$\left[\mathcal{L}_P = \mathrm{Im}\mathcal{A},\ \mathcal{L}_D = \mathcal{L}_P^\perp \right]$$

♣ Let us choose

$$G_\mu(X,S) = S + \mu \nabla K(X) = S - \mu X^{-1}$$

Then the Newton system becomes

$$\begin{array}{rcl} \Delta X_i \in \mathcal{L}_P & \Leftrightarrow & \Delta X_i = \mathcal{A}\Delta x_i \\ \Delta S_i \in \mathcal{L}_D & \Leftrightarrow & \mathcal{A}^* \Delta S_i = 0 \\ & & \mathcal{A}^* U = [\mathrm{Tr}(A_1 U); ...; \mathrm{Tr}(A_n U)] \end{array} \qquad (N_i)$$

$$(!) \quad \Delta S_i + \mu_{i+1} \nabla^2 K(X_i) \Delta X_i = -[S_i + \mu_{i+1} \nabla K(X_i)]$$

♠ Substituting $\Delta X_i = \mathcal{A}\Delta x_i$ and applying $\mathcal{A}^*$ to both sides in (!), we get

$$(*) \quad \mu_{i+1} \underbrace{[\mathcal{A}^* \nabla^2 K(X_i) \mathcal{A}]}_{\mathcal{H}} \Delta x_i = -\big[\underbrace{\mathcal{A}^* S_i}_{=c} + \mu_{i+1} \mathcal{A}^* \nabla K(X_i) \big]$$

$$\begin{array}{l} \Delta X_i = \mathcal{A}\Delta x_i \\ S_{i+1} = \mu_{i+1} \left[\nabla K(X_i) - \nabla^2 K(X_i) \mathcal{A}\Delta x_i \right] \end{array}$$

The mappings $h \mapsto \mathcal{A}h$, $H \mapsto \nabla^2 K(X_i) H$ have trivial kernels $\Rightarrow$ $\mathcal{H}$ is nonsingular $\Rightarrow$ (N_i) has a unique solution given by

$$\begin{array}{l} \Delta x_i = -\mathcal{H}^{-1} \left[\mu_{i+1}^{-1} c + \mathcal{A}^* \nabla K(X_i) \right] \\ \Delta X_i = \mathcal{A}\Delta x_i \\ S_{i+1} = S_i + \Delta S_i = -\mu_{i+1} \left[\nabla K(X_i) + \nabla^2 K(X_i) \mathcal{A}\Delta x_i \right] \end{array}$$

$$\begin{array}{rl}
\mathrm{Opt}(P) = \min\limits_x \left\{c^T x : \mathcal{A}x := \sum_{j=1}^n x_j A_j \succeq B\right\} & (P)\\
\Leftrightarrow \mathrm{Opt}(\mathcal{P}) = \min\limits_X \left\{\mathrm{Tr}(CX) : X \in [\mathcal{L}_P - B] \cap \mathbf{S}^\nu_+\right\} & (\mathcal{P})\\
\mathrm{Opt}(D) = \max\limits_S \left\{\mathrm{Tr}(BS) : S \in [\mathcal{L}_D + C] \cap \mathbf{S}^\nu_+\right\} & (D)\\
\left[\mathcal{L}_P = \mathrm{Im}\mathcal{A},\ \mathcal{L}_D = \mathcal{L}_P^\perp\right] &\\
\Rightarrow \left\{\begin{array}{l}
\Delta x_i = -\mathcal{H}^{-1}\left[\mu_{i+1}^{-1} c + \mathcal{A}^*\nabla K(X_i)\right]\\
\Delta X_i = \mathcal{A}\Delta x_i\\
S_{i+1} = S_i + \Delta S_i = -\mu_{i+1}\left[\nabla K(X_i) + \nabla^2 K(X_i)\mathcal{A}\Delta x_i\right]
\end{array}\right. &
\end{array}$$

♠ $X_i = \mathcal{A}x_i - B$ for a (uniquely defined by X_i) strictly feasible solution x_i to (P). Setting

$$F(x) = K(\mathcal{A}x - B),$$

we have $\mathcal{A}^*\nabla K(X_i) = \nabla F(x_i),\ \mathcal{H} = \nabla^2 F(x_i)$

⇒ *The above recurrence can be written solely in terms of x_i and F:*

$$\left\{\begin{array}{l}
\mu_i \mapsto \mu_{i+1} < \mu_i\\
x_{i+1} = x_i - [\nabla^2 F(x_i)]^{-1}\left[\mu_{i+1}^{-1} c + \nabla F(x_i)\right]
\end{array}\right.$$
$$X_{i+1} = \mathcal{A}x_{i+1} - B \qquad (\#)$$
$$S_{i+1} = -\mu_{i+1}\left[\nabla K(X_i) + \nabla^2 K(X_i)\mathcal{A}[x_{i+1} - x_i]\right]$$

Recurrence (#) is called the primal path-following method.

$$\mathrm{Opt}(P) = \min_x \left\{ c^T x : \mathcal{A}x := \sum_{j=1}^n x_j A_j \succeq B \right\} \qquad (P)$$
$$\Leftrightarrow \mathrm{Opt}(\mathcal{P}) = \min_X \left\{ \mathrm{Tr}(CX) : X \in [\mathcal{L}_P - B] \cap \mathbf{S}_+^\nu \right\} \qquad (\mathcal{P})$$
$$\mathrm{Opt}(D) = \max_S \left\{ \mathrm{Tr}(BS) : S \in [\mathcal{L}_D + C] \cap \mathbf{S}_+^\nu \right\} \qquad (D)$$
$$\left[\mathcal{L}_P = \mathrm{Im}\mathcal{A},\ \mathcal{L}_D = \mathcal{L}_P^\perp\right]$$

♠ The primal path-following method can be explained as follows:

• The barrier $K(X) = -\ln \mathrm{Det} X$ induces the barrier $F(x) = K(\mathcal{A}x - B)$ for the interior P^o of the feasible domain of (P).

• The primal central path

$$X_*(\mu) = \mathrm{argmin}_{X = \mathcal{A}x - B \succ 0} \left[\mathrm{Tr}(CX) + \mu K(X)\right]$$

induces the path

$$x_*(\mu) \in P^o:\ X_*(\mu) = \mathcal{A}x_*(\mu) + \mu F(x).$$

Observing that

$$\mathrm{Tr}(C[\mathcal{A}x - B]) + \mu K(\mathcal{A}x - B) = c^T x + \mu F(x) + \mathrm{const},$$

we have

$$x_*(\mu) = \mathrm{argmin}_{x \in P^o} F_\mu(x),\ F_\mu(x) = c^T x + \mu F(x).$$

• The method works as follows: given $x_i \in P^o, \mu_i > 0$, we

— replace μ_i with $\mu_{i+1} < \mu_i$

— convert x_i into x_{i+1} by applying to the function $F_{\mu_{i+1}}(\cdot)$ a single step of the *Newton minimization method*

$$x_i \mapsto x_{i+1} - [\nabla^2 F_{\mu_{i+1}}(x_i)]^{-1} \nabla F_{\mu_{i+1}}(x_i)$$

$$\begin{array}{rl}
\mathrm{Opt}(P) = \min\limits_{x}\left\{c^Tx : \mathcal{A}x := \sum_{j=1}^n x_jA_j \succeq B\right\} & (P)\\
\Leftrightarrow \mathrm{Opt}(\mathcal{P}) = \min\limits_{X}\left\{\mathrm{Tr}(CX) : X \in [\mathcal{L}_P - B] \cap \mathbf{S}^\nu_+\right\} & (\mathcal{P})\\
\mathrm{Opt}(D) = \max\limits_{S}\left\{\mathrm{Tr}(BS) : S \in [\mathcal{L}_D + C] \cap \mathbf{S}^\nu_+\right\} & (D)\\
\left[\mathcal{L}_P = \mathrm{Im}\mathcal{A},\ \mathcal{L}_D = \mathcal{L}_P^\perp\right]
\end{array}$$

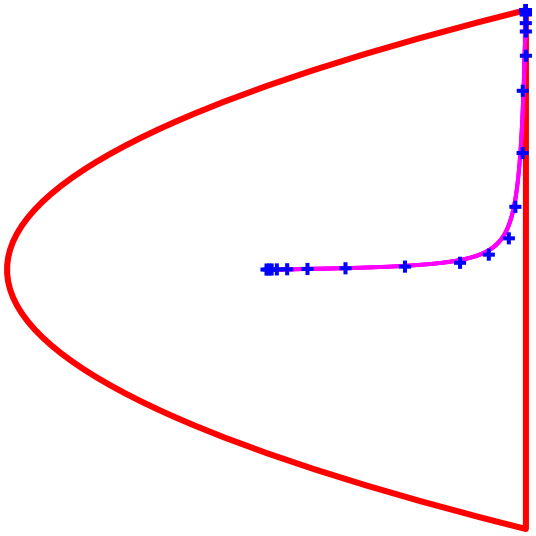
Fact. *Let $(X_0 = \mathcal{A}x_0 - B, S_0, \mu_0)$ be close to the primal-dual central path, and let (P) be solved by the Primal path-following method where the path parameter μ is updated according to*

$$\mu_{i+1} = \left(1 - \frac{0.1}{\sqrt{m}}\right)\mu_i. \qquad (*)$$

Then the method is well defined and all triples $(X_i = \mathcal{A}x_i - B, S_i, \mu_i)$ are close to the path.

♠ With the rule $(*)$ it takes $O(\sqrt{m})$ steps to reduce the path parameter μ by an absolute constant factor. Since the method stays close to the path, the duality gap $\mathrm{Tr}(X_iS_i)$ of i-th iterate does not exceed $2m\mu_i$.

⇒ *The number of steps to make the duality gap $\le \epsilon$ does not exceed $O(1)\sqrt{m}\ln\left(1 + \frac{2m\mu_0}{\epsilon}\right)$.*



$$\min_{x,y}\left\{-x-0.025y : \begin{bmatrix} 1 & x & y \\ x & 1 & y \\ y & y & 1 \end{bmatrix} \succeq 0\right\}$$

Red: feasible set of a toy SDP ($\mathbf{K} = \mathbf{S}^3_+$).

Magenta: the primal central path

Blue "+": iterates x_i of the Primal Path-Following method.

Itr#	Objective	Gap	Itr#	Objective	Gap
1	-0.010003	1.5e+02	20	-1.024917	2.9e-04
5	-0.158040	9.4e+00	25	-1.024997	8.9e-06
10	-0.937196	2.9e-01	30	-1.025000	2.8e-07
15	-1.022733	9.2e-03	32	-1.025000	7.0e-08

Objective and Duality Gap along the iterations

♣ The Primal path-following method is yielded by Conceptual Path-Following Scheme when the Augmented Complementary Slackness condition is represented as

$$G_\mu(X,S) := S + \mu\nabla K(X) = 0.$$

Passing to the representation

$$G\mu(X,S) := X + \mu\nabla K(S) = 0,$$

we arrive at the *Dual path-following method* with the same theoretical properties as those of the primal method. the Primal and the Dual path-following methods imply the best known so far complexity bounds for LP and SDP.

♠ In spite of being "theoretically perfect", Primal and Dual path-following methods in practice are inferior as compared with the methods based on less straightforward and more symmetric forms of the Augmented Complementary Slackness condition.

♠ The Augmented Complementary Slackness condition is

$$XS = SX = \mu I \qquad (*)$$

Fact: For $X, S \in \mathbf{S}^\nu_{++}$, $(*)$ is equivalent to

$$XS + SX = 2\mu I$$

Indeed, all we need to prove is that if $X, S \in \mathbf{S}^\nu_{++}$ and $XS + SX = 2\mu I$, then $XS = SX$. Under our premise we have $SX^2 = 2\mu X - XSX$ is symmetric $\Rightarrow$ $SX^2 = [SX^2]^T = X^2S$. Thus, S commutes with X^2 and therefore commutes with every polynomial of X^2, in particular with X (X is a polynomial of X^2 due to $X \succ 0$).

Fact: *Let $Q \in \mathbf{S}^\nu$ be nonsingular, and let $X, S \succ 0$. Then $XS = \mu I$ iff*

$$QXSQ^{-1} + Q^{-1}SXQ = 2\mu I$$

Indeed, it suffices to apply the previous fact to the matrices $\widehat{X} = QXQ \succ 0$, $\widetilde{S} = Q^{-1}SQ^{-1} \succ 0$.

♠ In practical path-following methods, at step i the Augmented Complementary Slackness condition is written down as

$$G_{\mu_{i+1}}(X, S) := Q_i X S Q_i^{-1} + Q_i^{-1} S X Q_i - 2\mu_{i+1} I = 0$$

with properly chosen varying from step to step nonsingular matrices $Q_i \in \mathbf{S}^\nu$.

Explanation: Let $Q \in \mathbf{S}^\nu$ be nonsingular. The *Q-scaling* $X \mapsto QXQ$ is a one-to-one linear mapping of $\mathbf{S}^\nu$ onto itself, the inverse being the mapping $X \mapsto Q^{-1}XQ^{-1}$. *Q-scaling is a symmetry of the positive semidefinite cone – it maps the cone onto itself.*

$\Rightarrow$ Given a primal-dual pair of semidefinite programs

$$\mathrm{Opt}(\mathcal{P}) = \min_X \left\{ \mathrm{Tr}(CX) : X \in [\mathcal{L}_P - B] \cap \mathbf{S}^\nu_+ \right\} \quad (\mathcal{P})$$
$$\mathrm{Opt}(\mathcal{D}) = \max_S \left\{ \mathrm{Tr}(BS) : S \in [\mathcal{L}_D + C] \cap \mathbf{S}^\nu_+ \right\} \quad (\mathcal{D})$$

and a nonsingular matrix $Q \in \mathbf{S}^\nu$, one can pass in $(\mathcal{P})$ from variable X to variables $\widehat{X} = QXQ$, while passing in $(\mathcal{D})$ from variable S to variable $\widetilde{S} = Q^{-1}SQ^{-1}$. The resulting problems are

$$\mathrm{Opt}(\mathcal{P}) = \min_{\widehat{X}} \left\{ \mathrm{Tr}(\widetilde{C}\widehat{X}) : \widehat{X} \in [\widehat{\mathcal{L}}_P - \widehat{B}] \cap \mathbf{S}^\nu_+ \right\} \quad (\widehat{\mathcal{P}})$$
$$\mathrm{Opt}(\mathcal{D}) = \max_{\widetilde{S}} \left\{ \mathrm{Tr}(\widehat{B}\widetilde{S}) : \widetilde{S} \in [\widetilde{\mathcal{L}}_D + \widetilde{C}] \cap \mathbf{S}^\nu_+ \right\} \quad (\widetilde{\mathcal{D}})$$
$$\left[\ \widehat{B} = QBQ, \widehat{\mathcal{L}}_P = \{QXQ : X \in \mathcal{L}_P\}, \widetilde{C} = Q^{-1}CQ^{-1}, \widetilde{\mathcal{L}}_D = \{Q^{-1}SQ^{-1} : S \in \mathcal{L}_D\} \ \right]$$

• $\widehat{\mathcal{P}}$ and $\widetilde{\mathcal{D}}$ are dual to each other, the primal-dual central path of this pair is the image of the primal-dual path of $(\mathcal{P})$, $(\mathcal{D})$ under the *primal-dual Q-scaling*

$$(X, S) \mapsto (\widehat{X} = QXQ, \widetilde{S} = Q^{-1}SQ^{-1})$$

$\mathcal{Q}$ preserves closeness to the path, etc.

Writing down the Augmented Complementary Slackness condition as

$$QXSQ^{-1} + Q^{-1}SXQ = 2\mu I \qquad (!)$$

we in fact

• pass from $(\mathcal{P})$, $(\mathcal{D})$ to the equivalent primal-dual pair of problems $(\widehat{\mathcal{P}})$, $(\widetilde{\mathcal{D}})$

• write down the Augmented Complementary Slackness condition for the latter pair in the simplest primal-dual symmetric form

$$\widehat{X}\widetilde{S} + \widetilde{S}\widehat{X} = 2\mu I,$$

• "scale back" to the original primal-dual variables X, S, thus arriving at (!).

Note: In the LP case $\mathbf{S}^\nu$ is composed of diagonal matrices, so that (!) is exactly the same as the "unscaled" condition $XS = \mu I$.

$$G_{\mu_{i+1}}(X,S) := Q_i X S Q_i^{-1} + Q_i^{-1} S X Q_i - 2\mu_{i+1} I = 0 \qquad (!)$$

With (!), the Newton system becomes

$$\Delta X \in \mathcal{L}_P,\ \Delta S \in \mathcal{L}_D$$
$$Q_i \Delta X S_i Q_i^{-1} + Q_i^{-1} S_i \Delta X Q_i + Q_i X_i \Delta S Q_i^{-1} + Q_i^{-1} \Delta S X_i Q_i$$
$$= 2\mu_{i+1} I - Q_i X_i S_i Q_i^{-1} - Q_i^{-1} S_i X_i Q_i$$

♣ Theoretical analysis of path-following methods simplifies a lot when the scaling (!) is *commutative*, meaning that the matrices $\widehat{X}_i = Q_i X_i Q_i$ and $\widehat{S}_i = Q_i^{-1} S_i Q_i^{-1}$ commute.
Popular choices of commutative scalings are:

- $Q_i = S_i^{1/2}$ ("XS-method," $\widetilde{S} = I$)
- $Q_i = X_i^{-1/2}$ ("SX-method, $\widehat{X} = I$)
- $Q_i = \left(X^{-1/2}(X^{1/2} S X^{1/2})^{-1/2} X^{1/2} S\right)^{1/2}$ (famous *Nesterov-Todd* method, $\widehat{X} = \widetilde{S}$).

$$\begin{array}{rc}
\mathrm{Opt}(P) = \min\limits_x \left\{ c^T x : \mathcal{A}x := \sum_{j=1}^n x_j A_j \succeq B \right\} & (P) \\
\Leftrightarrow \mathrm{Opt}(\mathcal{P}) = \min\limits_X \left\{ \mathrm{Tr}(CX) : X \in [\mathcal{L}_P - B] \cap \mathbf{S}^\nu_+ \right\} & (\mathcal{P}) \\
\mathrm{Opt}(D) = \max\limits_S \left\{ \mathrm{Tr}(BS) : S \in [\mathcal{L}_D + C] \cap \mathbf{S}^\nu_+ \right\} & (D) \\
\left[\mathcal{L}_P = \mathrm{Im}\mathcal{A},\, \mathcal{L}_D = \mathcal{L}_P^\perp \right]
\end{array}$$

Fact: *Let a strictly-feasible primal-dual pair (P), (D) of semidefinite programs be solved by a primal-dual path-following method based on commutative scalings. Assume that the method is initialized by a close to the path triple $(X_0, S_0, \mu_0 = \mathrm{Tr}(X_0 S_0)/m)$ and let the policy for updating μ be*

$$\mu_{i+1} = \left(1 - \frac{0.1}{\sqrt{m}}\right) \mu_i.$$

The the trajectory is well defined and stays close to the path.
As a result, every $O(\sqrt{m})$ steps of the method reduce duality gap by an absolute constant factor, and it takes $O(1)\sqrt{m} \ln\left(1 + \frac{m\mu_0}{\epsilon}\right)$ steps to make the duality gap $\leq \epsilon$.

♠ To improve the practical performance of primal-dual path-following methods, in actual computations

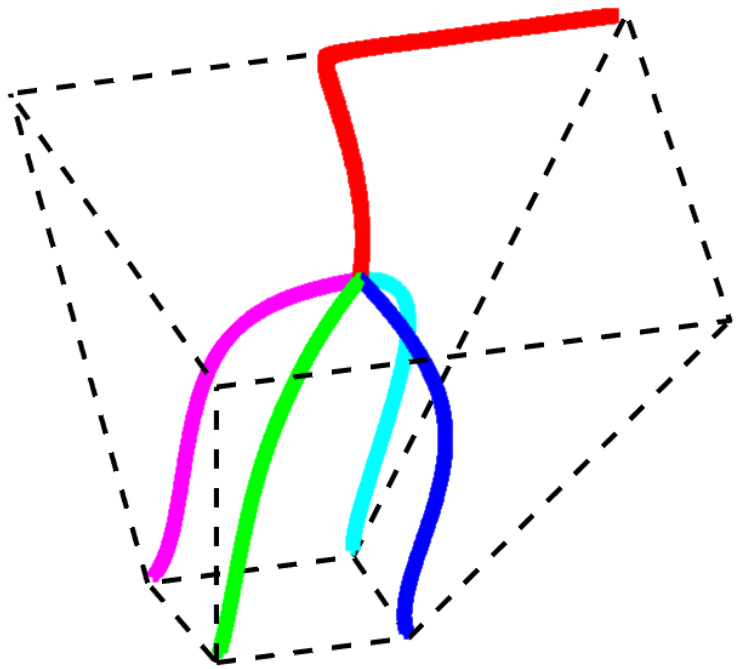
- the path parameter is updated in a more aggressive fashion than $\mu \mapsto \left(1 - \frac{0.1}{\sqrt{m}}\right)\mu$;
- the method is allowed to travel in a wider neighborhood of the primal-dual central path than the neighborhood given by our "close to the path" restriction $\mathrm{dist}(X, S, \mu) \leq 0.1$;
- instead of updating $X_{i+1} = X_i + \Delta X_i$, $S_{i+1} = S_i + \Delta S_i$, one uses the more flexible updating

$$X_{i+1} = X_i + \alpha_i \Delta X_i, \ S_{i+1} = S_i + \alpha_i \Delta S_i$$

with α_i given by appropriate line search.

♣ In practice, IPM's produce high accuracy solutions in few tens (like 30) iterations $\Rightarrow$ *practical scope of IPMs is restricted by our abilities to solve Newton systems.* Roughly peaking – *if the structure and size of the Newton system allows to assemple and solve the system in reasonable time, we are able to process the problem in reasonable time as well.*

♣ The constructions and the complexity results we have presented are incomplete — they do not take into account the necessity to come close to the central path before starting path-tracing and do not take care of the case when the pair (P), (D) is not strictly feasible. All these "gaps" can be easily closed via the same path-following technique as applied to appropriate augmented versions of the problem of interest.



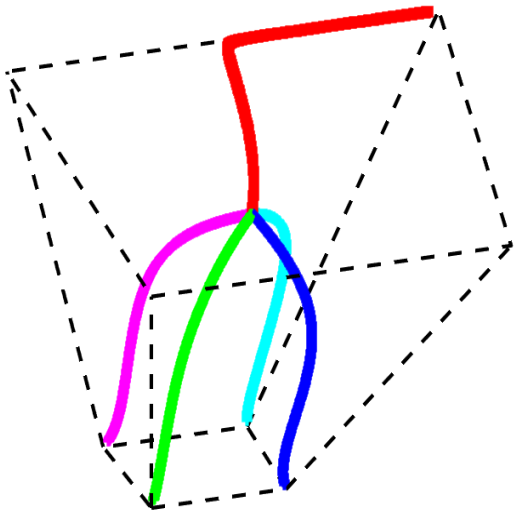
$$\min_x\{c^T x : x \in X\} \qquad [X = \{x \in \mathbf{R}^3 : Ax \le b\}]$$

Illustration: Primal path-following method on an LP $\min_x\{c^T x : x \in X\}$ with bounded feasible domain $X = \{x \in \mathbf{R}^3 : Ax \le b\}$.

- Log-barrier $K(y) = -\ln \mathrm{Det}\{\mathrm{Diag}\{y\}\}$ for nonnegative orthant induces barrier $F(x) = -\sum_i \ln([b - Ax]_i)$ for X.
- Every vector $g \in \mathbf{R}^3$ induces *central path*

$$x_g(\mu) = \operatorname*{argmin}_{x \in \mathrm{int}\, X}[g^T x + \mu F(x)]$$

As $\mu \to +0$, the path converges to a minimizer of $g^T x$ over X; as $\mu \to +\infty$, the path converges to the *analytic center* $x_F = \operatorname{argmin}_{x \in \mathrm{int}\, X} F(x)$ of X.

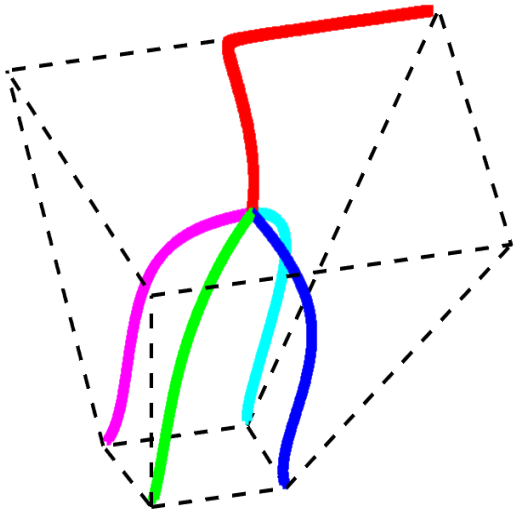


$\min_x\{c^T x : x \in X\}$ $[X = \{x \in \mathbf{R}^3 : Ax \leq b \in \mathbf{R}^6\}]$

♠ **In the primal space:**
Given a starting point $\bar{x} \in \operatorname{int} X$, we solve the LP of interest as follows:

• first, we trace the central path $x_g(\mu)$ given by $g = -\nabla F(\bar{x})$ *and push* μ *to* $+\infty$ in order to come close to x_F. The path x_g passes through $\bar{x} = x_g(1)$ $\Rightarrow$ no difficulties to start its tracing;

• after coming close to x_F, we come close to an easy-to-find point on the "path of interest" $x_c(\mu)$ and switch to tracing this path as $\mu \to +0$, thus approaching the optimal solution to the LP.

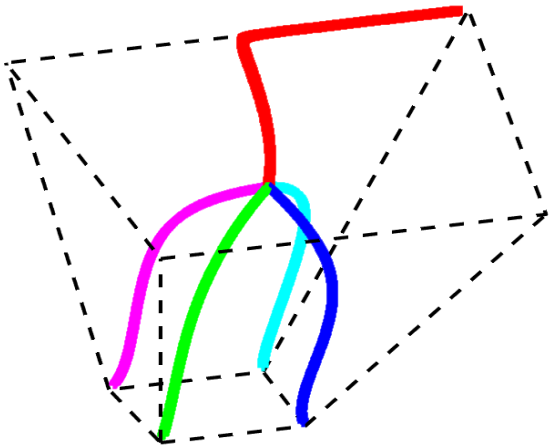


On the picture:

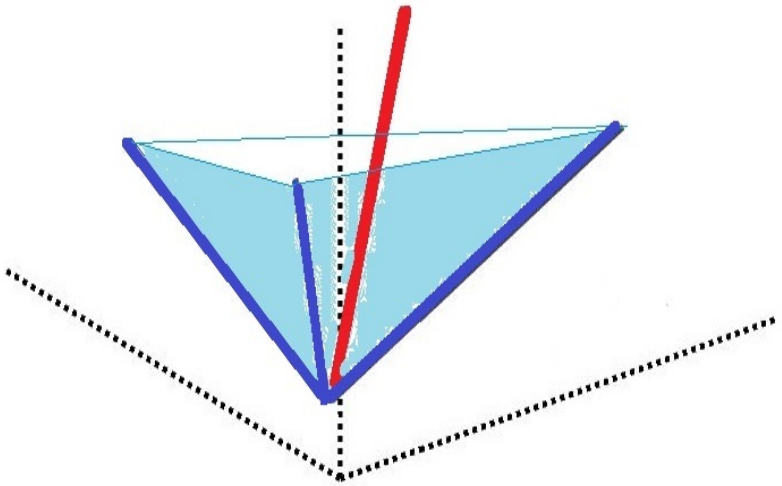
- magenta, green, cyan, blue: paths from 4 starting points to the analytic center
- red: path of interest — from the analytic center to the optimal solution.

Note: In our LP, there is a vertex of the feasible set with nearly the same value of the objective as at the optimal vertex. Were these values exactly equal, the path of interest would converge to something in-between these vertices – in LP with multiple optimal solutions, the central path converges to the analytic center of the optimal face, the barrier for the face being obtained from $F(x) = -\sum_i \ln([b - Ax]_i)$ by dropping out the terms which are $\equiv +\infty$ on the face). With the actual data, the path goes to this "in-between" point until it "feels" the difference between the "nearly optimal" and the optimal vertex and then makes a sharp turn and moves towards the optimal vertex.

♠ With Primal Path-following method, we first trace the "centering" path by increasing from iteration to iteration the path parameter by an appropriate factor, and then start to trace the path of interest, decreasing the parameter by the same factor.



primal paths



dual path

$$\min_x\{c^T x : x \in X\} \qquad [X = \{x \in \mathbf{R}^3 : Ax \leq b\}]$$

In dual space (right)

• When tracing the primal path leading from the analytic center of X to the optimal solution (red path on the left), we generate also the dual central path leading to the dual optimal solution.

• The dual central path lives in the intersection of the dual feasible plane

$$\mathcal{D} = \{\lambda : A^T\lambda = c\} \subset \mathbf{R}^6$$

of dimension 3 with the nonnegative orthant $\{\lambda \in \mathbf{R}^6 : \lambda \geq 0\}$. Projecting $\mathbf{R}^6_\lambda$ onto the space of the last three entries in λ (restricted on $\mathcal{D}$, this projection is a one-to-one mapping),
— $\mathbf{R}^6_+$ becomes the nonnegative orthant in $\mathbf{R}^3$ (extreme rays – black dotted rays))
— the dual feasible set becomes the shift (light blue) of some cone with three extreme rays
— the dual central path becomes the straight line (bold red) going from ∞ to the unique extreme point of the dual feasible set.