.

LECTURES
ON
MODERN CONVEX OPTIMIZATION

Arkadi Nemirovski
nemirovs@isye.gatech.edu
http://www.isye.gatech.edu/faculty-staff/profile.php?entry=an63

Department ISYE, Georgia Institute of Technology,

Fall Semester 005

Preface

Mathematical Programming deals with optimization programs of the form

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & \\ & g_i(x) \le 0,\ i = 1, ..., m, \\ & [x \subset \mathbf{R}^n] \end{array} \qquad \text{(P)}$$

and includes the following general areas:

1. *Modelling:* methodologies for posing various applied problems as optimization programs;
2. *Optimization Theory,* focusing on existence, uniqueness and on characterization of optimal solutions to optimization programs;
3. *Optimization Methods:* development and analysis of computational algorithms for various classes of optimization programs;
4. *Implementation, testing and application* of modelling methodologies and computational algorithms.

Essentially, Mathematical Programming was born in 1948, when George Dantzig has invented *Linear Programming* – the class of optimization programs (P) with linear objective $f(\cdot)$ and constraints $g_i(\cdot)$. This breakthrough discovery included

- the *methodological* idea that a natural desire of a human being to look for the best possible decisions can be posed in the form of an optimization program (P) and thus subject to mathematical and computational treatment;
- the *theory* of LP programs, primarily the LP duality (this is in part due to the great mathematician John von Neumann);
- the first *computational method* for LP – the Simplex method, which over the years turned out to be an extremely powerful computational tool.

As it often happens with first-rate discoveries (and to some extent is characteristic for such discoveries), today the above ideas and constructions look quite traditional and simple. Well, the same is with the wheel.

In 50 plus years since its birth, Mathematical Programming was rapidly progressing along all outlined avenues, "in width" as well as "in depth". I have no intention (and time) to trace the history of the subject decade by decade; instead, let me outline the major achievements in Optimization during the last 20 years or so, those which, I believe, allow to speak about *modern* optimization as opposed to the "classical" one as it existed circa 1980. The reader should be aware that the summary to follow is highly subjective and reflects the personal preferences of the author. Thus, *in my opinion* the major achievements in Mathematical Programming during last 15-20 years can be outlined as follows:

♠ Realizing what are the generic optimization programs one can solve well ("efficiently solvable" programs) and when such a possibility is, mildly speaking, problematic ("computationally intractable" programs). At this point, I do not intend to explain what does it mean exactly that "a generic optimization program is efficiently solvable"; we will arrive at this issue further in the course. However, I intend to answer the question (right now, not well posed!) "what are generic optimization programs we can solve well":

(!) *As far as numerical processing of programs* (P) *is concerned, there exists a "solvable case" – the one of* convex *optimization programs, where the objective f and the constraints g_i are convex functions.*

Under minimal additional "computability assumptions" (which are satisfied in basically all applications), a convex optimization program is "computationally tractable" – the computational effort required to solve the problem to a given accuracy "grows moderately" with the dimensions of the problem and the required number of accuracy digits.

In contrast to this, a general-type non-convex problems are too difficult for numerical solution – the computational effort required to solve such a problem by the best known so far numerical methods grows prohibitively fast with the dimensions of the problem and the number of accuracy digits, and there are serious theoretical reasons to guess that this is an intrinsic feature of non-convex problems rather than a drawback of the existing optimization techniques.

Just to give an example, consider a pair of optimization problems. The first is

$$\begin{array}{ll} \text{minimize} & -\sum_{i=1}^n x_i \\ \text{subject to} & \\ & \begin{array}{rcl} x_i^2 - x_i & = & 0,\ i = 1, ..., n; \\ x_i x_j & = & 0 \quad \forall (i,j) \in \Gamma, \end{array} \end{array} \tag{A}$$

Γ being a given set of pairs (i, j) of indices i, j. This is a fundamental combinatorial problem of computing the *stability number* of a graph; the corresponding "covering story" is as follows:

> Assume that we are given n letters which can be sent through a telecommunication channel, say, $n = 256$ usual bytes. When passing trough the channel, an input letter can be corrupted by errors; as a result, two distinct input letters can produce the same output and thus not necessarily can be distinguished at the receiving end. Let Γ be the set of "dangerous pairs of letters" – pairs (i, j) of distinct letters i, j which can be converted by the channel into the same output. If we are interested in error-free transmission, we should restrict the set S of letters we actually use to be *independent* – such that no pair (i, j) with $i, j \in S$ belongs to Γ. And in order to utilize best of all the capacity of the channel, we are interested to use a maximal – with maximum possible number of letters – independent sub-alphabet. It turns out that the minus optimal value in (A) is exactly the cardinality of such a maximal independent sub-alphabet.

Our second problem is

$$\begin{array}{ll} \text{minimize} & -2\sum_{i=1}^k \sum_{j=1}^m c_{ij} x_{ij} + x_{00} \\ \text{subject to} & \\ & \lambda_{\min}\left(\begin{bmatrix} x_1 & & & \sum_{j=1}^m b_{pj} x_{1j} \\ & \ddots & & \cdots \\ & & x_k & \sum_{j=1}^m b_{pj} x_{kj} \\ \sum_{j=1}^m b_{pj} x_{1j} & \cdots & \sum_{j=1}^m b_{pj} x_{kj} & x_{00} \end{bmatrix}\right) \geq 0, \\ & \qquad p = 1, ..., N, \\ & \qquad \sum_{i=1}^k x_i = 1, \end{array} \tag{B}$$

where $\lambda_{\min}(A)$ denotes the minimum eigenvalue of a symmetric matrix A. This problem is responsible for the design of a *truss* (a mechanical construction comprised of linked with each other thin elastic bars, like an electric mast, a bridge or the Eiffel Tower) capable to withstand best of all to k given loads.

When looking at the analytical forms of (A) and (B), it seems that the first problem is easier than the second: the constraints in (A) are simple explicit quadratic equations, while the constraints in (B) involve much more complicated functions of the design variables – the eigenvalues of certain matrices depending on the design vector. The truth, however, is that the first problem is, in a sense, "as difficult as an optimization problem can be", and

the worst-case computational effort to solve this problem within absolute inaccuracy 0.5 by all known optimization methods is about 2^n operations; for $n = 256$ (just 256 design variables corresponding to the "alphabet of bytes"), the quantity $2^n \approx 10^{77}$, for all practical purposes, is the same as $+\infty$. In contrast to this, the second problem is quite "computationally tractable". E.g., for $k = 6$ (6 loads of interest) and $m = 100$ (100 degrees of freedom of the construction) the problem has about 600 variables (twice the one of the "byte" version of (A)); however, it can be reliably solved within 6 accuracy digits in a couple of minutes. The dramatic difference in computational effort required to solve (A) and (B) finally comes from the fact that (A) is a non-convex optimization problem, while (B) is convex.

Note that realizing what is easy and what is difficult in Optimization is, aside of theoretical importance, extremely important methodologically. Indeed, mathematical models of real world situations in any case are incomplete and therefore are flexible to some extent. When you know in advance what you can process efficiently, you perhaps can use this flexibility to build a tractable (in our context – a convex) model. The "traditional" Optimization did not pay much attention to complexity and focused on easy-to-analyze *purely asymptotical* "rate of convergence" results. From this viewpoint, the most desirable property of f and g_i is smoothness (plus, perhaps, certain "nondegeneracy" at the optimal solution), and not their convexity; choosing between the above problems (A) and (B), a "traditional" optimizer would, perhaps, prefer the first of them. I suspect that a non-negligible part of "applied failures" of Mathematical Programming came from the traditional (I would say, heavily misleading) "order of preferences" in model-building. Surprisingly, some advanced users (primarily in Control) have realized the crucial role of convexity much earlier than some members of the Optimization community. Here is a real story. About 7 years ago, we were working on certain Convex Optimization method, and I sent an e-mail to people maintaining CUTE (a benchmark of test problems for constrained continuous optimization) requesting for the list of convex programs from their collection. The answer was: "We do not care which of our problems are convex, and this be a lesson for those developing Convex Optimization techniques." In their opinion, I am stupid; in my opinion, they are obsolete. Who is right, this I do not know...

♠ *Discovery of interior-point polynomial time methods for "well-structured" generic convex programs and throughout investigation of these programs.*

By itself, the "efficient solvability" of generic convex programs is a theoretical rather than a practical phenomenon. Indeed, assume that all we know about (P) is that the program is convex, its objective is called f, the constraints are called g_j and that we can compute f and g_i, along with their derivatives, at any given point at the cost of M arithmetic operations. In this case the computational effort for finding an ϵ-solution turns out to be *at least* $O(1)nM\ln(\frac{1}{\epsilon})$. Note that this is a *lower* complexity bound, and the best known so far *upper* bound is much worse: $O(1)n(n^3 + M)\ln(\frac{1}{\epsilon})$. Although the bounds grow "moderately" – polynomially – with the design dimension n of the program and the required number $\ln(\frac{1}{\epsilon})$ of accuracy digits, from the practical viewpoint the upper bound becomes prohibitively large already for n like 1000. This is in striking contrast with Linear Programming, where one can solve routinely problems with tens and hundreds of thousands of variables and constraints. The reasons for this huge difference come from the fact that

> *When solving an LP program, our a priory knowledge is far beyond the fact that the objective is called f, the constraints are called g_i, that they are convex and we can compute their values at derivatives at any given point. In LP, we know in advance what is the analytical structure of f and g_i, and we heavily exploit this knowledge when processing the problem.* In fact, all successful LP methods *never* never compute the values and the derivatives of f and g_i – they do something completely different.

One of the most important recent developments in Optimization is realizing the simple fact that *a jump from linear f and g_i's to "completely structureless" convex f and g_i's is too long: in-between these two extremes, there are many interesting and important generic convex programs*. These "in-between" programs, although non-linear, still possess nice analytical structure, and one can use this structure to develop dedicated optimization methods, the methods which turn out to be incomparably more efficient than those exploiting solely the convexity of the program.

The aforementioned "dedicated methods" are *Interior Point polynomial time algorithms*, and the most important "well-structured" generic convex optimization programs are those of *Linear*, *Conic Quadratic* and *Semidefinite Programming*; the last two entities merely did not exist as established research subjects just 15 years ago. In my opinion, *the discovery of Interior Point methods and of non-linear "well-structured" generic convex programs, along with the subsequent progress in these novel research areas, is one of the most impressive achievements in Mathematical Programming*.

♠ I have outlined the most revolutionary, in my appreciation, changes in the *theoretical core* of Mathematical Programming in the last 15-20 years. During this period, we have witnessed perhaps less dramatic, but still quite important progress in the methodological and application-related areas as well. The major novelty here is certain shift from the traditional for Operations Research applications in Industrial Engineering (production planning, etc.) to applications in "genuine" Engineering. I believe it is completely fair to say that the theory and methods of Convex Optimization, especially those of Semidefinite Programming, have become a kind of new paradigm in Control and are becoming more and more frequently used in Mechanical Engineering, Design of Structures, Medical Imaging, etc.

The aim of the course is to outline some of the novel research areas which have arisen in Optimization during the past decade or so. I intend to focus solely on Convex Programming, specifically, on

- *Conic Programming*, with emphasis on the most important particular cases – those of Linear, Conic Quadratic and Semidefinite Programming (LP, CQP and SDP, respectively).

  Here the focus will be on

  - basic Duality Theory for conic programs;
  - investigation of "expressive abilities" of CQP and SDP;
  - overview of the theory of Interior Point polynomial time methods for LP, CQP and SDP.

- *"Efficient* (polynomial time) *solvability"* of generic convex programs.

- *"Low cost"* optimization methods for extremely large-scale optimization programs.

Acknowledgements. The first four lectures of the five comprising the core of the course are based upon the recent book

Ben-Tal, A., Nemirovski, A., *Lectures on Modern Convex Optimization: Analysis, Algorithms, Engineering Applications*, MPS-SIAM Series on Optimization, SIAM, Philadelphia, 2001.

I am greatly indebted to my colleagues, primarily to Yuri Nesterov, Aharon Ben-Tal, Stephen Boyd, Claude Lemarechal and Kees Roos, who over the years have influenced significantly my

understanding of our subject as expressed in this course. Needless to say, I am the only person responsible for the drawbacks in what follows.

Arkadi Nemirovski,
Haifa, Israel, May 2002
Atlanta, USA, August 2003
Atlanta, USA, August 2005

Contents

1 From Linear to Conic Programming **13**
1.1 Linear programming: basic notions . . . 13
1.2 Duality in linear programming . . . 14
1.2.1 Certificates for solvability and insolvability . . . 14
1.2.2 Dual to an LP program: the origin . . . 18
1.2.3 The LP Duality Theorem . . . 21
1.3 From Linear to Conic Programming . . . 23
1.4 Orderings of $\mathbf{R}^m$ and cones . . . 23
1.5 "Conic programming" – what is it? . . . 26
1.6 Conic Duality . . . 27
1.6.1 Geometry of the primal and the dual problems . . . 29
1.7 Conic Duality Theorem . . . 32
1.7.1 Is something wrong with conic duality? . . . 35
1.7.2 Consequences of the Conic Duality Theorem . . . 37
1.8 Exercises . . . 42
1.8.1 Around General Theorem on Alternative . . . 42
1.8.2 Around cones . . . 43
1.8.3 Around conic problems . . . 46
1.8.4 Feasible and level sets of conic problems . . . 46

2 Conic Quadratic Programming **49**
2.1 Conic Quadratic problems: preliminaries . . . 49
2.2 Examples of conic quadratic problems . . . 51
2.2.1 Contact problems with static friction [11] . . . 51
2.3 What can be expressed via conic quadratic constraints? . . . 53
2.3.1 More examples of CQ-representable functions/sets . . . 68
2.4 More applications: Robust Linear Programming . . . 71
2.4.1 Robust Linear Programming: the paradigm . . . 72
2.4.2 Robust Linear Programming: examples . . . 73
2.4.3 Robust counterpart of uncertain LP with a CQr uncertainty set . . . 83
2.4.4 CQ-representability of the optimal value in a CQ program as a function of the data . . . 86
2.4.5 Affinely Adjustable Robust Counterpart . . . 87
2.5 Does Conic Quadratic Programming exist? . . . 95
2.6 Exercises . . . 99
2.6.1 Around randomly perturbed linear constraints . . . 99

2.6.2 Around Robust Antenna Design . . . 101

3 Semidefinite Programming 105
3.1 Semidefinite cone and Semidefinite programs . . . 105
3.1.1 Preliminaries . . . 105
3.2 What can be expressed via LMI's? . . . 108
3.3 Applications of Semidefinite Programming in Engineering . . . 122
3.3.1 Dynamic Stability in Mechanics . . . 122
3.3.2 Design of chips and Boyd's time constant . . . 124
3.3.3 Lyapunov stability analysis/synthesis . . . 126
3.4 Semidefinite relaxations of intractable problems . . . 133
3.4.1 Semidefinite relaxations of combinatorial problems . . . 133
3.4.2 Matrix Cube Theorem and interval stability analysis/synthesis . . . 145
3.4.3 Robust Quadratic Programming . . . 152
3.5 $\mathcal{S}$-Lemma and Approximate $\mathcal{S}$-Lemma . . . 155
3.5.1 $\mathcal{S}$-Lemma . . . 155
3.5.2 Inhomogeneous $\mathcal{S}$-Lemma . . . 158
3.5.3 Approximate $\mathcal{S}$-Lemma . . . 159
3.6 Extremal ellipsoids . . . 162
3.6.1 Ellipsoidal approximations of unions/intersections of ellipsoids . . . 166
3.6.2 Approximating sums of ellipsoids . . . 168
3.7 Exercises . . . 178
3.7.1 Around positive semidefiniteness, eigenvalues and $\succeq$-ordering . . . 178
3.7.2 SD representations of epigraphs of convex polynomials . . . 187
3.7.3 Around the Lovasz capacity number and semidefinite relaxations of combinatorial problems . . . 189
3.7.4 Around Lyapunov Stability Analysis . . . 194
3.7.5 Around Nesterov's $\frac{\pi}{2}$ Theorem . . . 195
3.7.6 Around ellipsoidal approximations . . . 196

4 Polynomial Time Interior Point algorithms for LP, CQP and SDP 201
4.1 Complexity of Convex Programming . . . 201
4.1.1 Combinatorial Complexity Theory . . . 201
4.1.2 Complexity in Continuous Optimization . . . 203
4.1.3 Computational tractability of convex optimization problems . . . 204
4.1.4 "What is inside" Theorem 4.1.1: Black-box represented convex programs and the Ellipsoid method . . . 206
4.1.5 Difficult continuous optimization problems . . . 215
4.2 Interior Point Polynomial Time Methods for LP, CQP and SDP . . . 215
4.2.1 Motivation . . . 215
4.2.2 Interior Point methods . . . 216
4.2.3 But... . . . 219
4.3 Interior point methods for LP, CQP, and SDP: building blocks . . . 220
4.3.1 Canonical cones and canonical barriers . . . 220
4.3.2 Elementary properties of canonical barriers . . . 222
4.4 Primal-dual pair of problems and primal-dual central path . . . 224
4.4.1 The problem(s) . . . 224

4.4.2 The central path(s) . . . 224
4.5 Tracing the central path . . . 230
4.5.1 The path-following scheme . . . 230
4.5.2 Speed of path-tracing . . . 232
4.5.3 The primal and the dual path-following methods . . . 232
4.5.4 The SDP case . . . 235
4.6 Complexity bounds for LP, CQP, SDP . . . 248
4.6.1 Complexity of $\mathcal{LP}_b$ . . . 248
4.6.2 Complexity of $\mathcal{CQP}_b$ . . . 249
4.6.3 Complexity of $\mathcal{SDP}_b$ . . . 249
4.7 Concluding remarks . . . 250
4.8 Exercises: Around the Ellipsoid method . . . 252

5 Simple methods for extremely large-scale problems 257
5.1 Motivation . . . 257
5.2 Information-based complexity of Convex Programming . . . 259
5.3 Methods with Euclidean geometry: Subgradient Descent and Bundle-Level . . . 262
5.3.1 The simplest of the cheapest – Subgradient Descent . . . 262
5.3.2 From SD to Bundle-Level: Adding memory . . . 265
5.3.3 Restricted Memory Bundle-Level . . . 268
5.4 The Bundle-Mirror scheme . . . 273
5.4.1 Mirror Descent – Building Blocks . . . 273
5.4.2 Non-Euclidean SD – Mirror Descent . . . 275
5.4.3 Mirror-Level Algorithm . . . 281
5.4.4 NERML – Non-Euclidean Restricted Memory Level algorithm . . . 284
5.5 Implementation issues and illustrations . . . 287
5.5.1 Implementing SD and MD . . . 287
5.5.2 Illustration: PET Image Reconstruction problem . . . 290
5.6 Appendix: strong convexity of $\omega(\cdot)$ for standard setups . . . 296

Bibliography 299

A Prerequisites from Linear Algebra and Analysis 301
A.1 Space $\mathbf{R}^n$: algebraic structure . . . 301
A.1.1 A point in $\mathbf{R}^n$ . . . 301
A.1.2 Linear operations . . . 301
A.1.3 Linear subspaces . . . 302
A.1.4 Linear independence, bases, dimensions . . . 303
A.1.5 Linear mappings and matrices . . . 304
A.2 Space $\mathbf{R}^n$: Euclidean structure . . . 306
A.2.1 Euclidean structure . . . 306
A.2.2 Inner product representation of linear forms on $\mathbf{R}^n$ . . . 307
A.2.3 Orthogonal complement . . . 307
A.2.4 Orthonormal bases . . . 308
A.3 Affine subspaces in $\mathbf{R}^n$ . . . 310
A.3.1 Affine subspaces and affine hulls . . . 310
A.3.2 Intersections of affine subspaces, affine combinations and affine hulls . . . 311

A.3.3 Affinely spanning sets, affinely independent sets, affine dimension 312
A.3.4 Dual description of linear subspaces and affine subspaces 314
A.3.5 Structure of the simplest affine subspaces 316
A.4 Space $\mathbf{R}^n$: metric structure and topology 317
A.4.1 Euclidean norm and distances 317
A.4.2 Convergence 318
A.4.3 Closed and open sets 319
A.4.4 Local compactness of $\mathbf{R}^n$ 320
A.5 Continuous functions on $\mathbf{R}^n$ 320
A.5.1 Continuity of a function 320
A.5.2 Elementary continuity-preserving operations 321
A.5.3 Basic properties of continuous functions on $\mathbf{R}^n$ 321
A.6 Differentiable functions on $\mathbf{R}^n$ 322
A.6.1 The derivative 322
A.6.2 Derivative and directional derivatives 324
A.6.3 Representations of the derivative 325
A.6.4 Existence of the derivative 326
A.6.5 Calculus of derivatives 327
A.6.6 Computing the derivative 327
A.6.7 Higher order derivatives 329
A.6.8 Calculus of C^k mappings 331
A.6.9 Examples of higher-order derivatives 332
A.6.10 Taylor expansion 333
A.7 Symmetric matrices 334
A.7.1 Spaces of matrices 334
A.7.2 Main facts on symmetric matrices 335
A.7.3 Variational characterization of eigenvalues 336
A.7.4 Positive semidefinite matrices and the semidefinite cone 339

B Convex sets in $\mathbf{R}^n$ 343
B.1 Definition and basic properties 343
B.1.1 A convex set 343
B.1.2 Examples of convex sets 343
B.1.3 Inner description of convex sets: Convex combinations and convex hull . . 346
B.1.4 Cones 347
B.1.5 "Calculus" of convex sets 348
B.1.6 Topological properties of convex sets 348
B.2 Main theorems on convex sets 352
B.2.1 Caratheodory Theorem 352
B.2.2 Radon Theorem 353
B.2.3 Helley Theorem 354
B.2.4 Homogeneous Farkas Lemma 355
B.2.5 Separation Theorem 357
B.2.6 Polar of a convex set and Milutin-Dubovitski Lemma 363
B.2.7 Extreme points and Krein-Milman Theorem 366
B.2.8 Structure of polyhedral sets 369

C Convex functions **379**
C.1 Convex functions: first acquaintance . . . 379
C.1.1 Definition and Examples . . . 379
C.1.2 Elementary properties of convex functions . . . 380
C.1.3 What is the value of a convex function outside its domain? . . . 381
C.2 How to detect convexity . . . 382
C.2.1 Operations preserving convexity of functions . . . 382
C.2.2 Differential criteria of convexity . . . 384
C.3 Gradient inequality . . . 387
C.4 Boundedness and Lipschitz continuity of a convex function . . . 388
C.5 Maxima and minima of convex functions . . . 391
C.6 Subgradients and Legendre transformation . . . 395
C.6.1 Proper functions and their representation . . . 395
C.6.2 Subgradients . . . 401
C.6.3 Legendre transformation . . . 402

D Convex Programming, Lagrange Duality, Saddle Points **407**
D.1 Mathematical Programming Program . . . 407
D.2 Convex Programming program and Lagrange Duality Theorem . . . 408
D.2.1 Convex Theorem on Alternative . . . 408
D.2.2 Lagrange Function and Lagrange Duality . . . 411
D.2.3 Optimality Conditions in Convex Programming . . . 413
D.3 Saddle Points . . . 417
D.3.1 Definition and Game Theory interpretation . . . 417
D.3.2 Existence of Saddle Points . . . 419

Lecture 1

From Linear to Conic Programming

1.1 Linear programming: basic notions

A *Linear Programming* (LP) program is an optimization program of the form

$$\min\left\{c^T x \;\middle|\; Ax \geq b\right\}, \tag{LP}$$

where

- $x \in \mathbf{R}^n$ is the *design vector*
- $c \in \mathbf{R}^n$ is a given vector of coefficients of the *objective function* $c^T x$
- A is a given $m \times n$ *constraint matrix*, and $b \in \mathbf{R}^m$ is a given *right hand side of the constraints*.

(LP) is called

– *feasible*, if its *feasible set*

$$\mathcal{F} = \{x \mid Ax - b \geq 0\}$$

is nonempty; a point $x \in \mathcal{F}$ is called a *feasible solution* to (LP);

– *bounded below*, if it is either infeasible, or its objective $c^T x$ is bounded below on $\mathcal{F}$.

For a feasible bounded below problem (LP), the quantity

$$c^* \equiv \inf_{x: Ax-b\geq 0} c^T x$$

is called the *optimal value* of the problem. For an infeasible problem, we set $c_* = +\infty$, while for feasible *un*bounded below problem we set $c_* = -\infty$.

(LP) is called *solvable*, if it is feasible, bounded below and the optimal value is attained, i.e., there exists $x \in \mathcal{F}$ with $c^T x = c^*$. An x of this type is called an *optimal solution* to (LP).

A priori it is unclear whether a feasible and bounded below LP program is solvable: why should the infimum be achieved? It turns out, however, that a feasible and bounded below program (LP) *always* is solvable. This nice fact (we shall establish it later) is specific for LP. Indeed, a very simple *nonlinear* optimization program

$$\min\left\{\frac{1}{x} \;\middle|\; x \geq 1\right\}$$

is feasible and bounded below, but it is not solvable.

1.2 Duality in linear programming

The most important and interesting feature of linear programming as a mathematical entity (i.e., aside of computations and applications) is the wonderful *LP duality theory* we are about to consider. We motivate this topic by first addressing the following question:

Given an LP program

$$c^* = \min_x \left\{ c^T x \,\middle|\, Ax - b \geq 0 \right\}, \tag{LP}$$

how to find a systematic way to bound from below its optimal value c^* ?

Why this is an important question, and how the answer helps to deal with LP, this will be seen in the sequel. For the time being, let us just believe that the question is worthy of the effort.

A trivial answer to the posed question is: solve (LP) and look what is the optimal value. There is, however, a smarter and a much more instructive way to answer our question. Just to get an idea of this way, let us look at the following example:

$$\min \left\{ x_1 + x_2 + ... + x_{2002} \,\middle|\, \begin{array}{rcl} x_1 + 2x_2 + ... + 2001x_{2001} + 2002x_{2002} - 1 & \geq & 0, \\ 2002x_1 + 2001x_2 + ... + 2x_{2001} + x_{2002} - 100 & \geq & 0, \\ ..... & ... & ... \end{array} \right\}.$$

We claim that the optimal value in the problem is $\geq \frac{101}{2003}$. How could one certify this bound? This is immediate: add the first two constraints to get the inequality

$$2003(x_1 + x_2 + ... + x_{1998} + x_{2002}) - 101 \geq 0,$$

and divide the resulting inequality by 2003. LP duality is nothing but a straightforward generalization of this simple trick.

1.2.1 Certificates for solvability and insolvability

Consider a (finite) system of scalar inequalities with n unknowns. To be as general as possible, we do not assume for the time being the inequalities to be linear, and we allow for both non-strict and strict inequalities in the system, as well as for equalities. Since an equality can be represented by a pair of non-strict inequalities, our system can always be written as

$$f_i(x) \ \Omega_i \ 0, \ i = 1, ..., m, \tag{$\mathcal{S}$}$$

where every Ω_i is either the relation " $>$ " or the relation " $\geq$ ".

The basic question about $(\mathcal{S})$ is

(?) Whether $(\mathcal{S})$ has a solution or not.

Knowing how to answer the question (?), we are able to answer many other questions. E.g., to verify whether a given real a is a lower bound on the optimal value c^* of (LP) is the same as to verify whether the system

$$\left\{ \begin{array}{rcl} -c^T x + a & > & 0 \\ Ax - b & \geq & 0 \end{array} \right.$$

has no solutions.

The general question above is too difficult, and it makes sense to pass from it to a seemingly simpler one:

(??) How to certify that $(\mathcal{S})$ has, or does not have, a solution.

Imagine that you are very smart and know the correct answer to (?); how could you convince somebody that your answer is correct? What could be an "evident for everybody" certificate of the validity of your answer?

If your claim is that $(\mathcal{S})$ is solvable, a certificate could be just to point out a solution x^* to $(\mathcal{S})$. Given this certificate, one can substitute x^* into the system and check whether x^* indeed is a solution.

Assume now that your claim is that $(\mathcal{S})$ has no solutions. What could be a "simple certificate" of this claim? How one could certify a *negative* statement? This is a highly nontrivial problem not just for mathematics; for example, in criminal law: how should someone accused in a murder prove his innocence? The "real life" answer to the question "how to certify a negative statement" is discouraging: such a statement normally *cannot* be certified (this is where the rule "a person is presumed innocent until proven guilty" comes from). In mathematics, however, the situation is different: in some cases there exist "simple certificates" of negative statements. E.g., in order to certify that $(\mathcal{S})$ has no solutions, it suffices to demonstrate that a consequence of $(\mathcal{S})$ is a contradictory inequality such as

$$-1 \geq 0.$$

For example, assume that λ_i, $i = 1, ..., m$, are nonnegative weights. Combining inequalities from $(\mathcal{S})$ with these weights, we come to the inequality

$$\sum_{i=1}^{m} \lambda_i f_i(x) \ \Omega \ 0 \qquad (\mathrm{Cons}(\lambda))$$

where Ω is either " $>$ " (this is the case when the weight of at least one strict inequality from $(\mathcal{S})$ is positive), or " $\geq$ " (otherwise). Since the resulting inequality, due to its origin, is a consequence of the system $(\mathcal{S})$, i.e., it is satisfied by every solution to $\mathcal{S}$), it follows that if $(\mathrm{Cons}(\lambda))$ has no solutions at all, we can be sure that $(\mathcal{S})$ has no solution. Whenever this is the case, we may treat the corresponding vector λ as a "simple certificate" of the fact that $(\mathcal{S})$ is infeasible.

Let us look what does the outlined approach mean when $(\mathcal{S})$ is comprised of *linear* inequalities:

$$(\mathcal{S}): \quad \{a_i^T x \ \Omega_i \ b_i, \ i = 1, ..., m\} \quad \left[\Omega_i = \begin{cases} " > " \\ " \geq " \end{cases}\right]$$

Here the "combined inequality" is linear as well:

$$(\mathrm{Cons}(\lambda)): \qquad (\sum_{i=1}^{m} \lambda a_i)^T x \ \Omega \ \sum_{i=1}^{m} \lambda b_i$$

(Ω is " $>$ " whenever $\lambda_i > 0$ for at least one i with $\Omega_i =$ " $>$ ", and Ω is " $\geq$ " otherwise). Now, when can a *linear* inequality

$$d^T x \ \Omega \ e$$

be contradictory? Of course, it can happen only when $d = 0$. Whether in this case the inequality is contradictory, it depends on what is the relation Ω: if $\Omega =$ " $>$ ", then the inequality is contradictory if and only if $e \geq 0$, and if $\Omega =$ " $\geq$ ", it is contradictory if and only if $e > 0$. We have established the following simple result:

Proposition 1.2.1 *Consider a system of linear inequalities*

$$(\mathcal{S}): \quad \begin{cases} a_i^T x > b_i, \ i = 1, ..., m_{\rm s}, \\ a_i^T x \geq b_i, \ i = m_{\rm s}+1, ..., m. \end{cases}$$

with n-dimensional vector of unknowns x. Let us associate with $(\mathcal{S})$ two systems of linear inequalities and equations with m-dimensional vector of unknowns λ:

$$\mathcal{T}_{\rm I}: \quad \begin{cases} (a) \quad \lambda \geq 0; \\ (b) \quad \sum_{i=1}^m \lambda_i a_i = 0; \\ (c_{\rm I}) \quad \sum_{i=1}^m \lambda_i b_i \geq 0; \\ \hline (d_{\rm I}) \quad \sum_{i=1}^{m_{\rm s}} \lambda_i > 0. \end{cases}$$

$$\mathcal{T}_{\rm II}: \quad \begin{cases} (a) \quad \lambda \geq 0; \\ (b) \quad \sum_{i=1}^m \lambda_i a_i = 0; \\ \hline (c_{\rm II}) \quad \sum_{i=1}^m \lambda_i b_i > 0. \end{cases}$$

Assume that at least one of the systems $\mathcal{T}_{\rm I}$, $\mathcal{T}_{\rm II}$ is solvable. Then the system $(\mathcal{S})$ is infeasible.

Proposition 1.2.1 says that in some cases it is easy to certify infeasibility of a linear system of inequalities: a "simple certificate" is a solution to another system of linear inequalities. Note, however, that the existence of a certificate of this latter type is to the moment only a *sufficient*, but not a *necessary*, condition for the infeasibility of $(\mathcal{S})$. A fundamental result in the theory of linear inequalities is that the sufficient condition in question is in fact also necessary:

Theorem 1.2.1 [General Theorem on Alternative] *In the notation from Proposition 1.2.1, system $(\mathcal{S})$ has no solutions if and only if either $\mathcal{T}_{\rm I}$, or $\mathcal{T}_{\rm II}$, or both these systems, are solvable.*

There are numerous proofs of the Theorem on Alternative; in my taste, the most instructive one is to reduce the Theorem to its particular case – the *Homogeneous Farkas Lemma*:

> [Homogeneous Farkas Lemma] *A homogeneous nonstrict linear inequality*
>
> $$a^T x \leq 0$$
>
> *is a consequence of a system of homogeneous nonstrict linear inequalities*
>
> $$a_i^T x \leq 0, \ i = 1, ..., m$$
>
> *if and only if it can be obtained from the system by taking weighted sum with nonnegative weights:*
>
> $$\begin{array}{ll} (a) & a_i^T x \leq 0, \ i = 1, ..., m \Rightarrow a^T x \leq 0, \\ & \qquad\qquad \Updownarrow \\ (b) & \qquad \exists \lambda_i \geq 0 : a = \sum_i \lambda_i a_i. \end{array} \qquad (1.2.1)$$

The reduction of GTA to HFL is easy. As about the HFL, there are, essentially, two ways to prove the statement:

- The "quick and dirty" one based on separation arguments (see Section B.2.5 and/or Exercise B.13), which is as follows:

1. First, we demonstrate that if A is a nonempty closed convex set in $\mathbf{R}^n$ and a is a point from $\mathbf{R}^n \backslash A$, then a can be *strongly separated* from A by a linear form: there exists $x \in \mathbf{R}^n$ such that
$$x^T a < \inf_{b \in A} x^T b. \tag{1.2.2}$$
To this end, it suffices to verify that
(a) In A, there exists a point closest to a w.r.t. the standard Euclidean norm $\|b\|_2 = \sqrt{b^T b}$, i.e., that the optimization program
$$\min_{b \in A} \|a - b\|_2$$
has a solution b_*;
(b) Setting $x = b_* - a$, one ensures (1.2.2).
Both (a) and (b) are immediate.
2. Second, we demonstrate that *the set*
$$A = \{b : \exists \lambda \geq 0 : b = \sum_{i=1}^m \lambda_i a_i\}$$
– *the cone spanned by the vectors* $a_1, ..., a_m$ – *is convex* (which is immediate) *and closed* (the proof of this crucial fact also is not difficult).
3. Combining the above facts, we immediately see that
— either $a \in A$, i.e., (1.2.1.b) holds,
— or there exists x such that $x^T a < \inf_{\lambda \geq 0} x^T \sum_i \lambda_i a_i$.
The latter inf is finite if and only if $x^T a_i \geq 0$ for all i, and in this case the inf is 0, so that the "or" statement says exactly that there exists x with $a_i^T x \geq 0$, $a^T x < 0$, or, which is the same, that (1.2.1.a) *does not* hold.
Thus, among the statements (1.2.1.a) and the negation of (1.2.1.b) at least one (and, as it is immediately seen, *at most* one as well) always is valid, which is exactly the equivalence (1.2.1).

- "Advanced" proofs based purely on Linear Algebra facts (see Section B.2.4). The advantage of these purely Linear Algebra proofs is that they, in contrast to the outlined separation-based proof, do not use the completeness of $\mathbf{R}^n$ as a metric space and thus work when we pass from systems with *real* coefficients and unknowns to systems with *rational* (or algebraic) coefficients. As a result, an advanced proof allows to establish the Theorem on Alternative for the case when the coefficients *and* unknowns in $(\mathcal{S})$, $\mathcal{T}_I$, $\mathcal{T}_{II}$ are restricted to belong to a given "real field" (e.g., are rational).

We formulate here explicitly two very useful principles following from the Theorem on Alternative:

A. *A system of linear inequalities*
$$a_i^T x \ \Omega_i \ b_i, \ i = 1, ..., m$$
has no solutions if and only if one can combine the inequalities of the system in a <u>linear</u> *fashion (i.e., multiplying the inequalities by nonnegative weights, adding the results and passing, if necessary, from an inequality* $a^T x > b$ *to the inequality* $a^T x \geq b$*) to get a contradictory inequality, namely, either the inequality* $0^T x \geq 1$*, or the inequality* $0^T x > 0$*.*

B. *A linear inequality*
$$a_0^T x \ \Omega_0 \ b_0$$

is a consequence of a solvable system of linear inequalities

$$a_i^T x \ \Omega_i \ b_i, \ i = 1, ..., m$$

if and only if it can be obtained by combining, in a linear fashion, the inequalities of the system and the trivial inequality $0 > -1$.

It should be stressed that the above principles are highly nontrivial and very deep. Consider, e.g., the following system of 4 linear inequalities with two variables u, v:

$$\begin{array}{c} -1 \le u \le 1 \\ -1 \le v \le 1. \end{array}$$

From these inequalities it follows that

$$u^2 + v^2 \le 2, \qquad (!)$$

which in turn implies, by the Cauchy inequality, the linear inequality $u + v \le 2$:

$$u + v = 1 \times u + 1 \times v \le \sqrt{1^2 + 1^2}\sqrt{u^2 + v^2} \le (\sqrt{2})^2 = 2. \qquad (!!)$$

The concluding inequality is linear and is a consequence of the original system, but in the demonstration of this fact both steps (!) and (!!) are "highly nonlinear". It is absolutely unclear a priori why the same consequence can, as it is stated by Principle **A**, be derived from the system in a linear manner as well [of course it can – it suffices just to add two inequalities $u \le 1$ and $v \le 1$].

Note that the Theorem on Alternative and its corollaries **A** and **B** heavily exploit the fact that we are speaking about *linear* inequalities. E.g., consider the following 2 quadratic and 2 linear inequalities with two variables:

$$\begin{array}{lrcl} (a) & u^2 & \ge & 1; \\ (b) & v^2 & \ge & 1; \\ (c) & u & \ge & 0; \\ (d) & v & \ge & 0; \end{array}$$

along with the quadratic inequality

$$(e) \quad uv \ge 1.$$

The inequality (e) is clearly a consequence of (a) – (d). However, if we extend the system of inequalities (a) – (b) by all "trivial" (i.e., identically true) linear and quadratic inequalities with 2 variables, like $0 > -1$, $u^2 + v^2 \ge 0$, $u^2 + 2uv + v^2 \ge 0$, $u^2 - uv + v^2 \ge 0$, etc., and ask whether (e) can be derived in a *linear* fashion from the inequalities of the extended system, the answer will be negative. Thus, Principle **A** fails to be true already for quadratic inequalities (which is a great sorrow – otherwise there were no difficult problems at all!)

We are about to use the Theorem on Alternative to obtain the basic results of the LP duality theory.

1.2.2 Dual to an LP program: the origin

As already mentioned, the motivation for constructing the problem dual to an LP program

$$c^* = \min_x \left\{ c^T x \,\middle|\, Ax - b \ge 0 \right\} \quad \left[A = \begin{bmatrix} a_1^T \\ a_2^T \\ ... \\ a_m^T \end{bmatrix} \in \mathbf{R}^{m \times n} \right] \qquad \text{(LP)}$$

is the desire to generate, in a systematic way, lower bounds on the optimal value c^* of (LP). An evident way to bound from below a given function $f(x)$ in the domain given by system of inequalities

$$g_i(x) \geq b_i,\ i = 1, ..., m, \tag{1.2.3}$$

is offered by what is called the *Lagrange duality* and is as follows:

Lagrange Duality:

• *Let us look at all inequalities which can be obtained from* (1.2.3) *by linear aggregation, i.e., at the inequalities of the form*

$$\sum_i y_i g_i(x) \geq \sum_i y_i b_i \tag{1.2.4}$$

with the "aggregation weights" $y_i \geq 0$. Note that the inequality (1.2.4)*, due to its origin, is valid on the entire set X of solutions of* (1.2.3)*.*

• *Depending on the choice of aggregation weights, it may happen that the left hand side in* (1.2.4) *is $\leq f(x)$ for all $x \in \mathbf{R}^n$. Whenever it is the case, the right hand side $\sum_i y_i b_i$ of* (1.2.4) *is a lower bound on f in X.*

Indeed, on X the quantity $\sum_i y_i b_i$ is a lower bound on $y_i g_i(x)$, and for y in question the latter function of x is everywhere $\leq f(x)$.

It follows that

• *The optimal value in the problem*

$$\max_y \left\{ \sum_i y_i b_i : \begin{array}{cc} y \geq 0, & (a) \\ \sum_i y_i g_i(x) \leq f(x)\ \forall x \in \mathbf{R}^n & (b) \end{array} \right\} \tag{1.2.5}$$

is a lower bound on the values of f on the set of solutions to the system (1.2.3)*.*

Let us look what happens with the Lagrange duality when f and g_i are homogeneous linear functions: $f = c^T x$, $g_i(x) = a_i^T x$. In this case, the requirement (1.2.5.b) merely says that $c = \sum_i y_i a_i$ (or, which is the same, $A^T y = c$ due to the origin of A). Thus, problem (1.2.5) becomes the Linear Programming problem

$$\max_y \left\{ b^T y : A^T y = c,\ y \geq 0 \right\}, \tag{LP*}$$

which is nothing but the LP dual of (LP).

By the construction of the dual problem,

[Weak Duality] *The optimal value in* (LP*) *is less than or equal to the optimal value in* (LP)*.*

In fact, the "less than or equal to" in the latter statement is "equal", provided that the optimal value c^* in (LP) is a number (i.e., (LP) is feasible and below bounded). To see that this indeed is the case, note that a real a is a lower bound on c^* if and only if $c^T x \geq a$ whenever $Ax \geq b$, or, which is the same, if and only if the system of linear inequalities

$$(\mathcal{S}_a): \qquad -c^T x > -a, Ax \geq b$$

has no solution. We know by the Theorem on Alternative that the latter fact means that some other system of linear equalities (more exactly, at least one of a certain pair of systems) does have a solution. More precisely,

(*) ($\mathcal{S}_a$) *has no solutions if and only if at least one of the following two systems with* $m+1$ *unknowns:*

$$\mathcal{T}_{\rm I}: \quad \left\{\begin{array}{lrcl} (a) & \lambda=(\lambda_0,\lambda_1,...,\lambda_m) & \geq & 0;\\ (b) & -\lambda_0 c+\sum\limits_{i=1}^m \lambda_i a_i & = & 0;\\ \hline (c_{\rm I}) & -\lambda_0 a+\sum\limits_{i=1}^m \lambda_i b_i & \geq & 0;\\ (d_{\rm I}) & \lambda_0 & > & 0, \end{array}\right.$$

or

$$\mathcal{T}_{\rm II}: \quad \left\{\begin{array}{lrcl} (a) & \lambda=(\lambda_0,\lambda_1,...,\lambda_m) & \geq & 0;\\ (b) & -\lambda_0 c-\sum\limits_{i=1}^m \lambda_i a_i & = & 0;\\ \hline (c_{\rm II}) & -\lambda_0 a-\sum\limits_{i=1}^m \lambda_i b_i & > & 0 \end{array}\right.$$

– has a solution.

Now assume that (LP) *is feasible.* We claim that *under this assumption* ($\mathcal{S}_a$) *has no solutions if and only if* $\mathcal{T}_{\rm I}$ *has a solution.*

The implication "$\mathcal{T}_{\rm I}$ has a solution $\Rightarrow$ ($\mathcal{S}_a$) has no solution" is readily given by the above remarks. To verify the inverse implication, assume that ($\mathcal{S}_a$) has no solutions and the system $Ax\leq b$ has a solution, and let us prove that then $\mathcal{T}_{\rm I}$ has a solution. If $\mathcal{T}_{\rm I}$ has no solution, then by (*) $\mathcal{T}_{\rm II}$ has a solution and, moreover, $\lambda_0=0$ for (every) solution to $\mathcal{T}_{\rm II}$ (since a solution to the latter system with $\lambda_0>0$ solves $\mathcal{T}_{\rm I}$ as well). But the fact that $\mathcal{T}_{\rm II}$ has a solution λ with $\lambda_0=0$ is independent of the values of a and c; if this fact would take place, it would mean, by the same Theorem on Alternative, that, e.g., the following instance of ($\mathcal{S}_a$):

$$0^Tx\geq -1, Ax\geq b$$

has no solutions. The latter means that the system $Ax\geq b$ has no solutions – a contradiction with the assumption that (LP) is feasible. ∎

Now, if $\mathcal{T}_{\rm I}$ has a solution, this system has a solution with $\lambda_0=1$ as well (to see this, pass from a solution λ to the one λ/λ_0; this construction is well-defined, since $\lambda_0>0$ for every solution to $\mathcal{T}_{\rm I}$). Now, an $(m+1)$-dimensional vector $\lambda=(1,y)$ is a solution to $\mathcal{T}_{\rm I}$ if and only if the m-dimensional vector y solves the system of linear inequalities and equations

$$\begin{array}{rcl} y & \geq & 0;\\ A^Ty\equiv\sum\limits_{i=1}^m y_i a_i & = & c;\\ b^Ty & \geq & a \end{array} \qquad \text{(D)}$$

Summarizing our observations, we come to the following result.

Proposition 1.2.2 *Assume that system* (D) *associated with the LP program* (LP) *has a solution* (y,a). *Then* a *is a lower bound on the optimal value in* (LP). *Vice versa, if* (LP) *is feasible and* a *is a lower bound on the optimal value of* (LP), *then* a *can be extended by a properly chosen* m*-dimensional vector* y *to a solution to* (D).

We see that the entity responsible for lower bounds on the optimal value of (LP) is the system (D): every solution to the latter system induces a bound of this type, and *in the case when* (LP) *is feasible*, all lower bounds can be obtained from solutions to (D). Now note that if (y, a) is a solution to (D), then the pair $(y, b^T y)$ also is a solution to the same system, and the lower bound $b^T y$ on c^* is not worse than the lower bound a. Thus, as far as lower bounds on c^* are concerned, we lose nothing by restricting ourselves to the solutions (y, a) of (D) with $a = b^T y$; the best lower bound on c^* given by (D) is therefore the optimal value of the problem $\max_y \left\{ b^T y \,\middle|\, A^T y = c, y \geq 0 \right\}$, which is nothing but the dual to (LP) problem (LP*). Note that (LP*) is also a Linear Programming program.

All we know about the dual problem to the moment is the following:

Proposition 1.2.3 *Whenever y is a feasible solution to* (LP*)*, the corresponding value of the dual objective $b^T y$ is a lower bound on the optimal value c^* in* (LP)*. If* (LP) *is feasible, then for every $a \leq c^*$ there exists a feasible solution y of* (LP*) *with $b^T y \geq a$.*

1.2.3 The LP Duality Theorem

Proposition 1.2.3 is in fact equivalent to the following

Theorem 1.2.2 [Duality Theorem in Linear Programming] *Consider a linear programming program*

$$\min_x \left\{ c^T x \,\middle|\, Ax \geq b \right\} \tag{LP}$$

along with its dual

$$\max_y \left\{ b^T y \,\middle|\, A^T y = c, y \geq 0 \right\} \tag{LP*}$$

Then

1) *The duality is symmetric: the problem dual to dual is equivalent to the primal;*

2) *The value of the dual objective at every dual feasible solution is $\leq$ the value of the primal objective at every primal feasible solution*

3) *The following 5 properties are equivalent to each other:*

(i) *The primal is feasible and bounded below.*

(ii) *The dual is feasible and bounded above.*

(iii) *The primal is solvable.*

(iv) *The dual is solvable.*

(v) *Both primal and dual are feasible.*

Whenever (i) $\equiv$ (ii) $\equiv$ (iii) $\equiv$ (iv) $\equiv$ (v) *is the case, the optimal values of the primal and the dual problems are equal to each other.*

Proof. 1) is quite straightforward: writing the dual problem (LP*) in our standard form, we get

$$\min_y \left\{ -b^T y \,\middle|\, \begin{bmatrix} I_m \\ A^T \\ -A^T \end{bmatrix} y - \begin{pmatrix} 0 \\ -c \\ c \end{pmatrix} \geq 0 \right\},$$

where I_m is the m-dimensional unit matrix. Applying the duality transformation to the latter problem, we come to the problem

$$\max_{\xi,\eta,\zeta}\left\{0^T\xi + c^T\eta + (-c)^T\zeta \;\middle|\; \begin{array}{rcl} \xi & \geq & 0 \\ \eta & \geq & 0 \\ \zeta & \geq & 0 \\ \xi - A\eta + A\zeta & = & -b \end{array}\right\},$$

which is clearly equivalent to (LP) (set $x = \eta - \zeta$).

2) is readily given by Proposition 1.2.3.

3):

(i)⇒(iv): If the primal is feasible and bounded below, its optimal value c^* (which of course is a lower bound on itself) can, by Proposition 1.2.3, be (non-strictly) majorized by a quantity b^Ty^*, where y^* is a feasible solution to (LP*). In the situation in question, of course, $b^Ty^* = c^*$ (by already proved item 2)); on the other hand, in view of the same Proposition 1.2.3, the optimal value in the dual is $\leq c^*$. We conclude that the optimal value in the dual is attained and is equal to the optimal value in the primal.

(iv)⇒(ii): evident;

(ii)⇒(iii): This implication, in view of the primal-dual symmetry, follows from the implication (i)⇒(iv).

(iii)⇒(i): evident.

We have seen that (i)≡(ii)≡(iii)≡(iv) and that the first (and consequently each) of these 4 equivalent properties implies that the optimal value in the primal problem is equal to the optimal value in the dual one. All which remains is to prove the equivalence between (i)–(iv), on one hand, and (v), on the other hand. This is immediate: (i)–(iv), of course, imply (v); vice versa, in the case of (v) the primal is not only feasible, but also bounded below (this is an immediate consequence of the feasibility of the dual problem, see 2)), and (i) follows. ∎

An immediate corollary of the LP Duality Theorem is the following *necessary and sufficient* optimality condition in LP:

Theorem 1.2.3 [Necessary and sufficient optimality conditions in linear programming] *Consider an LP program* (LP) *along with its dual* (LP*). *A pair* (x, y) *of primal and dual feasible solutions is comprised of optimal solutions to the respective problems if and only if*

$$y_i[Ax - b]_i = 0,\ i = 1, ..., m, \qquad \text{[complementary slackness]}$$

likewise as if and only if

$$c^Tx - b^Ty = 0 \qquad \text{[zero duality gap]}$$

Indeed, the "zero duality gap" optimality condition is an immediate consequence of the fact that the value of primal objective at every primal feasible solution is $\geq$ the value of the dual objective at every dual feasible solution, while the optimal values in the primal and the dual are equal to each other, see Theorem 1.2.2. The equivalence between the "zero duality gap" and the "complementary slackness" optimality conditions is given by the following

computation: whenever x is primal feasible and y is dual feasible, the products $y_i[Ax-b]_i$, $i=1,...,m$, are nonnegative, while the sum of these products is precisely the duality gap:

$$y^T[Ax-b] = (A^Ty)^Tx - b^Ty = c^Tx - b^Ty.$$

Thus, the duality gap can vanish at a primal-dual feasible pair (x,y) if and only if all products $y_i[Ax-b]_i$ for this pair are zeros.

1.3 From Linear to Conic Programming

Linear Programming models cover numerous applications. Whenever applicable, LP allows to obtain useful quantitative and qualitative information on the problem at hand. The specific analytic structure of LP programs gives rise to a number of general results (e.g., those of the LP Duality Theory) which provide us in many cases with valuable insight and understanding. At the same time, this analytic structure underlies some specific computational techniques for LP; these techniques, which by now are perfectly well developed, allow to solve routinely quite large (tens/hundreds of thousands of variables and constraints) LP programs. Nevertheless, there are situations in reality which cannot be covered by LP models. To handle these "essentially nonlinear" cases, one needs to extend the basic theoretical results and computational techniques known for LP beyond the bounds of Linear Programming.

For the time being, the widest class of optimization problems to which the basic results of LP were extended, is the class of *convex* optimization programs. There are several equivalent ways to define a general convex optimization problem; the one we are about to use is not the traditional one, but it is well suited to encompass the range of applications we intend to cover in our course.

When passing from a generic LP problem

$$\min_x \left\{ c^Tx \;\middle|\; Ax \geq b \right\} \quad [A: m \times n] \tag{LP}$$

to its nonlinear extensions, we should expect to encounter some nonlinear components in the problem. The traditional way here is to say: "Well, in (LP) there are a linear objective function $f(x) = c^Tx$ and inequality constraints $f_i(x) \geq b_i$ with linear functions $f_i(x) = a_i^Tx$, $i=1,...,m$. Let us allow some/all of these functions $f, f_1, ..., f_m$ to be nonlinear." In contrast to this traditional way, we intend to keep the objective and the constraints linear, but introduce "nonlinearity" in the inequality sign $\geq$.

1.4 Orderings of $\mathbf{R}^m$ and cones

The constraint inequality $Ax \geq b$ in (LP) is an inequality between *vectors*; as such, it requires a definition, and the definition is well-known: given two vectors $a, b \in \mathbf{R}^m$, we write $a \geq b$, if the coordinates of a majorate the corresponding coordinates of b:

$$a \geq b \Leftrightarrow \{a_i \geq b_i,\ i=1,...,m\}. \tag{" $\geq$ "}$$

In the latter relation, we again meet with the inequality sign $\geq$, but now it stands for the "arithmetic $\geq$" – a well-known relation between real numbers. The above "coordinate-wise" partial ordering of vectors in $\mathbf{R}^m$ satisfies a number of basic properties of the standard ordering of reals; namely, for all vectors $a, b, c, d, ... \in \mathbf{R}^m$ one has

1. *Reflexivity:* $a \geq a$;
2. *Anti-symmetry:* if both $a \geq b$ and $b \geq a$, then $a = b$;
3. *Transitivity:* if both $a \geq b$ and $b \geq c$, then $a \geq c$;
4. *Compatibility with linear operations:*
   (a) *Homogeneity:* if $a \geq b$ and λ is a nonnegative real, then $\lambda a \geq \lambda b$
   ("One can multiply both sides of an inequality by a nonnegative real")
   (b) *Additivity:* if both $a \geq b$ and $c \geq d$, then $a + c \geq b + d$
   ("One can add two inequalities of the same sign").

It turns out that

- *A significant part of the nice features of LP programs comes from the fact that the vector inequality $\geq$ in the constraint of* (LP) *satisfies the properties* 1. – 4.*;*
- *The standard inequality " $\geq$ " is neither the only possible, nor the only interesting way to define the notion of a vector inequality fitting the axioms* 1. – 4.

As a result,

> *A generic optimization problem which looks exactly the same as* (LP)*, up to the fact that the inequality $\geq$ in* (LP) *is now replaced with and ordering which differs from the component-wise one, inherits a significant part of the properties of LP problems. Specifying properly the ordering of vectors, one can obtain from* (LP) *generic optimization problems covering many important applications which cannot be treated by the standard LP.*

To the moment what is said is just a declaration. Let us look how this declaration comes to life.

We start with clarifying the "geometry" of a "vector inequality" satisfying the axioms 1. – 4. Thus, we consider vectors from a finite-dimensional Euclidean space $\mathbf{E}$ with an inner product $\langle \cdot, \cdot \rangle$ and assume that $\mathbf{E}$ is equipped with a partial ordering (called also *vector inequality*), let it be denoted by $\succeq$: in other words, we say what are the pairs of vectors a, b from $\mathbf{E}$ linked by the inequality $a \succeq b$. We call the ordering "good", if it obeys the axioms 1. – 4., and are interested to understand what are these good orderings.

Our first observation is:

> A. *A good vector inequality $\succeq$ is completely identified by the set $\mathbf{K}$ of $\succeq$-nonnegative vectors:*
>
> $$\mathbf{K} = \{a \in \mathbf{E} \mid a \succeq 0\}.$$
>
> *Namely,*
>
> $$a \succeq b \Leftrightarrow a - b \succeq 0 \quad [\Leftrightarrow a - b \in \mathbf{K}].$$
>
> Indeed, let $a \succeq b$. By 1. we have $-b \succeq -b$, and by 4.(b) we may add the latter inequality to the former one to get $a - b \succeq 0$. Vice versa, if $a - b \succeq 0$, then, adding to this inequality the one $b \succeq b$, we get $a \succeq b$.

The set $\mathbf{K}$ in Observation A cannot be arbitrary. It is easy to verify that it must be a *pointed cone*, i.e., it must satisfy the following conditions:

1. $\mathbf{K}$ is nonempty and closed under addition:
$$a, a' \in \mathbf{K} \Rightarrow a + a' \in \mathbf{K};$$
2. $\mathbf{K}$ is a conic set:
$$a \in \mathbf{K}, \lambda \geq 0 \Rightarrow \lambda a \in \mathbf{K}.$$
3. $\mathbf{K}$ is pointed:
$$a \in \mathbf{K} \text{ and } -a \in \mathbf{K} \Rightarrow a = 0.$$
Geometrically: $\mathbf{K}$ does not contain straight lines passing through the origin.

Exercise 1.1 *Prove that the outlined properties of* $\mathbf{K}$ *are necessary and sufficient for the vector inequality* $a \succeq b \Leftrightarrow a - b \in \mathbf{K}$ *to be good.*

Thus, every pointed cone $\mathbf{K}$ in $\mathbf{E}$ induces a partial ordering on $\mathbf{E}$ which satisfies the axioms 1. – 4. We denote this ordering by $\geq_{\mathbf{K}}$:

$$a \geq_{\mathbf{K}} b \Leftrightarrow a - b \geq_{\mathbf{K}} 0 \Leftrightarrow a - b \in \mathbf{K}.$$

What is the cone responsible for the standard coordinate-wise ordering $\geq$ on $\mathbf{E} = \mathbf{R}^m$ we have started with? The answer is clear: this is the cone comprised of vectors with nonnegative entries – the *nonnegative orthant*

$$\mathbf{R}^m_+ = \{x = (x_1, ..., x_m)^T \in \mathbf{R}^m : x_i \geq 0,\ i = 1, ..., m\}.$$

(Thus, in order to express the fact that a vector a is greater than or equal to, in the component-wise sense, to a vector b, we were supposed to write $a \geq_{\mathbf{R}^m_+} b$. However, we are not going to be that formal and shall use the standard shorthand notation $a \geq b$.)

The nonnegative orthant $\mathbf{R}^m_+$ is not just a pointed cone; it possesses two useful additional properties:

I. The cone is closed: if a sequence of vectors a^i from the cone has a limit, the latter also belongs to the cone.

II. The cone possesses a nonempty interior: there exists a vector such that a ball of positive radius centered at the vector is contained in the cone.

These additional properties are very important. For example, **I** is responsible for the possibility to pass to the term-wise limit in an inequality:

$$a^i \geq b^i \quad \forall i, \quad a^i \to a, b^i \to b \text{ as } i \to \infty \Rightarrow a \geq b.$$

It makes sense to restrict ourselves with good partial orderings coming from cones $\mathbf{K}$ sharing the properties **I**, **II**. Thus,

From now on, speaking about vector inequalities $\geq_{\mathbf{K}}$*, we always assume that the underlying set* $\mathbf{K}$ *is a pointed and closed cone with a nonempty interior.*

Note that the closedness of $\mathbf{K}$ makes it possible to pass to limits in $\geq_{\mathbf{K}}$-inequalities:

$$a^i \geq_{\mathbf{K}} b^i,\ a^i \to a, b^i \to b \text{ as } i \to \infty \Rightarrow a \geq_{\mathbf{K}} b.$$

The nonemptiness of the interior of $\mathbf{K}$ allows to define, along with the "non-strict" inequality $a \geq_{\mathbf{K}} b$, also the strict inequality according to the rule

$$a >_{\mathbf{K}} b \Leftrightarrow a - b \in \operatorname{int} \mathbf{K},$$

where $\operatorname{int} K$ is the interior of the cone $\mathbf{K}$. E.g., the strict coordinate-wise inequality $a >_{\mathbf{R}^m_+} b$ (shorthand: $a > b$) simply says that the coordinates of a are strictly greater, in the usual arithmetic sense, than the corresponding coordinates of b.

Examples. The partial orderings we are especially interested in are given by the following cones:

- The nonnegative orthant $\mathbf{R}^m_+$ in $\mathbf{R}^n$;
- The *Lorentz* (or the *second-order*, or the less scientific name the *ice-cream*) cone

$$\mathbf{L}^m = \left\{ x = (x_1, ..., x_{m-1}, x_m)^T \in \mathbf{R}^m : x_m \geq \sqrt{\sum_{i=1}^{m-1} x_i^2} \right\}$$

- The *semidefinite cone* $\mathbf{S}^m_+$. This cone "lives" in the space $\mathbf{E} = \mathbf{S}^m$ of $m \times m$ symmetric matrices (equipped with the Frobenius inner product $\langle A, B \rangle = \operatorname{Tr}(AB) = \sum\limits_{i,j} A_{ij} B_{ij}$) and consists of all $m \times m$ matrices A which are *positive semidefinite*, i.e.,

$$A = A^T; \quad x^T A x \geq 0 \quad \forall x \in \mathbf{R}^m.$$

1.5 "Conic programming" – what is it?

Let $\mathbf{K}$ be a cone in $\mathbf{E}$ (convex, pointed, closed and with a nonempty interior). Given an *objective* $c \in \mathbf{R}^n$, a linear mapping $x \mapsto Ax : \mathbf{R}^n \to \mathbf{E}$ and a *right hand side* $b \in \mathbf{E}$, consider the optimization problem

$$\min_x \left\{ c^T x \,\middle|\, Ax \geq_{\mathbf{K}} b \right\} \qquad \text{(CP)}.$$

We shall refer to (CP) as to a *conic* problem associated with the cone $\mathbf{K}$. Note that the only difference between this program and an LP problem is that the latter deals with the particular choice $\mathbf{E} = \mathbf{R}^m$, $\mathbf{K} = \mathbf{R}^m_+$. With the formulation (CP), we get a possibility to cover a much wider spectrum of applications which cannot be captured by LP; we shall look at numerous examples in the sequel.

1.6 Conic Duality

Aside of algorithmic issues, the most important theoretical result in Linear Programming is the LP Duality Theorem; can this theorem be extended to conic problems? What is the extension?

The source of the LP Duality Theorem was the desire to get in a systematic way a lower bound on the optimal value c^* in an LP program

$$c^* = \min_x \left\{ c^T x \;\middle|\; Ax \geq b \right\}. \tag{LP}$$

The bound was obtained by looking at the inequalities of the type

$$\langle \lambda, Ax \rangle \equiv \lambda^T Ax \geq \lambda^T b \tag{Cons(λ)}$$

with weight vectors $\lambda \geq 0$. By its origin, an inequality of this type is a consequence of the system of constraints $Ax \geq b$ of (LP), i.e., it is satisfied at every solution to the system. Consequently, whenever we are lucky to get, as the left hand side of (Cons(λ)), the expression $c^T x$, i.e., whenever a nonnegative weight vector λ satisfies the relation

$$A^T \lambda = c,$$

the inequality (Cons(λ)) yields a lower bound $b^T \lambda$ on the optimal value in (LP). And the dual problem

$$\max \left\{ b^T \lambda \mid \lambda \geq 0, A^T \lambda = c \right\}$$

was nothing but the problem of finding the best lower bound one can get in this fashion.

The same scheme can be used to develop the dual to a conic problem

$$\min \left\{ c^T x \mid Ax \geq_{\mathbf{K}} b \right\}, \; \mathbf{K} \subset \mathbf{E}. \tag{CP}$$

Here the only step which needs clarification is the following one:

> (?) *What are the "admissible" weight vectors λ, i.e., the vectors such that the scalar inequality*
>
> $$\langle \lambda, Ax \rangle \geq \langle \lambda, b \rangle$$
>
> *is a consequence of the vector inequality $Ax \geq_{\mathbf{K}} b$?*

In the particular case of coordinate-wise partial ordering, i.e., in the case of $\mathbf{E} = \mathbf{R}^m$, $\mathbf{K} = \mathbf{R}^m_+$, the admissible vectors were those with nonnegative coordinates. These vectors, however, not necessarily are admissible for an ordering $\geq_{\mathbf{K}}$ when $\mathbf{K}$ is different from the nonnegative orthant:

Example 1.6.1 *Consider the ordering $\geq_{\mathbf{L}^3}$ on $\mathbf{E} = \mathbf{R}^3$ given by the 3-dimensional ice-cream cone:*

$$\begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} \geq_{\mathbf{L}^3} \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \Leftrightarrow a_3 \geq \sqrt{a_1^2 + a_2^2}.$$

The inequality

$$\begin{pmatrix} -1 \\ -1 \\ 2 \end{pmatrix} \geq_{\mathbf{L}^3} \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

is valid; however, aggregating this inequality with the aid of a positive weight vector $\lambda = \begin{pmatrix} 1 \\ 1 \\ 0.1 \end{pmatrix}$, *we get the false inequality*

$$-1.8 \geq 0.$$

Thus, not every nonnegative weight vector is admissible for the partial ordering $\geq_{\mathbf{L}^3}$.

To answer the question (?) is the same as to say what are the weight vectors λ such that

$$\forall a \geq_{\mathbf{K}} 0: \quad \langle \lambda, a \rangle \geq 0. \tag{1.6.1}$$

Whenever λ possesses the property (1.6.1), the scalar inequality

$$\langle \lambda, a \rangle \geq \langle \lambda, b \rangle$$

is a consequence of the vector inequality $a \geq_{\mathbf{K}} b$:

$$\begin{array}{rrcll} & a & \geq_{\mathbf{K}} & b & \\ \Leftrightarrow & a-b & \geq_{\mathbf{K}} & 0 & [\text{additivity of } \geq_{\mathbf{K}}] \\ \Rightarrow & \langle \lambda, a-b \rangle & \geq & 0 & [\text{by (1.6.1)}] \\ \Leftrightarrow & \langle \lambda, a \rangle & \geq & \lambda^T b. & \blacksquare \end{array}$$

Vice versa, if λ is an admissible weight vector for the partial ordering $\geq_{\mathbf{K}}$:

$$\forall (a, b : a \geq_{\mathbf{K}} b): \quad \langle \lambda, a \rangle \geq \langle \lambda, b \rangle$$

then, of course, λ satisfies (1.6.1).

Thus the weight vectors λ which are admissible for a partial ordering $\geq_{\mathbf{K}}$ are exactly the vectors satisfying (1.6.1), or, which is the same, the vectors from the set

$$\mathbf{K}_* = \{\lambda \in \mathbf{E} : \langle \lambda, a \rangle \geq 0 \quad \forall a \in \mathbf{K}\}.$$

The set $\mathbf{K}_*$ is comprised of vectors whose inner products with *all* vectors from $\mathbf{K}$ are nonnegative. $\mathbf{K}_*$ is called the *cone dual to* $\mathbf{K}$. The name is legitimate due to the following fact (see Section B.2.6.B):

Theorem 1.6.1 [Properties of the dual cone] *Let* $\mathbf{E}$ *be a finite-dimensional Euclidean space with inner product* $\langle \cdot, \cdot \rangle$ *and let* $K \subset \mathbf{E}$ *be a nonempty set. Then*

(i) *The set*

$$K_* = \{\lambda \in \mathbf{E}^m : \langle \lambda, a \rangle \geq 0 \quad \forall a \in K\}$$

is a closed cone.

(ii) *If* $\operatorname{int} K \neq \emptyset$, *then* K_* *is pointed.*

(iii) *If* K *is a closed convex pointed cone, then* $\operatorname{int} K_* \neq \emptyset$.

(iv) *If* K *is a closed cone, then so is* K_*, *and the cone dual to* K_* *is* K *itself:*

$$(K_*)_* = K.$$

An immediate corollary of the Theorem is as follows:

Corollary 1.6.1 *A set* $K \subset \mathbf{E}$ *is a closed convex pointed cone with a nonempty interior if and only if the set* K_* *is so.*

From the dual cone to the problem dual to (CP). Now we are ready to derive the dual problem of a conic problem (CP). As in the case of Linear Programming, we start with the observation that whenever x is a feasible solution to (CP) and λ is an admissible weight vector, i.e., $\lambda \in \mathbf{K}_*$, then x satisfies the scalar inequality

$$(A^*\lambda)^T x \equiv \langle \lambda, Ax \rangle \geq \langle \lambda, b \rangle \quad [1)]$$

– this observation is an immediate consequence of the definition of $\mathbf{K}_*$. It follows that whenever λ_* is an admissible weight vector satisfying the relation

$$A^*\lambda = c,$$

one has

$$c^T x = (A^*\lambda)^T x = \langle \lambda, Ax \rangle \geq \langle b, \lambda \rangle$$

for all x feasible for (CP), so that the quantity $\langle b, \lambda \rangle$ is a lower bound on the optimal value of (CP). The best bound one can get in this fashion is the optimal value in the problem

$$\max \{ \langle b, \lambda \rangle \mid A^*\lambda = c, \lambda \geq_{\mathbf{K}_*} 0 \} \tag{D}$$

and this program is called the program *dual* to (CP).

So far, what we know about the duality we have just introduced is the following

Proposition 1.6.1 [Weak Conic Duality Theorem] *The optimal value of* (D) *is a lower bound on the optimal value of* (CP).

1.6.1 Geometry of the primal and the dual problems

The structure of problem (D) looks quite different from the one of (CP). However, a more careful analysis demonstrates that the difference in structures comes just from how we represent the data: geometrically, the problems are completely similar. Indeed, in (D) we are asked to maximize a linear objective $\langle b, \lambda \rangle$ over the intersection of an affine plane $L_* = \{\lambda \mid A^*\lambda = c\}$ with the cone $\mathbf{K}_*$. And what about (CP)? Let us pass in this problem from the "true design variables" x to their images $y = Ax - b \in \mathbf{E}$. When x runs through $\mathbf{R}^n$, y runs through the affine plane $L = \{y = Ax - b \mid x \in \mathbf{R}^n\}$; $x \in \mathbf{R}^n$ is feasible for (CP) if and only if the corresponding $y = Ax - b$ belongs to the cone $\mathbf{K}$. Thus, in (CP) we also deal with the intersection of an affine plane, namely, L, and a cone, namely, $\mathbf{K}$. Now assume that our objective $c^T x$ can be expressed in terms of $y = Ax - b$:

$$c^T x = \langle d, Ax - b \rangle + \text{const}.$$

This assumption is clearly equivalent to the inclusion

$$c \in \text{Im} A^*. \tag{1.6.2}$$

[1)] For a linear operator $x \mapsto Ax : \mathbf{R}^n \to \mathbf{E}$, A^* is the *conjugate* operator given by the identity

$$\langle y, Ax \rangle = x^T A y \quad \forall (y \in \mathbf{E}, x \in \mathbf{R}^n).$$

When representing the operators by their matrices in *orthogonal* bases in the argument and the range spaces, the matrix representing the conjugate operator is exactly the transpose of the matrix representing the operator itself.

Indeed, in the latter case we have $c = A^*d$ for some d, whence

$$c^T x = \langle A^*d, x\rangle = \langle d, Ax\rangle = \langle d, Ax - b\rangle + \langle d, b\rangle \quad \forall x. \tag{1.6.3}$$

In the case of (1.6.2) the primal problem (CP) can be posed equivalently as the following problem:

$$\min_y \{\langle d, y\rangle \mid y \in L,\ y \geq_{\mathbf{K}} 0\},$$

where $L = \mathrm{Im}A - b$ and d is (any) vector satisfying the relation $A^*d = c$. Thus,

> *In the case of* (1.6.2) *the primal problem, geometrically, is the problem of minimizing a linear form over the intersection of the affine plane L with the cone $\mathbf{K}$, and the dual problem, similarly, is to maximize another linear form over the intersection of the affine plane L_* with the dual cone $\mathbf{K}_*$.*

Now, what happens if the condition (1.6.2) is *not* satisfied? The answer is very simple: in this case (CP) makes no sense – it is either unbounded below, or infeasible.

> Indeed, assume that (1.6.2) is not satisfied. Then, by Linear Algebra, the vector c is not orthogonal to the null space of A, so that there exists e such that $Ae = 0$ and $c^T e > 0$. Now let x be a feasible solution of (CP); note that all points $x - \mu e$, $\mu \geq 0$, are feasible, and $c^T(x - \mu e) \to \infty$ as $\mu \to \infty$. Thus, when (1.6.2) is not satisfied, problem (CP), whenever feasible, is unbounded below.

From the above observation we see that if (1.6.2) is not satisfied, then we may reject (CP) from the very beginning. Thus, from now on we assume that (1.6.2) is satisfied. In fact in what follows we make a bit stronger assumption:

> **A.** *The mapping A is of full column rank, i.e., it has trivial null space.*
>
> Assuming that the mapping $x \mapsto Ax$ has the trivial null space ("we have eliminated from the very beginning the redundant degrees of freedom – those not affecting the value of Ax"), the equation
>
> $$A^*d = q$$
>
> is solvable for every right hand side vector q.

In view of **A**, problem (CP) can be reformulated as a problem (P) of minimizing a linear objective $\langle d, y\rangle$ over the intersection of an affine plane L and a cone $\mathbf{K}$. Conversely, a problem (P) of this latter type can be posed in the form of (CP) – to this end it suffices to represent the plane L as the image of an affine mapping $x \mapsto Ax - b$ (i.e., to parameterize somehow the feasible plane) and to "translate" the objective $\langle d, y\rangle$ to the space of x-variables – to set $c = A^*d$, which yields

$$y = Ax - b \Rightarrow \langle d, y\rangle = c^T x + \mathrm{const}.$$

Thus, when dealing with a conic problem, we may pass from its "analytic form" (CP) to the "geometric form" (P) and vice versa.

What are the relations between the "geometric data" of the primal and the dual problems? We already know that the cone $\mathbf{K}_*$ associated with the dual problem is dual of the cone $\mathbf{K}$ associated with the primal one. What about the feasible planes L and L_*? The answer is

simple: they are orthogonal to each other! More exactly, the affine plane L is the translation, by vector $-b$, of the linear subspace

$$\mathcal{L} = \mathrm{Im}A \equiv \{y = Ax \mid x \in \mathbf{R}^n\}.$$

And L_* is the translation, by any solution λ_0 of the system $A^*\lambda = c$, *e.g., by the solution d to the system*, of the linear subspace

$$\mathcal{L}_* = \mathrm{Null}(A^*) \equiv \{\lambda \mid A^*\lambda = 0\}.$$

A well-known fact of Linear Algebra is that the linear subspaces $\mathcal{L}$ and $\mathcal{L}_*$ are orthogonal complements of each other:

$$\mathcal{L} = \{y \mid \langle y, \lambda\rangle = 0 \quad \forall \lambda \in \mathcal{L}_*\}; \;\; \mathcal{L}_* = \{\lambda \mid \langle y, \lambda\rangle = 0 \quad \forall y \in \mathcal{L}\}.$$

Thus, we come to a nice geometrical conclusion:

A conic problem[2)] (CP) *is the problem*

$$\min_y \{\langle d, y\rangle \mid y \in \mathcal{L} - b, \; y \geq_{\mathbf{K}} 0\} \tag{P}$$

of minimizing a linear objective $\langle d, y\rangle$ over the intersection of a cone $\mathbf{K}$ with an affine plane $L = \mathcal{L} - b$ given as a translation, by vector $-b$, of a linear subspace $\mathcal{L}$.

The dual problem is the problem

$$\max_\lambda \left\{\langle b, \lambda\rangle \mid \lambda \in \mathcal{L}^\perp + d, \; \lambda \geq_{\mathbf{K}_*} 0\right\}. \tag{D}$$

of maximizing the linear objective $\langle b, \lambda\rangle$ over the intersection of the dual cone $\mathbf{K}_*$ with an affine plane $L_* = \mathcal{L}^\perp + d$ given as a translation, by the vector d, of the orthogonal complement $\mathcal{L}^\perp$ of $\mathcal{L}$.

What we get is an extremely transparent geometric description of *the primal-dual pair* of conic problems (P), (D). Note that the duality is completely symmetric: the problem dual to (D) is (P)! Indeed, we know from Theorem 1.6.1 that $(\mathbf{K}_*)_* = \mathbf{K}$, and of course $(\mathcal{L}^\perp)^\perp = \mathcal{L}$. Switch from maximization to minimization corresponds to the fact that the "shifting vector" in (P) is $(-b)$, while the "shifting vector" in (D) is d. The geometry of the primal-dual pair (P), (D) is

[2)] recall that we have restricted ourselves to the problems satisfying the assumption **A**

illustrated on the below picture:

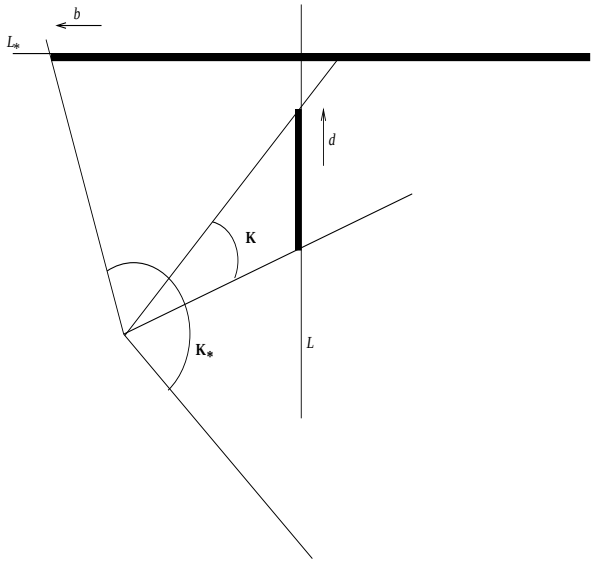


Figure 1.1. Primal-dual pair of conic problems
[bold: primal (vertical segment) and dual (horizontal ray) feasible sets]

Finally, note that in the case when (CP) is an LP program (i.e., in the case when $\mathbf{K}$ is the nonnegative orthant), the "conic dual" problem (D) is exactly the usual LP dual; this fact immediately follows from the observation that the cone dual to $\mathbf{R}^m_+$ is $\mathbf{R}^m_+$ itself.

We have explored the geometry of a primal-dual pair of conic problems: the "geometric data" of such a pair are given by a pair of dual to each other cones $\mathbf{K}$, $\mathbf{K}_*$ in $\mathbf{E}$ and a pair of affine planes $L = \mathcal{L} - b$, $L_* = \mathcal{L}^\perp + d$, where $\mathcal{L}$ is a linear subspace in $\mathbf{E}$ and $\mathcal{L}^\perp$ is its orthogonal complement. The first problem from the pair – let it be called (P) – is to minimize $\langle b, y\rangle$ over $y \in \mathbf{K} \cap L$, and the second (D) is to maximize $\langle d, \lambda\rangle$ over $\lambda \in \mathbf{K}_* \cap L_*$. Note that the "geometric data" $(\mathbf{K}, \mathbf{K}_*, L, L_*)$ of the pair do not specify completely the problems of the pair: given L, L_*, we can uniquely define $\mathcal{L}$, but not the shift vectors $(-b)$ and d: b is known up to shift by a vector from $\mathcal{L}$, and d is known up to shift by a vector from $\mathcal{L}^\perp$. However, this non-uniqueness is of absolutely no importance: replacing a chosen vector $d \in L_*$ by another vector $d' \in L_*$, we pass from (P) to a new problem (P$'$) which is completely equivalent to (P): indeed, both (P) and (P$'$) have the same feasible set, and *on the (common) feasible plane L of the problems their objectives $\langle d, y\rangle$ and $\langle d', y\rangle$ differ from each other by a constant:*

$$y \in L = \mathcal{L} - b, d - d' \in \mathcal{L}^\perp \Rightarrow \langle d - d', y + b\rangle = 0 \Rightarrow \langle d - d', y\rangle = \langle -(d - d'), b\rangle \quad \forall y \in L.$$

Similarly, shifting b along $\mathcal{L}$, we do modify the objective in (D), but in a trivial way – *on the feasible plane L_** of the problem the new objective differs from the old one by a constant.

1.7 Conic Duality Theorem

The Weak Duality (Proposition 1.6.1) we have established so far for conic problems is much weaker than the Linear Programming Duality Theorem. Is it possible to get results similar to those of the LP Duality Theorem in the general conic case as well? The answer is affirmative, *provided that the primal problem* (CP) *is strictly feasible*, i.e., that there exists x such that $Ax - b >_{\mathbf{K}} 0$, or, geometrically, $L \cap \operatorname{int} \mathbf{K} \neq \emptyset$.

The advantage of the geometrical definition of strict feasibility is that it is independent of the particular way in which the feasible plane is defined; hence, with this definition it is clear what does it mean that the dual problem (D) is strictly feasible.

Our main result is the following

Theorem 1.7.1 [Conic Duality Theorem] *Consider a conic problem*

$$c^* = \min_x \left\{ c^T x \mid Ax \geq_{\mathbf{K}} b \right\} \tag{CP}$$

along with its conic dual

$$b^* = \max \left\{ \langle b, \lambda \rangle \mid A^* \lambda = c, \lambda \geq_{\mathbf{K}_*} 0 \right\}. \tag{D}$$

1) *The duality is symmetric: the dual problem is conic, and the problem dual to dual is the primal.*

2) *The value of the dual objective at every dual feasible solution λ is $\leq$ the value of the primal objective at every primal feasible solution x, so that the duality gap*

$$c^T x - \langle b, \lambda \rangle$$

is nonnegative at every "primal-dual feasible pair" (x, λ).

3.a) *If the primal* (CP) *is bounded below and strictly feasible (i.e. $Ax >_{\mathbf{K}} b$ for some x), then the dual* (D) *is solvable and the optimal values in the problems are equal to each other: $c^* = b^*$.*

3.b) *If the dual* (D) *is bounded above and strictly feasible (i.e., exists $\lambda >_{\mathbf{K}_*} 0$ such that $A^*\lambda = c$), then the primal* (CP) *is solvable and $c^* = b^*$.*

4) *Assume that at least one of the problems* (CP), (D) *is bounded and strictly feasible. Then a primal-dual feasible pair (x, λ) is a pair of optimal solutions to the respective problems*

4.a) if and only if

$$\langle b, \lambda \rangle = c^T x \qquad \text{[zero duality gap]}$$

and

4.b) if and only if

$$\langle \lambda, Ax - b \rangle = 0 \qquad \text{[complementary slackness]}$$

Proof. 1): The result was already obtained when discussing the geometry of the primal and the dual problems.

2): This is the Weak Conic Duality Theorem.

3): Assume that (CP) is strictly feasible and bounded below, and let c^* be the optimal value of the problem. We should prove that the dual is solvable with the same optimal value. Since we already know that the optimal value of the dual is $\leq c^*$ (see 2)), all we need is to point out a dual feasible solution λ_* with $b^T \lambda_* \geq c^*$.

Consider the convex set

$$M = \{ y = Ax - b \mid x \in \mathbf{R}^n, c^T x \leq c^* \}.$$

Let us start with the case of $c \neq 0$. We claim that in this case

(i) The set M is nonempty;

(ii) the plane M does not intersect the interior K of the cone $\mathbf{K}$: $M \cap \operatorname{int} \mathbf{K} = \emptyset$.

(i) is evident (why?). To verify (ii), assume, on contrary, that there exists a point $\bar{x}$, $c^T \bar{x} \leq c^*$, such that $\bar{y} \equiv A\bar{x} - b >_{\mathbf{K}} 0$. Then, of course, $Ax - b >_{\mathbf{K}} 0$ for all x close enough to $\bar{x}$, i.e., all

points x in a small enough neighbourhood of $\bar{x}$ are also feasible for (CP). Since $c \neq 0$, there are points x in this neighbourhood with $c^T x < c^T \bar{x} \leq c^*$, which is impossible, since c^* is the optimal value of (CP).

Now let us make use of the following basic fact (see Section B.2.5):

Theorem 1.7.2 [Separation Theorem for Convex Sets] *Let S, T be nonempty non-intersecting convex subsets of a finite-dimensional Euclidean space $\mathbf{E}$ with inner product $\langle \cdot, \cdot \rangle$. Then S and T can be separated by a linear functional: there exists a nonzero vector $\lambda \in \mathbf{E}$ such that*

$$\sup_{u \in S} \langle \lambda, u \rangle \leq \inf_{u \in T} \langle \lambda, u \rangle.$$

Applying the Separation Theorem to $S = M$ and $T = K$, we conclude that there exists $\lambda \in \mathbf{E}$ such that

$$\sup_{y \in M} \langle \lambda, y \rangle \leq \inf_{y \in \operatorname{int} \mathbf{K}} \langle \lambda, y \rangle. \tag{1.7.1}$$

From the inequality it follows that the linear form $\langle \lambda, y \rangle$ of y is bounded below on $K = \operatorname{int} \mathbf{K}$. Since this interior is a conic set:

$$y \in K, \mu > 0 \Rightarrow \mu y \in K$$

(why?), this boundedness implies that $\langle \lambda, y \rangle \geq 0$ for all $y \in K$. Consequently, $\langle \lambda, y \rangle \geq 0$ for all y from the closure of K, i.e., for all $y \in \mathbf{K}$. We conclude that $\lambda \geq_{\mathbf{K}_*} 0$, so that the inf in (1.7.1) is nonnegative. On the other hand, the infimum of a linear form over a conic set clearly cannot be positive; we conclude that the inf in (1.7.1) is 0, so that the inequality reads

$$\sup_{u \in M} \langle \lambda, u \rangle \leq 0.$$

Recalling the definition of M, we get

$$[A^* \lambda]^T x \leq \langle \lambda, b \rangle \tag{1.7.2}$$

for all x from the half-space $c^T x \leq c^*$. But the linear form $[A^* \lambda]^T x$ can be bounded above on the half-space if and only if the vector $A^* \lambda$ is proportional, with a nonnegative coefficient, to the vector c:

$$A^* \lambda = \mu c$$

for some $\mu \geq 0$. We claim that $\mu > 0$. Indeed, assuming $\mu = 0$, we get $A^* \lambda = 0$, whence $\langle \lambda, b \rangle \geq 0$ in view of (1.7.2). It is time now to recall that (CP) is strictly feasible, i.e., $A\bar{x} - b >_{\mathbf{K}} 0$ for some $\bar{x}$. Since $\lambda \geq_{\mathbf{K}_*} 0$ and $\lambda \neq 0$, the product $\langle \lambda, A\bar{x} - b \rangle$ should be strictly positive (why?), while in fact we know that the product is $-\langle \lambda, b \rangle \leq 0$ (since $A^* \lambda = 0$ and, as we have seen, $\langle \lambda, b \rangle \geq 0$).

Thus, $\mu > 0$. Setting $\lambda_* = \mu^{-1} \lambda$, we get

$$\begin{array}{rll} \lambda_* & \geq_{\mathbf{K}_*} 0 & [\text{since } \lambda \geq_{\mathbf{K}_*} 0 \text{ and } \mu > 0] \\ A^* \lambda_* & = c & [\text{since } A^* \lambda = \mu c] \\ c^T x & \leq \langle \lambda_*, b \rangle \quad \forall x : c^T x \leq c^* & [\text{see (1.7.2)}] \end{array}.$$

We see that λ_* is feasible for (D), the value of the dual objective at λ_* being at least c^*, as required.

It remains to consider the case $c = 0$. Here, of course, $c^* = 0$, and the existence of a dual feasible solution with the value of the objective $\geq c^* = 0$ is evident: the required solution is $\lambda = 0$. 3.a) is proved.

3.b): the result follows from 3.a) in view of the primal-dual symmetry.

4): Let x be primal feasible, and λ be dual feasible. Then

$$c^T x - \langle b, \lambda \rangle = (A^*\lambda)^T x - \langle b, \lambda \rangle = \langle Ax - b, \lambda \rangle.$$

We get a useful identity as follows:

> (!) *For every primal-dual feasible pair (x, λ) the duality gap $c^T x - \langle b, \lambda \rangle$ is equal to the inner product of the primal slack vector $y = Ax - b$ and the dual vector λ.*
>
> Note that (!) in fact does not require "full" primal-dual feasibility: x may be arbitrary (i.e., y should belong to the primal feasible plane $\mathrm{Im}A - b$), and λ should belong to the dual feasible plane $A^*\lambda = c$, but y and λ not necessary should belong to the respective cones.

In view of (!) the complementary slackness holds if and only if the duality gap is zero; thus, all we need is to prove 4.a).

The "primal residual" $c^T x - c^*$ and the "dual residual" $b^* - \langle b, \lambda \rangle$ are nonnegative, provided that x is primal feasible, and λ is dual feasible. It follows that the duality gap

$$c^T x - \langle b, \lambda \rangle = [c^T x - c^*] + [b^* - \langle b, \lambda \rangle] + [c^* - b^*]$$

is nonnegative (recall that $c^* \geq b^*$ by 2)), and it is zero if and only if $c^* = b^*$ and both primal and dual residuals are zero (i.e., x is primal optimal, and λ is dual optimal). All these arguments hold without any assumptions of strict feasibility. We see that the condition "the duality gap at a primal-dual feasible pair is zero" is always sufficient for primal-dual optimality of the pair; and if $c^* = b^*$, this sufficient condition is also necessary. Since in the case of 4) we indeed have $c^* = b^*$ (this is stated by 3)), 4.a) follows. ∎

A useful consequence of the Conic Duality Theorem is the following

Corollary 1.7.1 *Assume that both* (CP) *and* (D) *are strictly feasible. Then both problems are solvable, the optimal values are equal to each other, and each one of the conditions 4.a), 4.b) is necessary and sufficient for optimality of a primal-dual feasible pair.*

> Indeed, by the Weak Conic Duality Theorem, if one of the problems is feasible, the other is bounded, and it remains to use the items 3) and 4) of the Conic Duality Theorem.

1.7.1 Is something wrong with conic duality?

The statement of the Conic Duality Theorem is weaker than that of the LP Duality theorem: in the LP case, feasibility (even non-strict) and boundedness of either primal, or dual problem implies solvability of both the primal and the dual and equality between their optimal values. In the general conic case something "nontrivial" is stated only in the case of strict feasibility (and boundedness) of one of the problems. It can be demonstrated by examples that this phenomenon reflects the nature of things, and is not due to our ability to analyze it. The case of non-polyhedral cone $\mathbf{K}$ is truly more complicated than the one of the nonnegative orthant $\mathbf{K}$; as a result, a "word-by-word" extension of the LP Duality Theorem to the conic case is false.

Example 1.7.1 Consider the following conic problem with 2 variables $x = (x_1, x_2)^T$ and the 3-dimensional ice-cream cone $\mathbf{K}$:

$$\min\left\{x_1 \mid Ax - b \equiv \begin{bmatrix} x_1 - x_2 \\ 1 \\ x_1 + x_2 \end{bmatrix} \geq_{\mathbf{L}^3} 0\right\}.$$

Recalling the definition of $\mathbf{L}^3$, we can write the problem equivalently as

$$\min\left\{x_1 \mid \sqrt{(x_1 - x_2)^2 + 1} \leq x_1 + x_2\right\},$$

i.e., as the problem

$$\min\left\{x_1 \mid 4x_1x_2 \geq 1, x_1 + x_2 > 0\right\}.$$

Geometrically the problem is to minimize x_1 over the intersection of the 3D ice-cream cone with a 2D plane; the inverse image of this intersection in the "design plane" of variables x_1, x_2 is part of the 2D nonnegative orthant bounded by the hyperbola $x_1x_2 \geq 1/4$. The problem is clearly strictly feasible (a strictly feasible solution is, e.g., $x = (1, 1)^T$) and bounded below, with the optimal value 0. This optimal value, however, is not achieved – the problem is unsolvable!

Example 1.7.2 Consider the following conic problem with two variables $x = (x_1, x_2)^T$ and the 3-dimensional ice-cream cone $\mathbf{K}$:

$$\min\left\{x_2 \mid Ax - b = \begin{bmatrix} x_1 \\ x_2 \\ x_1 \end{bmatrix} \geq_{\mathbf{L}^3} 0\right\}.$$

The problem is equivalent to the problem

$$\left\{x_2 \mid \sqrt{x_1^2 + x_2^2} \leq x_1\right\},$$

i.e., to the problem

$$\min\left\{x_2 \mid x_2 = 0, x_1 \geq 0\right\}.$$

The problem is clearly solvable, and its optimal set is the ray $\{x_1 \geq 0, x_2 = 0\}$.

Now let us build the conic dual to our (solvable!) primal. It is immediately seen that the cone dual to an ice-cream cone is this ice-cream cone itself. Thus, the dual problem is

$$\max_{\lambda}\left\{0 \mid \begin{bmatrix} \lambda_1 + \lambda_3 \\ \lambda_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \lambda \geq_{\mathbf{L}^3} 0\right\}.$$

In spite of the fact that primal is solvable, the dual is infeasible: indeed, assuming that λ is dual feasible, we have $\lambda \geq_{\mathbf{L}^3} 0$, which means that $\lambda_3 \geq \sqrt{\lambda_1^2 + \lambda_2^2}$; since also $\lambda_1 + \lambda_3 = 0$, we come to $\lambda_2 = 0$, which contradicts the equality $\lambda_2 = 1$.

We see that the weakness of the Conic Duality Theorem as compared to the LP Duality one reflects pathologies which indeed may happen in the general conic case.

1.7.2 Consequences of the Conic Duality Theorem

Sufficient condition for infeasibility. Recall that a necessary and sufficient condition for infeasibility of a (finite) system of scalar linear inequalities (i.e., for a vector inequality with respect to the partial ordering $\geq$) is the possibility to combine these inequalities in a linear fashion in such a way that the resulting scalar linear inequality is contradictory. In the case of cone-generated vector inequalities a slightly weaker result can be obtained:

Proposition 1.7.1 [Conic Theorem on Alternative] *Consider a linear vector inequality*

$$Ax - b \geq_{\mathbf{K}} 0. \tag{I}$$

(i) *If there exists λ satisfying*

$$\lambda \geq_{\mathbf{K}_*} 0, A^*\lambda = 0, \langle \lambda, b\rangle > 0, \tag{II}$$

then (I) *has no solutions.*

(ii) *If* (II) *has no solutions, then* (I) *is "almost solvable" – for every positive ϵ there exists b' such that $\|b' - b\|_2 < \epsilon$ and the perturbed system*

$$Ax - b' \geq_{\mathbf{K}} 0$$

is solvable.

Moreover,

(iii) (II) *is solvable if and only if* (I) *is not "almost solvable".*

Note the difference between the simple case when $\geq_{\mathbf{K}}$ is the usual partial ordering $\geq$ and the general case. In the former, one can replace in (ii) "nearly solvable" by "solvable"; however, in the general conic case "almost" is unavoidable.

Example 1.7.3 Let system (I) be given by

$$Ax - b \equiv \begin{bmatrix} x+1 \\ x-1 \\ \sqrt{2}x \end{bmatrix} \geq_{\mathbf{L}^3} 0.$$

Recalling the definition of the ice-cream cone $\mathbf{L}^3$, we can write the inequality equivalently as

$$\sqrt{2}x \geq \sqrt{(x+1)^2 + (x-1)^2} \equiv \sqrt{2x^2+2}, \tag{i}$$

which of course is unsolvable. The corresponding system (II) is

$$\begin{array}{cc} \lambda_3 \geq \sqrt{\lambda_1^2 + \lambda_2^2} & \left[\Leftrightarrow \lambda \geq_{\mathbf{L}^3_*} 0\right] \\ \lambda_1 + \lambda_2 + \sqrt{2}\lambda_3 = 0 & \left[\Leftrightarrow A^T\lambda = 0\right] \\ \lambda_2 - \lambda_1 > 0 & \left[\Leftrightarrow b^T\lambda > 0\right] \end{array} \tag{ii}$$

From the second of these relations, $\lambda_3 = -\frac{1}{\sqrt{2}}(\lambda_1 + \lambda_2)$, so that from the first inequality we get $0 \leq (\lambda_1 - \lambda_2)^2$, whence $\lambda_1 = \lambda_2$. But then the third inequality in (ii) is impossible! We see that here both (i) and (ii) have no solutions.

The geometry of the example is as follows. (i) asks to find a point in the intersection of the 3D ice-cream cone and a line. This line is an asymptote of the cone (it belongs to a 2D plane which crosses the cone in such way that the boundary of the cross-section is a branch of a hyperbola, and the line is one of two asymptotes of the hyperbola). Although the intersection is empty ((i) is unsolvable), small shifts of the line make the intersection nonempty (i.e., (i) is unsolvable and "almost solvable" at the same time). And it turns out that one cannot certify the fact that (i) itself is unsolvable by providing a solution to (ii).

Proof of the Proposition. (i) is evident (why?).

Let us prove (ii). To this end it suffices to verify that if (I) is not "almost solvable", then (II) is solvable. Let us fix a vector $\sigma >_{\mathbf{K}} 0$ and look at the conic problem

$$\min_{x,t} \{t \mid Ax + t\sigma - b \geq_{\mathbf{K}} 0\} \tag{CP}$$

in variables (x, t). Clearly, the problem is strictly feasible (why?). Now, if (I) is not almost solvable, then, first, the matrix of the problem $[A; \sigma]$ satisfies the full column rank condition **A** (otherwise the image of the mapping $(x, t) \mapsto Ax + t\sigma - b$ would coincide with the image of the mapping $x \mapsto Ax - b$, which is not he case – the first of these images does intersect $\mathbf{K}$, while the second does not). Second, the optimal value in (CP) is strictly positive (otherwise the problem would admit feasible solutions with t close to 0, and this would mean that (I) is almost solvable). From the Conic Duality Theorem it follows that the dual problem of (CP)

$$\max_{\lambda} \{\langle b, \lambda\rangle \mid A^*\lambda = 0, \langle \sigma, \lambda\rangle = 1, \lambda \geq_{\mathbf{K}_*} 0\}$$

has a feasible solution with positive $\langle b, \lambda\rangle$, i.e., (II) is solvable.

It remains to prove (iii). Assume first that (I) is not almost solvable; then (II) must be solvable by (ii). Vice versa, assume that (II) is solvable, and let λ be a solution to (II). Then λ solves also all systems of the type (II) associated with small enough perturbations of b instead of b itself; by (i), it implies that all inequalities obtained from (I) by small enough perturbation of b are unsolvable. ∎

When is a scalar linear inequality a consequence of a given linear vector inequality? The question we are interested in is as follows: given a linear vector inequality

$$Ax \geq_{\mathbf{K}} b \tag{V}$$

and a scalar inequality

$$c^T x \geq d \tag{S}$$

we want to check whether (S) is a consequence of (V). If $\mathbf{K}$ is the nonnegative orthant, the answer is given by the Inhomogeneous Farkas Lemma:

> *Inequality* (S) *is a consequence of a feasible system of linear inequalities $Ax \geq b$ if and only if* (S) *can be obtained from* (V) *and the trivial inequality $1 \geq 0$ in a linear fashion (by taking weighted sum with nonnegative weights).*

In the general conic case we can get a slightly weaker result:

Proposition 1.7.2 (i) *If* (S) *can be obtained from* (V) *and from the trivial inequality $1 \geq 0$ by admissible aggregation, i.e., there exist weight vector $\lambda \geq_{\mathbf{K}_*} 0$ such that*

$$A^*\lambda = c, \langle \lambda, b\rangle \geq d,$$

then (S) *is a consequence of* (V).

(ii) *If* (S) *is a consequence of a strictly feasible linear vector inequality* (V), *then* (S) *can be obtained from* (V) *by an admissible aggregation.*

The difference between the case of the partial ordering $\geq$ and a general partial ordering $\geq_{\mathbf{K}}$ is in the word "strictly" in (ii).

Proof of the proposition. (i) is evident (why?). To prove (ii), assume that (V) is strictly feasible and (S) is a consequence of (V) and consider the conic problem

$$\min_{x,t}\left\{t \mid \bar{A}\begin{pmatrix}x\\t\end{pmatrix}-\bar{b}\equiv\begin{bmatrix}Ax-b\\d-c^Tx+t\end{bmatrix}\geq_{\bar{\mathbf{K}}}0\right\},$$
$$\bar{\mathbf{K}}=\{(x,t)\mid x\in\mathbf{K},t\geq 0\}$$

The problem is clearly strictly feasible (choose x to be a strictly feasible solution to (V) and then choose t to be large enough). The fact that (S) is a consequence of (V) says exactly that the optimal value in the problem is nonnegative. By the Conic Duality Theorem, the dual problem

$$\max_{\lambda,\mu}\left\{\langle b,\lambda\rangle-d\mu \mid A^*\lambda-c=0,\mu=1,\begin{pmatrix}\lambda\\\mu\end{pmatrix}\geq_{\bar{\mathbf{K}}_*}0\right\}$$

has a feasible solution with the value of the objective ≥ 0. Since, as it is easily seen, $\bar{\mathbf{K}}_*=\{(\lambda,\mu)\mid \lambda\in\mathbf{K}_*,\mu\geq 0\}$, the indicated solution satisfies the requirements

$$\lambda\geq_{\mathbf{K}_*}0, A^*\lambda=c, \langle b,\lambda\rangle\geq d,$$

i.e., (S) can be obtained from (V) by an admissible aggregation. ∎

"Robust solvability status". Examples 1.7.2 – 1.7.3 make it clear that in the general conic case we may meet "pathologies" which do not occur in LP. E.g., a feasible and bounded problem may be unsolvable, the dual to a solvable conic problem may be infeasible, etc. Where the pathologies come from? Looking at our "pathological examples", we arrive at the following guess: the source of the pathologies is that *in these examples, the "solvability status" of the primal problem is non-robust – it can be changed by small perturbations of the data.* This issue of *robustness* is very important in modelling, and it deserves a careful investigation.

Data of a conic problem. When asked "What are the data of an LP program $\min\{c^Tx \mid Ax-b\geq 0\}$", everybody will give the same answer: "the objective c, the constraint matrix A and the right hand side vector b". Similarly, for a conic problem

$$\min\left\{c^Tx \mid Ax-b\geq_{\mathbf{K}}0\right\}, \tag{CP}$$

its data, *by definition*, is the triple (c,A,b), while the sizes of the problem – the dimension n of x and the dimension m of $\mathbf{K}$, same as the underlying cone $\mathbf{K}$ itself, are considered as the structure of (CP).

Robustness. A question of primary importance is whether the properties of the program (CP) (feasibility, solvability, etc.) are stable with respect to perturbations of the data. The reasons which make this question important are as follows:

- In actual applications, especially those arising in Engineering, the data are normally inexact: their true values, even when they "exist in the nature", are not known exactly when the problem is processed. Consequently, the results of the processing say something definite about the "true" problem only if these results are robust with respect to small data perturbations i.e., the properties of (CP) we have discovered are shared not only by the particular ("nominal") problem we were processing, but also by all problems with nearby data.

- Even when the exact data are available, we should take into account that processing them computationally we unavoidably add "noise" like rounding errors (you simply cannot load something like 1/7 to the standard computer). As a result, a real-life computational routine can recognize only those properties of the input problem which are stable with respect to small perturbations of the data.

Due to the above reasons, we should study not only whether a given problem (CP) is feasible/bounded/solvable, etc., but also whether these properties are robust – remain unchanged under small data perturbations. As it turns out, the Conic Duality Theorem allows to recognize "robust feasibility/boundedness/solvability...".

Let us start with introducing the relevant concepts. We say that (CP) is

- *robust feasible*, if all "sufficiently close" problems (i.e., those of the same structure $(n, m, \mathbf{K})$ and with data close enough to those of (CP)) are feasible;
- *robust infeasible*, if all sufficiently close problems are infeasible;
- *robust bounded below*, if all sufficiently close problems are bounded below (i.e., their objectives are bounded below on their feasible sets);
- *robust unbounded*, if all sufficiently close problems are not bounded;
- *robust solvable*, if all sufficiently close problems are solvable.

Note that a problem which is not robust feasible, *not necessarily* is robust infeasible, since among close problems there may be both feasible and infeasible (look at Example 1.7.2 – slightly shifting and rotating the plane $\mathrm{Im}\, A - b$, we may get whatever we want – a feasible bounded problem, a feasible unbounded problem, an infeasible problem...). This is why we need two kinds of definitions: one of "robust presence of a property" and one more of "robust absence of the same property".

Now let us look what are necessary and sufficient conditions for the most important robust forms of the "solvability status".

Proposition 1.7.3 [Robust feasibility] (CP) *is robust feasible if and only if it is strictly feasible, in which case the dual problem* (D) *is robust bounded above.*

Proof. The statement is nearly tautological. Let us fix $\delta >_{\mathbf{K}} 0$. If (CP) is robust feasible, then for small enough $t > 0$ the perturbed problem $\min\{c^T x \mid Ax - b - t\delta \geq_{\mathbf{K}} 0\}$ should be feasible; a feasible solution to the perturbed problem clearly is a strictly feasible solution to (CP). The inverse implication is evident (a strictly feasible solution to (CP) remains feasible for all problems with close enough data). It remains to note that if all problems sufficiently close to (CP) are feasible, then their duals, by the Weak Conic Duality Theorem, are bounded above, so that (D) is robust above bounded. ∎

Proposition 1.7.4 [Robust infeasibility] (CP) *is robust infeasible if and only if the system*

$$\langle b, \lambda\rangle = 1, A^*\lambda = 0, \lambda \geq_{\mathbf{K}_*} 0$$

is robust feasible, or, which is the same (by Proposition 1.7.3), if and only if the system

$$\langle b, \lambda\rangle = 1, A^*\lambda = 0, \lambda >_{\mathbf{K}_*} 0 \tag{1.7.3}$$

has a solution.

Proof. First assume that (1.7.3) is solvable, and let us prove that all problems sufficiently close to (CP) are infeasible. Let us fix a solution $\bar{\lambda}$ to (1.7.3). Since A is of full column rank, simple Linear Algebra says that the systems $[A']^*\lambda = 0$ are solvable for all matrices A' from a small enough neighbourhood U of A; moreover, the corresponding solution $\lambda(A')$ can be chosen to satisfy $\lambda(A) = \bar{\lambda}$ and to be continuous in $A' \in U$. Since $\lambda(A')$ is continuous and $\lambda(A) >_{\mathbf{K}_*} 0$, we have $\lambda(A')$ is $>_{\mathbf{K}_*} 0$ in a neighbourhood of A; shrinking U appropriately, we may assume that $\lambda(A') >_{\mathbf{K}_*} 0$ for all $A' \in U$. Now, $b^T\bar{\lambda} = 1$; by continuity reasons, there exists a neighbourhood V of b and a neighbourhood U' of A such that $b' \in V$ and all $A' \in U'$ one has $\langle b', \lambda(A')\rangle > 0$.

Thus, we have seen that there exist a neighbourhood U' of A and a neighbourhood V of b, along with a function $\lambda(A')$, $A' \in U'$, such that

$$\langle b'\lambda(A')\rangle > 0, [A']^*\lambda(A') = 0, \lambda(A') \geq_{\mathbf{K}_*} 0$$

for all $b' \in V$ and $A' \in U$. By Proposition 1.7.1.(i) it means that all the problems

$$\min\left\{[c']^T x \mid A'x - b' \geq_{\mathbf{K}} 0\right\}$$

with $b' \in V$ and $A' \in U'$ are infeasible, so that (CP) is robust infeasible.

Now let us assume that (CP) is robust infeasible, and let us prove that then (1.7.3) is solvable. Indeed, by the definition of robust infeasibility, there exist neighbourhoods U of A and V of b such that all vector inequalities

$$A'x - b' \geq_{\mathbf{K}} 0$$

with $A' \in U$ and $b' \in V$ are unsolvable. It follows that whenever $A' \in U$ and $b' \in V$, the vector inequality

$$A'x - b' \geq_{\mathbf{K}} 0$$

is <u>not</u> almost solvable (see Proposition 1.7.1). We conclude from Proposition 1.7.1.(ii) that for every $A' \in U$ and $b' \in V$ there exists $\lambda = \lambda(A', b')$ such that

$$\langle b', \lambda(A', b')\rangle > 0, [A']^*\lambda(A', b') = 0, \lambda(A', b') \geq_{\mathbf{K}_*} 0.$$

Now let us choose $\lambda_0 >_{\mathbf{K}_*} 0$. For all small enough positive ϵ we have $A_\epsilon = A + \epsilon b[A^*\lambda_0]^T \in U$. Let us choose an ϵ with the latter property to be so small that $\epsilon\langle b, \lambda_0\rangle > -1$ and set $A' = A_\epsilon$, $b' = b$. According to the previous observation, there exists $\lambda = \lambda(A', b)$ such that

$$\langle b, \lambda\rangle > 0, [A']^*\lambda \equiv A^*[\lambda + \epsilon\langle b, \lambda\rangle\lambda_0] = 0, \lambda \geq_{\mathbf{K}_*} 0.$$

Setting $\bar{\lambda} = \lambda + \epsilon\langle b, \lambda\rangle\lambda_0$, we get $\bar{\lambda} >_{\mathbf{K}_*} 0$ (since $\lambda \geq_{\mathbf{K}_*} 0, \lambda_0 >_{\mathbf{K}_*} 0$ and $\langle b, \lambda\rangle > 0$), while $A^*\bar{\lambda} = 0$ and $\langle b, \bar{\lambda}\rangle = \langle b, \lambda\rangle(1 + \epsilon\langle b, \lambda_0\rangle) > 0$. Multiplying $\bar{\lambda}$ by appropriate positive factor, we get a solution to (1.7.3). ∎

Now we are able to formulate our main result on "robust solvability".

Proposition 1.7.5 *For a conic problem* (CP) *the following conditions are equivalent to each other*

(i) (CP) *is robust feasible and robust bounded (below);*

(ii) (CP) *is robust solvable;*

(iii) (D) *is robust solvable;*

(iv) (D) *is robust feasible and robust bounded (above);*

(v) *Both* (CP) *and* (D) *are strictly feasible.*

In particular, under every one of these equivalent assumptions, both (CP) *and* (D) *are solvable with equal optimal values.*

Proof. (i) ⇒ (v): If (CP) is robust feasible, it also is strictly feasible (Proposition 1.7.3). If, in addition, (CP) is robust bounded below, then (D) is robust solvable (by the Conic Duality Theorem); in particular, (D) is robust feasible and therefore strictly feasible (again Proposition 1.7.3).

(v) ⇒ (ii): The implication is given by the Conic Duality Theorem.

(ii) ⇒ (i): trivial.

We have proved that (i)≡(ii)≡(v). Due to the primal-dual symmetry, we also have proved that (iii)≡(iv)≡(v). ■

1.8 Exercises

1.8.1 Around General Theorem on Alternative

Exercise 1.2 *Derive General Theorem on Alternative from Homogeneous Farkas Lemma*
Hint: Verify that the system

$$(\mathcal{S}): \quad \left\{ \begin{array}{rcl} a_i^T x & > & b_i,\ i = 1, ..., m_{\rm s}, \\ a_i^T x & \geq & b_i,\ i = m_{\rm s}+1, ..., m. \end{array} \right.$$

in variables x has no solution if and only if the homogeneous inequality

$$\epsilon \leq 0$$

in variables x, ϵ, t is a consequence of the system of homogeneous inequalities

$$\left\{ \begin{array}{rcl} a_i^T x - b_i t - \epsilon & \geq & 0,\ i = 1, ..., m_{\rm s}, \\ a_i^T x - b_i t & \geq & 0,\ i = m_{\rm s}+1, ..., m, \\ t & \geq & \epsilon, \end{array} \right.$$

in these variables.

There exist several particular cases of GTA (which in fact are equivalent to GTA); the goal of the next exercise is to prove the corresponding statements.

Exercise 1.3 *Derive the following statements from the General Theorem on Alternative:*

1. [Gordan's Theorem on Alternative] *One of the inequality systems*

$$\text{(I)} \quad Ax < 0,\ x \in \mathbf{R}^n,$$

$$\text{(II)} \quad A^T y = 0,\ 0 \neq y \geq 0,\ y \in \mathbf{R}^m,$$

(A being an $m \times n$ matrix, x are variables in (I)*, y are variables in* (II)*) has a solution if and only if the other one has no solutions.*

2. [Inhomogeneous Farkas Lemma] *A linear inequality in variables x*

$$a^T x \leq p \tag{1.8.1}$$

is a consequence of a solvable system of linear inequalities

$$Nx \leq q \tag{1.8.2}$$

if and only if it is a "linear consequence" of the system and the trivial inequality

$$0^T x \leq 1,$$

i.e., if it can be obtained by taking weighted sum, with nonnegative coefficients, of the inequalities from the system and this trivial inequality.

Algebraically: (1.8.1) is a consequence of solvable system (1.8.2) if and only if

$$a = N^T\nu$$

for some nonnegative vector ν such that

$$\nu^T q \le p.$$

3. [Motzkin's Theorem on Alternative] *The system*

$$Sx < 0,\ \ Nx \le 0$$

in variables x has no solutions if and only if the system

$$S^T\sigma + N^T\nu = 0,\ \sigma \ge 0,\ \ \nu \ge 0,\ \ \sigma \ne 0$$

in variables σ, ν has a solution.

Exercise 1.4 *Consider the linear inequality*

$$x + y \le 2$$

and the system of linear inequalities

$$\begin{cases} x \le 1 \\ -x \le -100 \end{cases}$$

Our inequality clearly is a consequence of the system – it is satisfied at every solution to it (simply because there are no solutions to the system at all). According to the Inhomogeneous Farkas Lemma, the inequality should be a linear consequence of the system and the trivial inequality $0 \le 1$, i.e., there should exist nonnegative ν_1, ν_2 such that

$$\begin{pmatrix} 1 \\ 1 \end{pmatrix} = \nu_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \nu_2 \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \quad \nu_1 - 1000\nu_2 \le 2,$$

which clearly is not the case. What is the reason for the observed "contradiction"?

1.8.2 Around cones

Attention! In what follows, if otherwise is not explicitly stated, "cone" is a shorthand for "closed pointed cone with a nonempty interior", $\mathbf{K}$ denotes a cone, and $\mathbf{K}_*$ is the cone dual to $\mathbf{K}$.

Exercise 1.5 *Let $\mathbf{K}$ be a cone, and let $\bar{x} >_{\mathbf{K}} 0$. Prove that $x >_{\mathbf{K}} 0$ if and only if there exists positive real t such that $x \ge_{\mathbf{K}} t\bar{x}$.*

Exercise 1.6 1) *Prove that if $0 \ne x \ge_{\mathbf{K}} 0$ and $\lambda >_{\mathbf{K}_*} 0$, then $\lambda^T x > 0$.*
2) *Assume that $\lambda \ge_{\mathbf{K}_*} 0$. Prove that $\lambda >_{\mathbf{K}_*} 0$ if and only if $\lambda^T x > 0$ whenever $0 \ne x \ge_{\mathbf{K}} 0$.*
3) *Prove that $\lambda >_{\mathbf{K}_*} 0$ if and only if the set*

$$\{x \ge_{\mathbf{K}} 0 \mid \lambda^T x \le 1\}$$

is compact.

Calculus of cones

Exercise 1.7 *Prove the following statements:*

1) [stability with respect to direct multiplication] *Let* $\mathbf{K}_i \subset \mathbf{R}^{n_i}$ *be cones,* $i = 1, ..., k$. *Prove that the direct product of the cones:*

$$\mathbf{K} = \mathbf{K}_1 \times ... \times \mathbf{K}_k = \{(x_1, ..., x_k) \mid x_i \in \mathbf{K}_i,\ i = 1, ..., k\}$$

is a cone in $\mathbf{R}^{n_1+...+n_k} = \mathbf{R}^{n_1} \times ... \times \mathbf{R}^{n_k}$.

Prove that the cone dual to $\mathbf{K}$ *is the direct product of the cones dual to* $\mathbf{K}_i$, $i = 1, .., k$.

2) [stability with respect to taking inverse image] *Let* $\mathbf{K}$ *be a cone in* $\mathbf{R}^n$ *and* $u \mapsto Au$ *be a linear mapping from certain* $\mathbf{R}^k$ *to* $\mathbf{R}^n$ *with trivial null space (*$\mathrm{Null}(A) = \{0\}$*) and such that* $\mathrm{Im}A \cap \mathrm{int}\,\mathbf{K} \neq \emptyset$. *Prove that the inverse image of* $\mathbf{K}$ *under the mapping:*

$$A^{-1}(\mathbf{K}) = \{u \mid Au \in \mathbf{K}\}$$

is a cone in $\mathbf{R}^k$.

Prove that the cone dual to $A^{-1}(\mathbf{K})$ *is* $A^T\mathbf{K}_*$, *i.e.*

$$(A^{-1}(\mathbf{K}))_* = \{A^T\lambda \mid \lambda \in \mathbf{K}_*\}.$$

3) [stability with respect to taking linear image] *Let* $\mathbf{K}$ *be a cone in* $\mathbf{R}^n$ *and* $y = Ax$ *be a linear mapping from* $\mathbf{R}^n$ *<u>onto</u>* $\mathbf{R}^N$ *(i.e., the image of* A *is the entire* $\mathbf{R}^N$*). Assume* $\mathrm{Null}(A) \cap \mathbf{K} = \{0\}$.

Prove that then the set

$$A\mathbf{K} = \{Ax \mid x \in \mathbf{K}\}$$

is a cone in $\mathbf{R}^N$.

Prove that the cone dual to $A\mathbf{K}$ *is*

$$(A\mathbf{K})_* = \{\lambda \in \mathbf{R}^N \mid A^T\lambda \in \mathbf{K}_*\}.$$

Demonstrate by example that if in the above statement the assumption $\mathrm{Null}(A) \cap \mathbf{K} = \{0\}$ *is weakened to* $\mathrm{Null}(A) \cap \mathrm{int}\,\mathbf{K} = \emptyset$, *then the set* $A(\mathbf{K})$ *may happen to be non-closed.*

Hint. Look what happens when the 3D ice-cream cone is projected onto its tangent plane.

Primal-dual pairs of cones and orthogonal pairs of subspaces

Exercise 1.8 *Let* A *be a* $m \times n$ *matrix of full column rank and* $\mathbf{K}$ *be a cone in* $\mathbf{R}^m$.

1) *Prove that <u>at least</u> one of the following facts always takes place:*

(i) *There exists a nonzero* $x \in \mathrm{Im}\,A$ *which is* $\geq_{\mathbf{K}} 0$;

(ii) *There exists a nonzero* $\lambda \in \mathrm{Null}(A^T)$ *which is* $\geq_{\mathbf{K}_*} 0$.

Geometrically: given a primal-dual pair of cones $\mathbf{K}$, $\mathbf{K}_*$ *and a pair* $L, L^\perp$ *of linear subspaces which are orthogonal complements of each other, we either can find a nontrivial ray in the intersection* $L \cap \mathbf{K}$, *or in the intersection* $L^\perp \cap \mathbf{K}_*$, *or both.*

2) *Prove that there exists* $\lambda \in \mathrm{Null}(A^T)$ *which is* $>_{\mathbf{K}_*} 0$ *(this is the strict version of* (ii)*) if and only if* (i) *is false. Prove that, similarly, there exists* $x \in \mathrm{Im}A$ *which is* $>_{\mathbf{K}} 0$ *(this is the strict version of* (i)*) if and only if* (ii) *is false.*

Geometrically: if $\mathbf{K}, \mathbf{K}_*$ *is a primal-dual pair of cones and* $L, L^\perp$ *are linear subspaces which are orthogonal complements of each other, then the intersection* $L \cap \mathbf{K}$ *is trivial (i.e., is the singleton* $\{0\}$*) if and only if the intersection* $L^\perp \cap \mathrm{int}\,\mathbf{K}_*$ *is nonempty.*

Several interesting cones

Given a cone $\mathbf{K}$ along with its dual $\mathbf{K}_*$, let us call a *complementary pair* every pair $x \in \mathbf{K}$, $\lambda \in \mathbf{K}_*$ such that

$$\lambda^T x = 0.$$

Recall that in "good cases" (e.g., under the premise of item 4 of the Conic Duality Theorem) a pair of feasible solutions (x, λ) of a primal-dual pair of conic problems

$$\min\left\{c^T x \mid Ax - b \geq_{\mathbf{K}} 0\right\}$$

$$\max\left\{b^T \lambda \mid A^T \lambda = c, \lambda \geq_{\mathbf{K}_*} 0\right\}$$

is primal-dual optimal if and only if the "primal slack" $y = Ax - b$ and λ are complementary.

Exercise 1.9 [Nonnegative orthant] *Prove that the n-dimensional nonnegative orthant $\mathbf{R}^n_+$ is a cone and that it is self-dual:*

$$(\mathbf{R}^n_+)_* = \mathbf{R}^n_+.$$

What are complementary pairs?

Exercise 1.10 [Ice-cream cone] *Let $\mathbf{L}^n$ be the n-dimensional ice-cream cone:*

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

1) *Prove that $\mathbf{L}^n$ is a cone.*
2) *Prove that the ice-cream cone is self-dual:*

$$(\mathbf{L}^n)_* = \mathbf{L}^n.$$

3) *Characterize the complementary pairs.*

Exercise 1.11 [Positive semidefinite cone] *Let $\mathbf{S}^n_+$ be the cone of $n \times n$ positive semidefinite matrices in the space $\mathbf{S}^n$ of symmetric $n \times n$ matrices. Assume that $\mathbf{S}^n$ is equipped with the Frobenius inner product*

$$\langle X, Y \rangle = \mathrm{Tr}(XY) = \sum_{i,j=1}^{n} X_{ij} Y_{ij}.$$

1) *Prove that $\mathbf{S}^n_+$ indeed is a cone.*
2) *Prove that the semidefinite cone is self-dual:*

$$(\mathbf{S}^n_+)_* = \mathbf{S}^n_+,$$

i.e., that the Frobenius inner products of a symmetric matrix Λ with all positive semidefinite matrices X of the same size are nonnegative if and only if the matrix Λ itself is positive semidefinite.

3) *Prove the following characterization of the complementary pairs:*

> *Two matrices $X \in \mathbf{S}^n_+$, $\Lambda \in (\mathbf{S}^n_+)_* \equiv \mathbf{S}^n_+$ are complementary (i.e., $\langle \Lambda, X \rangle = 0$) if and only if their matrix product is zero: $\Lambda X = X\Lambda = 0$. In particular, matrices from a complementary pair commute and therefore share a common orthonormal eigenbasis.*

1.8.3 Around conic problems

Several primal-dual pairs

Exercise 1.12 [The min-max Steiner problem] *Consider the problem as follows:*

Given N points $b_1, ..., b_N$ in $\mathbf{R}^n$, find a point $x \in \mathbf{R}^n$ which minimizes the maximum (Euclidean) distance from itself to the points $b_1, ..., b_N$, i.e., solve the problem

$$\min_x \max_{i=1,...,N} \|x - b_i\|_2.$$

Imagine, e.g., that $n = 2$, $b_1, ..., b_N$ are locations of villages and you are interested to locate a fire station for which the worst-case distance to a possible fire is as small as possible.

1) *Pose the problem as a conic quadratic one – a conic problem associated with a direct product of ice-cream cones.*

2) *Build the dual problem.*

3) *What is the geometric interpretation of the dual? Are the primal and the dual strictly feasible? Solvable? With equal optimal values? What is the meaning of the complementary slackness?*

Exercise 1.13 [The weighted Steiner problem] *Consider the problem as follows:*

Given N points $b_1, ..., b_N$ in $\mathbf{R}^n$ along with positive weights ω_i, $i = 1, ..., N$, find a point $x \in \mathbf{R}^n$ which minimizes the weighted sum of its (Euclidean) distances to the points $b_1, ..., b_N$, i.e., solve the problem

$$\min_x \sum_{i=1}^N \omega_i \|x - b_i\|_2.$$

Imagine, e.g., that $n = 2$, $b_1, ..., b_N$ are locations of N villages and you are interested to place a telephone station for which the total cost of cables linking the station and the villages is as small as possible. The weights can be interpreted as the per mile cost of the cables (they may vary from village to village due to differences in populations and, consequently, in the required capacities of the cables).

1) *Pose the problem as a conic quadratic one.*

2) *Build the dual problem.*

3) *What is the geometric interpretation of the dual? Are the primal and the dual strictly feasible? Solvable? With equal optimal values? What is the meaning of the complementary slackness?*

1.8.4 Feasible and level sets of conic problems

Attention! Remember that by our Assumption **A** matrix A below is of full column rank!

Consider a feasible conic problem

$$\min \left\{ c^T x \mid Ax - b \geq_{\mathbf{K}} 0 \right\}. \tag{CP}$$

In many cases it is important to know whether the problem has

1) *bounded feasible set* $\{x \mid Ax - b \geq_{\mathbf{K}} 0\}$

2) *bounded level sets*

$$\{x \mid Ax - b \geq_{\mathbf{K}} 0, c^T x \leq a\}$$

for all real a.

Exercise 1.14 *Let* (CP) *be feasible. Then the following four properties are equivalent:*

(i) *the feasible set of the problem is bounded;*

(ii) *the set of primal slacks* $Y = \{y \mid y \geq_{\mathbf{K}} 0, y = Ax - b\}$ *is bounded.*

(iii) $\mathrm{Im}\, A \cap \mathbf{K} = \{0\}$;

(iv) *the system of vector (in)equalities*

$$A^T\lambda = 0, \lambda >_{\mathbf{K}_*} 0$$

is solvable.

Corollary. The property of (CP) *to have a bounded feasible set is independent of the particular value of* b, *provided that with this* b (CP) *is feasible!*

Exercise 1.15 *Let problem* (CP) *be feasible. Prove that the following two conditions are equivalent:*

(i) (CP) *has bounded level sets;*

(ii) *The dual problem*

$$\max\left\{b^T\lambda \mid A^T\lambda = c, \lambda \geq_{\mathbf{K}_*} 0\right\}$$

is strictly feasible.

Corollary. The property of (CP) *to have bounded level sets is independent of the particular value of* b, *provided that with this* b (CP) *is feasible!*

Lecture 2

Conic Quadratic Programming

Several "generic" families of conic problems are of special interest, both from the viewpoint of theory and applications. The cones underlying these problems are simple enough, so that one can describe explicitly the dual cone; as a result, the general duality machinery we have developed becomes "algorithmic", as in the Linear Programming case. Moreover, in many cases this "algorithmic duality machinery" allows to understand more deeply the original model, to convert it into equivalent forms better suited for numerical processing, etc. The relative simplicity of the underlying cones also enables one to develop efficient computational methods for the corresponding conic problems. The most famous example of a "nice" generic conic problem is, doubtless, Linear Programming; however, it is not the only problem of this sort. Two other nice generic conic problems of extreme importance are *Conic Quadratic* and *Semidefinite* programs. We are about to consider the first of these two problems.

2.1 Conic Quadratic problems: preliminaries

Recall the definition of the m-dimensional ice-cream ($\equiv$second-order$\equiv$Lorentz) cone $\mathbf{L}^m$:

$$\mathbf{L}^m = \{x = (x_1, ..., x_m) \in \mathbf{R}^m : x_m \geq \sqrt{x_1^2 + ... + x_{m-1}^2}\}, \quad m \geq 2.$$

A *conic quadratic problem* is a conic problem

$$\min_x \left\{ c^T x : Ax - b \geq_{\mathbf{K}} 0 \right\} \tag{CP}$$

for which the cone $\mathbf{K}$ is a direct product of several ice-cream cones:

$$\begin{aligned} \mathbf{K} &= \mathbf{L}^{m_1} \times \mathbf{L}^{m_2} \times ... \times \mathbf{L}^{m_k} \\ &= \left\{ y = \begin{pmatrix} y[1] \\ y[2] \\ ... \\ y[k] \end{pmatrix} : y[i] \in \mathbf{L}^{m_i},\ i = 1, ..., k \right\}. \end{aligned} \tag{2.1.1}$$

In other words, a conic quadratic problem is an optimization problem with linear objective and finitely many *"ice-cream constraints"*

$$A_i x - b_i \geq_{\mathbf{L}^{m_i}} 0,\ i = 1, ..., k,$$

where

$$[A;b] = \left[\begin{array}{c} [A_1;b_1] \\ \hline [A_2;b_2] \\ \hline ... \\ \hline [A_k;b_k] \end{array}\right]$$

is the partition of the data matrix $[A;b]$ corresponding to the partition of y in (2.1.1). Thus, a conic quadratic program can be written as

$$\min_x \left\{ c^T x : A_i x - b_i \geq_{\mathbf{L}^{m_i}} 0,\ i = 1, ..., k \right\}. \tag{2.1.2}$$

Recalling the definition of the relation $\geq_{\mathbf{L}^m}$ and partitioning the data matrix $[A_i, b_i]$ as

$$[A_i; b_i] = \left[\begin{array}{cc} D_i & d_i \\ p_i^T & q_i \end{array}\right]$$

where D_i is of the size $(m_i - 1) \times \dim x$, we can write down the problem as

$$\min_x \left\{ c^T x : \|D_i x - d_i\|_2 \leq p_i^T x - q_i,\ i = 1, ..., k \right\}; \tag{QP}$$

this is the "most explicit" form is the one we prefer to use. In this form, D_i are matrices of the same row dimension as x, d_i are vectors of the same dimensions as the column dimensions of the matrices D_i, p_i are vectors of the same dimension as x and q_i are reals.

It is immediately seen that (2.1.1) is indeed a cone, in fact a self-dual one: $\mathbf{K}_* = \mathbf{K}$. Consequently, the problem dual to (CP) is

$$\max_\lambda \left\{ b^T \lambda : A^T \lambda = c,\ \lambda \geq_{\mathbf{K}} 0 \right\}.$$

Denoting $\lambda = \left(\begin{array}{c} \lambda_1 \\ \lambda_2 \\ ... \\ \lambda_k \end{array}\right)$ with m_i-dimensional blocks λ_i (cf. (2.1.1)), we can write the dual problem as

$$\max_{\lambda_1, ..., \lambda_m} \left\{ \sum_{i=1}^k b_i^T \lambda_i : \sum_{i=1}^k A_i^T \lambda_i = c,\ \lambda_i \geq_{\mathbf{L}^{m_i}} 0,\ i = 1, ..., k \right\}.$$

Recalling the meaning of $\geq_{\mathbf{L}^{m_i}} 0$ and representing $\lambda_i = \left(\begin{array}{c} \mu_i \\ \nu_i \end{array}\right)$ with scalar component ν_i, we finally come to the following form of the problem dual to (QP):

$$\max_{\mu_i, \nu_i} \left\{ \sum_{i=1}^k [\mu_i^T d_i + \nu_i q_i] : \sum_{i=1}^k [D_i^T \mu_i + \nu_i p_i] = c,\ \|\mu_i\|_2 \leq \nu_i,\ i = 1, ..., k \right\}. \tag{QD}$$

The design variables in (QD) are vectors μ_i of the same dimensions as the vectors d_i and reals ν_i, $i = 1, ..., k$.

Since from now on we will treat (QP) and (QD) as the standard forms of a conic quadratic problem and its dual, we now interpret for these two problems our basic assumption **A** from Lecture 2 and notions like feasibility, strict feasibility, boundedness, etc. Assumption **A** now reads (why?):

There is no nonzero x which is orthogonal to all rows of all matrices D_i and to all vectors p_i, $i = 1, ..., k$

and we always make this assumption by default. Now, among notions like feasibility, solvability, etc., the only notion which does need an interpretation is strict feasibility, which now reads as follows (why?):

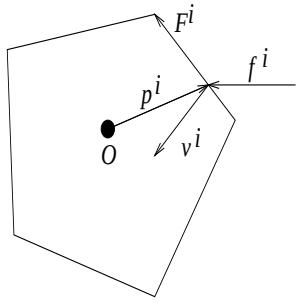
Strict feasibility of (QP) *means that there exist $\bar{x}$ such that $\|D_i\bar{x} - d_i\|_2 < p_i^T\bar{x} - q_i$ for all i.*

Strict feasibility of (QD) *means that there exists a feasible solution $\{\bar{\mu}_i, \bar{\nu}_i\}_{i=1}^k$ to the problem such that $\|\bar{\mu}_i\|_2 < \bar{\nu}_i$ for all $i = 1, ..., k$.*

2.2 Examples of conic quadratic problems

2.2.1 Contact problems with static friction [11]

Consider a rigid body in $\mathbf{R}^3$ and a robot with N fingers. When can the robot hold the body? To pose the question mathematically, let us look what happens at the point p^i of the body which is in contact with i-th finger of the robot:



Geometry of i-th contact
[p^i is the contact point; f^i is the contact force; v^i is the inward normal to the surface]

Let v^i be the unit inward normal to the surface of the body at the point p^i where i-th finger touches the body, f^i be the contact force exerted by i-th finger, and F^i be the friction force caused by the contact. Physics (Coulomb's law) says that the latter force is tangential to the surface of the body:

$$(F^i)^T v^i = 0 \tag{2.2.1}$$

and its magnitude cannot exceed μ times the magnitude of the normal component of the contact force, where μ is the friction coefficient:

$$\|F^i\|_2 \leq \mu(f^i)^T v^i. \tag{2.2.2}$$

Assume that the body is subject to additional external forces (e.g., gravity); as far as their mechanical consequences are concerned, all these forces can be represented by a single force – their sum – F^{ext} along with the *torque* T^{ext} – the sum of vector products of the external forces and the points where they are applied.

In order for the body to be in static equilibrium, the total force acting at the body and the total torque should be zero:

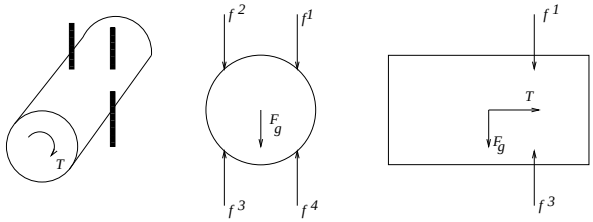
$$\begin{array}{rcl} \sum\limits_{i=1}^{N}(f^i+F^i)+F^{\rm ext} & = & 0 \\ \sum\limits_{i=1}^{N} p^i\times(f^i+F^i)+T^{\rm ext} & = & 0, \end{array} \qquad (2.2.3)$$

where $p\times q$ stands for the vector product of two 3D vectors p and q [1)].

The question "whether the robot is capable to hold the body" can be interpreted as follows. Assume that $f^i, F^{\rm ext}, T^{\rm ext}$ are given. If the friction forces F^i can adjust themselves to satisfy the friction constraints (2.2.1) – (2.2.2) and the equilibrium equations (2.2.3), i.e., if the system of constraints (2.2.1), (2.2.2), (2.2.3) with respect to unknowns F^i is solvable, then, and only then, the robot holds the body ("the body is in a *stable grasp*").

Thus, the question of stable grasp is the question of solvability of the system (S) of constraints (2.2.1), (2.2.2) and (2.2.3), which is a system of conic quadratic and linear constraints in the variables f^i, $F^{\rm ext}$, $T^{\rm ext}$, $\{F^i\}$. It follows that typical grasp-related optimization problems can be posed as CQPs. Here is an example:

> The robot should hold a cylinder by four fingers, all acting in the vertical direction. The external forces and torques acting at the cylinder are the gravity F_g and an externally applied torque T along the cylinder axis, as shown in the picture:



Perspective, front and side views

> The magnitudes ν_i of the forces f_i may vary in a given segment $[0, F_{\max}]$.
>
> What can be the largest magnitude τ of the external torque T such that a stable grasp is still possible?

Denoting by u^i the directions of the fingers, by v^i the directions of the inward normals to cylinder's surface at the contact points, and by u the direction of the axis of the cylinder, we

---

[1)] Here is the definition: if $p=\begin{pmatrix} p_1\\ p_2\\ p_3\end{pmatrix}$, $q=\begin{pmatrix} q_1\\ q_2\\ q_3\end{pmatrix}$ are two 3D vectors, then

$$[p,q]=\begin{pmatrix} {\rm Det}\begin{pmatrix} p_2 & p_3\\ q_2 & q_3\end{pmatrix}\\ {\rm Det}\begin{pmatrix} p_3 & p_1\\ q_3 & q_1\end{pmatrix}\\ {\rm Det}\begin{pmatrix} p_1 & p_2\\ q_1 & q_2\end{pmatrix}\end{pmatrix}$$

The vector $[p,q]$ is orthogonal to both p and q, and $\|[p,q]\|_2=\|p\|_2\|q\|_2\sin(\widehat{pq})$.

can pose the problem as the optimization program

$$\max \tau$$

s.t.

$$\begin{array}{rcll}
\sum_{i=1}^{4}(\nu_i u^i + F^i) + F_g & = & 0 & \text{[total force equals 0]} \\
\sum_{i=1}^{4} p^i \times (\nu_i u^i + F^i) + \tau u & = & 0 & \text{[total torque equals 0]} \\
(v^i)^T F^i & = & 0,\ i = 1,...,4 & [F^i \text{ are tangential to the surface}] \\
\|F^i\|_2 & \leq & [\mu[u^i]^T v^i]\nu_i,\ i = 1,...,4 & \text{[Coulomb's constraints]} \\
0 \leq \nu_i & \leq & F_{\max},\ i = 1,...,4 & \text{[bounds on } \nu_i\text{]}
\end{array}$$

in the design variables $\tau, \nu_i, F_i,\ i = 1,...,4$. This is a conic quadratic program, although not in the standard form (QP). To convert the problem to this standard form, it suffices, e.g., to replace all linear equalities by pairs of linear inequalities and further represent linear inequalities $\alpha^T x \leq \beta$ as conic quadratic constraints

$$Ax - b \equiv \left[\begin{array}{c} 0 \\ \beta - \alpha^T x \end{array} \right] \geq_{\mathbf{L}^2} 0.$$

2.3 What can be expressed via conic quadratic constraints?

Optimization problems arising in applications are not normally in their "catalogue" forms, and thus an important skill required from those interested in applications of Optimization is the ability to recognize the fundamental structure underneath the original formulation. The latter is frequently in the form

$$\min_x \{f(x) : x \in X\}, \tag{2.3.1}$$

where f is a "loss function", and the set X of admissible design vectors is typically given as

$$X = \bigcap_{i=1}^{m} X_i; \tag{2.3.2}$$

every X_i is the set of vectors admissible for a particular design restriction which in many cases is given by

$$X_i = \{x \in \mathbf{R}^n : g_i(x) \leq 0\}, \tag{2.3.3}$$

where $g_i(x)$ is i-th *constraint function*[2)].

It is well-known that the objective f in always can be assumed linear, otherwise we could move the original objective to the list of constraints, passing to the equivalent problem

$$\min_{t,x} \left\{ t : (t,x) \in \widehat{X} \equiv \{(x,t) : x \in X, t \geq f(f)\} \right\}.$$

Thus, we may assume that the original problem is of the form

$$\min_x \left\{ c^T x : x \in X = \bigcap_{i=1}^{m} X_i \right\}. \tag{P}$$

[2)] Speaking about a "real-valued function on $\mathbf{R}^n$", we assume that the function is allowed to take real values and the value $+\infty$ and is defined on the entire space. The set of those x where the function is finite is called the domain of the function, denoted by $\mathrm{Dom}\, f$.

In order to recognize that X is in one of our "catalogue" forms, one needs a kind of dictionary, where different forms of the same structure are listed. We shall build such a dictionary for the conic quadratic programs. Thus, our goal is to understand when a given set X can be represented by *conic quadratic inequalities* (c.q.i.'s), i.e., one or several constraints of the type $\|Dx - d\|_2 \leq p^T x - q$. The word "represented" needs clarification, and here it is:

> *We say that a set $X \subset \mathbf{R}^n$ can be represented via conic quadratic inequalities* (for short: is CQr – Conic Quadratic representable), *if there exists a system S of finitely many vector inequalities of the form $A_j \begin{pmatrix} x \\ u \end{pmatrix} - b_j \geq_{\mathbf{L}^{m_j}} 0$ in variables $x \in \mathbf{R}^n$ and additional variables u such that X is the projection of the solution set of S onto the x-space, i.e., $x \in X$ if and only if one can extend x to a solution (x, u) of the system S:*
>
> $$x \in X \Leftrightarrow \exists u : A_j \begin{pmatrix} x \\ u \end{pmatrix} - b_j \geq_{\mathbf{L}^{m_j}} 0, \ j = 1, ..., N.$$
>
> *Every such system S is called a conic quadratic representation* (for short: a CQR) *of the set X* [3)]

The idea behind this definition is clarified by the following observation:

> *Consider an optimization problem*
>
> $$\min_x \left\{ c^T x : x \in X \right\}$$
>
> *and assume that X is CQr. Then the problem is equivalent to a conic quadratic program. The latter program can be written down explicitly, provided that we are given a* CQR *of X.*
>
> Indeed, let S be a CQR of X, and u be the corresponding vector of additional variables. The problem
>
> $$\min_{x,u} \left\{ c^T x : (x, u) \text{ satisfy } S \right\}$$
>
> with design variables x, u is equivalent to the original problem (P), on one hand, and is a conic quadratic program, on the other hand.

Let us call a problem of the form (P) with CQ-representable X a *good* problem.

How to recognize good problems, i.e., how to recognize CQ-representable sets? Well, how we recognize continuity of a given function, like $f(x, y) = \exp\{\sin(x + \exp\{y\})\}$? Normally it is not done by a straightforward verification of the definition of continuity, but by using two kinds of tools:

A. We know a number of simple functions – a constant, $f(x) = x$, $f(x) = \sin(x)$, $f(x) = \exp\{x\}$, etc. – which indeed are continuous: "once for the entire life" we have verified it directly, by demonstrating that the functions fit the definition of continuity;

B. We know a number of basic continuity-preserving operations, like taking products, sums, superpositions, etc.

[3)] Note that here we do not impose on the representing system of conic quadratic inequalities S the requirement to satisfy assumption **A**; e.g., the entire space is CQr – it is a solution set of the "system" $|0^T x| \leq 1$ comprised of a single conic quadratic inequality.

When we see that a function is obtained from "simple" functions – those of type A – by operations of type B (as it is the case in the above example), we immediately infer that the function is continuous.

This approach which is common in Mathematics is the one we are about to follow. In fact, we need to answer two kinds of questions:

(?) What are CQ-representable sets

(??) What are CQ-representable functions $g(x)$, i.e., functions which possess CQ-representable *epigraphs*

$$\mathrm{Epi}\{g\} = \{(x,t) \in \mathbf{R}^n \times \mathbf{R} : g(x) \le t\}.$$

Our interest in the second question is motivated by the following

> **Observation:** *If a function g is CQ-representable, then so are all it level sets $\{x : g(x) \le a\}$, and every CQ-representation of (the epigraph of) g explicitly induces CQ-representations of the level sets.*
>
> Indeed, assume that we have a CQ-representation of the epigraph of g:
>
> $$g(x) \le t \Leftrightarrow \exists u : \|\alpha_j(x,t,u)\|_2 \le \beta_j(x,t,u),\ j = 1,...,N,$$
>
> where α_j and β_j are, respectively, vector-valued and scalar *affine* functions of their arguments. In order to get from this representation a CQ-representation of a level set $\{x : g(x) \le a\}$, it suffices to fix in the conic quadratic inequalities $\|\alpha_j(x,t,u)\|_2 \le \beta_j(x,t,u)$ the variable t at the value a.

We list below our "raw materials" – simple functions and sets admitting CQR's.

Elementary CQ-representable functions/sets

1. A constant function $g(x) \equiv a$.
Indeed, the epigraph of the function $\{(x,t) \mid a \le t\}$ is given by a linear inequality, and a linear inequality $0 \le p^T z - q$ is at the same time conic quadratic inequality $\|0\|_2 \le p^T z - q$.

2. An affine function $g(x) = a^T x + b$.
Indeed, the epigraph of an affine function is given by a linear inequality.

3. The Euclidean norm $g(x) = \|x\|_2$.
Indeed, the epigraph of g is given by the conic quadratic inequality $\|x\|_2 \le t$ in variables x, t.

4. The squared Euclidean norm $g(x) = x^T x$.
Indeed, $t = \frac{(t+1)^2}{4} - \frac{(t-1)^2}{4}$, so that

$$x^T x \le t \Leftrightarrow x^T x + \frac{(t-1)^2}{4} \le \frac{(t+1)^2}{4} \Leftrightarrow \left\| \begin{pmatrix} x \\ \frac{t-1}{2} \end{pmatrix} \right\|_2 \le \frac{t+1}{2}$$

(check the second $\Leftrightarrow$!), and the last relation is a conic quadratic inequality.

5. The fractional-quadratic function $g(x,s) = \begin{cases} \frac{x^Tx}{s}, & s>0 \\ 0, & s=0, x=0 \\ +\infty, & \text{otherwise} \end{cases}$ (x vector, s scalar).

Indeed, with the convention that $(x^Tx)/0$ is 0 or $+\infty$, depending on whether $x=0$ or not, and taking into account that $ts = \frac{(t+s)^2}{4} - \frac{(t-s)^2}{4}$, we have:

$$\{\tfrac{x^Tx}{s} \le t, s \ge 0\} \Leftrightarrow \{x^Tx \le ts, t \ge 0, s \ge 0\} \Leftrightarrow \{x^Tx + \tfrac{(t-s)^2}{4} \le \tfrac{(t+s)^2}{4}, t \ge 0, s \ge 0\}$$
$$\Leftrightarrow \left\| \begin{pmatrix} x \\ \frac{t-s}{2} \end{pmatrix} \right\|_2 \le \tfrac{t+s}{2}$$

(check the third $\Leftrightarrow$!), and the last relation is a conic quadratic inequality.

The level sets of the CQr functions 1 – 5 provide us with a spectrum of "elementary" CQr sets. We add to this spectrum one more set:

6. (A branch of) Hyperbola $\{(t,s) \in \mathbf{R}^2 : ts \ge 1, t > 0\}$.

Indeed,

$$\{ts \ge 1, t>0\} \Leftrightarrow \{\tfrac{(t+s)^2}{4} \ge 1 + \tfrac{(t-s)^2}{4} \ \&\ t>0\} \Leftrightarrow \{\left\| \begin{pmatrix} \frac{t-s}{2} \\ 1 \end{pmatrix} \right\|_2^2 \le \tfrac{(t+s)^2}{4}\}$$
$$\Leftrightarrow \{\left\| \begin{pmatrix} \frac{t-s}{2} \\ 1 \end{pmatrix} \right\|_2 \le \tfrac{t+s}{2}\}$$

(check the last $\Leftrightarrow$!), and the latter relation is a conic quadratic inequality.

Next we study simple operations preserving CQ-representability of functions/sets.

Operations preserving CQ-representability of sets

A. Intersection: If sets $X_i \subset \mathbf{R}^n$, $i=1,...,N$, are CQr, so is their intersection $X = \bigcap_{i=1}^N X_i$.

Indeed, let S_i be CQ-representation of X_i, and u_i be the corresponding vector of additional variables. Then the system S of constraints of the variables $(x, u_1, ..., u_N)$:

$$\{(x, u_i) \text{ satisfies } S_i\},\ i = 1, ..., N$$

is a system of conic quadratic inequalities, and this system clearly is a CQ-representation of X.

Corollary 2.3.1 *A polyhedral set – a set in $\mathbf{R}^n$ given by finitely many linear inequalities $a_i^Tx \le b_i$, $i=1,...,m$ – is CQr.*

Indeed, a polyhedral set is the intersection of finitely many level sets of affine functions, and all these functions (and thus – their level sets) are CQr.

Corollary 2.3.2 *If every one of the sets X_i in problem (P) is CQr, then the problem is good – it can be rewritten in the form of a conic quadratic problem, and such a transformation is readily given by CQR's of the sets X_i, $i=1,...,m$.*

Corollary 2.3.3 *Adding to a good problem finitely many CQr constraints $x \in X_i$, (e.g., finitely many scalar linear inequalities), we again get a good problem.*

B. Direct product: If sets $X_i \subset \mathbf{R}^{n_i}$, $i=1,...,k$, are CQr, then so is their direct product $X_1 \times ... \times X_k$.

Indeed, if $S_i = \{\|\alpha_j^i(x_i,u_i)\|_2 \le \beta_j^i(x_i,u_i)\}_{j=1}^{N_j}$, $i=1,...,k$, are CQR's of the sets X_i, then the union over i of this system of inequalities, regarded as a system with design variables $x=(x_1,...,x_k)$ and additional variables $u=(u_1,...,u_k)$ is a CQR for the direct product of $X_1,...,X_k$.

C. Affine image ("Projection"): If a set $X \subset \mathbf{R}^n$ is CQr and $x \mapsto y = \ell(x) = Ax + b$ is an affine mapping of $\mathbf{R}^n$ to $\mathbf{R}^k$, then the image $\ell(X)$ of the set X under the mapping is CQr.
Indeed, passing to an appropriate bases in $\mathbf{R}^n$ and $\mathbf{R}^k$, we may assume that the null space of A is comprised of the last $n-p$ vectors of the basis of $\mathbf{R}^n$, and that the image of A is spanned by the first p vectors of the basis in $\mathbf{R}^k$. In other words, we may assume that a vector $x \in \mathbf{R}^n$ can be partitioned as $x = \begin{pmatrix} x' \\ x'' \end{pmatrix}$ (x' is p-, and x'' is $(n-p)$-dimensional), and that a vector $y \in \mathbf{R}^k$ can be partitioned as $y = \begin{pmatrix} y' \\ y'' \end{pmatrix}$ (y' is p-, and y'' is $(k-p)$-dimensional) in such a way that $A \begin{pmatrix} x' \\ x'' \end{pmatrix} = \begin{pmatrix} Qx' \\ 0 \end{pmatrix}$ with a nonsingular $p \times p$ matrix Q. Thus,

$$\left\{ \begin{pmatrix} y' \\ y'' \end{pmatrix} = A \begin{pmatrix} x' \\ x'' \end{pmatrix} + b \right\} \Leftrightarrow \left\{ x = \begin{pmatrix} Q^{-1}(y' - b') \\ w \end{pmatrix} \text{ for some } w \ \& \ y'' = b'' \right\}.$$

Now let $S = \{\|\alpha_j(x,u)\|_2 \le \beta_j(x,u)\}_{j=1}^N$ be CQ-representation of X, where u is the corresponding vector of additional variables and α_j, β_j are affine in (x,u). Then the system of c.q.i.'s in the design variables $y = \begin{pmatrix} y' \\ y'' \end{pmatrix} \in \mathbf{R}^k$ and additional variables $w \in \mathbf{R}^{n-p}, u$:

$$S^+ = \{\|\alpha_j(\begin{pmatrix} Q^{-1}(y' - b') \\ w \end{pmatrix}, u)\|_2 \le \beta_j(\begin{pmatrix} Q^{-1}(y' - b') \\ w \end{pmatrix}, u)\}_{j=1}^N \ \& \ \{\|y'' - b''\|_2 \le 0\}$$

is a CQR of $\ell(X)$. Indeed, $y = \begin{pmatrix} y' \\ y'' \end{pmatrix} \in \ell(X)$ if and only if $y'' = b''$ and there exists $w \in \mathbf{R}^{n-p}$ such that the point $x = \begin{pmatrix} Q^{-1}(y' - b') \\ w \end{pmatrix}$ belongs to X, and the latter happens if and only if there exist u such that the point $(x,u) = (\begin{pmatrix} Q^{-1}(y' - b') \\ w \end{pmatrix}, u)$ solves S.

Corollary 2.3.4 *A nonempty set X is CQr if and only if its characteristic function*

$$\chi(x) = \begin{cases} 0, & x \in X \\ +\infty, & \text{otherwise} \end{cases}$$

is CQr.

Indeed, Epi$\{\chi\}$ is the direct product of X and the nonnegative ray; therefore if X is CQr, so is $\chi(\cdot)$ (see B. and Corollary 2.3.1). Vice versa, if χ is CQr, then X is CQr by C., since X is the projection of the Epi$\{\chi\}$ on the space of x-variables.

D. Inverse affine image: Let $X \subset \mathbf{R}^n$ be a CQr set, and let $\ell(y) = Ay + b$ be an affine mapping from $\mathbf{R}^k$ to $\mathbf{R}^n$. Then the inverse image $\ell^{-1}(X) = \{y \in \mathbf{R}^k : Ay + b \in X\}$ of X under the mapping is also CQr.
Indeed, let $S = \{\|\alpha_j(x,u)\|_2 \le \beta_j(x,u)\}_{i=1}^N$ be a CQR for X. Then the system of c.q.i.'s

$$S = \{\|\alpha_j(Ay + b, u)\|_2 \le \beta_j(Ay + b, u)\}_{i=1}^N$$

with variables y, u clearly is a CQR for $\ell^{-1}(X)$.

Corollary 2.3.5 *Consider a good problem* (P) *and assume that we restrict its design variables to be given affine functions of a new design vector y. Then the induced problem with the design vector y is also good.*

In particular, adding to a good problem arbitrarily many linear equality constraints, we end up with a good problem (Indeed, we may use the linear equations to express affinely the original design variables via part of them, let this part be y; the problem with added linear constraints can now be posed as a problem with design vector y).

It should be stressed that the above statements are not just existence theorems – they are "algorithmic": given CQR's of the "operands" (say, m sets $X_1, ..., X_m$), we may build *completely mechanically* a CQR for the "result of the operation" (e.g., for the intersection $\bigcap_{i=1}^{m} X_i$).

Operations preserving CQ-representability of functions

Recall that a function $g(x)$ is called CQ-representable, if its epigraph $\mathrm{Epi}\{g\} = \{(x,t) : g(x) \le t\}$ is a CQ-representable set; a CQR of the epigraph of g is called conic quadratic representation of g. Recall also that a level set of a CQr function is CQ-representable. Here are transformations preserving CQ-representability of functions:

E. Taking maximum: If functions $g_i(x)$, $i = 1, ..., m$, are CQr, then so is their maximum $g(x) = \max_{i=1,...,m} g_i(x)$.
Indeed, $\mathrm{Epi}\{g\} = \bigcap_i \mathrm{Epi}\{g_i\}$, and the intersection of finitely many CQr sets again is CQr.

F. Summation with nonnegative weights: If functions $g_i(x)$, $x \in \mathbf{R}^n$, are CQr, $i = 1, ..., m$, and α_i are nonnegative weights, then the function $g(x) = \sum_{i=1}^{m} \alpha_i g_i(x)$ is also CQr.
Indeed, consider the set

$$\Pi = \{(x_1, t_1; x_2, t_2; ...; x_m, t_m; t) : x_i \in \mathbf{R}^n, t_i \in \mathbf{R}, t \in \mathbf{R}, g_i(x_i) \le t_i, i = 1, ..., m; \sum_{i=1}^{m} \alpha_i t_i \le t\}.$$

The set is CQr. Indeed, the set is the direct product of the epigraphs of g_i intersected with the half-space given by the linear inequality $\sum_{i=1}^{m} \alpha_i t_i \le t$. Now, a direct product of CQr sets is also CQr, a half-space is CQr (it is a level set of an affine function, and such a function is CQr), and the intersection of CQr sets is also CQr. Since Π is CQr, so is its projection on subspace of variables $x_1, x_2, ..., x_m, t$, i.e., the set

$$\{(x_1, ..., x_m, t) : \exists t_1, ..., t_m : g_i(x_i) \le t_i, i = 1, ..., m, \sum_{i=1}^{m} \alpha_i t_i \le t\} = \{(x_1, ..., x_m, t) : \sum_{i=1}^{m} \alpha_i g_i(x) \le t\}.$$

Since the latter set is CQr, so is its inverse image under the mapping

$$(x, t) \mapsto (x, x, ...x, t),$$

and this inverse image is exactly the epigraph of g.

G. Direct summation: If functions $g_i(x_i)$, $x_i \in \mathbf{R}^{n_i}$, $i = 1, ..., m$, are CQr, so is their direct sum

$$g(x_1, ..., x_m) = g_1(x_1) + ... + g_m(x_m).$$

Indeed, the functions $\hat{g}_i(x_1, ..., x_m) = g_i(x_i)$ are clearly CQr – their epigraphs are inverse images of the epigraphs of g_i under the affine mappings $(x_1, ..., x_m, t) \mapsto (x_i, t)$. It remains to note that $g = \sum_i \hat{g}_i$.

H. Affine substitution of argument: If a function $g(x)$, $x \in \mathbf{R}^n$, is CQr and $y \mapsto Ay + b$ is an affine mapping from $\mathbf{R}^k$ to $\mathbf{R}^n$, then the superposition $g^{\rightarrow}(y) = g(Ay + b)$ is CQr.
Indeed, the epigraph of $g^{\rightarrow}$ is the inverse image of the epigraph of g under the affine mapping $(y, t) \mapsto (Ay + b, t)$.

I. Partial minimization: Let $g(x)$ be CQr. Assume that x is partitioned into two sub-vectors: $x = (v, w)$, and let $\hat{g}$ be obtained from g by partial minimization in w:

$$\hat{g}(v) = \inf_w g(v, w),$$

and assume that for every v the minimum in w is achieved. Then $\hat{g}$ is CQr.
Indeed, under the assumption that the minimum in w always is achieved, $\mathrm{Epi}\{\hat{g}\}$ is the image of the epigraph of $\mathrm{Epi}\{g\}$ under the projection $(v, w, t) \mapsto (v, t)$.

More operations preserving CQ-representability

Let us list a number of more "advanced" operations with sets/functions preserving CQ-representability.

J. Arithmetic summation of sets. Let X_i, $i = 1, ..., k$, be nonempty convex sets in $\mathbf{R}^n$, and let $X_1 + X_2 + ... + X_k$ be the arithmetic sum of these sets:

$$X_1 + ... + X_k = \{x = x^1 + ... + x^k : x^i \in X_i,\ i = 1, ..., k\}.$$

We claim that

If all X_i are CQr, so is their sum.

Indeed, the direct product

$$X = X_1 \times X_2 \times ... \times X_k \subset \mathbf{R}^{nk}$$

is CQr by B.; it remains to note that $X_1 + ... + X_k$ is the image of X under the linear mapping

$$(x^1, ..., x^k) \mapsto x^1 + ... + x^k : \mathbf{R}^{nk} \to \mathbf{R}^n,$$

and by C. the image of a CQr set under an affine mapping is also CQr (see C.)

J.1. inf-convolution. The operation with functions related to the arithmetic summation of sets is the inf-convolution defined as follows. Let $f_i : \mathbf{R}^n \to \mathbf{R} \cup \{\infty\}$, $i = 1, ..., n$, be functions. Their inf-convolution is the function

$$f(x) = \inf\{f_1(x^1) + ... + f_k(x^k) : x^1 + ... + x^k = x\}. \qquad (*)$$

We claim that

If all f_i are CQr, their inf*-convolution is $> -\infty$ everywhere and for every x for which the* inf *in the right hand side of (*) is finite, this infimum is achieved, then f is CQr.*

Indeed, under the assumption in question the epigraph $\mathrm{Epi}\{f\} = \mathrm{Epi}\{f_1\} + ... + \mathrm{Epi}\{f_k\}$.

K. Taking conic hull of a convex set. Let $X \in \mathbf{R}^n$ be a nonempty convex set. Its *conic hull* is the set

$$X^+ = \{(x, t) \in \mathbf{R}^n \times \mathbf{R} : t > 0, t^{-1}x \in X\}.$$

Geometrically: we add to the coordinates of vectors from $\mathbf{R}^n$ a new coordinate equal to 1:

$$(x_1, ..., x_n)^T \mapsto (x_1, ..., x_n, 1)^T,$$

thus getting an affine embedding of $\mathbf{R}^n$ in $\mathbf{R}^{n+1}$. We take the image of X under this mapping – "lift" X by one along the $(n+1)$st axis – and then form the set X^+ by taking all (open) rays emanating from the origin and crossing the "lifted" X.

The conic hull is not closed (e.g., it does not contain the origin, which clearly is in its closure). The closed convex hull of X is the closure of its conic hull:

$$\widehat{X}^+ = \mathrm{cl}\, X^+ = \left\{(x,t) \in \mathbf{R}^n \times \mathbf{R} : \exists\{(x_i,t_i)\}_{i=1}^\infty : t_i > 0, t_i^{-1}x_i \in X, t = \lim_i t_i, x = \lim_i x_i\right\}.$$

Note that if X is a closed convex set, then the conic hull X^+ of X is nothing but the intersection of the closed convex hull $\widehat{X}^+$ and the open half-space $\{t > 0\}$ (check!); thus, the closed conic hull of a closed convex set X is larger than the conic hull by some part of the hyperplane $\{\tau = 0\}$. When X is closed *and bounded*, then the difference between the hulls is pretty small: $\widehat{X}^+ = X^+ \cup \{0\}$ (check!). Note also that if X is a closed convex set, you can obtain it from its (closed) convex hull by taking intersection with the hyperplane $\{t = 1\}$:

$$x \in X \Leftrightarrow (x,1) \in \widehat{X}^+ \Leftrightarrow (x,1) \in X^+.$$

Proposition 2.3.1 (i) *If a set X is CQr:*

$$X = \{x : \exists u : Ax + Bu + b \geq_{\mathbf{K}} 0\}, \tag{2.3.4}$$

where $\mathbf{K}$ is a direct product of the ice-cream cones, then the conic hull X^+ is CQr as well:

$$X^+ = \{(x,t) : \exists(u,s) : Ax + Bu + tb \geq_{\mathbf{K}} 0, \begin{bmatrix} 2 \\ s-t \\ s+t \end{bmatrix} \geq_{\mathbf{L}^3} 0\}. \tag{2.3.5}$$

(ii) *If the set X given by* (2.3.4) *is closed, then the CQr set*

$$\widetilde{X}^+ = \{(x,t) : \exists u : Ax + Bu + tb \geq_{\mathbf{K}} 0\} \bigcap \{(x,t) : t \geq 0\} \tag{2.3.6}$$

is "between" the conic hull X^+ and the closed conic hull $\widehat{X}^+$ of X:

$$X^+ \subset \widetilde{X}^+ \subset \widehat{X}^+.$$

(iii) *If the CQR* (2.3.4) *is such that $Bu \in \mathbf{K}$ implies that $Bu = 0$, then $\widetilde{X}^+ = \widehat{X}^+$, so that $\widehat{X}^+$ is CQr.*

(i): We have

$$\begin{array}{rcl} X^+ & \equiv & \{(x,t) : t > 0, x/t \in X\} \\ & = & \{(x,t) : \exists u : A(x/t) + Bu + b \geq_{\mathbf{K}} 0, t > 0\} \\ & = & \{(x,t) : \exists v : Ax + Bv + tb \geq_{\mathbf{K}} 0, t > 0\} \\ & = & \{(x,t) : \exists v, s : Ax + Bv + tb \geq_{\mathbf{K}} 0, t, s \geq 0, ts \geq 1\}, \end{array}$$

and we arrive at (2.3.5).

(ii): We should prove that the set $\widetilde{X}^+$ (which by construction is CQr) is between X^+ and $\widehat{X}^+$. The inclusion $X^+ \subset \widetilde{X}^+$ is readily given by (2.3.5). Next, let us prove that $\widetilde{X}^+ \subset \widehat{X}^+$. Let us choose a point $\bar{x} \in X$, so that for a properly chosen $\bar{u}$ it holds

$$A\bar{x} + B\bar{u} + b \geq_{\mathbf{K}} 0,$$

i.e., $(\bar{x}, 1) \in \widetilde{X}^+$. Since $\widetilde{X}^+$ is convex (this is true for every CQr set), we conclude that whenever (x,t) belongs to $\widetilde{X}^+$, so does every pair $(x_\epsilon = x + \epsilon\bar{x}, t_\epsilon = t + \epsilon)$ with $\epsilon > 0$:

$$\exists u = u_\epsilon : Ax_\epsilon + Bu_\epsilon + t_\epsilon b \geq_{\mathbf{K}} 0.$$

It follows that $t_\epsilon^{-1}x_\epsilon \in X$, whence $(x_\epsilon, t_\epsilon) \in X^+ \subset \widehat{X}^+$. As $\epsilon \to +0$, we have $(x_\epsilon, t_\epsilon) \to (x,t)$, and since $\widehat{X}^+$ is closed, we get $(x,t) \in \widehat{X}^+$. Thus, $\widetilde{X}^+ \subset \widehat{X}^+$.

(ii): Assume that $Bu \in \mathbf{K}$ only if $Bu = 0$, and let us show that $\widetilde{X}^+ = \widehat{X}^+$. We just have to prove that $\widetilde{X}^+$ is closed, which indeed is the case due to the following

Lemma 2.3.1 *Let Y be a CQr set with CQR*

$$Y = \{y : \exists v : Py + Qv + r \geq_{\mathbf{K}} 0\}$$

such that $Qv \in \mathbf{K}$ only when $Qv = 0$. Then

(i) *There exists a constant $C < \infty$ such that*

$$Py + Qv + r \in \mathbf{K} \Rightarrow \|Qv\|_2 \leq C(1 + \|Py + r\|_2); \tag{2.3.7}$$

(ii) *Y is closed.*

Proof of Lemma. (i): Assume, on the contrary to what should be proved, that there exists a sequence $\{y_i, v_i\}$ such that

$$Py_i + Qv_i + r \in \mathbf{K},\ \|Qv_i\|_2 \geq \alpha_i(1 + \|Py_i + r\|_2),\ _i \to \infty \text{ as } i \to \infty. \tag{2.3.8}$$

By Linear Algebra, for every b such that the linear system $Qv = b$ is solvable, it admits a solution v such that $\|v\|_2 \leq C_1\|b\|_2$ with $C_1 < \infty$ depending on Q only; therefore we can assume, in addition to (2.3.8), that

$$\|v_i\|_2 \leq C_1\|Qv_i\|_2 \tag{2.3.9}$$

for all i. Now, from (2.3.8) it clearly follows that

$$\|Qv_i\|_2 \to \infty \text{ as } i \to \infty; \tag{2.3.10}$$

setting

$$\widehat{v}_i = \frac{1}{\|Qv_i\|_2} v_i,$$

we have

$$\begin{array}{lcll}
(a) & \|Q\widehat{v}_i\|_2 = 1 \quad \forall i, & & \\
(b) & \|\widehat{v}_i\| \leq C_1 \quad \forall i, & & [\text{by (2.3.9)}] \\
(c) & Q\widehat{v}_i + \|Qv_i\|_2^{-1}(Py_i + r) \in \mathbf{K} \quad \forall i, & & \\
(d) & \|Qv_i\|_2^{-1}\|Py_i + r\|_2 \leq \alpha_i^{-1} \to 0 \text{ as } i \to \infty & & [\text{by (2.3.8)}]
\end{array}$$

Taking into account (b) and passing to a subsequence, we can assume that $\widehat{v}_i \to \widehat{v}$ as $i \to \infty$; by (c, d) $Q\widehat{v} \in \mathbf{K}$, while by (a) $\|Q\widehat{v}\|_2 = 1$, i.e., $Q\widehat{v} \neq 0$, which is the desired contradiction.

(ii) To prove that Y is closed, assume that $y_i \in Y$ and $y_i \to y$ as $i \to \infty$, and let us verify that $y \in Y$. Indeed, since $y_i \in Y$, there exist v_i such that $Py_i + Qv_i + r \in \mathbf{K}$. Same as above, we can assume that (2.3.9) holds. Since $y_i \to y$, the sequence $\{Qv_i\}$ is bounded by (2.3.7), so that the sequence $\{v_i\}$ is bounded by (2.3.9). Passing to a subsequence, we can assume that $v_i \to v$ as $i \to \infty$; passing to the limit, as $i \to \infty$, in the inclusion $Py_i + Qv_i + r \in \mathbf{K}$, we get $Py + Qv + r \in \mathbf{K}$, i.e., $y \in Y$. ∎

K.1. "Projective transformation" of a CQr function. The operation with functions related to taking conic hull of a convex set is the "projective transformation" which converts a function $f(x) : \mathbf{R}^n \to \mathbf{R} \cup \{\infty\}$ [4] into the function

$$f^+(x, s) = sf(x/s) : \{s > 0\} \times \mathbf{R}^n \to \mathbf{R} \cup \{\infty\}.$$

The epigraph of f^+ is the conic hull of the epigraph of f with the origin excluded:

$$\begin{array}{rcl}
\{(x, s, t) : s > 0, t \geq f^+(x, s)\} & = & \left\{(x, s, t) : s > 0, s^{-1}t \geq f(s^{-1}x)\right\} \\
& = & \left\{(x, s, t) : s > 0, s^{-1}(x, t) \in \mathrm{Epi}\{f\}\right\}.
\end{array}$$

The closure $\mathrm{cl}\,\mathrm{Epi}\{f^+\}$ is the epigraph of certain function, let it be denoted $\widehat{f}^+(x, s)$; this function is called the *projective transformation* of f. E.g., the fractional-quadratic function from Example 5 is

[4] Recall that "a function" for us means a *proper* function – one which takes a finite value at least at one point

the projective transformation of the function $f(x) = x^T x$. Note that the function $\widehat{f}^+(x,s)$ does not necessarily coincide with $f^+(x,s)$ even in the open half-space $s > 0$; this is the case if and only if the epigraph of f is closed (or, which is the same, f is *lower semicontinuous*: whenever $x_i \to x$ and $f(x_i) \to a$, we have $f(x) \le a$). We are about to demonstrate that the projective transformation "nearly" preserves CQ-representability:

Proposition 2.3.2 *Let $f : \mathbf{R}^n \to \mathbf{R} \cup \{\infty\}$ be a lower semicontinuous function which is CQr:*

$$\mathrm{Epi}\{f\} \equiv \{(x,t) : t \ge f(x)\} = \{(t,x) : \exists u : Ax + tp + Bu + b \ge_{\mathbf{K}} 0\}, \tag{2.3.11}$$

where $\mathbf{K}$ is a direct product of ice-cream cones. Assume that the CQR is such that $Bu \ge_{\mathbf{K}} 0$ implies that $Bu = 0$. Then the projective transformation $\widehat{f}^+$ of f is CQr, namely,

$$\mathrm{Epi}\{\widehat{f}^+\} = \{(x,t,s) : s \ge 0, \exists u : Ax + tp + Bu + sb \ge_{\mathbf{K}} 0\}.$$

Indeed, let us set

$$G = \{(x,t,s) : \exists u : s \ge 0, Ax + tp + Bu + sb \ge_{\mathbf{K}} 0\}.$$

As we remember from the previous combination rule, G is exactly the closed conic hull of the epigraph of f, i.e., $G = \mathrm{Epi}\{\widehat{f}^+\}$.

L. The polar of a convex set. Let $X \subset \mathbf{R}^n$ be a convex set containing the origin. The *polar* of X is the set

$$X_* = \left\{y \in \mathbf{R}^n : y^T x \le 1 \ \forall x \in X\right\}.$$

In particular,

- the polar of the singleton $\{0\}$ is the entire space;
- the polar of the entire space is the singleton $\{0\}$;
- the polar of a linear subspace is its orthogonal complement (why?);
- the polar of a closed convex pointed cone K with a nonempty interior is $-K_*$, *minus* the dual cone (why?).

Polarity is "symmetric": if X is a closed convex set containing the origin, then so is X_*, and twice taken polar is the original set: $(X_*)_* = X$.

We are about to prove that the polarity $X \mapsto X_*$ "nearly" preserves CQ-representability:

Proposition 2.3.3 *Let $X \subset \mathbf{R}^n$, $0 \in X$, be a CQr set:*

$$X = \{x : \exists u : Ax + Bu + b \ge_{\mathbf{K}} 0\}, \tag{2.3.12}$$

where $\mathbf{K}$ is a direct product of ice-cream cones.

Assume that there exists $\bar{x}, \bar{u}$ such that

$$A\bar{x} + B\bar{u} + b >_{\mathbf{K}} 0.$$

Then the polar of X is the CQr set

$$X_* = \left\{y : \exists \xi : A^T\xi + y = 0, B^T\xi = 0, b^T\xi \le 1, \xi \ge_{\mathbf{K}} 0\right\} \tag{2.3.13}$$

Indeed, consider the following conic quadratic problem:

$$\min_{x,u} \left\{-y^T x : Ax + Bu + b \ge_{\mathbf{K}} 0\right\}. \tag{P_y}$$

A vector y belongs to X_* if and only if (P_y) is bounded below and its optimal value is at least -1. Since (P_y) is strictly feasible, from the Conic Duality Theorem it follows that these properties of (P_y) hold if and only if the dual problem

$$-b^T\xi \to \max : A^T\xi = -y, B^T\xi = 0, \xi \geq_{\mathbf{K}} 0$$

(recall that $\mathbf{K}$ is self-dual) has a feasible solution with the value of the dual objective at least -1. Thus,

$$X_* = \left\{y : \exists \xi : A^T\xi + y = 0, B^T\xi = 0, b^T\xi \leq 1, \xi \geq_{\mathbf{K}} 0\right\},$$

as claimed in (2.3.13). It remains to note that X_* is obtained from the CQr set $\mathbf{K}$ by operations preserving CQ-representability: intersection with the CQr set $\{\xi : B^T\xi = 0, b^T\xi \leq 1\}$ and subsequent affine mapping $\xi \mapsto -A^T\xi$.

L.1. The Legendre transformation of a CQr function. The operation with functions related to taking polar of a convex set is the *Legendre* (or *conjugate*) *transformation*. The Legendre transformation ($\equiv$ the conjugate) of a function $f(x) : \mathbf{R}^n \to \mathbf{R} \cup \{\infty\}$ is the function

$$f_*(y) = \sup_x \left[y^T x - f(x)\right].$$

In particular,

- the conjugate of a constant $f(x) \equiv c$ is the function
$$f_*(y) = \begin{cases} -c, & y = 0 \\ +\infty, & y \neq 0 \end{cases};$$
- the conjugate of an affine function $f(x) \equiv a^T x + b$ is the function
$$f_*(y) = \begin{cases} -b, & y = a \\ +\infty, & y \neq a \end{cases};$$
- the conjugate of a convex quadratic form $f(x) \equiv \frac{1}{2}x^T D^T D x + b^T x + c$ with rectangular D such that $\mathrm{Null}(D^T) = \{0\}$ is the function
$$f_*(y) = \begin{cases} \frac{1}{2}(y-b)^T D^T (DD^T)^{-2} D(y-b) - c, & y - b \in \mathrm{Im}D^T \\ +\infty, & \text{otherwise} \end{cases};$$

It is worth mentioning that the Legendre transformation is symmetric: if f is a proper convex lower semicontinuous function (i.e., $\emptyset \neq \mathrm{Epi}\{f\}$ is convex and closed), then so is f_*, and taken twice, the Legendre transformation recovers the original function: $(f_*)_* = f$.

We are about to prove that the Legendre transformation "nearly" preserves CQ-representability:

Proposition 2.3.4 *Let $f : \mathbf{R}^n \to \mathbf{R} \cup \{\infty\}$ be CQr:*

$$\{(x,t) : t \geq f(x)\} = \{(t,x) : \exists u : Ax + tp + Bu + b \geq_{\mathbf{K}} 0\},$$

where $\mathbf{K}$ is a direct product of ice-cream cones. Assume that there exist $\bar{x}, \bar{t}, \bar{u}$ such that

$$A\bar{x} + \bar{t}p + B\bar{u} + b >_{\mathbf{K}} 0.$$

Then the Legendre transformation of f is CQr:

$$\mathrm{Epi}\{f_*\} = \left\{(y,s) : \exists \xi : A^T\xi = -y, B^T\xi = 0, p^T\xi = 1, s \geq b^T\xi, \xi \geq_{\mathbf{K}} 0\right\}. \tag{2.3.14}$$

Indeed, we have

$$\mathrm{Epi}\{f_*\} = \left\{(y,s) : y^T x - f(x) \le s \ \forall x\right\} = \left\{(y,s) : y^T x - t \le s \ \forall (x,t) \in \mathrm{Epi}\{f\}\right\}. \tag{2.3.15}$$

Consider the conic quadratic program

$$\min_{x,t,u} \left\{-y^T x + t : Ax + tp + Bu + b \ge \mathbf{K}_0\right\}. \tag{P_y}$$

By (2.3.15), a pair (y,s) belongs to $\mathrm{Epi}\{f_*\}$ if and only if (P_y) is bounded below with optimal value $\ge -s$. Since (P_y) is strictly feasible, this is the case if and only if the dual problem

$$\max_{\xi} \left\{-b^T\xi : A^T\xi = -y, B^T\xi = 0, p^T\xi = 1, \xi \ge_{\mathbf{K}} 0\right\}$$

has a feasible solution with the value of the dual objective $\ge -s$. Thus,

$$\mathrm{Epi}\{f_*\} = \left\{(y,s) : \exists \xi : A^T\xi = -y, B^T\xi = 0, p^T\xi = 1, s \ge b^T\xi, \xi \ge_{\mathbf{K}} 0\right\}$$

as claimed in (2.3.14). It remains to note that the right hand side set in (2.3.14) is CQr (as a set obtained from the CQr set $\mathbf{K} \times \mathbf{R}_s$ by operations preserving CQ-representability – intersection with the set $\{\xi : B^T\xi = 0, p^T\xi = 1, b^T\xi \le s\}$ and subsequent affine mapping $\xi \mapsto -A^T\xi$).

Corollary 2.3.6 *The support function*

$$\mathrm{Supp}_X(x) = \sup_{y \in X} x^T y$$

of a nonempty CQr set X is CQr.

Indeed, $\mathrm{Supp}_X(\cdot)$ is the conjugate of the characteristic function of X, and it remains to refer to Corollary 2.3.4.

M. Taking convex hull of several sets. The convex hull of a set $Y \subset \mathbf{R}^n$ is the smallest convex set which contains Y:

$$\mathrm{Conv}(Y) = \left\{x = \sum_{i=1}^{k_x} \alpha_i x_i : x_i \in Y, \alpha_i \ge 0, \sum_i \alpha_i = 1\right\}$$

The closed convex hull $\overline{\mathrm{Conv}}(Y) = \mathrm{cl}\,\mathrm{Conv}(Y)$ of Y is the smallest *closed* convex set containing Y.

Following Yu. Nesterov, let us prove that taking convex hull "nearly" preserves CQ-representability:

Proposition 2.3.5 *Let $X_1, ..., X_k \subset \mathbf{R}^n$ be closed convex CQr sets:*

$$X_i = \{x : A_i x + B_i u_i + b_i \ge_{\mathbf{K}_i} 0, \ i = 1, ..., k\}, \tag{2.3.16}$$

where $\mathbf{K}_i$ is a direct product of ice-cream cones.

Then the CQr set

$$\begin{array}{rcl} Y & = & \{x : \exists \xi^1, ..., \xi^k, t_1, ..., t_k, \eta^1, ..., \eta^k : \\ & & \left[\begin{array}{c} A_1\xi^1 + B_1\eta^1 + t_1 b_1 \\ A_2\xi^2 + B_2\eta^2 + t_2 b_2 \\ ... \\ A_k\xi^k + B_k\eta^k + t_k b_k \end{array}\right] \ge_{\mathbf{K}} 0, \\ & & \begin{array}{rcl} t_1, ..., t_k & \ge & 0, \\ \xi^1 + ... + \xi^k & = & x \\ t_1 + ... + t_k & = & 1\}, \end{array} \\ \mathbf{K} & = & \mathbf{K}_1 \times ... \times \mathbf{K}_k \end{array} \tag{2.3.17}$$

is between the convex hull and the closed convex hull of the set $X_1 \cup ... \cup X_k$:

$$\mathrm{Conv}(\bigcup_{i=1}^{k} X_i) \subset Y \subset \overline{\mathrm{Conv}}(\bigcup_{i=1}^{k} X_i).$$

If, in addition to CQ-representability,

(i) *all X_i are bounded,*

or

(ii) *$X_i = Z_i + W$, where Z_i are closed and bounded sets and W is a convex closed set,*

then

$$\mathrm{Conv}(\bigcup_{i=1}^{k} X_i) = Y = \overline{\mathrm{Conv}}(\bigcup_{i=1}^{k} X_i)$$

is CQr.

First, the set Y clearly contains $\mathrm{Conv}(\bigcup_{i=1}^{k} X_i)$. Indeed, since the sets X_i are convex, the convex hull of their union is

$$\left\{ x = \sum_{i=1}^{k} t_i x^i : x^i \in X_i, t_i \geq 0, \sum_{i=1}^{k} t_i = 1 \right\}$$

(why?); for a point

$$x = \sum_{i=1}^{k} t_i x^i \qquad \left[x^i \in X_i, t_i \geq 0, \sum_{i=1}^{k} t_i = 1 \right],$$

there exist u^i, $i = 1, ..., k$, such that

$$A_i x^i + B_i u^i + b_i \geq_{\mathbf{K}_i} 0.$$

We get

$$\begin{array}{rcl} x & = & (t_1 x^1) + ... + (t_k x^k) \\ & = & \xi^1 + ... + \xi^k, \\ & & [\xi^i = t_i x^i]; \\ t_1, ..., t_k & \geq & 0; \\ t_1 + ... + t_k & = & 1; \\ A_i \xi^i + B_i \eta^i + t_i b_i & \geq_{\mathbf{K}_i} & 0,\ i = 1, ..., k, \\ & & [\eta^i = t_i u^i], \end{array} \tag{2.3.18}$$

so that $x \in Y$ (see the definition of Y).

To complete the proof that Y is between the convex hull and the closed convex hull of $\bigcup_{i=1}^{k} X_i$, it remains to verify that if $x \in Y$ then x is contained in the closed convex hull of $\bigcup_{i=1}^{k} X_i$. Let us somehow choose $\bar{x}^i \in X_i$; for properly chosen $\bar{u}^i$ we have

$$A_i \bar{x}^i + B_i \bar{u}^i + b_i \geq_{\mathbf{K}_i} 0,\ i = 1, ..., k. \tag{2.3.19}$$

Since $x \in Y$, there exist t_i, ξ^i, η^i satisfying the relations

$$\begin{array}{rcl} x & = & \xi^1 + ... + \xi^k, \\ t_1, ..., t_k & \geq & 0, \\ t_1 + ... + t_k & = & 1, \\ A_i \xi^i + B_i \eta^i + t_i b_i & \geq_{\mathbf{K}_i} & 0,\ i = 1, ..., k. \end{array} \tag{2.3.20}$$

In view of the latter relations and (2.3.19), we have for $0 < \epsilon < 1$:

$$A_i[(1-\epsilon)\xi^i + \epsilon k^{-1}\bar{x}^i] + B_i[(1-\epsilon)\eta^i + \epsilon k^{-1}\bar{u}^i] + [(1-\epsilon)t_i + \epsilon k^{-1}]b_i \geq_{\mathbf{K}_i} 0;$$

setting

$$\begin{array}{rcl} t_{i,\epsilon} & = & (1-\epsilon)t_i + \epsilon k^{-1}; \\ x^i_\epsilon & = & t^{-1}_{i,\epsilon}\left[(1-\epsilon)\xi^i + \epsilon k^{-1}\bar{x}^i\right]; \\ u^i_\epsilon & = & t^{-1}_{i,\epsilon}\left[(1-\epsilon)\eta^i + \epsilon k^{-1}\bar{u}^i\right], \end{array}$$

we get

$$\begin{array}{rcl} A_i x^i_\epsilon + B_i u^i_\epsilon + b_i & \geq_{\mathbf{K}_i} & 0 \Rightarrow \\ x^i_\epsilon & \in & X_i, \\ t_{1,\epsilon}, ..., t_{k,\epsilon} & \geq & 0, \\ t_{1,\epsilon} + ... + t_{k,\epsilon} & = & 1 \\ \Rightarrow & & \\ x_\epsilon & \equiv & \sum\limits_{i=1}^k t_{i,\epsilon} x^i_\epsilon \\ & \in & \mathrm{Conv}(\bigcup\limits_{i=1}^k X_i). \end{array}$$

On the other hand, we have by construction

$$x_\epsilon = \sum_{i=1}^k \left[(1-\epsilon)\xi^i + \epsilon k^{-1}\bar{x}^i\right] \to x = \sum_{i=1}^k \xi^i \text{ as } \epsilon \to +0,$$

so that x belongs to the closed convex hull of $\bigcup\limits_{i=1}^k X_i$, as claimed.

It remains to verify that in the cases of (i), (ii) the convex hull of $\bigcup\limits_{i=1}^k X_i$ is the same as the closed convex hull of this union. (i) is a particular case of (ii) corresponding to $W = \{0\}$, so that it suffices to prove (ii). Assume that

$$\begin{array}{c} x_t = \sum\limits_{i=1}^k \mu_{ti}[z_{ti} + p_{ti}] \to x \text{ as } i \to \infty \\ \left[z_{ti} \in Z_i, p_{ti} \in W, \mu_{ti} \geq 0, \sum\limits_i \mu_{ti} = 1\right] \end{array}$$

and let us prove that x belongs to the convex hull of the union of X_i. Indeed, since Z_i are closed and bounded, passing to a subsequence, we may assume that

$$z_{ti} \to z_i \in Z_i \text{ and } \mu_{ti} \to \mu_i \text{ as } t \to \infty.$$

It follows that the vectors

$$p_t = \sum_{i=1}^m \mu_{ti} p_{ti} = x_t - \sum_{i=1}^k \mu_{ti} z_{ti}$$

converge as $t \to \infty$ to some vector p, and since W is closed and convex, $p \in W$. We now have

$$x = \lim_{i\to\infty}\left[\sum_{i=1}^k \mu_{ti} z_{ti} + p_t\right] = \sum_{i=1}^k \mu_i z_i + p = \sum_{i=1}^k \mu_i [z_i + p],$$

so that x belongs to the convex hull of the union of X_i (as a convex combination of points $z_i + p \in X_i$).

P. The recessive cone of a CQr set. Let X be a closed convex set. The *recessive cone* $\mathrm{Rec}(X)$ of X is the set

$$\mathrm{Rec}(X) = \{h : x + th \in X \quad \forall (x \in X, t \geq 0)\}.$$

It can be easily verified that $\mathrm{Rec}(X)$ is a closed cone, and that

$$\mathrm{Rec}(X) = \{h : \bar{x} + th \in X \quad \forall t \geq 0\} \quad \forall \bar{x} \in X,$$

i.e., that $\mathrm{Rec}(X)$ is the set of all directions h such that the ray emanating from a point of X and directed by h is contained in X.

Proposition 2.3.6 *Let X be a nonempty CQr set with CQR*

$$X = \{x \in \mathbf{R}^n : \exists u : Ax + Bu + b \geq_{\mathbf{K}} 0\},$$

where $\mathbf{K}$ is a direct product of ice-cream cones, and let the CQR be such that $Bu \in \mathbf{K}$ only if $Bu = 0$. Then X is closed, and the recessive cone of X is CQr:

$$\mathrm{Rec}(X) = \{h : \exists v : Ah + Bv \geq_{\mathbf{K}} 0\}. \tag{2.3.21}$$

Proof. The fact that X is closed is given by Lemma 2.3.1. In order to prove (2.3.21), let us temporary denote by R the set in the left hand side of this relation; we should prove that $R = \mathrm{Rec}(X)$. The inclusion $R \subset \mathrm{Rec}(X)$ is evident. To prove the inverse inclusion, let $\bar{x} \in X$ and $h \in \mathrm{Rec}(X)$, so that for every $i = 1, 2, ...$ there exists u_i such that

$$A(\bar{x} + ih) + Bu_i + b \in \mathbf{K}. \tag{2.3.22}$$

By Lemma 2.3.1,

$$\|Bu_i\|_2 \leq C(1 + \|A(\bar{x} + ih) + b\|_2) \tag{2.3.23}$$

for certain $C < \infty$ and all i. Besides this, we can assume w.l.o.g. that

$$\|u_i\|_2 \leq C_1 \|Bu_i\|_2 \tag{2.3.24}$$

(cf. the proof of Lemma 2.3.1). By (2.3.23) – (2.3.24), the sequence $\{v_i = i^{-1}u_i\}$ is bounded; passing to a subsequence, we can assume that $v_i \to v$ as $i \to \infty$. By (2.3.22, we have for all i

$$i^{-1}A(\bar{x} + ih) + Bv_i + i^{-1}b \in \mathbf{K},$$

whence, passing to limit as $i \to \infty$, $Ah + Bv \in \mathbf{K}$. Thus, $h \in R$. ∎

O. Theorem on superposition. Let $f_\ell : \mathbf{R}^m \to \mathbf{R} \cup \{+\infty\}$, $\ell = 1, ..., m$ be CQr functions:

$$t \geq f_\ell(x) \Leftrightarrow \exists u^\ell : A_\ell(x, t, u^\ell) \succeq_{\mathbf{K}_\ell} 0,$$

where $\mathbf{K}_\ell$ is a direct product of ice-cream cones, and let

$$f : \mathbf{R}^m \to \mathbf{R} \cup \{+\infty\}$$

be CQr:

$$t \geq f(y) \Leftrightarrow \exists v : A(y, t, v) \succeq_{\mathbf{K}} 0,$$

where $\mathbf{K}$ is a direct product of ice-cream cones.

Assume that f is monotone with respect to the usual partial ordering of $\mathbf{R}^m$:

$$y' \geq y'' \Rightarrow f(y') \geq f(y''),$$

and that the superposition

$$g(x) = \begin{cases} f(f_1(x), ..., f_m(x)) & f_\ell(x) < \infty, \ell = 1, ..., m \\ +\infty & \text{otherwise} \end{cases}$$

is a proper function (i.e., it is finite at least at one point).

Theorem 2.3.1 *Under the above setting, the superposition g is CQr with CQR*

$$t \geq g(x) \Leftrightarrow \exists t_1, ..., t_m, u^1, ..., u^m, v : \left\{ \begin{array}{l} A_\ell(x, t_\ell, u^\ell) \succeq_{\mathbf{K}_\ell} 0, \ \ell = 1, ..., m \\ A(t_1, ..., t_m, t, v) \succeq_{\mathbf{K}} 0 \end{array} \right. \tag{2.3.25}$$

Proof of this simple statement is left to the reader.

Remark 2.3.1 If part of the "inner" functions, say, $f_1, ..., f_k$, are affine, it suffices to require the monotonicity of the "outer" function f with respect to the variables $y_{k+1}, ..., y_m$ only. A CQR for the superposition in this case becomes

$$t \geq g(x) \Leftrightarrow \exists t_{k+1}, ..., t_m, u^{k+1}, ..., u^m, v : \left\{ \begin{array}{l} A_\ell(x, t_\ell, u^\ell) \succeq_{\mathbf{K}_\ell} 0, \ \ell = k+1, ..., m \\ A(f_1(x), f_2(x), ..., f_k(x), t_{k+1}, t_{k+2}, ..., t_m, t, v) \succeq_{\mathbf{K}} 0 \end{array} \right. \tag{2.3.26}$$

2.3.1 More examples of CQ-representable functions/sets

We are sufficiently equipped to build the dictionary of CQ-representable functions/sets. Having built already the "elementary" part of the dictionary, we can add now a more "advanced" part.

7. Convex quadratic form $g(x) = x^T Q x + q^T x + r$ (Q is a positive semidefinite symmetric matrix) is CQr.
Indeed, Q is positive semidefinite symmetric and therefore can be decomposed as $Q = D^T D$, so that $g(x) = \|Dx\|_2^2 + q^T x + r$. We see that g is obtained from our "raw materials" – the squared Euclidean norm and an affine function – by affine substitution of argument and addition.

Here is an explicit CQR of g:

$$\{(x,t) : x^T D^T D x + q^T x + r \leq t\} = \{(x,t) : \left\| \begin{array}{c} Dx \\ \frac{t+q^T x+r}{2} \end{array} \right\|_2 \leq \frac{t - q^T x - r}{2}\} \tag{2.3.27}$$

8. The cone $K = \{(x, \sigma_1, \sigma_2) \in \mathbf{R}^n \times \mathbf{R} \times \mathbf{R} : \sigma_1, \sigma_2 \geq 0, \sigma_1 \sigma_2 \geq x^T x\}$ is CQr.
Indeed, the set is just the epigraph of the fractional-quadratic function $x^T x / s$, see Example 5; we simply write σ_1 instead of s and σ_2 instead of t.

Here is an explicit CQR for the set:

$$K = \{(x, \sigma_1, \sigma_2) : \left\| \begin{pmatrix} x \\ \frac{\sigma_1 - \sigma_2}{2} \end{pmatrix} \right\|_2 \leq \frac{\sigma_1 + \sigma_2}{2}\} \tag{2.3.28}$$

Surprisingly, our set is just the ice-cream cone, more precisely, its inverse image under the one-to-one linear mapping

$$\begin{pmatrix} x \\ \sigma_1 \\ \sigma_2 \end{pmatrix} \mapsto \begin{pmatrix} x \\ \frac{\sigma_1 - \sigma_2}{2} \\ \frac{\sigma_1 + \sigma_2}{2} \end{pmatrix}.$$

9. The "half-cone" $K_+^2 = \{(x_1, x_2, t) \in \mathbf{R}^3 : x_1, x_2 \geq 0, 0 \leq t \leq \sqrt{x_1 x_2}\}$ is CQr.
Indeed, our set is the intersection of the cone $\{t^2 \leq x_1 x_2, x_1, x_2 \geq 0\}$ from the previous example and the half-space $t \geq 0$.

Here is the explicit CQR of K_+:

$$K_+ = \{(x_1, x_2, t) : t \geq 0, \left\| \begin{pmatrix} t \\ \frac{x_1 - x_2}{2} \end{pmatrix} \right\|_2 \leq \frac{x_1 + x_2}{2}\}. \tag{2.3.29}$$

10. The hypograph of the geometric mean – the set $K^2 = \{(x_1, x_2, t) \in \mathbf{R}^3 : x_1, x_2 \geq 0, t \leq \sqrt{x_1 x_2}\}$ – is CQr.
Note the difference with the previous example – here t is not required to be nonnegative!

Here is the explicit CQR for K^2 (cf. Example 9):

$$K^2 = \left\{(x_1, x_2, t) : \exists \tau : t \leq \tau; \tau \geq 0, \left\| \begin{pmatrix} \tau \\ \frac{x_1 - x_2}{2} \end{pmatrix} \right\|_2 \leq \frac{x_1 + x_2}{2} \right\}.$$

11. The hypograph of the geometric mean of 2^l variables – the set $K^{2^l} = \{(x_1, ..., x_{2^l}, t) \in \mathbf{R}^{2^l+1} : x_i \geq 0, i = 1, ..., 2^l, t \leq (x_1 x_2 ... x_{2^l})^{1/2^l}\}$ – is CQr. To see it and to get its CQR, it suffices to iterate the construction of Example 10. Indeed, let us add to our initial variables a number of additional x-variables:

– let us call our 2^l original x-variables the variables of *level 0* and write $x_{0,i}$ instead of x_i. Let us add one new variable of *level 1* per every two variables of level 0. Thus, we add 2^{l-1} variables $x_{1,i}$ of level 1.

– similarly, let us add one new variable of level 2 per every two variables of level 1, thus adding 2^{l-2} variables $x_{2,i}$; then we add one new variable of level 3 per every two variables of level 2, and so on, until level l with a single variable $x_{l,1}$ is built.

Now let us look at the following system S of constraints:

$$\begin{array}{ll} \text{layer 1:} & x_{1,i} \leq \sqrt{x_{0,2i-1} x_{0,2i}}, x_{1,i}, x_{0,2i-1}, x_{0,2i} \geq 0, \ \ i = 1, ..., 2^{l-1} \\ \text{layer 2:} & x_{2,i} \leq \sqrt{x_{1,2i-1} x_{1,2i}}, x_{2,i}, x_{1,2i-1}, x_{1,2i} \geq 0, \ \ i = 1, ..., 2^{l-2} \\ & ..... \\ \text{layer } l\text{:} & x_{l,1} \leq \sqrt{x_{l-1,1} x_{l-1,2}}, x_{l,1}, x_{l-1,1}, x_{l-1,2} \geq 0 \\ \hline (*) & t \leq x_{l,1} \end{array}$$

The inequalities of the first layer say that the variables of the zero and the first level should be nonnegative and every one of the variables of the first level should be $\leq$ the geometric mean of the corresponding pair of our original x-variables. The inequalities of the second layer add the requirement that the variables of the second level should be nonnegative, and every one of them should be $\leq$ the geometric mean of the corresponding pair of the first level variables, etc. It is clear that if all these inequalities and (*) are satisfied, then t is $\leq$ the geometric mean of $x_1, ..., x_{2^l}$. Vice versa, given nonnegative $x_1, ..., x_{2^l}$ and a real t which is $\leq$ the geometric mean of $x_1, ..., x_{2^l}$, we always can extend these data to a solution of S. In other words, K^{2^l} is the projection of the solution set of S onto the plane of our original variables $x_1, ..., x_{2^l}, t$. It remains to note that the set of solutions of S is CQr (as the intersection of CQr sets $\{(v, p, q, r) \in \mathbf{R}^N \times \mathbf{R}^3_+ : r \leq \sqrt{qp}\}$, see Example 9), so that its projection is also CQr. To get a CQR of K^{2^l}, it suffices to replace the inequalities in S with their conic quadratic equivalents, explicitly given in Example 9.

12. The convex increasing power function $x_+^{p/q}$ of rational degree $p/q \geq 1$ is CQr.
Indeed, given positive integers p, q, $p > q$, let us choose the smallest integer l such that $p \leq 2^l$, and consider the CQr set

$$K^{2^l} = \{(y_1, ..., y_{2^l}, s) \in \mathbf{R}^{2^l+1}_+ : s \leq (y_1 y_2 ... y_{2^l})^{1/2^l}\}. \tag{2.3.30}$$

Setting $r = 2^l - p$, consider the following affine parameterization of the variables from $\mathbf{R}^{2^l+1}$ by two variables ξ, t:

– s and r first variables y_i are all equal to ξ (note that we still have $2^l - r = p \geq q$ "unused" variables y_i);

– q next variables y_i are all equal to t;

– the remaining y_i's, if any, are all equal to 1.

The inverse image of K^{2^l} under this mapping is CQr and it is the set

$$K = \{(\xi, t) \in \mathbf{R}^2_+ : \xi^{1-r/2^l} \leq t^{q/2^l}\} = \{(\xi, t) \in \mathbf{R}^2_+ : t \geq \xi^{p/q}\}.$$

It remains to note that the epigraph of $x_+^{p/q}$ can be obtained from the CQr set K by operations preserving the CQr property. Specifically, the set $L = \{(x,\xi,t) \in \mathbf{R}^3 : \xi \geq 0, \xi \geq x, t \geq \xi^{p/q}\}$ is the intersection of $\mathbf{K} \times \mathbf{R}$ and the half-space $\{(x,\xi,t) : \xi \geq x\}$ and thus is CQr along with K, and $\mathrm{Epi}\{x_+^{p/q}\}$ is the projection of the CQr set L on the plane of x,t-variables.

13. The decreasing power function $g(x) = \begin{cases} x^{-p/q}, & x > 0 \\ +\infty, & x \leq 0 \end{cases}$ (p, q are positive integers) is CQr.
Same as in Example 12, we choose the smallest integer l such that $2^l \geq p+q$, consider the CQr set (2.3.30) and parameterize affinely the variables y_i, s by two variables (x,t) as follows:
– s and the first $(2^l - p - q)$ y_i's are all equal to one;
– p of the remaining y_i's are all equal to x, and the q last of y_i's are all equal to t.

It is immediately seen that the inverse image of K^{2^l} under the indicated affine mapping is the epigraph of g.

14. The even power function $g(x) = x^{2p}$ on the axis (p positive integer) is CQr.
Indeed, we already know that the sets $P = \{(x,\xi,t) \in \mathbf{R}^3 : x^2 \leq \xi\}$ and $K' = \{(x,\xi,t) \in \mathbf{R}^3 : 0 \leq \xi, \xi^p \leq t\}$ are CQr (both sets are direct products of $\mathbf{R}$ and the sets with already known to us CQR's). It remains to note that the epigraph of g is the projection of $P \cap Q$ onto the (x,t)-plane.

Example 14 along with our combination rules allows to build a CQR for a polynomial $p(x)$ of the form

$$p(x) = \sum_{l=1}^{L} p_l x^{2l}, \quad x \in \mathbf{R},$$

with nonnegative coefficients.

15. The concave monomial $x_1^{\pi_1}...x_n^{\pi_n}$. Let $\pi_1 = \frac{p_1}{p}, ..., \pi_n = \frac{p_n}{p}$ be positive rational numbers with $\pi_1 + ... + \pi_n \leq 1$. The function

$$f(x) = -x_1^{\pi_1}...x_n^{\pi_n} : \mathbf{R}_+^n \to \mathbf{R}$$

is CQr.
The construction is similar to the one of Example 12. Let l be such that $2^l \geq p$. We recall that the set

$$Y = \{(y_1, ..., y_{2^l}, s) : y_1, ..., y_{2^l}, s) : y_1, ..., y_{2^l} \geq 0, 0 \leq s \leq (y_1...y_{2^l})^{1/2^l}\}$$

is CQr, and therefore so is its inverse image under the affine mapping

$$(x_1, ..., x_n, s) \mapsto (\underbrace{x_1, ..., x_1}_{p_1}, \underbrace{x_2, ..., x_2}_{p_2}, ..., \underbrace{x_n, ..., x_n}_{p_n}, \underbrace{s, ..., s}_{2^l - p}, \underbrace{1, ..., 1}_{p - p_1 - ... - p_n}, s),$$

i.e., the set

$$\begin{aligned} Z &= \{(x_1, ..., x_n, s) : x_1, ..., x_n \geq 0, 0 \leq s \leq (x_1^{p_1}...x_n^{p_n} s^{2^l - p})^{1/2^l}\} \\ &= \{(x_1, ..., x_n, s) : x_1, ..., x_n \geq 0, 0 \leq s \leq x_1^{p_1/p}...x_n^{p_n/p}\}. \end{aligned}$$

Since the set Z is CQr, so is the set

$$Z' = \{(x_1, ..., x_n, t, s) : x_1, ..., x_n \geq 0, s \geq 0, 0 \leq s - t \leq x_1^{\pi_1}...x_n^{\pi_n}\},$$

which is the intersection of the half-space $\{s \geq 0\}$ and the inverse image of Z under the affine mapping $(x_1, ..., x_n, t, s) \mapsto (x_1, ..., x_n, s - t)$. It remains to note that the epigraph of f is the projection of Z' onto the plane of the variables $x_1, ..., x_n, t$.

16. The convex monomial $x_1^{-\pi_1}...x_n^{-\pi_n}$. Let $\pi_1, ..., \pi_n$ be positive rational numbers. The function

$$f(x) = x_1^{-\pi_1}...x_n^{-\pi_n} : \{x \in \mathbf{R}^n : x > 0\} \to \mathbf{R}$$

is CQr.

The verification is completely similar to the one in Example 15.

17. The p-norm $\|x\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p} : \mathbf{R}^n \to \mathbf{R}$ ($p \geq 1$ is a rational number). We claim that the function $\|x\|_p$ is CQr.

It is immediately seen that

$$\|x\|_p \leq t \Leftrightarrow t \geq 0 \ \& \ \exists v_1, ..., v_n \geq 0 : |x_i| \leq t^{(p-1)/p} v_i^{1/p}, \ i = 1, ..., n, \sum_{i=1}^n v_i \leq t. \qquad (2.3.31)$$

Indeed, if the indicated v_i exist, then $\sum_{i=1}^n |x_i|^p \leq t^{p-1} \sum_{i=1}^n v_i \leq t^p$, i.e., $\|x\|_p \leq t$. Vice versa, assume that $\|x\|_p \leq t$. If $t = 0$, then $x = 0$, and the right hand side relations in (2.3.31) are satisfied for $v_i = 0$, $i = 1, ..., n$. If $t > 0$, we can satisfy these relations by setting $v_i = |x_i|^p t^{1-p}$.

(2.3.31) says that the epigraph of $\|x\|_p$ is the projection onto the (x, t)-plane of the set of solutions to the system of inequalities

$$\begin{gathered} t \geq 0 \\ v_i \geq 0, \ i = 1, ..., n \\ x_i \leq t^{(p-1)/p} v_i^{1/p}, \ i = 1, ..., n \\ -x_i \leq t^{(p-1)/p} v_i^{1/p}, \ i = 1, ..., n \\ v_1 + ... + v_n \leq t \end{gathered}$$

Each of these inequalities defines a CQr set (in particular, for the nonlinear inequalities this is due to Example 15). Thus, the solution set of the system is CQr (as an intersection of finitely many CQr sets), whence its projection on the (x, t)-plane – i.e., the epigraph of $\|x\|_p$ – is CQr.

17b. The function $\|x_+\|_p = \left(\sum_{i=1}^n \max^p[x_i, 0]\right)^{1/p} : \mathbf{R}^n \to \mathbf{R}$ ($p \geq 1$ a rational number) is CQr.

Indeed,

$$t \geq \|x_+\|_p \Leftrightarrow \exists y_1, ..., y_n : 0 \leq y_i, x_i \leq y_i, \ i = 1, ..., n, \|y\|_p \leq t.$$

Thus, the epigraph of $\|x_+\|_p$ is a projection of the CQr set (see Example 17) given by the system of inequalities in the right hand side.

From the above examples it is seen that the "expressive abilities" of c.q.i.'s are indeed strong: they allow to handle a wide variety of very different functions and sets.

2.4 More applications: Robust Linear Programming

Equipped with abilities to treat a wide variety of CQr functions and sets, we can consider now an important generic application of Conic Quadratic Programming, specifically, in the *Robust Linear Programming*.

2.4.1 Robust Linear Programming: the paradigm

Consider an LP program

$$\min_x \left\{ c^T x : Ax - b \geq 0 \right\}. \tag{LP}$$

In real world applications, the data c, A, b of (LP) is not always known exactly; what is typically known is a domain $\mathcal{U}$ in the space of data – an "uncertainty set" – which for sure contains the "actual" (unknown) data. There are cases in reality where, in spite of this data uncertainty, our decision x must satisfy the "actual" constraints, whether we know them or not. Assume, e.g., that (LP) is a model of a technological process in Chemical Industry, so that entries of x represent the amounts of different kinds of materials participating in the process. Typically the process includes a number of decomposition-recombination stages. A model of this problem must take care of natural balance restrictions: the amount of every material to be used at a particular stage cannot exceed the amount of the same material yielded by the preceding stages. In a meaningful production plan, these balance inequalities must be satisfied even though they involve coefficients affected by unavoidable uncertainty of the exact contents of the raw materials, of time-varying parameters of the technological devices, etc.

If indeed all we know about the data is that they belong to a given set $\mathcal{U}$, but we still have to satisfy the actual constraints, the only way to meet the requirements is to restrict ourselves to *robust feasible* candidate solutions – those satisfying *all possible* realizations of the uncertain constraints, i.e., vectors x such that

$$Ax - b \geq 0 \quad \forall [A; b] \text{ such that } \exists c : (c, A, b) \in \mathcal{U}. \tag{2.4.1}$$

In order to choose among these robust feasible solutions the best possible, we should decide how to "aggregate" the various realizations of the objective into a single "quality characteristic". To be methodologically consistent, we use the same worst-case-oriented approach and take as an objective function $f(x)$ the maximum, over all possible realizations of the objective $c^T x$:

$$f(x) = \sup\{c^T x \mid c : \exists [A; b] : (c, A, b) \in \mathcal{U}\}.$$

With this methodology, we can associate with our *uncertain LP program* (i.e., with the family

$$\mathcal{LP}(\mathcal{U}) = \left\{ \min_{x: Ax \geq b} c^T x | (c, A, b) \in \mathcal{U} \right\}$$

of all usual ("certain") LP programs with the data belonging to $\mathcal{U}$) its robust counterpart. In the latter problem we are seeking for a robust feasible solution with the smallest possible value of the "guaranteed objective" $f(x)$. In other words, the robust counterpart of $\mathcal{LP}(\mathcal{U})$ is the optimization problem

$$\min_{t,x} \left\{ t : c^T x \leq t, Ax - b \geq 0 \quad \forall (c, A, b) \in \mathcal{U} \right\}. \tag{R}$$

Note that (R) is a usual – "certain" – optimization problem, but typically it is not an LP program: the structure of (R) depends on the geometry of the uncertainty set $\mathcal{U}$ and can be very complicated.

As we shall see in a while, in many cases it is reasonable to specify the uncertainty set $\mathcal{U}$ as an *ellipsoid* – the image of the unit Euclidean ball under an affine mapping – or, more generally, as a CQr set. As we shall see in a while, in this case the robust counterpart of an uncertain LP problem is (equivalent to) an explicit conic quadratic program. Thus, Robust

Linear Programming with CQr uncertainty sets can be viewed as a "generic source" of conic quadratic problems.

Let us look at the robust counterpart of an uncertain LP program

$$\left\{\min_x \left\{c^T x : a_i^T x - b_i \geq 0,\ i = 1, ..., m\right\} | (c, A, b) \in \mathbf{U}\right\}$$

in the case of a "simple" ellipsoidal uncertainty – one where the data (a_i, b_i) of i-th inequality constraint

$$a_i^T x - b_i \geq 0,$$

and the objective c are allowed to run independently of each other through respective ellipsoids E_i, E. Thus, we assume that the uncertainty set is

$$\mathcal{U} = \left\{(a_1, b_1; ...; a_m, b_m; c) : \exists(\{u_i, u_i^T u_i \leq 1\}_{i=0}^m) :\ \ c = c_* + P_0 u_0,\ \begin{pmatrix} a_i \\ b_i \end{pmatrix} = \begin{pmatrix} a_i^* \\ b_i^* \end{pmatrix} + P_i u^i, i = 1, ..., m\right\},$$

where c_*, a_i^*, b_i^* are the "nominal data" and $P_i u_i$, $i = 0, 1, ..., m$, represent the data perturbations; the restrictions $u_i^T u_i \leq 1$ enforce these perturbations to vary in ellipsoids.

In order to realize that the robust counterpart of our uncertain LP problem is a conic quadratic program, note that x is robust feasible if and only if for every $i = 1, ..., m$ we have

$$\begin{aligned} 0 &\leq \min_{u_i : u_i^T u_i \leq 1} \left[a_i^T[u] x - b_i[u] : \begin{pmatrix} a_i[u] \\ b_i[u] \end{pmatrix} = \begin{pmatrix} a_i^* \\ b_i^* \end{pmatrix} + P_i u_i\right] \\ &= (a_i^* x)^T x - b_i^* + \min_{u_i : u_i^T u_i \leq 1} u_i^T P_i^T \begin{pmatrix} x \\ -1 \end{pmatrix} \\ &= (a_i^*)^T x - b_i^* - \left\| P_i^T \begin{pmatrix} x \\ -1 \end{pmatrix} \right\|_2 \end{aligned}$$

Thus, x is robust feasible if and only if it satisfies the system of c.q.i.'s

$$\left\| P_i^T \begin{pmatrix} x \\ -1 \end{pmatrix} \right\|_2 \leq [a_i^*]^T x - b_i^*,\ \ i = 1, ..., m.$$

Similarly, a pair (x, t) satisfies all realizations of the inequality $c^T x \leq t$ "allowed" by our ellipsoidal uncertainty set $\mathcal{U}$ if and only if

$$c_*^T x + \|P_0^T x\|_2 \leq t.$$

Thus, the robust counterpart (R) becomes the conic quadratic program

$$\min_{x,t} \left\{t : \|P_0^T x\|_2 \leq -c_*^T x + t;\ \left\| P_i^T \begin{pmatrix} x \\ -1 \end{pmatrix} \right\|_2 \leq [a_i^*]^T x - b_i^*,\ \ i = 1, ..., m\right\} \qquad \text{(RLP)}$$

2.4.2 Robust Linear Programming: examples

Example 1: Robust synthesis of antenna array. Consider a monochromatic transmitting antenna placed at the origin. Physics says that

1. The directional distribution of energy sent by the antenna can be described in terms of *antenna's diagram* which is a complex-valued function $D(\delta)$ of a 3D direction δ. The directional distribution of energy sent by the antenna is proportional to $|D(\delta)|^2$.

2. When the antenna is comprised of several antenna elements with diagrams $D_1(\delta)$,..., $D_k(\delta)$, the diagram of the antenna is just the sum of the diagrams of the elements.

In a typical Antenna Design problem, we are given several antenna elements with diagrams $D_1(\delta)$,...,$D_k(\delta)$ and are allowed to multiply these diagrams by complex *weights* x_i (which in reality corresponds to modifying the output powers and shifting the phases of the elements). As a result, we can obtain, as a diagram of the array, any function of the form

$$D(\delta) = \sum_{i=1}^{k} x_i D_i(\delta),$$

and our goal is to find the weights x_i which result in a diagram as close as possible, in a prescribed sense, to a given "target diagram" $D_*(\delta)$.

Consider an example of a planar antenna comprised of a central circle and 9 concentric rings of the same area as the circle (Fig. 2.1.(a)) in the XY-plane ("Earth's surface"). Let the wavelength be $\lambda = 50$cm, and the outer radius of the outer ring be 1 m (twice the wavelength).

One can easily see that the diagram of a ring $\{a \le r \le b\}$ in the plane XY (r is the distance from a point to the origin) as a function of a 3-dimensional direction δ depends on the altitude (the angle θ between the direction and the plane) only. The resulting function of θ turns out to be *real-valued*, and its analytic expression is

$$D_{a,b}(\theta) = \frac{1}{2} \int_a^b \left[\int_0^{2\pi} r \cos\left(2\pi r \lambda^{-1} \cos(\theta)\cos(\phi)\right) d\phi \right] dr.$$

Fig. 2.1.(b) represents the diagrams of our 10 rings for $\lambda = 50$cm.

Assume that our goal is to design an array with a real-valued diagram which should be axial symmetric with respect to the Z-axis and should be "concentrated" in the cone $\pi/2 \ge \theta \ge \pi/2 - \pi/12$. In other words, our target diagram is a real-valued function $D_*(\theta)$ of the altitude θ with $D_*(\theta) = 0$ for $0 \le \theta \le \pi/2 - \pi/12$ and $D_*(\theta)$ somehow approaching 1 as θ approaches $\pi/2$. The target diagram $D_*(\theta)$ used in this example is given in Fig. 2.1.(c) (the dashed curve).

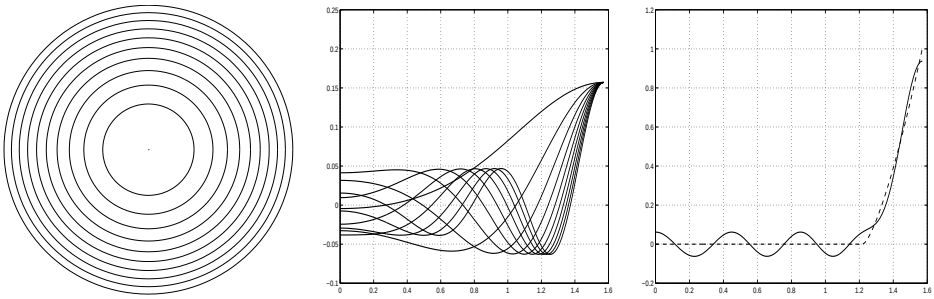
Finally, let us measure the discrepancy between a synthesized diagram and the target one by the Tschebyshev distance, taken along the equidistant 120-point grid of altitudes, i.e., by the quantity

$$\tau = \max_{\ell=1,...,120} \left| D_*(\theta_\ell) - \sum_{j=1}^{10} x_j \underbrace{D_{r_{j-1},r_j}(\theta_\ell)}_{D_j(\theta_\ell)} \right|, \quad \theta_\ell = \frac{\ell\pi}{240}.$$

Our design problem is simplified considerably by the fact that the diagrams of our "building blocks" and the target diagram are real-valued; thus, we need no complex numbers, and the problem we should finally solve is

$$\min_{\tau\in\mathbf{R}, x\in\mathbf{R}^{10}} \left\{ \tau : -\tau \le D_*(\theta_\ell) - \sum_{j=1}^{10} x_j D_j(\theta_\ell) \le \tau,\ \ell = 1,...,120 \right\}. \qquad \text{(Nom)}$$

This is a simple LP program; its optimal solution x^* results in the diagram depicted at Fig. 2.1.(c). The uniform distance between the actual and the target diagrams is ≈ 0.0621 (recall that the target diagram varies from 0 to 1).



(a) (b) (c)

Figure 2.1. Synthesis of antennae array

(a): 10 array elements of equal areas in the XY-plane
the outer radius of the largest ring is 1m, the wavelength is 50cm

(b): "building blocks" – the diagrams of the rings as functions of the altitude angle θ

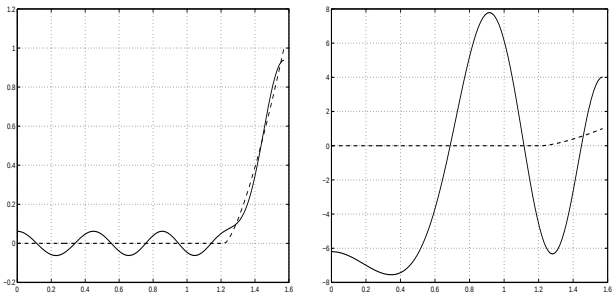
(c): the target diagram (dashed) and the synthesied diagram (solid)

Now recall that our design variables are characteristics of certain physical devices. In reality, of course, we cannot tune the devices to have precisely the optimal characteristics x_j^*; the best we may hope for is that the actual characteristics $x_j^{\rm fct}$ will coincide with the desired values x_j^* within a small margin, say, 0.1% (this is a fairly high accuracy for a physical device):

$$x_j^{\rm fct} = p_j x_j^*,\ 0.999 \le p_j \le 1.001.$$

It is natural to assume that the factors p_j are random with the mean value equal to 1; it is perhaps not a great sin to assume that these factors are independent of each other.

Since the actual weights differ from their desired values x_j^*, the actual (random) diagram of our array of antennae will differ from the "nominal" one we see on Fig.2.1.(c). How large could be the difference? Look at the picture:



"Dream and reality": the nominal (left, solid) and an actual (right, solid) diagrams [dashed: the target diagram]

The diagram shown to the right is not even the worst case: we just have taken as p_j a sample of 10 independent numbers distributed uniformly in $[0.999, 1.001]$ and have plotted the diagram corresponding to $x_j = p_j x_j^*$. Pay attention not only to the shape (completely opposite to what we need), but also to the scale: the target diagram varies from 0 to 1, and the nominal diagram (the one corresponding to the exact optimal x_j) differs from the target by no more than by 0.0621 (this is the optimal value in the "nominal" problem (Nom)). The actual diagram varies from ≈ -8 to ≈ 8, and its uniform distance from the target is 7.79 (125 times the nominal optimal value!). We see that our nominal optimal design is completely meaningless: it looks as if we were trying to get the worse possible result, not the best possible one...

How could we get something better? Let us try to apply the Robust Counterpart approach. To this end we take into account from the very beginning that if we want the amplification coefficients to be certain x_j, then the actual amplification coefficients will be $x_j^{\rm fct} = p_j x_j$, $0.999 \leq p_j \leq 1.001$, and the actual discrepancies will be

$$\delta_\ell(x) = D_*(\theta_\ell) - \sum_{j=1}^{10} p_j x_j D_j(\theta_\ell).$$

Thus, we in fact are solving an uncertain LP problem where the uncertainty affects the coefficients of the constraint matrix (those corresponding to the variables x_j): these coefficients may vary within 0.1% margin of their nominal values.

In order to apply to our uncertain LP program the Robust Counterpart approach, we should specify the uncertainty set $\mathcal{U}$. The most straightforward way is to say that our uncertainty is "an interval" one – every uncertain coefficient in a given inequality constraint may (independently of all other coefficients) vary through its own uncertainty segment "nominal value $\pm 0.1\%$". This approach, however, is too conservative: we have completely ignored the fact that our p_j's are of stochastic nature and are independent of each other, so that it is highly improbable that all of them will *simultaneously* fluctuate in "dangerous" directions. In order to utilize the statistical independence of perturbations, let us look what happens with a particular inequality

$$-\tau \leq \delta_\ell(x) \equiv D_*(\theta_\ell) - \sum_{j=1}^{10} p_j x_j D_j(\theta_\ell) \leq \tau \tag{2.4.2}$$

when p_j's are random. For a fixed x, the quantity $\delta_\ell(x)$ is a random variable with the mean

$$\delta_\ell^*(x) = D_*(\theta_\ell) - \sum_{j=1}^{10} x_j D_j(\theta_\ell)$$

and the standard deviation

$$\sigma_\ell(x) = \sqrt{E\{(\delta_\ell(x) - \delta_\ell^*(x))^2\}} = \sqrt{\sum_{j=1}^{10} x_j^2 D_j^2(\theta_\ell) E\{(p_j - 1)^2\}} \leq \kappa \nu_\ell(x),$$
$$\nu_\ell(x) = \sqrt{\sum_{j=1}^{10} x_j^2 D_j^2(\theta_\ell)}, \kappa = 0.001.$$

Thus, "a typical value" of $\delta_\ell(x)$ differs from $\delta_\ell^*(x)$ by a quantity of order of $\sigma_\ell(x)$. Now let us act as an engineer which believes that a random variable differs from its mean by at most three times its standard deviation; since we are not obliged to be that concrete, let us choose a "safety parameter" ω and *ignore all events which result in* $|\delta_\ell(x) - \delta_\ell^*(x)| > \omega\nu_\ell(x)$ [5)]. As for the remaining events – those with $|\delta_\ell(x) - \delta_\ell^*(x)| \leq \omega\nu_\ell(x)$ – we take upon ourselves full responsibility. With this approach, a "reliable deterministic version" of the uncertain constraint (2.4.2) becomes the pair of inequalities

$$-\tau \leq \delta_\ell^*(x) - \omega\nu_\ell(x),$$
$$\delta_\ell^*(x) + \omega\nu_\ell(x) \leq \tau;$$

[5*)] It would be better to use here σ_ℓ instead of ν_ℓ; however, we did not assume that we know the distribution of p_j, this is why we replace unknown σ_ℓ with its known upper bound ν_ℓ

Replacing all uncertain inequalities in (Nom) with their "reliable deterministic versions" and recalling the definition of $\delta_\ell^*(x)$ and $\nu_\ell(x)$, we end up with the optimization problem

$$\begin{array}{lll} \text{minimize} & & \tau \\ \text{s.t.} & & \\ & \|Q_\ell x\|_2 & \le [D_*(\theta_\ell) - \sum_{j=1}^{10} x_j D_j(\theta_\ell)] + \tau, \ \ell = 1, ..., 120 \\ & \|Q_\ell x\|_2 & \le -[D_*(\theta_\ell) - \sum_{j=1}^{10} x_j D_j(\theta_\ell)] + \tau, \ \ell = 1, ..., 120 \\ & & [Q_\ell = \omega\kappa \mathrm{Diag}(D_1(\theta_\ell), D_2(\theta_\ell), ..., D_{10}(\theta_\ell))] \end{array} \tag{Rob}$$

It is immediately seen that (Rob) *is nothing but the robust counterpart of* (Nom) *corresponding to a simple ellipsoidal uncertainty*, namely, the one as follows:

The only data of a constraint

$$\sum_{j=1}^{10} A_{\ell j} x_j \le p_\ell \tau + q_\ell$$

(all constraints in (Nom) are of this form) affected by the uncertainty are the coefficients $A_{\ell j}$ of the left hand side, and the difference $dA[\ell]$ between the vector of these coefficients and the nominal value $(D_1(\theta_\ell), ..., D_{10}(\theta_\ell))^T$ of the vector of coefficients belongs to the ellipsoid

$$\{dA[\ell] = \omega\kappa Q_\ell u : u \in \mathbf{R}^{10}, u^T u \le 1\}.$$

Thus, the above "engineering reasoning" leading to (Rob) was nothing but a reasonable way to specify the uncertainty ellipsoids!

The bottom line of our "engineering reasoning" deserves to be formulated as a separate statement and to be equipped with a "reliability bound":

Proposition 2.4.1 *Consider a randomly perturbed linear constraint*

$$a_0(x) + \epsilon_1 a_1(x) + ... + \epsilon_n a_n(x) \ge 0, \tag{2.4.3}$$

where $a_j(x)$ are deterministic affine functions of the design vector x, and ϵ_j are independent random perturbations with zero means and such that $|\epsilon_j| \le \sigma_j$. Assume that x satisfies the "reliable" version of (2.4.3)*, specifically, the deterministic constraint*

$$a_0(x) - \kappa\sqrt{\sigma_1^2 a_1^2(x) + ... + \sigma_n^2 a_n^2(x)} \ge 0 \tag{2.4.4}$$

($\kappa > 0$). Then x satisfies a realization of (2.4.3) *with probability at least $1 - \exp\{-\kappa^2/4\}$. If all ϵ_i are symmetrically distributed, this bound can be improved to $1 - \exp\{-\kappa^2/2\}$.*

Proof. All we need is to verify the following *Bernstein's bound on probabilities of large deviations*:

If a_i are deterministic reals and ϵ_i are independent random variables with zero means and such that $|\epsilon_i| \le \sigma_i$ for given deterministic σ_i, then for every $\kappa \ge 0$ one has

$$p(\kappa) \equiv \mathrm{Prob}\left\{\sum_i \epsilon_i a_i > \kappa\sqrt{\sum_i a_i^2\sigma_i^2}\right\} \le \exp\{-\kappa^2/4\}.$$

When ϵ_i are symmetrically distributed, the right hand side can be replaced with $\exp\{-\kappa^2/2\}$.

Verification is easy: for $\gamma > 0$ we have

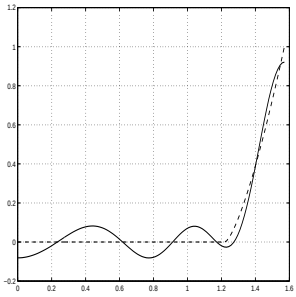
$$\begin{array}{rcl}
\exp\{\gamma\kappa\sigma\}p(\kappa) & \le & \mathbf{E}\left\{\exp\{\gamma\sum_i c_i\epsilon_i\}\right\} = \prod_i \mathbf{E}\left\{\exp\{\gamma c_i\epsilon_i\}\right\} \\
& = & \prod_i \mathbf{E}\left\{1 + \gamma c_i\epsilon_i + \sum_{k\ge 2}\frac{\gamma^k c_i^k \xi^k}{k!}\right\} = \prod_i \mathbf{E}\left\{1 + \gamma c_i\epsilon_i + \sum_{k\ge 2}\frac{\gamma^k c_i^k \xi^k}{k!}\right\} \\
& \le & \prod_i\left[1 + \sum_{k\ge 2}\frac{\gamma^k|c_i\sigma_i|^k}{k!}\right] \\
& \le & \prod_i \exp\{\gamma^2 c_i^2\sigma_i^2\} \\
& & [\text{since } \exp\{t\} - t \le \exp\{t^2\} \text{ for } t \ge 0] \\
& = & \exp\{\gamma^2\sigma^2\}.
\end{array} \tag{2.4.5}$$

Thus,

$$p(\kappa) \le \min_{\gamma>0}\exp\{\gamma^2\sigma^2 - \gamma\kappa\sigma\} = \exp\left\{-\frac{\kappa^2}{4}\right\}.$$

In the case of symmetrically distributed ϵ_i, the factors in the right hand side of the second inequality in (2.4.5) can be replaced with $\sum_{\ell=0}^{\infty}\frac{(\gamma\sigma_i a_i)^{2\ell}}{(2\ell)!}$, and these factors are $\le \exp\{-\gamma^2\sigma_i^2 a_i^2/2\}$. As a result, the concluding bound in (2.4.5) becomes $\exp\{\gamma^2\sigma^2/2\}$, which results in $p(\kappa) \le \exp\{-\kappa^2/2\}$. ∎

Now let us look what are the diagrams yielded by the Robust Counterpart approach – i.e., those given by the robust optimal solution. These diagrams are also random (neither the nominal nor the robust solution cannot be implemented exactly!). However, it turns out that they are incomparably closer to the target (and to each other) than the diagrams associated with the optimal solution to the "nominal" problem. Look at a typical "robust" diagram:



A "Robust" diagram. Uniform distance from the target is 0.0822.
[the safety parameter for the uncertainty ellipsoids is $\omega = 1$]

With the safety parameter $\omega = 1$, the robust optimal value is 0.0817; although it is by 30% larger than the nominal optimal value 0.0635, the robust optimal value has a definite advantage that it indeed says something reliable about the quality of actual diagrams we can obtain when implementing the robust optimal solution: in a sample of 40 realizations of the diagrams corresponding to the robust optimal solution, the uniform distances from the target were varying from 0.0814 to 0.0830.

We have built the robust optimal solution under the assumption that the "implementation errors" do not exceed 0.1%. What happens if in reality the errors are larger – say, 1%? It turns out that nothing dramatic happens: now in a sample of 40 diagrams given by the "old" robust optimal solution (affected by 10 times larger "implementation errors") the uniform distances from the target were varying from 0.0834 to 0.116. Imagine what will happen with the nominal solution under the same circumstances...

The last issue to be addressed here is: why is the nominal solution so unstable? And why with the robust counterpart approach we were able to get a solution which is incomparably better, as far as "actual implementation" is concerned? The answer becomes clear when looking at the nominal and the robust optimal weights:

j	1	2	3	4	5	6	7	8	9	10
$x_j^{\rm nom}$	1624.4	-14701	55383	-107247	95468	19221	-138622	144870	-69303	13311
$x_j^{\rm rob}$	-0.3010	4.9638	-3.4252	-5.1488	6.8653	5.5140	5.3119	-7.4584	-8.9140	13.237

It turns out that the nominal problem is "ill-posed" – although its optimal solution is far away from the origin, there is a "massive" set of "nearly optimal" solutions, and among the latter ones we can choose solutions of quite moderate magnitude. Indeed, here are the optimal values obtained when we add to the constraints of (Nom) the box constraints $|x_j| \le L$, $j = 1, ..., 10$:

L	1	10	10^2	10^3	10^4	10^5	10^6	10^7
Opt_Val	0.09449	0.07994	0.07358	0.06955	0.06588	0.06272	0.06215	0.06215

Since the "implementation inaccuracies" for a solution are the larger the larger the solution is, there is no surprise that our "huge" nominal solution results in a very unstable actual design. In contrast to this, the Robust Counterpart penalizes the (properly measured) magnitude of x (look at the terms $\|Q_\ell x\|_2$ in the constraints of (Rob)) and therefore yields a much more stable design. Note that this situation is typical for many applications: the nominal solution is on the boundary of the nominal feasible domain, and there are "nearly optimal" solutions to the nominal problem which are in the "deep interior" of this domain. When solving the nominal problem, we do not take any care of a reasonable tradeoff between the "depth of feasibility" and the optimality: *any* improvement in the objective is sufficient to make the solution just marginally feasible for the nominal problem. And a solution which is only marginally feasible in the nominal problem can easily become "very infeasible" when the data are perturbed. This would not be the case for a "deeply interior" solution. With the Robust Counterpart approach, we do use certain tradeoff between the "depth of feasibility" and the optimality – we are trying to find something like the "deepest feasible nearly optimal solution"; as a result, we normally gain a lot in stability; and if, as in our example, there are "deeply interior nearly optimal" solutions, we do not loose that much in optimality.

Example 2: NETLIB Case Study. `NETLIB` is a collection of about 100 not very large LPs, mostly of real-world origin, used as the standard benchmark for LP solvers. In the study to

be described, we used this collection in order to understand how "stable" are the feasibility properties of the standard – "nominal" – optimal solutions with respect to small uncertainty in the data. To motivate the methodology of this "Case Study", here is the constraint # 372 of the problem `PILOT4` from `NETLIB`:

$$\begin{array}{rcl} a^T x & \equiv & -15.79081x_{826} - 8.598819x_{827} - 1.88789x_{828} - 1.362417x_{829} - 1.526049x_{830} \\ & & -0.031883x_{849} - 28.725555x_{850} - 10.792065x_{851} - 0.19004x_{852} - 2.757176x_{853} \\ & & -12.290832x_{854} + 717.562256x_{855} - 0.057865x_{856} - 3.785417x_{857} - 78.30661x_{858} \\ & & -122.163055x_{859} - 6.46609x_{860} - 0.48371x_{861} - 0.615264x_{862} - 1.353783x_{863} \\ & & -84.644257x_{864} - 122.459045x_{865} - 43.15593x_{866} - 1.712592x_{870} - 0.401597x_{871} \\ & & +x_{880} - 0.946049x_{898} - 0.946049x_{916} \\ & \geq & b \equiv 23.387405 \end{array} \qquad \text{(C)}$$

The related *nonzero* coordinates in the optimal solution x^* of the problem, as reported by `CPLEX` (one of the best commercial LP solvers), are as follows:

$$\begin{array}{lll} x^*_{826} = 255.6112787181108 & x^*_{827} = 6240.488912232100 & x^*_{828} = 3624.613324098961 \\ x^*_{829} = 18.20205065283259 & x^*_{849} = 174397.0389573037 & x^*_{870} = 14250.00176680900 \\ x^*_{871} = 25910.00731692178 & x^*_{880} = 104958.3199274139 & \end{array}$$

The indicated optimal solution makes (C) an equality within machine precision.

Observe that most of the coefficients in (C) are "ugly reals" like -15.79081 or -84.644257. We have all reasons to believe that coefficients of this type characterize certain technological devices/processes, and as such *they could hardly be known to high accuracy*. It is quite natural to assume that the "ugly coefficients" are in fact uncertain – they coincide with the "true" values of the corresponding data within accuracy of 3-4 digits, not more. The only exception is the coefficient `1` of $\mathtt{x_{880}}$ – it perhaps reflects the structure of the problem and is therefore exact – "certain".

Assuming that the uncertain entries of a are, say, 0.1%-accurate approximations of unknown entries of the "true" vector of coefficients $\tilde{a}$, we looked what would be the effect of this uncertainty on the validity of the "true" constraint $\tilde{a}^T x \geq b$ at x^*. Here is what we have found:

• The minimum (over all vectors of coefficients $\tilde{a}$ compatible with our "0.1%-uncertainty hypothesis") value of $\tilde{a}^T x^* - b$, is < -104.9; in other words, the violation of the constraint can be as large as 450% of the right hand side!

• Treating the above worst-case violation as "too pessimistic" (why should the true values of all uncertain coefficients differ from the values indicated in (C) in the "most dangerous" way?), consider a more realistic measure of violation. Specifically, assume that the true values of the uncertain coefficients in (C) are obtained from the "nominal values" (those shown in (C)) by random perturbations $a_j \mapsto \tilde{a}_j = (1 + \xi_j)a_j$ with independent and, say, uniformly distributed on $[-0.001, 0.001]$ "relative perturbations" ξ_j. What will be a "typical" relative violation

$$V = \frac{\max[b - \tilde{a}^T x^*, 0]}{b} \times 100\%$$

of the "true" (now random) constraint $\tilde{a}^T x \geq b$ at x^*? The answer is nearly as bad as for the worst scenario:

Prob$\{V > 0\}$	Prob$\{V > 150\%\}$	Mean(V)
0.50	0.18	125%

Table 2.1. Relative violation of constraint # 372 in `PILOT4`
(1,000-element sample of 0.1% perturbations of the uncertain data)

We see that *quite small (just 0.1%) perturbations of "obviously uncertain" data coefficients can make the "nominal" optimal solution x^* heavily infeasible and thus – practically meaningless.*

Inspired by this preliminary experiment, we have carried out the "diagnosis" and the "treatment" phases as follows.

"Diagnosis". Given a "perturbation level" ϵ (for which we have used the values 1%, 0.1%, 0.01%), for every one of the NETLIB problems, we have measured its "stability index" at this perturbation level, specifically, as follows.

1. We computed the optimal solution x^* of the program by CPLEX.

2. For every one of the inequality constraints

$$a^T x \leq b$$

of the program,

- We looked at the right hand side coefficients a_j and split them into "certain" – those which can be represented, within machine accuracy, as rational fractions p/q with $|q| \leq 100$, and "uncertain" – all the rest. Let J be the set of all uncertain coefficients of the constraint under consideration.
- We defined the *reliability index* of the constraint as the quantity

$$\frac{a^T x^* + \epsilon\sqrt{\sum\limits_{j\in J} a_j^2 (x_j^*)^2} - b}{\max[1, |b|]} \times 100\% \qquad \text{(I)}$$

Note that the quantity $\epsilon\sqrt{\sum\limits_{j\in J} a_j^2 (x_j^*)^2}$, as we remember from the Antenna story, is of order of typical difference between $a^T x^*$ and $\tilde{a}^T x^*$, where $\tilde{a}$ is obtained from a by random perturbation $a_j \mapsto \tilde{a}_j = p_j a_j$ of uncertain coefficients, with independent random p_j uniformly distributed in the segment $[-\epsilon, \epsilon]$. In other words, *the reliability index is of order of typical violation* (measured in percents of the right hand side) *of the constraint, as evaluated at x^*, under independent random perturbations of uncertain coefficients, ϵ being the relative magnitude of the perturbations.*

3. We treat the nominal solution as *unreliable*, and the problem - as *bad*, the level of perturbations being ϵ, if the worst, over the inequality constraints, reliability index of the constraint is worse than 5%.

The results of the Diagnosis phase of our Case Study were as follows. From the total of 90 NETLIB problems we have processed,

• in 27 problems the nominal solution turned out to be unreliable at the largest ($\epsilon = 1\%$) level of uncertainty;

• 19 of these 27 problems are already bad at the 0.01%-level of uncertainty, and in 13 of these 19 problems, 0.01% perturbations of the uncertain data can make the nominal solution more than 50%-infeasible for some of the constraints.

The details are given in Table 2.2.

Problem	Size[a)]	$\epsilon = 0.01\%$		$\epsilon = 0.1\%$		$\epsilon = 1\%$	
		Nbad[b)]	Index[c)]	Nbad	Index	Nbad	Index
80BAU3B	2263×9799	37	84	177	842	364	8,420
25FV47	822×1571	14	16	28	162	35	1,620
ADLITTLE	57×97			2	6	7	58
AFIRO	28×32			1	5	2	50
BNL2	2325×3489					24	34
BRANDY	221×249					1	5
CAPRI	272×353			10	39	14	390
CYCLE	1904×2857	2	110	5	1,100	6	11,000
D2Q06C	2172×5167	107	1,150	134	11,500	168	115,000
E226	224×282					2	15
FFFFF800	525×854					6	8
FINNIS	498×614	12	10	63	104	97	1,040
GREENBEA	2393×5405	13	116	30	1,160	37	11,600
KB2	44×41	5	27	6	268	10	2,680
MAROS	847×1443	3	6	38	57	73	566
NESM	751×2923					37	20
PEROLD	626×1376	6	34	26	339	58	3,390
PILOT	1442×3652	16	50	185	498	379	4,980
PILOT4	411×1000	42	210,000	63	2,100,000	75	21,000,000
PILOT87	2031×4883	86	130	433	1,300	990	13,000
PILOTJA	941×1988	4	46	20	463	59	4,630
PILOTNOV	976×2172	4	69	13	694	47	6,940
PILOTWE	723×2789	61	12,200	69	122,000	69	1,220,000
SCFXM1	331×457	1	95	3	946	11	9,460
SCFXM2	661×914	2	95	6	946	21	9,460
SCFXM3	991×1371	3	95	9	946	32	9,460
SHARE1B	118×225	1	257	1	2,570	1	25,700

Table 2.2. Bad NETLIB problems.

[a)] # of linear constraints (excluding the box ones) plus 1 and # of variables

[b)] # of constraints with index $> 5\%$

[c)] The worst, over the constraints, reliability index, in %

Our diagnosis leads to the following conclusion:

> $\diamondsuit$ *In real-world applications of Linear Programming one cannot ignore the possibility that a small uncertainty in the data* (intrinsic for most real-world LP programs) *can make the usual optimal solution of the problem completely meaningless from a practical viewpoint.*
>
> Consequently,
>
> $\diamondsuit$ *In applications of LP, there exists a real need of a technique capable of detecting cases when data uncertainty can heavily affect the quality of the nominal solution, and in these cases to generate a "reliable" solution, one which is immune against uncertainty.*

"Treatment". At the treatment phase of our Case Study, we used the Robust Counterpart methodology, as outlined in Example 1, to pass from "unreliable" nominal solutions of bad NETLIB problems to "uncertainty-immunized" robust solutions. The primary goals here were to understand whether "treatment" is at all possible (the Robust Counterpart may happen to be infeasible) and how "costly" it is – by which margin the robust solution is worse, in terms of

the objective, than the nominal solution. The answers to both these questions turned out to be quite encouraging:

- Reliable solutions do exist, except for the four cases corresponding to the highest ($\epsilon = 1\%$) uncertainty level (see the right column in Table 2.3).
- The price of immunization in terms of the objective value is surprisingly low: when $\epsilon \leq 0.1\%$, it never exceeds 1% and it is less than 0.1% in 13 of 23 cases. Thus, passing to the robust solutions, we gain a lot in the ability of the solution to withstand data uncertainty, while losing nearly nothing in optimality.

The details are given in Table 2.3.

Problem	Nominal optimal value	Objective at robust solution		
		$\epsilon = 0.01\%$	$\epsilon = 0.1\%$	$\epsilon = 1\%$
80BAU3B	987224.2	987311.8 (+ 0.01%)	989084.7 (+ 0.19%)	1009229 (+ 2.23%)
25FV47	5501.846	5501.862 (+ 0.00%)	5502.191 (+ 0.01%)	5505.653 (+ 0.07%)
ADLITTLE	225495.0		225594.2 (+ 0.04%)	228061.3 (+ 1.14%)
AFIRO	-464.7531		-464.7500 (+ 0.00%)	-464.2613 (+ 0.11%)
BNL2	1811.237		1811.237 (+ 0.00%)	1811.338 (+ 0.01%)
BRANDY	1518.511			1518.581 (+ 0.00%)
CAPRI	1912.621		1912.738 (+ 0.01%)	1913.958 (+ 0.07%)
CYCLE	1913.958	1913.958 (+ 0.00%)	1913.958 (+ 0.00%)	1913.958 (+ 0.00%)
D2Q06C	122784.2	122793.1 (+ 0.01%)	122893.8 (+ 0.09%)	Infeasible
E226	-18.75193			-18.75173 (+ 0.00%)
FFFFF800	555679.6			555715.2 (+ 0.01%)
FINNIS	172791.1	172808.8 (+ 0.01%)	173269.4 (+ 0.28%)	178448.7 (+ 3.27%)
GREENBEA	-72555250	-72526140 (+ 0.04%)	-72192920 (+ 0.50%)	-68869430 (+ 5.08%)
KB2	-1749.900	-1749.877 (+ 0.00%)	-1749.638 (+ 0.01%)	-1746.613 (+ 0.19%)
MAROS	-58063.74	-58063.45 (+ 0.00%)	-58011.14 (+ 0.09%)	-57312.23 (+ 1.29%)
NESM	14076040			14172030 (+ 0.68%)
PEROLD	-9380.755	-9380.755 (+ 0.00%)	-9362.653 (+ 0.19%)	Infeasible
PILOT	-557.4875	-557.4538 (+ 0.01%)	-555.3021 (+ 0.39%)	Infeasible
PILOT4	-64195.51	-64149.13 (+ 0.07%)	-63584.16 (+ 0.95%)	-58113.67 (+ 9.47%)
PILOT87	301.7109	301.7188 (+ 0.00%)	302.2191 (+ 0.17%)	Infeasible
PILOTJA	-6113.136	-6113.059 (+ 0.00%)	-6104.153 (+ 0.15%)	-5943.937 (+ 2.77%)
PILOTNOV	-4497.276	-4496.421 (+ 0.02%)	-4488.072 (+ 0.20%)	-4405.665 (+ 2.04%)
PILOTWE	-2720108	-2719502 (+ 0.02%)	-2713356 (+ 0.25%)	-2651786 (+ 2.51%)
SCFXM1	18416.76	18417.09 (+ 0.00%)	18420.66 (+ 0.02%)	18470.51 (+ 0.29%)
SCFXM2	36660.26	36660.82 (+ 0.00%)	36666.86 (+ 0.02%)	36764.43 (+ 0.28%)
SCFXM3	54901.25	54902.03 (+ 0.00%)	54910.49 (+ 0.02%)	55055.51 (+ 0.28%)
SHARE1B	-76589.32	-76589.32 (+ 0.00%)	-76589.32 (+ 0.00%)	-76589.29 (+ 0.00%)

Table 2.3. Objective values for nominal and robust solutions to bad NETLIB problems

2.4.3 Robust counterpart of uncertain LP with a CQr uncertainty set

We have seen that the robust counterpart of uncertain LP with simple "constraint-wise" ellipsoidal uncertainty is a conic quadratic problem. This fact is a special case of the following

Proposition 2.4.2 *Consider an uncertain LP*

$$\mathcal{LP}(\mathcal{U}) = \left\{ \min_{x:Ax\geq b} c^T x : (c,A,b) \in \mathcal{U} \right\}$$

and assume that the uncertainty set $\mathcal{U}$ is CQr:

$$\mathcal{U} = \left\{ \zeta = (c,A,B) \in \mathbf{R}^n \times \mathbf{R}^{m\times n} \times \mathbf{R}^m | \exists u : \mathcal{A}(\zeta,u) \equiv P\zeta + Qu + r \geq_{\mathbf{K}} 0 \right\},$$

where $\mathcal{A}(\zeta,u)$ is an affine mapping and $\mathbf{K}$ is a direct product of ice-cream cones. Assume, further, that the above CQR of $\mathcal{U}$ is strictly feasible:

$$\exists(\bar{\zeta},\bar{u}): \quad \mathcal{A}(\bar{\zeta},\bar{u}) >_{\mathbf{K}} 0.$$

Then the robust counterpart of $\mathcal{LP}(\mathcal{U})$ is equivalent to an explicit conic quadratic problem.

Proof. Introducing an additional variable t and denoting by $z = (t,x)$ the extended vector of design variables, we can write down the instances of our uncertain LP in the form

$$\min_z \left\{ d^T z : \alpha_i^T(\zeta) z - \beta_i(\zeta) \geq 0,\ i = 1,...,m+1 \right\} \qquad (\mathrm{LP}[\zeta])$$

with an appropriate vector d; here the functions

$$\alpha_i(\zeta) = A_i\zeta + a_i, \quad \beta_i(\zeta) = b_i^T\zeta + c_i$$

are affine in the data vector ζ. The robust counterpart of our uncertain LP is the optimization program

$$\min_z \left\{ d^T z \to \min : \alpha_i^T(\zeta) z - \beta_i(\zeta) \geq 0 \quad \forall \zeta \in \mathcal{U}\ \forall i = 1,...,m+1 \right\}. \qquad (\mathrm{RC_{ini}})$$

Let us fix i and ask ourselves what does it mean that a vector z satisfies the infinite system of linear inequalities

$$\alpha_i^T(\zeta) z - \beta_i(\zeta) \geq 0 \quad \forall \zeta \in \mathcal{U}. \qquad (\mathrm{C}_i)$$

Clearly, a given vector z possesses this property if and only if the optimal value in the optimization program

$$\min_{\tau,\zeta} \left\{ \tau : \tau \geq \alpha_i^T(\zeta) z - \beta_i(\zeta),\ \zeta \in \mathcal{U} \right\}$$

is nonnegative. Recalling the definition of $\mathcal{U}$, we see that the latter problem is equivalent to the conic quadratic program

$$\min_{\tau,\zeta} \left\{ \tau : \tau \geq \alpha_i^T(\zeta) z - \beta_i(\zeta) \equiv [\underbrace{A_i\zeta + a_i}_{\alpha_i(\zeta)}]^T z - [\underbrace{b_i^T\zeta + c_i}_{\beta_i(\zeta)}],\ \mathcal{A}(\zeta,u) \equiv P\zeta + Qu + r \geq_{\mathbf{K}} 0 \right\} \qquad (\mathrm{CQ}_i[z])$$

in variables τ,ζ,u. Thus, z satisfies (C_i) if and only if the optimal value in $(\mathrm{CQ}_i[z])$ is nonnegative.

Since by assumption the system of conic quadratic inequalities $\mathcal{A}(\zeta,u) \geq_{\mathbf{K}} 0$ is strictly feasible, the conic quadratic program $(\mathrm{CQ}_i[z])$ is strictly feasible. By the Conic Duality Theorem, <u>if</u> (a) the optimal value in $(\mathrm{CQ}_i[z])$ is nonnegative, <u>then</u> (b) the dual to $(\mathrm{CQ}_i[z])$ problem admits a feasible solution with a nonnegative value of the dual objective. By Weak Duality, (b) implies

(a). Thus, the fact that the optimal value in $(\mathrm{CQ}_i[z])$ is nonnegative *is equivalent* to the fact that the dual problem admits a feasible solution with a nonnegative value of the dual objective:

$$
\begin{array}{c}
z \text{ satisfies } (\mathrm{C}_i) \\
\Updownarrow \\
\mathrm{Opt}(\mathrm{CQ}_i[z]) \geq 0 \\
\Updownarrow \\
\left\{\begin{array}{c}
\exists \lambda \in \mathbf{R}, \xi \in \mathbf{R}^N (N \text{ is the dimension of } \mathbf{K}): \\
\lambda[a_i^T z - c_i] - \xi^T r \geq 0, \\
\lambda = 1, \\
-\lambda A_i^T z + b_i + P^T \xi = 0, \\
Q^T \xi = 0, \\
\lambda \geq 0, \\
\xi \geq_{\mathbf{K}} 0.
\end{array}\right. \\
\Updownarrow \\
\left\{\begin{array}{c}
\exists \xi \in \mathbf{R}^N : \\
a_i^T z - c_i - \xi^T r \geq 0 \\
-A_i^T z + b_i + P^T \xi = 0, \\
Q^T \xi = 0, \\
\xi \geq_{\mathbf{K}} 0.
\end{array}\right.
\end{array}
$$

We see that *the set of vectors z satisfying (C_i) is CQr:*

$$
\begin{array}{c}
z \text{ satisfies } (\mathrm{C}_i) \\
\Updownarrow \\
\left\{\begin{array}{c}
\exists \xi \in \mathbf{R}^N : \\
a_i^T z - c_i - \xi^T r \geq 0, \\
-A_i^T z + b_i + P^T \xi = 0, \\
Q^T \xi = 0, \\
\xi \geq_{\mathbf{K}} 0.
\end{array}\right.
\end{array}
$$

Consequently, the set of robust feasible z – those satisfying (C_i) for all $i = 1, ..., m+1$ – is CQr (as the intersection of finitely many CQr sets), whence the robust counterpart of our uncertain LP, being the problem of minimizing a linear objective over a CQr set, is equivalent to a conic quadratic problem. Here is this problem:

$$
\begin{array}{c}
\text{minimize } d^T z \\
\left\{\begin{array}{c}
a_i^T z - c_i - \xi_i^T r \geq 0, \\
-A_i^T z + b_i + P^T \xi_i = 0, \\
Q^T \xi_i = 0, \\
\xi_i \geq_{\mathbf{K}} 0
\end{array}\right. , i = 1, ..., m+1
\end{array}
$$

with design variables $z, \xi_1, ..., \xi_{m+1}$. Here A_i, a_i, c_i, b_i come from the affine functions $\alpha_i(\zeta) = A_i\zeta + a_i$ and $\beta_i(\zeta) = b_i^T\zeta + c_i$, while P, Q, r come from the description of $\mathcal{U}$:

$$
\mathcal{U} = \{\zeta : \exists u : \quad P\zeta + Qu + r \geq_{\mathbf{K}} 0\}.
$$

■

Remark 2.4.1 Looking at the proof of Proposition 2.4.2, we see that the assumption that the uncertainty set $\mathcal{U}$ is CQr plays no crucial role. What indeed is important is that $\mathcal{U}$ is the projection on the ζ-space of the solution set of a strictly feasible conic inequality associated with certain cone $\mathbf{K}$. Whenever this is the case, the above construction demonstrates that the robust counterpart of $\mathcal{LP}(\mathcal{U})$ is a conic problem associated with the cone which is a direct product of several cones dual to $\mathbf{K}$. E.g., when the uncertainty set is polyhedral (i.e., it is given by finitely many scalar linear inequalities: $\mathbf{K} = \mathbf{R}^m_+$), the robust counterpart of $\mathcal{LP}(\mathcal{U})$ is an explicit LP program (and in this case we can eliminate the assumption that the conic inequality defining $\mathcal{U}$ is strictly feasible (why?)). Consider, e.g., an uncertain LP with *interval* uncertainty in the data:

$$\left\{ \min_x \left\{ c^T x : Ax \geq b \right\} : \begin{array}{l} |c_j - c_j^*| \leq \epsilon_j,\ j = 1, ..., n \\ A_{ij} \in [A_{ij}^* - \epsilon_{ij}, A_{ij}^* + \epsilon_{ij}], i = 1, ..., m, j = 1, ..., n \\ |b_i - b_i^*| \leq \delta_i,\ i = 1, ..., m \end{array} \right\}.$$

The (LP equivalent of the) Robust Counterpart of the program is

$$\min_{x,y} \left\{ \sum_j [c_j^* x_j + \epsilon_j y_j] : \begin{array}{l} \sum\limits_j A_{ij}^* x_j - \sum\limits_j \epsilon_{ij} y_j \geq b_i^* + \delta_i,\ i = 1, ..., m \\ -y_j \leq x_j \leq y_j,\ j = 1, ..., n \end{array} \right\}$$

(why ?)

2.4.4 CQ-representability of the optimal value in a CQ program as a function of the data

Let us ask ourselves the following question: consider a conic quadratic program

$$\min_x \left\{ c^T x : Ax - b \geq_{\mathbf{K}} 0 \right\}, \tag{2.4.6}$$

where $\mathbf{K}$ is a direct product of ice-cream cones and A is a matrix with trivial null space. The optimal value of the problem clearly is a function of the data (c, A, b) of the problem. What can be said about CQ-representability of this function? In general, not much: the function is not even convex. There are, however, two modifications of our question which admit good answers. Namely, under mild regularity assumptions

(a) With c, A fixed, the optimal value is a CQ-representable function of the right hand side vector b;

(b) with A, b fixed, the minus optimal value is a CQ-representable function of c.

Here are the exact forms of our claims:

Proposition 2.4.3 *Let c, A be fixed, and let $\mathcal{B}$ be a CQr set in $\mathbf{R}^{\dim b}$ such that for every $b \in \mathcal{B}$ problem* (2.4.6) *is strictly feasible. Then the optimal value of the problem is a CQr function on $\mathcal{B}$.*

The statement is quite evident: if b is such that (2.4.6) is strictly feasible, then the optimal value $\mathrm{Opt}(b)$ in the problem is either $-\infty$, or is achieved (by Conic Duality Theorem). In both cases,

$$\mathrm{Opt}(b) \leq t \Leftrightarrow \exists x : \left\{ \begin{array}{c} c^T x \leq t, \\ Ax - b \geq_{\mathbf{K}} 0 \end{array} \right. ,$$

which is, essentially, a CQR for certain function which coincides with $\mathrm{Opt}(b)$ on the set $\mathcal{B}$ of values of b; in this CQR, b, t are the "variables of interest", and x plays the role of the additional variable. The CQR of the function $\mathrm{Opt}(b)$ with the domain $\mathcal{B}$ is readily given, via calculus of CQR's, by the representation

$$\{b \in \mathcal{B} \& \mathrm{Opt}(b) \le t\} \Leftrightarrow \exists x : \left\{ \begin{array}{c} c^T x \le t, \\ Ax - b \ge_{\mathbf{K}} 0, \\ b \in \mathcal{B} \end{array} \right.$$

(recall that $\mathcal{B}$ was assumed to be CQr). ∎

The claim (b) is essentially a corollary of (a) – via duality, the optimal value in (2.4.6) is, up to pathological cases, the same as the optimal value in the dual problem, in which c becomes the right hand side vector. Here is the exact formulation:

Proposition 2.4.4 *Let A, b be such that* (2.4.6) *is strictly feasible. Then the minus optimal value $-\mathrm{Opt}(c)$ of the problem is a CQr function of c.*

Proof. For every c and t, the relation $-\mathrm{Opt}(c) \le t$ says *exactly* that (2.4.6) is below bounded with the optimal value $\ge -t$. By the Conic Duality Theorem (note that (2.4.6) is strictly feasible!) this is the case if and only if the dual problem admits a feasible solution with the value of the dual objective $\ge -t$, so that

$$\mathrm{Opt}(c) \le t \Leftrightarrow \exists y : \left\{ \begin{array}{c} b^T y \ge t, \\ A^T y = c, \\ y \ge_{\mathbf{K}} 0. \end{array} \right.$$

The resulting description of the epigraph of the function $-\mathrm{Opt}(c)$ is a CQR for this function, with c, t playing the role of the "variables of interest" and y being the additional variable. ∎
A careful reader could have realized that Proposition 2.4.2 is nothing but a straightforward application of Proposition 2.4.4.

Remark 2.4.2 Same as in the case of Proposition 2.4.2, in Propositions 2.4.3, 2.4.4 the assumption that uncertainty set $\mathcal{U}$ is CQr plays no crucial role. Thus, *Propositions 2.4.3, 2.4.4 remain valid for an arbitrary conic program, up to the fact that in this general case we should speak about the representability of the epigraphs of* $\mathrm{Opt}(b)$ *and* $-\mathrm{Opt}(c)$ *via conic inequalities associated with direct products of cones* $\mathbf{K}$*, and their duals, rather than about CQ-representability. In particular, Propositions 2.4.2, 2.4.3, 2.4.4 remain valid in the case of Semidefinite representability to be discussed in Lecture 3.*

2.4.5 Affinely Adjustable Robust Counterpart

The rationale behind our Robust Optimization paradigm is based on the following tacit assumptions:

1. All constraints are "a must", so that a meaningful solution should satisfy all realizations of the constraints from the uncertainty set.
2. All decisions are made in advance and thus cannot tune themselves to the "true" values of the data. Thus, candidate solutions must be fixed vectors, and not functions of the true data.

Here we preserve the first of these two assumptions and try to relax the second of them. The motivation is twofold:

- There are situations in dynamical decision-making when the decisions should be made at subsequent time instants, and decision made at instant t in principle can depend on the part of uncertain data which becomes known at this instant.
- There are situations in LP when some of the decision variables do not correspond to actual decisions; they are artificial "analysis variables" added to the problem in order to convert it to a desired form, say, a Linear Programming one. The analysis variables clearly may adjust themselves to the true values of the data.

  To give an example, consider the problem where we look for the best, in the discrete L_1-norm, approximation of a given sequence b by a linear combination of given sequences a_j, $j = 1, ..., n$, so that the problem with no data uncertainty is

$$\min_{x,t}\left\{t : \sum_{t=1}^{T} |b_t - \sum_j a_{tj}x_j| \le t\right\} \qquad \text{(P)}$$

$$\Updownarrow$$

$$\min_{t,x,y}\left\{t : \sum_{t=1}^{T} y_t \le t, -y_t \le b_t - \sum_j a_{tj}x_j \le y_t, 1 \le t \le T\right\} \quad \text{(LP)}$$

  Note that (LP) is an equivalent LP reformulation of (P), and y are typical analysis variables; whether x's do or do not represent "actual decisions", y's definitely do not represent them. Now assume that the data become uncertain. Perhaps we have reasons to require from (t, x)s to be independent of actual data and to satisfy the constraint in (P) for all realizations of the data. This requirement means that the variables t, x in (LP) must be data-independent, but we have absolutely no reason to insist on data-independence of y's: (t, x) is robust feasible for (P) if and only if (t, x), for all realizations of the data from the uncertainty set, can be extended, by a properly chosen *and perhaps depending on the data* vector y, to a feasible solution of (the corresponding realization of) (LP). In other words, *equivalence between* (P) *and* (LP) *is restricted to the case of certain data only; when the data become uncertain, the robust counterpart of* (LP) *is more conservative than the one of* (P).

In order to take into account a possibility for (part of) the variables to adjust themselves to the true values of (part of) the data, we could act as follows.

Adjustable and non-adjustable decision variables. Consider an uncertain LP program. Without loss of generality, we may assume that the data are affinely parameterized by properly chosen "perturbation vector" ζ running through a given *perturbation set* $\mathcal{Z}$; thus, our uncertain LP can be represented as the family of LP instances

$$\mathcal{LP} = \left\{\min_x \left\{c^T[\zeta]x : A[\zeta]x - b[\zeta] \ge 0\right\} : \zeta \in \mathcal{Z}\right\}$$

Now assume that decision variable x_j is allowed to depend on part of the true data. Since the true data are affine functions of ζ, this is the same as to assume that x_j can depend on "a part" $P_j\zeta$ of the perturbation vector, where P_j is a given matrix. The case of $P_j = 0$ correspond to "here and now" decisions x_j – those which should be done in advance; we shall call these decision variables *non-adjustable*. The case of nonzero P_j ("*adjustable decision variable*") corresponds to allowing certain dependence of x_j on the data, and the case when P_j has trivial kernel means that x_j is allowed to depend on the entire true data.

Adjustable Robust Counterpart of $\mathcal{LP}$. With our assumptions, a natural modification of the Robust Optimization methodology results in the following *adjustable Robust Counterpart* of $\mathcal{LP}$:

$$\min_{t,\{\phi_j(\cdot)\}_{j=1}^n}\left\{t : \begin{array}{l} \sum_{j=1}^{n} c_j[\zeta]\phi_j(P_j\zeta) \le t \ \forall \zeta \in \mathcal{Z} \\ \sum_{j=1}^{n} \phi_j(P_j\zeta)A_j[\zeta] - b[\zeta] \ge 0 \ \forall \zeta \in \mathcal{Z} \end{array}\right\} \qquad \text{(ARC)}$$

Here $c_j[\zeta]$ is j-th entry of the objective vector, and $A_j[\zeta]$ is j-th column of the constraint matrix.

It should be stressed that *the variables in* (ARC) *corresponding to adjustable decision variables in the original problem are* <u>*not*</u> *reals; they are "decision rules" – real-valued functions of the corresponding portion $P_j\zeta$ of the data.* This fact makes (ARC) *infinite-dimensional* optimization problem and thus problem which is extremely difficult for numerical processing. Indeed, in general it is unclear how to *represent* in a tractable way a general-type function of three (not speaking of three hundred) variables; and how could we hope to find, in an efficient manner, *optimal* decision rules when we even do not know how to write them down? Thus, in general (ARC) has no actual meaning – basically all we can do with the problem is to write it down on paper and then look at it...

Affinely Adjustable Robust Counterpart of $\mathcal{LP}$. A natural way to overcome the outlined difficulty is to restrict the decision rules to be "easily representable", specifically, *to be affine functions of the allowed portions of data*:

$$\phi_j(P_j\zeta) = \mu_j + \nu_j^T P_j\zeta.$$

With this approach, *our new decision variables become reals μ_j and vectors ν_j*, and (ARC) becomes the following problem (called *Affinely Adjustable Robust Counterpart of $\mathcal{LP}$*):

$$\min_{t,\{\mu_j,\nu_j\}_{j=1}^n} \left\{ t : \begin{array}{l} \sum\limits_j c_j[z][\mu_j + \nu_j^T P_j\zeta] \le t \ \forall \zeta \in \mathcal{Z} \\ \sum\limits_j [\mu_j + \nu_j^T P_j] A_j[\zeta] - b[\zeta] \ge 0 \ \forall \zeta \in \mathcal{Z} \end{array} \right\} \qquad \text{(AARC)}$$

Note that the AARC is "in-between" the usual non-adjustable RC (no dependence of variables on the true data at all) and the ARC (arbitrary dependencies of the decision variables on the allowed portions of the true data). Note also that the *only* reason to restrict ourselves with affine decision rules is the desire to end up with a "tractable" robust counterpart, and even this natural goal for the time being is not achieved. Indeed, the constraints in (AARC) are affine in our new decision variables t, μ_j, ν_j, which is a good news. At the same time, they are *semi-infinite*, same as in the case of the non-adjustable Robust Counterpart, but, in contrast to this latter case, in general are *quadratic* in perturbations rather than to be linear in them. This indeed makes a difference: as we know from Proposition 2.4.2, the usual – non-adjustable – RC of an uncertain LP with CQr uncertainty set is equivalent to an explicit Conic Quadratic problem and as such is computationally tractable (in fact, the latter remain true for the case of non-adjustable RC of uncertain LP with arbitrary "computationally tractable" uncertainty set). In contrast to this, AARC can become intractable for uncertainty sets as simple as boxes. There are, however, good news on AARCs:

- First, there exist a generic "good case" where the AARC is tractable. This is the "fixed recourse" case, where the coefficients of adjustable variables x_j – those with $P_j \neq 0$ – are certain (not affected by uncertainty). In this case, the left hand sides of the constraints in (AARC) are affine in ζ, and thus AARC, same as the usual non-adjustable RC, is computationally tractable whenever the perturbation set $\mathcal{Z}$ is so; in particular, Proposition 2.4.2 remains valid for both RC and AARC.
- Second, we shall see in Lecture 3 that even when AARC is intractable, it still admits tight, in certain precise sense, tractable approximations.

Example: Uncertain Inventory Management Problem

The model. Consider a single product inventory system comprised of a warehouse and I factories. The planning horizon is T periods. At a period t:

- d_t is the demand for the product. All the demand must be satisfied;
- $v(t)$ is the amount of the product in the warehouse at the beginning of the period ($v(1)$ is given);
- $p_i(t)$ is the i-th order of the period – the amount of the product to be produced during the period by factory i and used to satisfy the demand of the period (and, perhaps, to replenish the warehouse);

- $P_i(t)$ is the maximal production capacity of factory i;
- $c_i(t)$ is the cost of producing a unit of the product at a factory i.

Other parameters of the problem are:

- $V_{\min}$ - the minimal allowed level of inventory at the warehouse;
- $V_{\max}$ - the maximal storage capacity of the warehouse;
- Q_i - the maximal cumulative production capacity of i'th factory throughout the planning horizon.

The goal is to minimize the total production cost over all factories and the entire planning period. When all the data are certain, the problem can be modelled by the following linear program:

$$\begin{array}{ll} \min\limits_{p_i(t),v(t),F} & F \\ \text{s.t.} & \sum\limits_{t=1}^{T}\sum\limits_{i=1}^{I} c_i(t)p_i(t) \le F \\ & 0 \le p_i(t) \le P_i(t),\ i=1,\dots,I,\ t=1,\dots,T \\ & \sum\limits_{t=1}^{T} p_i(t) \le Q(i),\ i=1,\dots,I \\ & v(t+1) = v(t) + \sum\limits_{i=1}^{I} p_i(t) - d_t,\ t=1,\dots,T \\ & V_{\min} \le v(t) \le V_{\max},\ t=2,\dots,T+1. \end{array} \tag{2.4.7}$$

Eliminating v-variables, we get an inequality constrained problem:

$$\begin{array}{ll} \min\limits_{p_i(t),F} & F \\ \text{s.t.} & \sum\limits_{t=1}^{T}\sum\limits_{i=1}^{I} c_i(t)p_i(t) \le F \\ & 0 \le p_i(t) \le P_i(t),\ i=1,\dots,I,\ t=1,\dots,T \\ & \sum\limits_{t=1}^{T} p_i(t) \le Q(i),\ i=1,\dots,I \\ & V_{\min} \le v(1) + \sum\limits_{s=1}^{t}\sum\limits_{i=1}^{I} p_i(s) - \sum\limits_{s=1}^{t} d_s \le V_{\max},\ t=1,\dots,T. \end{array} \tag{2.4.8}$$

Assume that the decision on supplies $p_i(t)$ is made at the beginning of period t, and that we are allowed to make these decisions on the basis of demands d_r observed at periods $r \in I_t$, where I_t is a given subset of $\{1,...,t\}$. Further, assume that we should specify our supply policies before the planning period starts ("at period 0"), and that when specifying these policies, we do not know exactly the future demands; all we know is that

$$d_t \in [d_t^* - \theta d_t^*, d_t^* + \theta d_t^*],\ t=1,\dots,T, \tag{2.4.9}$$

with given positive θ and positive nominal demand d_t^*. We have now an uncertain LP, where the uncertain data are the actual demands d_t, the decision variables are the supplies $p_i(t)$, and these decision variables are allowed to depend on the data $\{d_\tau : \tau \in I_t\}$ which become known when $p_i(t)$ should be specified. Note that our uncertain LP is a "fixed recourse" one – the uncertainty affects solely the right hand side. Thus, the AARC of the problem is computationally tractable, which is good. Let us build the AARC. Restricting our decision-making policy with *affine decision rules*

$$p_i(t) = \pi_{i,t}^0 + \sum_{r\in I_t} \pi_{i,t}^r d_r, \tag{2.4.10}$$

where the coefficients $\pi^r_{i,t}$ are our new non-adjustable design variables, we get from (2.4.8) the following uncertain Linear Programming problem in variables $\pi^s_{i,t}$, F:

$$\begin{array}{ll}\min\limits_{\pi,F} & F\\ \text{s.t.} & \sum\limits_{t=1}^{T}\sum\limits_{i=1}^{I} c_i(t)\left(\pi^0_{i,t}+\sum\limits_{r\in I_t}\pi^r_{i,t}d_r\right)\le F\\ & 0\le \pi^0_{i,t}+\sum\limits_{r\in I_t}\pi^r_{i,t}d_r\le P_i(t),\ i=1,\dots,I,\ t=1,\dots,T\\ & \sum\limits_{t=1}^{T}\left(\pi^0_{i,t}+\sum\limits_{r\in I_t}\pi^r_{i,t}d_r\right)\le Q(i),\ i=1,\dots,I\\ & V_{\min}\le v(1)+\sum\limits_{s=1}^{t}\left(\sum\limits_{i=1}^{I}\pi^0_{i,s}+\sum\limits_{r\in I_s}\pi^r_{i,s}d_r\right)-\sum\limits_{s=1}^{t}d_s\le V_{\max},\\ & \qquad t=1,\dots,T\\ & \forall\{d_t\in[d^*_t-\theta d^*_t,d^*_t+\theta d^*_t],\ t=1,\dots,T\},\end{array}\tag{2.4.11}$$

or, which is the same,

$$\begin{array}{ll}\min\limits_{\pi,F} & F\\ \text{s.t.} & \sum\limits_{t=1}^{T}\sum\limits_{i=1}^{I} c_i(t)\pi^0_{i,t}+\sum\limits_{r=1}^{T}\left(\sum\limits_{i=1}^{I}\sum\limits_{t:r\in I_t} c_i(t)\pi^r_{i,t}\right)d_r-F\le 0\\ & \pi^0_{i,t}+\sum\limits_{r\in I_t}^{t}\pi^r_{i,t}d_r\le P_i(t),\ i=1,\dots,I,\ t=1,\dots,T\\ & \pi^0_{i,t}+\sum\limits_{r\in I_t}\pi^r_{i,t}d_r\ge 0,\ i=1,\dots,I,\ t=1,\dots,T\\ & \sum\limits_{t=1}^{T}\pi^0_{i,t}+\sum\limits_{r=1}^{T}\left(\sum\limits_{t:r\in I_t}\pi^r_{i,t}\right)d_r\le Q_i,\ i=1,\dots,I\\ & \sum\limits_{s=1}^{t}\sum\limits_{i=1}^{I}\pi^0_{i,s}+\sum\limits_{r=1}^{t}\left(\sum\limits_{i=1}^{I}\sum\limits_{s\le t,r\in I_s}\pi^r_{i,s}-1\right)d_r\le V_{\max}-v(1)\\ & \qquad t=1,\dots,T\\ & -\sum\limits_{s=1}^{t}\sum\limits_{i=1}^{I}\pi^0_{i,s}-\sum\limits_{r=1}^{t}\left(\sum\limits_{i=1}^{I}\sum\limits_{s\le t,r\in I_s}\pi^r_{i,s}-1\right)d_r\le v(1)-V_{\min}\\ & \qquad t=1,\dots,T\\ & \forall\{d_t\in[d^*_t-\theta d^*_t,d^*_t+\theta d^*_t],\ t=1,\dots,T\}.\end{array}\tag{2.4.12}$$

Now, using the following equivalences

$$\begin{array}{c}\sum\limits_{t=1}^{T}d_tx_t\le y,\ \forall d_t\in[d^*_t(1-\theta),d^*_t(1+\theta)]\\ \Updownarrow\\ \sum\limits_{t:x_t<0}d^*_t(1-\theta)x_t+\sum\limits_{t:x_t>0}d^*_t(1+\theta)x_t\le y\\ \Updownarrow\\ \sum\limits_{t=1}^{T}d^*_tx_t+\theta\sum\limits_{t=1}^{T}d^*_t|x_t|\le y,\end{array}$$

and defining additional variables

$$\alpha_r\equiv\sum_{t:r\in I_t}c_i(t)\pi^r_{i,t};\quad \delta^r_i\equiv\sum_{t:r\in I_t}\pi^r_{i,t};\quad \xi^r_t\equiv\sum_{i=1}^{I}\sum_{s\le t,r\in I_s}\pi^r_{i,s}-1,$$

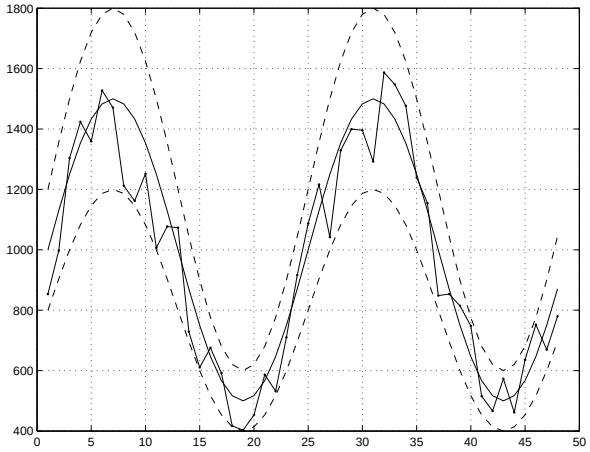
we can straightforwardly convert the AARC (2.4.12) into an equivalent LP (cf. Remark 2.4.1):

$$\begin{array}{c}\min\limits_{\pi,F,\alpha,\beta,\gamma,\delta,\zeta,\xi,\eta} F\\ \sum\limits_{i=1}^{I}\sum\limits_{t:r\in I_t} c_i(t)\pi^r_{i,t}=\alpha_r,\ -\beta_r\le\alpha_r\le\beta_r,\ 1\le r\le T,\ \sum\limits_{t=1}^{T}\sum\limits_{i=1}^{I} c_i(t)\pi^0_{i,t}+\sum\limits_{r=1}^{T}\alpha_r d^*_r+\theta\sum\limits_{r=1}^{T}\beta_r d^*_r\le F;\\ -\gamma^r_{i,t}\le\pi^r_{i,t}\le\gamma^r_{i,t},\ r\in I_t,\ \pi^0_{i,t}+\sum\limits_{r\in I_t}\pi^r_{i,t}d^*_r+\theta\sum\limits_{r\in I_t}\gamma^r_{i,t}d^*_r\le P_i(t),\ 1\le i\le I, 1\le t\le T;\\ \pi^0_{i,t}+\sum\limits_{r\in I_t}\pi^r_{i,t}d^*_r-\theta\sum\limits_{r\in I_t}\gamma^r_{i,t}d^*_r\ge 0,\ \sum\limits_{t:r\in I_t}\pi^r_{i,t}=\delta^r_i,\ -\zeta^r_i\le\delta^r_i\le\zeta^r_i,\ 1\le i\le I, 1\le r\le T,\\ \sum\limits_{t=1}^{T}\pi^0_{i,t}+\sum\limits_{r=1}^{T}\delta^r_i d^*_r+\theta\sum\limits_{r=1}^{T}\zeta^r_i d^*_r\le Q_i,\ 1\le i\le I;\\ \sum\limits_{i=1}^{I}\sum\limits_{s\le t, r\in I_s}\pi^r_{i,s}-\xi^r_t=1,\ -\eta^r_t\le\xi^r_t\le\eta^r_t,\ 1\le r\le t\le T,\\ \sum\limits_{s=1}^{t}\sum\limits_{i=1}^{I}\pi^0_{i,s}+\sum\limits_{r=1}^{t}\xi^r_t d^*_r+\theta\sum\limits_{r=1}^{t}\eta^r_t d^*_r\le V_{\max}-v(1),\ 1\le t\le T,\\ \sum\limits_{s=1}^{t}\sum\limits_{i=1}^{I}\pi^0_{i,s}+\sum\limits_{r=1}^{t}\xi^r_t d^*_r-\theta\sum\limits_{r=1}^{t}\eta^r_t d^*_r\ge v(1)-V_{\min},\ 1\le t\le T.\end{array} \tag{2.4.13}$$

An illustrative example. There are $I=3$ factories producing a seasonal product, and one warehouse. The decisions concerning production are made every two weeks, and we are planning production for 48 weeks, thus the time horizon is $T=24$ periods. The nominal demand d^* is seasonal, reaching its maximum in winter, specifically,

$$d^*_t=1000\left(1+\frac{1}{2}\sin\left(\frac{\pi\,(t-1)}{12}\right)\right),\quad t=1,\ldots,24.$$

We assume that the uncertainty level θ is 20%, i.e., $d_t\in[0.8d^*_t, 1.2d^*_t]$, as shown on the picture.



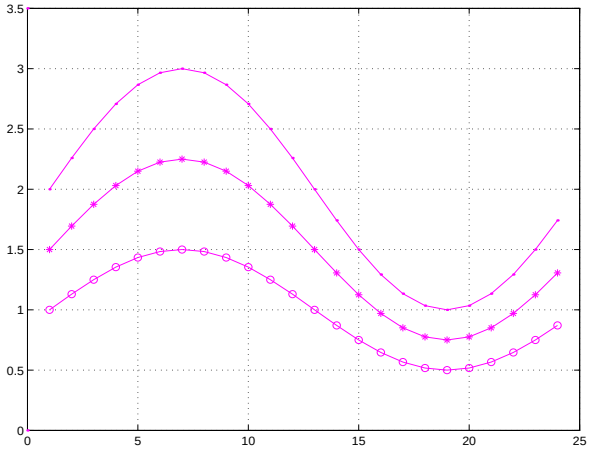
- Nominal demand (solid)
- "demand tube" – nominal demand ±20% (dashed)
- a sample realization of actual demand (dotted)

Demand

The production costs per unit of the product depend on the factory and on time and follow the same seasonal pattern as the demand, i.e., rise in winter and fall in summer. The production cost for a factory i at a period t

is given by:

$$c_i(t) = \alpha_i \left(1 + \tfrac{1}{2}\sin\left(\tfrac{\pi\,(t-1)}{12}\right)\right), \quad t = 1, \ldots, 24.$$
$$\alpha_1 = 1$$
$$\alpha_2 = 1.5$$
$$\alpha_3 = 2$$



Production costs for the 3 factories

The maximal production capacity of each one of the factories at each two-weeks period is $P_i(t) = 567$ units, and the integral production capacity of each one of the factories for a year is $Q_i = 13600$. The inventory at the warehouse should not be less then 500 units, and cannot exceed 2000 units.

With this data, the AARC (2.4.13) of the uncertain inventory problem is an LP, the dimensions of which vary, depending on the "information basis" (see below), from 919 variables and 1413 constraints (empty information basis) to 2719 variables and 3213 constraints (on-line information basis).

The experiments. In every one of the experiments, the corresponding management policy was tested against a given number (100) of simulations; in every one of the simulations, the actual demand d_t of period t was drawn at random, according to the uniform distribution on the segment $[(1-\theta)d_t^*, (1+\theta)d_t^*]$ where θ was the "uncertainty level" characteristic for the experiment. The demands of distinct periods were independent of each other.

We have conducted two series of experiments:

1. The aim of the first series of experiments was to check the influence of the demand uncertainty θ on the total production costs corresponding to the *robustly adjustable* management policy – the policy (2.4.10) yielded by the optimal solution to the AARC (2.4.13). We compared this cost to the "ideal" one, i.e., the cost we would have paid in the case when all the demands were known to us in advance and we were using the corresponding optimal management policy as given by the optimal solution of (2.4.7).
2. The aim of the second series of experiments was to check the influence of the "information basis" allowed for the management policy, on the resulting management cost. Specifically, in our model as described in the previous section, when making decisions $p_i(t)$ at time period t, we can make these decisions depending on the demands of periods $r \in I_t$, where I_t is a given subset of the segment $\{1, 2, ..., t\}$. The larger are these subsets, the more flexible can be our decisions, and hopefully the less are the corresponding management costs. In order to quantify this phenomenon, we considered 4 "information bases" of the decisions:
   (a) $I_t = \{1, ..., t\}$ (the richest "on-line" information basis);
   (b) $I_t = \{1, ..., t-1\}$ (this *standard information basis* seems to be the most natural "information basis": past is known, present and future are unknown);
   (c) $I_t = \{1, ..., t-4\}$ (the information about the demand is received with a four-day delay);
   (d) $I_t = \emptyset$ (i.e., no adjusting of future decisions to actual demands at all. This "information basis" corresponds exactly to the management policy yielded by the usual RC of our uncertain LP.).

The results of our experiments are as follows:

1. The influence of the uncertainty level on the management cost. Here we tested the robustly adjustable management policy *with the standard information basis* against different levels of uncertainty, specifically, the levels of 20%, 10%, 5% and 2.5%. For every uncertainty level, we have computed the average (over 100 simulations) management costs when using the corresponding robustly adaptive management policy. We saved the simulated demand trajectories and then used these trajectories to compute the ideal management costs. The results are summarized in the table below. As expected, the less is the uncertainty, the closer are our management costs to the ideal ones. What is surprising, is the low "price of robustness": even at the 20% uncertainty level, the average management cost for the robustly adjustable policy was just by 3.4% worse than the corresponding ideal cost; the similar quantity for 2.5%-uncertainty in the demand was just 0.3%.

	AARC		Ideal case		
Uncertainty	Mean	Std	Mean	Std	price of robustness
2.5%	33974	190	33878	194	0.3%
5%	34063	432	33864	454	0.6%
10%	34471	595	34009	621	1.6%
20%	35121	1458	33958	1541	3.4%

Management costs vs. uncertainty level

2. The influence of the information basis. The influence of the information basis on the performance of the robustly adjustable management policy is displayed in the following table:

information basis for decision $p_i(t)$ is demand in periods	Management cost	
	Mean	Std
$1, ..., t$	34583	1475
$1, \ldots, t-1$	35121	1458
$1, \ldots, t-4$	Infeasible	
$\emptyset$	Infeasible	

These experiments were carried out at the uncertainty level of 20%. We see that the poorer is the information basis of our management policy, the worse are the results yielded by this policy. In particular, with 20% level of uncertainty, there does not exist a robust *non-adjustable* management policy: the usual RC of our uncertain LP is infeasible. In other words, in our illustrating example, passing from a priori decisions yielded by RC to "adjustable" decisions yielded by AARC is indeed crucial.

An interesting question is what is the uncertainty level which still allows for a priori decisions. It turns out that the RC is infeasible even at the 5% uncertainty level. Only at the uncertainty level as small as 2.5% the RC becomes feasible and yields the following management costs:

	RC		Ideal cost		
Uncertainty	Mean	Std	Mean	Std	price of robustness
2.5%	35287	0	33842	172	4.3%

Note that even at this unrealistically small uncertainty level the price of robustness for the policy yielded by the RC is by 4.3% larger than the ideal cost (while for the robustly adjustable management this difference is just 0.3%.

Comparison with Dynamic Programming. An Inventory problem we have considered is a typical example of sequential decision-making under dynamical uncertainty, where the information basis for the decision x_t made at time t is the part of the uncertainty revealed at instant t. This example allows for an instructive comparison of the AARC-based approach with Dynamic Programming, which is the traditional technique for sequential decision-making under dynamical uncertainty. Restricting ourselves with the case where the decision-making problem can be modelled as a Linear Programming problem with the data affected by dynamical uncertainty, we could say that (minimax-oriented) *Dynamic Programming is a specific technique for solving the ARC of this uncertain LP.* Therefore *when applicable,* Dynamic Programming has a significant advantage as compared to the above AARC-based approach,

since it does *not* impose on the adjustable variables an "ad hoc" restriction (motivated solely by the desire to end up with a tractable problem) to be affine functions of the uncertain data. At the same time, the above "if applicable" is highly restrictive: the computational effort in Dynamical Programming explodes exponentially with the dimension of the state space of the dynamical system in question. For example, the simple Inventory problem we have considered has 4-dimensional state space (the current amount of product in the warehouse plus remaining total capacities of the three factories), which is already computationally too demanding for accurate implementation of Dynamic Programming. The main advantage of the AARC-based dynamical decision-making as compared with Dynamic Programming (as well as with Multi-Stage Stochastic Programming) comes from the "built-in" computational tractability of the approach, which prevents the "curse of dimensionality" and allows to process routinely fairly complicated models with high-dimensional state spaces and many stages.

By the way, it is instructive to compare the AARC approach with Dynamic Programming when the latter is applicable. For example, let us reduce the number of factories in our Inventory problem from 3 to 1, increasing the production capacity of this factory from the previous 567 to 1800 units per period, and let us make the cumulative capacity of the factory equal to 24×1800, so that the restriction on cumulative production becomes redundant. The resulting dynamical decision-making problem has just one-dimensional state space (all which matters for the future is the current amount of product in the warehouse). Therefore we can easily find by Dynamic Programming the "minimax optimal" inventory management cost (minimum over *arbitrary* casual[6)] decision rules, maximum over the realizations of the demands from the uncertainty set). With 20% uncertainty, this minimax optimal inventory management cost turns out to be $\mathrm{Opt}_* = 31269.69$. The guarantees for the AARC-based inventory policy can be only worse than for the minimax optimal one: we should pay a price for restricting the decision rules to be affine in the demands. How large is this price? Computation shows that the optimal value in the AARC is $\mathrm{Opt}_{\mathrm{AARC}} = 31514.17$, i.e., it is just by 0.8% larger than the minimax optimal cost Opt_*. And all this – at the uncertainty level as large as 20%! We conclude that the AARC is perhaps not as bad as one could think...

2.5 Does Conic Quadratic Programming exist?

Of course it does. What is meant is *whether SQP exists as an independent entity?*. Specifically, we ask:

> (?) *Whether a conic quadratic problem can be "efficiently approximated" by a Linear Programming one?*

To pose the question formally, let us say that a system of linear inequalities

$$Py + tp + Qu \geq 0 \tag{LP}$$

approximates the conic quadratic inequality

$$\|y\|_2 \leq t \tag{CQI}$$

within accuracy ϵ (or, which is the same, is an ϵ-*approximation* of (CQI)), if

(i) Whenever (y, t) satisfies (CQI), there exists u such that (y, t, u) satisfies (LP);

(ii) Whenever (y, t, u) satisfies (LP), (y, t) "nearly satisfies" (CQI), namely,

$$\|y\|_2 \leq (1 + \epsilon)t. \tag{CQI$_\epsilon$}$$

Note that given a conic quadratic program

$$\min_x \left\{c^T x : \|A_i x - b_i\|_2 \leq c_i^T x - d_i,\ i = 1, ..., m\right\} \tag{CQP}$$

[6)] That is, decision of instant t depends solely on the demands at instants $\tau < t$

with $m_i \times n$-matrices A_i and ϵ-approximations

$$P^i y_i + t_i p^i + Q^i u_i \geq 0$$

of conic quadratic inequalities

$$\|y_i\|_2 \leq t_i \qquad [\dim y_i = m_i],$$

one can approximate (CQP) by the Linear Programming program

$$\min_{x,u} \left\{ c^T x : P^i(A_i x - b_i) + (c_i^T x - d_i)p^i + Q^i u_i \geq 0,\ i = 1, ..., m \right\};$$

if ϵ is small enough, this program, for every practical purpose, is "the same" as (CQP) [7].

Now, *in principle*, any closed cone of the form

$$\{(y, t) : t \geq \phi(y)\}$$

can be approximated, in the aforementioned sense, by a system of linear inequalities within any accuracy $\epsilon > 0$. The question of crucial importance, however, is how large should be the approximating system – how many linear constraints and additional variables it requires. With naive approach to approximating $\mathbf{L}^{n+1}$ – "take tangent hyperplanes along a fine finite grid of boundary directions and replace the Lorentz cone with the resulting polyhedral one" – the number of linear constraints in, say, 0.5-approximation blows up exponentially as n grows, rapidly making the approximation completely meaningless. Surprisingly, there is a much smarter way to approximate $\mathbf{L}^{n+1}$:

Theorem 2.5.1 *Let n be the dimension of y in* (CQI)*, and let $0 < \epsilon < 1/2$. There exists (and can be explicitly written) a system of no more than $O(1)n \ln \frac{1}{\epsilon}$ linear inequalities of the form* (LP) *with $\dim u \leq O(1)n \ln \frac{1}{\epsilon}$ which is an ϵ-approximation of* (CQI)*. Here $O(1)$'s are appropriate absolute constants.*

To get an impression of the constant factors in the Theorem, look at the numbers $I(n, \epsilon)$ of linear inequalities and $V(n, \epsilon)$ of additional variables u in an ϵ-approximation (LP) of the conic quadratic inequality (CQI) with $\dim y = n$:

n	$\epsilon = 10^{-1}$		$\epsilon = 10^{-6}$		$\epsilon = 10^{-14}$	
	$I(n,\epsilon)$	$V(N,\epsilon)$	$I(n,\epsilon)$	$V(n,\epsilon)$	$I(n,\epsilon)$	$V(n,\epsilon)$
4	6	17	31	69	70	148
16	30	83	159	345	361	745
64	133	363	677	1458	1520	3153
256	543	1486	2711	5916	6169	12710
1024	2203	6006	10899	23758	24773	51050

You can see that $I(n, \epsilon) \approx 0.7n \ln \frac{1}{\epsilon}$, $V(n, \epsilon) \approx 2n \ln \frac{1}{\epsilon}$.

The smart approximation described in Theorem 2.5.1 is incomparably better than the outlined naive approximation. On a closest inspection, the "power" of the smart approximation comes from the fact that here we approximate the Lorentz cone by a *projection of a simple higher-dimensional polyhedral cone*. When projecting a polyhedral cone living in $\mathbf{R}^N$ onto a linear subspace of dimension $<< N$, you get a polyhedral cone with the number of facets which can be by an exponential in N factor larger than the number of facets of the original cone. Thus, *the projection of a simple* (with small number of facets) *polyhedral cone onto a subspace of smaller dimension can be a very complicated* (with an astronomical number of facets) *polyhedral cone*, and this is the fact exploited in the approximation scheme to follow.

Proof of Theorem 2.5.1

Let $\epsilon > 0$ and a positive integer n be given. We intend to build a polyhedral ϵ-approximation of the Lorentz cone $\mathbf{L}^{n+1}$. Without loss of generality we may assume that n is an integer power of 2: $n = 2^\kappa$, $\kappa \in \mathbf{N}$.

[7] Note that standard computers do not distinguish between 1 and 1 ± 10^{-17}. It follows that "numerically speaking", with $\epsilon \sim 10^{-17}$, (CQI) is the same as (CQI_ϵ).

1^0. "Tower of variables". The first step of our construction is quite straightforward: we introduce extra variables to represent a conic quadratic constraint

$$\sqrt{y_1^2 + ... + y_n^2} \leq t \tag{CQI}$$

of dimension $n+1$ by a system of conic quadratic constraints of dimension 3 each. Namely, let us call our original y-variables "variables of generation 0" and let us split them into pairs $(y_1, y_2), ..., (y_{n-1}, y_n)$. We associate with every one of these pairs its "successor" – an additional variable " of generation 1". We split the resulting $2^{\kappa-1}$ variables of generation 1 into pairs and associate with every pair its successor – an additional variable of "generation 2", and so on; after $\kappa - 1$ steps we end up with two variables of the generation $\kappa - 1$. Finally, the only variable of generation κ is the variable t from (CQI).

To introduce convenient notation, let us denote by y_i^ℓ i-th variable of generation ℓ, so that $y_1^0, ..., y_n^0$ are our original y-variables $y_1, ..., y_n$, $y_1^\kappa \equiv t$ is the original t-variable, and the "parents" of y_i^ℓ are the variables $y_{2i-1}^{\ell-1}, y_{2i}^{\ell-1}$.

Note that the total number of all variables in the "tower of variables" we end up with is $2n-1$.

It is clear that the system of constraints

$$\sqrt{[y_{2i-1}^{\ell-1}]^2 + [y_{2i}^{\ell-1}]^2} \leq y_i^\ell, \ i = 1, ..., 2^{\kappa-\ell}, \ \ell = 1, ..., \kappa \tag{2.5.1}$$

is a representation of (CQI) in the sense that a collection $(y_1^0 \equiv y_1, ..., y_n^0 \equiv y_n, y_1^\kappa \equiv t)$ can be extended to a solution of (2.5.1) if and only if (y, t) solves (CQI). Moreover, let $\Pi_\ell(x_1, x_2, x_3, u^\ell)$ be polyhedral ϵ_ℓ-approximations of the cone

$$\mathbf{L}^3 = \{(x_1, x_2, x_3) : \sqrt{x_1^2 + x_2^2} \leq x_3\},$$

$\ell = 1, ..., \kappa$. Consider the system of linear constraints in variables y_i^ℓ, u_i^ℓ:

$$\Pi_\ell(y_{2i-1}^{\ell-1}, y_{2i}^{\ell-1}, y_i^\ell, u_i^\ell) \geq 0, \ i = 1, ..., 2^{\kappa-\ell}, \ \ell = 1, ..., \kappa. \tag{2.5.2}$$

Writing down this system of linear constraints as $\Pi(y, t, u) \geq 0$, where Π is linear in its arguments, $y = (y_1^0, ..., y_n^0)$, $t = y_1^\kappa$, and u is the collection of all u_i^ℓ, $\ell = 1, ..., \kappa$ and all y_i^ℓ, $\ell = 1, ..., \kappa - 1$, we immediately conclude that Π is a polyhedral ϵ-approximation of $\mathbf{L}^{n+1}$ with

$$1 + \epsilon = \prod_{\ell=1}^{\kappa} (1 + \epsilon_\ell). \tag{2.5.3}$$

In view of this observation, we may focus on building polyhedral approximations of the Lorentz cone $\mathbf{L}^3$.

2^0. Polyhedral approximation of $\mathbf{L}^3$ we intend to use is given by the system of linear inequalities as follows (positive integer ν is the parameter of the construction):

$$\begin{array}{ll}
(a) & \left\{ \begin{array}{rcl} \xi^0 & \geq & |x_1| \\ \eta^0 & \geq & |x_2| \end{array} \right. \\
(b) & \left\{ \begin{array}{rcl} \xi^j & = & \cos\left(\frac{\pi}{2^{j+1}}\right) \xi^{j-1} + \sin\left(\frac{\pi}{2^{j+1}}\right) \eta^{j-1} \\ \eta^j & \geq & \left| -\sin\left(\frac{\pi}{2^{j+1}}\right) \xi^{j-1} + \cos\left(\frac{\pi}{2^{j+1}}\right) \eta^{j-1} \right| \end{array} \right. , \quad j = 1, ..., \nu \\
(c) & \left\{ \begin{array}{rcl} \xi^\nu & \leq & x_3 \\ \eta^\nu & \leq & \mathrm{tg}\left(\frac{\pi}{2^{\nu+1}}\right) \xi^\nu \end{array} \right.
\end{array} \tag{2.5.4}$$

Note that (2.5.4) can be straightforwardly written down as a system of linear homogeneous inequalities $\Pi^{(\nu)}(x_1, x_2, x_3, u) \geq 0$, where u is the collection of $2(\nu+1)$ variables ξ^j, η^i, $j = 0, ..., \nu$.

Proposition 2.5.1 *$\Pi^{(\nu)}$ is a polyhedral $\delta(\nu)$-approximation of $\mathbf{L}^3 = \{(x_1, x_2, x_3) : \sqrt{x_1^2 + x_2^2} \leq x_3\}$ with*

$$\delta(\nu) = \frac{1}{\cos\left(\frac{\pi}{2^{\nu+1}}\right)} - 1. \tag{2.5.5}$$

Proof. We should prove that

(i) If $(x_1, x_2, x_3) \in \mathbf{L}^3$, then the triple (x_1, x_2, x_3) can be extended to a solution to (2.5.4);

(ii) If a triple (x_1, x_2, x_3) can be extended to a solution to (2.5.4), then $\|(x_1, x_2)\|_2 \leq (1 + \delta(\nu))x_3$.

(i): Given $(x_1, x_2, x_3) \in \mathbf{L}^3$, let us set $\xi^0 = |x_1|, \eta^0 = |x_2|$, thus ensuring (2.5.4.$a$). Note that $\|(\xi^0, \eta^0)\|_2 = \|(x_1, x_2)\|_2$ and that the point $P^0 = (\xi^0, \eta^0)$ belongs to the first quadrant.

Now, for $j = 1, ..., \nu$ let us set

$$\begin{array}{rcl} \xi^j & = & \cos\left(\frac{\pi}{2^{j+1}}\right)\xi^{j-1} + \sin\left(\frac{\pi}{2^{j+1}}\right)\eta^{j-1} \\ \eta^j & = & \left|-\sin\left(\frac{\pi}{2^{j+1}}\right)\xi^{j-1} + \cos\left(\frac{\pi}{2^{j+1}}\right)\eta^{j-1}\right| \end{array},$$

thus ensuring (2.5.4.b), and let $P^j = (\xi^j, \eta^j)$. The point P^i is obtained from P^{j-1} by the following construction: we rotate clockwise P^{j-1} by the angle $\phi_j = \frac{\pi}{2^{j+1}}$, thus getting a point Q^{j-1}; if this point is in the upper half-plane, we set $P^j = Q^{j-1}$, otherwise P^j is the reflection of Q^{j-1} with respect to the x-axis. From this description it is clear that

(I) $\|P^j\|_2 = \|P^{j-1}\|_2$, so that all vectors P^j are of the same Euclidean norm as P^0, i.e., of the norm $\|(x_1, x_2)\|_2$;

(II) Since the point P^0 is in the first quadrant, the point Q^0 is in the angle $-\frac{\pi}{4} \leq \arg(P) \leq \frac{\pi}{4}$, so that P^1 is in the angle $0 \leq \arg(P) \leq \frac{\pi}{4}$. The latter relation, in turn, implies that Q^1 is in the angle $-\frac{\pi}{8} \leq \arg(P) \leq \frac{\pi}{8}$, whence P^2 is in the angle $0 \leq \arg(P) \leq \frac{\pi}{8}$. Similarly, P^3 is in the angle $0 \leq \arg(P) \leq \frac{\pi}{16}$, and so on: P^j is in the angle $0 \leq \arg(P) \leq \frac{\pi}{2^{j+1}}$.

By (I), $\xi^\nu \leq \|P^\nu\|_2 = \|(x_1, x_2)\|_2 \leq x_3$, so that the first inequality in (2.5.4.c) is satisfied. By (II), P^ν is in the angle $0 \leq \arg(P) \leq \frac{\pi}{2^{\nu+1}}$, so that the second inequality in (2.5.4.c) also is satisfied. We have extended a point from $\mathbf{L}^3$ to a solution to (2.5.4).

(ii): Let (x_1, x_2, x_3) can be extended to a solution $(x_1, x_2, x_3, \{\xi^j, \eta^j\}_{j=0}^\nu)$ to (2.5.4). Let us set $P^j = (\xi^j, \eta^j)$. From (2.5.4.a, b) it follows that all vectors P^j are nonnegative. We have $\| P^0 \|_2 \geq \| (x_1, x_2) \|_2$ by (2.5.4.a). Now, (2.5.4.b) says that the coordinates of P^j are $\geq$ absolute values of the coordinates of P^{j-1} taken in certain orthonormal system of coordinates, so that $\|P^j\|_2 \geq \|P^{j-1}\|_2$. Thus, $\|P^\nu\|_2 \geq \|(x_1, x_2)^T\|_2$. On the other hand, by (2.5.4.c) one has $\|P^\nu\|_2 \leq \frac{1}{\cos(\frac{\pi}{2^{\nu+1}})}\xi^\nu \leq \frac{1}{\cos(\frac{\pi}{2^{\nu+1}})}x_3$, so that $\|(x_1, x_2)^T\|_2 \leq \delta(\nu)x_3$, as claimed. ∎

Specifying in (2.5.2) the mappings $\Pi_\ell(\cdot)$ as $\Pi^{(\nu_\ell)}(\cdot)$, we conclude that for every collection of positive integers $\nu_1, ..., \nu_\kappa$ one can point out a polyhedral β-approximation $\Pi_{\nu_1,...,\nu_\kappa}(y, t, u)$ of $\mathbf{L}^n$, $n = 2^\kappa$:

$$\begin{array}{ll} (a_{\ell,i}) & \left\{\begin{array}{rcl} \xi_{\ell,i}^0 & \geq & |y_{2i-1}^{\ell-1}| \\ \eta_{\ell,i}^0 & \geq & |y_{2i}^{\ell-1}| \end{array}\right. \\ (b_{\ell,i}) & \left\{\begin{array}{rcl} \xi_{\ell,i}^j & = & \cos\left(\frac{\pi}{2^{j+1}}\right)\xi_{\ell,i}^{j-1} + \sin\left(\frac{\pi}{2^{j+1}}\right)\eta_{\ell,i}^{j-1} \\ \eta_{\ell,i}^j & \geq & \left|-\sin\left(\frac{\pi}{2^{j+1}}\right)\xi_{\ell,i}^{j-1} + \cos\left(\frac{\pi}{2^{j+1}}\right)\eta_{\ell,i}^{j-1}\right| \end{array}\right., \quad j = 1, ..., \nu_\ell \\ (c_{\ell,i}) & \left\{\begin{array}{rcl} \xi_{\ell,i}^{\nu_\ell} & \leq & y_i^\ell \\ \eta_{\ell,i}^{\nu_\ell} & \leq & \mathrm{tg}\left(\frac{\pi}{2^{\nu_\ell+1}}\right)\xi_{\ell,i}^{\nu_\ell} \end{array}\right. \\ & \qquad i = 1, ..., 2^{\kappa-\ell},\ \ell = 1, ..., \kappa. \end{array} \tag{2.5.6}$$

The approximation possesses the following properties:

1. The dimension of the u-vector (comprised of all variables in (2.5.6) except $y_i = y_i^0$ and $t = y_1^\kappa$) is

$$p(n, \nu_1, ..., \nu_\kappa) \leq n + O(1)\sum_{\ell=1}^{\kappa} 2^{\kappa-\ell}\nu_\ell;$$

2. The image dimension of $\Pi_{\nu_1,...,\nu_\kappa}(\cdot)$ (i.e., the # of linear inequalities plus twice the # of linear equations in (2.5.6)) is

$$q(n, \nu_1, ..., \nu_\kappa) \leq O(1)\sum_{\ell=1}^{\kappa} 2^{\kappa-\ell}\nu_\ell;$$

3. The quality β of the approximation is

$$\beta = \beta(n;\nu_1,...,\nu_\kappa) = \prod_{\ell=1}^{\kappa} \frac{1}{\cos\left(\frac{\pi}{2^{\nu_\ell+1}}\right)} - 1.$$

3^0. Back to the general case. Given $\epsilon \in (0,1]$ and setting

$$\nu_\ell = \lfloor O(1)\ell \ln \frac{2}{\epsilon} \rfloor,\ \ell = 1,...,\kappa,$$

with properly chosen absolute constant $O(1)$, we ensure that

$$\begin{array}{rcl} \beta(\nu_1,...,\nu_\kappa) & \leq & \epsilon, \\ p(n,\nu_1,...,\nu_\kappa) & \leq & O(1)n\ln\frac{2}{\epsilon}, \\ q(n,\nu_1,...,\nu_\kappa) & \leq & O(1)n\ln\frac{2}{\epsilon}, \end{array}$$

as required. ■

2.6 Exercises

2.6.1 Around randomly perturbed linear constraints

Consider a linear constraint

$$a^T x \geq b \quad [x \in \mathbf{R}^n]. \tag{2.6.1}$$

We have seen that if the coefficients a_j of the left hand side are subject to random perturbations:

$$a_j = a_j^* + \epsilon_j, \tag{2.6.2}$$

where ϵ_j are independent random variables with zero means taking values in segments $[-\sigma_j, \sigma_j]$, then "a reliable version" of the constraint is

$$\sum_j a_j^* x_j - \underbrace{\omega\sqrt{\sum_j \sigma_j^2 x_j^2}}_{\alpha(x)} \geq b, \tag{2.6.3}$$

where $\omega > 0$ is a "safety parameter". "Reliability" means that if certain x satisfies (2.6.3), then x is "$\exp\{-\omega^2/4\}$-reliable solution to (2.6.1)", that is, the probability that x fails to satisfy a realization of the randomly perturbed constraint (2.6.1) does not exceed $\exp\{-\omega^2/4\}$ (see Proposition 2.4.1). Of course, there exists a possibility to build an "absolutely safe" version of (2.6.1) – (2.6.2) (an analogy of the Robust Counterpart), that is, to require that $\min\limits_{|\epsilon_j|\leq\sigma_j} \sum\limits_j (a_j^* + \epsilon_j)x_j \geq b$, which is exactly the inequality

$$\sum_j a_j^* x_j - \underbrace{\sum_j \sigma_j |x_j|}_{\beta(x)} \geq b. \tag{2.6.4}$$

Whenever x satisfies (2.6.4), x satisfies *all* realizations of (2.6.1), and not "all, up to exceptions of small probability". Since (2.6.4) ensures more guarantees than (2.6.3), it is natural to expect from the latter inequality to be "less conservative" than the former one, that is, to expect that the solution set of (2.6.3) is larger than the solution set of (2.6.4). Whether this indeed is the case? The answer depends on the value of the safety parameter ω: when $\omega \leq 1$, the "safety term" $\alpha(x)$ in (2.6.3) is, for every x, not greater than the safety term $\beta(x)$ in (2.6.4), so that every solution to (2.6.4) satisfies (2.6.3). When $\sqrt{n} > \omega > 1$, the "safety terms" in our inequalities become "non-comparable": depending on x, it may happen that $\alpha(x) \leq \beta(x)$ (which is typical when $\omega << \sqrt{n}$), same as it may happen that $\alpha(x) > \beta(x)$. Thus, in the

range $1 < \omega < \sqrt{n}$ no one of inequalities (2.6.3), (2.6.4) is more conservative than the other one. Finally, when $\omega \geq \sqrt{n}$, we always have $\alpha(x) \geq \beta(x)$ (why?), so that for "large" values of ω (2.6.3) is even more conservative than (2.6.4). The bottom line is that (2.6.3) *is not completely satisfactory candidate to the role of "reliable version" of linear constraint* (2.6.1) *affected by random perturbations* (2.6.2)*: depending on the safety parameter, this candidate not necessarily is less conservative than the "absolutely reliable" version* (2.6.4).

The goal of the subsequent exercises is to build and to investigate an improved version of (2.6.3).

Exercise 2.1 *1) Given x, assume that there exist u, v such that*

$$\begin{array}{ll} (a) & x = u + v \\ (b) & \sum\limits_j a_j^* x_j - \sum\limits_j \sigma_j |u_j| - \omega \sqrt{\sum\limits_j \sigma_j^2 v_j^2} \geq b \end{array} \tag{2.6.5}$$

Prove that then the probability for x to violate a realization of (2.6.1) *is* $\leq \exp\{-\omega^2/4\}$ *(and is* $\leq \exp\{-\omega^2/2\}$ *in the case of symmetrically distributed ϵ_j).*

2) Verify that the requirement "x can be extended, by properly chosen u, v, to a solution of (2.6.5)*" is weaker than every one of the requirements*

(a) x satisfies (2.6.3)

(b) x satisfies (2.6.4)

The conclusion of Exercise 2.1 is:

> *A good "reliable version" of randomly perturbed constraint* (2.6.1) – (2.6.2) *is system* (2.6.5) *of linear and conic quadratic constraints in variables x, u, v:*
>
> • *whenever x can be extended to a solution of system* (2.6.5)*, x is* $\exp\{-\omega^2/4\}$*-reliable solution to* (2.6.1) (when the perturbations are symmetrically distributed, you can replace $\exp\{-\omega^2/4\}$ with $\exp\{-\omega^2/2\}$);
>
> • *at the same time, "as far as x is concerned", system* (2.6.5) *is less conservative than every one of the inequalities* (2.6.3), (2.6.4)*: if x solves one of these inequalities, x can be extended to a feasible solution of the system.*

Recall that both (2.6.3) and (2.6.4) are Robust Counterparts

$$\min_{a \in \mathcal{U}} a^T x \geq b \tag{2.6.6}$$

of (2.6.1) corresponding to certain choices of the uncertainty set $\mathcal{U}$: (2.6.3) corresponds to the ellipsoidal uncertainty set

$$\mathcal{U} = \{a : a_j = a_j^* + \sigma_j \zeta_j, \sum_j \zeta_j^2 \leq \omega^2\},$$

while (2.6.3) corresponds to the box uncertainty set

$$\mathcal{U} = \{a : a_j = a_j^* + \sigma_j \zeta_j, \max_j |\zeta_j| \leq 1\}.$$

What about (2.6.5)? Here is the answer:

> (!) *System* (2.6.5) *is (equivalent to) the Robust Counterpart* (2.6.6) *of* (2.6.1)*, the uncertainty set being the intersection of the above ellipsoid and box:*
>
> $$\mathcal{U}_* = \{a : a_j = a_j^* + \sigma_j \zeta_j, \sum_j \zeta_j^2 \leq \omega^2, \max_j |\zeta_j| \leq 1\}.$$
>
> *Specifically, x can be extended to a feasible solution of* (2.6.5) *if and only if one has*
>
> $$\min_{a \in \mathcal{U}_*} a^T x \geq b.$$

Exercise 2.2 *Prove* (!) *by demonstrating that*

$$\max_z \left\{ p^T z : \sum_j z_j^2 \le R^2, |z_j| \le 1 \right\} = \min_{u,v} \left\{ \sum_j |u_j| + R\|v\|_2 : u + v = p \right\}.$$

Exercise 2.3 *Extend the above constructions and results to the case of uncertain linear inequality*

$$a^T x \ge b$$

with certain b and the vector of coefficients a randomly perturbed according to the scheme

$$a = a^* + B\epsilon,$$

where B is deterministic, and the entries $\epsilon_1, ..., \epsilon_N$ of ϵ are independent random variables with zero means and such that $|\epsilon_i| \le \sigma_i$ for all i (σ_i are deterministic).

2.6.2 Around Robust Antenna Design

Consider Antenna Design problem as follows:

> *Given locations $p_1, ..., p_k \in \mathbf{R}^3$ of k coherent harmonic oscillators, design antenna array which sends as much energy as possible in a given direction (which w.l.o.g. may be taken as the positive direction of the x-axis).*

Of course, this is informal setting. The goal of subsequent exercises is to build and process the corresponding model.

Background. In what follows, you can take for granted the following facts:

1. The diagram of "standardly invoked" harmonic oscillator placed at a point $p \in \mathbf{R}^3$ is the following function of a 3D *unit* direction δ:
$$D_p(\delta) = \cos\left(\frac{2\pi p^T \delta}{\lambda}\right) + i \sin\left(\frac{2\pi p^T \delta}{\lambda}\right) \qquad [\delta \in \mathbf{R}^3, \delta^T\delta = 1] \tag{2.6.7}$$
where λ is the wavelength, and i is the imaginary unit.
2. The diagram of an array of oscillators placed at points $p_1, ..., p_k$, is the function
$$D(\delta) = \sum_{\ell=1}^k z_\ell D_{p_\ell}(\delta),$$
where z_ℓ are the "element weights" (which form the antenna design and can be arbitrary complex numbers).
3. A natural for engineers way to measure the "concentration" of the energy sent by antenna around a given direction e (which from now on is the positive direction of the x-axis) is
   - to choose a $\theta > 0$ and to define the corresponding *sidelobe angle* Δ_θ as the set of all unit 3D directions δ which are at the angle $\ge \theta$ with the direction e;
   - to measure the "energy concentration" by the *index* $\rho = \frac{|D(e)|}{\max\limits_{\delta \in \Delta_\theta} |D(\delta)|}$, where $D(\cdot)$ is the diagram of the antenna.
4. To make the index easily computable, let us replace in its definition the maximum over the entire sidelobe angle with the maximum over a given "fine finite grid" $\Gamma \subset \Delta_\theta$, thus arriving at the quantity
$$\rho = \frac{|D(e)|}{\max\limits_{\delta \in \Gamma_\theta} |D(\delta)|}$$
which we from now on call the *concentration index*.

Developments. Now we can formulate the Antenna Design problem as follows:

(*) *Given*

- *locations* $p_1, ..., p_k$ *of harmonic oscillators,*
- *wavelength* λ*,*
- *finite set* Γ *of unit 3D directions,*

choose complex weights $z_\ell = x_\ell + iy_\ell$, $\ell = 1, ..., k$ *which maximize the index*

$$\rho = \frac{|\sum_\ell z_\ell D_\ell(e)|}{\max_{\delta\in\Gamma} |\sum_\ell z_\ell D_\ell(\delta)|} \tag{2.6.8}$$

where $D_\ell(\cdot)$ *are given by* (2.6.7).

Exercise 2.4 *1) Whether the objective* (2.6.8) *is a concave (and thus "easy to maximize") function?*
2) Prove that (*) *is equivalent to the convex optimization program*

$$\max_{x_\ell, y_\ell \in \mathbf{R}} \left\{ \Re\left(\sum_\ell (x_\ell + iy_\ell) D_\ell(e) \right) : |\sum_\ell (x_\ell + iy_\ell) D_\ell(\delta)| \le 1,\ \delta \in \Gamma \right\}. \tag{2.6.9}$$

In order to carry out our remaining tasks, it makes sense to approximate (2.6.9) with a Linear Programming problem. To this end, it suffices to approximate the modulus of a complex number z (i.e., the Euclidean norm of a 2D vector) by the quantity

$$\pi_J(z) = \max_{j=1,...,J} \Re(\omega_j z) \qquad [\omega_j = \cos(\tfrac{2\pi j}{J}) + i \sin(\tfrac{2\pi j}{J})]$$

(geometrically: we approximate the unit disk in $\mathbf{C} = \mathbf{R}^2$ by circumscribed perfect J-side polygon).

Exercise 2.5 *What is larger –* $\pi_J(z)$ *or* $|z|$*? Within which accuracy the "polyhedral norm"* $\pi_J(\cdot)$ *approximates the modulus?*

With the outlined approximation of the modulus, (2.6.9) becomes the optimization program

$$\max_{x_\ell, y_\ell \in \mathbf{R}} \left\{ \Re\left(\sum_\ell (x_\ell + iy_\ell) D_\ell(e) \right) : \Re\left(\omega_j \sum_\ell (x_\ell + iy_\ell) D_\ell(\delta) \right) \le 1,\ 1 \le j \le J, \delta \in \Gamma \right\}. \tag{2.6.10}$$

Exercise 2.6 *1) Verify that* (2.6.10) *is a Linear Programming program and solve it numerically for the following two setups:*

Data A:

- $k = 16$ oscillators placed at the points $p_\ell = (\ell - 1)e$, $\ell = 1, ..., 16$;
- wavelength $\lambda = 2.5$;
- $J = 10$;
- sidelobe grid Γ: since with the oscillators located along the x-axis, the diagram of the array is symmetric with respect to rotations around the x-axis, it suffices to look at the "sidelobe directions" from the xy-plane. To get Γ, we form the set of all directions which are at the angle *at least* $\theta = 0.3$rad away from the positive direction of the x-axis, and take 64-point equidistant grid in the resulting "arc of directions", so that

$$\Gamma = \left\{ \delta_s = \begin{bmatrix} \cos(\alpha + sd\alpha) \\ \sin(\alpha + sd\alpha) \\ 0 \end{bmatrix} \right\}_{s=0}^{63} \qquad [\alpha = 0.3, d\alpha = \tfrac{2(\pi - \alpha)}{63}]$$

Data B: exactly as Data A, except for the wavelength, which is now $\lambda = 5$.

2) Assume that in reality the weights are affected by "implementation errors":

$$x_\ell = x_\ell^*(1 + \sigma\xi_\ell),\ y_\ell = x_\ell^*(1 + \sigma\eta_\ell),$$

where x_ℓ^*, y_ℓ^* are the "nominal optimal weights" obtained when solving (2.6.10)*, x_ℓ, y_ℓ are actual weights, $\sigma > 0$ is the "perturbation level", and ξ_ℓ, η_ℓ are mutually independent random perturbations uniformly distributed in* $[-1, 1]$.

2.1) Check by simulation what happens with the concentration index of the actual diagram as a result of implementation errors. Carry out the simulations for the perturbation level σ taking values 1.e-4, 5.e-4, 1.e-3.

2.2) If you are not satisfied with the behaviour of nominal design(s) in the presence implementation errors, use the Robust Counterpart methodology to replace the nominal designs with the robust ones. What is the "price of robustness" in terms of the index? What do you gain in stability of the diagram w.r.t. implementation errors?

Lecture 3

Semidefinite Programming

In this lecture we study *Semidefinite Programming* – a generic conic program with an extremely wide area of applications.

3.1 Semidefinite cone and Semidefinite programs

3.1.1 Preliminaries

Let $\mathbf{S}^m$ be the space of symmetric $m \times m$ matrices, and $\mathbf{M}^{m,n}$ be the space of rectangular $m \times n$ matrices with real entries. In the sequel, we always think of these spaces as of Euclidean spaces equipped with the *Frobenius* inner product

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

and we may use in connection with these spaces all notions based upon the Euclidean structure, e.g., the (Frobenius) norm of a matrix

$$\|X\|_2 = \sqrt{\langle X, X\rangle} = \sqrt{\sum_{i,j=1}^m X_{ij}^2} = \sqrt{\mathrm{Tr}(X^TX)}$$

and likewise the notions of orthogonality, orthogonal complement of a linear subspace, etc. Of course, the Frobenius inner product of symmetric matrices can be written without the transposition sign:

$$\langle X, Y\rangle = \mathrm{Tr}(XY),\ X, Y \in \mathbf{S}^m.$$

Let us focus on the space $\mathbf{S}^m$. After it is equipped with the Frobenius inner product, we may speak about a cone dual to a given cone $\mathbf{K} \subset \mathbf{S}^m$:

$$\mathbf{K}_* = \{Y \in \mathbf{S}^m \mid \langle Y, X\rangle \geq 0 \quad \forall X \in \mathbf{K}\}.$$

Among the cones in $\mathbf{S}^m$, the one of special interest is the *semidefinite cone* $\mathbf{S}^m_+$, the cone of all symmetric positive semidefinite matrices[1)]. It is easily seen that $\mathbf{S}^m_+$ indeed is a cone, and moreover it is self-dual:

$$(\mathbf{S}^m_+)_* = \mathbf{S}^m_+.$$

Another simple fact is that the interior $\mathbf{S}^m_{++}$ of the semidefinite cone $\mathbf{S}^m_+$ is exactly the set of all *positive definite* symmetric $m \times m$ matrices, i.e., symmetric matrices A for which $x^TAx > 0$ for all nonzero vectors x, or, which is the same, symmetric matrices with positive eigenvalues.

[1)] Recall that a symmetric $n \times n$ matrix A is called positive semidefinite if $x^TAx \geq 0$ for all $x \in \mathbf{R}^m$; an equivalent definition is that all eigenvalues of A are nonnegative

The semidefinite cone gives rise to a family of conic programs "minimize a linear objective over the intersection of the semidefinite cone and an affine plane"; these are the *semidefinite programs* we are about to study.

Before writing down a generic semidefinite program, we should resolve a small difficulty with notation. Normally we use lowercase Latin and Greek letters to denote vectors, and the uppercase letters – to denote matrices; e.g., our usual notation for a conic problem is

$$\min_x \left\{ c^T x : Ax - b \geq_{\mathbf{K}} 0 \right\}. \tag{CP}$$

In the case of semidefinite programs, where $\mathbf{K} = \mathbf{S}^m_+$, the usual notation leads to a conflict with the notation related to the space where $\mathbf{S}^m_+$ lives. Look at (CP): without additional remarks it is unclear what is A – is it a $m \times m$ matrix from the space $\mathbf{S}^m$ or is it a linear mapping acting from the space of the design vectors – some $\mathbf{R}^n$ – to the space $\mathbf{S}^m$? When speaking about a conic problem on the cone $\mathbf{S}^m_+$, we should have in mind the second interpretation of A, while the standard notation in (CP) suggests the first (wrong!) interpretation. In other words, we meet with the necessity to distinguish between *linear mappings* acting to/from $\mathbf{S}^m$ and elements of $\mathbf{S}^m$ (which themselves are linear mappings from $\mathbf{R}^m$ to $\mathbf{R}^m$). In order to resolve the difficulty, we make the following

Notational convention: To denote a linear mapping acting from a linear space to a space of matrices (or from a space of matrices to a linear space), we use uppercase script letters like $\mathcal{A}$, $\mathcal{B}$,... Elements of usual vector spaces $\mathbf{R}^n$ are, as always, denoted by lowercase Latin/Greek letters $a, b, ..., z, \alpha, ..., \zeta$, while elements of a space of matrices usually are denoted by uppercase Latin letters $A, B, ..., Z$. According to this convention, a semidefinite program of the form (CP) should be written as

$$\min_x \left\{ c^T x : \mathcal{A}x - B \geq_{\mathbf{S}^m_+} 0 \right\}. \tag{$*$}$$

We also simplify the sign $\geq_{\mathbf{S}^m_+}$ to $\succeq$ and the sign $>_{\mathbf{S}^m_+}$ to $\succ$ (same as we write $\geq$ instead of $\geq_{\mathbf{R}^m_+}$ and $>$ instead of $>_{\mathbf{R}^m_+}$). Thus, $A \succeq B$ ($\Leftrightarrow B \preceq A$) means that A and B are symmetric matrices of the same size and $A - B$ is positive semidefinite, while $A \succ B$ ($\Leftrightarrow B \prec A$) means that A, B are symmetric matrices of the same size with positive definite $A - B$.

Our last convention is how to write down expressions of the type $A\mathcal{A}xB$ ($\mathcal{A}$ is a linear mapping from some $\mathbf{R}^n$ to $\mathbf{S}^m$, $x \in \mathbf{R}^n$, $A, B \in \mathbf{S}^m$); what we are trying to denote is the result of the following operation: we first take the value $\mathcal{A}x$ of the mapping $\mathcal{A}$ at a vector x, thus getting an $m \times m$ matrix $\mathcal{A}x$, and then multiply this matrix from the left and from the right by the matrices A, B. In order to avoid misunderstandings, we write expressions of this type as

$$A[\mathcal{A}x]B$$

or as $A\mathcal{A}(x)B$, or as $A\mathcal{A}[x]B$.

How to specify a mapping $\mathcal{A} : \mathbf{R}^n \to \mathbf{S}^m$. A natural data specifying a linear mapping $A : \mathbf{R}^n \to \mathbf{R}^m$ is a collection of n elements of the "destination space" – n vectors $a_1, a_2, ..., a_n \in \mathbf{R}^m$ – such that

$$Ax = \sum_{j=1}^n x_j a_j, \quad x = (x_1, ..., x_n)^T \in \mathbf{R}^n.$$

Similarly, a natural data specifying a linear mapping $\mathcal{A} : \mathbf{R}^n \to \mathbf{S}^m$ is a collection $A_1, ..., A_n$ of n matrices from $\mathbf{S}^m$ such that

$$\mathcal{A}x = \sum_{j=1}^n x_j A_j, \quad x = (x_1, ..., x_n)^T \in \mathbf{R}^n. \tag{3.1.1}$$

In terms of these data, the semidefinite program (*) can be written as

$$\min_x \left\{ c^T x : x_1 A_1 + x_2 A_2 + ... + x_n A_n - B \succeq 0 \right\}. \tag{SDPr}$$

It is a simple exercise to verify that if $\mathcal{A}$ is represented as in (3.1.1), then the conjugate to $\mathcal{A}$ linear mapping $\mathcal{A}^* : \mathbf{S}^m \to \mathbf{R}^n$ is given by

$$\mathcal{A}^*\Lambda = (\mathrm{Tr}(\Lambda A_1), ..., \mathrm{Tr}(\Lambda A_n))^T : \mathbf{S}^m \to \mathbf{R}^n. \tag{3.1.2}$$

Linear Matrix Inequality constraints and semidefinite programs. In the case of conic quadratic problems, we started with the simplest program of this type – the one with a single conic quadratic constraint $Ax - b \geq_{\mathbf{L}^m} 0$ – and then defined a conic quadratic program as a program with finitely many constraints of this type, i.e., as a conic program on a *direct product* of the ice-cream cones. In contrast to this, when defining a semidefinite program, we impose on the design vector just one *Linear Matrix Inequality* (LMI) $\mathcal{A}x - B \succeq 0$. Now we indeed should not bother about more than a single LMI, due to the following simple fact:

A system of finitely many LMI's

$$\mathcal{A}_i x - B_i \succeq 0, \ i = 1, ..., k,$$

is equivalent to the single LMI

$$\mathcal{A}x - B \succeq 0,$$

with

$$\mathcal{A}x = \mathrm{Diag}\left(\mathcal{A}_1 x, \mathcal{A}_2 x, ..., \mathcal{A}_k x\right), B = \mathrm{Diag}(B_1, ..., B_k);$$

here for a collection of symmetric matrices $Q_1, ..., Q_k$

$$\mathrm{Diag}(Q_1, ..., Q_k) = \begin{pmatrix} Q_1 & & \\ & \ddots & \\ & & Q_k \end{pmatrix}$$

is the block-diagonal matrix with the diagonal blocks $Q_1, ..., Q_k$.

Indeed, a block-diagonal symmetric matrix is positive (semi)definite if and only if all its diagonal blocks are so.

Dual to a semidefinite program (SDP). Specifying the general concept of conic dual of a conic program in the case when the latter is a semidefinite program (*) and taking into account (3.1.2) along with the fact that the semidefinite cone is self-dual, we see that the dual to (*) is the semidefinite program

$$\max_{\Lambda} \left\{ \langle B, \Lambda \rangle \equiv \mathrm{Tr}(B\Lambda) : \mathrm{Tr}(A_i \Lambda) = c_i, \ i = 1, ..., n; \Lambda \succeq 0 \right\}. \tag{SDDl}$$

Conic Duality in the case of Semidefinite Programming. Let us see what we get from the Conic Duality Theorem in the case of semidefinite programs. First note that our default assumption **A** on a conic program in the form of (CP) (Lecture 1) as applied to (SDPr) says that no nontrivial linear combination of the matrices $A_1, ..., A_n$ is 0. Strict feasibility of (SDPr) means that there exists x such that $\mathcal{A}x - B$ is positive definite, and strict feasibility of (SDDl) means that there exists a positive definite Λ satisfying $\mathcal{A}^*\Lambda = c$. According to the Conic Duality Theorem, if both primal and dual are strictly feasible, both are solvable, the optimal values are equal to each other, and the complementary slackness condition

$$[\mathrm{Tr}(\Lambda[\mathcal{A}x - B]) \equiv] \qquad \langle \Lambda, \mathcal{A}x - B \rangle = 0$$

is necessary and sufficient for a pair of a primal feasible solution x and a dual feasible solution Λ to be optimal for the corresponding problems.

It is easily seen that for a pair X, Y of positive semidefinite symmetric matrices one has

$$\mathrm{Tr}(XY) = 0 \Leftrightarrow XY = YX = 0;$$

in particular, in the case of strictly feasible primal and dual problems, the "primal slack" $S_* = \mathcal{A}x^* - B$ corresponding to a primal optimal solution commutes with (any) dual optimal solution Λ_*, and the product of these two matrices is 0. Besides this, S_* and Λ_*, as a pair of commuting symmetric matrices, share a common eigenbasis, and the fact that $S_*\Lambda_* = 0$ means that the eigenvalues of the matrices in this basis are "complementary": for every common eigenvector, either the eigenvalue of S_*, or the one of Λ_*, or both, are equal to 0 (cf. with complementary slackness in the LP case).

3.2 What can be expressed via LMI's?

As in the previous lecture, the first thing to realize when speaking about the "semidefinite programming universe" is how to recognize that a convex optimization program

$$\min_x \left\{ c^T x : x \in X = \bigcap_{i=1}^m X_i \right\} \tag{P}$$

can be cast as a semidefinite program. Just as in the previous lecture, this question actually asks whether a given convex set/convex function is positive semidefinite representable (in short: SDr). The definition of the latter notion is completely similar to the one of a CQr set/function:

We say that a convex set $X \subset \mathbf{R}^n$ is SDr, if there exists an affine mapping $(x,u) \to \mathcal{A}\begin{pmatrix} x \\ u \end{pmatrix} - B : \mathbf{R}^n_x \times \mathbf{R}^k_u \to \mathbf{S}^m$ such that

$$x \in X \Leftrightarrow \exists u : \mathcal{A}\begin{pmatrix} x \\ u \end{pmatrix} - B \succeq 0;$$

in other words, X is SDr, if there exists LMI

$$\mathcal{A}\begin{pmatrix} x \\ u \end{pmatrix} - B \succeq 0,$$

in the original design vector x and a vector u of additional design variables such that X is a projection of the solution set of the LMI onto the x-space. An LMI with this property is called Semidefinite Representation (SDR) *of the set X.*

A convex function $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ is called SDr, if its epigraph

$$\{(x,t) \mid t \geq f(x)\}$$

is a SDr set. A SDR of the epigraph of f is called semidefinite representation of f.

By exactly the same reasons as in the case of conic quadratic problems, one has:

1. *If f is a SDr function, then all its level sets $\{x \mid f(x) \leq a\}$ are SDr; the SDR of the level sets are explicitly given by (any) SDR of f;*
2. *If all the sets X_i in problem* (P) *are SDr with known SDR's, then the problem can explicitly be converted to a semidefinite program.*

In order to understand which functions/sets are SDr, we may use the same approach as in Lecture 2. "The calculus", i.e., the list of basic operations preserving SD-representability, is exactly the same as in the case of conic quadratic problems; we just may repeat word by word the relevant reasoning from Lecture 2, each time replacing "CQr" with "SDr". Thus, the only issue to be addressed is the derivation of a catalogue of "simple" SDr functions/sets. Our first observation in this direction is as follows:

1-17. [2] *If a function/set is CQr, it is also SDr, and any CQR of the function/set can be explicitly converted to its SDR.*

Indeed, the notion of a CQr/SDr function is a "derivative" of the notion of a CQr/SDr set: by definition, a function is CQr/SDr if and only if its epigraph is so. Now, CQr sets are exactly those sets which can be obtained as projections of the solution sets of systems of conic quadratic inequalities, i.e., as projections of inverse images, under affine mappings, of direct products of ice-cream cones. Similarly, SDr sets are projections of the inverse images, under affine mappings, of positive semidefinite cones. Consequently,

(i) in order to verify that a CQr set is SDr as well, it suffices to show that an inverse image, under an affine mapping, of a direct product of ice-cream cones – a set of the form

$$Z = \{z \mid Az - b \in \mathbf{K} = \prod_{i=1}^{l} \mathbf{L}^{k_i}\}$$

is the inverse image of a semidefinite cone under an affine mapping. To this end, in turn, it suffices to demonstrate that

(ii) a direct product $\mathbf{K} = \prod_{i=1}^{l} \mathbf{L}^{k_i}$ of ice-cream cones is an inverse image of a semidefinite cone under an affine mapping.

Indeed, representing $\mathbf{K}$ as $\{y \mid \mathcal{A}y - b \in \mathbf{S}^m_+\}$, we get

$$Z = \{z \mid Az - b \in \mathbf{K}\} = \{z \mid \hat{\mathcal{A}}z - \hat{B} \in \mathbf{S}^m_+\},$$

where $\hat{\mathcal{A}}z - \hat{B} = \mathcal{A}(Az - b) - B$ is affine.

In turn, in order to prove (ii) it suffices to show that

(iii) Every ice-cream cone $\mathbf{L}^k$ is an inverse image of a semidefinite cone under an affine mapping.

> In fact the implication (iii) $\Rightarrow$ (ii) is given by our calculus, since a direct product of SDr sets is again SDr[3].

We have reached the point where no more reductions are necessary, and here is the demonstration of (iii). To see that the Lorentz cone $\mathbf{L}^k$, $k > 1$, is SDr, it suffices to observe that

$$\begin{pmatrix} x \\ t \end{pmatrix} \in \mathbf{L}^k \Leftrightarrow \mathcal{A}(x,t) = \begin{pmatrix} tI_{k-1} & x \\ x^T & t \end{pmatrix} \succeq 0 \tag{3.2.1}$$

(x is $k-1$-dimensional, t is scalar, I_{k-1} is the $(k-1) \times (k-1)$ unit matrix). (3.2.1) indeed resolves the problem, since the matrix $\mathcal{A}(x,t)$ is linear in (x,t)!

---

[2] We refer to Examples 1-17 of CQ-representable functions/sets from Section 2.3

[3] Just to recall where the calculus comes from, here is a direct verification:

Given a direct product $\mathbf{K} = \prod_{i=1}^{l} \mathbf{L}^{k_i}$ of ice-cream cones and given that every factor in the product is the inverse image of a semidefinite cone under an affine mapping:

$$\mathbf{L}^{k_i} = \{x_i \in \mathbf{R}^{k_i} \mid \mathcal{A}_i x_i - B_i \succeq 0\},$$

we can represent $\mathbf{K}$ as the inverse image of a semidefinite cone under an affine mapping, namely, as

$$\mathbf{K} = \{x = (x_1, ..., x_l) \in \mathbf{R}^{k_1} \times ... \times \mathbf{R}^{k_l} \mid \mathrm{Diag}(\mathcal{A}_1 x_i - B_1, ..., \mathcal{A}_l x_l - B_l) \succeq 0\}.$$

It remains to verify (3.2.1), which is immediate. If $(x,t) \in \mathbf{L}^k$, i.e., if $\|x\|_2 \le t$, then for every $y = \begin{pmatrix} \xi \\ \tau \end{pmatrix} \in \mathbf{R}^k$ (ξ is $(k-1)$-dimensional, τ is scalar) we have

$$\begin{array}{rcl} y^T\mathcal{A}(x,t)y & = & \tau^2 t + 2\tau x^T\xi + t\xi^T\xi \ge \tau^2 t - 2|\tau|\|x\|_2\|\xi\|_2 + t\|\xi\|_2^2 \\ & \ge & t\tau^2 - 2t|\tau|\|\xi\|_2 + t\|\xi\|_2^2 \\ & \ge & t(|\tau| - \|\xi\|_2)^2 \ge 0, \end{array}$$

so that $\mathcal{A}(x,t) \succeq 0$. Vice versa, if $\mathcal{A}(t,x) \succeq 0$, then of course $t \ge 0$. Assuming $t = 0$, we immediately obtain $x = 0$ (since otherwise for $y = \begin{pmatrix} x \\ 0 \end{pmatrix}$ we would have $0 \le y^T\mathcal{A}(x,t)y = -2\|x\|_2^2$); thus, $A(x,t) \succeq 0$ implies $\|x\|_2 \le t$ in the case of $t = 0$. To see that the same implication is valid for $t > 0$, let us set $y = \begin{pmatrix} -x \\ t \end{pmatrix}$ to get

$$0 \le y^T\mathcal{A}(x,t)y = tx^Tx - 2tx^Tx + t^3 = t(t^2 - x^Tx),$$

whence $\|x\|_2 \le t$, as claimed. ∎

We see that the "expressive abilities" of semidefinite programming are even richer than those of Conic Quadratic programming. In fact the gap is quite significant. The first new possibility is the ability to handle eigenvalues, and the importance of this possibility can hardly be overestimated.

SD-representability of functions of eigenvalues of symmetric matrices. Our first eigenvalue-related observation is as follows:

18. The largest eigenvalue $\lambda_{\max}(X)$ regarded as a function of $m \times m$ symmetric matrix X is SDr. Indeed, the epigraph of this function

$$\{(X,t) \in \mathbf{S}^m \times \mathbf{R} \mid \lambda_{\max}(X) \le t\}$$

is given by the LMI

$$tI_m - X \succeq 0,$$

where I_m is the unit $m \times m$ matrix.

Indeed, the eigenvalues of $tI_m - X$ are t minus the eigenvalues of X, so that the matrix $tI_m - X$ is positive semidefinite – all its eigenvalues are nonnegative – if and only if t majorates all eigenvalues of X.

The latter example admits a natural generalization. Let M, A be two symmetric $m \times m$ matrices, and let M be positive definite. A real λ and a nonzero vector e are called eigenvalue and eigenvector of the *pencil* $[M, A]$, if $Ae = \lambda Me$ (in particular, the usual eigenvalues/eigenvectors of A are exactly the eigenvalues/eigenvectors of the pencil $[I_m, A]$). Clearly, λ is an eigenvalue of $[M, A]$ if and only if the matrix $\lambda M - A$ is singular, and nonzero vectors from the kernel of the latter matrix are exactly the eigenvectors of $[M, A]$ associated with the eigenvalue λ. The eigenvalues of the pencil $[M, A]$ are the usual eigenvalues of the matrix $M^{-1/2}AM^{-1/2}$, as can be concluded from:

$$\mathrm{Det}(\lambda M - A) = 0 \Leftrightarrow \mathrm{Det}(M^{1/2}(\lambda I_m - M^{-1/2}AM^{-1/2})M^{1/2}) = 0 \Leftrightarrow \mathrm{Det}(\lambda I_m - M^{-1/2}AM^{-1/2}) = 0.$$

The announced extension of Example 18 is as follows:

18a. [The maximum eigenvalue of a pencil]: Let M be a positive definite symmetric $m \times m$ matrix, and let $\lambda_{\max}(X : M)$ be the largest eigenvalue of the pencil $[M, X]$, where X is a symmetric $m \times m$ matrix. The inequality

$$\lambda_{\max}(X : M) \le t$$

is equivalent to the matrix inequality

$$tM - X \succeq 0.$$

In particular, $\lambda_{\max}(X : M)$, regarded as a function of X, is SDr.

18b. The spectral norm $|X|$ of a symmetric $m \times m$ matrix X, i.e., the maximum of absolute values of the eigenvalues of X, is SDr. Indeed, a SDR of the epigraph

$$\{(X,t) \mid |X| \le t\} = \{(X,t) \mid \lambda_{\max}(X) \le t, \lambda_{\max}(-X) \le t\}$$

of $|X|$ is given by the pair of LMI's

$$tI_m - X \succeq 0,\ tI_m + X \succeq 0.$$

In spite of their simplicity, the indicated results are extremely useful. As a more complicated example, let us build a SDr for the sum of the k *largest* eigenvalues of a symmetric matrix.

From now on, speaking about $m \times m$ symmetric matrix X, we denote by $\lambda_i(X)$, $i = 1, ..., m$, its eigenvalues *counted with their multiplicities* and *arranged in a non-ascending order:*

$$\lambda_1(X) \ge \lambda_2(X) \ge ... \ge \lambda_m(X).$$

The vector of the eigenvalues (in the indicated order) will be denoted $\lambda(X)$:

$$\lambda(X) = (\lambda_1(X), ..., \lambda_m(X))^T \in \mathbf{R}^m.$$

The question we are about to address is which functions of the eigenvalues are SDr. We already know that this is the case for the largest eigenvalue $\lambda_1(X)$. Other eigenvalues cannot be SDr since they are not convex functions of X. And convexity, of course, is a necessary condition for SD-representability (cf. Lecture 2). It turns out, however, that the m functions

$$S_k(X) = \sum_{i=1}^{k} \lambda_i(X),\ k = 1, ..., m,$$

are convex and, moreover, are SDr:

18c. Sums of largest eigenvalues of a symmetric matrix. Let X be $m \times m$ symmetric matrix, and let $k \le m$. Then the function $S_k(X)$ is SDr. Specifically, the epigraph

$$\{(X,t) \mid S_k(x) \le t\}$$

of the function admits the SDR

$$\begin{array}{lrcl} (a) & t - ks - \mathrm{Tr}(Z) & \ge & 0 \\ (b) & Z & \succeq & 0 \\ (c) & Z - X + sI_m & \succeq & 0 \end{array} \qquad (3.2.2)$$

where $Z \in \mathbf{S}^m$ and $s \in \mathbf{R}$ are additional variables.

We should prove that

(i) If a given pair X, t can be extended, by properly chosen s, Z, to a solution of the system of LMI's (3.2.2), then $S_k(X) \le t$;

(ii) Vice versa, if $S_k(X) \le t$, then the pair X, t can be extended, by properly chosen s, Z, to a solution of (3.2.2).

To prove (i), we use the following basic fact[4)]:

[4)] which is n immediate corollary of the fundamental *Variational Characterization of Eigenvalues* of symmetric matrices, see Section A.7.3: for a symmetric $m \times m$ matrix A,

$$\lambda_i(A) = \min_{E \in \mathcal{E}_i} \max_{e \in E: e^T e = 1} e^T A e,$$

where $\mathcal{E}_i$ is the collection of all linear subspaces of the dimension $n - i + 1$ in $\mathbf{R}^m$,

(W) *The vector $\lambda(X)$ is a $\succeq$-monotone function of $X \in \mathbf{S}^m$:*

$$X \succeq X' \Rightarrow \lambda(X) \geq \lambda(X').$$

Assuming that (X, t, s, Z) is a solution to (3.2.2), we get $X \preceq Z + sI_m$, so that

$$\lambda(X) \leq \lambda(Z + sI_m) = \lambda(Z) + s \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix},$$

whence

$$S_k(X) \leq S_k(Z) + sk.$$

Since $Z \succeq 0$ (see (3.2.2.b)), we have $S_k(Z) \leq \mathrm{Tr}(Z)$, and combining these inequalities we get

$$S_k(X) \leq \mathrm{Tr}(Z) + sk.$$

The latter inequality, in view of (3.2.2.a)), implies $S_k(X) \leq t$, and (i) is proved.

To prove (ii), assume that we are given X, t with $S_k(X) \leq t$, and let us set $s = \lambda_k(X)$. Then the k largest eigenvalues of the matrix $X - sI_m$ are nonnegative, and the remaining are nonpositive. Let Z be a symmetric matrix with the same eigenbasis as X and such that the k largest eigenvalues of Z are the same as those of $X - sI_m$, and the remaining eigenvalues are zeros. The matrices Z and $Z - X + sI_m$ are clearly positive semidefinite (the first by construction, and the second since in the eigenbasis of X this matrix is diagonal with the first k diagonal entries being 0 and the remaining being the same as those of the matrix $sI_m - X$, i.e., nonnegative). Thus, the matrix Z and the real s we have built satisfy (3.2.2.b, c). In order to see that (3.2.2.a) is satisfied as well, note that by construction $\mathrm{Tr}(Z) = S_k(x) - sk$, whence $t - sk - \mathrm{Tr}(Z) = t - S_k(x) \geq 0$.

In order to proceed, we need the following highly useful technical result:

Lemma 3.2.1 [Lemma on the Schur Complement] *Let*

$$A = \begin{pmatrix} B & C^T \\ C & D \end{pmatrix}$$

be a symmetric matrix with $k \times k$ block B and $\ell \times \ell$ block D. Assume that B is positive definite. Then A is positive (semi)definite if and only if the matrix

$$D - CB^{-1}C^T$$

is positive (semi)definite (this matrix is called the Schur complement of B in A).

Proof. The positive semidefiniteness of A is equivalent to the fact that

$$0 \leq (x^T, y^T) \begin{pmatrix} B & C^T \\ C & D \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = x^T Bx + 2x^T C^T y + y^T Dy \quad \forall x \in \mathbf{R}^k, y \in \mathbf{R}^\ell,$$

or, which is the same, to the fact that

$$\inf_{x \in \mathbf{R}^k} \left[x^T Bx + 2x^T C^T y + y^T Dy \right] \geq 0 \quad \forall y \in \mathbf{R}^\ell.$$

Since B is positive definite by assumption, the infimum in x can be computed explicitly for every fixed y: the optimal x is $-B^{-1}C^T y$, and the optimal value is

$$y^T Dy - y^T CB^{-1}C^T y = y^T [D - CB^{-1}C^T] y.$$

The positive definiteness/semidefiniteness of A is equivalent to the fact that the latter expression is, respectively, positive/nonnegative for every $y \neq 0$, i.e., to the positive definiteness/semidefiniteness of the Schur complement of B in A. ∎

18d. "Determinant" of a symmetric positive semidefinite matrix. Let X be a symmetric positive semidefinite $m \times m$ matrix. Although its determinant

$$\mathrm{Det}(X) = \prod_{i=1}^{m} \lambda_i(X)$$

is neither a convex nor a concave function of X (if $m \geq 2$), it turns out that the function $\mathrm{Det}^q(X)$ is concave in X whenever $0 \leq q \leq \frac{1}{m}$. Function of this type are important in many volume-related problems (see below); we are about to prove that

if q is a rational number,, $0 \leq q \leq \frac{1}{m}$, then the function

$$f_q(X) = \begin{cases} -\mathrm{Det}^q(X), & X \succeq 0 \\ +\infty, & \textit{otherwise} \end{cases}$$

is SDr.

Consider the following system of LMI's:

$$\begin{pmatrix} X & \Delta \\ \Delta^T & D(\Delta) \end{pmatrix} \succeq 0, \tag{D}$$

where Δ is $m \times m$ lower triangular matrix comprised of additional variables, and $D(\Delta)$ is the diagonal matrix with the same diagonal entries as those of Δ. Let $\mathrm{diag}(\Delta)$ denote the vector of the diagonal entries of the square matrix Δ.

As we know from Lecture 2 (see Example 15), the set

$$\{(\delta, t) \in \mathbf{R}^m_+ \times \mathbf{R} \mid \ t \leq (\delta_1 ... \delta_m)^q\}$$

admits an explicit CQR. Consequently, this set admits an explicit SDR as well. The latter SDR is given by certain LMI $S(\delta, t; u) \succeq 0$, where u is the vector of additional variables of the SDR, and $S(\delta, t, u)$ is a matrix affinely depending on the arguments. We claim that

(!) *The system of LMI's* (D) & $S(\mathrm{diag}(\Delta), t; u) \succeq 0$ *is a SDR for the set*

$$\{(X, t) \mid \ X \succeq 0, t \leq \mathrm{Det}^q(X)\},$$

which is basically the epigraph of the function f_q (the latter is obtained from our set by reflection with respect to the plane $t = 0$).

To support our claim, recall that by Linear Algebra a matrix X is positive semidefinite if and only if it can be factorized as $X = \widehat{\Delta}\widehat{\Delta}^T$ with a lower triangular $\widehat{\Delta}$, $\mathrm{diag}(\widehat{\Delta}) \geq 0$; the resulting matrix $\widehat{\Delta}$ is called the Choleski factor of X. No note that if $X \succeq 0$ and $t \leq \mathrm{Det}^q(X)$, then

(1) We can extend X by appropriately chosen lower triangular matrix Δ to a solution of (D) in such a way that if $\delta = \mathrm{diag}(\Delta)$, then $\prod_{i=1}^{m} \delta_i = \mathrm{Det}(X)$.

Indeed, let $\widehat{\Delta}$ be the Choleski factor of X. Let $\widehat{D}$ be the diagonal matrix with the same diagonal entries as those of $\widehat{\Delta}$, and let $\Delta = \widehat{\Delta}\widehat{D}$, so that the diagonal entries δ_i of Δ are squares of the diagonal entries $\widehat{\delta}_i$ of the matrix $\widehat{\Delta}$. Thus, $D(\Delta) = \widehat{D}^2$. It follows that for every $\epsilon > 0$ one has $\Delta[D(\Delta) + \epsilon I]^{-1}\Delta^T = \widehat{\Delta}\widehat{D}[\widehat{D}^2 + \epsilon I]^{-1}\widehat{D}\widehat{\Delta}^T \preceq \widehat{\Delta}\widehat{\Delta}^T = X$. We see that by the Schur Complement Lemma all matrices of the form $\begin{pmatrix} X & \Delta \\ \Delta^T & D(\Delta) + \epsilon I \end{pmatrix}$ with $\epsilon > 0$ are positive semidefinite, whence $\begin{pmatrix} X & \Delta \\ \Delta^T & D(\Delta) \end{pmatrix} \succeq 0$. Thus, (D) is indeed satisfied by (X, Δ). And of course $X = \widehat{\Delta}\widehat{\Delta}^T \Rightarrow \mathrm{Det}(X) = \mathrm{Det}^2(\widehat{\Delta}) = \prod_{i=1}^{m} \widehat{\delta}_i^2 = \prod_{i=1}^{m} \delta_i$.

(2) Since $\delta = \mathrm{diag}(\Delta) \succeq 0$ and $\prod_{i=1}^m \delta_i = \mathrm{Det}(X)$, we get $t \le \mathrm{Det}^q(X) = \left(\prod_{i=1}^m \delta_i\right)^q$, so that we can extend (t, δ) by a properly chosen u to a solution of the LMI $S(\mathrm{diag}(\Delta), t; u) \succeq 0$.

We conclude that if $X \succeq 0$ and $t \le \mathrm{Det}^q(X)$, then one can extend the pair X, t by properly chosen Δ and u to a solution of the LMI (D) & $S(\mathrm{diag}(\Delta), t; u) \succeq 0$, which is the first part of the proof of (!).

To complete the proof of (!), it suffices to demonstrate that if for a given pair X, t there exist Δ and u such that (D) and the LMI $S(\mathrm{diag}(\Delta), t; u) \succeq 0$ are satisfied, then X is positive semidefinite and $t \le \mathrm{Det}^q(X)$. This is immediate: denoting $\delta = \mathrm{diag}(\Delta)$ $[\ge 0]$ and applying the Schur Complement Lemma, we conclude that $X \succeq \Delta[D(\Delta) + \epsilon I]^{-1}\Delta^T$ for every $\epsilon > 0$. Applying **(W)**, we get $\lambda(X) \ge \lambda(\Delta[D(\Delta) + \epsilon I]^{-1}\Delta^T)$, whence of course $\mathrm{Det}(X) \ge \mathrm{Det}(\Delta[D(\Delta) + \epsilon I]^{-1}\Delta^T) = \prod_{i=1}^m \delta_i^2/(\delta_i + \epsilon)$. Passing to limit as $\epsilon \to 0$, we get $\prod_{i=1}^m \delta_i \le \mathrm{Det}(X)$. On the other hand, the LMI $S(\delta, t; u) \succeq 0$ takes place, which means that $t \le \left(\prod_{i=1}^m \delta_i\right)^q$. Combining the resulting inequalities, we come to $t \le \mathrm{Det}^q(X)$, as required. ∎

18e. Negative powers of the determinant. Let q be a positive rational. Then the function

$$f(X) = \begin{cases} \mathrm{Det}^{-q}(X), & X \succ 0 \\ +\infty, & \text{otherwise} \end{cases}$$

of symmetric $m \times m$ matrix X is SDr.

The construction is completely similar to the one used in Example 18d. As we remember from Lecture 2, Example 16, the function $g(\delta) = (\delta_1 ... \delta_m)^{-q}$ of positive vector $\delta = (\delta_1, ..., \delta_m)^T$ is CQr and is therefore SDr as well. Let an SDR of the function be given by LMI $\mathcal{R}(\delta, t, u) \succeq 0$. The same arguments as in Example 18d demonstrate that the pair of LMI's (D) & $\mathcal{R}(\mathrm{Dg}(\Delta), t, u) \succeq 0$ is an SDR for f.

In examples 18, 18b – 18d we were discussed SD-representability of particular functions of eigenvalues of a symmetric matrix. Here is a general statement of this type:

Proposition 3.2.1 *Let $g(x_1, ..., x_m) : \mathbf{R}^m \to \mathbf{R} \cup \{+\infty\}$ be a symmetric (i.e., invariant with respect to permutations of the coordinates $x_1, ..., x_m$) SD-representable function:*

$$t \ge f(x) \Leftrightarrow \exists u : \mathcal{S}(x, t, u) \succeq 0,$$

with $\mathcal{S}$ affinely depending on x, t, u. Then the function

$$f(X) = g(\lambda(X))$$

of symmetric $m \times m$ matrix X is SDr, with SDR given by the relation

$$\begin{array}{ll} (a) & t \ge f(X) \\ & \Updownarrow \\ & \exists x_1, ..., x_m, u : \\ (b) & \left\{ \begin{array}{c} \mathcal{S}(x_1, ..., x_m, t, u) \succeq 0 \\ x_1 \ge x_2 \ge ... \ge x_m \\ S_j(X) \le x_1 + ... + x_j, \ j = 1, ..., m-1 \\ \mathrm{Tr}(X) = x_1 + ... + x_m \end{array} \right. \end{array} \qquad (3.2.3)$$

(recall that the functions $S_j(X) = \sum_{i=1}^k \lambda_i(X)$ are SDr, see Example 18c). Thus, the solution set of (b) is SDr (as an intersection of SDr sets), which implies SD-representability of the projection of this set onto the (X, t)-plane; by (3.2.3) the latter projection is exactly the epigraph of f).

The proof of Proposition 3.2.1 is based upon an extremely useful result known as *Birkhoff's Theorem*[5)].

As a corollary of Proposition 3.2.1, we see that the following functions of a symmetric $m \times m$ matrix X are SDr:

- $f(X) = -\mathrm{Det}^q(X)$, $X \succeq 0$, $q \leq \frac{1}{m}$ is a positive rational (this fact was already established directly);
  [here $g(x_1, ..., x_m) = (x_1...x_m)^q : \mathbf{R}^n_+ \to \mathbf{R}$; a CQR (and thus – a SDR) of g is presented in Example 15 of Lecture 2]
- $f(x) = \mathrm{Det}^{-q}(X)$, $X \succ 0$, q is a positive rational (cf. Example 18e)
  [here $g(x_1, ..., x_m) = (x_1, ..., x_m)^{-q} : \mathbf{R}^m_{++} \to \mathbf{R}$; a CQR of g is presented in Example 16 of Lecture 2]
- $\|X\|_p = \left(\sum\limits_{i=1}^m |\lambda_i(X)|^p\right)^{1/p}$, $p \geq 1$ is rational
  $[g(x) = \|x\|_p \equiv \left(\sum\limits_{i=1}^m |x_i|^p\right)^{1/p}$, see Lecture 2, Example 17a]
- $\|X_+\|_p = \left(\sum\limits_{i=1}^m \max^p[\lambda_i(X), 0]\right)^{1/p}$, $p \geq 1$ is rational
  [here $g(x) = \|x_+\|_p \equiv \left(\sum\limits_{i=1}^m |\max^p[x_i, 0]\right)^{1/p}$, see Lecture 2, Example 17b]

SD-representability of functions of singular values. Consider the space $\mathbf{M}^{k,l}$ of $k \times l$ rectangular matrices and assume that $k \leq l$. Given a matrix $A \in \mathbf{M}^{k,l}$, consider the symmetric positive semidefinite $k \times k$ matrix $(AA^T)^{1/2}$; its eigenvalues are called *singular values* of A and are denoted by $\sigma_1(A), ...\sigma_k(A)$: $\sigma_i(A) = \lambda_i((AA^T)^{1/2})$. According to the convention on how we enumerate eigenvalues of a symmetric matrix, the singular values form a non-ascending sequence:

$$\sigma_1(A) \geq \sigma_2(A) \geq ... \geq \sigma_k(A).$$

The importance of the singular values comes from the Singular Value Decomposition Theorem which states that a $k \times l$ matrix A ($k \leq l$) can be represented as

$$A = \sum_{i=1}^k \sigma_i(A) e_i f_i^T,$$

where $\{e_i\}_{i=1}^k$ and $\{f_i\}_{i=1}^k$ are orthonormal sequences in $\mathbf{R}^k$ and $\mathbf{R}^l$, respectively; this is a surrogate of the eigenvalue decomposition of a symmetric $k \times k$ matrix

$$A = \sum_{i=1}^k \lambda_i(A) e_i e_i^T,$$

where $\{e_i\}_{i=1}^k$ form an orthonormal eigenbasis of A.

Among the singular values of a rectangular matrix, the most important is the largest $\sigma_1(A)$. This is nothing but the *operator* (or *spectral*) *norm* of A:

$$|A| = \max\{\|Ax\|_2 \mid \|x\|_2 \leq 1\}.$$

[5)] The Birkhoff Theorem, which, aside of other applications, implies a number of crucial facts about eigenvalues of symmetric matrices, by itself even does not mention the word "eigenvalue" and reads: *The extreme points of the polytope $\mathcal{P}$ of double stochastic $m \times m$ matrices – those with nonnegative entries and unit sums of entries in every row and every column – are exactly the permutation matrices (those with a single nonzero entry, equal to 1, in every row and every column).*

For a symmetric matrix, the singular values are exactly the modulae of the eigenvalues, and our new definition of the norm coincides with the one already given in 18b.

It turns out that the sum of a given number of the largest singular values of A

$$\Sigma_p(A) = \sum_{i=1}^{p} \sigma_i(A)$$

is a convex and, moreover, a SDr function of A. In particular, the operator norm of A is SDr:

19. The sum $\Sigma_p(X)$ of p largest singular values of a rectangular matrix $X \in \mathbf{M}^{k,l}$ is SDr. In particular, the operator norm of a rectangular matrix is SDr:

$$|X| \le t \Leftrightarrow \begin{pmatrix} tI_l & -X^T \\ -X & tI_k \end{pmatrix} \succeq 0.$$

Indeed, the result in question follows from the fact that the sums of p largest eigenvalues of a symmetric matrix are SDr (Example 18c) due to the following

Observation. *The singular values $\sigma_i(X)$ of a rectangular $k \times l$ matrix X $(k \le l)$ for $i \le k$ are equal to the eigenvalues $\lambda_i(\bar{X})$ of the $(k+l) \times (k+l)$ symmetric matrix*

$$\bar{X} = \begin{pmatrix} 0 & X^T \\ X & 0 \end{pmatrix}.$$

Since $\bar{X}$ linearly depends on X, SDR's of the functions $S_p(\cdot)$ induce SDR's of the functions $\Sigma_p(X) = S_p(\bar{X})$ (Rule on affine substitution, Lecture 2; recall that all "calculus rules" established in Lecture 2 for CQR's are valid for SDR's as well).

Let us justify our observation. Let $X = \sum_{i=1}^{k} \sigma_i(X) e_i f_i^T$ be a singular value decomposition of X. We claim that the $2k$ $(k+l)$-dimensional vectors $g_i^+ = \begin{pmatrix} f_i \\ e_i \end{pmatrix}$ and $g_i^- = \begin{pmatrix} f_i \\ -e_i \end{pmatrix}$ are orthogonal to each other, and they are eigenvectors of $\bar{X}$ with the eigenvalues $\sigma_i(X)$ and $-\sigma_i(X)$, respectively. Moreover, $\bar{X}$ vanishes on the orthogonal complement of the linear span of these vectors. In other words, we claim that the eigenvalues of $\bar{X}$, arranged in the non-ascending order, are as follows:

$$\sigma_1(X), \sigma_2(X), ..., \sigma_k(X), \underbrace{0, ..., 0}_{2(l-k)}, -\sigma_k(X), -\sigma_{k-1}(X), ..., -\sigma_1(X);$$

this, of course, proves our Observation.

Now, the fact that the $2k$ vectors $g_i^{\pm}$, $i = 1, ..., k$, are mutually orthogonal and nonzero is evident. Furthermore (we write σ_i instead of $\sigma_i(X)$),

$$\begin{aligned} \begin{pmatrix} 0 & X^T \\ X & 0 \end{pmatrix} \begin{pmatrix} f_i \\ e_i \end{pmatrix} &= \begin{pmatrix} 0 & \sum_{j=1}^{k} \sigma_j f_j e_j^T \\ \sum_{j=1}^{k} \sigma_j e_j f_j^T & 0 \end{pmatrix} \begin{pmatrix} f_i \\ e_i \end{pmatrix} \\ &= \begin{pmatrix} \sum_{j=1}^{k} \sigma_j f_j (e_j^T e_i) \\ \sum_{j=1}^{k} \sigma_j e_j (f_j^T f_i) \end{pmatrix} \\ &= \sigma_i \begin{pmatrix} f_i \\ e_i \end{pmatrix} \end{aligned}$$

(we have used that both $\{f_j\}$ and $\{e_j\}$ are orthonormal systems). Thus, g_i^+ is an eigenvector of $\bar{X}$ with the eigenvalue $\sigma_i(X)$. Similar computation shows that g_i^- is an eigenvector of $\bar{X}$ with the eigenvalue $-\sigma_i(X)$.

It remains to verify that if $h = \begin{pmatrix} f \\ e \end{pmatrix}$ is orthogonal to all $g_i^\pm$ (f is l-dimensional, e is k-dimensional), then $\bar{X}h = 0$. Indeed, the orthogonality assumption means that $f^T f_i \pm e^T e_i = 0$ for all i, whence $e^T e_i = 0$ and $f^T f_i = 0$ for all i. Consequently,

$$\begin{pmatrix} 0 & X^T \\ X & 0 \end{pmatrix} \begin{pmatrix} f \\ e \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^k f_j(e_j^T e) \\ \sum_{i=1}^k e_j(f_j^T f) \end{pmatrix} = 0. \qquad \blacksquare$$

Looking at Proposition 3.2.1, we see that the fact that specific functions of eigenvalues of a symmetric matrix X, namely, the sums $S_k(X)$ of k largest eigenvalues of X, are SDr, underlies the possibility to build SDR's for a wide class of functions of the eigenvalues. The role of the sums of k largest singular values of a *rectangular* matrix X is equally important:

Proposition 3.2.2 *Let $g(x_1, ..., x_k) : \mathbf{R}^k_+ \to \mathbf{R} \cup \{+\infty\}$ be a symmetric monotone function:*

$$0 \le y \le x \in \mathrm{Dom}\, f \Rightarrow f(y) \le f(x).$$

Assume that g is SDr:

$$t \ge g(x) \Leftrightarrow \exists u : \mathcal{S}(x, t, u) \succeq 0,$$

with $\mathcal{S}$ affinely depending on x, t, u. Then the function

$$f(X) = g(\sigma(X))$$

of $k \times l$ ($k \le l$) rectangular matrix X is SDr, with SDR given by the relation

$$\begin{array}{ll} (a) & t \ge f(X) \\ & \Updownarrow \\ & \exists x_1, ..., x_k, u : \\ (b) & \left\{ \begin{array}{c} \mathcal{S}(x_1, ..., x_k, t, u) \succeq 0 \\ x_1 \ge x_2 \ge ... \ge x_k \\ \Sigma_j(X) \le x_1 + ... + x_j, \ j = 1, ..., m \end{array} \right. \end{array} \tag{3.2.4}$$

Note the difference between the symmetric (Proposition 3.2.1) and the non-symmetric (Proposition 3.2.2) situations: in the former the function $g(x)$ was assumed to be SDr and symmetric only, while in the latter the monotonicity requirement is added.
The proof of Proposition 3.2.2 is outlined in Section 3.7

"Nonlinear matrix inequalities". There are several cases when matrix inequalities $F(x) \succeq 0$, where F is a nonlinear function of x taking values in the space of symmetric $m \times m$ matrices, can be "linearized" – expressed via LMI's.

20a. General quadratic matrix inequality. Let X be a rectangular $k \times l$ matrix and

$$F(X) = (AXB)(AXB)^T + CXD + (CXD)^T + E$$

be a "quadratic" matrix-valued function of X; here $A, B, C, D, E = E^T$ are rectangular matrices of appropriate sizes. Let m be the row size of the values of F. Consider the "$\succeq$-epigraph" of the (matrix-valued!) function F – the set

$$\{(X, Y) \in \mathbf{M}^{k,l} \times \mathbf{S}^m \mid F(X) \preceq Y\}.$$

We claim that this set is SDr with the SDR

$$\left(\begin{array}{c|c} I_r & (AXB)^T \\ \hline AXB & Y - E - CXD - (CXD)^T \end{array} \right) \succeq 0 \qquad [B : l \times r]$$

Indeed, by the Schur Complement Lemma our LMI is satisfied if and only if the Schur complement of the North-Western block is positive semidefinite, which is exactly our original "quadratic" matrix inequality.

20b. General "fractional-quadratic" matrix inequality. Let X be a rectangular $k\times l$ matrix, and V be a positive definite symmetric $l\times l$ matrix. Then we can define the matrix-valued function

$$F(X,V)=XV^{-1}X^T$$

taking values in the space of $k\times k$ symmetric matrices. We claim that the closure of the $\succeq$-epigraph of this (matrix-valued!) function, i.e., the set

$$E=\mathrm{cl}\,\{(X,V;Y)\in\mathbf{M}^{k,l}\times\mathbf{S}^l_{++}\times\mathbf{S}^k\mid F(X,V)\equiv XV^{-1}X^T\preceq Y\}$$

is SDr, and an SDR of this set is given by the LMI

$$\begin{pmatrix}V & X^T\\ X & Y\end{pmatrix}\succeq 0. \tag{R}$$

Indeed, by the Schur Complement Lemma a triple (X,V,Y) with *positive definite* V belongs to the "epigraph of F" – satisfies the relation $F(X,V)\preceq Y$ – if and only if it satisfies (R). Now, if a triple (X,V,Y) belongs to E, i.e., it is the limit of a sequence of triples from the epigraph of F, then it satisfies (R) (as a limit of triples satisfying (R)). Vice versa, if a triple (X,V,Y) satisfies (R), then V is positive semidefinite (as a diagonal block in a positive semidefinite matrix). The "regularized" triples $(X,V_\epsilon=V+\epsilon I_l,Y)$ associated with $\epsilon>0$ satisfy (R) along with the triple (X,V,R); since, as we just have seen, $V\succeq 0$, we have $V_\epsilon\succ 0$, for $\epsilon>0$. Consequently, the triples (X,V_ϵ,Y) belong to E (this was our very first observation); since the triple (X,V,Y) is the limit of the regularized triples which, as we have seen, all belong to the epigraph of F, the triple (X,Y,V) belongs to the closure E of this epigraph. ∎

20c. Matrix inequality $Y\preceq(C^TX^{-1}C)^{-1}$. In the case of scalars x,y the inequality $y\le(cx^{-1}c)^{-1}$ in variables x,y is just an awkward way to write down the linear inequality $y\le c^{-2}x$, but it naturally to the matrix analogy of the original inequality, namely, $Y\preceq(C^TX^{-1}C)^{-1}$, with rectangular $m\times n$ matrix C and variable symmetric $n\times n$ matrix Y and $m\times m$ matrix X. In order for the matrix inequality to make sense, we should assume that the rank of C equals n (and thus $m\ge n$). Under this assumption, the matrix $(C^TX^{-1}C)^{-1}$ makes sense at least for a positive definite X. We claim that the closure of the solution set of the resulting inequality – the set

$$\mathcal{X}=\mathrm{cl}\,\{(X,Y)\in\mathbf{S}^m\times\mathbf{S}^n\mid X\succ 0, Y\preceq(C^TX^{-1}C)^{-1}\}$$

is SDr:

$$\mathcal{X}=\{(X,Y)\mid\exists Z:\,Y\preceq Z, Z\succeq 0, X\succeq CZC^T\}.$$

Indeed, let us denote by $\mathcal{X}'$ the set in the right hand side of the latter relation; we should prove that $\mathcal{X}'=\mathcal{X}$. By definition, $\mathcal{X}$ is the closure of its intersection with the domain $X\succ 0$. It is clear that $\mathcal{X}'$ also is the closure of its intersection with the domain $X\succ 0$. Thus, all we need to prove is that a pair (Y,X) with $X\succ 0$ belongs to $\mathcal{X}$ if and only if it belongs to $\mathcal{X}'$.
"If" part: Assume that $X\succ 0$ and $(Y,X)\in\mathcal{X}'$. Then there exists Z such that $Z\succeq 0$, $Z\succeq Y$ and $X\succeq CZC^T$. Let us choose a sequence $Z_i\succ Z$ such that $Z_i\to Z$, $i\to\infty$. Since $CZ_iC^T\to CZC^T\preceq X$ as $i\to\infty$, we can find a sequence of matrices X_i such that $X_i\to X$, $i\to\infty$, and $X_i\succ CZ_iC^T$ for all i. By the Schur Complement Lemma, the matrices $\begin{pmatrix}X_i & C\\ C^T & Z_i^{-1}\end{pmatrix}$ are positive definite; applying this lemma again, we conclude that $Z_i^{-1}\succeq C^TX_i^{-1}C$. Note that the left and the right hand side matrices in the latter inequality are positive definite. Now let us use the following simple fact

Lemma 3.2.2 *Let U, V be positive definite matrices of the same size. Then*

$$U \preceq V \Leftrightarrow U^{-1} \succeq V^{-1}.$$

Proof. Note that we can multiply an inequality $A \preceq B$ by a matrix Q from the left and Q^T from the right:

$$A \preceq B \Rightarrow QAQ^T \preceq QBQ^T \quad [A, B \in \mathbf{S}^m, Q \in \mathbf{M}^{k,m}]$$

(why?) Thus, if $0 \prec U \preceq V$, then $V^{-1/2}UV^{-1/2} \preceq V^{-1/2}VV^{-1/2} = I$ (note that $V^{-1/2} = [V^{-1/2}]^T$), whence clearly $V^{1/2}U^{-1}V^{1/2} = [V^{-1/2}UV^{-1/2}]^{-1} \succeq I$. Thus, $V^{1/2}U^{-1}V^{1/2} \succeq I$; multiplying this inequality from the left and from the right by $V^{-1/2} = [V^{-1/2}]^T$, we get $U^{-1} \succeq V^{-1}$. ∎

Applying Lemma 3.2.2 to the inequality $Z_i^{-1} \succeq C^T X_i^{-1} C [\succ 0]$, we get $Z_i \preceq (C^T X_i^{-1} C)^{-1}$. As $i \to \infty$, the left hand side in this inequality converges to Z, and the right hand side converges to $(C^T X^{-1} C)^{-1}$. Hence $Z \preceq (C^T X^{-1} C)^{-1}$, and since $Y \preceq Z$, we get $Y \preceq (C^T X^{-1} C)^{-1}$, as claimed.

"Only if" part: Let $X \succ 0$ and $Y \preceq (C^T X^{-1} C)^{-1}$; we should prove that there exists $Z \succeq 0$ such that $Z \succeq Y$ and $X \succeq CZC^T$. We claim that the required relations are satisfied by $Z = (C^T X^{-1} C)^{-1}$. The only nontrivial part of the claim is that $X \succeq CZC^T$, and here is the required justification: by its origin $Z \succ 0$, and by the Schur Complement Lemma the matrix $\begin{pmatrix} Z^{-1} & C^T \\ C & X \end{pmatrix}$ is positive semidefinite, whence, by the same Lemma, $X \succeq C(Z^{-1})^{-1}C^T = CZC^T$.

Nonnegative polynomials. Consider the problem of the best polynomial approximation – given a function f on certain interval, we want to find its best uniform (or Least Squares, etc.) approximation by a polynomial of a given degree. This problem arises typically as a subproblem in all kinds of signal processing problems. In some situations the approximating polynomial is required to be nonnegative (think, e.g., of the case where the resulting polynomial is an estimate of an unknown probability density); how to express the nonnegativity restriction? As it was shown by Yu. Nesterov [14], it can be done via semidefinite programming:

The set of all nonnegative (on the entire axis, or on a given ray, or on a given segment) *polynomials of a given degree is SDr.*

In this statement (and everywhere below) we identify a polynomial $p(t) = \sum_{i=0}^{k} p_i t^i$ of degree (not exceeding) k with the $(k+1)$-dimensional vector $\mathrm{Coef}(p) = \mathrm{Coef}(p) = (p_0, p_1, ..., p_k)^T$ of the coefficients of p. Consequently, a set of polynomials of the degree $\le k$ becomes a set in $\mathbf{R}^{k+1}$, and we may ask whether this set is or is not SDr.

Let us look what are the SDR's of different sets of nonnegative polynomials. The key here is to get a SDR for the set $P_{2k}^+(\mathbf{R})$ of polynomials of (at most) a given degree $2k$ which are nonnegative on the entire axis[6)]

21a. Polynomials nonnegative on the entire axis: The set $P_{2k}^+(\mathbf{R})$ is SDr – it is the image of the semidefinite cone $\mathbf{S}_+^{k+1}$ under the affine mapping

$$X \mapsto \mathrm{Coef}(e^T(t) X e(t)) : \mathbf{S}^{k+1} \to \mathbf{R}^{2k+1}, \quad e(t) = \begin{pmatrix} 1 \\ t \\ t^2 \\ ... \\ t^k \end{pmatrix} \tag{C}$$

[6)] It is clear why we have restricted the degree to be even: a polynomial of an odd degree cannot be nonnegative on the entire axis!

First note that the fact that $P^+ \equiv P^+_{2k}(\mathbf{R})$ is an affine image of the semidefinite cone indeed implies the SD-representability of P^+, see the "calculus" of conic representations in Lecture 2. Thus, all we need is to show that P^+ is exactly the same as the image, let it be called P, of $\mathbf{S}^{k+1}_+$ under the mapping (C).

(1) The fact that P *is contained* in P^+ is immediate. Indeed, let X be a $(k+1)\times(k+1)$ positive semidefinite matrix. Then X is a sum of dyadic matrices:

$$X = \sum_{i=1}^{k+1} p^i (p^i)^T, p^i = (p_{i0}, p_{i1}, ..., p_{ik})^T \in \mathbf{R}^{k+1}$$

(why?) But then

$$e^T(t) X e(t) = \sum_{i=1}^{k+1} e^T(t) p^i [p^i]^T e(t) = \sum_{i=1}^{k+1} \left(\sum_{j=0}^{k} p_{ij} t^j \right)^2$$

is the sum of squares of other polynomials and therefore is nonnegative on the axis. Thus, the image of X under the mapping (C) belongs to P^+.

Note that reversing our reasoning, we get the following result:

> (!) *If a polynomial $p(t)$ of degree $\le 2k$ can be represented as a sum of squares of other polynomials, then the vector* Coef(p) *of the coefficients of p belongs to the image of $\mathbf{S}^{k+1}_+$ under the mapping* (C).

With (!), the remaining part of the proof – the demonstration that the image of $\mathbf{S}^{k+1}_+$ *contains* P^+, is readily given by the following well-known algebraic fact:

> (!!) *A polynomial is nonnegative on the axis <u>if and only if</u> it is a sum of squares of polynomials.*
>
> The proof of (!!) is so nice that we cannot resist the temptation to present it here. The "if" part is evident. To prove the "only if" one, assume that $p(t)$ is nonnegative on the axis, and let the degree of p (it must be even) be $2k$. Now let us look at the roots of p. The real roots $\lambda_1, ..., \lambda_r$ must be of even multiplicities $2m_1, 2m_2, ...2m_r$ each (otherwise p would alter its sign in a neighbourhood of a root, which contradicts the nonnegativity). The complex roots of p can be arranged in conjugate pairs $(\mu_1, \mu_1^*), (\mu_2, \mu_2^*), ..., (\mu_s, \mu_s^*)$, and the factor of p
>
> $$(t-\mu_i)(t-\mu_i^*) = (t - \Re\mu_i)^2 + (\Im\mu_i)^2$$
>
> corresponding to such a pair is a sum of two squares. Finally, the leading coefficient of p is positive. Consequently, we have
>
> $$p(t) = \omega^2 [(t-\lambda_1)^2]^{m_1} ... [(t-\lambda_r)^2]^{m_r} [(t-\mu_1)(t-\mu_1^*)] ... [(t-\mu_s)(t-\mu_s^*)]$$
>
> *is a product of sums of squares.* But such a product is itself a sum of squares (open the parentheses)!
>
> > In fact we can say more: a nonnegative polynomial p is a sum of just <u>two</u> squares! To see this, note that, as we have seen, p is a product of sums of *two* squares and take into account the following fact (Louville):
> > *The product of sums of two squares is again a sum of two squares:*
> >
> > $$(a^2+b^2)(c^2+d^2) = (ac-bd)^2 + (ad+bc)^2$$
> >
> > (cf. with: "the modulus of a product of two complex numbers is the product of their modulae").

Equipped with the SDR of the set $P^+_{2k}(\mathbf{R})$ of polynomials nonnegative on the entire axis, we can immediately obtain SDR's for the polynomials nonnegative on a given ray/segment:

21b. Polynomials nonnegative on a ray/segment.

1) *The set $P_k^+(\mathbf{R}_+)$ of (coefficients of) polynomials of degree $\le k$ which are nonnegative on the nonnegative ray, is SDr.*

Indeed, this set is the inverse image of the SDr set $P_{2k}^+(\mathbf{R})$ under the linear mapping of the spaces of (coefficients of) polynomials given by the mapping

$$p(t) \mapsto p^+(t) \equiv p(t^2)$$

(recall that the inverse image of an SDr set is SDr).

2) *The set $P_k^+([0,1])$ of (coefficients of) polynomials of degree $\le k$ which are nonnegative on the segment $[0,1]$, is SDr.*

Indeed, a polynomial $p(t)$ of degree $\le k$ is nonnegative on $[0,1]$ if and only if the rational function

$$g(t) = p\left(\frac{t^2}{1+t^2}\right)$$

is nonnegative on the entire axis, or, which is the same, if and only if the polynomial

$$p^+(t) = (1+t^2)^k g(t)$$

of degree $\le 2k$ is nonnegative on the entire axis. The coefficients of p^+ depend linearly on the coefficients of p, and we conclude that $P_k^+([0,1])$ is the inverse image of the SDr set $P_{2k}^+(\mathbf{R})$ under certain linear mapping.

Our last example in this series deals with trigonometric polynomials

$$p(\phi) = a_0 + \sum_{\ell=1}^{k} [a_\ell \cos(\ell\phi) + b_\ell \sin(\ell\phi)]$$

Identifying such a polynomial with its vector of coefficients $\mathrm{Coef}(p) \in \mathbf{R}^{2k+1}$, we may ask how to express the set $S_k^+(\Delta)$ of those trigonometric polynomials of degree $\le k$ which are nonnegative on a segment $\Delta \subset [0, 2\pi]$.

21c. Trigonometric polynomials nonnegative on a segment. The set $S_k^+(\Delta)$ is SDr.

Indeed, $\sin(\ell\phi)$ and $\cos(\ell\phi)$ are polynomials of $\sin(\phi)$ and $\cos(\phi)$, and the latter functions, in turn, are rational functions of $\zeta = \tan(\phi/2)$:

$$\cos(\phi) = \frac{1-\zeta^2}{1+\zeta^2}, \sin(\phi) = \frac{2\zeta}{1+\zeta^2} \quad [\zeta = \tan(\phi/2)].$$

Consequently, a trigonometric polynomial $p(\phi)$ of degree $\le k$ can be represented as a rational function of $\zeta = \tan(\phi/2)$:

$$p(\phi) = \frac{p^+(\zeta)}{(1+\zeta^2)^k} \quad [\zeta = \tan(\phi/2)],$$

where the coefficients of the algebraic polynomial p^+ of degree $\le 2k$ are linear functions of the coefficients of p. Now, the requirement for p to be nonnegative on a given segment $\Delta \subset [0, 2\pi]$ is equivalent to the requirement for p^+ to be nonnegative on a "segment" Δ^+ (which, depending on Δ, may be either the usual finite segment, or a ray, or the entire axis). We see that $S_k^+(\Delta)$ is inverse image, under certain linear mapping, of the SDr set $P_{2k}^+(\Delta^+)$, so that $S_k^+(\Delta)$ itself is SDr.

Finally, we may ask which part of the above results can be saved when we pass from nonnegative polynomials of one variable to those of two or more variables. Unfortunately, not too much. E.g., among nonnegative polynomials of a given degree with $r > 1$ variables, exactly those of them *who are sums of squares* can be obtained as the image of a positive semidefinite cone under certain linear mapping similar to (D). The difficulty is that in the multi-dimensional case the nonnegativity of a polynomial is not equivalent to its representability as a sum of squares, thus, the positive semidefinite cone gives only part of the polynomials we are interested to describe.

3.3 Applications of Semidefinite Programming in Engineering

Due to its tremendous expressive abilities, Semidefinite Programming allows to pose and process numerous highly nonlinear convex optimization programs arising in applications, in particular, in Engineering. We are about to outline briefly just few instructive examples.

3.3.1 Dynamic Stability in Mechanics

"Free motions" of the so called *linearly elastic* mechanical systems, i.e., their behaviour when no external forces are applied, are governed by systems of differential equations of the type

$$M\frac{d^2}{dt^2}x(t) = -Ax(t), \tag{N}$$

where $x(t) \in \mathbf{R}^n$ is the state vector of the system at time t, M is the (generalized) "mass matrix", and A is the "stiffness" matrix of the system. Basically, (N) is the Newton law for a system with the potential energy $\frac{1}{2}x^TAx$.

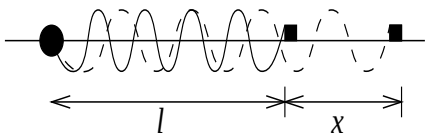
As a simple example, consider a system of k points of masses $\mu_1, ..., \mu_k$ linked by springs with given elasticity coefficients; here x is the vector of the displacements $x_i \in \mathbf{R}^d$ of the points from their equilibrium positions e_i ($d = 1/2/3$ is the dimension of the model). The Newton equations become

$$\mu_i\frac{d^2}{dt^2}x_i(t) = -\sum_{j\neq i}\nu_{ij}(e_i - e_j)(e_i - e_j)^T(x_i - x_j), i = 1, ..., k,$$

with ν_{ij} given by

$$\nu_{ij} = \frac{\kappa_{ij}}{\|e_i - e_j\|_2^3},$$

where $\kappa_{ij} > 0$ are the elasticity coefficients of the springs. The resulting system is of the form (N) with a diagonal matrix M and a positive semidefinite symmetric matrix A. The well-known simplest system of this type is a *pendulum* (a single point capable to slide along a given axis and linked by a spring to a fixed point on the axis):



$\frac{d^2}{dt^2}x(t) = -\nu x(t), \ \ \nu = \frac{\kappa}{l}.$

Another example is given by *trusses* – mechanical constructions, like a railway bridge or the Eiffel Tower, built from linked to each other thin elastic bars.

Note that in the above examples both the mass matrix M and the stiffness matrix A are symmetric positive semidefinite; in "nondegenerate" cases they are even positive definite, and this is what we assume

from now on. Under this assumption, we can pass in (N) from the variables $x(t)$ to the variables $y(t) = M^{1/2}x(t)$; in these variables the system becomes

$$\frac{d^2}{dt^2}y(t) = -\hat{A}y(t),\ \hat{A} = M^{-1/2}AM^{-1/2}. \tag{N$'$}$$

It is well known that the space of solutions of system (N$'$) (where $\hat{A}$ is symmetric positive definite) is spanned by fundamental (perhaps complex-valued) solutions of the form $\exp\{\mu t\}f$. A nontrivial (with $f \neq 0$) function of this type is a solution to (N$'$) if and only if

$$(\mu^2 I + \hat{A})f = 0,$$

so that the allowed values of μ^2 are the minus eigenvalues of the matrix $\hat{A}$, and f's are the corresponding eigenvectors of $\hat{A}$. Since the matrix $\hat{A}$ is symmetric positive definite, the only allowed values of μ are purely imaginary, with the imaginary parts $\pm\sqrt{\lambda_j(\hat{A})}$. Recalling that the eigenvalues/eigenvectors of $\hat{A}$ are exactly the eigenvalues/eigenvectors of the pencil $[M, A]$, we come to the following result:

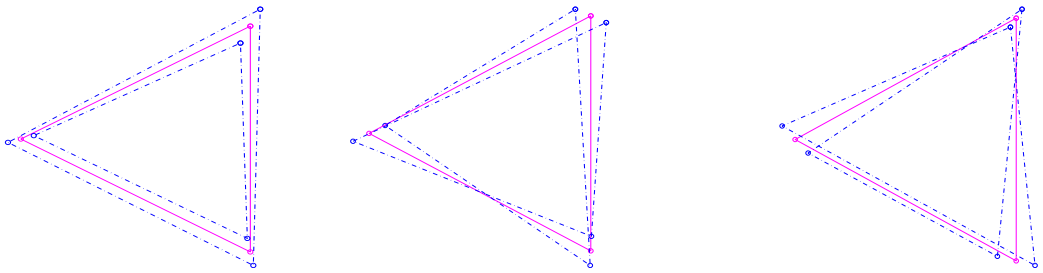
(!) *In the case of positive definite symmetric M, A, the solutions to* (N) *– the "free motions" of the corresponding mechanical system $\mathcal{S}$ – are of the form*

$$x(t) = \sum_{j=1}^{n}[a_j \cos(\omega_j t) + b_j \sin(\omega_j t)]e_j,$$

where a_j, b_j are free real parameters, e_j are the eigenvectors of the pencil $[M, A]$:

$$(\lambda_j M - A)e_j = 0$$

and $\omega_j = \sqrt{\lambda_j}$. Thus, the "free motions" of the system $\mathcal{S}$ are mixtures of harmonic oscillations along the eigenvectors of the pencil $[M, A]$, and the frequencies of the oscillations ("the eigenfrequencies of the system") are the square roots of the corresponding eigenvalues of the pencil.



$\omega = 1.274$ $\omega = 0.957$ $\omega = 0.699$

"Nontrivial" modes of a spring triangle (3 unit masses linked by springs)

Shown are 3 "eigenmotions" (modes) of a spring triangle with nonzero frequencies; at each picture, the dashed lines depict two instant positions of the oscillating triangle.
There are 3 more "eigenmotions" with zero frequency, corresponding to shifts and rotation of the triangle.

From the engineering viewpoint, the "dynamic behaviour" of mechanical constructions such as buildings, electricity masts, bridges, etc., is the better the larger are the eigenfrequencies of the system[7)]. This is why a typical design requirement in mechanical engineering is a lower bound

$$\lambda_{\min}(A : M) \geq \lambda_* \quad [\lambda_* > 0] \tag{3.3.1}$$

[7)] Think about a building and an earthquake or about sea waves and a light house: in this case the external load acting at the system is time-varying and can be represented as a sum of harmonic oscillations of different (and low) frequencies; if some of these frequencies are close to the eigenfrequencies of the system, the system can be crushed by resonance. In order to avoid this risk, one is interested to move the eigenfrequencies of the system away from 0 as far as possible.

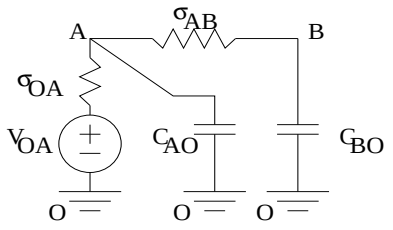
on the *smallest* eigenvalue $\lambda_{\min}(A : M)$ of the pencil $[M, A]$ comprised of the mass and the stiffness matrices of the would-be system. In the case of positive definite symmetric mass matrices (3.3.1) is equivalent to the matrix inequality

$$A - \lambda_* M \succeq 0. \tag{3.3.2}$$

If M and A are affine functions of the design variables (as is the case in, e.g., Truss Design), the matrix inequality (3.3.2) is a *linear* matrix inequality on the design variables, and therefore it can be processed via the machinery of semidefinite programming. Moreover, in the cases when A is affine in the design variables, and M is constant, (3.3.2) is an LMI in the design variables and λ_*, and we may play with λ_*, e.g., solve a problem of the type "given the mass matrix of the system to be designed and a number of (SDr) constraints on the design variables, build a system with the minimum eigenfrequency as large as possible", which is a semidefinite program, provided that the stiffness matrix is affine in the design variables.

3.3.2 Design of chips and Boyd's time constant

Consider an RC-electric circuit, i.e., a circuit comprised of three types of elements: (1) resistors; (2) capacitors; (3) resistors in a series combination with outer sources of voltage:



A simple circuit

Element OA:	outer supply of voltage V_{OA} and resistor with conductance σ_{OA}
Element AO:	capacitor with capacitance C_{AO}
Element AB:	resistor with conductance σ_{AB}
Element BO:	capacitor with capacitance C_{BO}

E.g., a chip is, electrically, a complicated circuit comprised of elements of the indicated type. When designing chips, the following characteristics are of primary importance:

- *Speed.* In a chip, the outer voltages are switching at certain frequency from one constant value to another. Every switch is accompanied by a "transition period"; during this period, the potentials/currents in the elements are moving from their previous values (corresponding to the static steady state for the "old" outer voltages) to the values corresponding to the new static steady state. Since there are elements with "inertia" – capacitors – this transition period takes some time[8]). In order to ensure stable performance of the chip, the transition period should be much less than the time between subsequent switches in the outer voltages. Thus, the duration of the transition period is responsible for the speed at which the chip can perform.

- *Dissipated heat.* Resistors in the chip dissipate heat which should be eliminated, otherwise the chip will not function. This requirement is very serious for modern "high-density" chips. Thus, a characteristic of vital importance is the dissipated heat power.

The two objectives: high speed (i.e., a small transition period) and small dissipated heat – usually are conflicting. As a result, a chip designer faces the tradeoff problem like "how to get a chip with a given speed and with the minimal dissipated heat". It turns out that different optimization problems related

[8]) From purely mathematical viewpoint, the transition period takes infinite time – the currents/voltages approach asymptotically the new steady state, but never actually reach it. From the engineering viewpoint, however, we may think that the transition period is over when the currents/voltages become close enough to the new static steady state.

to the tradeoff between the speed and the dissipated heat in an RC circuit belong to the "semidefinite universe". We restrict ourselves with building an SDR for the speed.

Simple considerations, based on Kirchoff laws, demonstrate that the transition period in an RC circuit is governed by a linear system of differential equations as follows:

$$C\frac{d}{dt}w(t) = -Sw(t) + Rv. \tag{3.3.3}$$

Here

- The state vector $w(\cdot)$ is comprised of the potentials at all but one nodes of the circuit (the potential at the remaining node – "the ground" – is normalized to be identically zero);
- Matrix $C \succeq 0$ is readily given by the topology of the circuit and the capacitances of the capacitors and is linear in these capacitances. Similarly, matrix $S \succeq 0$ is readily given by the topology of the circuit and the conductances of the resistors and is linear in these conductances. Matrix R is given solely by the topology of the circuit;
- v is the vector of outer voltages; recall that this vector is set to certain constant value at the beginning of the transition period.

As we have already mentioned, the matrices C and S, due to their origin, are positive semidefinite; in nondegenerate cases, they are even positive definite, which we assume from now on.

Let $\widehat{w}$ be the steady state of (3.3.3), so that $S\widehat{w} = Rv$. The difference $\delta(t) = w(t) - \widehat{w}$ is a solution to the homogeneous differential equation

$$C\frac{d}{dt}\delta(t) = -S\delta(t). \tag{3.3.4}$$

Setting $\gamma(t) = C^{1/2}\delta(t)$ (cf. Section 3.3.1), we get

$$\frac{d}{dt}\gamma(t) = -(C^{-1/2}SC^{-1/2})\gamma(t). \tag{3.3.5}$$

Since C and S are positive definite, all eigenvalues λ_i of the symmetric matrix $C^{-1/2}SC^{-1/2}$ are positive. It is clear that the space of solutions to (3.3.5) is spanned by the "eigenmotions"

$$\gamma_i(t) = \exp\{-\lambda_i t\}e_i,$$

where $\{e_i\}$ is an orthonormal eigenbasis of the matrix $C^{-1/2}SC^{-1/2}$. We see that all solutions to (3.3.5) (and thus - to (3.3.4) as well) are exponentially fast converging to 0, or, which is the same, the state $w(t)$ of the circuit exponentially fast approaches the steady state $\widehat{w}$. The "time scale" of this transition is, essentially, defined by the quantity $\lambda_{\min} = \min\limits_i \lambda_i$; a typical "decay rate" of a solution to (3.3.5) is nothing but $T = \lambda_{\min}^{-1}$. S. Boyd has proposed to use T to quantify the length of the transition period, and to use the reciprocal of it – i.e., the quantity $\lambda_{\min}$ itself – as the quantitative measure of the speed. Technically, the main advantage of this definition is that the speed turns out to be the minimum eigenvalue of the matrix $C^{-1/2}SC^{-1/2}$, i.e., the minimum eigenvalue of the matrix pencil $[C : S]$. Thus, the speed in Boyd's definition turns out to be efficiently computable (which is not the case for other, more sophisticated, "time constants" used by engineers). Even more important, with Boyd's approach, a typical design specification "the speed of a circuit should be at least such and such" is modelled by the matrix inequality

$$S \succeq \lambda_* C. \tag{3.3.6}$$

As it was already mentioned, S and C are linear in the capacitances of the capacitors and conductances of the resistors; in typical circuit design problems, the latter quantities are affine functions of the design parameters, and (3.3.6) becomes an LMI in the design parameters.

3.3.3 Lyapunov stability analysis/synthesis

Uncertain dynamical systems. Consider a time-varying *uncertain* linear dynamical system

$$\frac{d}{dt}x(t) = A(t)x(t),\ x(0) = x_0. \tag{ULS}$$

Here $x(t) \in \mathbf{R}^n$ represents the state of the system at time t, and $A(t)$ is a time-varying $n \times n$ matrix. We assume that the system is *uncertain* in the sense that we have no idea of what is x_0, and all we know about $A(t)$ is that this matrix, at any time t, belongs to a given uncertainty set $\mathcal{U}$. Thus, (ULS) represents a wide family of linear dynamic systems rather than a single system; and it makes sense to call a *trajectory* of the uncertain linear system (ULS) every function $x(t)$ which is an "actual trajectory" of a system from the family, i.e., is such that

$$\frac{d}{dt}x(t) = A(t)x(t)$$

for all $t \geq 0$ and certain matrix-valued function $A(t)$ taking all its values in $\mathcal{U}$.

Note that we can model a *nonlinear* dynamic system

$$\frac{d}{dt}x(t) = f(t, x(t)) \quad [x \in \mathbf{R}^n] \tag{NLS}$$

with a given right hand side $f(t, x)$ and a given equilibrium $x(t) \equiv 0$ (i.e., $f(t, 0) = 0,\ t \geq 0$) as an uncertain *linear* system. Indeed, let us define the set $\mathcal{U}_f$ as the closed convex hull of the set of $n \times n$ matrices $\left\{\frac{\partial}{\partial x}f(t, x) \mid t \geq 0, x \in \mathbf{R}^n\right\}$. Then for every point $x \in \mathbf{R}^n$ we have

$$f(t, x) = f(t, 0) + \int_0^s \left[\tfrac{\partial}{\partial x}f(t, sx)\right] x ds = A_x(t)x,$$
$$A_x(t) = \int_0^1 \tfrac{\partial}{\partial x}f(t, sx) ds \in \mathcal{U}.$$

We see that every trajectory of the original nonlinear system (NLS) is also a trajectory of the uncertain linear system (ULS) associated with the uncertainty set $\mathcal{U} = \mathcal{U}_f$ (this trick is called "global linearization"). Of course, the set of trajectories of the resulting uncertain linear system can be much wider than the set of trajectories of (NLS); however, all "good news" about the uncertain system (like "all trajectories of (ULS) share such and such property") are automatically valid for the trajectories of the "nonlinear system of interest" (NLS), and only "bad news" about (ULS) ("such and such property is not shared by some trajectories of (ULS)") may say nothing about the system of interest (NLS).

Stability and stability certificates. The basic question about a dynamic system is the one of its stability. For (ULS), this question sounds as follows:

(?) *Is it true that* (ULS) *is stable, i.e., that*

$$x(t) \to 0 \text{ as } t \to \infty$$

for every trajectory of the system?

A sufficient condition for the stability of (ULS) is the existence of a *quadratic Lyapunov function*, i.e., a quadratic form $\mathcal{L}(x) = x^T X x$ with symmetric positive definite matrix X such that

$$\frac{d}{dt}\mathcal{L}(x(t)) \leq -\alpha\mathcal{L}(x(t)) \tag{3.3.7}$$

for certain $\alpha > 0$ and all trajectories of (ULS):

Lemma 3.3.1 [Quadratic Stability Certificate] *Assume* (ULS) *admits a quadratic Lyapunov function* $\mathcal{L}$. *Then* (ULS) *is stable.*

Proof. If (3.3.7) is valid with some $\alpha > 0$ for all trajectories of (ULS), then, by integrating this differential inequality, we get

$$\mathcal{L}(x(t)) \leq \exp\{-\alpha \mathcal{L}(x(0))\} \to 0 \text{ as } t \to \infty.$$

Since $\mathcal{L}(\cdot)$ is a positive definite quadratic form, $\mathcal{L}(x(t)) \to 0$ implies that $x(t) \to 0$. ∎

Of course, the statement of Lemma 3.3.1 also holds for non-quadratic Lyapunov functions: all we need is (3.3.7) plus the assumption that $\mathcal{L}(x)$ is smooth, nonnegative and is bounded away from 0 outside every neighbourhood of the origin. The advantage of a *quadratic* Lyapunov function is that we more or less know how to find such a function, if it exists:

Proposition 3.3.1 [Existence of Quadratic Stability Certificate] *Let* $\mathcal{U}$ *be the uncertainty set associated with uncertain linear system* (ULS). *The system admits quadratic Lyapunov function if and only if the optimal value of the "semi-infinite*[9)] *semidefinite program"*

$$\begin{array}{lrcl} \textit{minimize} & s & & \\ \textit{s.t.} & & & \\ & sI_n - A^T X - XA & \succeq & 0, \quad \forall A \in \mathcal{U} \\ & X & \succeq & I_n \end{array} \tag{Ly}$$

with the design variables $s \in \mathbf{R}$ *and* $X \in \mathbf{S}^n$, *is negative. Moreover, every feasible solution to the problem with negative value of the objective provides a quadratic Lyapunov stability certificate for* (ULS).

We shall refer to a positive definite matrix $X \succeq I_n$ which can be extended, by properly chosen $s < 0$, to a feasible solution of (Ly), as to a *Lyapunov stability certificate* for (ULS), the uncertainty set being $\mathcal{U}$.
Proof of Proposition 3.3.1. The derivative $\frac{d}{dt}\left[x^T(t)Xx(t)\right]$ of the quadratic function x^TXx along a trajectory of (ULS) is equal to

$$\left[\frac{d}{dt}x(t)\right]^T Xx(t) + x^T(t)X\left[\frac{d}{dt}x(t)\right] = x^T(t)[A^T(t)X + XA(t)]x(t).$$

If x^TXx is a Lyapunov function, then the resulting quantity must be at most $-\alpha x^T(t)Xx(t)$, i.e., we should have

$$x^T(t)\left[-\alpha X - A^T(t)X - XA(t)\right]x(t) \geq 0$$

for every possible value of $A(t)$ at any time t and for every possible value $x(t)$ of a trajectory of the system at this time. Since possible values of $x(t)$ fill the entire $\mathbf{R}^n$ and possible values of $A(t)$ fill the entire $\mathcal{U}$, we conclude that

$$-\alpha X - A^T X - XA \succeq 0 \quad \forall A \in \mathcal{U}.$$

By definition of a quadratic Lyapunov function, $X \succ 0$ and $\alpha > 0$; by normalization (dividing both X and α by the smallest eigenvalue of X), we get a pair $\hat{s} > 0, \hat{X} \geq I_n$ such that

$$-\hat{s}\hat{X} - A^T\hat{X} - \hat{X}A \succeq 0 \quad \forall A \in \mathcal{U}.$$

Since $\hat{X} \succeq I_n$, we conclude that

$$-\hat{s}I_n - A^T\hat{X} - \hat{X}A \succeq \hat{s}\hat{X} - A^T\hat{X} - \hat{X}A \succeq 0 \quad \forall A \in \mathcal{U};$$

thus, $(s = -\hat{s}, \hat{X})$ is a feasible solution to (Ly) with negative value of the objective. We have demonstrated that if (ULS) admits a quadratic Lyapunov function, then (Ly) has a feasible solution with negative value of the objective. Reversing the reasoning, we can verify the inverse implication. ∎

[9)] i.e., with infinitely many LMI constraints

Lyapunov stability analysis. According to Proposition 3.3.1, the existence of a Lyapunov stability certificate is a *sufficient*, but, in general, not a necessary stability condition for (ULS). When the condition is not satisfied (i.e., if the optimal value in (Ly) is nonnegative), then all we can say is that *the stability of* (ULS) *cannot be certified by a quadratic Lyapunov function*, although (ULS) still may be stable.[10)] In this sense, the stability analysis based on quadratic Lyapunov functions is conservative. This drawback, however, is in a sense compensated by the fact that this kind of stability analysis is "implementable": in many cases we can efficiently solve (Ly), thus getting a quadratic "stability certificate", provided that it exists, in a constructive way. Let us look at two such cases.

Polytopic uncertainty set. The first "tractable case" of (Ly) is when $\mathcal{U}$ is a polytope *given as a convex hull of finitely many points*:

$$\mathcal{U} = \mathrm{Conv}\{A_1, ..., A_N\}.$$

In this case (Ly) is equivalent to the semidefinite program

$$\min_{s,X} \left\{ s : sI_n - A_i^T X - X A_i \succeq 0,\ i = 1, ..., N; X \succeq I_n \right\} \tag{3.3.8}$$

(why?).

The assumption that $\mathcal{U}$ is a polytope *given as a convex hull of a finite set* is crucial for a possibility to get a "computationally tractable" equivalent reformulation of (Ly). If $\mathcal{U}$ is, say, a polytope *given by a list of linear inequalities* (e.g., all we know about the entries of $A(t)$ is that they reside in certain intervals; this case is called "interval uncertainty"), (Ly) may become as hard as a problem can be: it may happen that just to check whether a *given* pair (s, X) is feasible for (Ly) is already a "computationally intractable" problem. The same difficulties may occur when $\mathcal{U}$ is a general-type ellipsoid in the space of $n \times n$ matrices. There exists, however, a specific type of "uncertainty ellipsoids" $\mathcal{U}$ for which (Ly) is "easy". Let us look at this case.

Norm-bounded perturbations. In numerous applications the $n \times n$ matrices A forming the uncertainty set $\mathcal{U}$ are obtained from a fixed "nominal" matrix A_* by adding perturbations of the form $B\Delta C$, where $B \in \mathbf{M}^{n,k}$ and $C \in \mathbf{M}^{l,n}$ are given rectangular matrices and $\Delta \in \mathbf{M}^{k,l}$ is "the perturbation" varying in a "simple" set $\mathcal{D}$:

$$\mathcal{U} = \{A = A_* + B\Delta C \mid \Delta \in \mathcal{D} \subset \mathbf{M}^{k,l}\} \quad \left[B \in \mathbf{M}^{n,k}, 0 \neq C \in \mathbf{M}^{l,n}\right] \tag{3.3.9}$$

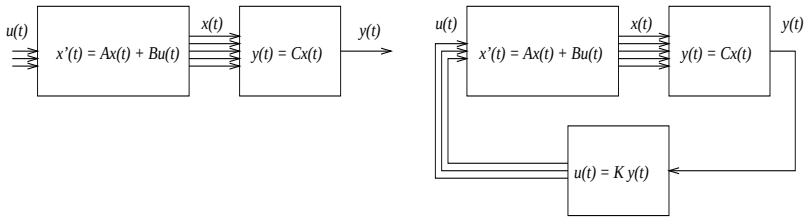
As an instructive example, consider a *controlled* linear time-invariant dynamic system

$$\begin{array}{rcl} \frac{d}{dt}x(t) & = & Ax(t) + Bu(t) \\ y(t) & = & Cx(t) \end{array} \tag{3.3.10}$$

(x is the state, u is the control and y is the output we can observe) "closed" by a feedback

$$u(t) = Ky(t).$$

[10)] The only case when the existence of a quadratic Lyapunov function is a criterion (i.e., a *necessary and sufficient* condition) for stability is the simplest case of certain time-invariant linear system $\frac{d}{dt}x(t) = Ax(t)$ ($\mathcal{U} = \{A\}$). This is the case which led Lyapunov to the general concept of what is now called "a Lyapunov function" and what is the basic approach to establishing convergence of different time-dependent processes to their equilibria. Note also that in the case of time-invariant linear system there exists a straightforward algebraic stability criterion – all eigenvalues of A should have negative real parts. The advantage of the Lyapunov approach is that it can be extended to more general situations, which is not the case for the eigenvalue criterion.



Open loop (left) and closed loop (right) controlled systems

The resulting "closed loop system" is given by

$$\frac{d}{dt}x(t) = \hat{A}x(t), \quad \hat{A} = A + BKC. \tag{3.3.11}$$

Now assume that A, B and C are constant and known, but the feedback K is drifting around certain nominal feedback K^*: $K = K^* + \Delta$. As a result, the matrix $\hat{A}$ of the closed loop system also drifts around its nominal value $A^* = A + BK^*C$, and the perturbations in $\hat{A}$ are exactly of the form $B\Delta C$.

Note that we could get essentially the same kind of drift in $\hat{A}$ assuming, instead of additive perturbations, multiplicative perturbations $C = (I_l + \Delta)C^*$ in the observer (or multiplicative disturbances in the actuator B).

Now assume that the input perturbations Δ are of spectral norm $|\Delta|$ not exceeding a given ρ (*norm-bounded perturbations*):

$$\mathcal{D} = \{\Delta \in \mathbf{M}^{k,l} \mid |\Delta| \le \rho\}. \tag{3.3.12}$$

Proposition 3.3.2 [5] *In the case of uncertainty set* (3.3.9), (3.3.12) *the "semi-infinite" semidefinite program* (Ly) *is equivalent to the usual semidefinite program*

$$\begin{array}{rcl} & & \text{minimize} \quad \alpha \\ & & \text{s.t.} \\ \begin{pmatrix} \alpha I_n - A_*^T X - X A_* - \lambda C^T C & \rho X B \\ \rho B^T X & \lambda I_k \end{pmatrix} & \succeq & 0 \\ X & \succeq & I_n \end{array} \tag{3.3.13}$$

in the design variables α, λ, X.

When shrinking the set of perturbations (3.3.12) *to the ellipsoid*

$$\mathcal{E} = \{\Delta \in \mathbf{M}^{k,l} \mid \|\Delta\|_2 \equiv \sqrt{\sum_{i=1}^{k}\sum_{j=1}^{l}\Delta_{ij}^2} \le \rho\}, \quad {}^{11)} \tag{3.3.14}$$

we basically do not vary (Ly)*: in the case of the uncertainty set* (3.3.9), (Ly) *is still equivalent to* (3.3.13).

Proof. It suffices to verify the following general statement:

Lemma 3.3.2 *Consider the matrix inequality*

$$Y - Q^T\Delta^T P^T Z^T R - R^T Z P \Delta Q \succeq 0 \tag{3.3.15}$$

where Y is symmetric $n \times n$ matrix, Δ is a $k \times l$ matrix and P, Q, Z, R are rectangular matrices of appropriate sizes (i.e., $q \times k$, $l \times n$, $p \times q$ and $p \times n$, respectively). Given Y, P, Q, Z, R, with $Q \neq 0$ (this is the only nontrivial case), this matrix inequality is satisfied

[11] This indeed is a "shrinkage": $|\Delta| \le \|\Delta\|_2$ for every matrix Δ (prove it!)

for all Δ with $|\Delta| \le \rho$ if and only if it is satisfied for all Δ with $\|\Delta\|_2 \le \rho$, and this is the case if and only if

$$\begin{pmatrix} Y - \lambda Q^T Q & -\rho R^T Z P \\ -\rho P^T Z^T R & \lambda I_k \end{pmatrix} \succeq 0$$

for a properly chosen real λ.

The statement of Proposition 3.4.14 is just a particular case of Lemma 3.3.2. For example, in the case of uncertainty set (3.3.9), (3.3.12) a pair (α, X) is a feasible solution to (Ly) if and only if $X \succeq I_n$ and (3.3.15) is valid for $Y = \alpha X - A_*^T X - X A_*$, $P = B$, $Q = C$, $Z = X$, $R = I_n$; Lemma 3.3.2 provides us with an LMI reformulation of the latter property, and this LMI is exactly what we see in the statement of Proposition 3.4.14.

Proof of Lemma. (3.3.15) is valid for all Δ with $|\Delta| \le \rho$ (let us call this property of (Y, P, Q, Z, R) "Property 1") if and only if

$$2[\xi^T R^T Z P]\Delta[Q\xi] \le \xi^T Y \xi \quad \forall \xi \in \mathbf{R}^n \quad \forall (\Delta : |\Delta| \le \rho),$$

or, which is the same, if and only if

$$\max_{\Delta : |\Delta| \le \rho} 2\left[[P^T Z^T R\xi]^T \Delta [Q\xi]\right] \le \xi^T Y \xi \quad \forall \xi \in \mathbf{R}^n. \tag{Property 2}$$

The maximum over Δ, $|\Delta| \le \rho$, of the quantity $\eta^T \Delta \zeta$, clearly is equal to ρ times the product of the Euclidean norms of the vectors η and ζ (why?). Thus, Property 2 is equivalent to

$$\xi^T Y \xi - 2\rho \|Q\xi\|_2 \|P^T Z^T R\xi\|_2 \ge 0 \quad \forall \xi \in \mathbf{R}^n. \tag{Property 3}$$

Now is the trick: Property 3 is clearly equivalent to the following

Property 4: *Every pair $\zeta = (\xi, \eta) \in \mathbf{R}^n \times \mathbf{R}^k$ which satisfies the quadratic inequality*

$$\xi^T Q^T Q \xi - \eta^T \eta \ge 0 \tag{I}$$

satisfies also the quadratic inequality

$$\xi^T Y \xi - 2\rho \eta^T P^T Z^T R \xi \ge 0. \tag{II}$$

Indeed, for a fixed ξ the minimum over η satisfying (I) of the left hand side in (II) is nothing but the left hand side in (Property 3).

It remains to use the $\mathcal{S}$-Lemma:

$\mathcal{S}$-Lemma. *Let A, B be symmetric $n \times n$ matrices, and assume that the quadratic inequality*

$$x^T A x \ge 0 \tag{A}$$

is strictly feasible: there exists $\bar{x}$ such that $\bar{x}^T A \bar{x} > 0$. Then the quadratic inequality

$$x^T B x \ge 0 \tag{B}$$

is a consequence of (A) *if and only if it is a linear consequence of* (A)*, i.e., if and only if there exists a nonnegative λ such that*

$$B \succeq \lambda A.$$

(for a proof, see Section 3.5). Property 4 says that the quadratic inequality (II) with variables ξ, η is a consequence of (I); by the $\mathcal{S}$-Lemma (recall that $Q \neq 0$, so that (I) is strictly feasible!) this is equivalent to the existence of a nonnegative λ such that

$$\begin{pmatrix} Y & -\rho R^T Z P \\ -\rho P^T Z^T R & \end{pmatrix} - \lambda \begin{pmatrix} Q^T Q & \\ & -I_k \end{pmatrix} \succeq 0,$$

which is exactly the statement of Lemma 3.3.2 for the case of $|\Delta| \le \rho$. The case of perturbations with $\|\Delta\|_2 \le \rho$ is completely similar, since the equivalence between Properties 2 and 3 is valid independently of which norm of Δ – $|\cdot|$ or $\|\cdot\|_2$ – is used.

Lyapunov Stability Synthesis. We have seen that under reasonable assumptions on the underlying uncertainty set the question of whether a given uncertain linear system (ULS) admits a quadratic Lyapunov function can be reduced to a semidefinite program. Now let us switch from the *analysis* question: "whether a stability of an uncertain linear system may be certified by a quadratic Lyapunov function" to the *synthesis* question which is as follows. Assume that we are given an *uncertain open loop controlled system*

$$\begin{array}{rcl} \frac{d}{dt}x(t) & = & A(t)x(t) + B(t)u(t) \\ y(t) & = & C(t)x(t); \end{array} \qquad \text{(UOS)}$$

all we know about the collection $(A(t), B(t), C(t))$ of time-varying $n \times n$ matrix $A(t)$, $n \times k$ matrix $B(t)$ and $l \times n$ matrix $C(t)$ is that this collection, at every time t, belongs to a given uncertainty set $\mathcal{U}$. The question is whether we can equip our uncertain "open loop" system (UOS) with a linear feedback

$$u(t) = Ky(t)$$

in such a way that the resulting uncertain closed loop system

$$\frac{d}{dt}x(t) = [A(t) + B(t)KC(t)]\, x(t) \qquad \text{(UCS)}$$

will be stable and, moreover, such that its stability can be certified by a quadratic Lyapunov function. In other words, now we are simultaneously looking for a "stabilizing controller" and a quadratic Lyapunov certificate of its stabilizing ability.

> With the "global linearization" trick we may use the results on uncertain controlled linear systems to build stabilizing linear controllers for *nonlinear* controlled systems
>
> $$\begin{array}{rcl} \frac{d}{dt}x(t) & = & f(t, x(t), u(t)) \\ y(t) & = & g(t, x(t)) \end{array}$$
>
> Assuming $f(t, 0, 0) = 0$, $g(t, 0) = 0$ and denoting by $\mathcal{U}$ the closed convex hull of the set
>
> $$\left\{ \left(\frac{\partial}{\partial x} f(t, x, u), \frac{\partial}{\partial u} f(t, x, u), \frac{\partial}{\partial x} g(t, x) \right) \middle| t \geq 0, x \in \mathbf{R}^n, u \in \mathbf{R}^k \right\},$$
>
> we see that every trajectory of the original nonlinear system is a trajectory of the uncertain linear system (UOS) associated with the set $\mathcal{U}$. Consequently, if we are able to find a stabilizing controller for (UOS) and certify its stabilizing property by a quadratic Lyapunov function, then the resulting controller/Lyapunov function will stabilize the nonlinear system and will certify the stability of the closed loop system, respectively.

Exactly the same reasoning as in the previous section leads us to the following

Proposition 3.3.3 *Let $\mathcal{U}$ be the uncertainty set associated with an uncertain open loop controlled system* (UOS)*. The system admits a stabilizing controller along with a quadratic Lyapunov stability certificate for the resulting closed loop system if and only if the optimal value in the optimization problem*

$$\begin{array}{ll} \text{minimize} & s \\ \text{s.t.} & \\ & [A + BKC]^T X + X[A + BKC] \preceq sI_n \quad \forall (A, B, C) \in \mathcal{U} \\ & X \succeq I_n, \end{array} \qquad \text{(LyS)}$$

in design variables s, X, K, is negative. Moreover, every feasible solution to the problem with negative value of the objective provides stabilizing controller along with quadratic Lyapunov stability certificate for the resulting closed loop system.

A bad news about (LyS) is that it is much more difficult to rewrite this problem as a semidefinite program than in the analysis case (i.e., the case of $K = 0$), since (LyS) is a semi-infinite system of *nonlinear* matrix

inequalities. There is, however, an important particular case where this difficulty can be eliminated. This is the case of a feedback via the *full* state vector – the case when $y(t) = x(t)$ (i.e., $C(t)$ is the unit matrix). In this case, all we need in order to get a stabilizing controller along with a quadratic Lyapunov certificate of its stabilizing ability, is to solve a system of *strict* matrix inequalities

$$\begin{array}{rcl} [A+BK]^T X + X[A+BK] & \preceq & Z \prec 0 \quad \forall (A,B) \in \mathcal{U} \\ X & \succ & 0 \end{array} . \qquad (*)$$

Indeed, given a solution (X, K, Z) to this system, we always can convert it by normalization of X to a solution of (LyS). Now let us make the change of variables

$$Y = X^{-1}, L = KX^{-1}, W = X^{-1}ZX^{-1} \quad \left[\Leftrightarrow X = Y^{-1}, K = LY^{-1}, Z = Y^{-1}WY^{-1}\right].$$

With respect to the new variables Y, L, K system (*) becomes

$$\left\{ \begin{array}{rcl} [A+BLY^{-1}]^T Y^{-1} + Y^{-1}[A+BLY^{-1}] & \preceq & Y^{-1}WY^{-1} \prec 0 \\ Y^{-1} & \succ & 0 \end{array} \right.$$

$$\Updownarrow$$

$$\left\{ \begin{array}{rcl} L^T B^T + YA^T + BL + AY & \preceq & W \prec 0, \quad \forall (A,B) \in \mathcal{U} \\ Y & \succ & 0 \end{array} \right.$$

(we have multiplied all original matrix inequalities from the left and from the right by Y). What we end up with, is a system of strict *linear* matrix inequalities with respect to our new design variables L, Y, W; the question of whether this system is solvable can be converted to the question of whether the optimal value in a problem of the type (LyS) is negative, and we come to the following

Proposition 3.3.4 *Consider an uncertain controlled linear system with a full observer:*

$$\begin{array}{rcl} \frac{d}{dt}x(t) & = & A(t)x(t) + B(t)u(t) \\ y(t) & = & x(t) \end{array}$$

and let $\mathcal{U}$ be the corresponding uncertainty set (which now is comprised of pairs (A, B) of possible values of $(A(t), B(t))$, since $C(t) \equiv I_n$ is certain).

The system can be stabilized by a linear controller

$$u(t) = Ky(t) \quad [\equiv Kx(t)]$$

in such a way that the resulting uncertain closed loop system

$$\frac{d}{dt}x(t) = [A(t) + B(t)K]x(t)$$

admits a quadratic Lyapunov stability certificate if and only if the optimal value in the optimization problem

$$\begin{array}{lrcll} \text{minimize} & s & & & \\ \text{s.t.} & & & & \\ & BL + AY + L^T B^T + YA^T & \preceq & sI_n \quad \forall (A,B) \in \mathcal{U} \\ & Y & \succeq & I \end{array} \qquad (\text{Ly}^*)$$

in the design variables $s \in \mathbf{R}$, $Y \in \mathbf{S}^n$, $L \in \mathbf{M}^{k,n}$, is negative. Moreover, every feasible solution to (Ly*) *with negative value of the objective provides a stabilizing linear controller along with related quadratic Lyapunov stability certificate.*

In particular, in the polytopic case:

$$\mathcal{U} = \text{Conv}\{(A_1, B_1), ..., (A_N, B_N)\}$$

the Quadratic Lyapunov Stability Synthesis reduces to solving the semidefinite program

$$\min_{s,Y,L} \left\{ s : B_i L + A_i Y + YA_i^T + L^T B_i^T \preceq sI_n, \ i = 1, ..., N; Y \succeq I_n \right\}.$$

3.4 Semidefinite relaxations of intractable problems

One of the most challenging and promising applications of Semidefinite Programming is in building tractable approximations of "computationally intractable" optimization problems. Let us look at several applications of this type.

3.4.1 Semidefinite relaxations of combinatorial problems

Combinatorial problems and their relaxations. Numerous problems of planning, scheduling, routing, etc., can be posed as *combinatorial optimization problems*, i.e., optimization programs with discrete design variables (integer or zero-one). There are several "universal forms" of combinatorial problems, among them Linear Programming with integer variables and Linear Programming with 0-1 variables; a problem given in one of these forms can always be converted to any other universal form, so that in principle it does not matter which form to use. Now, the majority of combinatorial problems are difficult – we do not know theoretically efficient (in certain precise meaning of the notion) algorithms for solving these problems. What we do know is that nearly all these difficult problems are, in a sense, equivalent to each other and are *NP-complete*. The exact meaning of the latter notion will be explained in Lecture 4; for the time being it suffices to say that NP-completeness of a problem P means that the problem is "as difficult as a combinatorial problem can be" – if we knew an efficient algorithm for P, we would be able to convert it to an efficient algorithm for *any other* combinatorial problem. NP-complete problems may look extremely "simple", as it is demonstrated by the following example:

> (Stones) *Given n stones of positive integer weights (i.e., given n positive integers $a_1, ..., a_n$), check whether you can partition these stones into two groups of equal weight, i.e., check whether a linear equation*
>
> $$\sum_{i=1}^{n} a_i x_i = 0$$
>
> *has a solution with $x_i = \pm 1$.*

Theoretically difficult combinatorial problems happen to be difficult to solve in practice as well. An important ingredient in basically all algorithms for combinatorial optimization is a technique for building bounds for the unknown optimal value of a given (sub)problem. A typical way to estimate the optimal value of an optimization program

$$f^* = \min_x \{f(x) : x \in X\}$$

from above is to present a feasible solution $\bar{x}$; then clearly $f^* \leq f(\bar{x})$. And a typical way to bound the optimal value from below is to pass from the problem to its *relaxation*

$$f_* = \min_x \{f(x) : x \in X'\}$$

increasing the feasible set: $X \subset X'$. Clearly, $f_* \leq f^*$, so, whenever the relaxation is efficiently solvable (to ensure this, we should take care of how we choose X'), it provides us with a "computable" lower bound on the actual optimal value.

When building a relaxation, one should take care of two issues: on one hand, we want the relaxation to be "efficiently solvable". On the other hand, we want the relaxation to be "tight", otherwise the lower bound we get may be by far "too optimistic" and therefore not useful. For a long time, the only practical relaxations were the LP ones, since these were the only problems one could solve efficiently. With recent progress in optimization techniques, nonlinear relaxations become more and more "practical"; as a result, we are witnessing a growing theoretical and computational activity in the area of nonlinear relaxations of combinatorial problems. These developments mostly deal with *semidefinite* relaxations. Let us look how they emerge.

Shor's Semidefinite Relaxation scheme

As it was already mentioned, there are numerous "universal forms" of combinatorial problems. E.g., a combinatorial problem can be posed as minimizing a quadratic objective under quadratic inequality constraints:

$$\begin{array}{ll} \text{minimize in } x\in\mathbf{R}^n & f_0(x)=x^TA_0x+2b_0^Tx+c_0 \\ \text{s.t.} & \\ & f_i(x)=x^TA_ix+2b_i^Tx+c_i \quad \leq \quad 0,\ i=1,...,m. \end{array} \tag{3.4.1}$$

To see that this form is "universal", note that it covers the classical universal combinatorial problem – a generic LP program with Boolean (0-1) variables:

$$\min_x \left\{c^Tx : a_i^Tx-b_i\leq 0,\ i=1,...,m; x_j\in\{0,1\},\ j=1,...,n\right\} \tag{B}$$

Indeed, the fact that a variable x_j must be Boolean can be expressed by the quadratic equality

$$x_j^2-x_j=0,$$

which can be represented by a pair of opposite quadratic inequalities and a linear inequality $a_i^Tx-b_i\leq 0$ is a particular case of quadratic inequality. Thus, (B) is equivalent to the problem

$$\min_{x,s} \left\{c^Tx : a_i^Tx-b_i\leq 0,\ i=1,...,m; x_j^2-x_j\leq 0, -x_j^2+x_j\leq 0\ j=1,...,n\right\},$$

and this problem is of the form (3.4.1).

To bound from below the optimal value in (3.4.1), we may use the same technique we used for building the dual problem (it is called the *Lagrange relaxation*). We choose somehow "weights" $\lambda_i\geq 0$, $i=1,...,m$, and add the constraints of (3.4.1) with these weights to the objective, thus coming to the function

$$\begin{array}{rcl} f_\lambda(x) & = & f_0(x)+\sum\limits_{i=1}^m \lambda_i f_i(x) \\ & = & x^TA(\lambda)x+2b^T(\lambda)x+c(\lambda), \end{array} \tag{3.4.2}$$

where

$$\begin{array}{rcl} A(\lambda) & = & A_0+\sum\limits_{i=1}^m \lambda_iA_i \\ b(\lambda) & = & b_0+\sum\limits_{i=1}^m \lambda_ib_i \\ c(\lambda) & = & c_0+\sum\limits_{i=1}^m \lambda_ic_i \end{array}$$

By construction, the function $f_\lambda(x)$ is $\leq$ the actual objective $f_0(x)$ on the feasible set of the problem (3.4.1). Consequently, the unconstrained infimum of this function

$$a(\lambda)=\inf_{x\in\mathbf{R}^n} f_\lambda(x)$$

is a lower bound for the optimal value in (3.4.1). We come to the following simple result (cf. the Weak Duality Theorem:)

(*) *Assume that $\lambda\in\mathbf{R}^m_+$ and $\zeta\in\mathbf{R}$ are such that*

$$f_\lambda(x)-\zeta\geq 0 \quad \forall x\in\mathbf{R}^n \tag{3.4.3}$$

(i.e., that $\zeta\leq a(\lambda)$). Then ζ is a lower bound for the optimal value in (3.4.1).

It remains to clarify what does it mean that (3.4.3) holds. Recalling the structure of f_λ, we see that it means that the inhomogeneous quadratic form

$$g_\lambda(x)=x^TA(\lambda)x+2b^T(\lambda)x+c(\lambda)-\zeta$$

is nonnegative on the entire space. Now, an inhomogeneous quadratic form

$$g(x) = x^T Ax + 2b^T x + c$$

is nonnegative everywhere if and only if certain associated *homogeneous* quadratic form is nonnegative everywhere. Indeed, given $t \neq 0$ and $x \in \mathbf{R}^n$, the fact that $g(t^{-1}x) \geq 0$ means exactly the nonnegativity of the *homogeneous* quadratic form $G(x,t)$

$$G(x,t) = x^T Ax + 2tb^T x + ct^2$$

with $(n+1)$ variables x, t. We see that if g is nonnegative, then G is nonnegative whenever $t \neq 0$; by continuity, G then is nonnegative everywhere. Thus, if g is nonnegative, then G is, and of course vice versa (since $g(x) = G(x,1)$). Now, to say that G is nonnegative everywhere is literally the same as to say that the matrix

$$\begin{pmatrix} c & b^T \\ b & A \end{pmatrix} \tag{3.4.4}$$

is positive semidefinite.

It is worthy to catalogue our simple observation:

Simple Lemma. *A quadratic inequality with a (symmetric) $n \times n$ matrix A*

$$x^T Ax + 2b^T x + c \geq 0$$

is identically true – is valid for all $x \in \mathbf{R}^n$ – if only if the matrix (3.4.4) *is positive semidefinite.*

Applying this observation to $g_\lambda(x)$, we get the following equivalent reformulation of (*):

If $(\lambda, \zeta) \in \mathbf{R}^m_+ \times \mathbf{R}$ satisfy the LMI

$$\begin{pmatrix} \sum\limits_{i=1}^m \lambda_i c_i - \zeta & b_0^T + \sum\limits_{i=1}^m \lambda_i b_i^T \\ b_0 + \sum\limits_{i=1}^m \lambda_i b_i & A_0 + \sum\limits_{i=1}^m \lambda_i A_i \end{pmatrix} \succeq 0,$$

then ζ is a lower bound for the optimal value in (3.4.1).

Now, what is the best lower bound we can get with this scheme? Of course, it is the optimal value of the semidefinite program

$$\max_{\zeta,\lambda} \left\{ \zeta : \begin{pmatrix} c_0 + \sum\limits_{i=1}^m \lambda_i c_i - \zeta & b_0^T + \sum\limits_{i=1}^m \lambda_i b_i^T \\ b_0 + \sum\limits_{i=1}^m \lambda_i b_i & A_0 + \sum\limits_{i=1}^m \lambda_i A_i \end{pmatrix} \succeq 0, \lambda \geq 0 \right\}. \tag{3.4.5}$$

We have proved the following simple

Proposition 3.4.1 *The optimal value in* (3.4.5) *is a lower bound for the optimal value in* (3.4.1).

The outlined scheme is extremely transparent, but it looks different from a relaxation scheme as explained above – where is the extension of the feasible set of the original problem? In fact the scheme is of this type. To see it, note that the value of a quadratic form at a point $x \in \mathbf{R}^n$ can be written as the Frobenius inner product of a matrix defined by the problem data and the dyadic matrix $X(x) = \begin{pmatrix} 1 \\ x \end{pmatrix} \begin{pmatrix} 1 \\ x \end{pmatrix}^T$:

$$x^T Ax + 2b^T x + c = \begin{pmatrix} 1 \\ x \end{pmatrix}^T \begin{pmatrix} c & b^T \\ b & A \end{pmatrix} \begin{pmatrix} 1 \\ x \end{pmatrix} = \mathrm{Tr}\left(\begin{pmatrix} c & b^T \\ b & A \end{pmatrix} X(x) \right).$$

Consequently, (3.4.1) can be written as

$$\min_x \left\{ \mathrm{Tr}\left(\begin{pmatrix} c_0 & b_0^T \\ b_0 & A_0 \end{pmatrix} X(x) \right) : \mathrm{Tr}\left(\begin{pmatrix} c_i & b_i^T \\ b_i & A_i \end{pmatrix} X(x) \right) \le 0,\ i = 1, ..., m \right\}. \tag{3.4.6}$$

Thus, we may think of (3.4.2) as a problem with *linear* objective and *linear* equality constraints and with the design vector X which is a symmetric $(n+1) \times (n+1)$ matrix *running through the nonlinear manifold $\mathcal{X}$ of dyadic matrices $X(x)$, $x \in \mathbf{R}^n$*. Clearly, all points of $\mathcal{X}$ are positive semidefinite matrices with North-Western entry 1. Now let $\bar{\mathcal{X}}$ be the set of all such matrices. Replacing $\mathcal{X}$ by $\bar{\mathcal{X}}$, we get a relaxation of (3.4.6) (the latter problem is, essentially, our original problem (3.4.1)). This relaxation is the semidefinite program

$$\begin{array}{c} \min_X \left\{ \mathrm{Tr}(\bar{A}_0 X) : \mathrm{Tr}(\bar{A}_i X) \le 0, i = 1, ..., m; X \succeq 0; X_{11} = 1 \right\} \\ \left[A_i = \begin{pmatrix} c_i & b_i^T \\ b_i & A_i \end{pmatrix},\ i = 1, ..., m \right]. \end{array} \tag{3.4.7}$$

We get the following

Proposition 3.4.2 *The optimal value of the semidefinite program* (3.4.7) *is a lower bound for the optimal value in* (3.4.1).

One can easily verify that problem (3.4.5) is just the semidefinite dual of (3.4.7); thus, when deriving (3.4.5), we were in fact implementing the idea of relaxation. This is why in the sequel we call both (3.4.7) and (3.4.5) *semidefinite relaxations* of (3.4.1).

When the semidefinite relaxation is exact? In general, the optimal value in (3.4.7) is just a lower bound on the optimal value of (3.4.1). There are, however, two cases when this bound is exact. These are

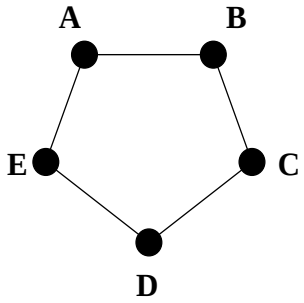
- *Convex case,* where all quadratic forms in (3.4.1) are convex (i.e., $Q_i \succeq 0$, $i = 0, 1, ..., m$). Here, strict feasibility of the problem (i.e., the existence of a feasible solution $\bar{x}$ with $f_i(\bar{x}) < 0$, $i = 1, ..., m$) plus below boundedness of it imply that (3.4.7) is solvable with the optimal value equal to the one of (3.4.1). This statement is a particular case of the well-known *Lagrange Duality Theorem* in Convex Programming.

- The case of $m = 1$. Here the optimal value in (3.4.1) is equal to the one in (3.4.7), provided that (3.4.1) is strictly feasible. This highly surprising fact (no convexity is assumed!) is called *Inhomogeneous $\mathcal{S}$-Lemma*; we shall prove it in Section .

Let us look at several interesting examples of Semidefinite relaxations.

Stability number, Shannon capacity and Lovasz capacity of a graph

Stability number of a graph. Consider a (non-oriented) graph – a finite set of *nodes* linked by arcs[12)], like the simple 5-node graph C_5:



Graph C_5

One of the fundamental characteristics of a graph Γ is its *stability number* $\alpha(\Gamma)$ defined as the maximum cardinality of an *independent* subset of nodes – a subset such that no two nodes from it are linked by an arc. E.g., the stability number for the graph C_5 is 2, and a maximal independent set is, e.g., $\{\mathrm{A};\mathrm{C}\}$.

The problem of computing the stability number of a given graph is NP-complete, this is why it is important to know how to bound this number.

Shannon capacity of a graph. An upper bound on the stability number of a graph which interesting by its own right is the *Shannon capacity* $\Theta(\Gamma)$ defined as follows.

Let us treat the nodes of Γ as letters of certain alphabet, and the arcs as possible errors in certain communication channel: you can send trough the channel one letter per unit time, and what arrives on the other end of the channel can be either the letter you have sent, or any letter adjacent to it. Now assume that you are planning to communicate with an addressee through the channel by sending n-letter words (n is fixed). You fix in advance a dictionary D_n of words to be used and make this dictionary known to the addressee. What you are interested in when building the dictionary is to get a *good* one, meaning that no word from it could be transformed by the channel into another word from the dictionary. If your dictionary satisfies this requirement, you may be sure that the addressee will never misunderstand you: whatever word from the dictionary you send and whatever possible transmission errors occur, the addressee is able either to get the correct message, or to realize that the message was corrupted during transmission, but there is no risk that your "yes" will be read as "no!". Now, in order to utilize the channel "at full capacity", you are interested to get as large dictionary as possible. How many words it can include? The answer is clear: this is precisely the stability number of the graph Γ^n as follows: the nodes of Γ^n are ordered n-element collections of the nodes of Γ – all possible n-letter words in your alphabet; two distinct nodes $(i_1,...,i_n)$ $(j_1,...,j_n)$ are adjacent in Γ^n if and only if for every l the l-th letters i_l and j_l in the two words either coincide, or are adjacent in Γ (i.e., two distinct n-letter words are adjacent, if the transmission can convert one of them into the other one). Let us denote the maximum number of words in a "good" dictionary D_n (i.e., the stability number of Γ^n) by $f(n)$, The function $f(n)$ possesses the following nice property:

$$f(k)f(l) \le f(k+l),\ k,l = 1,2,... \qquad (*)$$

Indeed, given the best (of the cardinality $f(k)$) good dictionary D_k and the best good dictionary D_l, let us build a dictionary comprised of all $(k+l)$-letter words as follows: the initial k-letter fragment of a word belongs to D_k, and the remaining l-letter fragment belongs to D_l. The resulting dictionary is clearly good and contains $f(k)f(l)$ words, and (*) follows.

[12)] One of the formal definitions of a (non-oriented) graph is as follows: a n-node graph is just a $n\times n$ symmetric matrix A with entries $0,1$ and zero diagonal. The rows (and the columns) of the matrix are identified with the nodes $1,2,...,n$ of the graph, and the nodes i,j are adjacent (i.e., linked by an arc) exactly for those i,j with $A_{ij}=1$.

Now, this is a simple exercise in analysis to see that for a nonnegative function f with property (*) one has

$$\lim_{k\to\infty}(f(k))^{1/k} = \sup_{k\geq 1}(f(k))^{1/k} \in [0,+\infty].$$

In our situation $\sup_{k\geq 1}(f(k))^{1/k} < \infty$, since clearly $f(k) \leq n^k$, n being the number of letters (the number of nodes in Γ). Consequently, the quantity

$$\Theta(\Gamma) = \lim_{k\to\infty}(f(k))^{1/k}$$

is well-defined; moreover, for every k the quantity $(f(k))^{1/k}$ is a lower bound for $\Theta(\Gamma)$. The number $\Theta(\Gamma)$ is called the *Shannon capacity* of Γ. Our immediate observation is that

(!) *The Shannon capacity $\Theta(\Gamma)$ majorates the stability number of Γ:*

$$\alpha(\Gamma) \leq \Theta(\Gamma).$$

Indeed, as we remember, $(f(k))^{1/k}$ is a lower bound for $\Theta(\Gamma)$ for every $k = 1, 2, ...$; setting $k = 1$ and taking into account that $f(1) = \alpha(\Gamma)$, we get the desired result.

We see that the Shannon capacity number is an upper bound on the stability number; and this bound has a nice interpretation in terms of the Information Theory. The bad news is that we do not know how to compute the Shannon capacity. E.g., what is it for the toy graph C_5?

The stability number of C_5 clearly is 2, so that our first observation is that

$$\Theta(C_5) \geq \alpha(C_5) = 2.$$

To get a better estimate, let us look the graph $(C_5)^2$ (as we remember, $\Theta(\Gamma) \geq (f(k))^{1/k} = (\alpha(\Gamma^k))^{1/k}$ for every k). The graph $(C_5)^2$ has 25 nodes, so that we do not draw it; it, however, is not that difficult to find its stability number, which turns out to be 5. A good 5-element dictionary ($\equiv$ a 5-node independent set in $(C_5)^2$) is, e.g.,

$$AA, BC, CE, DB, ED.$$

Thus, we get

$$\Theta(C_5) \geq \sqrt{\alpha((C_5)^2)} = \sqrt{5}.$$

Attempts to compute the subsequent lower bounds $(f(k))^{1/k}$, as long as they are implementable (think how many vertices there are in $(C_5)^4$!), do not yield any improvements, and for more than 20 years it remained unknown whether $\Theta(C_5) = \sqrt{5}$ or is $> \sqrt{5}$. And this is for a toy graph! The breakthrough in the area of upper bounds for the stability number is due to L. Lovasz who in early 70's found a new – computable! – bound of this type.

Lovasz capacity number. Given a n-node graph Γ, let us associate with it an affine matrix-valued function $\mathcal{L}(x)$ taking values in the space of $n \times n$ symmetric matrices, namely, as follows:

• For every pair i, j of indices ($1 \leq i, j \leq n$) such that the nodes i and j are <u>not</u> linked by an arc, the ij-th entry of $\mathcal{L}$ is equal to 1;

• For a pair $i < j$ of indices such that the nodes i, j are linked by an arc, the ij-th and the ji-th entries in $\mathcal{L}$ are equal to x_{ij} – to the variable associated with the arc (i, j).

Thus, $\mathcal{L}(x)$ is indeed an affine function of N design variables x_{ij}, where N is the number of arcs in the graph. E.g., for graph C_5 the function $\mathcal{L}$ is as follows:

$$\mathcal{L} = \begin{pmatrix} 1 & x_{AB} & 1 & 1 & x_{EA} \\ x_{AB} & 1 & x_{BC} & 1 & 1 \\ 1 & x_{BC} & 1 & x_{CD} & 1 \\ 1 & 1 & x_{CD} & 1 & x_{DE} \\ x_{EA} & 1 & 1 & x_{DE} & 1 \end{pmatrix}.$$

Now, the Lovasz capacity number $\vartheta(\Gamma)$ is defined as the optimal value of the optimization program

$$\min_x \{\lambda_{\max}(\mathcal{L}(x))\},$$

i.e., as the optimal value in the semidefinite program

$$\min_{\lambda,x} \{\lambda : \lambda I_n - \mathcal{L}(x) \succeq 0\}. \tag{L}$$

Proposition 3.4.3 [Lovasz] *The Lovasz capacity number is an upper bound for the Shannon capacity:*

$$\vartheta(\Gamma) \geq \Theta(\Gamma)$$

and, consequently, for the stability number:

$$\vartheta(\Gamma) \geq \Theta(\Gamma) \geq \alpha(\Gamma).$$

For the graph C_5, the Lovasz capacity can be easily computed analytically and turns out to be exactly $\sqrt{5}$. Thus, a small byproduct of Lovasz's result is a solution to the problem which remained open for two decades.

Let us look how the Lovasz bound on the stability number can be obtained from the general relaxation scheme. To this end note that the stability number of an n-node graph Γ is the optimal value of the following optimization problem with 0-1 variables:

$$\max_x \left\{e^T x : x_i x_j = 0 \text{ whenever } i,j \text{ are adjacent nodes }, x_i \in \{0,1\},\ i = 1,...,n\right\},$$
$$e = (1,...,1)^T \in \mathbf{R}^n.$$

Indeed, 0-1 n-dimensional vectors can be identified with sets of nodes of Γ: the coordinates x_i of the vector x representing a set A of nodes are ones for $i \in A$ and zeros otherwise. The quadratic equality constraints $x_i x_j = 0$ for such a vector express equivalently the fact that the corresponding set of nodes is independent, and the objective $e^T x$ counts the cardinality of this set.

As we remember, the 0-1 restrictions on the variables can be represented equivalently by quadratic equality constraints, so that the stability number of Γ is the optimal value of the following problem with quadratic (in fact linear) objective and quadratic equality constraints:

$$\begin{array}{lrcl} \text{maximize} & e^T x & & \\ \text{s.t.} & & & \\ & x_i x_j & = & 0,\ (i,j) \text{ is an arc} \\ & x_i^2 - x_i & = & 0,\ i = 1,...,n. \end{array} \tag{3.4.8}$$

The latter problem is in the form of (3.4.1), with the only difference that the objective should be maximized rather than minimized. Switching from maximization of $e^T x$ to minimization of $(-e)^T x$ and passing to (3.4.5), we get the problem

$$\max_{\zeta,\mu} \left\{\zeta : \begin{pmatrix} -\zeta & -\frac{1}{2}(e+\mu)^T \\ -\frac{1}{2}(e+\mu) & A(\mu,\lambda) \end{pmatrix} \succeq 0\right\},$$

where μ is n-dimensional and $A(\mu,\lambda)$ is as follows:

- The diagonal entries of $A(\mu,\lambda)$ are $\mu_1,...,\mu_n$;
- The off-diagonal cells ij corresponding to non-adjacent nodes i,j ("empty cells") are zeros;
- The off-diagonal cells ij, $i<j$, and the symmetric cells ji corresponding to adjacent nodes i,j ("arc cells") are filled with free variables λ_{ij}.

Note that the optimal value in the resulting problem is a *lower* bound for *minus* the optimal value of (3.4.8), i.e., for minus the stability number of Γ.

Passing in the resulting problem from the variable ζ to a new variable $\xi = -\zeta$ and again switching from maximization of $\zeta = -\xi$ to minimization of ξ, we end up with the semidefinite program

$$\min_{\xi,\lambda,\mu}\left\{\xi : \begin{pmatrix} \xi & -\frac{1}{2}(e+\mu)^T \\ -\frac{1}{2}(e+\mu) & A(\mu,\lambda)\end{pmatrix} \succeq 0\right\}. \tag{3.4.9}$$

The optimal value in this problem is the minus optimal value in the previous one, which, in turn, is a lower bound on the minus stability number of Γ; consequently, the optimal value in (3.4.9) is an *upper* bound on the stability number of Γ.

We have built a semidefinite relaxation (3.4.9) of the problem of computing the stability number of Γ; the optimal value in the relaxation is an upper bound on the stability number. To get the Lovasz relaxation, let us further fix the μ-variables at the level 1 (this may only increase the optimal value in the problem, so that it still will be an upper bound for the stability number)[13]. With this modification, we come to the problem

$$\min_{\xi,\lambda}\left\{\xi : \begin{pmatrix} \xi & -e^T \\ -e & A(e,\lambda)\end{pmatrix} \succeq 0\right\}.$$

In every feasible solution to the problem, ξ should be ≥ 1 (it is an upper bound for $\alpha(\Gamma) \geq 1$). When $\xi \geq 1$, the LMI

$$\begin{pmatrix} \xi & -e^T \\ e & A(e,\lambda)\end{pmatrix} \succeq 0$$

by the Schur Complement Lemma is equivalent to the LMI

$$A(e,\lambda) \succeq (-e)\xi^{-1}(-e)^T,$$

or, which is the same, to the LMI

$$\xi A(e,\lambda) - ee^T \succeq 0.$$

The left hand side matrix in the latter LMI is equal to $\xi I_n - B(\xi,\lambda)$, where the matrix $B(\xi,\lambda)$ is as follows:

- The diagonal entries of $B(\xi,\lambda)$ are equal to 1;
- The off-diagonal "empty cells" are filled with ones;
- The "arc cells" from a symmetric pair off-diagonal pair ij and ji $(i<j)$ are filled with $\xi\lambda_{ij}$.

Passing from the design variables λ to the new ones $x_{ij} = \xi\lambda_{ij}$, we conclude that problem (3.4.9) with μ's set to ones is equivalent to the problem

$$\min_{\xi,x}\left\{\xi \to \min \mid \xi I_n - \mathcal{L}(x) \succeq 0\right\},$$

whose optimal value is exactly the Lovasz capacity number of Γ.

As a byproduct of our derivation, we get the easy part of the Lovasz Theorem – the inequality $\vartheta(\Gamma) \geq \alpha(\Gamma)$; this inequality, however, could be easily obtained directly from the definition of $\vartheta(\Gamma)$. The advantage of our derivation is that it demonstrates what is the origin of $\vartheta(\Gamma)$.

How good is the Lovasz capacity number? The Lovasz capacity number plays a crucial role in numerous graph-related problems; there is an important sub-family of graphs – *perfect graphs* – for which this number coincides with the stability number. However, for a general-type graph Γ, $\vartheta(\Gamma)$ may be a fairly poor bound for $\alpha(\Gamma)$. Lovasz has proved that for any graph Γ with n nodes, $\vartheta(\Gamma)\vartheta(\hat{\Gamma}) \geq n$, where $\hat{\Gamma}$ is the *complement* to Γ (i.e., two distinct nodes are adjacent in $\hat{\Gamma}$ if and only if they are *not* adjacent in Γ). It follows that for n-node graph Γ one always has $\max[\vartheta(\Gamma),\vartheta(\hat{\Gamma})] \geq \sqrt{n}$. On the other

[13] In fact setting $\mu_i = 1$, we do not vary the optimal value at all.

hand, it turns out that for a random n-node graph Γ (the arcs are drawn at random and independently of each other, with probability 0.5 to draw an arc linking two given distinct nodes) $\max[\alpha(\Gamma), \alpha(\hat{\Gamma})]$ is "typically" (with probability approaching 1 as n grows) of order of $\ln n$. It follows that for random n-node graphs a typical value of the ratio $\vartheta(\Gamma)/\alpha(\Gamma)$ is at least of order of $n^{1/2}/\ln n$; as n grows, this ratio blows up to ∞.

A natural question arises: are there "difficult" (NP-complete) combinatorial problems admitting "good" semidefinite relaxations – those with the quality of approximation not deteriorating as the sizes of instances grow? Let us look at two breakthrough results in this direction.

The MAXCUT problem and maximizing quadratic form over a box

The MAXCUT problem. The *maximum cut* problem is as follows:

Problem 3.4.1 [MAXCUT] *Let Γ be an n-node graph, and let the arcs (i,j) of the graph be associated with nonnegative "weights" a_{ij}. The problem is to find a cut of the largest possible weight, i.e., to partition the set of nodes in two parts S, S' in such a way that the total weight of all arcs "linking S and S'" (i.e., with one incident node in S and the other one in S') is as large as possible.*

In the MAXCUT problem, we may assume that the weights $a_{ij} = a_{ji} \geq 0$ are defined for every pair i, j of indices; it suffices to set $a_{ij} = 0$ for pairs i, j of non-adjacent nodes.

In contrast to the *minimum cut* problem (where we should minimize the weight of a cut instead of maximizing it), which is, basically, a nice LP program of finding the maximum flow in a network and is therefore efficiently solvable, the MAXCUT problem is as difficult as a combinatorial problem can be – it is NP-complete.

Theorem of Goemans and Williamson [7]. It is easy to build a semidefinite relaxation of MAXCUT. To this end let us pose MAXCUT as a quadratic problem with quadratic equality constraints. Let Γ be a n-node graph. A cut (S, S') – a partitioning of the set of nodes in two disjoint parts S, S' – can be identified with a n-dimensional vector x with coordinates ± 1 – $x_i = 1$ for $i \in S$, $x_i = -1$ for $i \in S'$. The quantity $\frac{1}{2}\sum_{i,j=1}^{n} a_{ij}x_ix_j$ is the total weight of arcs with both ends either in S or in S' minus the weight of the cut (S, S'); consequently, the quantity

$$\frac{1}{2}\left[\frac{1}{2}\sum_{i,j=1}^{n} a_{ij} - \frac{1}{2}\sum_{i,j=1}^{n} a_{ij}x_ix_j\right] = \frac{1}{4}\sum_{i,j=1}^{n} a_{ij}(1 - x_ix_j)$$

is exactly the weight of the cut (S, S').

We conclude that the MAXCUT problem can be posed as the following quadratic problem with quadratic equality constraints:

$$\max_x \left\{\frac{1}{4}\sum_{i,j=1}^{n} a_{ij}(1 - x_ix_j) : x_i^2 = 1,\ i = 1, ..., n\right\}. \tag{3.4.10}$$

For this problem, the semidefinite relaxation (3.4.7) after evident simplifications becomes the semidefinite program

$$\begin{array}{ll} \text{maximize} & \frac{1}{4}\sum_{i,j=1}^{n} a_{ij}(1 - X_{ij}) \\ \text{s.t.} & \\ & X = [X_{ij}]_{i,j=1}^{n} = X^T \succeq 0 \\ & X_{ii} = 1,\ i = 1, ..., n; \end{array} \tag{3.4.11}$$

the optimal value in the latter problem is an upper bound for the optimal value of MAXCUT.

The fact that (3.4.11) is a relaxation of (3.4.10) can be established directly, independently of any "general theory": (3.4.10) is the problem of maximizing the objective

$$\frac{1}{4}\sum_{i,j=1}^n a_{ij} - \frac{1}{2}\sum_{i,j=1}^n a_{ij}x_ix_j \equiv \frac{1}{4}\sum_{i,j=1}^n a_{ij} - \frac{1}{4}\mathrm{Tr}(AX(x)), \quad X(x) = xx^T$$

over all rank 1 matrices $X(x) = xx^T$ given by n-dimensional vectors x with entries ± 1. All these matrices are symmetric positive semidefinite with unit entries on the diagonal, i.e., they belong the feasible set of (3.4.11). Thus, (3.4.11) indeed is a relaxation of (3.4.10).

The quality of the semidefinite relaxation (3.4.11) is given by the following brilliant result of Goemans and Williamson (1995):

Theorem 3.4.1 *Let OPT be the optimal value of the* MAXCUT *problem* (3.4.10)*, and SDP be the optimal value of the semidefinite relaxation* (3.4.11)*. Then*

$$OPT \le SDP \le \alpha \cdot OPT, \ \alpha = 1.138... \tag{3.4.12}$$

Proof. The left inequality in (3.4.12) is what we already know – it simply says that semidefinite program (3.4.11) is a relaxation of MAXCUT. To get the right inequality, Goemans and Williamson act as follows. Let $X = [X_{ij}]$ be a feasible solution to the semidefinite relaxation. Since X is positive semidefinite, it is the covariance matrix of a Gaussian random vector ξ with zero mean, so that $\mathbf{E}\{\xi_i\xi_j\} = X_{ij}$. Now consider the random vector $\zeta = \mathrm{sign}[\xi]$ comprised of signs of the entries in ξ. A realization of ζ is almost surely a vector with coordinates ± 1, i.e., it is a cut. What is the *expected* weight of this cut? A straightforward computation demonstrates that $\mathbf{E}\{\zeta_i\zeta_j\} = \frac{2}{\pi}\mathrm{asin}(X_{ij})$ [14]. It follows that

$$\mathbf{E}\left\{\frac{1}{4}\sum_{i,j=1}^n a_{ij}(1-\zeta_i\zeta_i)\right\} = \frac{1}{4}\sum_{i,j=1}^n a_{ij}\left(1 - \frac{2}{\pi}\mathrm{asin}(X_{ij})\right). \tag{3.4.13}$$

Now, it is immediately seen that

$$-1 \le t \le 1 \Rightarrow 1 - \frac{2}{\pi}\mathrm{asin}(t) \ge \alpha^{-1}(1-t), \quad \alpha = 1.138...$$

In view of $a_{ij} \ge 0$, the latter observation combines with (3.4.13) to imply that

$$\mathbf{E}\left\{\frac{1}{4}\sum_{i,j=1}^n a_{ij}(1-\zeta_i\zeta_i)\right\} \ge \alpha^{-1}\frac{1}{4}\sum_{i,j=1}^n a_{ij}(1-X_{ij}).$$

The left hand side in this inequality, by evident reasons, is $\le OPT$. We have proved that the value of the objective in (3.4.11) at every feasible solution X to the problem is $\le \alpha \cdot OPT$, whence $SDP \le \alpha \cdot OPT$ as well. ∎

Note that the proof of Theorem 3.4.1 provides a randomized algorithm for building a suboptimal, within the factor $\alpha^{-1} = 0.878...$, solution to MAXCUT: we find a (nearly) optimal solution X to the semidefinite relaxation (3.4.11) of MAXCUT, generate a sample of, say, 100 realizations of the associated random cuts ζ and choose the one with the maximum weight.

Nesterov's $\frac{\pi}{2}$ Theorem

In the MAXCUT problem, we are in fact maximizing the homogeneous quadratic form

$$x^TAx \equiv \sum_{i=1}^n \left(\sum_{j=1}^n a_{ij}\right)x_i^2 - \sum_{i,j=1}^n a_{ij}x_ix_j$$

[14] Recall that $X_{ij} \succeq 0$ is normalized by the requirement $X_{ii} = 1$ for all i. Omitting this normalization, we would get $\mathbf{E}\{\zeta_i\zeta_j\} = \frac{2}{\pi}\mathrm{asin}\left(\frac{X_{ij}}{\sqrt{X_{ii}X_{jj}}}\right)$.

over the set S_n of n-dimensional vectors x with coordinates ± 1. It is easily seen that the matrix A of this form is positive semidefinite and possesses a specific feature that the off-diagonal entries are nonpositive, while the sum of the entries in every row is 0. What happens when we are maximizing over S_n a quadratic form x^TAx with a general-type (symmetric) matrix A? An extremely nice result in this direction was obtained by Yu. Nesterov. The cornerstone of Nesterov's construction relates to the case when A is positive semidefinite, and this is the case we shall focus on. Note that the problem of maximizing a quadratic form x^TAx with positive semidefinite (and, say, integer) matrix A over S_n, same as MAXCUT, is NP-complete.

The semidefinite relaxation of the problem

$$\max_x \left\{x^TAx : x \in S_n \quad [\Leftrightarrow x_i \in \{-1,1\},\ i=1,...,n]\right\} \tag{3.4.14}$$

can be built exactly in the same way as (3.4.11) and turns out to be the semidefinite program

$$\begin{array}{lrcl} \text{maximize} & \mathrm{Tr}(AX) & & \\ \text{s.t.} & & & \\ & X = X^T = [X_{ij}]_{i,j=1}^n & \succeq & 0 \\ & X_{ii} & = & 1,\ i=1,...,n. \end{array} \tag{3.4.15}$$

The optimal value in this problem, let it again be called SDP, is $\geq$ the optimal value OPT in the original problem (3.4.14). The ratio OPT/SDP, however, cannot be too large:

Theorem 3.4.2 [Nesterov's $\frac{\pi}{2}$ Theorem] *Let A be positive semidefinite. Then*

$$OPT \leq SDP \leq \frac{\pi}{2} SDP \quad [\frac{\pi}{2} = 1.570...]$$

The proof utilizes the central idea of Goemans and Williamson in the following brilliant reasoning:

The inequality $SDP \geq OPT$ is valid since (3.4.15) is a relaxation of (3.4.14). Let X be a feasible solution to the relaxed problem; let, same as in the MAXCUT construction, ξ be a Gaussian random vector with zero mean and the covariance matrix X, and let $\zeta = \mathrm{sign}[\xi]$. As we remember,

$$\mathbf{E}\left\{\zeta^TA\zeta\right\} = \sum_{i,j} A_{ij} \frac{2}{\pi} \mathrm{asin}(X_{ij}) = \frac{2}{\pi} \mathrm{Tr}(A, \mathrm{asin}[X]), \tag{3.4.16}$$

where for a function f on the axis and a matrix X $f[X]$ denotes the matrix with the entries $f(X_{ij})$. Now – the crucial (although simple) observation:

> *For a positive semidefinite symmetric matrix X with diagonal entries ± 1 (in fact, for any positive semidefinite X with $|X_{ij}| \leq 1$) one has*
>
> $$\mathrm{asin}[X] \succeq X. \tag{3.4.17}$$
>
> The proof is immediate: denoting by $[X]^k$ the matrix with the entries X_{ij}^k and making use of the Taylor series for the asin (this series converges uniformly on $[-1,1]$), for a matrix X with all entries belonging to $[-1,1]$ we get
>
> $$\mathrm{asin}[X] - X = \sum_{k=1}^{\infty} \frac{1 \times 3 \times 5 \times ... \times (2k-1)}{2^k k!(2k+1)} [X]^{2k+1},$$
>
> and all we need is to note is that all matrices in the left hand side are $\succeq 0$ along with X [15)]

[15)] The fact that the entry-wise product of two positive semidefinite matrices is positive semidefinite is a standard fact from Linear Algebra. The easiest way to understand it is to note that if P, Q are positive semidefinite symmetric matrices of the same size, then they are Gram matrices: $P_{ij} = p_i^Tp_j$ for certain system of vectors p_i from certain (no matter from which exactly) $\mathbf{R}^N$ and $Q_{ij} = q_i^Tq_j$ for a system of vectors q_i from certain $\mathbf{R}^M$. But then the entry-wise product of P and Q – the matrix with the entries $P_{ij}Q_{ij} = (p_i^Tp_j)(q_i^Tq_j)$ – also is a Gram matrix, namely, the Gram matrix of the matrices $p_iq_i^T \in \mathbf{M}^{N,M} = \mathbf{R}^{NM}$. Since every Gram matrix is positive semidefinite, the entry-wise product of P and Q is positive semidefinite.

Combining (3.4.16), (3.4.17) and the fact that A is positive semidefinite, we conclude that

$$[OPT \geq] \quad \mathbf{E}\left\{\zeta^T A\zeta\right\} = \frac{2}{\pi}\mathrm{Tr}(A\mathrm{asin}[X]) \geq \frac{2}{\pi}\mathrm{Tr}(AX).$$

The resulting inequality is valid for every feasible solution X of (3.4.15), whence $SDP \leq \frac{\pi}{2}OPT$. ∎

The $\frac{\pi}{2}$ Theorem has a number of far-reaching consequences (see Nesterov's papers [15, 16]), for example, the following two:

Theorem 3.4.3 *Let T be an SDr compact subset in $\mathbf{R}^n_+$. Consider the set*

$$\mathcal{T} = \{x \in \mathbf{R}^n : (x_1^2, ..., x_n^2)^T \in T\},$$

and let A be a symmetric $n \times n$ matrix. Then the quantities $m_*(A) = \min\limits_{x\in\mathcal{T}} x^T Ax$ and $m^*(A) = \max\limits_{x\in\mathcal{T}} x^T Ax$ admit efficiently computable bounds

$$\begin{array}{rcl} s_*(A) & \equiv & \min\limits_X \left\{\mathrm{Tr}(AX) : X \succeq 0, (X_{11}, ..., X_{nn})^T \in T\right\}, \\ s^*(A) & \equiv & \max\limits_X \left\{\mathrm{Tr}(AX) : X \succeq 0, (X_{11}, ..., X_{nn})^T \in T\right\}, \end{array}$$

such that

$$s_*(A) \leq m_*(A) \leq m^*(A) \leq s^*(A)$$

and

$$m^*(A) - m_*(A) \leq s^*(A) - s_*(A) \leq \frac{\pi}{4-\pi}(m^*(A) - m_*(A))$$

(in the case of $A \succeq 0$ and $0 \in T$, the factor $\frac{\pi}{4-\pi}$ can be replaced with $\frac{\pi}{2}$).

Thus, the "variation" $\max\limits_{x\in\mathcal{T}} x^T Ax - \min\limits_{x\in\mathcal{T}} x^T Ax$ of the quadratic form $x^T Ax$ on $\mathcal{T}$ can be efficiently bounded from above, and the bound is tight within an absolute constant factor.

Note that if T is given by a strictly feasible SDR, then both $(-s_*(A))$ and $s^*(A)$ are SDr functions of A (Proposition 2.4.4).

Theorem 3.4.4 *Let $p \in [2, \infty]$, $r \in [1, 2]$, and let A be an $m \times n$ matrix. Consider the problem of computing the operator norm $\|A\|_{p,r}$ of the linear mapping $x \mapsto Ax$, considered as the mapping from the space $\mathbf{R}^n$ equipped with the norm $\|\cdot\|_p$ to the space $\mathbf{R}^m$ equipped with the norm $\|\cdot\|_r$:*

$$\|A\|_{p,r} = \max\left\{\|Ax\|_r : \|x\|_p \leq 1\right\};$$

note that it is difficult (NP-hard) to compute this norm, except for the case of $p = r = 2$. The "computationally intractable" quantity $\|A\|_{p,r}$ admits an efficiently computable upper bound

$$\omega_{p,r}(A) = \min_{\lambda\in\mathbf{R}^m, \mu\in\mathbf{R}^n} \left\{\frac{1}{2}\left[\|\mu\|_{\frac{p}{p-2}} + \|\lambda\|_{\frac{r}{2-r}}\right] : \begin{pmatrix} \mathrm{Diag}\{\mu\} & A^T \\ A & \mathrm{Diag}\{\lambda\} \end{pmatrix} \succeq 0\right\};$$

this bound is exact for a nonnegative matrix A, and for an arbitrary A the bound is tight within the factor $\frac{\pi}{2\sqrt{3}-2\pi/3} = 2.293...$:

$$\|A\|_{p,r} \leq \omega_{p,r}(A) \leq \frac{\pi}{2\sqrt{3} - 2\pi/3}\|A\|_{p,r}.$$

Moreover, when $p \in [1, \infty)$ and $r \in [1, 2]$ are rational (or $p = \infty$ and $r \in [1, 2]$ is rational), the bound $\omega_{p,r}(A)$ is an SDr function of A.

3.4.2 Matrix Cube Theorem and interval stability analysis/synthesis

Consider the problem of Lyapunov Stability Analysis in the case of interval uncertainty:

$$\mathcal{U} = \mathcal{U}_\rho = \{A \in \mathbf{M}^{n,n} \mid |A_{ij} - A^*_{ij}| \le \rho D_{ij},\ i,j = 1,...,n\}, \tag{3.4.18}$$

where A^* is the "nominal" matrix, $D \neq 0$ is a matrix with nonnegative entries specifying the "scale" for perturbations of different entries, and $\rho \ge 0$ is the "level of perturbations". We deal with a polytopic uncertainty, and as we remember from Section 3.3.3, to certify the stability is the same as to find a feasible solution of the associated semidefinite program (3.3.8) with a negative value of the objective. The difficulty, however, is that the number N of LMI constraints in this problem is the number of vertices of the polytope (3.4.18), i.e., $N = 2^m$, where m is the number of uncertain entries in our interval matrix ($\equiv$the number of positive entries in D). For 5×5 interval matrices with "full uncertainty" $m = 25$, i.e., $N = 2^{25} = 33,554,432$, which is "a bit" too many; for "fully uncertain" 10×10 matrices, $N = 2^{100} > 1.2 \times 10^{30}$... Thus, the "brute force" approach fails already for "pretty small" matrices affected by interval uncertainty.

In fact, the difficulty we have encountered lies in the NP-hardness of the following problem:

> *Given a candidate Lyapunov stability certificate $X \succ 0$ and $\rho > 0$, check whether X indeed certifies stability of all instances of $\mathcal{U}_\rho$, i.e., whether X solves the semi-infinite system of LMI's*
>
> $$A^T X + XA \preceq -I \quad \forall A \in \mathcal{U}_\rho. \tag{3.4.19}$$
>
> (in fact, we are interested in the system "$A^T X + XA \prec 0 \, \forall A \in \mathcal{U}_\rho$", but this is a minor difference – the "system of interest" is homogeneous in X, and therefore every feasible solution of it can be converted to a solution of (3.4.19) just by scaling $X \mapsto tX$).

The above problem, in turn, is a particular case of the following problem:

> "Matrix Cube": *Given matrices $A_0, A_1, ..., A_m \in \mathbf{S}^n$ with $A_0 \succeq 0$, find the largest $\rho = R[A_1, ..., A_m : A_0]$ such that the set*
>
> $$\mathcal{A}_\rho = \left\{ A = A_0 + \sum_{i=1}^m z_i A_i \mid \|z\|_\infty \le \rho \right\} \tag{3.4.20}$$
>
> *– the image of the m-dimensional cube $\{z \in \mathbf{R}^m \mid \|z\|_\infty \le \rho\}$ under the affine mapping $z \mapsto A_0 + \sum_{i=1}^m z_i A_i$ – is contained in the semidefinite cone $\mathbf{S}^n_+$.*

This is the problem we will focus on.

The Matrix Cube Theorem. The problem "Matrix Cube" (MC for short) is NP-hard; this is true also for the "feasibility version" MC_ρ of MC, where we, given a $\rho \ge 0$, are interested to verify the inclusion $\mathcal{A}_\rho \subset \mathbf{S}^n_+$. However, we can point out a simple *sufficient condition* for the validity of the inclusion $\mathcal{A}_\rho \subset \mathbf{S}^n_+$:

Proposition 3.4.4 *Assume that the system of LMI's*

$$\begin{array}{ll} (a) & X^i \succeq \rho A_i,\ X^i \succeq -\rho A_i,\ i = 1,...,m; \\ (b) & \sum_{i=1}^m X_i \preceq A_0 \end{array} \tag{S_ρ}$$

in matrix variables $X^1, ..., X^m \in \mathbf{S}^n$ is solvable. Then $\mathcal{A}_\rho \subset \mathbf{S}^n_+$.

Proof. Let $X^1, ..., X^m$ be a solution of (S_ρ). From (a) it follows that whenever $\|z\|_\infty \le \rho$, we have $X^i \succeq z_i A_i$ for all i, whence by (b)

$$A_0 + \sum_{i=1}^m z_i A_i \succeq A_0 - \sum_i X_i \succeq 0. \qquad \blacksquare$$

Our main result is that the sufficient condition for the inclusion $\mathcal{A}_\rho \subset \mathbf{S}^n_+$ stated by Proposition 3.4.4 is not too conservative:

Theorem 3.4.5 *If the system of LMI's* (S_ρ) *is not solvable, then*

$$\mathcal{A}_{\vartheta(\mu)\rho} \not\subset \mathbf{S}^n_+; \tag{3.4.21}$$

here

$$\mu = \max_{1\le i\le m} \mathrm{Rank}(A_i)$$

(note "$i \ge 1$" in the max*!), and*

$$\vartheta(k) \le \frac{\pi\sqrt{k}}{2},\ k \ge 1; \quad \vartheta(2) = \frac{\pi}{2}. \tag{3.4.22}$$

Proof. Below $\zeta \sim \mathcal{N}(0, I_n)$ means that ζ is a random Gaussian n-dimensional vector with zero mean and the unit covariance matrix, and $p_n(\cdot)$ stands for the density of the corresponding probability distribution:

$$p_n(u) = (2\pi)^{-n/2} \exp\left\{-\frac{u^T u}{2}\right\}, \quad u \in \mathbf{R}^n.$$

Let us set

$$\vartheta(k) = \frac{1}{\min\left\{\int |\alpha_i u_1^2 + ... + \alpha_k u_k^2| p_k(u)du \,\middle|\, \alpha \in \mathbf{R}^k, \|\alpha\|_1 = 1\right\}}. \tag{3.4.23}$$

It suffices to verify that

(i): With the just defined $\vartheta(\cdot)$, insolvability of (S_ρ) does imply (3.4.21);

(ii): $\vartheta(\cdot)$ satisfies (3.4.22).

Let us prove (i).

1^0. Assume that (S_ρ) has no solutions. It means that the optimal value of the semidefinite problem

$$\min_{t,\{X^i\}} \left\{ t \,\middle|\, \begin{array}{c} X^i \succeq \rho A_i,\ X^i \succeq -\rho A_i,\ i = 1, ..., m; \\ \sum_{i=1}^m X_i \preceq A_0 + tI \end{array} \right\} \tag{3.4.24}$$

is positive. Since the problem is strictly feasible, its optimal value is positive if and only if the optimal value of the dual problem

$$\max_{W,\{U^i,V^i\}} \left\{ \rho \sum_{i=1}^m \mathrm{Tr}([U^i - V^i]A_i) - \mathrm{Tr}(WA_0) \,\middle|\, \begin{array}{c} U^i + V^i = W,\ i = 1, ..., m, \\ \mathrm{Tr}(W) = 1, \\ U^i, V^i, W \succeq 0 \end{array} \right\}$$

is positive. Thus, there exists matrices U^i, V^i, W such that

$$\begin{array}{ll} (a) & U^i, V^i, W \succeq 0, \\ (b) & U^i + V^i = W,\ i = 1, 2, ...m, \\ (c) & \rho \sum_{i=1}^m \mathrm{Tr}([U^i - V^i]A_i) > \mathrm{Tr}(WA_0). \end{array} \tag{3.4.25}$$

2^0. Now let us use simple

Lemma 3.4.1 *Let* $W, A \in \mathbf{S}^n$, $W \succeq 0$. *Then*

$$\max_{U,V\succeq 0, U+V=W} \mathrm{Tr}([U - V]A) = \max_{X=X^T:\|\lambda(X)\|_\infty\le 1} \mathrm{Tr}(XW^{1/2}AW^{1/2}) = \|\lambda(W^{1/2}AW^{1/2})\|_1. \tag{3.4.26}$$

Proof of Lemma. We clearly have

$$U, V \succeq 0, U+V=W \Leftrightarrow U=W^{1/2}PW^{1/2}, V=W^{1/2}QW^{1/2}, P, Q \succeq 0, P+Q=I,$$

whence

$$\max_{U,V:U,V\succeq 0,U+V=W} \mathrm{Tr}([U-V]A) = \max_{P,Q:P,Q\succeq 0,P+Q=I} \mathrm{Tr}([P-Q]W^{1/2}AW^{1/2}).$$

When P, Q are linked by the relation $P+Q=I$ and vary in $\{P \succeq 0, Q \succeq 0\}$, the matrix $X=P-Q$ runs through the entire "interval" $\{-I \preceq X \preceq I\}$ (why?); we have proved the first equality in (3.4.26). When proving the second equality, we may assume w.l.o.g. that the matrix $W^{1/2}AW^{1/2}$ is diagonal, so that $\mathrm{Tr}(XW^{1/2}AW^{1/2}) = \lambda^T(W^{1/2}AW^{1/2})\mathrm{Dg}(X)$, where $\mathrm{Dg}(X)$ is the diagonal of X. When X runs through the "interval" $\{-I \preceq X \preceq I\}$, the diagonal of X runs through the entire unit cube $\{\|x\|_\infty \le 1\}$, which immediately yields the second equality in (3.4.26). ∎

By Lemma 3.4.1, from (3.4.25) it follows that there exists $W \succeq 0$ such that

$$\rho \sum_{i=1}^m \|\lambda(W^{1/2}A_iW^{1/2})\|_1 > \mathrm{Tr}(W^{1/2}A_0W^{1/2}). \tag{3.4.27}$$

3^0. Now let us use the following observation:

Lemma 3.4.2 *With $\xi \sim \mathcal{N}(0, I_n)$, for every k and every symmetric $n \times n$ matrix A with $\mathrm{Rank}(A) \le k$ one has*

$$\begin{array}{ll} (a) & \mathbf{E}\left\{\xi^TA\xi\right\} = \mathrm{Tr}(A), \\ (a) & \mathbf{E}\left\{|\xi^TA\xi|\right\} \ge \frac{1}{\vartheta(\mathrm{Rank}(A))}\|\lambda(A)\|_1; \end{array} \tag{3.4.28}$$

here $\mathbf{E}$ stands for the expectation w.r.t. the distribution of ξ.

Proof of Lemma. (3.4.28.a) is evident:

$$\mathbf{E}\left\{\xi^TA\xi\right\} = \sum_{i,j=1}^m A_{ij}\mathbf{E}\left\{\xi_i\xi_j\right\} = \mathrm{Tr}(A).$$

To prove (3.4.28.b), by homogeneity it suffices to consider the case when $\|\lambda(A)\|_1 = 1$, and by rotational invariance of the distribution of ξ – the case when A is diagonal, and the first $\mathrm{Rank}(A)$ of diagonal entries of A are the nonzero eigenvalues of the matrix; with this normalization, the required relation immediately follows from the definition of $\vartheta(\cdot)$. ∎

4^0. Now we are ready to prove (i). Let $\xi \sim \mathcal{N}(0, I_n)$. We have

$$\begin{array}{rcl} \mathbf{E}\left\{\rho\vartheta(\mu)\sum\limits_{i=1}^k |\xi^TW^{1/2}A_iW^{1/2}\xi|\right\} & = & \sum\limits_{i=1}^m \rho\vartheta(\mu)\mathbf{E}\left\{|\xi^TW^{1/2}A_iW^{1/2}\xi|\right\} \\ & \ge & \rho\sum\limits_{i=1}^m \|\lambda(W^{1/2}A_iW^{1/2}\|_1 \\ \text{[by (3.4.28.}b\text{) due to } \mathrm{Rank}(W^{1/2}A_iW^{1/2}) \le \mathrm{Rank}(A_i) \le \mu,\ i \ge 1] & & \\ & > & \mathrm{Tr}(W^{1/2}A_0W^{1/2}) \\ & & \text{[by (3.4.27)]} \\ & = & \mathrm{Tr}(\xi^TW^{1/2}A_0W^{1/2}\xi), \\ & & \text{[by (3.4.28.}a\text{)]} \end{array}$$

whence

$$\mathbf{E}\left\{\rho\vartheta(\mu)\sum_{i=1}^k |\xi^TW^{1/2}A_iW^{1/2}\xi| - \xi^TW^{1/2}A_0W^{1/2}\xi\right\} > 0.$$

It follows that there exists $r \in \mathbf{R}^n$ such that

$$\vartheta(\mu)\rho\sum_{i=1}^{m}|r^TW^{1/2}A_iW^{1/2}r| > r^TW^{1/2}A_0W^{1/2}r,$$

so that setting $z_i = -\vartheta(\mu)\rho\mathrm{sign}(r^TW^{1/2}A_iW^{1/2}r)$, we get

$$r^TW^{1/2}\left(A_0+\sum_{i=1}^{m}z_iA_i\right)W^{1/2}r < 0.$$

We see that the matrix $A_0+\sum_{i=1}^{m}z_iA_i$ is *not* positive semidefinite, while by construction $\|z\|_\infty \le \vartheta(\mu)\rho$. Thus, (3.4.21) holds true. (i) is proved.

To prove (ii), let $\alpha \in \mathbf{R}^k$ be such that $\|\alpha\|_1 = 1$, and let

$$J = \int |\alpha_1u_1^2 + ... + \alpha_ku_k^2|p_k(u)du.$$

Let $\beta = \begin{bmatrix} \alpha \\ -\alpha \end{bmatrix}$, and let $\xi \sim \mathcal{N}(0, I_{2k})$. We have

$$\mathbf{E}\left\{\left|\sum_{i=1}^{2k}\beta_i\xi_i^2\right|\right\} \le \mathbf{E}\left\{\left|\sum_{i=1}^{k}\beta_i\xi_i^2\right|\right\} + \mathbf{E}\left\{\left|\sum_{i=1}^{k}\beta_{i+k}\xi_{i+k}^2\right|\right\} = 2J. \tag{3.4.29}$$

On the other hand, let $\eta_i = \frac{1}{\sqrt{2}}(\xi_i - \xi_{k+i})$, $\zeta_i = \frac{1}{\sqrt{2}}(\xi_i + \xi_{k+i})$, $i = 1, ..., k$, and let $\omega = \begin{pmatrix} \alpha_1\eta_1 \\ \vdots \\ \alpha_k\eta_k \end{pmatrix}$, $\widetilde{\omega} = \begin{pmatrix} |\alpha_1\eta_1| \\ \vdots \\ |\alpha_k\eta_k| \end{pmatrix}$, $\zeta = \begin{pmatrix} \zeta_1 \\ \vdots \\ \zeta_k \end{pmatrix}$. Observe that ζ and ω are independent and $\zeta \sim \mathcal{N}(0, I_k)$. We have

$$\mathbf{E}\left\{\left|\sum_{i=1}^{2k}\beta_i\xi_i^2\right|\right\} = 2\mathbf{E}\left\{\left|\sum_{i=1}^{k}\alpha_i\eta_i\zeta_i\right|\right\} = 2\mathbf{E}\left\{|\omega^T\zeta|\right\} = \mathbf{E}\left\{\|\omega\|_2\right\}\mathbf{E}\left\{|\zeta_1|\right\},$$

where the concluding equality follows from the fact that $\zeta \sim \mathcal{N}(0, I_k)$ is independent of ω. We further have

$$\mathbf{E}\left\{|\zeta_1|\right\} = \int |t|p_1(t)dt = \frac{2}{\sqrt{2\pi}}$$

and

$$\mathbf{E}\left\{\|\omega\|_2\right\} = \mathbf{E}\left\{\|\widetilde{\omega}\|_2\right\} \ge \|\mathbf{E}\left\{\widetilde{\omega}\right\}\|_2 = \left[\int |t|p_1(t)dt\right]\sqrt{\sum_{i=1}^{m}\alpha_i^2}.$$

Combining our observations, we come to

$$\mathbf{E}\left\{\left|\sum_{i=1}^{2k}\beta_i\xi_i^2\right|\right\} \ge 2\left(\frac{2}{\sqrt{2\pi}}\right)^2\|\alpha\|_2 \ge \frac{4}{\pi\sqrt{k}}\|\alpha\|_1 = \frac{4}{\pi\sqrt{k}}.$$

This relation combines with (3.4.29) to yield $J \ge \frac{2}{\pi\sqrt{k}}$. Recalling the definition of $\vartheta(k)$, we come to $\vartheta(k) \le \frac{\pi\sqrt{k}}{2}$, as required in (3.4.22).

It remains to prove that $\vartheta(2) = \frac{\pi}{2}$. From the definition of $\vartheta(\cdot)$ it follows that

$$\vartheta^{-1}(2) = \min_{0\le\theta\le 1} \int |\theta u_1^2 - (1-\theta)u_2^2| p_2(u)du \equiv \min_{0\le\theta\le 1} f(\theta).$$

The function $f(\theta)$ is clearly convex and satisfies the identity $f(\theta) = f(1-\theta)$, $0 \le \theta \le 1$, so that its minimum is attained at $\theta = \frac{1}{2}$. A direct computation says that $f(\frac{1}{2}) = \frac{2}{\pi}$. ∎

Corollary 3.4.1 *Let the ranks of all matrices $A_1, ..., A_m$ in* MC *be $\le \mu$. Then the optimal value in the semidefinite problem*

$$\rho[A_1, ..., A_m : A_0] = \max_{\rho, X^i} \left\{ \rho \mid \begin{array}{c} X^i \succeq \rho A_i,\ X^i \succeq -\rho A_i,\ i = 1, ..., m, \\ \sum_{i=1}^m X^i \preceq A_0 \end{array} \right\} \tag{3.4.30}$$

is a lower bound on $R[A_1, ..., A_m : A_0]$, and the "true" quantity is at most $\vartheta(\mu)$ times (see (3.4.23), (3.4.22)*) larger than the bound:*

$$\rho[A_1, ..., A_m : A_0] \le R[A_1, ..., A_m : A_0] \le \vartheta(\mu)\rho[A_1, ..., A_m : A_0]. \tag{3.4.31}$$

Application: Lyapunov Stability Analysis for an interval matrix. Now we are equipped to attack the problem of certifying the stability of uncertain linear dynamic system with interval uncertainty. The problem we are interested in is as follows:

"Interval Lyapunov": *Given a stable $n \times n$ matrix A_** [16] *and an $n \times n$ matrix $D \ne 0$ with nonnegative entries, find the supremum $R[A_*, D]$ of those $\rho \ge 0$ for which all instances of the "interval matrix"*

$$\mathcal{U}_\rho = \{A \in \mathbf{M}^{n,n} : |A_{ij} - (A_*)_{ij}| \le \rho D_{ij},\ i, j = 1, ..., n\}$$

share a common quadratic Lyapunov function, i.e., the semi-infinite system of LMI's

$$X \succeq I; \quad A^T X + XA \preceq -I\ \forall A \in \mathcal{U}_\rho \tag{Ly$_\rho$}$$

in matrix variable $X \in \mathbf{S}^n$ is solvable.

Observe that $X \succeq I$ solves (Ly$_\rho$) if and only if the matrix cube

$$\begin{array}{rcl} \mathcal{A}_\rho[X] & = & \Big\{ B = \underbrace{\left[-I - A_*^T X - XA_*\right]}_{A_0[X]} \\ & & \quad + \sum_{(i,j)\in\mathcal{D}} z_{ij} \underbrace{\left[[D_{ij}E^{ij}]^T X + X[D_{ij}E^{ij}]\right]}_{A_{ij}[X]} \Big| \, |z_{ij}| \le \rho,\ (i,j) \in \mathcal{D} \Big\} \\ \mathcal{D} & = & \{(i,j) : D_{ij} > 0\} \end{array}$$

is contained in $\mathbf{S}^n_+$; here E^{ij} are the "basic $n \times n$ matrices" (ij-th entry of E^{ij} is 1, all other entries are zero). Note that *the ranks of the matrices $A_{ij}[X]$, $(i,j) \in \mathcal{D}$, are at most 2.* Therefore from Proposition 3.4.4 and Theorem 3.4.5 we get the following result:

Proposition 3.4.5 *Let $\rho \ge 0$. Then*

(i) *If the system of LMI's*

$$\begin{array}{c} X \succeq I, \\ X^{ij} \succeq -\rho D_{ij}\left[[E^{ij}]^T X + XE^{ij}\right],\ X^{ij} \succeq \rho D_{ij}\left[[E^{ij}]^T X + XE^{ij}\right],\ (i,j) \in \mathcal{D} \\ \sum_{(i,j)\in\mathcal{D}}^n X^{ij} \preceq -I - A_*^T X - XA_* \end{array} \tag{A$_\rho$}$$

[16] I.e., with all eigenvalues from the open left half-plane, or, which is the same, such that $A_*^T X + XA_* \prec 0$ for certain $X \succ 0$.

in matrix variables X, X^{ij}, $(i,j) \in \mathcal{D}$, is solvable, then so is the system (Ly_ρ)*, and the X-component of a solution of the former system solves the latter system.*

(ii) *If the system of LMI's* (A_ρ) *is <u>not</u> solvable, then so is the system* $(\mathrm{Ly}_{\frac{\pi\rho}{2}})$.

In particular, the supremum $\rho[A_*, D]$ of those ρ for which (A_ρ) *is solvable is a lower bound for $R[A_*, D]$, and the "true" quantity is at most $\frac{\pi}{2}$ times larger than the bound:*

$$\rho[A_*, D] \le R[A_*, D] \le \frac{\pi}{2}\rho[A_*, D].$$

Computing $\rho[A_*, D]$. The quantity $\rho[A_*, D]$, in contrast to $R[A_*, D]$, is "efficiently computable": applying dichotomy in ρ, we can find a high-accuracy approximation of $\rho[A_*, D]$ via solving a small series of semidefinite feasibility problems (A_ρ). Note, however, that problem (A_ρ), although "computationally tractable", is not that simple: in the case of "full uncertainty" ($D_{ij} > 0$ for all i, j) it has $n^2 + n$ matrix variables of the size $n \times n$ each. It turns out that applying semidefinite duality, one can reduce dramatically the sizes of the problem specifying $\rho[A_*, D]$. The resulting (equivalent!) description of the bound is:

$$\frac{1}{\rho[A_*, D]} = \inf_{\lambda, Y, X, \{\eta_i\}} \left\{ \lambda \middle| \begin{array}{c} X \succeq I, \\ \begin{pmatrix} Y - \sum\limits_{\ell=1}^{m} \eta_\ell e_{j_\ell} e_{j_\ell}^T & [Xe_{i_1}; Xe_{i_2}; ...; Xe_{i_m}] \\ [Xe_{i_1}; Xe_{i_2}; ...; Xe_{i_m}]^T & \mathrm{Diag}(\eta_1, ..., \eta_m) \end{pmatrix} \succeq 0, \\ A_0[X] \equiv -I - A_*^T X + X A_* \succ 0, \\ Y \preceq \lambda A_0[X] \end{array} \right\}, \qquad (3.4.32)$$

where (i_1, j_1), ..., (i_m, j_m) are the positions of the uncertain entries in our uncertain matrix (i.e., the pairs (i, j) such that $D_{ij} > 0$) and $e_1, ..., e_n$ are the standard basic orths in $\mathbf{R}^n$.

Note that the optimization program in (3.4.32) has just two symmetric matrix variables X, Y, a single scalar variable λ and $m \le n^2$ scalar variables η_i, i.e., totally at most $2n^2 + n + 2$ scalar design variables, which, for large m, is much less than the design dimension of (A_ρ).

Remark 3.4.1 Note that our results on the Matrix Cube problem can be applied to the interval version of the Lyapunov Stability *Synthesis* problem, where we are interested to find the supremum R of those ρ for which an uncertain controllable system

$$\frac{d}{dt}x(t) = A(t)x(t) + B(t)u(t)$$

with interval uncertainty

$$(A(t), B(t)) \in \mathcal{U}_\rho = \{(A, B) : |A_{ij} - (A_*)_{ij}| \le \rho D_{ij},\ |B_{i\ell} - (B_*)_{i\ell}| \le \rho C_{i\ell}\ \ \forall i, j, \ell\}$$

admits a linear feedback

$$u(t) = Kx(t)$$

such that all instances $A(t) + B(t)K$ of the resulting closed loop system share a common quadratic Lyapunov function. Here our constructions should be applied to the semi-infinite system of LMI's

$$Y \succeq I, \quad BL + AY + L^T B^T + Y A^T \preceq -I \quad \forall (A, B) \in \mathcal{U}_\rho$$

in variables L, Y (see Proposition 3.3.4), and them yield an efficiently computable lower bound on R which is at most $\frac{\pi}{2}$ times less than R.

We have seen that the Matrix Cube Theorem allows to build tight computationally tractable approximations to semi-infinite systems of LMI's responsible for stability of uncertain linear dynamical systems affected by interval uncertainty. The same is true for many other semi-infinite systems of LMI's arising in Control in the presence of interval uncertainty, since in a typical Control-related LMI, a perturbation of a single entry in the underlying data results in a small-rank perturbation of the LMI – a situation well-suited for applying the Matrix Cube Theorem.

Nesterov's Theorem revisited. Our results on the Matrix Cube problem give an alternative proof of Nesterov's $\frac{\pi}{2}$ Theorem (Theorem 3.4.2). Recall that in this theorem we are comparing the true maximum

$$OPT = \max_d \{d^T Ad \mid \|d\|_\infty \leq 1\}$$

of a *positive semidefinite* ($A \succeq 0$) quadratic form on the unit n-dimensional cube and the semidefinite upper bound

$$SDP = \max_X \{\mathrm{Tr}(AX) \mid X \succeq 0, X_{ii} \leq 1, \, i = 1, ..., n\} \tag{3.4.33}$$

on OPT; the theorem says that

$$OPT \leq SDP \leq \frac{\pi}{2} OPT. \tag{3.4.34}$$

To derive (3.4.34) from the Matrix Cube-related considerations, assume that $A \succ 0$ rather than $A \succeq 0$ (by continuity reasons, to prove (3.4.34) for the case of $A \succ 0$ is the same as to prove the relation for all $A \succeq 0$) and let us start with the following simple observation:

Lemma 3.4.3 *Let $A \succ 0$ and*

$$OPT = \max_d \left\{d^T Ad \mid \|d\|_\infty \leq 1\right\}.$$

Then

$$\frac{1}{OPT} = \max \left\{\rho : \begin{pmatrix} 1 & d^T \\ d & A^{-1} \end{pmatrix} \succeq 0 \quad \forall (d : \|d\|_\infty \leq \rho^{1/2})\right\} \tag{3.4.35}$$

and

$$\frac{1}{OPT} = \max \left\{\rho : A^{-1} \succeq X \quad \forall (X \in \mathbf{S}^n : |X_{ij}| \leq \rho \forall i, j)\right\}. \tag{3.4.36}$$

Proof. To get (3.4.35), note that by the Schur Complement Lemma, all matrices of the form $\begin{pmatrix} 1 & d^T \\ d & A^{-1} \end{pmatrix}$ with $\|d\|_\infty \leq \rho^{1/2}$ are $\succeq 0$ if and only if $d^T (A^{-1})^{-1} d = d^T Ad \leq 1$ for all d, $\|d\|_\infty \leq \rho^{1/2}$, i.e., if and only if $\rho \cdot OPT \leq 1$; we have derived (3.4.35). We now have

$$\begin{array}{lcr}
(a) & \frac{1}{OPT} \geq \rho & \\
& \Updownarrow & \text{[by (3.4.35)]} \\
& \begin{pmatrix} 1 & d^T \\ d & A^{-1} \end{pmatrix} \succeq 0 \quad \forall (d : \|d\|_\infty \leq \rho^{1/2}) & \\
& \Updownarrow & \text{[the Schur Complement Lemma]} \\
& A^{-1} \succeq \rho dd^T \;\; \forall (d, \|d\|_\infty \leq 1) & \\
& \Updownarrow & \\
& x^T A^{-1} x \geq \rho (d^T x)^2 \;\; \forall x \, \forall (d : \|d\|_\infty \leq 1) & \\
& \Updownarrow & \\
& x^T A^{-1} x \geq \rho \|x\|_1^2 \;\; \forall x & \\
& \Updownarrow & \\
(b) & A^{-1} \succeq \rho Y \;\; \forall (Y = Y^T : |Y_{ij}| \leq 1 \forall i, j) &
\end{array}$$

where the concluding $\Updownarrow$ is given by the evident relation

$$\|x\|_1^2 = \max_Y \left\{x^T Y x : Y = Y^T, |Y_{ij}| \leq 1 \; \forall i, j\right\}.$$

The equivalence $(a) \Leftrightarrow (b)$ is exactly (3.4.36). ■

By (3.4.36), $\frac{1}{OPT}$ is exactly the maximum R of those ρ for which the matrix cube

$$\mathcal{C}_\rho = \{A^{-1} + \sum_{1 \leq i \leq j \leq n} z_{ij} S^{ij} \big| \max_{i,j} |z_{ij}| \leq \rho\}$$

is contained in $\mathbf{S}^n_+$; here S^{ij} are the "basic symmetric matrices" (S^{ii} has a single nonzero entry, equal to 1, in the cell ii, and S^{ij}, $i < j$, has exactly two nonzero entries, equal to 1, in the cells ij and ji). Since

the ranks of the matrices S^{ij} do not exceed 2, Proposition 3.4.4 and Theorem 3.4.5 say that the optimal value in the semidefinite program

$$\rho(A) = \max_{\rho, X^{ij}} \left\{ \rho \;\middle|\; \begin{array}{c} X^{ij} \succeq \rho S^{ij},\ X^{ij} \succeq -\rho S^{ij},\ 1 \le i \le j \le n, \\ \sum\limits_{i \le j} X^{ij} \preceq A^{-1} \end{array} \right\} \tag{S}$$

is a lower bound for R, and this bound coincides with R up to the factor $\frac{\pi}{2}$; consequently, $\frac{1}{\rho(A)}$ is an upper bound on OPT, and this bound is at most $\frac{\pi}{2}$ times larger than OPT. It remains to note that a direct computation demonstrates that $\frac{1}{\rho(A)}$ is exactly the quantity SDP given by (3.4.33).

3.4.3 Robust Quadratic Programming

The concept of robust counterpart of an optimization problem with uncertain data (see Section 2.4.1) is in no sense restricted to Linear Programming. Whenever we have an optimization problem depending on certain data, we may ask what happens when the data are uncertain and all we know is an uncertainty set the data belong to. Given such an uncertainty set, we may require from candidate solutions to be *robust feasible* – to satisfy the realizations of the constraints for all data running through the uncertainty set. The robust counterpart of an uncertain problem is the problem of minimizing the objective[17] over the set of robust feasible solutions.

Now, we have seen in Section 2.4.1 that the "robust form" of an uncertain linear inequality with the coefficients varying in an ellipsoid is a conic quadratic inequality; as a result, the robust counterpart of an uncertain LP problem with ellipsoidal uncertainty (or, more general, with a CQr uncertainty set) is a conic quadratic problem. What is the "robust form" of an uncertain *conic quadratic* inequality

$$\|Ax + b\|_2 \le c^T x + d \qquad [A \in \mathbf{M}^{m,n}, b \in \mathbf{R}^m, c \in \mathbf{R}^n, d \in \mathbf{R}] \tag{3.4.37}$$

with uncertain data $(A, b, c, d) \in \mathcal{U}$? The question is how to describe the set of all robust feasible solutions of this inequality, i.e., the set of x's such that

$$\|Ax + b\|_2 \le c^T x + d \quad \forall (A, b, c, d) \in \mathcal{U}. \tag{3.4.38}$$

We intend to focus on the case when the uncertainty is "side-wise" – the data (A, b) of the left hand side and the data (c, d) of the right hand side of the inequality (3.4.37) *independently of each other* run through respective uncertainty sets $\mathcal{U}_\rho^{\rm left}$, $\mathcal{U}^{\rm right}$ (ρ is the left hand side uncertainty level). It suffices to assume the right hand side uncertainty set to be SDr with a strictly feasible SDR:

$$\mathcal{U}^{\rm right} = \{(c, d) \mid \exists u : \mathcal{P}c + Qd + \mathcal{R}u \succeq S\}. \tag{3.4.39}$$

As about the left hand side uncertainty set, we assume that it is an intersection of concentric ellipsoids, specifically, that

$$\mathcal{U}_\rho^{\rm left} = \left\{ [A, b] = [A_*, b_*] + \sum_{\ell=1}^{L} \zeta_\ell [A_\ell, b_\ell] : \zeta^T Q_j \zeta \le \rho^2,\ j = 1, ..., J \right\}, \tag{3.4.40}$$

where $Q_1, ..., Q_J$ are positive semidefinite matrices with positive definite sum.

Since the left hand side and the right hand side data independently of each other run through respective uncertainty sets, a point x is robust feasible if and only if there exists a real τ such that

$$\begin{array}{lrcl} (a) & \tau & \le & c^T x + d \quad \forall (c, d) \in \mathcal{U}^{\rm right}, \\ (b) & \|Ax + b\|_2 & \le & \tau \quad \forall [A, b] \in \mathcal{U}_\rho^{\rm left}. \end{array} \tag{3.4.41}$$

[17] Without loss of generality, we may assume that the objective is "certain" – is not affected by the data uncertainty. Indeed, we can always ensure this situation by passing to an equivalent problem with linear (and standard) objective:

$$\min_x \{f(x) : x \in X\} \mapsto \min_{t,x} \{t : f(x) - t \le 0, x \in X\}.$$

We know from the previous Lecture that the set of (τ, x) satisfying (3.4.41.a) is SDr (see Proposition 2.4.2 and Remark 2.4.1); it is easy to verify that the corresponding SDR is as follows:

$$\begin{array}{ll} (a) & (x,\tau) \text{ satisfies (3.4.41.}a\text{)} \\ & \Updownarrow \\ & \exists \Lambda : \\ (b) & \Lambda \succeq 0,\ \mathcal{P}^*\Lambda = x,\ \mathrm{Tr}(Q\Lambda) = 1,\ \mathcal{R}^*\Lambda = 0,\ \mathrm{Tr}(S\Lambda) \geq \tau. \end{array} \tag{3.4.42}$$

As about building SDR of the set of pairs (τ, x) satisfying (3.4.41.b), this is much more difficult (and in many cases even hopeless) task, since (3.4.38) in general turns out to be NP-hard and as such cannot be posed as an explicit semidefinite program. We can, however, build a kind of "inner approximation" of the set in question. To this end we shall use the ideas of semidefinite relaxation. Specifically, let us set

$$a[x] = A_* x + b_*, \quad A[x] = [A_1 x + b_1, ..., A_L x + b_L],$$

so that

$$(A_* + \sum_{\ell=1}^{L} \zeta_\ell A_\ell)x + (b_* + \sum_{\ell=1}^{L} \zeta_\ell b_\ell) = a[x] + A[x]\zeta.$$

In view of the latter identity, relation (3.4.41.b) reads

$$\|a[x] + \rho A[x]\zeta\|_2 \leq \tau \quad \forall(\zeta : \zeta^T Q_j \zeta \leq 1,\ j = 1, ..., J),$$

or, which is the same (set $\zeta = t^{-1}\xi$), as

$$\|ta[x] + \rho A[x]\xi\|_2 \leq \tau t^2 \quad \forall((t,\xi) : \xi^T Q_j \xi \leq t^2,\ j = 1, ..., J),$$

which in turn is equivalent to

$$\{\tau \geq 0\}_{\mathrm{I}} \ \& \ \left\{ \begin{array}{c} t^2(\tau^2 - a^T[x]a[x]) - 2t\rho a^T[x]A[x]\xi - \rho^2 \xi^T A^T[x]A[x]\xi \geq 0 \\ \forall((\xi,t) : \xi^T Q_j \xi \leq t^2,\ j = 1, ..., J) \end{array} \right\}_{\mathrm{II}}.$$

Predicate $\{\cdot\}_{\mathrm{II}}$ requires from certain quadratic form of t, ξ to be nonnegative when a number of other quadratic forms of these variables are nonnegative. An evident sufficient condition for this is that the former quadratic form is $\succeq$ a linear combination, with nonnegative coefficients, of the latter forms. When $\tau \geq 0$, this sufficient condition for the predicate $\{\cdot\}_{\mathrm{II}}$ to be valid can be reduced to the existence of nonnegative weights λ_j such that the quadratic form

$$t^2(\tau^2 - a^T[x]a[x]) - 2t\rho a^T[x]A[x]\xi - \rho^2 \xi^T A^T[x]A[x]\xi - \tau \sum_j \lambda_j (t^2 - \xi^T Q_j \xi)$$

in variables t, ξ is positive semidefinite. This condition is the same as the existence of nonnegative λ_j such that

$$\tau \begin{pmatrix} \tau - \sum\limits_j \lambda_j & \\ & \sum\limits_j \lambda_j Q_j \end{pmatrix} - [a[x], \rho A[x]]^T [a[x], \rho A[x]] \succeq 0.$$

Invoking the Schur Complement Lemma, the latter condition, in turn, is equivalent to the existence of nonnegative λ_j such that the matrix $\begin{pmatrix} \tau - \sum\limits_j \lambda_j & & a^T[x] \\ & \sum\limits_j \lambda_j Q_j & \rho A^T[x] \\ a[x] & \rho A[x] & \tau I \end{pmatrix}$ is positive semidefinite. We have established the implication as follows:

$$\begin{array}{lll} (a) & \{\tau \succeq 0\} \ \& & \left\{ \exists(\lambda_j \geq 0) : \begin{pmatrix} \tau - \sum\limits_j \lambda_j & & a^T[x] \\ & \sum\limits_j \lambda_j Q_j & \rho A^T[x] \\ a[x] & \rho A[x] & \tau I \end{pmatrix} \succeq 0 \right\} \\ & \Downarrow & \\ (b) & (x,\tau) \text{ satisfies (3.4.41.}b\text{)} & \end{array} \tag{3.4.43}$$

Combining our observations, we arrive at the first – easy – part of the following statement:

Proposition 3.4.6 *Let the data in the conic quadratic inequality* (3.4.37) *be affected by side-wise uncertainty* (3.4.39), (3.4.40). *Then*

(i) *The system* (S[ρ]) *of LMIs* (3.4.42.b), (3.4.43.a) *in variables* $x, \tau, \Lambda, \{\lambda_j\}$ *is a "conservative approximation" of the Robust Counterpart of* (3.4.37) *in the sense that whenever* x *can be extended to a feasible solution of* (S[ρ]), x *is robust feasible for* (3.4.37), *the uncertainty set being* $\mathcal{U}_\rho^{\rm left} \times \mathcal{U}^{\rm right}$.

(ii) *The tightness of* (S[ρ]) *as an approximation to the robust counterpart of* (3.4.37) *can be quantified as follows: if* x *cannot be extended to a feasible solution of* (S[ρ]), *then* x *is not robust feasible for* (3.4.37), *the uncertainty set being* $\mathcal{U}_{\vartheta\rho}^{\rm left} \times \mathcal{U}^{\rm right}$. *Here the "tightness factor"* ϑ *can be bounded as follows:*

1. *In the case of* $J = 1$ *(i.e., perturbations* ζ *are varying in an ellipsoid rather than in an intersection of concentric ellipsoids), one has* $\vartheta = 1$ *(i.e.,* (S[ρ]) *is exactly equivalent to the robust counterpart of* $\mathcal{U}_{\vartheta\rho}^{\rm left} \times \mathcal{U}^{\rm right}$*);*

2. *In the case of "box uncertainty"* $J = \dim \zeta$, $\zeta^T Q_j \zeta = \zeta_j^2$, *one has* $\vartheta = \frac{\pi}{2} = 1.570...$*;*

3. *In the general case,*

$$\vartheta = \sqrt{2 \ln \left(6 \sum_{j=1}^{J} \mathrm{Rank}(Q_j) \right)}.$$

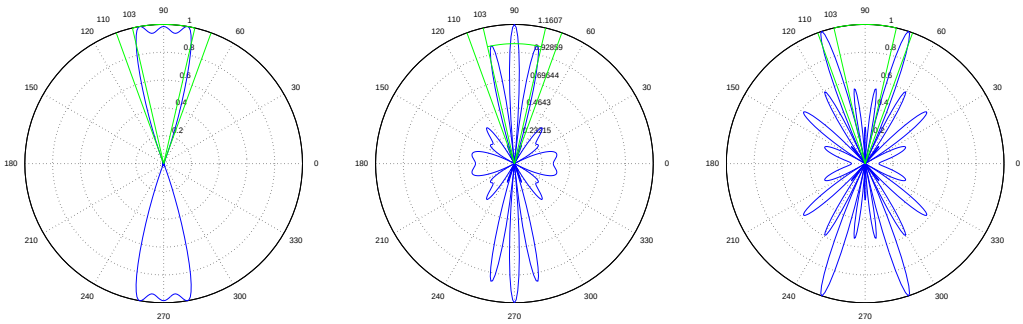
For the proof of the "difficult part" (ii) of the Proposition, see [4].

Example: Antenna Synthesis revisited. To illustrate the potential of the Robust Optimization methodology as applied to conic quadratic problems, consider the Circular Antenna Design problem from Section 2.4.1. Assume that now we deal with 40 ring-type antenna elements, and that our goal is to minimize the (discretized) L_2-distance from the synthesized diagram $\sum_{j=1}^{40} x_j D_{r_{j-1},r_j}(\cdot)$ to the "ideal" diagram $D_*(\cdot)$ which is equal to 1 in the range $77^o \le \theta \le 90^o$ and is equal to 0 in the range $0^o \le \theta \le 70^o$. The associated problem is just the Least Squares problem

$$\min_{\tau,x} \left\{ \tau : \underbrace{\sqrt{\frac{\sum\limits_{\theta\in\Theta_{\rm cns}} D_x^2(\theta) + \sum\limits_{\theta\in\Theta_{\rm obj}} (D_x(\theta)-1)^2}{\mathrm{card}(\Theta_{\rm cns}\cup\Theta_{\rm obj})}}_{\|D_*-D_x\|_2} \le \tau \right\}, \quad D_x(\theta) = \sum_{j=1}^{40} x_j D_{r_{j-1},r_j}(\theta) \tag{3.4.44}$$

where $\Theta_{\rm cns}$ and $\Theta_{\rm obj}$ are the intersections of the 240-point grid on the segment $0 \le \theta \le 90^o$ with the "angle of interest" $77^o \le \theta \le 90^o$ and the "sidelobe angle" $0^o \le \theta \le 70^o$, respectively.

The *Nominal Least Squares design* obtained from the optimal solution to this problem is completely unstable w.r.t. small implementation errors $x_j \mapsto (1+\xi_j)x_j$, $|\xi_j| \le \rho$:



Dream
no errors
$\|D_* - D\|_2 = 0.014$

Reality
0.1% errors
$\|D_* - D\|_2 \in [0.17, 0.89]$

Reality
2% errors
$\|D_* - D\|_2 \in [2.9, 19.6]$

Nominal Least Squares design: dream and reality
Data over a 100-diagram sample

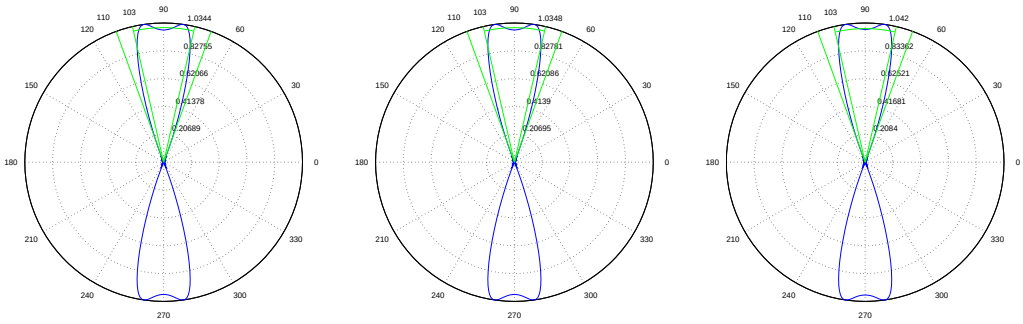
In order to take into account implementation errors, we should treat (3.4.44) as an uncertain conic quadratic problem

$$\left\{\min_{\tau,x}\{\tau : \|Ax - b\|_2 \le \tau\} \,\middle|\, A \in \mathcal{U}\right\}$$

with the uncertainty set of the form

$$\mathcal{U} = \{A = A_* + A_*\mathrm{Diag}(\xi) \mid \|\xi\|_\infty \le \rho\},$$

which is a particular case of the ellipsoidal uncertainty (specifically, what was called "box uncertainty" in Proposition 3.4.6). In the experiments to be reported, we use $\rho = 0.02$. The approximate Robust Counterpart (S[ρ]) of our uncertain conic quadratic problem yields the Robust design as follows:



Dream
no errors
$\|D_* - D\|_2 = 0.025$

Reality
0.1% errors
$\|D_* - D\|_2 \approx 0.025$

Reality
2% errors
$\|D_* - D\|_2 \approx 0.025$

Robust Least Squares design: dream and reality
Data over a 100-diagram sample

3.5 $\mathcal{S}$-Lemma and Approximate $\mathcal{S}$-Lemma

3.5.1 $\mathcal{S}$-Lemma

Let us look again at the Lagrange relaxation of a quadratically constrained quadratic problem, but in the very special case when all the forms involved are homogeneous, and the right hand sides of the inequality constraints are zero:

$$\begin{array}{ll} \text{minimize} & x^T B x \\ \text{s.t.} & x^T A_i x \ge 0,\ i = 1, ..., m \end{array} \tag{3.5.1}$$

($B, A_1, ..., A_m$ are given symmetric $m \times m$ matrices). Assume that the problem is feasible. In this case (3.5.1) is, at a first glance, a trivial problem: due to homogeneity, its optimal value is either $-\infty$ or 0, depending on whether there exists or does not exist a feasible vector x such that $x^T Bx < 0$. The challenge here is to detect which one of these two alternatives takes place, i.e., to understand whether or not a homogeneous quadratic inequality $x^T Bx \geq 0$ is a consequence of the system of homogeneous quadratic inequalities $x^T A_i x \geq 0$, or, which is the same, to understand when the implication

$$\begin{array}{ll} (a) & x^T A_i x \geq 0,\ i = 1, ..., m \\ & \Downarrow \\ (b) & x^T Bx \geq 0 \end{array} \tag{3.5.2}$$

holds true.

In the case of homogeneous *linear* inequalities it is easy to recognize when an inequality $x^T b \geq 0$ is a consequence of the system of inequalities $x^T a_i \geq 0$, $i = 1, ..., m$: by Farkas Lemma, it is the case if and only if the inequality is a *linear* consequence of the system, i.e., if b is representable as a linear combination, with nonnegative coefficients, of the vectors a_i. Now we are asking a similar question about homogeneous quadratic inequalities: when (b) is a consequence of (a)?

In general, there is no analogy of the Farkas Lemma for homogeneous quadratic inequalities. Note, however, that the easy "if" part of the Lemma can be extended to the quadratic case: <u>if</u> the target inequality (b) can be obtained by linear aggregation of the inequalities (a) and a trivial – identically true – inequality, <u>then</u> the implication in question is true. Indeed, a linear aggregation of the inequalities (a) is an inequality of the type

$$x^T (\sum_{i=1}^m \lambda_i A_i) x \geq 0$$

with nonnegative weights λ_i, and a trivial – identically true – homogeneous quadratic inequality is of the form

$$x^T Q x \geq 0$$

with $Q \succeq 0$. The fact that (b) can be obtained from (a) and a trivial inequality by linear aggregation means that B can be represented as $B = \sum\limits_{i=1}^m \lambda_i A_i + Q$ with $\lambda_i \geq 0$, $Q \succeq 0$, or, which is the same, if $B \succeq \sum\limits_{i=1}^m \lambda_i A_i$ for certain nonnegative λ_i. If this is the case, then (3.5.2) is trivially true. We have arrived at the following simple

Proposition 3.5.1 *Assume that there exist nonnegative λ_i such that $B \succeq \sum\limits_i \lambda_i A_i$. Then the implication* (3.5.2) *is true.*

Proposition 3.5.1 is no more than a *sufficient* condition for the implication (3.5.2) to be true, and in general this condition is <u>not</u> necessary. There is, however, an extremely fruitful particular case when the condition is both necessary and sufficient – this is the case of $m = 1$, i.e., a single quadratic inequality in the premise of (3.5.2):

Theorem 3.5.1 [$\mathcal{S}$-Lemma] *Let A, B be symmetric $n \times n$ matrices, and assume that the quadratic inequality*

$$x^T Ax \geq 0 \tag{A}$$

is strictly feasible: there exists $\bar{x}$ such that $\bar{x}^T A \bar{x} > 0$. Then the quadratic inequality

$$x^T Bx \geq 0 \tag{B}$$

is a consequence of (A) *if and only if it is a linear consequence of* (A)*, i.e., if and only if there exists a nonnegative λ such that*

$$B \succeq \lambda A.$$

We are about to present an "intelligent" proof of the $\mathcal{S}$-Lemma based on the ideas of semidefinite relaxation.

In view of Proposition 3.5.1, all we need is to prove the "only if" part of the $\mathcal{S}$-Lemma, i.e., to demonstrate that if the optimization problem

$$\min_x \left\{x^T Bx : x^T Ax \geq 0\right\}$$

is strictly feasible and its optimal value is ≥ 0, then $B \succeq \lambda A$ for certain $\lambda \geq 0$. By homogeneity reasons, it suffices to prove exactly the same statement for the optimization problem

$$\min_x \left\{x^T Bx : x^T Ax \geq 0, x^T x = n\right\}. \tag{P}$$

The standard semidefinite relaxation of (P) is the problem

$$\min_X \left\{\mathrm{Tr}(BX) : \mathrm{Tr}(AX) \geq 0, \mathrm{Tr}(X) = n, X \succeq 0\right\}. \tag{P$'$}$$

If we could show that when passing from the original problem (P) *to the relaxed problem* (P$'$) *the optimal value* (which was nonnegative for (P)) *remains nonnegative, we would be done.* Indeed, observe that (P$'$) is clearly bounded below (its feasible set is compact!) and is strictly feasible (which is an immediate consequence of the strict feasibility of (A)). Thus, by the Conic Duality Theorem the problem dual to (P$'$) is solvable with the same optimal value (let it be called $n\theta^*$) as the one in (P$'$). The dual problem is

$$\max_{\mu,\lambda} \left\{n\mu : \lambda A + \mu I \preceq B, \lambda \geq 0\right\},$$

and the fact that its optimal value is $n\theta^*$ means that there exists a nonnegative λ such that

$$B \succeq \lambda A + n\theta^* I.$$

If we knew that the optimal value $n\theta^*$ in (P$'$) *is nonnegative*, we would conclude that $B \succeq \lambda A$ for certain nonnegative λ, which is exactly what we are aiming at. Thus, all we need is to prove that *under the premise of the $\mathcal{S}$-Lemma the optimal value in* (P$'$) *is nonnegative,* and here is the proof:

Observe first that problem (P$'$) is feasible with a compact feasible set, and thus is solvable. Let X^* be an optimal solution to the problem. Since $X^* \geq 0$, there exists a matrix D such that $X^* = DD^T$. Note that we have

$$\begin{array}{c} 0 \leq \mathrm{Tr}(AX^*) = \mathrm{Tr}(ADD^T) = \mathrm{Tr}(D^T AD), \\ n\theta^* = \mathrm{Tr}(BX^*) = \mathrm{Tr}(BDD^T) = \mathrm{Tr}(D^T BD), \\ n = \mathrm{Tr}(X^*) = \mathrm{Tr}(DD^T) = \mathrm{Tr}(D^T D). \end{array} \tag{*}$$

It remains to use the following observation

> (!) *Let P, Q be symmetric matrices such that $\mathrm{Tr}(P) \geq 0$ and $\mathrm{Tr}(Q) < 0$. Then there exists a vector e such that $e^T Pe \geq 0$ and $e^T Qe < 0$.*

Indeed, let us believe that (!) is valid, and let us prove that $\theta^* \geq 0$. Assume, on the contrary, that $\theta^* < 0$. Setting $P = D^T BD$ and $Q = D^T AD$ and taking into account (*), we see that the matrices P, Q satisfy the premise in (!), whence, by (!), there exists a vector e such that $0 \leq e^T Pe = [De]^T A[De]$ and $0 > e^T Qe = [De]^T B[De]$, which contradicts the premise of the $\mathcal{S}$-Lemma.

It remains to prove (!). Given P and Q as in (!), note that Q, as every symmetric matrix, admits a representation

$$Q = U^T \Lambda U$$

with an orthonormal U and a diagonal Λ. Note that $\theta \equiv \mathrm{Tr}(\Lambda) = \mathrm{Tr}(Q) < 0$. Now let ξ be a random n-dimensional vector with independent entries taking values ± 1 with probabilities $1/2$. We have

$$[U^T\xi]^T Q[U^T\xi] = [U^T\xi]^T U^T \Lambda U[U^T\xi] = \xi^T \Lambda \xi = \mathrm{Tr}(\Lambda) = \theta \quad \forall \xi,$$

while

$$[U^T\xi]^T P[U^T\xi] = \xi^T[UPU^T]\xi,$$

and the expectation of the latter quantity over ξ is clearly $\mathrm{Tr}(UPU^T) = \mathrm{Tr}(P) \geq 0$. Since the expectation is nonnegative, there is at least one realization $\bar{\xi}$ of our random vector ξ such that

$$0 \leq [U^T\bar{\xi}]^T P[U^T\bar{\xi}].$$

We see that the vector $e = U^T\bar{\xi}$ is a required one: $e^TQe = \theta < 0$ and $e^TPe \geq 0$. ∎

3.5.2 Inhomogeneous $\mathcal{S}$-Lemma

Proposition 3.5.2 [Inhomogeneous $\mathcal{S}$-Lemma] *Consider optimization problem with quadratic objective and a single quadratic constraint:*

$$f_* = \min_x \left\{ f_0(x) \equiv x^TA_0x + 2b_0^Tx + c_0 : f_1(x) \equiv x^TA_1x + 2b_1^Tx + c_1 \leq 0 \right\} \tag{3.5.3}$$

Assume that the problem is strictly feasible and below bounded. Then the Semidefinite relaxation (3.4.5) *of the problem is solvable with the optimal value f_*.*

Proof. By Proposition 3.4.1, the optimal value in (3.4.5) can be only $\leq f_*$. Thus, it suffices to verify that (3.4.5) admits a feasible solution with the value of the objective $\geq f_*$, that is, that there exists $\lambda_* \geq 0$ such that

$$\begin{pmatrix} c_0 + \lambda_* c_1 - f_* & b_0^T + \lambda_* b_1^T \\ b_0 + \lambda_* b_1 & A_0 + \lambda_* A_1 \end{pmatrix} \succeq 0. \tag{3.5.4}$$

To this end, let us associate with (3.5.3) a pair of homogeneous quadratic forms of the extended vector of variables $y = (t, x)$, where $t \in \mathbf{R}$, specifically, the forms

$$y^TPy \equiv x^TA_1x + 2tb_1^Tx + c_1t^2,\ y^TQy = -x^TA_0y - 2tb_0^Tx - (c_0 - f_*)t^2.$$

We claim that, first, there exist $\epsilon_0 > 0$ and $\bar{y}$ with $\bar{y}^TP\bar{y} < -\epsilon_0\bar{y}^T\bar{y}$ and, second, that for every $\epsilon \in (0, \epsilon_0]$ the implication

$$y^TPy \leq -\epsilon y^Ty \Rightarrow y^TQy \leq 0 \tag{3.5.5}$$

holds true. The first claim is evident: by assumption, there exists $\bar{x}$ such that $f_1(\bar{x}) < 0$; setting $\bar{y} = (1, \bar{x})$, we see that $\bar{y}^TP\bar{y} = f_1(\bar{x}) < 0$, whence $\bar{y}^TP\bar{y} < -\epsilon_0\bar{y}^T\bar{y}$ for appropriately chosen $\epsilon_0 > 0$. To support the second claim, assume that $y = (t, x)$ is such that $y^TPy \leq -\epsilon y^Ty$, and let us prove that then $y^TQy \leq 0$.

- Case 1: $t \neq 0$. Setting $y' = t^{-1}y = (1, x')$, we have $f_1(x') = [y']^TPy' = t^{-2}y^TPy \leq 0$, whence $f_0(x') \geq f_*$, or, which is the same, $[y']^TQy' \leq 0$, so that $y^TQy \leq 0$, as required in (3.5.5).
- Case 2: $t = 0$. In this case, $-\epsilon x^Tx = -\epsilon y^Ty \geq y^TPy = x^TA_1x$ and $y^TQy = -x^TA_0x$, and we should prove that the latter quantity is nonpositive. Assume, on the contrary, that this quantity is positive, that is, $x^TA_0x < 0$. Then $x \neq 0$ and therefore $x^TA_1x \leq -\epsilon x^Tx < 0$. From $x^TA_1x < 0$ and $x^TA_0x < 0$ it follows that $f_1(sx) \to -\infty$ and $f_0(sx) \to -\infty$ as $s \to +\infty$, which contradicts the assumption that (3.5.3) is below bounded. Thus, $y^TQy \leq 0$.

Our observations combine with $\mathcal{S}$-Lemma to imply that for every $\epsilon \in (0, \epsilon_0]$ there exists $\lambda = \lambda_\epsilon \geq 0$ such that

$$B \preceq \lambda_\epsilon(A + \epsilon I), \tag{3.5.6}$$

whence, in particular,

$$\bar{y}^TB\bar{y} \leq \lambda_\epsilon\bar{y}^T[A + \epsilon I]\bar{y}.$$

The latter relation, due to $\bar{y}^TA\bar{y} < 0$, implies that λ_ϵ remains bounded as $\epsilon \to +0$. Thus, we have $\lambda_{\epsilon_i} \to \lambda_* \geq 0$ as $i \to \infty$ for a properly chosen sequence $\epsilon_i \to +0$ of values of ϵ, and (3.5.6) implies that $B \preceq \lambda_*A$. Recalling what are A and B, we arrive at (3.5.4). ∎

3.5.3 Approximate $\mathcal{S}$-Lemma

In general, the $\mathcal{S}$-Lemma fails to be true when there is more than a single quadratic form in (3.5.2) (that is, when $m > 1$). Similarly, Inhomogeneous $\mathcal{S}$-Lemma fails to be true for general quadratic quadratically constrained problems with more than a single quadratic constraint. There exists, however, a useful *approximate* version of the Inhomogeneous $\mathcal{S}$-Lemma in the "multi-constrained" case which is as follows:

Proposition 3.5.3 [Approximate $\mathcal{S}$-Lemma] *Consider the following quadratic quadratically constrained optimization problem:*

$$\mathrm{Opt} = \max_x \left\{ f(x) = x^T A x + 2b^T x : x^T A_i x \le c_i,\ i = 1, ..., m \right\}, \tag{3.5.7}$$

where $c_i > 0$, $A_i \succeq 0$, $i = 1, ..., m$ (A can be arbitrary symmetric matrix) and $\sum_{i=1}^m A_i \succ 0$. Let SDP *be the optimal value in the Semidefinite relaxation of this problem:*

$$\mathrm{SDP} = \min_{\omega,\lambda} \left\{ \omega : \begin{pmatrix} \omega - \sum_{i=1}^m c_i \lambda_i & -b^T \\ -b & \sum_{i=1}^m \lambda_i A_i - A \end{pmatrix} \succeq 0, \lambda \ge 0 \right\} \tag{3.5.8}$$

(note that the problem of interest is a maximization one, whence the difference between the relaxation and (3.4.5)). *Then*

$$\mathrm{Opt} \le \mathrm{SDP} \le \Theta \mathrm{Opt}, \tag{3.5.9}$$

where $\Theta = 1$ in the case of $m = 1$ and

$$\Theta = 2 \ln \left(6 \sum_{i=1}^m \mathrm{Rank}(A_i) \right)$$

in the case of $m > 1$. Moreover, in the latter case there exists x_* such that

$$\begin{array}{rcl} b^T x_* & \ge & 0, \\ x_*^T A x_* + 2b^T x_* & \ge & \mathrm{SDP}, \\ x_*^T A_i x^* & \le & \Theta c_i,\ i = 1, ..., m. \end{array} \tag{3.5.10}$$

Proof. The case of $m = 1$ is given by the Inhomogeneous $\mathcal{S}$-Lemma. Thus, let $m > 1$.

1^0. We clearly have

$$\begin{array}{rcl} \mathrm{Opt} & = & \max\limits_{t,x} \left\{ x^T A x + 2tb^T x : t^2 \le 1, x^T A_i x \le c_i,\ i = 1, ..., m \right\} \\ & = & \max\limits_{z=(t,x)} \left\{ z^T B z : z^T B_i z \le c_i,\ i = 0, 1, ..., m \right\}, \\ & & B = \left[\begin{array}{c|c} & b^T \\ \hline b & A \end{array} \right], \\ & & B_0 = \left[\begin{array}{c|c} 1 & \\ \hline & \end{array} \right], \\ & & B_i = \left[\begin{array}{c|c} & \\ \hline & A_i \end{array} \right],\ i = 1, ..., m, \\ & & c_0 = 1. \end{array} \tag{3.5.11}$$

Note that (3.5.8) is nothing but the semidefinite dual of the semidefinite program

$$\max_Z \left\{ \mathrm{Tr}(BZ) : \mathrm{Tr}(B_i Z) \le c_i,\ i = 0, 1, ..., m, Z \succeq 0 \right\}. \tag{3.5.12}$$

Since $c_i > 0$ for $i \ge 0$, (3.5.12) is strictly feasible, and since $\sum_{i=0}^m B_i \succ 0$, the feasible set of the problem is bounded (so that the problem is solvable). Since (3.5.12) is strictly feasible and bounded, by Semidefinite Duality Theorem we have

$$\mathrm{SDP} = \max_Z \left\{ \mathrm{Tr}(BZ) : \mathrm{Tr}(B_i Z) \le c_i,\ i = 0, 1, ..., m, Z \succeq 0 \right\}. \tag{3.5.13}$$

2^0. Let Z_* be an optimal solution to (3.5.12). Let us set

$$\begin{array}{rcl}\widehat{B} & = & Z_*^{1/2} B Z_*^{1/2} \\ \widehat{B}_i & = & Z_*^{1/2} B_i Z_*^{1/2},\ i = 0, ..., m,\end{array}$$

and let $\widehat{B} = U^T D U$ be the eigenvalue decomposition of $\widehat{B}$ (so that U is orthogonal and D is diagonal). Finally, let us set $D_i = U \widehat{B}_i U^T$. Thus, we have arrived at the symmetric matrices $D, D_0, D_1, ..., D_m$ such that

1) $D = U Z_*^{1/2} B Z_*^{1/2} U^T$ is diagonal, and $\mathrm{Tr}(D) = \mathrm{Tr}(Z_* B) = \mathrm{SDP}$;
2) For $i = 0, 1, ..., m$, the matrices $D_i = U Z_*^{1/2} B_i Z_*^{1/2} U^T$ are symmetric positive semidefinite, $\mathrm{Rank}(D_i) \le \mathrm{Rank}(B_i)$, and $\mathrm{Tr}(D_i) = \mathrm{Tr}(Z_* B_i) \le c_i$.

Now let ξ be a random vector with independent coordinates taking values ± 1 with probability $1/2$, and let $\eta = Z_*^{1/2} U^T \xi$. Observe that

$$\begin{array}{lrcll} (a) & \eta^T B \eta & \equiv & \xi^T D \xi & \\ & & \equiv & \mathrm{SDP} & [\text{by 1)}] \\ (b) & \mathbf{E}\left\{\eta^T B_i \eta\right\} & = & \mathbf{E}\left\{\xi^T D_i \xi\right\} = \mathrm{Tr}(D_i) & \\ & & \le & c_i,\ i = 0, 1, ..., m & [\text{by 2)}] \end{array} \tag{3.5.14}$$

3^0. We claim that from (3.5.14.b) it follows that

$$\begin{array}{lrcl} (a) & \mathrm{Prob}\left\{\eta^T B_0 \eta > 1\right\} & \le & \frac{2}{3} \\ (b) & \mathrm{Prob}\left\{\eta^T B_i \eta > \theta c_i\right\} & \le & \mathrm{Rank}(A_i) \exp\{-\theta/2\},\ i = 1, ..., m \end{array} \tag{3.5.15}$$

Indeed, $\eta^T B_0 \eta = \xi^T D_0 \xi$. Note that $D_0 = d_0 d_0^T$ for certain vector d_0 (since $B_0 = b_0 b_0^T$ for certain b_0). Besides this, $\mathrm{Tr}(D_0) \le c_0 = 1$ by (3.5.14.b). Thus,

$$\eta^T B_0 \eta = (\sum_j p_j \epsilon_j)^2,$$

where deterministic p_j satisfy $\sum\limits_j p_j^2 \le 1$, and ϵ_j are independent random variables taking values ± 1 with probabilities $1/2$. Now (3.5.15.a) is given by the following fact [4]:

With p_j and ϵ_j as above, one has $\mathrm{Prob}\left\{|\sum\limits_j p_j \epsilon_j| \le 1\right\} \ge \frac{1}{3}$.

To verify (3.5.15.b), let us fix $i \ge 1$. Since D_i is positive semidefinite along with B_i, we have

$$D_i = \sum_{j=1}^k d_j d_j^T \qquad [k = \mathrm{Rank}(D_i) \le \mathrm{Rank}(B_i) = \mathrm{Rank}(A_i)].$$

By Bernstein's Inequality (see the proof of Proposition 2.4.1), we have

$$\mathrm{Prob}\left\{|d_j^T \xi| \ge \sqrt{\theta} \|d_j\|_2\right\} \le 2 \exp\{-\theta/2\}.$$

In the case of $\xi^T D_i \xi \ge \theta \sum\limits_{j=1}^k \|d_j\|_2^2$ we clearly have $\xi^T d_j d_j^T \xi \ge \theta \|d_j\|_2^2$ for certain (depending on ξ) value of j, so that

$$\begin{array}{rcl} \mathrm{Prob}\left\{\xi^T D_i \xi > \theta \sum\limits_{j=1}^k \|d_j\|_2^2\right\} & \le & \sum\limits_{j=1}^k \mathrm{Prob}\left\{|d_j^T \xi| \ge \sqrt{\theta} \|d_j\|_2\right\} \\ & \le & 2k \exp\{-\theta/2\}. \end{array}$$

The resulting inequality implies (3.5.15.b) due to the facts that $\eta^T B_i \eta = \xi^T D_i \xi$ and that

$$\sum_{j=1}^{k} \|d_j\|_2^2 = \mathrm{Tr}(\sum_j d_j d_j^T) = \mathrm{Tr}(D_i) \le c_i.$$

4^0. Let $K = \sum_{i=1}^{m} \mathrm{Rank}(A_i)$. For every $\theta > \Theta = 2\ln(6K)$ we have $\theta \ge 1$ and $\frac{2}{3} + 2K\exp\{-\theta/2\} < 1$. In view of the latter fact and (3.5.15), there exists a realization $\bar\eta = (\bar t, \bar x)$ of η such that

$$\begin{array}{lrcl} (a) & \bar t^2 & \equiv & \bar\eta^T B_0 \bar\eta \le 1, \\ (b) & \bar x^T A_i \bar x & \equiv & \bar\eta^T B_i \bar\eta \le \theta c_i,\ i = 1, ..., m. \end{array} \tag{3.5.16}$$

while

$$\bar x^T A \bar x + 2\bar t b^T \bar x = \bar\eta^T Bbar\eta = \mathrm{SDP}$$

by (3.5.14.a). Replacing, if necessary, $\bar t$ with $-\bar t$ and $\bar x$ with $-\bar x$, we ensure the validity of (3.5.16.b) along with the relation

$$\bar x^T A \bar x + \underbrace{2b^T \bar x}_{\ge 0} \ge \mathrm{SDP}. \tag{3.5.17}$$

Since $\sum_i A_i \succ 0$, relations (3.5.16) imply that $(\bar t, \bar x)$ remain bounded as $\theta \to +\Theta$, whence (3.5.16) and (3.5.17) are valid for properly chosen $\bar t, \bar x$ and $\theta = \Theta$; setting $x_* = \bar x$, we arrive at (3.5.10). By (3.5.10), $\Theta^{-1/2} x_*$ is a feasible solution of (3.5.7) and

$$\Theta^{-1} x_*^T A x_* + 2\Theta^{-1/2} b^T x_* \ge \Theta^{-1}\mathrm{SDP}, \tag{3.5.18}$$

whence $\mathrm{Opt} \ge \Theta^{-1}\mathrm{SDP}$. ∎

Application: Approximating the Affinely Adjustable Robust Counterpart of an Uncertain Linear Programming problem

The notion of Affinely Adjustable Robust Counterpart (AARC) of uncertain LP was introduced and motivated in Section 2.4.5. As applied to uncertain LP

$$\mathcal{LP} = \left\{ \min_x \left\{ c^T[\zeta] x : A[\zeta] x - b[\zeta] \ge 0 \right\} : \zeta \in \mathcal{Z} \right\} \tag{3.5.19}$$

affinely parameterized by perturbation vector ζ and with variables x_j allowed to be affine functions of $P_j\zeta$:

$$x_j = \mu_j + \nu_j^T P_j \zeta, \tag{3.5.20}$$

the AARC is the following semi-infinite optimization program in variables t, μ_j, ν_j:

$$\min_{t, \{\mu_j, \nu_j\}_{j=1}^n} \left\{ t : \begin{array}{c} \sum_j c_j[z][\mu_j + \nu_j^T P_j \zeta] \le t \ \forall \zeta \in \mathcal{Z} \\ \sum_j [\mu_j + \nu_j^T P_j] A_j[\zeta] - b[\zeta] \ge 0 \ \forall \zeta \in \mathcal{Z} \end{array} \right\} \tag{AARC}$$

It was explained that in the case of *fixed recourse* ($c_j[\zeta]$ and $A_j[\zeta]$ are independent of ζ for all j for which x_j is adjustable, that is, $P_j \ne 0$), (AARC) is equivalent to an explicit conic quadratic program, provided that the perturbation set $\mathcal{Z}$ is CQr with strictly feasible CQR. In fact CQ-representability plays no crucial role here (see Remark 2.4.1); in particular, when $\mathcal{Z}$ is SDr with a strictly feasible SDR, (AARC), in the case of fixed recourse, is equivalent to an explicit semidefinite program. What indeed plays a crucial role is the assumption of fixed recourse; it can be shown that when this assumption does not hold, (AARC) can be computationally intractable. Our current goal is to demonstrate that even in this difficult case

(AARC) admits a "tight" computationally tractable approximation, provided that $\mathcal{Z}$ is an intersection of ellipsoids centered at the origin:

$$\mathcal{Z} = \mathcal{Z}_\rho \equiv \left\{\zeta : \zeta^T Q_i \zeta \le \rho^2,\ i = 1, ..., m\right\} \qquad \left[Q_i \succeq 0, \sum_i Q_i \succ 0\right] \tag{3.5.21}$$

Indeed, since $A_j[\zeta]$ are affine in ζ, every semi-infinite constraint in (AARC) is of the form

$$\zeta^T A[z]\zeta + 2b^T[z]\zeta \le c[z] \quad \forall \zeta \in \mathcal{Z}_\rho \tag{3.5.22}$$

where $z = (t, \{\mu_j, \nu_j\}_{j=1}^n)$ is the vector of variables in (AARC), and $A[z]$, $b[z]$, $c[z]$ are affine in z matrix, vector and scalar. Applying Approximate $\mathcal{S}$-Lemma, a *sufficient* condition for (3.5.22) to be valid for a given z is the relation

$$\min_{\omega,\lambda}\left\{\omega : \begin{pmatrix} \omega - \sum\limits_{i=1}^m \rho^2\lambda_i & -b^T[z] \\ -b[z] & \sum\limits_{i=1}^m \lambda_i Q_i - A[z] \end{pmatrix} \succeq 0, \lambda \ge 0\right\} \le c[z],$$

or, which is the same, the possibility to extend z, by properly chosen λ, to a solution to the system of constraints

$$\lambda \ge 0, \begin{pmatrix} c[z] - \sum\limits_{i=1}^m \rho^2\lambda_i & -b^T[z] \\ -b[z] & \sum\limits_{i=1}^m \lambda_i Q_i - A[z] \end{pmatrix}, \tag{3.5.23}$$

which is a system of LMIs in variables (z, λ). Replacing every one of the semi-infinite constraints in (AARC) with corresponding system (3.5.23), we end up with an explicit semidefinite program which is a "conservative approximation" of (AARC): both problems have the same objective, and the z-component of a feasible solution to the approximation is feasible solution of (AARC). At the same time, the approximation is tight up to the quite moderate factor

$$\theta = \sqrt{2\ln\left(6\sum_{i=1}^m \mathrm{Rank}(Q_i)\right)}:$$

whenever z can<u>not</u> be extended to a feasible solution of the approximation, the "moreover" part of the Approximate $\mathcal{S}$-Lemma says that z becomes <u>in</u>feasible for (AARC) after the original level of perturbations is increased by the factor θ.

3.6 Extremal ellipsoids

We already have met, on different occasions, with the notion of an ellipsoid – a set E in $\mathbf{R}^n$ which can be represented as the image of the unit Euclidean ball under an affine mapping:

$$E = \{x = Au + c \mid u^T u \le 1\} \quad [A \in \mathbf{M}^{n,q}] \tag{Ell}$$

Ellipsoids are very convenient mathematical entities:

- it is easy to specify an ellipsoid – just to point out the corresponding matrix A and vector c;
- the family of ellipsoids is closed with respect to affine transformations: the image of an ellipsoid under an affine mapping again is an ellipsoid;
- there are many operations, like minimization of a linear form, computation of volume, etc., which are easy to carry out when the set in question is an ellipsoid, and is difficult to carry out for more general convex sets.

By the indicated reasons, ellipsoids play important role in different areas of applied mathematics; in particular, people use ellipsoids to approximate more complicated sets. Just as a simple motivating example, consider a discrete-time linear time invariant controlled system:

$$\begin{array}{rcl} x(t+1) & = & Ax(t)+Bu(t),\ t=0,1,... \\ x(0) & = & 0 \end{array}$$

and assume that the control is norm-bounded:

$$\|u(t)\|_2 \le 1 \quad \forall t.$$

The question is what is the set X_T of all states "reachable in a given time T", i.e., the set of all possible values of $x(T)$. We can easily write down the answer:

$$X_T = \{x = Bu_{T-1} + ABu_{T-2} + A^2Bu_{T-3} + ... + A^{T-1}Bu_0 \mid \|u_t\|_2 \le 1,\ t = 0, ..., T-1\},$$

but this answer is not "explicit"; just to check whether a given vector x belongs to X_T requires to solve a nontrivial conic quadratic problem, the complexity of the problem being the larger the larger is T. In fact the geometry of X_T may be very complicated, so that there is no possibility to get a "tractable" explicit description of the set. This is why in many applications it makes sense to use "simple" – ellipsoidal - approximations of X_T; as we shall see, approximations of this type can be computed in a recurrent and computationally efficient fashion.

It turns out that the natural framework for different problems of the "best possible" approximation of convex sets by ellipsoids is given by semidefinite programming. In this section we intend to consider a number of basic problems of this type.

Preliminaries on ellipsoids. According to our definition, an ellipsoid in $\mathbf{R}^n$ is the image of the unit Euclidean ball in certain $\mathbf{R}^q$ under an affine mapping; e.g., for us a segment in $\mathbf{R}^{100}$ is an ellipsoid; indeed, it is the image of one-dimensional Euclidean ball under affine mapping. In contrast to this, in geometry an ellipsoid in $\mathbf{R}^n$ is usually defined as the image of the n-dimensional unit Euclidean ball under an invertible affine mapping, i.e., as the set of the form (Ell) with additional requirements that $q = n$, i.e., that the matrix A is square, and that it is nonsingular. In order to avoid confusion, let us call these "true" ellipsoids *full-dimensional*. Note that a full-dimensional ellipsoid E admits two nice representations:

- First, E can be represented in the form (Ell) with *positive definite symmetric* A:

$$E = \{x = Au + c \mid u^T u \le 1\} \quad [A \in \mathbf{S}^n_{++}] \tag{3.6.1}$$

  Indeed, it is clear that if a matrix A represents, via (Ell), a given ellipsoid E, the matrix AU, U being an orthogonal $n \times n$ matrix, represents E as well. It is known from Linear Algebra that by multiplying a nonsingular square matrix from the right by a properly chosen orthogonal matrix, we get a positive definite symmetric matrix, so that we always can parameterize a full-dimensional ellipsoid by a positive definite symmetric A.

- Second, E can be given by a strictly convex quadratic inequality:

$$E = \{x \mid (x-c)^T D(x-c) \le 1\} \quad [D \in \mathbf{S}^n_{++}]. \tag{3.6.2}$$

  Indeed, one may take $D = A^{-2}$, where A is the matrix from the representation (3.6.1).

Note that the set (3.6.2) makes sense and is convex when the matrix D is positive semidefinite rather than positive definite. When $D \succeq 0$ is not positive definite, the set (3.6.1) is, geometrically, an "elliptic cylinder" – a shift of the direct product of a full-dimensional ellipsoid in the range space of D and the complementary to this range linear subspace – the kernel of D.

In the sequel we deal a lot with *volumes* of full-dimensional ellipsoids. Since an invertible affine transformation $x \mapsto Ax + b : \mathbf{R}^n \to \mathbf{R}^n$ multiplies the volumes of n-dimensional domains by $|\mathrm{Det}A|$,

the volume of a full-dimensional ellipsoid E given by (3.6.1) is $\kappa_n \mathrm{Det}A$, where κ_n is the volume of the n-dimensional unit Euclidean ball. In order to avoid meaningless constant factors, it makes sense to pass from the usual n-dimensional volume $\mathrm{mes}_n(G)$ of a domain G to its *normalized volume*

$$\mathrm{Vol}(G) = \kappa_n^{-1}\mathrm{mes}_n(G),$$

i.e., to choose, as the unit of volume, the volume of the unit ball rather than the one of the cube with unit edges. From now on, speaking about volumes of n-dimensional domains, we always mean their normalized volume (and omit the word "normalized"). With this convention, the volume of a full-dimensional ellipsoid E given by (3.6.1) is just

$$\mathrm{Vol}(E) = \mathrm{Det}A,$$

while for an ellipsoid given by (3.6.1) the volume is

$$\mathrm{Vol}(E) = [\mathrm{Det}D]^{-1/2}.$$

Outer and inner ellipsoidal approximations. It was already mentioned that our current goal is to realize how to solve basic problems of "the best" ellipsoidal approximation E of a given set S. There are two types of these problems:

- *Outer approximation*, where we are looking for the "smallest" ellipsoid E *containing* the set S;
- *Inner approximation*, where we are looking for the "largest" ellipsoid E *contained* in the set S.

In both these problems, a natural way to say when one ellipsoid is "smaller" than another one is to compare the volumes of the ellipsoids. The main advantage of this viewpoint is that it results in *affine-invariant* constructions: an invertible affine transformation multiplies volumes of all domains by the same constant and therefore preserves ratios of volumes of the domains.

Thus, what we are interested in are the *largest volume ellipsoid(s) contained in a given set S* and the *smallest volume ellipsoid(s) containing a given set S*. In fact these *extremal ellipsoids* are unique, provided that S is a *solid* – a closed and bounded convex set with a nonempty interior, and are not too bad approximations of the set:

Theorem 3.6.1 [Löwner – Fritz John] *Let $S \subset \mathbf{R}^n$ be a solid. Then*

(i) *There exists and is uniquely defined the largest volume full-dimensional ellipsoid E_{in} contained in S. The concentric to E_{in} n times larger (in linear sizes) ellipsoid contains S; if S is central-symmetric, then already $\sqrt{n}$ times larger than E_{in} concentric to E_{in} ellipsoid contains S.*

(ii) *There exists and is uniquely defined the smallest volume full-dimensional ellipsoid E_{out} containing S. The concentric to E_{out} n times smaller (in linear sizes) ellipsoid is contained in S; if S is central-symmetric, then already $\sqrt{n}$ times smaller than E_{out} concentric to E_{out} ellipsoid is contained in S.*

The proof is the subject of Exercise 3.37.

The existence of extremal ellipsoids is, of course, a good news; but how to compute these ellipsoids? The possibility to compute efficiently (nearly) extremal ellipsoids heavily depends on the description of S. Let us start with two simple examples.

Inner ellipsoidal approximation of a polytope. Let S be a polyhedral set *given by a number of linear equalities:*

$$S = \{x \in \mathbf{R}^n \mid a_i^T x \le b_i,\ i = 1, ..., m\}.$$

Proposition 3.6.1 *Assume that S is a full-dimensional polytope (i.e., is bounded and possesses a nonempty interior). Then the largest volume ellipsoid contained in S is*

$$E = \{x = Z_* u + z_* \mid u^T u \le 1\},$$

where Z_*, z_* are given by an optimal solution to the following semidefinite program:

$$\begin{array}{lrcl}
\text{maximize} & t & & \\
\text{s.t.} & & & \\
(a) & t & \le & (\mathrm{Det}Z)^{1/n}, \\
(b) & Z & \succeq & 0, \\
(c) & \|Za_i\|_2 & \le & b_i - a_i^T z,\ i = 1, ..., m,
\end{array} \tag{In}$$

with the design variables $Z \in \mathbf{S}^n, z \in \mathbf{R}^n, t \in \mathbf{R}$.

Note that (In) *indeed is a semidefinite program: both* (In.a) *and* (In.c) *can be represented by LMIs, see Examples* 18d *and* **1-17** *in Section 3.2.*

Proof. Indeed, an ellipsoid (3.6.1) is contained in S if and only if

$$a_i^T(Au + c) \le b_i \quad \forall u : u^T u \le 1,$$

or, which is the same, if and only if

$$\|Aa_i\|_2 + a_i^T c = \max_{u: u^T u \le 1} [a_i^T Au + a_i^T c] \le b_i.$$

Thus, (In.$b-c$) just express the fact that the ellipsoid $\{x = Zu + z \mid u^T u \le 1\}$ is contained in S, so that (In) is nothing but the problem of maximizing (a positive power of) the volume of an ellipsoid over ellipsoids contained in S. ∎

We see that if S is a polytope *given by a set of linear inequalities*, then the problem of the best inner ellipsoidal approximation of S is an explicit semidefinite program and as such can be efficiently solved. In contrast to this, if S is a polytope *given as a convex hull of finite set:*

$$S = \mathrm{Conv}\{x_1, ..., x_m\},$$

then the problem of the best inner ellipsoidal approximation of S is "computationally intractable" – in this case, it is difficult just to check whether a given candidate ellipsoid is contained in S.

Outer ellipsoidal approximation of a finite set. Let S be a polyhedral set *given as a convex hull of a finite set of points:*

$$S = \mathrm{Conv}\{x_1, ..., x_m\}.$$

Proposition 3.6.2 *Assume that S is a full-dimensional polytope (i.e., possesses a nonempty interior). Then the smallest volume ellipsoid containing S is*

$$E = \{x \mid (x - c_*)^T D_* (x - c_*) \le 1\},$$

where c_*, D_* are given by an optimal solution (t_*, Z_*, z_*, s_*) to the semidefinite program

$$\begin{array}{lrcl}
\text{maximize} & t & & \\
\text{s.t.} & & & \\
(a) & t & \le & (\mathrm{Det}Z)^{1/n}, \\
(b) & Z & \succeq & 0, \\
(c) & \begin{pmatrix} s & z^T \\ z & Z \end{pmatrix} & \succeq & 0, \\
(d) & x_i^T Z x_i - 2x_i^T z + s & \le & 1,\ i = 1, ..., m,
\end{array} \tag{Out}$$

with the design variables $Z \in \mathbf{S}^n, z \in \mathbf{R}^n, t, s \in \mathbf{R}$ via the relations

$$D_* = Z_*; c_* = Z_*^{-1} z_*.$$

Note that (Out) *indeed is a semidefinite program, cf. Proposition 3.6.1.*

Proof. Indeed, let us pass in the description (3.6.2) from the "parameters" D, c to the parameters $Z = D, z = Dc$, thus coming to the representation

$$E = \{x \mid x^T Z x - 2x^T z + z^T Z^{-1} z \leq 1\}. \tag{!}$$

The ellipsoid of the latter type contains the points $x_1, ..., x_m$ if and only if

$$x_i^T Z x_i - 2x_i^T z + z^T Z^{-1} z \leq 1, \ i = 1, ..., m,$$

or, which is the same, if and only if there exists $s \geq z^T Z^{-1} z$ such that

$$x_i^T Z x_i - 2x_i^T z + s \leq 1, \ i = 1, ..., m.$$

Recalling Lemma on the Schur Complement, we see that the constraints (Out.$b - d$) say exactly that the ellipsoid (!) contains the points $x_1, ..., x_m$. Since the volume of such an ellipsoid is $(\mathrm{Det}Z)^{-1/2}$, (Out) is the problem of maximizing a negative power of the volume of an ellipsoid containing the finite set $\{x_1, ..., x_m\}$, i.e., the problem of finding the smallest volume ellipsoid containing this finite set. It remains to note that an ellipsoid is convex, so that it is exactly the same – to say that it contains a finite set $\{x_1, ..., x_m\}$ and to say that it contains the convex hull of this finite set. ∎

We see that if S is a polytope *given as a convex hull of a finite set*, then the problem of the best outer ellipsoidal approximation of S is an explicit semidefinite program and as such can be efficiently solved. In contrast to this, if S is a polytope *given by a list of inequality constraints*, then the problem of the best outer ellipsoidal approximation of S is "computationally intractable" – in this case, it is difficult just to check whether a given candidate ellipsoid contains S.

3.6.1 Ellipsoidal approximations of unions/intersections of ellipsoids

Speaking informally, Proposition 3.6.1 deals with inner ellipsoidal approximation of the intersection of "degenerate" ellipsoids, namely, half-spaces (a half-space is just a very large Euclidean ball!) Similarly, Proposition 3.6.2 deals with the outer ellipsoidal approximation of the union of degenerate ellipsoids, namely, points (a point is just a ball of zero radius!). We are about to demonstrate that when passing from "degenerate" ellipsoids to the "normal" ones, we still have a possibility to reduce the corresponding approximation problems to explicit semidefinite programs. The key observation here is as follows:

Proposition 3.6.3 *[5] An ellipsoid*

$$E = E(Z, z) \equiv \{x = Zu + z \mid u^T u \leq 1\} \quad [Z \in \mathbf{M}^{n,q}]$$

is contained in the full-dimensional ellipsoid

$$W = W(Y, y) \equiv \{x \mid (x - y)^T Y^T Y (x - y) \leq 1\} \quad [Y \in \mathbf{M}^{n,n}, \mathrm{Det}Y \neq 0]$$

if and only if there exists λ such that

$$\begin{pmatrix} I_n & Y(z-y) & YZ \\ (z-y)^T Y^T & 1-\lambda & \\ Z^T Y^T & & \lambda I_q \end{pmatrix} \succeq 0 \tag{3.6.3}$$

as well as if and only if there exists λ such that

$$\begin{pmatrix} Y^{-1}(Y^{-1})^T & z-y & Z \\ (z-y)^T & 1-\lambda & \\ Z^T & & \lambda I_q \end{pmatrix} \succeq 0 \tag{3.6.4}$$

Proof. We clearly have

$$\begin{array}{c}
E \subset W \\
\Updownarrow \\
u^Tu \le 1 \Rightarrow (Zu+z-y)^TY^TY(Zu+z-y) \le 1 \\
\Updownarrow \\
u^Tu \le t^2 \Rightarrow (Zu+t(z-y))^TY^TY(Zu+t(z-y)) \le t^2 \\
\Updownarrow \mathcal{S}\text{-Lemma} \\
\exists \lambda \ge 0 : [t^2-(Zu+t(z-y))^TY^TY(Zu+t(z-y))] - \lambda[t^2-u^Tu] \ge 0 \quad \forall (u,t) \\
\Updownarrow \\
\exists \lambda \ge 0 : \begin{pmatrix} 1-\lambda-(z-y)^TY^TY(z-y) & -(z-y)^TY^TYZ \\ -Z^TY^TY(z-y) & \lambda I_q - Z^TY^TYZ \end{pmatrix} \succeq 0 \\
\Updownarrow \\
\exists \lambda \ge 0 : \begin{pmatrix} 1-\lambda & \\ & \lambda I_q \end{pmatrix} - \begin{pmatrix} (z-y)^TY^T \\ Z^TY^T \end{pmatrix} \begin{pmatrix} Y(z-y) & YZ \end{pmatrix} \succeq 0
\end{array}$$

Now note that in view of Lemma on the Schur Complement the matrix

$$\begin{pmatrix} 1-\lambda & \\ & \lambda I_q \end{pmatrix} - \begin{pmatrix} (z-y)^TY^T \\ Z^TY^T \end{pmatrix} \begin{pmatrix} Y(z-y) & YZ \end{pmatrix}$$

is positive semidefinite if and only if the matrix in (3.6.3) is so. Thus, $E \subset W$ if and only if there exists a nonnegative λ such that the matrix in (3.6.3), let it be called $P(\lambda)$, is positive semidefinite. Since the latter matrix can be positive semidefinite only when $\lambda \ge 0$, we have proved the first statement of the proposition. To prove the second statement, note that the matrix in (3.6.4), let it be called $Q(\lambda)$, is closely related to $P(\lambda)$:

$$Q(\lambda) = SP(\lambda)S^T, \quad S = \begin{pmatrix} Y^{-1} & & \\ & 1 & \\ & & I_q \end{pmatrix} \succ 0,$$

so that $Q(\lambda)$ is positive semidefinite if and only if $P(\lambda)$ is so. ∎

Here are some consequences of Proposition 3.6.3.

Inner ellipsoidal approximation of the intersection of full-dimensional ellipsoids. Let

$$W_i = \{x \mid (x-c_i)^TB_i^2(x-c_i) \le 1\} \quad [B_i \in \mathbf{S}^n_{++}],$$

$i = 1, ..., m$, be given full-dimensional ellipsoids in $\mathbf{R}^n$; assume that the intersection W of these ellipsoids possesses a nonempty interior. Then the problem of the best inner ellipsoidal approximation of W is the explicit semidefinite program

$$\begin{array}{ll}
\text{maximize} & t \\
\text{s.t.} & \\
& \begin{array}{ccc} t & \le & (\mathrm{Det}Z)^{1/n}, \\ \begin{pmatrix} I_n & B_i(z-c_i) & B_iZ \\ (z-c_i)^TB_i & 1-\lambda_i & \\ ZB_i & & \lambda_iI_n \end{pmatrix} & \succeq & 0,\ i=1,...,m, \\ Z & \succeq & 0 \end{array}
\end{array} \qquad \text{(InEll)}$$

with the design variables $Z \in \mathbf{S}^n, z \in \mathbf{R}^n, \lambda_i, t \in \mathbf{R}$. The largest ellipsoid contained in $W = \bigcap_{i=1}^m W_i$ is given by an optimal solution $Z_*, z_*, t_*, \{\lambda_i^*\}$) of (InEll) via the relation

$$E = \{x = Z_*u + z_* \mid u^Tu \le 1\}.$$

Indeed, by Proposition 3.6.3 the LMIs

$$\left(\begin{array}{ccc} I_n & B_i(z-c_i) & B_iZ \\ (z-c_i)^TB_i & 1-\lambda_i & \\ ZB_i & & \lambda_iI_n \end{array}\right) \succeq 0,\ i=1,...,m$$

express the fact that the ellipsoid $\{x = Zu + z \mid u^Tu \leq 1\}$ with $Z \succeq 0$ is contained in every one of the ellipsoids W_i, i.e., is contained in the intersection W of these ellipsoids. Consequently, (InEll) is exactly the problem of maximizing (a positive power of) the volume of an ellipsoid over the ellipsoids contained in W.

Outer ellipsoidal approximation of the union of ellipsoids. Let

$$W_i = \{x = A_iu + c_i \mid u^Tu \leq 1\} \quad [A_i \in \mathbf{M}^{n,k_i}],$$

$i = 1,...,m$, be given ellipsoids in $\mathbf{R}^n$; assume that the convex hull W of the union of these ellipsoids possesses a nonempty interior. Then the problem of the best outer ellipsoidal approximation of W is the explicit semidefinite program

$$\begin{array}{ll} \text{maximize} & t \\ \text{s.t.} & \\ & \begin{array}{rcl} t & \leq & (\mathrm{Det}Y)^{1/n}, \\ \left(\begin{array}{ccc} I_n & Yc_i-z & YA_i \\ (Yc_i-z)^T & 1-\lambda_i & \\ A_i^TY & & \lambda_iI_{k_i} \end{array}\right) & \succeq & 0,\ i=1,...,m, \\ Y & \succeq & 0 \end{array} \end{array} \qquad \text{(OutEll)}$$

with the design variables $Y \in \mathbf{S}^n, z \in \mathbf{R}^n, \lambda_i, t \in \mathbf{R}$. The smallest ellipsoid containing $W = \mathrm{Conv}(\bigcup_{i=1}^m W_i)$ is given by an optimal solution $(Y_*, z_*, t_*, \{\lambda_i^*\})$ of (OutEll) via the relation

$$E = \{x \mid (x-y_*)Y_*^2(x-y_*) \leq 1\}, \quad y_* = Y_*^{-1}z_*.$$

Indeed, by Proposition 3.6.3 for $Y \succ 0$ the LMIs

$$\left(\begin{array}{ccc} I_n & Yc_i-z & YA_i \\ (Yc_i-z)^T & 1-\lambda_i & \\ A_i^TY & & \lambda_iI_{k_i} \end{array}\right) \succeq 0,\ i=1,...,m$$

express the fact that the ellipsoid $E = \{x \mid (x-Y^{-1}z)^TY^2(x-Y^{-1}y) \leq 1\}$ contains every one of the ellipsoids W_i, i.e., contains the convex hull W of the union of these ellipsoids. The volume of the ellipsoid E is $(\mathrm{Det}Y)^{-1}$; consequently, (OutEll) is exactly the problem of maximizing a *negative* power (i.e., of minimizing a positive power) of the volume of an ellipsoid over the ellipsoids containing W.

3.6.2 Approximating sums of ellipsoids

Let us come back to our motivating example, where we were interested to build ellipsoidal approximation of the set X_T of all states $x(T)$ where a given discrete time invariant linear system

$$\begin{array}{rcl} x(t+1) & = & Ax(t) + Bu(t),\ t = 0,...,T-1 \\ x(0) & = & 0 \end{array}$$

can be driven in time T by a control $u(\cdot)$ satisfying the norm bound

$$\|u(t)\|_2 \leq 1,\ t = 0,...,T-1.$$

How could we build such an approximation recursively? Let X_t be the set of all states where the system can be driven in time $t \le T$, and assume that we have already built inner and outer ellipsoidal approximations $E^t_{\rm in}$ and $E^t_{\rm out}$ of the set X_t:

$$E^t_{\rm in} \subset X_t \subset E^t_{\rm out}.$$

Let also

$$E = \{x = Bu \mid u^T u \le 1\}.$$

Then the set

$$F^{t+1}_{\rm in} = AE^t_{\rm in} + E \equiv \{x = Ay + z \mid y \in E^t_{\rm in}, z \in E\}$$

clearly is contained in X_{t+1}, so that a natural recurrent way to define an inner ellipsoidal approximation of X_{t+1} is to take as $E^{t+1}_{\rm in}$ the largest volume ellipsoid contained in $F^{t+1}_{\rm in}$. Similarly, the set

$$F^{t+1}_{\rm out} = AE^t_{\rm out} + E \equiv \{x = Ay + z \mid y \in E^t_{\rm out}, z \in E\}$$

clearly covers X_{t+1}, and the natural recurrent way to define an outer ellipsoidal approximation of X_{t+1} is to take as $E^{t+1}_{\rm out}$ the smallest volume ellipsoid containing $F^{t+1}_{\rm out}$.

Note that the sets $F^{t+1}_{\rm in}$ and $F^{t+1}_{\rm out}$ are of the same structure: each of them is the *arithmetic sum* $\{x = v + w \mid v \in V, w \in W\}$ of two ellipsoids V and W. Thus, we come to the problem as follows: *Given two ellipsoids W, V, find the best inner and outer ellipsoidal approximations of their arithmetic sum $W + V$.* In fact, it makes sense to consider a little bit more general problem:

> *Given m ellipsoids $W_1, ..., W_m$ in $\mathbf{R}^n$, find the best inner and outer ellipsoidal approximations of the arithmetic sum*
>
> $$W = \{x = w_1 + w_1 + ... + w_m \mid w_i \in W_i,\ i = 1, ..., m\}$$
>
> *of the ellipsoids $W_1, ..., W_m$.*

In fact, we have posed two different problems: the one of inner approximation of W (let this problem be called (I)) and the other one, let it be called (O), of outer approximation. It seems that *in general* both these problems are difficult (at least when m is not once for ever fixed). There exist, however, "computationally tractable" *approximations* of both (I) and (O) we are about to consider.

In considerations to follow we assume, for the sake of simplicity, that the ellipsoids $W_1, ..., W_m$ are full-dimensional (which is not a severe restriction – a "flat" ellipsoid can be easily approximated by a "nearly flat" full-dimensional ellipsoid). Besides this, we may assume without loss of generality that all our ellipsoids W_i are centered at the origin. Indeed, we have $W_i = c_i + V_i$, where c_i is the center of W_i and $V_i = W_i - c_i$ is centered at the origin; consequently,

$$W_1 + ... + W_m = (c_1 + ... + c_m) + (V_1 + ... + V_m),$$

so that the problems (I) and (O) for the ellipsoids $W_1, ..., W_m$ can be straightforwardly reduced to similar problems for the centered at the origin ellipsoids $V_1, ..., V_m$.

Problem (O). Let the ellipsoids $W_1, ..., W_m$ be represented as

$$W_i = \{x \in \mathbf{R}^n \mid x^T B_i x \le 1\} \qquad [B_i \succ 0].$$

Our strategy to approximate (O) is very natural: we intend to build a parametric family of ellipsoids in such a way that, first, every ellipsoid from the family contains the arithmetic sum $W_1 + ... + W_m$ of given ellipsoids, and, second, the problem of finding the smallest volume ellipsoid *within the family* is a "computationally tractable" problem (specifically, is an explicit semidefinite program)[18]. The seemingly

[18] Note that we, in general, do not pretend that our parametric family includes all ellipsoids containing $W_1 + ... + W_m$, so that the ellipsoid we end with should be treated as nothing more than a "computable surrogate" of the smallest volume ellipsoid containing the sum of W_i's.

simplest way to build the desired family was proposed in [5] and is based on the idea of semidefinite relaxation. Let us start with the observation that an ellipsoid

$$W[Z] = \{x \mid x^T Z x \le 1\} \qquad [Z \succ 0]$$

contains $W_1 + ... + W_m$ if and only if the following implication holds:

$$\big\{\{x^i \in \mathbf{R}^n\}_{i=1}^m, [x^i]^T B_i x^i \le 1,\ i = 1, ..., m\big\} \Rightarrow (x^1 + ... + x^m)^T Z (x^1 + ... + x^m) \le 1. \qquad (*)$$

Now let B^i be $(nm) \times (nm)$ block-diagonal matrix with m diagonal blocks of the size $n \times n$ each, such that all diagonal blocks, except the i-th one, are zero, and the i-th block is the $n \times n$ matrix B_i. Let also $M[Z]$ denote $(mn) \times (mn)$ block matrix with m^2 blocks of the size $n \times n$ each, every of these blocks being the matrix Z. This is how B^i and $M[Z]$ look in the case of $m = 2$:

$$B^1 = \begin{bmatrix} B_1 & \\ & \end{bmatrix}, \quad B^2 = \begin{bmatrix} & \\ & B_2 \end{bmatrix}, \quad M[Z] = \begin{bmatrix} Z & Z \\ Z & Z \end{bmatrix}.$$

Validity of implication $(*)$ clearly is equivalent to the following fact:

(*.1) *For every (mn)-dimensional vector x such that*

$$x^T B^i x \equiv \mathrm{Tr}(B^i \underbrace{x x^T}_{X[x]}) \le 1,\ i = 1, ..., m,$$

one has

$$x^T M[Z] x \equiv \mathrm{Tr}(M[Z] X[x]) \le 1.$$

Now we can use the standard trick: the rank one matrix $X[x]$ is positive semidefinite, so that we for sure enforce the validity of the above fact when enforcing the following stronger fact:

(*.2) *For every $(mn) \times (mn)$ symmetric positive semidefinite matrix X such that*

$$\mathrm{Tr}(B^i X) \le 1,\ i = 1, ..., m,$$

one has

$$\mathrm{Tr}(M[Z] X) \le 1.$$

We have arrived at the following result.

(D) *Let a positive definite $n \times n$ matrix Z be such that the optimal value in the semidefinite program*

$$\max_X \big\{\mathrm{Tr}(M[Z]X) \big| \mathrm{Tr}(B^i X) \le 1,\ i = 1, ..., m,\ X \succeq 0\big\} \qquad \text{(SDP)}$$

is ≤ 1. Then the ellipsoid

$$W[Z] = \{x \mid x^T Z x \le 1\}$$

contains the arithmetic sum $W_1 + ... + W_m$ of the ellipsoids $W_i = \{x \mid x^T B_i x \le 1\}$.

We are basically done: the set of those symmetric matrices Z for which the optimal value in (SDP) is ≤ 1 is SD-representable; indeed, the problem is clearly strictly feasible, and Z affects, in a linear fashion, the objective of the problem only. On the other hand, the optimal value in a strictly feasible semidefinite *maximization* program is a SDr function of the objective ("semidefinite version" of Proposition 2.4.4). Consequently, the set of those Z for which the optimal value in (SDP) is ≤ 1 is SDr (as the inverse image, under affine mapping, of the level set of a SDr function). Thus, the "parameter" Z of those ellipsoids $W[Z]$ which satisfy the premise in **(D)** and thus contain $W_1 + ... + W_m$ varies in an SDr set $\mathcal{Z}$. Consequently, the problem of finding the smallest volume ellipsoid in the family $\{W[Z]\}_{Z \in \mathcal{Z}}$ is equivalent to the problem of maximizing a positive power of $\mathrm{Det}(Z)$ over the SDr set $\mathcal{Z}$, i.e., is equivalent to a semidefinite program.

It remains to build the aforementioned semidefinite program. By the Conic Duality Theorem the optimal value in the (clearly strictly feasible) maximization program (SDP) is ≤ 1 if and only if the dual problem

$$\min_{\lambda}\left\{\sum_{i=1}^{m}\lambda_i \Big| \sum_i \lambda_i B^i \succeq M[Z], \lambda_i \geq 0,\ i=1,...,m\right\}.$$

admits a feasible solution with the value of the objective ≤ 1, or, which is clearly the same (why?), admits a feasible solution with the value of the objective equal 1. In other words, whenever $Z \succeq 0$ is such that $M[Z]$ is $\preceq$ a convex combination of the matrices B^i, the set

$$W[Z] = \{x \mid x^T Z x \leq 1\}$$

(which is an ellipsoid when $Z \succ 0$) contains the set $W_1 + ... + W_m$. We have arrived at the following result (see [5], Section 3.7.4):

Proposition 3.6.4 *Given m centered at the origin full-dimensional ellipsoids*

$$W_i = \{x \in \mathbf{R}^n \mid x^T B_i x \leq 1\} \quad [B_i \succ 0],$$

$i = 1, ..., m$, in $\mathbf{R}^n$, let us associate with these ellipsoids the semidefinite program

$$\max_{t,Z,\lambda}\left\{t \Big| \begin{array}{c} t \leq \mathrm{Det}^{1/n}(Z) \\ \sum_{i=1}^{m}\lambda_i B^i \succeq M[Z] \\ \lambda_i \geq 0,\ i=1,...,m \\ Z \succeq 0 \\ \sum_{i=1}^{m}\lambda_i = 1 \end{array}\right\} \qquad (\tilde{\mathrm{O}})$$

where B^i is the $(mn)\times(mn)$ block-diagonal matrix with blocks of the size $n \times n$ and the only nonzero diagonal block (the i-th one) equal to B_i, and $M[Z]$ is the $(mn)\times(mn)$ matrix partitioned into m^2 blocks, every one of them being Z. Every feasible solution $(Z, ...)$ to this program with positive value of the objective produces ellipsoid

$$W[Z] = \{x \mid x^T Z x \leq 1\}$$

which contains $W_1 + ... + W_m$, and the volume of this ellipsoid is at most $t^{-n/2}$. The smallest volume ellipsoid which can be obtained in this way is given by (any) optimal solution of $(\tilde{\mathrm{O}})$.

How "conservative" is $(\tilde{\mathrm{O}})$? The ellipsoid $W[Z^*]$ given by the optimal solution of $(\tilde{\mathrm{O}})$ contains the arithmetic sum W of the ellipsoids W_i, but not necessarily is the smallest volume ellipsoid containing W; all we know is that this ellipsoid is the smallest volume one in certain subfamily of the family of all ellipsoids containing W. "In the nature" there exists the "true" smallest volume ellipsoid $W[Z^{**}] = \{x \mid x^T Z^{**} x \leq 1\}$, $Z^{**} \succ 0$, containing W. It is natural to ask how large could be the ratio

$$\vartheta = \frac{\mathrm{Vol}(W[Z^*])}{\mathrm{Vol}(W[Z^{**}])}.$$

The answer is as follows:

Proposition 3.6.5 *One has $\vartheta \leq \left(\frac{\pi}{2}\right)^{n/2}$.*

Note that the bound stated by Proposition 3.6.5 is not as bad as it looks: the natural way to compare the "sizes" of two n-dimensional bodies E', E'' is to look at the ratio of their *average linear sizes* $\left(\frac{\mathrm{Vol}(E')}{\mathrm{Vol}(E'')}\right)^{1/n}$ (it is natural to assume that shrinking a body by certain factor, say, 2, we reduce the "size" of the body exactly by this factor, and not by 2^n). With this approach, the "level of non-optimality" of $W[Z^*]$ is no more than $\sqrt{\pi/2} = 1.253...$, i.e., is within 25% margin.

Proof of Proposition 3.6.5: Since Z^{**} contains W, the implication (*.1) holds true, i.e., one has

$$\max_{x\in\mathbf{R}^{mn}}\{x^TM[Z^{**}]x \mid x^TB^ix\le 1,\ i=1,...,m\}\le 1.$$

Since the matrices B^i, $i=1,...,m$, commute and $M[Z^{**}]\succeq 0$, we can apply Proposition 3.7.1 (see Section 3.7.5) to conclude that there exist nonnegative μ_i, $i=1,...,m$, such that

$$M[Z^{**}]\preceq\sum_{i=1}^m\mu_iB^i,\quad \sum_i\mu_i\le\frac{\pi}{2}.$$

It follows that setting $\lambda_i=\left(\sum_j\mu_j\right)^{-1}\mu_i$, $Z=\left(\sum_j\mu_j\right)^{-1}Z^{**}$, $t=\mathrm{Det}^{1/n}(Z)$, we get a feasible solution of $(\tilde{\mathrm{O}})$. Recalling the origin of Z^*, we come to

$$\mathrm{Vol}(W[Z^*])\le\mathrm{Vol}(W[Z])=\left(\sum_j\mu_j\right)^{n/2}\mathrm{Vol}(W[Z^{**}])\le(\pi/2)^{n/2}\mathrm{Vol}(W[Z^{**}]),$$

as claimed. ∎

Problem (O), the case of "co-axial" ellipsoids. Consider the *co-axial case* – the one when there exist coordinates (not necessarily orthogonal) such that all m quadratic forms defining the ellipsoids W_i are diagonal in these coordinates, or, which is the same, there exists a nonsingular matrix C such that all the matrices C^TB_iC, $i=1,...,m$, are diagonal. Note that the case of $m=2$ always is co-axial – Linear Algebra says that every two homogeneous quadratic forms, at least one of the forms being positive outside of the origin, become diagonal in a properly chosen coordinates.

We are about to prove that

(E) *In the "co-axial" case,* $(\tilde{\mathrm{O}})$ *yields the smallest in volume ellipsoid containing* $W_1+...+W_m$.

Consider the co-axial case. Since we are interested in volume-related issues, and the ratio of volumes remains unchanged under affine transformations, we can assume w.l.o.g. that the matrices B_i defining the ellipsoids $W_i=\{x\mid x^TB_ix\le 1\}$ are positive definite *and diagonal*; let b^i_ℓ be the ℓ-th diagonal entry of B_i, $\ell=1,...,n$.

By the Fritz John Theorem, "in the nature" there exists a unique smallest volume ellipsoid W_* which contains $W_1+...+W_m$; from uniqueness combined with the fact that the sum of our ellipsoids is symmetric w.r.t. the origin it follows that this optimal ellipsoid W_* is centered at the origin:

$$W_*=\{x\mid x^TZ_*x\le 1\}$$

with certain positive definite matrix Z_*.

Our next observation is that the matrix Z_* is diagonal. Indeed, let E be a diagonal matrix with diagonal entries ±1. Since all B_i's are diagonal, the sum $W_1+...+W_m$ remains invariant under multiplication by E:

$$x\in W_1+...+W_m\Leftrightarrow Ex\in W_1+...+W_m.$$

It follows that the ellipsoid $E(W_*)=\{x\mid x^T(E^TZ_*E)x\le 1\}$ covers $W_1+...+W_m$ along with W_* and of course has the same volume as W_*; from the uniqueness of the optimal ellipsoid it follows that $E(W_*)=W_*$, whence $E^TZ_*E=Z_*$ (why?). Since the concluding relation should be valid for all diagonal matrices E with diagonal entries ±1, Z_* must be diagonal.

Now assume that the set

$$W(z)=\{x\mid x^T\mathrm{Diag}(z)x\le 1\}\tag{3.6.5}$$

given by a nonnegative vector z contains $W_1+...+W_m$. Then the following implication holds true:

$$\forall\{x^i_\ell\}_{\substack{i=1,...,m\\ \ell=1,...,n}}:\quad \sum_{\ell=1}^n b^i_\ell(x^i_\ell)^2\le 1,\ i=1,...,m\ \Rightarrow\ \sum_{\ell=1}^n z_\ell(x^1_\ell+x^2_\ell+...+x^m_\ell)^2\le 1.\tag{3.6.6}$$

Denoting $y_\ell^i = (x_\ell^i)^2$ and taking into account that $z_\ell \geq 0$, we see that the validity of (3.6.6) implies the validity of the implication

$$\forall \{y_\ell^i \geq 0\}_{\substack{i=1,\ldots,m \\ \ell=1,\ldots,n}} : \quad \sum_{\ell=1}^n b_\ell^i y_\ell^i \leq 1,\ i=1,...,m \ \Rightarrow \ \sum_{\ell=1}^n z_\ell \left(\sum_{i=1}^m y_\ell^i + 2 \sum_{1\leq i<j\leq m} \sqrt{y_\ell^i y_\ell^j} \right) \leq 1. \qquad (3.6.7)$$

Now let Y be an $(mn)\times(mn)$ symmetric matrix satisfying the relations

$$Y \succeq 0;\ \mathrm{Tr}(YB^i) \leq 1,\ i=1,...,m. \qquad (3.6.8)$$

Let us partition Y into m^2 square blocks, and let Y_ℓ^{ij} be the ℓ-th diagonal entry of the ij-th block of Y. For all i,j with $1 \leq i < j \leq m$, and all ℓ, $1 \leq \ell \leq n$, the 2×2 matrix $\begin{pmatrix} Y_\ell^{ii} & Y_\ell^{ij} \\ Y_\ell^{ij} & Y_\ell^{jj} \end{pmatrix}$ is a principal submatrix of Y and therefore is positive semidefinite along with Y, whence

$$Y_\ell^{ij} \leq \sqrt{Y_\ell^{ii} Y_\ell^{jj}}. \qquad (3.6.9)$$

In view of (3.6.8), the numbers $y_\ell^i \equiv Y_\ell^{ii}$ satisfy the premise in the implication (3.6.7), so that

$$\begin{array}{rcll} 1 & \geq & \sum\limits_{\ell=1}^n z_\ell \left[\sum\limits_{i=1}^m Y_\ell^{ii} + 2 \sum\limits_{1\leq i<j\leq m} \sqrt{Y_\ell^{ii} Y_\ell^{jj}} \right] & [\text{by (3.6.7)}] \\ & \geq & \sum\limits_{\ell=1}^n z_\ell \left[\sum\limits_{i=1}^m Y_\ell^{ii} + 2 \sum\limits_{1\leq i<j\leq m} Y_\ell^{ij} \right] & [\text{since } z\geq 0 \text{ and by (3.6.9)}] \\ & = & \mathrm{Tr}(YM[\mathrm{Diag}(z)]). & \end{array}$$

Thus, (3.6.8) implies the inequality $\mathrm{Tr}(YM[\mathrm{Diag}(z)]) \leq 1$, i.e., the implication

$$Y \succeq 0,\ \mathrm{Tr}(YB^i) \leq 1,\ i=1,...,m \ \Rightarrow \ \mathrm{Tr}(YM[\mathrm{Diag}(z)]) \leq 1$$

holds true. Since the premise in this implication is strictly feasible, the validity of the implication, by Semidefinite Duality, implies the existence of nonnegative λ_i, $\sum_i \lambda_i \leq 1$, such that

$$M[\mathrm{Diag}(z)] \preceq \sum_i \lambda_i B^i.$$

Combining our observations, we come to the conclusion as follows:

> *In the case of diagonal matrices B_i, if the set (3.6.5), given by a nonnegative vector z, contains $W_1 + ... + W_m$, then the matrix $\mathrm{Diag}(z)$ can be extended to a feasible solution of the problem $(\tilde{\mathrm{O}})$. Consequently, in the case in question the approximation scheme given by $(\tilde{\mathrm{O}})$ yields the minimum volume ellipsoid containing $W_1 + ... + W_m$* (since the latter ellipsoid, as we have seen, is of the form (3.6.5) with $z \geq 0$).

It remains to note that the approximation scheme associated with $(\tilde{\mathrm{O}})$ is affine-invariant, so that the above conclusion remains valid when we replace in its premise "the case of diagonal matrices B_i" with "the co-axial case".

Remark 3.6.1 In fact, **(E)** is an immediate consequence of the following fact (which, essentially, is proved in the above reasoning):

Let $A_1, ..., A_m$, B be symmetric matrices such that the off-diagonal entries of all A_i's are nonpositive, and the off-diagonal entries of B are nonnegative. Assume also that the system of inequalities

$$x^T A_i x \leq a_i,\ i=1,...,m \qquad (S)$$

is strictly feasible. Then the inequality

$$x^T B x \leq b$$

is a consequence of the system (S) *if and only if it is a "linear consequence" of* (S)*, i.e., if and only if there exist nonnegative weights* λ_i *such that*

$$B \preceq \sum_i \lambda_i A_i, \quad \sum_i \lambda_i a_i \le b.$$

In other words, in the case in question the optimization program

$$\max_x \left\{x^T B x \mid x^T A_i x \le a_i,\ i = 1, ..., m\right\}$$

and its standard semidefinite relaxation

$$\max_X \left\{\mathrm{Tr}(BX) \mid X \succeq 0,\ \mathrm{Tr}(A_i X) \le a_i,\ i = 1, ..., m\right\}$$

share the same optimal value.

Problem (I). Let us represent the given centered at the origin ellipsoids W_i as

$$W_i = \{x \mid x = A_i u \mid u^T u \le 1\} \qquad [\mathrm{Det}(A_i) \ne 0].$$

We start from the following observation:

(F) *An ellipsoid* $E[Z] = \{x = Zu \mid u^T u \le 1\}$ *(*$[\mathrm{Det}(Z) \ne 0]$*) is contained in the sum* $W_1 + ... + W_m$ *of the ellipsoids* W_i *if and only if one has*

$$\forall x: \quad \|Z^T x\|_2 \le \sum_{i=1}^m \|A_i^T x\|_2. \tag{3.6.10}$$

Indeed, assume, first, that there exists a vector x_* such that the inequality in (3.6.10) is violated at $x = x_*$, and let us prove that in this case $W[Z]$ is not contained in the set $W = W_1 + ... + W_m$. We have

$$\max_{x \in W_i} x_*^T x = \max\left[x_*^T A_i u \mid u^T u \le 1\right] = \|A_i^T x_*\|_2,\ i = 1, ..., m,$$

and similarly

$$\max_{x \in E[Z]} x_*^T x = \|Z^T x_*\|_2,$$

whence

$$\begin{aligned} \max_{x \in W} x_*^T x &= \max_{x^i \in W_i} x_*^T (x^1 + ... + x^m) = \sum_{i=1}^m \max_{x^i \in W_i} x_*^T x^i \\ &= \sum_{i=1}^m \|A_i^T x_*\|_2 < \|Z^T x_*\|_2 = \max_{x \in E[Z]} x_*^T x, \end{aligned}$$

and we see that $E[Z]$ cannot be contained in W. Vice versa, assume that $E[Z]$ is not contained in W, and let $y \in E[Z] \backslash W$. Since W is a convex compact set and $y \notin W$, there exists a vector x_* such that $x_*^T y > \max_{x \in W} x_*^T x$, whence, due to the previous computation,

$$\|Z^T x_*\|_2 = \max_{x \in E[Z]} x_*^T x \ge x_*^T y > \max_{x \in W} x_*^T x = \sum_{i=1}^m \|A_i^T x_*\|_2,$$

and we have found a point $x = x_*$ at which the inequality in (3.6.10) is violated. Thus, $E[Z]$ is *not* contained in W if and only if (3.6.10) is *not* true, which is exactly what should be proved.

A natural way to generate ellipsoids satisfying (3.6.10) is to note that whenever X_i are $n\times n$ matrices of spectral norms

$$|X_i|\equiv\sqrt{\lambda_{\max}(X_i^TX_i)}=\sqrt{\lambda_{\max}(X_iX_i^T)}=\max_x\{\|X_ix\|_2\mid\|x\|_2\le 1\}$$

not exceeding 1, the matrix

$$Z=Z(X_1,...,X_m)=A_1X_1+A_2X_2+...+A_mX_m$$

satisfies (3.6.10):

$$\|Z^Tx\|_2=\|[X_1^TA_1^T+...+X_m^TA_m^T]x\|_2\le\sum_{i=1}^m\|X_i^TA_i^Tx\|_2\le\sum_{i=1}^m|X_i^T|\|A_i^Tx\|_2\le\sum_{i=1}^m\|A_i^Tx\|_2.$$

Thus, every collection of square matrices X_i with spectral norms not exceeding 1 produces an ellipsoid satisfying (3.6.10) and thus contained in W, and we could use the largest volume ellipsoid of this form (i.e., the one corresponding to the largest $|\mathrm{Det}(A_1X_1+...+A_mX_m)|$) as a surrogate of the largest volume ellipsoid contained in W. Recall that we know how to express a bound on the spectral norm of a matrix via LMI:

$$|X|\le t\Leftrightarrow\left(\begin{array}{cc}tI_n&-X^T\\-X&tI_n\end{array}\right)\succeq 0\quad[X\in\mathbf{M}^{n,n}]$$

(item 16 of Section 3.2). The difficulty, however, is that the matrix $\sum\limits_{i=1}^m A_iX_i$ specifying the ellipsoid $E(X_1,...,X_m)$, although being linear in the "design variables" X_i, is not necessarily symmetric positive semidefinite, and we do not know how to maximize the determinant over general-type square matrices. We may, however, use the following fact from Linear Algebra:

Lemma 3.6.1 *Let $Y=S+C$ be a square matrix represented as the sum of a symmetric matrix S and a skew-symmetric (i.e., $C^T=-C$) matrix C. Assume that S is positive definite. Then*

$$|\mathrm{Det}(Y)|\ge\mathrm{Det}(S).$$

Proof. We have $Y=S+C=S^{1/2}(I+\Sigma)S^{1/2}$, where $\Sigma=S^{-1/2}CS^{-1/2}$ is skew-symmetric along with C. We have $|\mathrm{Det}(Y)|=\mathrm{Det}(S)|\mathrm{Det}(I+\Sigma)|$; it remains to note that all eigenvalues of the skew-symmetric matrix Σ are purely imaginary, so that the eigenvalues of $I+\Sigma$ are ≥ 1 in absolute value, whence $|\mathrm{Det}(I+\Sigma)|\ge 1$. ∎

In view of Lemma, it makes sense to impose on $X_1,...,X_m$, besides the requirement that their spectral norms are ≤ 1, also the requirement that the "symmetric part"

$$S(X_1,...,X_m)=\frac{1}{2}\left[\sum_{i=1}^m A_iX_i+\sum_{i=1}^m X_i^TA_i\right]$$

of the matrix $\sum\limits_i A_iX_i$ is positive semidefinite, and to maximize under these constraints the quantity $\mathrm{Det}(S(X_1,...,X_m))$ – a lower bound on the volume of the ellipsoid $E[Z(X_1,...,X_m)]$. With this approach, we come to the following result:

Proposition 3.6.6 *Let $W_i=\{x=A_iu\mid u^Tu\le 1\}$, $A_i\succ 0$, $i=1,..,m$. Consider the semidefinite program*

$$\begin{array}{lll}\text{maximize} & t & \\ \text{s.t.} & & \\ (a) & t\le\left(\mathrm{Det}\left(\frac{1}{2}\sum\limits_{i=1}^m[X_i^TA_i+A_iX_i]\right)\right)^{1/n} & \\ (b) & \sum\limits_{i=1}^m[X_i^TA_i+A_iX_i]\succeq 0 & \\ (c) & \left(\begin{array}{cc}I_n&-X_i^T\\-X_i&I_n\end{array}\right)\succeq 0,\ i=1,...,m & \end{array}\qquad(\tilde{I})$$

with design variables $X_1, ..., X_m \in \mathbf{M}^{n,n}, t \in \mathbf{R}$. *Every feasible solution* $(\{X_i\}, t)$ *to this problem produces the ellipsoid*

$$E(X_1, ..., X_m) = \{x = (\sum_{i=1}^m A_i X_i)u \mid u^T u \leq 1\}$$

contained in the arithmetic sum $W_1 + ... + W_m$ *of the original ellipsoids, and the volume of this ellipsoid is at least* t^n. *The largest volume ellipsoid which can be obtained in this way is associated with (any) optimal solution to* $(\tilde{\mathrm{I}})$.

In fact, problem (I) is *equivalent* to the problem

$$|\mathrm{Det}(\sum_{i=1}^m A_i X_i)| \to \max \mid |X_i| \leq 1, \ i = 1, ..., m \tag{3.6.11}$$

we have started with, since the latter problem always has an optimal solution $\{X_i^*\}$ with positive semidefinite symmetric matrix $G_* = \sum_{i=1}^m A_i X_i^*$. Indeed, let $\{X_i^+\}$ be an optimal solution of the problem. The matrix $G_+ = \sum_{i=1}^m A_i X_i^+$, as every $n \times n$ square matrix, admits a representation $G_+ = G_* U$, where G_+ is a positive semidefinite symmetric, and U is an orthogonal matrix. Setting $X_i^* = X_i U^T$, we convert $\{X_i^+\}$ into a new feasible solution of (3.6.11); for this solution $\sum_{i=1}^m A_i X_i^* = G_* \succeq 0$, and $\mathrm{Det}(G_+) = \mathrm{Det}(G_*)$, so that the new solution is optimal along with $\{X_i^+\}$.

Problem (I), the co-axial case. We are about to demonstrate that in the co-axial case, when in properly chosen coordinates in $\mathbf{R}^n$ the ellipsoids W_i can be represented as

$$W_i = \{x = A_i u \mid u^T u \leq 1\}$$

with positive definite *diagonal* matrices A_i, the above scheme yields *the best* (the largest volume) ellipsoid among those contained in $W = W_1 + ... + W_m$. Moreover, this ellipsoid can be pointed out explicitly – it is exactly the ellipsoid $E[Z]$ with $Z = Z(I_n, ..., I_n) = A_1 + ... + A_m$!

The announced fact is nearly evident. Assuming that A_i are positive definite and diagonal, consider the parallelotope

$$\widehat{W} = \{x \in \mathbf{R}^n \mid |x_j| \leq \ell_j = \sum_{i=1}^m [A_i]_{jj}, \ j = 1, ..., n\}.$$

This parallelotope clearly contains W (why?), and the largest volume ellipsoid contained in $\widehat{W}$ clearly is the ellipsoid

$$\{x \mid \sum_{j=1}^n \ell_j^{-2} x_j^2 \leq 1\},$$

i.e., is nothing else but the ellipsoid $E[A_1 + ... + A_m]$. As we know from our previous considerations, the latter ellipsoid is contained in W, and since it is the largest volume ellipsoid among those contained in the set $\widehat{W} \supset W$, it is the largest volume ellipsoid contained in W as well.

Example. In the example to follow we are interested to understand what is the domain D_T on the 2D plane which can be reached by a trajectory of the differential equation

$$\frac{d}{dt}\begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix} = \underbrace{\begin{pmatrix} -0.8147 & -0.4163 \\ 0.8167 & -0.1853 \end{pmatrix}}_{A} \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix} + \begin{pmatrix} u_1(t) \\ 0.7071 u_2(t) \end{pmatrix}, \quad \begin{pmatrix} x_1(0) \\ x_2(0) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

in T sec under a piecewise-constant control $u(t) = \begin{pmatrix} u_1(t) \\ u_2(t) \end{pmatrix}$ which switches from one constant value to another one every $\Delta t = 0.01$ sec and is subject to the norm bound

$$\|u(t)\|_2 \le 1 \quad \forall t.$$

The system is stable (the eigenvalues of A are $-0.5 \pm 0.4909i$). In order to build D_T, note that the states of the system at time instants $k\Delta t$, $k = 0, 1, 2, ...$ are the same as the states $x[k] = \begin{pmatrix} x_1(k\Delta t) \\ x_2(k\Delta t) \end{pmatrix}$ of the discrete time system

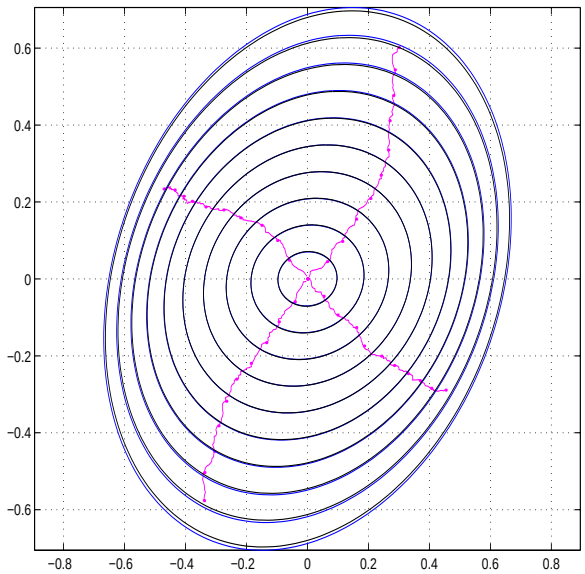
$$x[k+1] = \underbrace{\exp\{A\Delta t\}}_{S} x[k] + \underbrace{\left[\int_0^{\Delta t} \exp\{As\} \begin{pmatrix} 1 & 0 \\ 0 & 0.7071 \end{pmatrix} ds\right]}_{B} u[k], \; x[0] = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \tag{3.6.12}$$

where $u[k]$ is the value of the control on the "continuous time" interval $(k\Delta t, (k+1)\Delta t)$.

We build the inner $\mathcal{I}_k$ and the outer $\mathcal{O}_k$ ellipsoidal approximations of the domains $D^k = D_{k\Delta t}$ in a recurrent manner:

- the ellipses $\mathcal{I}_0$ and $\mathcal{O}_0$ are just the singletons (the origin);
- $\mathcal{I}_{k+1}$ is the best (the largest in the area) ellipsis contained in the set
$$S\mathcal{I}_k + BW, \quad W = \{u \in \mathbf{R}^2 \mid \|u\|_2 \le 1\},$$
which is the sum of two ellipses;
- $\mathcal{O}_{k+1}$ is the best (the smallest in the area) ellipsis containing the set
$$S\mathcal{O}_k + BW,$$
which again is the sum of two ellipses.

Here is the picture we get:



Outer and inner approximations of the "reachability domains"
$D^{10\ell} = D_{0.1\ell \text{ sec}}$, $\ell = 1, 2, ..., 10$, for system (3.6.12)

- Ten pairs of ellipses are the outer and inner approximations of the domains $D^1, ..., D^{10}$ (look how close the ellipses from a pair are close to each other!);
- Four curves are sample trajectories of the system (dots correspond to time instants 0.1ℓ sec in continuous time, i.e., time instants 10ℓ in discrete time, $\ell = 0, 1, ..., 10$).

3.7 Exercises

3.7.1 Around positive semidefiniteness, eigenvalues and $\succeq$-ordering

Criteria for positive semidefiniteness

Recall the criterion of positive definiteness of a symmetric matrix:

> [Sylvester] *A symmetric $m \times m$ matrix $A = [a_{ij}]_{i,j=1}^m$ is positive definite if and only if all angular minors*
>
> $$\mathrm{Det}\left([a_{ij}]_{i,j=1}^k\right), \quad k = 1, ..., m,$$
>
> *are positive.*

Exercise 3.1 *Prove that a symmetric $m \times m$ matrix A is positive semidefinite if and only if all its principal minors (i.e., determinants of square sub-matrices symmetric w.r.t. the diagonal) are nonnegative.*

Hint: look at the angular minors of the matrices $A + \epsilon I_n$ for small positive ϵ.

Demonstrate by an example that nonnegativity of angular minors of a symmetric matrix is not sufficient for the positive semidefiniteness of the matrix.

Exercise 3.2 [Diagonal-dominant matrices] *Let a symmetric matrix $A = [a_{ij}]_{i,j=1}^m$ satisfy the relation*

$$a_{ii} \geq \sum_{j \neq i} |a_{ij}|, \ i = 1, ..., m.$$

Prove that A is positive semidefinite.

Variational characterization of eigenvalues

The basic fact about eigenvalues of a symmetric matrix is the following

> **Variational Characterization of Eigenvalues** [Theorem A.7.3] *Let A be a symmetric $m \times m$ matrix and $\lambda(A) = (\lambda_1(A), ..., \lambda_m(A))$ be the vector of eigenvalues of A taken with their multiplicities and arranged in non-ascending order:*
>
> $$\lambda_1(A) \geq \lambda_2(A) \geq ... \geq \lambda_m(A).$$
>
> *Then for every $i = 1, ..., m$ one has:*
>
> $$\lambda_i(A) = \min_{E \in \mathcal{E}_i} \max_{v \in E, v^T v = 1} v^T A v,$$
>
> *where $\mathcal{E}_i$ is the family of all linear subspaces of $\mathbf{R}^m$ of dimension $m - i + 1$.*

Singular values of rectangular matrices also admit variational description:

> **Variational Characterization of Singular Values** *Let A be an $m \times n$ matrix, $m \leq n$, and let $\sigma(A) = \lambda((AA^T)^{1/2})$ be the vector of singular values of A. Then for every $i = 1, ..., m$ one has:*
>
> $$\sigma_i(A) = \min_{E \in \mathcal{E}_i} \max_{v \in E, v^T v = 1} \|Av\|_2,$$
>
> *where $\mathcal{E}_i$ is the family of all linear subspaces of $\mathbf{R}^n$ of dimension $n - i + 1$.*

Exercise 3.3 *Derive the Variational Characterization of Singular Values from the Variational Characterization of Eigenvalues.*

Exercise 3.4 *Derive from the Variational Characterization of Eigenvalues the following facts:*

(i) [Monotonicity of the vector of eigenvalues] *If $A \succeq B$, then $\lambda(A) \geq \lambda(B)$;*

(ii) *The functions $\lambda_1(X)$, $\lambda_m(X)$ of $X \in \mathbf{S}^m$ are convex and concave, respectively.*

(iii) *If Δ is a convex subset of the real axis, then the set of all matrices $X \in \mathbf{S}^m$ with spectrum from Δ is convex.*

Recall now the definition of *a function of symmetric matrix*. Let A be a symmetric $m \times m$ matrix and

$$p(t) = \sum_{i=0}^{k} p_i t^i$$

be a real polynomial on the axis. By definition,

$$p(A) = \sum_{i=0}^{k} p_i A^i \in \mathbf{S}^m.$$

This definition is compatible with the arithmetic of real polynomials: when you add/multiply polynomials, you add/multiply the "values" of these polynomials at every fixed symmetric matrix:

$$(p+q)(A) = p(A) + q(A); (p \cdot q)(A) = p(A)q(A).$$

A nice feature of this definition is that

> **(A)** *For $A \in \mathbf{S}^m$, the matrix $p(A)$ depends only on the restriction of p on the spectrum (set of eigenvalues) of A: if p and q are two polynomials such that $p(\lambda_i(A)) = q(\lambda_i(A))$ for $i = 1, ..., m$, then $p(A) = q(A)$.*
>
> Indeed, we can represent a symmetric matrix A as $A = U^T \Lambda U$, where U is orthogonal and Λ is diagonal with the eigenvalues of A on its diagonal. Since $UU^T = I$, we have $A^i = U^T \Lambda^i U$; consequently,
>
> $$p(A) = U^T p(\Lambda) U,$$
>
> and since the matrix $p(\Lambda)$ depends on the restriction of p on the spectrum of A only, the result follows.
>
> As a byproduct of our reasoning, we get an "explicit" representation of $p(A)$ in terms of the *spectral decomposition* $A = U^T \Lambda U$ (U is orthogonal, Λ is diagonal with the diagonal $\lambda(A)$):
>
> **(B)** *The matrix $p(A)$ is just $U^T \mathrm{Diag}(p(\lambda_1(A)), ..., p(\lambda_n(A)))U$.*

(A) allows to define arbitrary functions of matrices, not necessarily polynomials:

> *Let A be symmetric matrix and f be a real-valued function defined at least at the spectrum of A. By definition, the matrix $f(A)$ is defined as $p(A)$, where p is a polynomial coinciding with f on the spectrum of A.* (The definition makes sense, since by **(A)** $p(A)$ depends only on the restriction of p on the spectrum of A, i.e., every "polynomial continuation" $p(\cdot)$ of f from the spectrum of A to the entire axis results in the same $p(A)$).

The "calculus of functions of a symmetric matrix" is fully compatible with the usual arithmetic of functions, e.g:

$$(f+g)(A) = f(A) + g(A); (\mu f)(A) = \mu f(A); (f \cdot g)(A) = f(A)g(A); (f \circ g)(A) = f(g(A)),$$

provided that the functions in question are well-defined on the spectrum of the corresponding matrix. And of course the spectral decomposition of $f(A)$ is just $f(A) = U^T \mathrm{Diag}(f(\lambda_1(A)), ..., f(\lambda_m(A)))U$, where $A = U^T \mathrm{Diag}(\lambda_1(A), ..., \lambda_m(A))U$ is the spectral decomposition of A.

Note that "Calculus of functions of symmetric matrices" becomes very unusual when we are trying to operate with functions of several (non-commuting) matrices. E.g., it is generally *not* true that $\exp\{A + B\} = \exp\{A\}\exp\{B\}$ (the right hand side matrix may be even non-symmetric!). It is also generally not true that if f is monotone and $A \succeq B$, then $f(A) \succeq f(B)$, etc.

Exercise 3.5 *Demonstrate by an example that the relation $0 \preceq A \preceq B$ does not necessarily imply that $A^2 \preceq B^2$.*

By the way, the relation $0 \preceq A \preceq B$ does imply that $0 \preceq A^{1/2} \preceq B^{1/2}$.

Sometimes, however, we can get "weak" matrix versions of usual arithmetic relations. E.g.,

Exercise 3.6 *Let f be a nondecreasing function on the real line, and let $A \succeq B$. Prove that $\lambda(f(A)) \geq \lambda(f(B))$.*

The strongest (and surprising) "weak" matrix version of a usual ("scalar") inequality is as follows.

Let $f(t)$ be a *closed* convex function on the real line; by definition, it means that f is a function on the axis taking real values and the value $+\infty$ such that

– the set $\mathrm{Dom}\, f$ of the values of argument where f is finite is convex and nonempty;

– if a sequence $\{t_i \in \mathrm{Dom}\, f\}$ converges to a point t and the sequence $f(t_i)$ has a limit, then $t \in \mathrm{Dom}\, f$ and $f(t) \leq \lim\limits_{i\to\infty} f(t_i)$ (this property is called "lower semicontinuity").

E.g., the function $f(x) = \begin{cases} 0, & 0 \leq t \leq 1 \\ +\infty, & \text{otherwise} \end{cases}$ is closed. In contrast to this, the functions

$$g(x) = \begin{cases} 0, & 0 < t \leq 1 \\ 1, & t = 0 \\ +\infty, & \text{for all remaining } t \end{cases}$$

and

$$h(x) = \begin{cases} 0, & 0 < t < 1 \\ +\infty, & \text{otherwise} \end{cases}$$

are not closed, although they are convex: a closed function cannot "jump up" at an endpoint of its domain, as it is the case for g, and it cannot take value $+\infty$ at a point, if it takes values $\leq a < \infty$ in a neighbourhood of the point, as it is the case for h.

For a convex function f, its *Legendre transformation* f_* (also called the *conjugate*, or the *Fenchel dual* of f) is defined as

$$f_*(s) = \sup_t \left[ts - f(t)\right].$$

It turns out that the Legendre transformation of a closed convex function also is closed and convex, and that twice taken Legendre transformation of a closed convex function is this function.

The Legendre transformation (which, by the way, can be defined for convex functions on $\mathbf{R}^n$ as well) underlies many standard inequalities. Indeed, by definition of f_* we have

$$f_*(s) + f(t) \geq st \quad \forall s, t; \tag{L}$$

For specific choices of f, we can derive from the general inequality (L) many useful inequalities. E.g.,

- If $f(t) = \frac{1}{2}t^2$, then $f_*(s) = \frac{1}{2}s^2$, and (L) becomes the standard inequality

$$st \leq \frac{1}{2}t^2 + \frac{1}{2}s^2 \quad \forall s, t \in \mathbf{R};$$

- If $1 < p < \infty$ and $f(t) = \begin{cases} \frac{t^p}{p}, & t \geq 0 \\ +\infty, & t < 0 \end{cases}$, then $f_*(s) = \begin{cases} \frac{s^q}{q}, & s \geq 0 \\ +\infty, & s < 0 \end{cases}$, with q given by $\frac{1}{p} + \frac{1}{q} = 1$, and (L) becomes the Young inequality

$$\forall (s, t \geq 0): \quad ts \leq \frac{t^p}{p} + \frac{s^q}{q},\ 1 < p, q < \infty, \frac{1}{p} + \frac{1}{q} = 1.$$

Now, what happens with (L) if s, t are symmetric matrices? Of course, both sides of (L) still make sense and are matrices, but we have no hope to say something reasonable about the relation between these matrices (e.g., the right hand side in (L) is not necessarily symmetric). However,

Exercise 3.7 *Let f_* be a closed convex function with the domain* $\mathrm{Dom}\, f_* \subset \mathbf{R}_+$*, and let f be the Legendre transformation of f_*. Then for every pair of symmetric matrices X, Y of the same size with the spectrum of X belonging to* $\mathrm{Dom}\, f$ *and the spectrum of Y belonging to* $\mathrm{Dom}\, f_*$ *one has*

$$\lambda(f(X)) \geq \lambda\left(Y^{1/2}XY^{1/2} - f_*(Y)\right)$$ [19)]

Birkhoff's Theorem

Surprisingly enough, one of the most useful facts about eigenvalues of symmetric matrices is the following, essentially combinatorial, statement (it does not mention the word "eigenvalue" at all).

Birkhoff's Theorem. *Consider the set $\mathcal{S}_m$ of double-stochastic $m \times m$ matrices, i.e., square matrices $[p_{ij}]_{i,j=1}^m$ satisfying the relations*

$$\begin{array}{rcl} p_{ij} & \geq & 0,\ i,j = 1,...,m; \\ \sum\limits_{i=1}^m p_{ij} & = & 1,\ j = 1,...,m; \\ \sum\limits_{j=1}^m p_{ij} & = & 1,\ i = 1,...,m. \end{array}$$

A matrix P belongs to $\mathcal{S}_m$ if and only if it can be represented as a convex combination of $m \times m$ permutation matrices:

$$P \in \mathcal{S}_m \Leftrightarrow \exists(\lambda_i \geq 0, \sum_i \lambda_i = 1) : P = \sum_i \lambda_i \Pi^i,$$

where all Π^i are permutation matrices (i.e., with exactly one nonzero element, equal to 1, in every row and every column).

An immediate corollary of the Birkhoff Theorem is the following fact:

(C) *Let $f : \mathbf{R}^m \to \mathbf{R} \cup \{+\infty\}$ be a convex symmetric function (symmetry means that the value of the function remains unchanged when we permute the coordinates in an argument), let $x \in \mathrm{Dom}\, f$ and $P \in \mathcal{S}_m$. Then*

$$f(Px) \leq f(x).$$

The proof is immediate: by Birkhoff's Theorem, Px is a convex combination of a number of permutations x^i of x. Since f is convex, we have

$$f(Px) \leq \max_i f(x^i) = f(x),$$

the concluding equality resulting from the symmetry of f.

The role of **(C)** in numerous questions related to eigenvalues is based upon the following simple

Observation. *Let A be a symmetric $m \times m$ matrix. Then the diagonal* $\mathrm{Dg}(A)$ *of the matrix A is the image of the vector $\lambda(A)$ of the eigenvalues of A under multiplication by a double stochastic matrix:*

$$\mathrm{Dg}(A) = P\lambda(A) \text{ for some } P \in \mathcal{S}_m$$

Indeed, consider the spectral decomposition of A:

$$A = U^T \mathrm{Diag}(\lambda_1(A), ..., \lambda_m(A))U$$

[19)] In the scalar case, our inequality reads $f(x) \geq y^{1/2}xy^{1/2} - f_*(y)$, which is an equivalent form of (L) when $\mathrm{Dom}\, f_* \subset \mathbf{R}_+$.

with orthogonal $U = [u_{ij}]$. Then

$$A_{ii} = \sum_{j=1}^{m} u_{ji}^2 \lambda_j(A) \equiv (P\lambda(A))_i,$$

where the matrix $P = [u_{ji}^2]_{i,j=1}^m$ is double stochastic.

Combining the Observation and **(C)**, we conclude that if f is a convex symmetric function on $\mathbf{R}^m$, then for every $m \times m$ symmetric matrix A one has

$$f(\mathrm{Dg}(A)) \leq f(\lambda(A)).$$

Moreover, let $\mathcal{O}_m$ be the set of all orthogonal $m \times m$ matrices. For every $V \in \mathcal{O}_m$, the matrix $V^T AV$ has the same eigenvalues as A, so that for a convex symmetric f one has

$$f(\mathrm{Dg}(V^T AV)) \leq f(\lambda(V^T AV)) = f(\lambda(A)),$$

whence

$$f(\lambda(A)) \geq \max_{V \in \mathcal{O}_m} f(\mathrm{Dg}(V^T AV)).$$

In fact the inequality here is equality, since for properly chosen $V \in \mathcal{O}_m$ we have $\mathrm{Dg}(V^T AV) = \lambda(A)$. We have arrived at the following result:

(D) *Let f be a symmetric convex function on $\mathbf{R}^m$. Then for every symmetric $m \times m$ matrix A one has*

$$f(\lambda(A)) = \max_{V \in \mathcal{O}_m} f(\mathrm{Dg}(V^T AV)),$$

$\mathcal{O}_m$ being the set of all $m \times m$ orthogonal matrices.

In particular, the function

$$F(A) = f(\lambda(A))$$

is convex in $A \in \mathbf{S}^m$ (as the maximum of a family of convex in A functions $F_V(A) = f(\mathrm{Dg}(V^T AV))$, $V \in \mathcal{O}_m$.)

Exercise 3.8 *Let $g(t) : \mathbf{R} \to \mathbf{R} \cup \{+\infty\}$ be a convex function, and let $\mathcal{F}_n$ be the set of all matrices $X \in \mathbf{S}^n$ with the spectrum belonging to* $\mathrm{Dom}\, g$. *Prove that the function* $\mathrm{Tr}(g(X))$ *is convex on $\mathcal{F}_n$.*

Hint: Apply **(D)** to the function $f(x_1, ..., x_n) = g(x_1) + ... + g(x_n)$.

Exercise 3.9 *Let $A = [a_{ij}]$ be a symmetric $m \times m$ matrix. Prove that*

(i) *Whenever $p \geq 1$, one has $\sum_{i=1}^m |a_{ii}|^p \leq \sum_{i=1}^m |\lambda_i(A)|^p$;*

(ii) *Whenever A is positive semidefinite, $\prod_{i=1}^m a_{ii} \geq \mathrm{Det}(A)$;*

(iii) *For $x \in \mathbf{R}^m$, let the function $S_k(x)$ be the sum of k largest entries of x (i.e., the sum of the first k entries in the vector obtained from x by writing down the coordinates of x in the non-ascending order). Prove that $S_k(x)$ is a convex symmetric function of x and derive from this observation that*

$$S_k(\mathrm{Dg}(A)) \leq S_k(\lambda(A)).$$

Hint: note that $S_k(x) = \max_{1 \leq i_1 < i_2 < ... < i_k \leq m} \sum_{l=1}^k x_{i_l}$.

(iv) [Trace inequality] *Whenever $A, B \in \mathbf{S}^m$, one has*

$$\lambda^T(A)\lambda(B) \geq \mathrm{Tr}(AB).$$

Exercise 3.10 *Prove that if $A \in \mathbf{S}^m$ and $p, q \in [1, \infty]$ are such that $\frac{1}{p} + \frac{1}{q} = 1$, then*

$$\max_{B \in \mathbf{S}^m : \|\lambda(B)\|_q = 1} \mathrm{Tr}(AB) = \|\lambda(A)\|_p.$$

In particular, $\|\lambda(\cdot)\|_p$ is a norm on $\mathbf{S}^m$, and the conjugate of this norm is $\|\lambda(\cdot)\|_q$, $\frac{1}{p} + \frac{1}{q} = 1$.

Exercise 3.11 *Let* $X = \begin{pmatrix} X_{11} & X_{12} & ... & X_{1m} \\ X_{12}^T & X_{22} & ... & X_{2m} \\ \vdots & \vdots & \ddots & \cdots \\ X_{1m}^T & X_{2m}^T & ... & X_{mm} \end{pmatrix}$ *be an $n \times n$ symmetric matrix which is partitioned into m^2 blocks X_{ij} in a symmetric, w.r.t. the diagonal, fashion (so that the blocks X_{jj} are square), and let*

$$\widehat{X} = \begin{pmatrix} X_{11} & & & \\ & X_{22} & & \\ & & \ddots & \\ & & & X_{mm} \end{pmatrix}.$$

1) *Let $F : \mathbf{S}^n \to \mathbf{R} \cup \{+\infty\}$ be a convex "rotation-invariant" function: for all $Y \in \mathbf{S}^n$ and all orthogonal matrices U one has $F(U^T Y U) = F(Y)$. Prove that*

$$F(\widehat{X}) \le F(X).$$

Hint: Represent the matrix $\widehat{X}$ as a convex combination of the rotations $U^T X U$, $U^T U = I$, of X.

2) *Let $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ be a convex symmetric w.r.t. permutations of the entries in the argument function, and let $F(Y) = f(\lambda(Y))$, $Y \in \mathbf{S}^n$. Prove that*

$$F(\widehat{X}) \le F(X).$$

3) *Let $g : \mathbf{R} \to \mathbf{R} \cup \{+\infty\}$ be convex function on the real line which is finite on the set of eigenvalues of X, and let $\mathcal{F}_n \subset \mathbf{S}^n$ be the set of all $n \times n$ symmetric matrices with all eigenvalues belonging to the domain of g. Assume that the mapping*

$$Y \mapsto g(Y) : \mathcal{F}_n \to \mathbf{S}^n$$

is $\succeq$-convex:

$$g(\lambda' Y' + \lambda'' Y'') \preceq \lambda' g(Y') + \lambda'' g(Y'') \quad \forall (Y', Y'' \in \mathcal{F}_n, \lambda', \lambda'' \ge 0, \lambda' + \lambda'' = 1).$$

Prove that

$$(g(X))_{ii} \succeq g(X_{ii}),\ i = 1, ..., m,$$

where the partition of $g(X)$ into the blocks $(g(X))_{ij}$ is identical to the partition of X into the blocks X_{ij}.

Exercise 3.11 gives rise to a number of interesting inequalities. Let X, $\widehat{X}$ be the same as in the Exercise, and let $[Y]$ denote the northwest block, of the same size as X_{11}, of an $n \times n$ matrix Y. Then

1. $\left(\sum_{i=1}^m \|\lambda(X_{ii})\|_p^p \right)^{1/p} \le \|\lambda(X)\|_p$, $1 \le p < \infty$
   [Exercise 3.11.2), $f(x) = \|x\|_p$];
2. If $X \succ 0$, then $\mathrm{Det}(X) \le \prod_{i=1}^m \mathrm{Det}(X_{ii})$
   [Exercise 3.11.2), $f(x) = -(x_1 ... x_n)^{1/n}$ for $x \ge 0$];

3. $[X^2] \succeq X_{11}^2$
   [This inequality is nearly evident; it follows also from Exercise 3.11.3) with $g(t) = t^2$ (the $\succeq$-convexity of $g(Y)$ is stated in Exercise 3.21.1))];
4. If $X \succ 0$, then $X_{11}^{-1} \preceq [X^{-1}]$
   [Exercise 3.11.3) with $g(t) = t^{-1}$ for $t > 0$; the $\succeq$-convexity of $g(Y)$ on $\mathbf{S}^n_{++}$ is stated by Exercise 3.21.2)];
5. For every $X \succeq 0$, $[X^{1/2}] \preceq X_{11}^{1/2}$
   [Exercise 3.11.3) with $g(t) = -\sqrt{t}$; the $\succeq$-convexity of $g(Y)$ is stated by Exercise 3.21.4)].
   Extension: If $X \succeq 0$, then for every $\alpha \in (0,1)$ one has $[X^\alpha] \preceq X_{11}^\alpha$
   [Exercise 3.11.3) with $g(t) = -t^\alpha$; the function $-Y^\alpha$ of $Y \succeq 0$ is known to be $\succeq$-convex];
6. If $X \succ 0$, then $[\ln(X)] \preceq \ln(X_{11})$
   [Exercise 3.11.3) with $g(t) = -\ln t$, $t > 0$; the $\succeq$-convexity of $g(Y)$ is stated by Exercise 3.21.5)].

Exercise 3.12 1) *Let $A = [a_{ij}]_{i,j} \succeq 0$, let $\alpha \geq 0$, and let $B \equiv [b_{ij}]_{i,j} = A^\alpha$. Prove that*

$$b_{ii} \begin{cases} \leq a_{ii}^\alpha, & \alpha \leq 1 \\ \geq a_{ii}^\alpha, & \alpha \geq 1 \end{cases}$$

2) *Let $A = [a_{ij}]_{i,j} \succ 0$, and let $B \equiv [b_{ij}]_{i,j} = A^{-1}$. Prove that $b_{ii} \geq a_{ii}^{-1}$.*
3) *Let $[A]$ denote the northwest 2×2 block of a square matrix. Which of the implications*

$$\begin{array}{ll} (a) & A \succeq 0 \Rightarrow [A^4] \succeq [A]^4 \\ (b) & A \succeq 0 \Rightarrow [A^4]^{1/4} \succeq [A] \end{array}$$

are true?

Semidefinite representations of functions of eigenvalues

The goal of the subsequent series of exercises is to prove Proposition 3.2.1.

We start with a description (important by its own right) of the convex hull of permutations of a given vector. Let $x \in \mathbf{R}^m$, and let $X[x]$ be the set of all convex combinations of $m!$ vectors obtained from x by all permutations of the coordinates.

> **Claim:** ["Majorization principle"] *$X[x]$ is exactly the solution set of the following system of inequalities in variables $y \in \mathbf{R}^m$:*
>
> $$\begin{array}{c} S_j(y) \leq S_j(x), \ j = 1, ..., m-1 \\ y_1 + ... + y_m = x_1 + ... + x_m \end{array} \qquad (+)$$
>
> (recall that $S_j(y)$ is the sum of the largest j entries of a vector y).

Exercise 3.13 [Easy part of the claim] *Let Y be the solution set of $(+)$. Prove that $Y \supset X[x]$.*

Hint: Use **(C)** and the convexity of the functions $S_j(\cdot)$.

Exercise 3.14 [Difficult part of the claim] *Let Y be the solution set of $(+)$. Prove that $Y \subset X[x]$.*

Sketch of the proof: Let $y \in Y$. We should prove that $y \in X[x]$. By symmetry, we may assume that the vectors x and y are ordered: $x_1 \geq x_2 \geq ... \geq x_m$, $y_1 \geq y_2 \geq ... \geq y_m$. Assume that $y \notin X[x]$, and let us lead this assumption to a contradiction.

1) Since $X[x]$ clearly is a convex compact set and $y \notin X[x]$, there exists a linear functional $c(z) = \sum_{i=1}^m c_i z_i$ which separates y and $X[x]$:

$$c(y) > \max_{z \in X[x]} c(z).$$

Prove that such a functional can be chosen "to be ordered": $c_1 \geq c_2 \geq ... \geq c_m$.

2) Verify that

$$c(y) \equiv \sum_{i=1}^{m} c_i y_i = \sum_{i=1}^{m-1} (c_i - c_{i+1}) \sum_{j=1}^{i} y_j + c_m \sum_{j=1}^{m} y_j$$

(Abel's formula – a discrete version of integration by parts). Use this observation along with "orderedness" of $c(\cdot)$ and the inclusion $y \in Y$ to conclude that $c(y) \leq c(x)$, thus coming to the desired contradiction.

Exercise 3.15 *Use the Majorization principle to prove Proposition 3.2.1.*

The next pair of exercises is aimed at proving Proposition 3.2.2.

Exercise 3.16 *Let $x \in \mathbf{R}^m$, and let $X_+[x]$ be the set of all vectors x' dominated by a vector form $X[x]$:*

$$X_+[x] = \{y \mid \exists z \in X[x] : y \leq z\}.$$

1) Prove that $X_+[x]$ is a closed convex set.

2) Prove the following characterization of $X_+[x]$:

$X_+[x]$ is exactly the set of solutions of the system of inequalities $S_j(y) \leq S_j(x)$, $j = 1, ..., m$, in variables y.

Hint: Modify appropriately the constriction outlined in Exercise 3.14.

Exercise 3.17 *Derive Proposition 3.2.2 from the result of Exercise 3.16.2).*

Cauchy's inequality for matrices

The standard Cauchy's inequality says that

$$|\sum_i x_i y_i| \leq \sqrt{\sum_i x_i^2} \sqrt{\sum_i y_i^2} \tag{3.7.1}$$

for reals x_i, y_i, $i = 1, ..., n$; this inequality is exact in the sense that for every collection $x_1, ..., x_n$ there exists a collection $y_1, ..., y_n$ with $\sum_i y_i^2 = 1$ which makes (3.7.1) an equality.

Exercise 3.18 (i) *Prove that whenever $X_i, Y_i \in \mathbf{M}^{p,q}$, one has*

$$\sigma(\sum_i X_i^T Y_i) \leq \lambda \left(\left[\sum_i X_i^T X_i \right]^{1/2} \right) \| \lambda \left(\sum_i Y_i^T Y_i \right) \|_\infty^{1/2} \tag{$*$}$$

where $\sigma(A) = \lambda([AA^T]^{1/2})$ is the vector of singular values of a matrix A arranged in the non-ascending order.

Prove that for every collection $X_1, ..., X_n \in \mathbf{M}^{p,q}$ there exists a collection $Y_1, ..., Y_n \in \mathbf{M}^{p,q}$ with $\sum_i Y_i^T Y_i = I_q$ which makes $(*)$ an equality.

(ii) *Prove the following "matrix version" of the Cauchy inequality: whenever $X_i, Y_i \in \mathbf{M}^{p,q}$, one has*

$$|\sum_i \mathrm{Tr}(X_i^T Y_i)| \leq \mathrm{Tr} \left(\left[\sum_i X_i^T X_i \right]^{1/2} \right) \| \lambda(\sum_i Y_i^T Y_i) \|_\infty^{1/2}, \tag{$**$}$$

and for every collection $X_1, ..., X_n \in \mathbf{M}^{p,q}$ there exists a collection $Y_1, ..., Y_n \in \mathbf{M}^{p,q}$ with $\sum_i Y_i^T Y_i = I_q$ which makes $(**)$ an equality.

Here is another exercise of the same flavour:

Exercise 3.19 *For nonnegative reals $a_1, ..., a_m$ and a real $\alpha > 1$ one has*

$$\left(\sum_{i=1}^m a_i^\alpha\right)^{1/\alpha} \le \sum_{i=1}^m a_i.$$

Both sides of this inequality make sense when the nonnegative reals a_i are replaced with positive semidefinite $n \times n$ matrices A_i. What happens with the inequality in this case?

Consider the following four statements (where $\alpha > 1$ is a real and $m, n > 1$):

1)
$$\forall (A_i \in \mathbf{S}^n_+): \quad \left(\sum_{i=1}^m A_i^\alpha\right)^{1/\alpha} \preceq \sum_{i=1}^m A_i.$$

2)
$$\forall (A_i \in \mathbf{S}^n_+): \quad \lambda_{\max}\left(\left(\sum_{i=1}^m A_i^\alpha\right)^{1/\alpha}\right) \le \lambda_{\max}\left(\sum_{i=1}^m A_i\right).$$

3)
$$\forall (A_i \in \mathbf{S}^n_+): \quad \mathrm{Tr}\left(\left(\sum_{i=1}^m A_i^\alpha\right)^{1/\alpha}\right) \le \mathrm{Tr}\left(\sum_{i=1}^m A_i\right).$$

4)
$$\forall (A_i \in \mathbf{S}^n_+): \quad \mathrm{Det}\left(\left(\sum_{i=1}^m A_i^\alpha\right)^{1/\alpha}\right) \le \mathrm{Det}\left(\sum_{i=1}^m A_i\right).$$

Among these 4 statements, exactly 2 are true. Identify and prove the true statements.

$\succeq$-convexity of some matrix-valued functions

Consider a function $F(x)$ defined on a convex set $X \subset \mathbf{R}^n$ and taking values in $\mathbf{S}^m$. We say that such a function is $\succeq$-convex, if

$$F(\alpha x + (1-\alpha)y) \preceq \alpha F(x) + (1-\alpha)F(y)$$

for all $x, y \in X$ and all $\alpha \in [0,1]$. F is called $\succeq$-*concave*, if $-F$ is $\succeq$-convex.

A function $F : \mathrm{Dom}\, F \to \mathbf{S}^m$ defined on a set $\mathrm{Dom}\, F \subset \mathbf{S}^k$ is called $\succeq$-*monotone*, if

$$x, y \in \mathrm{Dom}\, F, x \succeq y \Rightarrow F(x) \succeq F(y);$$

F is called $\succeq$-*antimonotone*, if $-F$ is $\succeq$-monotone.

Exercise 3.20 1) *Prove that a function $F : X \to \mathbf{S}^m$, $X \subset \mathbf{R}^n$, is $\succeq$-convex if and only if its "epigraph"*

$$\{(x, Y) \in \mathbf{R}^n \to \mathbf{S}^m \mid x \in X, F(x) \preceq Y\}$$

is a convex set.

2) *Prove that a function $F : X \to \mathbf{S}^m$ with convex domain $X \subset \mathbf{R}^n$ is $\succeq$-convex if and only if for every $A \in \mathbf{S}^m_+$ the function $\mathrm{Tr}(AF(x))$ is convex on X.*

3) *Let $X \subset \mathbf{R}^n$ be a convex set with a nonempty interior and $F : X \to \mathbf{S}^m$ be a function continuous on X which is twice differentiable in $\mathrm{int}\, X$. Prove that F is $\succeq$-convex if and only if the second directional derivative of F*

$$D^2F(x)[h,h] \equiv \left.\frac{d^2}{dt^2}\right|_{t=0} F(x+th)$$

is $\succeq 0$ for every $x \in \operatorname{int} X$ and every direction $h \in \mathbf{R}^n$.

4) *Let $F : \operatorname{dom} F \to \mathbf{S}^m$ be defined and continuously differentiable on an open convex subset of $\mathbf{S}^k$. Prove that the necessary and sufficient condition for F to be $\succeq$-monotone is the validity of the implication*

$$h \in \mathbf{S}^k_+, x \in \operatorname{Dom} F \Rightarrow DF(x)[h] \succeq 0.$$

5) *Let F be $\succeq$-convex and $\mathcal{S} \subset \mathbf{S}^m$ be a convex set which is $\succeq$-antimonotone, i.e. whenever $Y' \preceq Y$ and $Y \in \mathcal{S}$, one has $Y' \in \mathcal{S}$. Prove that the set $F^{-1}(\mathcal{S}) = \{x \in X \mid F(x) \in \mathcal{S}\}$ is convex.*

6) *Let $G : \operatorname{Dom} G \to \mathbf{S}^k$ and $F : \operatorname{Dom} F \to \mathbf{S}^m$, let $G(\operatorname{Dom} G) \subset \operatorname{Dom} F$, and let $H(x) = F(G(x)) : \operatorname{Dom} G \to \mathbf{S}^m$.*

a) *Prove that if G and F are $\succeq$-convex and F is $\succeq$-monotone, then H is $\succeq$-convex.*

b) *Prove that if G and F are $\succeq$-concave and F is $\succeq$-monotone, then H is $\succeq$-concave.*

7) *Let $F_i : G \to \mathbf{S}^m$, and assume that for every $x \in G$ exists*

$$F(x) = \lim_{i \to \infty} F_i(x).$$

Prove that if all functions from the sequence $\{F_i\}$ are (a) $\succeq$-convex, or (b) $\succeq$-concave, or (c) $\succeq$-monotone, or (d) $\succeq$-antimonotone, then so is F.

The goal of the next exercise is to establish the $\succeq$-convexity of several matrix-valued functions.

Exercise 3.21 *Prove that the following functions are $\succeq$-convex:*

1) $F(x) = xx^T : \mathbf{M}^{p,q} \to \mathbf{S}^p$;

2) $F(x) = x^{-1} : \operatorname{int} \mathbf{S}^m_+ \to \operatorname{int} \mathbf{S}^m_+$;

3) $F(u, v) = u^T v^{-1} u : \mathbf{M}^{p,q} \times \operatorname{int} \mathbf{S}^p_+ \to \mathbf{S}^q$;

Prove that the following functions are $\succeq$-concave and monotone:

4) $F(x) = x^{1/2} : \mathbf{S}^m_+ \to \mathbf{S}^m$;

5) $F(x) = \ln x : \operatorname{int} \mathbf{S}^m_+ \to \mathbf{S}^m$;

6) $F(x) = \left(Ax^{-1}A^T\right)^{-1} : \operatorname{int} \mathbf{S}^n_+ \to \mathbf{S}^m$, *provided that A is an $m \times n$ matrix of rank m.*

3.7.2 SD representations of epigraphs of convex polynomials

Mathematically speaking, the central question concerning the "expressive abilities" of Semidefinite Programming is how wide is the family of convex sets which are SDr. By definition, an SDr set is the projection of the inverse image of $\mathbf{S}^m_+$ under affine mapping. In other words, every SDr set is a projection of a convex set given by a number of polynomial inequalities (indeed, the cone $\mathbf{S}^m_+$ is a convex set given by polynomial inequalities saying that all principal minors of matrix are nonnegative). Consequently, the inverse image of $\mathbf{S}^m_+$ under an affine mapping is also a convex set given by a number of (non-strict) polynomial inequalities. And it is known that every projection of such a set is also given by a number of polynomial inequalities (both strict and non-strict). We conclude that

> *A SD-representable set always is a convex set given by finitely many polynomial inequalities (strict and non-strict).*

A natural (and seemingly very difficult) question is whether the inverse is true – whether a convex set given by a number of polynomial inequalities is always SDr. This question can be simplified in many ways – we may fix the dimension of the set, we may assume the polynomials participating in inequalities to be convex, we may fix the degrees of the polynomials, etc.; to the best of our knowledge, all these questions are open.

The goal of the subsequent exercises is to answer affirmatively the simplest question of the above series:

> *Let $\pi(x)$ be a convex polynomial of one variable. Then its epigraph*
>
> $$\{(t, x) \in \mathbf{R}^2 \mid t \geq \pi(x)\}$$
>
> *is SDr.*

Let us fix a nonnegative integer k and consider the curve

$$p(x) = (1, x, x^2, ..., x^{2k})^T \in \mathbf{R}^{2k+1}.$$

Let Π_k be the closure of the convex hull of values of the curve. How can one describe Π_k?

A convenient way to answer this question is to pass to a *matrix representation* of all objects involved. Namely, let us associate with a vector $\xi = (\xi_0, \xi_1, ..., \xi_{2k}) \in \mathbf{R}^{2k+1}$ the $(k+1) \times (k+1)$ symmetric matrix

$$\mathcal{M}(\xi) = \begin{pmatrix} \xi_0 & \xi_1 & \xi_2 & \xi_3 & \cdots & \xi_k \\ \xi_1 & \xi_2 & \xi_3 & \xi_4 & \cdots & \xi_{k+1} \\ \xi_2 & \xi_3 & \xi_4 & \xi_5 & \cdots & \xi_{k+2} \\ \xi_3 & \xi_4 & \xi_5 & \xi_6 & \cdots & \xi_{k+3} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \xi_k & \xi_{k+1} & \xi_{k+2} & \xi_{k+3} & ... & \xi_{2k} \end{pmatrix},$$

so that

$$[\mathcal{M}(\xi)]_{ij} = \xi_{i+j}, \ i, j = 0, ..., k.$$

The transformation $\xi \mapsto \mathcal{M}(\xi) : \mathbf{R}^{2k+1} \to \mathbf{S}^{k+1}$ is a linear embedding; the image of Π_k under this embedding is the closure of the convex hull of values of the curve

$$P(x) = \mathcal{M}(p(x)).$$

It follows that the image $\widehat{\Pi}_k$ of Π_k under the mapping $\mathcal{M}$ possesses the following properties:

(i) $\widehat{\Pi}_k$ belongs to the image of $\mathcal{M}$, i.e., to the subspace H_k of $\mathbf{S}^{2k+1}$ comprised of *Hankel* matrices – matrices with entries depending on the sum of indices only:

$$H_k = \left\{ X \in \mathbf{S}^{2k+1} | i + j = i' + j' \Rightarrow X_{ij} = X_{i'j'} \right\};$$

(ii) $\widehat{\Pi}_k \subset \mathbf{S}^{k+1}_+$ (indeed, every matrix $\mathcal{M}(p(x))$ is positive semidefinite);

(iii) For every $X \in \widehat{\Pi}_k$ one has $X_{00} = 1$.

It turns out that properties (i) – (iii) characterize $\widehat{\Pi}_k$:

> **(G)** *A symmetric $(k+1) \times (k+1)$ matrix X belongs to $\widehat{\Pi}_k$ if and only if it possesses the properties* (i) – (iii)*: its entries depend on sum of indices only (i.e., $X \in H_k$), X is positive semidefinite and $X_{00} = 1$.*

(G) is a particular case of the classical results related to the so called "moment problem". The goal of the subsequent exercises is to give a simple alternative proof of this statement.

Note that the mapping $\mathcal{M}^* : \mathbf{S}^{k+1} \to \mathbf{R}^{2k+1}$ conjugate to the mapping $\mathcal{M}$ is as follows:

$$(\mathcal{M}^* X)_l = \sum_{i=0}^{l} X_{i,l-i}, \ l = 0, 1, ..., 2k,$$

and we know something about this mapping: Example 21a (Lecture 3) says that

> **(H)** *The image of the cone $\mathbf{S}^{k+1}_+$ under the mapping $\mathcal{M}^*$ is exactly the cone of coefficients of polynomials of degree $\leq 2k$ which are nonnegative on the entire real line.*

Exercise 3.22 *Derive* **(G)** *from* **(H)**.

(G), among other useful things, implies the result we need:

> **(I)** *Let $\pi(x) = \pi_0 + \pi_1 x + \pi_2 x^2 + ... + \pi_{2k} x^{2k}$ be a convex polynomial of degree $2k$. Then the epigraph of π is SDr:*
>
> $$\{(t, x) \in \mathbf{R}^2 \mid t \geq p(x)\} = \mathcal{X}[\pi],$$

where

$$\mathcal{X}[\pi] = \left\{(t,x)\,\middle|\,\exists x_2,...,x_{2k}: \begin{pmatrix} 1 & x & x_2 & x_3 & ... & x_k \\ x & x_2 & x_3 & x_4 & ... & x_{k+1} \\ x_2 & x_3 & x_4 & x_5 & ... & x_{k+2} \\ x_3 & x_4 & x_5 & x_6 & ... & x_{k+3} \\ ... & ... & ... & ... & \ddots & ... \\ x_k & x_{k+1} & x_{k+2} & x_{k+3} & ... & x_{2k} \end{pmatrix} \succeq 0, \right.$$
$$\left. \pi_0+\pi_1x+\pi_2x_2+\pi_3x_3+...+\pi_{2k}x_{2k} \le t \right\}$$

Exercise 3.23 *Prove* (I).

Note that the set $\mathcal{X}[\pi]$ makes sense for an arbitrary polynomial π, not necessary for a convex one. What is the projection of this set onto the (t,x)-plane? The answer is surprisingly nice: this is the convex hull of the epigraph of the polynomial π!

Exercise 3.24 *Let $\pi(x)=\pi_0+\pi_1x+...+\pi_{2k}x^{2k}$ with $\pi_{2k}>0$, and let*

$$G[\pi]=\mathrm{Conv}\{(t,x)\in\mathbf{R}^2 \mid t\ge p(x)\}$$

be the convex hull of the epigraph of π (the set of all convex combinations of points from the epigraph of π).

1) *Prove that $G[\pi]$ is a closed convex set.*
2) *Prove that*

$$G[\pi]=\mathcal{X}[\pi].$$

3.7.3 Around the Lovasz capacity number and semidefinite relaxations of combinatorial problems

Recall that the Lovasz capacity number $\Theta(\Gamma)$ of an n-node graph Γ is the optimal value of the following semidefinite program:

$$\min_{\lambda,x}\{\lambda : \lambda I_n-\mathcal{L}(x)\succeq 0\} \tag{L}$$

where the symmetric $n\times n$ matrix $\mathcal{L}(x)$ is defined as follows:

- the dimension of x is equal to the number of arcs in Γ, and the coordinates of x are indexed by these arcs;
- the element of $\mathcal{L}(x)$ in an "empty" cell ij (one for which the nodes i and j are not linked by an arc in Γ) is 1;
- the elements of $\mathcal{L}(x)$ in a pair of symmetric "non-empty" cells ij, ji (those for which the nodes i and j are linked by an arc) are equal to the coordinate of x indexed by the corresponding arc.

As we remember, the importance of $\Theta(\Gamma)$ comes from the fact that $\Theta(\Gamma)$ is a computable upper bound on the stability number $\alpha(\Gamma)$ of the graph. We have seen also that the Shor semidefinite relaxation of the problem of finding the stability number of Γ leads to a "seemingly stronger" upper bound on $\alpha(\Gamma)$, namely, the optimal value $\sigma(\Gamma)$ in the semidefinite program

$$\min_{\lambda,\mu,\nu}\left\{\lambda : \begin{pmatrix} \lambda & -\frac{1}{2}(e+\mu)^T \\ -\frac{1}{2}(e+\mu) & A(\mu,\nu)\end{pmatrix}\succeq 0\right\} \tag{Sh}$$

where $e=(1,...,1)^T\in\mathbf{R}^n$ and $A(\mu,\nu)$ is the matrix as follows:

- the dimension of ν is equal to the number of arcs in Γ, and the coordinates of ν are indexed by these arcs;
- the diagonal entries of $A(\mu,\nu)$ are $\mu_1,...,\mu_n$;

- the off-diagonal entries of $A(\mu,\nu)$ corresponding to "empty cells" are zeros;
- the off-diagonal entries of $A(\mu,\nu)$ in a pair of symmetric "non-empty" cells ij, ji are equal to the coordinate of ν indexed by the corresponding arc.

We have seen that (L) can be obtained from (Sh) when the variables μ_i are set to 1, so that $\sigma(\Gamma) \le \Theta(\Gamma)$. Thus,

$$\alpha(\Gamma) \le \sigma(\Gamma) \le \Theta(\Gamma). \tag{3.7.2}$$

Exercise 3.25 1) *Prove that if* (λ,μ,ν) *is a feasible solution to* (Sh)*, then there exists a symmetric* $n\times n$ *matrix* A *such that* $\lambda I_n - A \succeq 0$ *and at the same time the diagonal entries of* A *and the off-diagonal entries in the "empty cells" are* ≥ 1*. Derive from this observation that the optimal value in* (Sh) *is not less than the optimal value* $\Theta'(\Gamma)$ *in the following semidefinite program:*

$$\min_{\lambda,x}\{\lambda : \lambda I_n - X \succeq 0, X_{ij} \ge 1 \text{ whenever } i,j \text{ are not adjacent in } \Gamma\} \tag{Sc}.$$

2) *Prove that* $\Theta'(\Gamma) \ge \alpha(\Gamma)$.

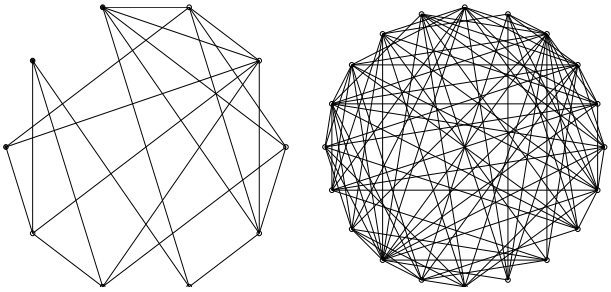
Hint: Demonstrate that if all entries of a symmetric $k\times k$ matrix are ≥ 1, then the maximum eigenvalue of the matrix is at least k. Derive from this observation and the Interlacing Eigenvalues Theorem (Exercise 3.4.(ii)) that if a symmetric matrix contains a principal $k\times k$ submatrix with entries ≥ 1, then the maximum eigenvalue of the matrix is at least k.

The upper bound $\Theta'(\Gamma)$ on the stability number of Γ is called the *Schrijver* capacity of graph Γ. Note that we have

$$\alpha(\Gamma) \le \Theta'(\Gamma) \le \sigma(\Gamma) \le \Theta(\Gamma).$$

A natural question is which inequalities in this chain may happen to be strict. In order to answer it, we have computed the quantities in question for about 2,000 random graphs with number of nodes varying 8 to 20. In our experiments, the stability number was computed – by brute force – for graphs with ≤ 12 nodes; for all these graphs, the integral part of $\Theta(\Gamma)$ was equal to $\alpha(\Gamma)$. Furthermore, $\Theta(\Gamma)$ was non-integer in 156 of our 2,000 experiments, and in 27 of these 156 cases the Schrijver capacity number $\Theta'(\Gamma)$ was strictly less than $\Theta(\Gamma)$. The quantities $\Theta'(\cdot), \sigma(\cdot), \Theta(\cdot)$ for 13 of these 27 cases are listed in the table below:

Graph #	# of nodes	α	Θ'	σ	Θ
1	20	?	4.373	4.378	4.378
2	20	?	5.062	5.068	5.068
3	20	?	4.383	4.389	4.389
4	20	?	4.216	4.224	4.224
5	13	?	4.105	4.114	4.114
6	20	?	5.302	5.312	5.312
7	20	?	6.105	6.115	6.115
8	20	?	5.265	5.280	5.280
9	9	3	3.064	3.094	3.094
10	12	4	4.197	4.236	4.236
11	8	3	3.236	3.302	3.302
12	12	4	4.236	4.338	4.338
13	10	3	3.236	3.338	3.338



Graphs # 13 (left) and # 8 (right); all nodes are on circumferences.

Exercise 3.26 *Compute the stability numbers of the graphs # 8 and # 13.*

Exercise 3.27 *Prove that* $\sigma(\Gamma) = \Theta(\Gamma)$.

The *chromatic number* $\xi(\Gamma)$ of a graph Γ is the minimal number of colours such that one can colour the nodes of the graph in such a way that no two adjacent (i.e., linked by an arc) nodes get the same colour[20]. The *complement* $\bar{\Gamma}$ of a graph Γ is the graph with the same set of nodes, and two distinct nodes in $\bar{\Gamma}$ are linked by an arc if and only if they are *not* linked by an arc in Γ.

Lovasz proved that for every graph

$$\Theta(\Gamma) \leq \xi(\bar{\Gamma}) \qquad (*)$$

so that

$$\alpha(\Gamma) \leq \Theta(\Gamma) \leq \xi(\bar{\Gamma})$$

(Lovasz's Sandwich Theorem).

Exercise 3.28 *Prove* (*).

Hint: Let us colour the vertices of Γ in $k = \xi(\bar{\Gamma})$ colours in such a way that no two vertices of the same colour are adjacent in $\bar{\Gamma}$, i.e., every two nodes of the same colour are adjacent in Γ. Set $\lambda = k$, and let x be such that

$$[\mathcal{L}(x)]_{ij} = \begin{cases} -(k-1), & i \neq j,\ i,j \text{ are of the same colour} \\ 1, & \text{otherwise} \end{cases}$$

Prove that (λ, x) is a feasible solution to (L).

Now let us switch from the Lovasz capacity number to semidefinite relaxations of combinatorial problems, specifically to those of maximizing a quadratic form over the vertices of the unit cube, and over the entire cube:

$$\begin{array}{ll} (a) & \max\limits_x \left\{x^T A x : x \in \mathrm{Vrt}(C_n) = \{x \in \mathbf{R}^n \mid x_i = \pm 1\ \forall i\}\right\} \\ (b) & \max\limits_x \left\{x^T A x : x \in C_n = \{x \in \mathbf{R}^n \mid -1 \leq x_i \leq 1,\ \forall i\}\right\} \end{array} \qquad (3.7.3)$$

The standard semidefinite relaxations of the problems are, respectively, the problems

$$\begin{array}{ll} (a) & \max\limits_X \{\mathrm{Tr}(AX) : X \succeq 0, X_{ii} = 1,\ i = 1, ..., n\}, \\ (b) & \max\limits_X \{\mathrm{Tr}(AX) : X \succeq 0, X_{ii} \leq 1,\ i = 1, ..., n\}; \end{array} \qquad (3.7.4)$$

the optimal value of a relaxation is an upper bound for the optimal value of the respective original problem.

[20] E.g., when colouring a geographic map, it is convenient not to use the same colour for a pair of countries with a common border. It was observed that to meet this requirement for actual maps, 4 colours are sufficient. The famous "4-colour" Conjecture claims that this is so for *every* geographic map. Mathematically, you can represent a map by a graph, where the nodes represent the countries, and two nodes are linked by an arc if and only if the corresponding countries have common border. A characteristic feature of such a graph is that it is *planar* – you may draw it on 2D plane in such a way that the arcs will not cross each other, meeting only at the nodes. Thus, mathematical form of the 4-colour Conjecture is that the chromatic number of any planar graph is at most 4. This is indeed true, but it took about 100 years to prove the conjecture!

Exercise 3.29 *Let $A \in \mathbf{S}^n$. Prove that*

$$\max_{x:x_i=\pm 1,\, i=1,\ldots,n} x^T A x \geq \mathrm{Tr}(A).$$

Develop an efficient algorithm which, given A, generates a point x with coordinates ± 1 such that $x^T A x \geq \mathrm{Tr}(A)$.

Exercise 3.30 *Prove that if the diagonal entries of A are nonnegative, then the optimal values in (3.7.4.a) and (3.7.4.b) are equal to each other. Thus, in the case in question, the relaxations "do not understand" whether we are maximizing over the vertices of the cube or over the entire cube.*

Exercise 3.31 *Prove that the problems dual to (3.7.4.a, b) are, respectively,*

$$\begin{array}{ll} (a) & \min\limits_{\Lambda} \left\{\mathrm{Tr}(\Lambda) : \Lambda \succeq A, \Lambda \text{ is diagonal}\right\}, \\ (b) & \min\limits_{\Lambda} \left\{\mathrm{Tr}(\Lambda) : \Lambda \succeq A, \Lambda \succeq 0, \Lambda \text{ is diagonal}\right\}; \end{array} \tag{3.7.5}$$

the optimal values in these problems are equal to those of the respective problems in (3.7.4) and are therefore upper bounds on the optimal values of the respective combinatorial problems from (3.7.3).

The latter claim is quite transparent, since the problems (3.7.5) can be obtained as follows:

• In order to bound from above the optimal value of a quadratic form $x^T A x$ on a given set S, we look at those quadratic forms $x^T \Lambda x$ which can be easily maximized over S. For the case of $S = \mathrm{Vrt}(C_n)$ these are quadratic forms with diagonal matrices Λ, and for the case of $S = C_n$ these are quadratic forms with diagonal and positive semidefinite matrices Λ; in both cases, the respective maxima are merely $\mathrm{Tr}(\Lambda)$.

• Having specified a family $\mathcal{F}$ of quadratic forms $x^T \Lambda x$ "easily optimizable over S", we then look at those forms from $\mathcal{F}$ which majorate *everywhere* the original quadratic form $x^T A x$, and take among these forms the one with the minimal $\max\limits_{x \in S} x^T \Lambda x$, thus coming to the problem

$$\min_{\Lambda} \left\{ \max_{x \in S} x^T \Lambda x : \Lambda \succeq A, \Lambda \in \mathcal{F} \right\}. \tag{!}$$

It is evident that the optimal value in this problem is an upper bound on $\max\limits_{x \in S} x^T A x$. It is also immediately seen that in the case of $S = \mathrm{Vrt}(C_n)$ the problem (!), with $\mathcal{F}$ specified as the set $\mathcal{D}$ of all diagonal matrices, is equivalent to (3.7.5.a), while in the case of $S = C_n$ (!), with $\mathcal{F}$ specified as the set $\mathcal{D}_+$ of positive semidefinite diagonal matrices, is nothing but (3.7.5.b).

Given the direct and quite transparent road leading to (3.7.5.a, b), we can try to move a little bit further along this road. To this end observe that there are trivial upper bounds on the maximum of *an arbitrary* quadratic form $x^T \Lambda x$ over $\mathrm{Vrt}(C_n)$ and C_n, specifically:

$$\max_{x \in \mathrm{Vrt}(C_n)} x^T \Lambda x \leq \mathrm{Tr}(\Lambda) + \sum_{i \neq j} |\Lambda_{ij}|, \quad \max_{x \in C_n} x^T \Lambda x \leq \sum_{i,j} |\Lambda_{ij}|.$$

For the above families $\mathcal{D}$, $\mathcal{D}_+$ of matrices Λ for which $x^T \Lambda x$ is "easily optimizable" over $\mathrm{Vrt}(C_n)$, respectively, C_n, the above bounds are equal to the precise values of the respective maxima. Now let us update (!) as follows: we eliminate the restriction $\Lambda \in \mathcal{F}$, replacing simultaneously the objective $\max\limits_{x \in S} x^T \Lambda x$ with its upper bound, thus coming to the pair of problems

$$\begin{array}{lll} (a) & \min\limits_{\Lambda} \left\{ \mathrm{Tr}(\Lambda) + \sum\limits_{i \neq j} |\Lambda_{ij}| : \Lambda \succeq A \right\} & [S = \mathrm{Vrt}(C_n)] \\ (b) & \min\limits_{\Lambda} \left\{ \sum\limits_{i,j} |\Lambda_{ij}| : \Lambda \succeq A \right\} & [S = C_n] \end{array} \tag{3.7.6}$$

From the origin of the problems it is clear that they still yield upper bounds on the optimal values of the respective problems (3.7.3.a,b), and that these bounds are at least as good as the bounds yielded by the standard relaxations (3.7.4.a,b):

$$\begin{array}{ll}(a) & \mathrm{Opt}(3.7.3.a) \le \mathrm{Opt}(3.7.6.a) \underbrace{\le}_{(*)} \mathrm{Opt}(3.7.4.a) = \mathrm{Opt}(3.7.5.a),\\ (b) & \mathrm{Opt}(3.7.3.b) \le \mathrm{Opt}(3.7.6.b) \underbrace{\le}_{(**)} \mathrm{Opt}(3.7.4.b) = \mathrm{Opt}(3.7.5.b),\end{array} \tag{3.7.7}$$

where $\mathrm{Opt}(\cdot)$ means the optimal value of the corresponding problem.

Indeed, consider the problem (3.7.6.a). Whenever Λ is a feasible solution of this problem, the quadratic form $x^T\Lambda x$ majorates everywhere the form x^TAx, so that $\max\limits_{x\in\mathrm{Vrt}(C_n)} x^TAx \le \max\limits_{x\in\mathrm{Vrt}(C_n)} x^T\Lambda x$; the latter quantity, in turn, is majorated by $\mathrm{Tr}(\Lambda) + \sum\limits_{i\ne j}|\Lambda_{ij}|$, whence the value of the objective of the problem (3.7.6.a) at every feasible solution of the problem majorates the quantity $\max\limits_{x\in\mathrm{Vrt}(C_n)} x^TAx$. Thus, the optimal value in (3.7.6.a) is an upper bound on the maximum of x^TAx over the vertices of the cube C_n. At the same time, when passing from the (dual form of the) standard relaxation (3.7.5.a) to our new bounding problem (3.7.6.a), we only extend the feasible set and do not vary the objective at the "old" feasible set; as a result of such a modification, the optimal value may only decrease. Thus, the upper bound on the maximum of x^TAx over $\mathrm{Vrt}(C_n)$ yielded by (3.7.6.a) is at least as good as those (equal to each other) bounds yielded by the standard relaxations (3.7.4.a), (3.7.5.a), as required in (3.7.7.a). Similar reasoning proves (3.7.7.b).

Note that problems (3.7.6) are equivalent to semidefinite programs and thus are of the same status of "computational tractability" as the standard SDP relaxations (3.7.5) of the combinatorial problems in question. At the same time, our new bounding problems are more difficult than the standard SDP relaxations. Can we justify this by getting an improvement in the quality of the bounds?

Exercise 3.32 *Find out whether the problems* (3.7.6.a,b) *yield better bounds than the respective problems* (3.7.5.a,b)*, i.e., whether the inequalities* (*), (**) *in* (3.7.7) *can be strict.*

Hint: Look at the problems dual to (3.7.6.a,b).

Exercise 3.33 *Let D be a given subset of $\mathbf{R}^n_+$. Consider the following pair of optimization problems:*

$$\begin{array}{ll}\max\limits_x \left\{x^TAx : (x_1^2, x_2^2, ..., x_n^2)^T \in D\right\} & (P)\\ \max\limits_X \{\mathrm{Tr}(AX) : X \succeq 0, \mathrm{Dg}(X) \in D\} & (R)\end{array}$$

($\mathrm{Dg}(X)$ is the diagonal of a square matrix X). Note that when $D = \{(1, ..., 1)^T\}$, (P) is the problem of maximizing a quadratic form over the vertices of C_n, while (R) is the standard semidefinite relaxation of (P); when $D = \{x \in \mathbf{R}^n \mid 0 \le x_i \le 1\ \forall i\}$, (P) is the problem of maximizing a quadratic form over the cube C_n, and (R) is the standard semidefinite relaxation of the latter problem.

1) *Prove that if D is semidefinite-representable, then (R) can be reformulated as a semidefinite program.*

2) *Prove that (R) is a relaxation of (P), i.e., that*

$$\mathrm{Opt}(P) \le \mathrm{Opt}(R).$$

3) [Nesterov; Ye] *Let $A \succeq 0$. Prove that then*

$$\mathrm{Opt}(P) \le \mathrm{Opt}(R) \le \frac{\pi}{2}\mathrm{Opt}(P).$$

Hint: Use Nesterov's Theorem (Theorem 3.4.2).

Exercise 3.34 *Let $A \in \mathbf{S}^m_+$. Prove that*

$$\max\{x^T Ax \mid x_i = \pm 1,\ i = 1, ..., m\} = \max\{\frac{2}{\pi} \sum_{i,j=1}^{m} a_{ij} \mathrm{asin}(X_{ij}) \mid X \succeq 0, X_{ii} = 1,\ i = 1, ..., m\}.$$

3.7.4 Around Lyapunov Stability Analysis

A natural mathematical model of a swing is the linear time invariant dynamic system

$$y''(t) = -\omega^2 y(t) - 2\mu y'(t) \tag{S}$$

with positive ω^2 and $0 \leq \mu < \omega$ (the term $2\mu y'(t)$ represents friction). A general solution to this equation is

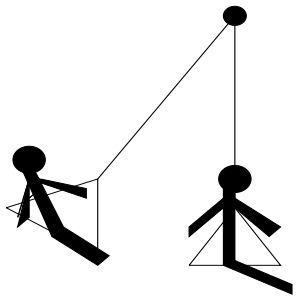
$$y(t) = a \cos(\omega' t + \phi_0) \exp\{-\mu t\},\ \omega' = \sqrt{\omega^2 - \mu^2},$$

with free parameters a and ϕ_0, i.e., this is a decaying oscillation. Note that the equilibrium

$$y(t) \equiv 0$$

is stable – every solution to (S) converges to 0, along with its derivative, exponentially fast.

After stability is observed, an immediate question arises: how is it possible to swing on a swing? Everybody knows from practice that it is possible. On the other hand, since the equilibrium is stable, it looks as if it was impossible to swing, without somebody's assistance, for a long time. The reason which makes swinging possible is highly nontrivial – *parametric resonance*. A swinging child does not sit on the swing in a once forever fixed position; what he does is shown below:



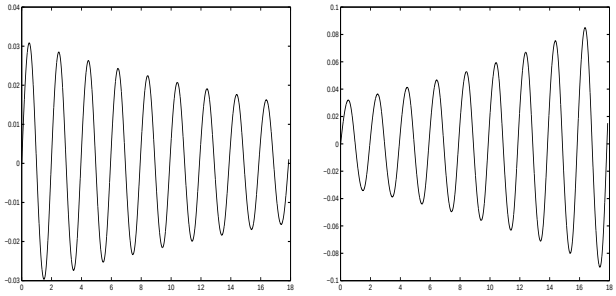
A swinging child

As a result, the "effective length" of the swing l – the distance from the point where the rope is fixed to the center of gravity of the system – is varying with time: $l = l(t)$. Basic mechanics says that $\omega^2 = g/l$, g being the gravity acceleration. Thus, the actual swing is a *time-varying* linear dynamic system:

$$y''(t) = -\omega^2(t) y(t) - 2\mu y'(t), \tag{S$'$}$$

and it turns out that for properly varied $\omega(t)$ the equilibrium $y(t) \equiv 0$ is *not* stable. A swinging child is just varying $l(t)$ in a way which results in an *unstable* dynamic system (S$'$), and this instability is in fact

what the child enjoys...



$$y''(t) = -\frac{g}{l+h\sin(2\omega t)}y(t) - 2\mu y'(t), \quad y(0) = 0, y'(0) = 0.1$$

$$\left[l = 1\,[\mathrm{m}], g = 10\,[\tfrac{\mathrm{m}}{\mathrm{sec}^2}], \mu = 0.15[\tfrac{1}{\mathrm{sec}}], \omega = \sqrt{g/l}\right]$$

Graph of $y(t)$

left: $h = 0.125$: this child is too small; he should grow up...

right: $h = 0.25$: this child can already swing...

Exercise 3.35 *Assume that you are given parameters l ("nominal length of the swing rope"), $h > 0$ and $\mu > 0$, and it is known that a swinging child can vary the "effective length" of the rope within the bounds $l \pm h$, i.e., his/her movement is governed by the uncertain linear time-varying system*

$$y''(t) = -a(t)y(t) - 2\mu y'(t), \quad a(t) \in \left[\frac{g}{l+h}, \frac{g}{l-h}\right].$$

Try to identify the domain in the 3D-space of parameters l, μ, h where the system is stable, as well as the domain where its stability can be certified by a quadratic Lyapunov function. What is "the difference" between these two domains?

3.7.5 Around Nesterov's $\frac{\pi}{2}$ Theorem

Exercise 3.36 *Prove the following statement:*

Proposition 3.7.1 [Nesterov;Ye] *Consider the optimization program*

$$\mathrm{Opt} = \max_x \left\{x^T A x : x^T B_i x \le 1,\ i = 1, ..., m\right\}, \tag{P}$$

along with its semidefinite relaxation

$$\mathrm{SDP} = \max_X \left\{\mathrm{Tr}(AX) : \mathrm{Tr}(B_i X) \le b_i,\ i = 1, ..., m, X \succeq 0\right\} \tag{SDP}$$

and the dual of the latter problem:

$$\min_\lambda \left\{\sum_i \lambda_i b_i : \sum_i \lambda_i B_i \succeq A,\ \lambda \ge 0\right\}. \tag{SDD}$$

Assume that

1. *The matrices $B_1, ..., B_m$ commute with each other;*
2. *There exists a combination of the matrices B_i with nonnegative coefficients which is positive definite;*
3. *$A \succeq 0$.*

Then $\mathrm{Opt} \ge 0$, *(SDP) and (SDD) are solvable with equal optimal values, and*

$$\mathrm{Opt} \le \mathrm{SDP} \le \frac{\pi}{2}\mathrm{Opt}. \tag{3.7.8}$$

Hint: Observe that since B_i are commuting symmetric matrices, they share a common orthogonal eigenbasis, so that w.l.o.g. we can assume that all B_i's are diagonal. In this latter case, use Theorem 3.4.3.

3.7.6 Around ellipsoidal approximations

Exercise 3.37 *Prove the Löwner – Fritz John Theorem (Theorem 3.6.1).*

More on ellipsoidal approximations of sums of ellipsoids. The goal of two subsequent exercises is to get in an alternative way the problem $(\tilde{O})$ "generating" a parametric family of ellipsoids containing the arithmetic sum of m given ellipsoids (Section 3.6.2).

Exercise 3.38 *Let P_i be nonsingular, and Λ_i be positive definite $n\times n$ matrices, $i=1,...,m$. Prove that for every collection $x^1,...,x^m$ of vectors from $\mathbf{R}^n$ one has*

$$[x^1+...+x^m]^T\left[\sum_{i=1}^m [P_i^T]^{-1}\Lambda_i^{-1}P_i^{-1}\right]^{-1}[x^1+...+x^m]\le\sum_{i=1}^m [x^i]^T P_i\Lambda_i P_i^T x^i. \tag{3.7.9}$$

Hint: Consider the $(nm+n)\times(nm+n)$ symmetric matrix

$$A=\left[\begin{array}{ccc|c} P_1\Lambda_1P_1^T & & & I_n\\ & \ddots & & \vdots\\ & & P_m\Lambda_mP_m^T & I_n\\ \hline I_n & \cdots & I_n & \sum\limits_{i=1}^m [P_i^T]^{-1}\Lambda_i^{-1}P_i^{-1}\end{array}\right]$$

and apply twice the Schur Complement Lemma: first time – to prove that the matrix is positive semidefinite, and the second time – to get from the latter fact the desired inequality.

Exercise 3.39 *Assume you are given m full-dimensional ellipsoids centered at the origin*

$$W_i=\{x\in\mathbf{R}^n\mid x^TB_ix\le 1\},\ i=1,...,m \qquad [B_i\succ 0]$$

in $\mathbf{R}^n$.

1) *Prove that for every collection Λ of positive definite $n\times n$ matrices Λ_i such that*

$$\sum_i\lambda_{\max}(\Lambda_i)\le 1$$

the ellipsoid

$$E_\Lambda=\{x\mid x^T\left[\sum_{i=1}^m B_i^{-1/2}\Lambda_i^{-1}B_i^{-1/2}\right]^{-1}x\le 1\}$$

contains the sum $W_1+...+W_m$ of the ellipsoids W_i.

2) *Prove that in order to find the smallest volume ellipsoid in the family $\{E_\Lambda\}_\Lambda$ defined in 1) it suffices to solve the semidefinite program*

$$\begin{array}{ll}
 & \text{maximize } t \text{ s.t.}\\
(a) & t\le \mathrm{Det}^{1/n}(Z),\\
(b) & \left[\begin{array}{cccc}\Lambda_1 & & & \\ & \Lambda_2 & & \\ & & \ddots & \\ & & & \Lambda_m\end{array}\right]\succeq\left[\begin{array}{cccc} B_1^{-1}ZB_1^{-1} & B_1^{-1/2}ZB_2^{-1/2} & ... & B_1^{-1/2}ZB_m^{-1/2}\\ B_2^{-1/2}ZB_1^{-1/2} & B_2^{-1/2}ZB_2^{-1/2} & ... & B_2^{-1/2}ZB_m^{-1/2}\\ \vdots & \vdots & \ddots & \vdots\\ B_m^{-1/2}ZB_1^{-1/2} & B_m^{-1/2}ZB_2^{-1/2} & ... & B_m^{-1/2}ZB_m^{-1/2}\end{array}\right],\\
(c) & Z\succeq 0,\\
(d) & \Lambda_i\preceq\lambda_iI_n,\ i=1,...,m,\\
(e) & \sum\limits_{i=1}^m\lambda_i\le 1
\end{array} \tag{3.7.10}$$

in variables $Z,\Lambda_i\in\mathbf{S}^n,t,\lambda_i\in\mathbf{R}$; the smallest volume ellipsoid in the family $\{E_\Lambda\}_\Lambda$ is E_{Λ^*}, where Λ^* is the "Λ-part" of an optimal solution of the problem.

Hint: Use example 20c from Lecture 3.

3) *Demonstrate that the optimal value in* (3.7.10) *remains unchanged when the matrices* Λ_i *are further restricted to be scalar:* $\Lambda_i = \lambda_i I_n$. *Prove that with this additional constraint problem* (3.7.10) *becomes equivalent to problem* $(\tilde{\mathrm{O}})$ *from Section 3.6.2.*

Remark 3.7.1 Exercise 3.39 demonstrates that the approximating scheme for solving problem (O) presented in Proposition 3.6.4 is equivalent to the following one:

> *Given* m *positive reals* λ_i *with unit sum, one defines the ellipsoid* $E(\lambda) = \{x \mid x^T \left[\sum_{i=1}^m \lambda_i^{-1} B_i^{-1}\right]^{-1} x \leq 1\}$. *This ellipsoid contains the arithmetic sum* W *of the ellipsoids* $\{x \mid x^T B_i x \leq 1\}$, *and in order to approximate the smallest volume ellipsoid containing* W, *we merely minimize* $\mathrm{Det}(E(\lambda))$ *over* λ *varying in the standard simplex* $\{\lambda \geq 0, \sum_i \lambda_i = 1\}$.

In this form, the approximation scheme in question was proposed by Schweppe (1975).

Exercise 3.40 *Let* A_i *be nonsingular* $n \times n$ *matrices,* $i = 1, ..., m$, *and let* $W_i = \{x = A_i u \mid u^T u \leq 1\}$ *be the associated ellipsoids in* $\mathbf{R}^n$. *Let* $\Delta_m = \{\lambda \in \mathbf{R}^m_+ \mid \sum_i \lambda_i = 1\}$. *Prove that*

1) *Whenever* $\lambda \in \Delta_m$ *and* $A \in \mathbf{M}^{n,n}$ *is such that*

$$AA^T \succeq F(\lambda) \equiv \sum_{i=1}^m \lambda_i^{-1} A_i A_i^T,$$

the ellipsoid $E[A] = \{x = Au \mid u^T u \leq 1\}$ *contains* $W = W_1 + ... + W_m$.

Hint: Use the result of Exercise 3.39.1)

2) *Whenever* $A \in \mathbf{M}^{n,n}$ *is such that*

$$AA^T \preceq F(\lambda) \quad \forall \lambda \in \Delta_m,$$

the ellipsoid $E[A]$ *is contained in* $W_1 + ... + W_m$, *and vice versa.*

Hint: Note that

$$\left(\sum_{i=1}^m |\alpha_i|\right)^2 = \min_{\lambda \in \Delta_m} \sum_{i=1}^m \frac{\alpha_i^2}{\lambda_i}$$

and use statement **(F)** from Section 3.6.2.

"Simple" ellipsoidal approximations of sums of ellipsoids. Let $W_i = \{x = A_i u \mid u^T u \leq 1\}$, $i = 1, ..., m$, be full-dimensional ellipsoids in $\mathbf{R}^n$ (so that A_i are nonsingular $n \times n$ matrices), and let $W = W_1 + ... + W_m$ be the arithmetic sum of these ellipsoids. Observe that W is the image of the set

$$\mathcal{B} = \left\{ u = \begin{bmatrix} u[1] \\ \vdots \\ u[m] \end{bmatrix} \in \mathbf{R}^{nm} \mid u^T[i] u[i] \leq 1, \ i = 1, ..., m \right\}$$

under the linear mapping

$$u \mapsto \mathcal{A}u = \sum_{i=1}^m A_i u[i] : \mathbf{R}^{nm} \to \mathbf{R}^n.$$

It follows that

> *Whenever an* nm*-dimensional ellipsoid* $\mathcal{W}$ *contains* $\mathcal{B}$, *the set* $\mathcal{A}(\mathcal{W})$, *which is an* n*-dimensional ellipsoid* (why?) *contains* W, *and whenever* $\mathcal{W}$ *is contained in* $\mathcal{B}$, *the ellipsoid* $\mathcal{A}(\mathcal{W})$ *is contained in* W.

In view of this observation, we can try to approximate W from inside and from outside by the ellipsoids $W_- \equiv \mathcal{A}(\mathcal{W}_-)$ and $W^+ = \mathcal{A}(\mathcal{W}^+)$, where $\mathcal{W}_-$ and $\mathcal{W}^+$ are, respectively, the largest and the smallest volume nm-dimensional ellipsoids contained in/containing $\mathcal{B}$.

Exercise 3.41 1) *Prove that*

$$\begin{aligned}\mathcal{W}_- &= \{u \in \mathbf{R}^{nm} \mid \sum_{i=1}^m u^T[i]u[i] \le 1\},\\ \mathcal{W}^+ &= \{u \in \mathbf{R}^{nm} \mid \sum_{i=1}^m u^T[i]u[i] \le m\},\end{aligned}$$

so that

$$\begin{aligned}W \supset W_- &\equiv \{x = \sum_{i=1}^m A_i u[i] \mid \sum_{i=1}^m u^T[i]u[i] \le 1\},\\ W \subset W_+ &\equiv \{x = \sum_{i=1}^m A_i u[i] \mid \sum_{i=1}^m u^T[i]u[i] \le m\} = \sqrt{m}W_-.\end{aligned}$$

2) *Prove that W_- can be represented as*

$$W_- = \{x = Bu \mid u \in \mathbf{R}^n, u^T u \le 1\}$$

with matrix $B \succ 0$ representable as

$$B = \sum_{i=1}^m A_i X_i$$

with square matrices X_i of norms $|X_i| \le 1$.

Derive from this observation that the "level of conservativeness" of the inner ellipsoidal approximation of W given by Proposition 3.6.6 is at most $\sqrt{m}$: if W_* is this inner ellipsoidal approximation and W_{**} is the largest volume ellipsoid contained in W, then

$$\left(\frac{\mathrm{Vol}(W_{**})}{\mathrm{Vol}(W_*)}\right)^{1/n} \le \left(\frac{\mathrm{Vol}(W)}{\mathrm{Vol}(W_*)}\right)^{1/n} \le \sqrt{m}.$$

Invariant ellipsoids

Exercise 3.42 *Consider a discrete time controlled dynamic system*

$$\begin{aligned}x(t+1) &= Ax(t) + bu(t),\ t = 0, 1, 2, ...\\ x(0) &= 0,\end{aligned}$$

where $x(t) \in \mathbf{R}^n$ is the state vector and $u(t) \in [-1, 1]$ is the control at time t. An ellipsoid centered at the origin

$$W = \{x \mid x^T Z x \le 1\} \quad [Z \succ 0]$$

is called invariant, if

$$x \in W \Rightarrow Ax \pm b \in W.$$

Prove that

1) If W is an invariant ellipsoid and $x(t) \in W$ for some t, then $x(t') \in W$ for all $t' \ge t$.

2) Assume that the vectors $b, Ab, A^2b, ..., A^{n-1}b$ are linearly independent. Prove that an invariant ellipsoid exists if and only if A is stable (the absolute values of all eigenvalues of A are < 1).

3) Assuming that A is stable, prove that an ellipsoid $\{x \mid x^T Z x \le 1\}$ $[Z \succ 0]$ is invariant if and only if there exists $\lambda \ge 0$ such that

$$\begin{pmatrix} 1 - b^T Z b - \lambda & -b^T Z A \\ -A^T Z b & \lambda Z - A^T Z A \end{pmatrix} \succeq 0.$$

How could one use this fact to approximate numerically the smallest volume invariant ellipsoid?

Greedy "infinitesimal" ellipsoidal approximations. Consider a linear time-varying controlled system

$$\frac{d}{dt}x(t) = A(t)x(t) + B(t)u(t) + v(t) \tag{3.7.11}$$

with continuous matrix-valued functions $A(t)$, $B(t)$, continuous vector-valued function $v(\cdot)$ and norm-bounded control:

$$\|u(\cdot)\|_2 \le 1. \tag{3.7.12}$$

Assume that the initial state of the system belongs to a given ellipsoid:

$$x(0) \in E(0) = \{x \mid (x - x^0)^T G^0 (x - x^0) \le 1\} \quad [G^0 = [G^0]^T \succ 0]. \tag{3.7.13}$$

Our goal is to build, in an "on-line" fashion, a system of ellipsoids

$$E(t) = \{x \mid (x - x_t)^T G_t (x - x_t) \le 1\} \quad [G_t = G_t^T \succ 0] \tag{3.7.14}$$

in such a way that if $u(\cdot)$ is a control satisfying (3.7.12) and $x(0)$ is an initial state satisfying (3.7.13), then for every $t \ge 0$ it holds

$$x(t) \in E(t).$$

We are interested to minimize the volumes of the resulting ellipsoids.

There is no difficulty with the path x_t of centers of the ellipsoids: it "obviously" should satisfy the requirements

$$\frac{d}{dt}x_t = A(t)x_t + v(t),\ t \ge 0; \quad x_0 = x^0. \tag{3.7.15}$$

Let us take this choice for granted and focus on how should we define the positive definite matrices G_t. Let us look for a continuously differentiable matrix-valued function G_t, taking values in the set of positive definite symmetric matrices, with the following property:

> **(L)** *For every $t \ge 0$ and every point $x^t \in E(t)$* (see (3.7.14)), *every trajectory $x(\tau)$, $\tau \ge t$, of the system*
>
> $$\frac{d}{d\tau}x(\tau) = A(\tau)x(\tau) + B(\tau)u(\tau) + v(\tau), \quad x(t) = x^t$$
>
> *with $\|u(\cdot)\|_2 \le 1$ satisfies $x(\tau) \in E(\tau)$ for all $\tau \ge t$.*

Note that **(L)** is a sufficient (but in general not necessary) condition for the system of ellipsoids $E(t)$, $t \ge 0$, "to cover" all trajectories of (3.7.11) – (3.7.12). Indeed, when formulating **(L)**, we act as if we were sure that the states $x(t)$ of our system run through the entire ellipsoid $E(t)$, which is not necessarily the case. The advantage of **(L)** is that this condition can be converted into an "infinitesimal" form:

Exercise 3.43 *Prove that if $G_t \succ 0$ is continuously differentiable and satisfies* **(L)**, *then*

$$\forall \left(t \ge 0, x, u : x^T G_t x = 1, u^T u \le 1\right): \quad 2u^T B^T(t) G_t x + x^T [\frac{d}{dt}G_t + A^T(t)G_t + G_t A(t)]x \le 0. \tag{3.7.16}$$

Vice versa, if G_t is a continuously differentiable function taking values in the set of positive definite symmetric matrices and satisfying (3.7.16) *and the initial condition $G_0 = G^0$, then the associated system of ellipsoids $\{E(t)\}$ satisfies* **(L)**.

The result of Exercise 3.43 provides us with a kind of description of the families of ellipsoids $\{E(t)\}$ we are interested in. Now let us take care of the volumes of these ellipsoids. The latter can be done via a "greedy" (locally optimal) policy: given $E(t)$, let us try to minimize, under restriction (3.7.16), the derivative of the volume of the ellipsoid at time t. Note that this locally optimal policy does not necessary yield the smallest volume ellipsoids satisfying **(L)** (achieving "instant reward" is not always the best way to happiness); nevertheless, this policy makes sense.

We have $2\ln \mathrm{vol}(E_t) = -\ln \mathrm{Det}(G_t)$, whence

$$2\frac{d}{dt}\ln \mathrm{vol}(E(t)) = -\mathrm{Tr}(G_t^{-1}\frac{d}{dt}G_t);$$

thus, our greedy policy requires to choose $H_t \equiv \frac{d}{dt}G_t$ as a solution to the optimization problem

$$\max_{H=H^T}\Big\{\mathrm{Tr}(G_t^{-1}H) : 2u^TB^T(t)G_tx + x^T[\tfrac{d}{dt}G_t - A^T(t)G_t - G_tA(t)]x \le 0 \\ \forall\,(x,u : x^TG_tx = 1, u^Tu \le 1)\Big\}.$$

Exercise 3.44 *Prove that the outlined greedy policy results in the solution G_t to the differential equation*

$$\tfrac{d}{dt}G_t = -A^T(t)G_t - G_tA(t) - \sqrt{\tfrac{n}{\mathrm{Tr}(G_tB(t)B^T(t))}}G_tB(t)B^T(t)G_t - \sqrt{\tfrac{\mathrm{Tr}(G_tB(t)B^T(t))}{n}}\,G_t,\ t \ge 0; \\ G_0 = G^0.$$

Prove that the solution to this equation is symmetric and positive definite for all $t > 0$, provided that $G^0 = [G^0]^T \succ 0$.

Exercise 3.45 *Modify the previous reasoning to demonstrate that the "locally optimal" policy for building <u>inner</u> ellipsoidal approximation of the set*

$$\begin{array}{rcl} X(t) & = & \big\{x(t) \mid \exists x^0 \in E(0) \equiv \{x \mid (x-x^0)^TG^0(x-x^0) \le 1\}, \exists u(\cdot), \|u(\cdot)\|_2 \le 1 : \\ & & \quad \tfrac{d}{d\tau}x(\tau) = A(\tau)x(\tau) + B(\tau)u(\tau) + v(\tau),\ 0 \le \tau \le t,\ x(0) = x^0\big\} \end{array}$$

results in the family of ellipsoids

$$\underline{E}(t) = \{x \mid (x-x_t)^TW_t(x-x_t) \le 1\},$$

where x_t is given by (3.7.15) *and W_t is the solution of the differential equation*

$$\frac{d}{dt}W_t = -A^T(t)W_t - W_tA(t) - 2W_t^{1/2}(W_t^{1/2}B(t)B^T(t)W_t^{1/2})^{1/2}W_t^{1/2},\ t \ge 0; \quad W_0 = G^0.$$

Lecture 4

Polynomial Time Interior Point algorithms for LP, CQP and SDP

4.1 Complexity of Convex Programming

When we attempt to solve any problem, we would like to know whether it is possible to find a correct solution in a "reasonable time". Had we known that the solution will not be reached in the next 30 years, we would think (at least) twice before starting to solve it. Of course, this in an extreme case, but undoubtedly, it is highly desirable to distinguish between "computationally tractable" problems – those that can be solved efficiently – and problems which are "computationally intractable". The corresponding *complexity theory* was first developed in Computer Science for combinatorial (discrete) problems, and later somehow extended onto the case of continuous computational problems, including those of Continuous Optimization. In this section, we outline the main concepts of the CCT – Combinatorial Complexity Theory – along with their adaptations to Continuous Optimization.

4.1.1 Combinatorial Complexity Theory

A generic combinatorial problem is a family $\mathcal{P}$ of *problem instances* of a "given structure", each instance $(p) \in \mathcal{P}$ being identified by a finite-dimensional *data vector* Data(p), specifying the particular values of the coefficients of "generic" analytic expressions. The data vectors are assumed to be *Boolean vectors* – with entries taking values $0, 1$ only, so that the data vectors are, actually, finite binary words.

The model of computations in CCT: an idealized computer capable to store only integers (i.e., finite binary words), and its operations are *bitwise*: we are allowed to multiply, add and compare integers. To add and to compare two ℓ-bit integers, it takes $O(\ell)$ "bitwise" elementary operations, and to multiply a pair of ℓ-bit integers it costs $O(\ell^2)$ elementary operations (the cost of multiplication can be reduced to $O(\ell \ln(\ell))$, but it does not matter) .

In CCT, a solution to an instance (p) of a generic problem $\mathcal{P}$ is a finite binary word y such that the pair (Data(p),y) satisfies certain "verifiable condition" $\mathcal{A}(\cdot, \cdot)$. Namely, it is assumed that there exists a code $\mathcal{M}$ for the above "Integer Arithmetic computer" such that executing the code on every input pair x, y of finite binary words, the computer after finitely many elementary operations terminates and outputs either "yes", if $\mathcal{A}(x, y)$ is satisfied, or "no", if $\mathcal{A}(x, y)$ is not satisfied. Thus, $\mathcal{P}$ is the problem

Given x, find y such that

$$\mathcal{A}(x, y) = \texttt{true}, \tag{4.1.1}$$

or detect that no such y exists.

For example, the problem *Stones*:

Given n positive integers $a_1, ..., a_n$, find a vector $x = (x_1, ..., x_n)^T$ with coordinates ± 1 such that $\sum_i x_i a_i = 0$, or detect that no such vector exists

is a generic combinatorial problem. Indeed, the data of the instance of the problem, same as candidate solutions to the instance, can be naturally encoded by finite sequences of integers. In turn, finite sequences of integers can be easily encoded by finite binary words. And, of course, for this problem you can easily point out a code for the "Integer Arithmetic computer" which, given on input two binary words $x = \mathrm{Data}(p)$, y encoding the data vector of an instance (p) of the problem and a candidate solution, respectively, verifies in finitely many "bit" operations whether y represents or does not represent a solution to (p).

A solution algorithm for a generic problem $\mathcal{P}$ is a code $\mathcal{S}$ for the Integer Arithmetic computer which, given on input the data vector $\mathrm{Data}(p)$ of an instance $(p) \in \mathcal{P}$, after finitely many operations either returns a solution to the instance, or a (correct!) claim that no solution exists. The running time $T_{\mathcal{S}}(p)$ of the solution algorithm on instance (p) is exactly the number of elementary (i.e., bit) operations performed in course of executing $\mathcal{S}$ on $\mathrm{Data}(p)$.

A solvability test for a generic problem $\mathcal{P}$ is defined similarly to a solution algorithm, but now all we want of the code is to say (correctly!) whether the input instance is or is not solvable, just "yes" or "no", without constructing a solution in the case of the "yes" answer.

The complexity of a solution algorithm/solvability test $\mathcal{S}$ is defined as

$$\mathrm{Compl}_{\mathcal{S}}(\ell) = \max\{T_{\mathcal{S}}(p) \mid (p) \in \mathcal{P}, \mathrm{length}(\mathrm{Data}(p)) \le \ell\},$$

where $\mathrm{length}(x)$ is the bit length (i.e., number of bits) of a finite binary word x. The algorithm/test is called *polynomial time*, if its complexity is bounded from above by a polynomial of ℓ.

Finally, a generic problem $\mathcal{P}$ is called to be *polynomially solvable*, if it admits a polynomial time solution algorithm. If $\mathcal{P}$ admits a polynomial time solvability test, we say that $\mathcal{P}$ is *polynomially verifiable*.

Classes P and NP. A generic problem $\mathcal{P}$ is said to belong to the class NP, if the corresponding condition $\mathcal{A}$, see (4.1.1), possesses the following two properties:

I. $\mathcal{A}$ is polynomially computable, i.e., the running time $T(x, y)$ (measured, of course, in elementary "bit" operations) of the associated code $\mathcal{M}$ is bounded from above by a polynomial of the bit length $\mathrm{length}(x) + \mathrm{length}(y)$ of the input:

$$T(x, y) \le \chi(\mathrm{length}(x) + \mathrm{length}(y))^{\chi} \quad \forall (x, y)\ ^{1)}$$

Thus, the first property of an NP problem states that *given the data Data(p) of a problem instance p and a candidate solution y, it is easy to check whether y is an actual solution of* (p) – to verify this fact, it suffices to compute $A(\mathrm{Data}(p), y)$, and this computation requires polynomial in $\mathrm{length}(\mathrm{Data}(p)) + \mathrm{length}(y)$ time.
The second property of an NP problem makes its instances even more easier:

II. A solution to an instance (p) of a problem cannot be "too long" as compared to the data of the instance: there exists χ such that

$$\mathrm{length}(y) > \chi \mathrm{length}^{\chi}(x) \Rightarrow \mathcal{A}(x, y) = \text{"}\texttt{no}\text{"}.$$

[1)] Here and in what follows, we denote by χ positive "characteristic constants" associated with the predicates/problems in question. The particular values of these constants are of no importance, the only thing that matters is their existence. Note that in different places of even the same equation χ may have different values.

A generic problem $\mathcal{P}$ is said to belong to the class P, if it belongs to the class NP and is polynomially solvable.

NP-completeness is defined as follows:

Definition 4.1.1 (i) Let $\mathcal{P}$, $\mathcal{Q}$ be two problems from NP. Problem $\mathcal{Q}$ is called to be *polynomially reducible to* $\mathcal{P}$, if there exists a polynomial time algorithm $\mathcal{M}$ (i.e., a code for the Integer Arithmetic computer with the running time bounded by a polynomial of the length of the input) with the following property. Given on input the data vector Data(q) of an instance $(q) \in \mathcal{Q}$, $\mathcal{M}$ converts this data vector to the data vector Data($p[q]$) of an instance of $\mathcal{P}$ such that $(p[q])$ is solvable if and only if (q) is solvable.

(ii) A generic problem $\mathcal{P}$ from NP is called NP-*complete*, if every other problem $\mathcal{Q}$ from NP is polynomially reducible to $\mathcal{P}$.

The importance of the notion of an NP-complete problem comes from the following fact:

> *If a particular* NP-*complete problem is polynomially verifiable (i.e., admits a polynomial time solvability test), then* every *problem from NP is polynomially solvable:* P = NP.

The question whether P=NP – whether NP-complete problems are or are not polynomially solvable, is qualified as "the most important open problem in Theoretical Computer Science" and remains open for about 30 years. One of the most basic results of Theoretical Computer Science is that NP-*complete problems do exist* (the Stones problem is an example). Many of these problems are of huge practical importance, and are therefore subject, over decades, of intensive studies of thousands excellent researchers. However, no polynomial time algorithm for any of these problems was found. Given the huge total effort invested in this research, we should conclude that it is "highly improbable" that NP-complete problems are polynomially solvable. Thus, at the "practical level" the fact that certain problem is NP-complete is sufficient to qualify the problem as "computationally intractable", at least at our present level of knowledge.

4.1.2 Complexity in Continuous Optimization

It is convenient to represent continuous optimization problems as *Mathematical Programming problems*, i.e. programs of the following form:

$$\min_x \left\{ p_0(x) : x \in X(p) \subset \mathbf{R}^{n(p)} \right\} \qquad (p)$$

where

- $n(p)$ is the *design dimension* of program (p);
- $X(p) \subset \mathbf{R}^n$ is the *feasible domain* of the program;
- $p_0(x) : \mathbf{R}^n \to \mathbf{R}$ is the *objective* of (p).

Families of optimization programs. We want to speak about methods for solving optimization programs (p) "of a given structure" (for example, Linear Programming ones). All programs (p) "of a given structure", like in the combinatorial case, form certain family $\mathcal{P}$, and we assume that every particular program in this family – every *instance* (p) of $\mathcal{P}$ – is specified by its particular *data* Data(p). However, now the data is a finite-dimensional *real* vector; one may think about the entries of this data vector as about particular values of coefficients of "generic" (specific for $\mathcal{P}$) analytic expressions for $p_0(x)$ and $X(p)$. The dimension of the vector Data(p) will be called the *size* of the instance:

$$\mathrm{Size}(p) = \dim \mathrm{Data}(p).$$

The model of computations. This is what is known as "Real Arithmetic Model of Computations", as opposed to "Integer Arithmetic Model" in the CCT. We assume that the computations are carried out by an idealized version of the usual computer which is capable to store countably many reals and can perform with them the standard *exact* real arithmetic operations – the four basic arithmetic operations, evaluating elementary functions, like cos and exp, and making comparisons.

Accuracy of approximate solutions. We assume that a generic optimization problem $\mathcal{P}$ is equipped with an "infeasibility measure" $\mathrm{Infeas}_{\mathcal{P}}(x,p)$ – a real-valued function of $p \in \mathcal{P}$ and $x \in \mathbf{R}^{n(p)}$ which quantifies the infeasibility of vector x as a candidate solution to (p). In our general considerations, all we require from this measure is that

- $\mathrm{Infeas}_{\mathcal{P}}(x,p) \geq 0$, and $\mathrm{Infeas}_{\mathcal{P}}(x,p) = 0$ when x is feasible for (p) (i.e., when $x \in X(p)$).

Given an infeasibility measure, we can proceed to define the notion of an ϵ-solution to an instance $(p) \in \mathcal{P}$, namely, as follows. Let

$$\mathrm{Opt}(p) \in \{-\infty\} \cup \mathbf{R} \cup \{+\infty\}$$

be the optimal value of the instance (i.e., the infimum of the values of the objective on the feasible set, if the instance is feasible, and $+\infty$ otherwise). A point $x \in \mathbf{R}^{n(p)}$ is called *an ϵ-solution* to (p), if

$$\mathrm{Infeas}_{\mathcal{P}}(x,p) \leq \epsilon \text{ and } p_0(x) - \mathrm{Opt}(p) \leq \epsilon,$$

i.e., if x is both "ϵ-feasible" and "ϵ-optimal" for the instance.

It is convenient to define the *number of accuracy digits* in an ϵ-solution to (p) as the quantity

$$\mathrm{Digits}(p,\epsilon) = \ln\left(\frac{\mathrm{Size}(p) + \|\mathrm{Data}(p)\|_1 + \epsilon^2}{\epsilon}\right).$$

Solution methods. A solution method $\mathcal{M}$ for a given family $\mathcal{P}$ of optimization programs is a code for the idealized Real Arithmetic computer. When solving an instance $(p) \in \mathcal{P}$, the computer first inputs the data vector $\mathrm{Data}(p)$ of the instance and a real $\epsilon > 0$ – the accuracy to which the instance should be solved, and then executes the code $\mathcal{M}$ on this input. We assume that the execution, on every input $(\mathrm{Data}(p), \epsilon > 0)$ with $(p) \in \mathcal{P}$, takes finitely many elementary operations of the computer, let this number be denoted by $\mathrm{Compl}_{\mathcal{M}}(p,\epsilon)$, and results in one of the following three possible outputs:

– an $n(p)$-dimensional vector $\mathrm{Res}_{\mathcal{M}}(p,\epsilon)$ which is an ϵ-solution to (p),
– a correct message "(p) is infeasible",
– a correct message "(p) is unbounded below".

We measure the efficiency of a method by its *running time* $\mathrm{Compl}_{\mathcal{M}}(p,\epsilon)$ – the number of elementary operations performed by the method when solving instance (p) within accuracy ϵ. By definition, the fact that $\mathcal{M}$ is "efficient" (*polynomial time*) on $\mathcal{P}$, means that there exists a *polynomial* $\pi(s,\tau)$ such that

$$\begin{gathered}\mathrm{Compl}_{\mathcal{M}}(p,\epsilon) \leq \pi\left(\mathrm{Size}(p), \mathrm{Digits}(p,\epsilon)\right) \\ \forall (p) \in \mathcal{P}\ \forall \epsilon > 0.\end{gathered} \tag{4.1.2}$$

Informally speaking, polynomiality of $\mathcal{M}$ means that when we increase the size of an instance and the required number of accuracy digits by absolute constant factors, the running time increases by no more than another absolute constant factor.

We call a family $\mathcal{P}$ of optimization problems *polynomially solvable* (or, which is the same, *computationally tractable*), if it admits a polynomial time solution method.

4.1.3 Computational tractability of convex optimization problems

A generic optimization problem $\mathcal{P}$ is called *convex*, if, for every instance $(p) \in \mathcal{P}$, both the objective $p_0(x)$ of the instance and the infeasibility measure $\mathrm{Infeas}_{\mathcal{P}}(x,p)$ are convex functions of $x \in \mathbf{R}^{n(p)}$. One of the major complexity results in Continuous Optimization is that *a generic convex optimization problem, under mild computability and regularity assumptions, is polynomially solvable* (and thus "computationally tractable"). To formulate the precise result, we start with specifying the aforementioned "mild assumptions".

Polynomial computability. Let $\mathcal{P}$ be a generic convex program, and let $\mathrm{Infeas}_{\mathcal{P}}(x,p)$ be the corresponding measure of infeasibility of candidate solutions. We say that our family is *polynomially computable*, if there exist two codes $\mathcal{C}_{\rm obj}$, $\mathcal{C}_{\rm cons}$ for the Real Arithmetic computer such that

1. For every instance $(p)\in\mathcal{P}$, the computer, when given on input the data vector of the instance (p) and a point $x\in\mathbf{R}^{n(p)}$ and executing the code $\mathcal{C}_{\rm obj}$, outputs the value $p_0(x)$ and a subgradient $e(x)\in\partial p_0(x)$ of the objective p_0 of the instance at the point x, and the running time (i.e., total number of operations) of this computation $T_{\rm obj}(x,p)$ is bounded from above by a polynomial of the size of the instance:

$$\forall\left((p)\in\mathcal{P},x\in\mathbf{R}^{n(p)}\right):\quad T_{\rm obj}(x,p)\le\chi\mathrm{Size}^{\chi}(p)\quad[\mathrm{Size}(p)=\dim\mathrm{Data}(p)]. \tag{4.1.3}$$

(recall that in our notation, χ is a common name of characteristic constants associated with $\mathcal{P}$).

2. For every instance $(p)\in\mathcal{P}$, the computer, when given on input the data vector of the instance (p), a point $x\in\mathbf{R}^{n(p)}$ and an $\epsilon>0$ and executing the code $\mathcal{C}_{\rm cons}$, reports on output whether $\mathrm{Infeas}_{\mathcal{P}}(x,p)\le\epsilon$, and if it is not the case, outputs a linear form a which separates the point x from all those points y where $\mathrm{Infeas}_{\mathcal{P}}(y,p)\le\epsilon$:

$$\forall\left(y,\mathrm{Infeas}_{\mathcal{P}}(y,p)\le\epsilon\right):\qquad a^Tx>a^Ty, \tag{4.1.4}$$

the running time $T_{\rm cons}(x,\epsilon,p)$ of the computation being bounded by a polynomial of the size of the instance and of the "number of accuracy digits":

$$\forall\left((p)\in\mathcal{P},x\in\mathbf{R}^{n(p)},\epsilon>0\right):\quad T_{\rm cons}(x,\epsilon,p)\le\chi\left(\mathrm{Size}(p)+\mathrm{Digits}(p,\epsilon)\right)^{\chi}. \tag{4.1.5}$$

Note that the vector a in (4.1.4) is not supposed to be nonzero; when it is 0, (4.1.4) simply says that there are no points y with $\mathrm{Infeas}_{\mathcal{P}}(y,p)\le\epsilon$.

Polynomial growth. We say that a generic convex program $\mathcal{P}$ is with *polynomial growth*, if the objectives and the infeasibility measures, as functions of x, grow polynomially with $\|x\|_1$, the degree of the polynomial being a power of $\mathrm{Size}(p)$:

$$\begin{array}{l}\forall\left((p)\in\mathcal{P},x\in\mathbf{R}^{n(p)}\right):\\ |p_0(x)|+\mathrm{Infeas}_{\mathcal{P}}(x,p)\le\left(\chi\left[\mathrm{Size}(p)+\|x\|_1+\|\mathrm{Data}(p)\|_1\right]\right)^{\left(\chi\mathrm{Size}^{\chi}(p)\right)}.\end{array} \tag{4.1.6}$$

Polynomial boundedness of feasible sets. We say that a generic convex program $\mathcal{P}$ has *polynomially bounded feasible sets*, if the feasible set $X(p)$ of every instance $(p)\in\mathcal{P}$ is bounded, and is contained in the Euclidean ball, centered at the origin, of "not too large" radius:

$$\begin{array}{l}\forall(p)\in\mathcal{P}:\\ X(p)\subset\left\{x\in\mathbf{R}^{n(p)}:\|x\|_2\le\left(\chi\left[\mathrm{Size}(p)+\|\mathrm{Data}(p)\|_1\right]\right)^{\chi\mathrm{Size}^{\chi}(p)}\right\}.\end{array} \tag{4.1.7}$$

Example. Consider generic optimization problems $\mathcal{LP}_b$, $\mathcal{CQP}_b$, $\mathcal{SDP}_b$ with instances in the conic form

$$\min_{x\in\mathbf{R}^{n(p)}}\left\{p_0(x)\equiv c_{(p)}^Tx:x\in X(p)\equiv\{x:A_{(p)}x-b_{(p)}\in\mathbf{K}(p),\|x\|_2\le R\}\right\}; \tag{4.1.8}$$

where $\mathbf{K}$ is a cone belonging to a characteristic for the generic program family $\mathcal{K}$ of cones, specifically,

- the family of nonnegative orthants for $\mathcal{LP}_b$,
- the family of *direct products* of Lorentz cones for $\mathcal{CQP}_b$,
- the family of semidefinite cones for $\mathcal{SDP}_b$.

The *data* of and instance (p) of the type (4.1.8) is the collection

$$\mathrm{Data}(p)=(n(p),c_{(p)},A_{(p)},b_{(p)},R,\langle\text{size(s) of }\mathbf{K}_{(p)}\rangle),$$

with naturally defined size(s) of a cone $\mathbf{K}$ from the family $\mathcal{K}$ associated with the generic program under consideration: the sizes of $\mathbf{R}^n_+$ and of $\mathbf{S}^n_+$ equal n, and the size of a direct product of Lorentz cones is the sequence of the dimensions of the factors.

The generic conic programs in question are equipped with the infeasibility measure

$$\mathrm{Infeas}(x,p) = \min\left\{t \geq 0 : t\mathbf{e}[\mathbf{K}_{(p)}] + A_{(p)}x - b_{(p)} \in \mathbf{K}_{(p)}\right\}, \tag{4.1.9}$$

where $\mathbf{e}[\mathbf{K}]$ is a naturally defined "central point" of $\mathbf{K} \in \mathcal{K}$, specifically,

- the n-dimensional vector of ones when $\mathbf{K} = \mathbf{R}^n_+$,
- the vector $\mathbf{e}_m = (0, ..., 0, 1)^T \in \mathbf{R}^m$ when $\mathbf{K}_{(p)}$ is the Lorentz cone $\mathbf{L}^m$, and the direct sum of these vectors, when $\mathbf{K}$ is a direct product of Lorentz cones,
- the unit matrix of appropriate size when $\mathbf{K}$ is a semidefinite cone.

In the sequel, we refer to the three generic problems we have just defined as to *Linear, Conic Quadratic and Semidefinite Programming problems with ball constraints*, respectively. It is immediately seen that the generic programs $\mathcal{LP}_b$, $\mathcal{CQP}_b$ and $\mathcal{SDP}_b$ are convex and possess the properties of polynomial computability, polynomial growth and polynomially bounded feasible sets (the latter property is ensured by making the ball constraint $\|x\|_2 \leq R$ a part of program's formulation).

Computational Tractability of Convex Programming. The role of the properties we have introduced becomes clear from the following result:

Theorem 4.1.1 *Let $\mathcal{P}$ be a family of convex optimization programs equipped with infeasibility measure $\mathrm{Infeas}_{\mathcal{P}}(\cdot,\cdot)$. Assume that the family is polynomially computable, with polynomial growth and with polynomially bounded feasible sets. Then $\mathcal{P}$ is polynomially solvable.*

In particular, the generic Linear, Conic Quadratic and Semidefinite programs with ball constraints $\mathcal{LP}_b$, $\mathcal{CQP}_b$, $\mathcal{SDP}_b$ are polynomially solvable.

4.1.4 "What is inside" Theorem 4.1.1: Black-box represented convex programs and the Ellipsoid method

Theorem 4.1.1 is a more or less straightforward corollary of a result related to the so called *Information-Based complexity of black-box represented convex programs*. This result is interesting by its own right, this is why we reproduce it here:

Consider a Convex Programming program

$$\min_x \{f(x) : x \in X\} \tag{4.1.10}$$

where

- X is a convex compact set in $\mathbf{R}^n$ with a nonempty interior
- f is a continuous convex function on X.

Assume that our "environment" when solving (4.1.10) is as follows:

1. We have access to a *Separation Oracle* $\mathrm{Sep}(X)$ for X – a routine which, given on input a point $x \in \mathbf{R}^n$, reports on output whether or not $x \in \mathrm{int}\, X$, and in the case of $x \notin \mathrm{int}\, X$, returns a *separator* – a nonzero vector e such that

$$e^T x \geq \max_{y \in X} e^T y \tag{4.1.11}$$

(the existence of such a separator is guaranteed by the Separation Theorem for convex sets);

2. We have access to a *First Order* oracle which, given on input a point $x \in \operatorname{int} X$, returns the value $f(x)$ and a subgradient $f'(x)$ of f at x (Recall that a subgradient $f'(x)$ of f at x is a vector such that
$$f(y) \geq f(x) + (y-x)^T f'(x) \tag{4.1.12}$$
for all y; convex function possesses subgradients at every relative interior point of its domain, see Section C.6.2.);
3. We are given two positive reals $R \geq r$ such that X is contained in the Euclidean ball, centered at the origin, of the radius R and contains a Euclidean ball of the radius r (not necessarily centered at the origin).

The result we are interested in is as follows:

Theorem 4.1.2 *In the outlined "working environment", for every given $\epsilon > 0$ it is possible to find an ϵ-solution to* (4.1.10)*, i.e., a point $x_\epsilon \in X$ with*
$$f(x_\epsilon) \leq \min_{x \in X} f(x) + \epsilon$$
in no more than $N(\epsilon)$ subsequent calls to the Separation and the First Order oracles plus no more than $O(1)n^2 N(\epsilon)$ arithmetic operations to process the answers of the oracles, with
$$N(\epsilon) = O(1)n^2 \ln\left(2 + \frac{\operatorname{Var}_X(f) R}{\epsilon \cdot r}\right). \tag{4.1.13}$$
Here
$$\operatorname{Var}_X(f) = \max_X f - \min_X f.$$

Proof of Theorem 4.1.2: the Ellipsoid method

Assume that we are interested to solve the convex program (4.1.10) and we have an access to a separation oracle $\mathrm{Sep}(X)$ for the feasible domain of (4.1.10) and to a first order oracle $\mathcal{O}(f)$ for the objective of (4.1.10). How could we solve the problem via these "tools"? An extremely transparent way is given by the *Ellipsoid method* which can be viewed as a multi-dimensional extension of the usual bisection.

Ellipsoid method: the idea. Assume that we have already found an n-dimensional ellipsoid
$$E = \{x = c + Bu \mid u^T u \leq 1\} \qquad [B \in \mathbf{M}^{n,n}, \mathrm{Det} B \neq 0]$$
which contains the optimal set X_* of (4.1.10) (note that $X_* \neq \emptyset$, since the feasible set X of (4.1.10) is assumed to be compact, and the objective f – to be convex on the entire $\mathbf{R}^n$ and therefore continuous, see Theorem C.4.1). How could we construct a smaller ellipsoid containing X_* ?

The answer is immediate.

1) Let us call the separation oracle $\mathrm{Sep}(X)$, the center c of the current ellipsoid being the input. There are two possible cases:

1.a) $\mathrm{Sep}(X)$ reports that $c \notin X$ and returns a separator a:
$$a \neq 0, \;\; a^T c \geq \sup_{y \in X} a^T y. \tag{4.1.14}$$
In this case we can replace our current "localizer" E of the optimal set X_* by a smaller one – namely, by the "half-ellipsoid"
$$\widehat{E} = \{x \in E \mid a^T x \leq a^T c\}.$$
Indeed, by assumption $X_* \subset E$; when passing from E to $\widehat{E}$, we cut off all points x of E where $a^T x > a^T c$, and by (4.1.14) all these points are outside of X and therefore outside of $X_* \subset X$. Thus, $X_* \subset \widehat{E}$.

1.b) $\mathrm{Sep}(X)$ reports that $c \in X$. In this case we call the first order oracle $\mathcal{O}(f)$, c being the input; the oracle returns the value $f(c)$ and a subgradient $a = f'(c)$ of f at c. Again, two cases are possible:

1.b.1) $a = 0$. In this case we are done – c is a minimizer of f on X. Indeed, $c \in X$, and (4.1.12) reads

$$f(y) \geq f(c) + 0^T(y - c) = f(c) \quad \forall y \in \mathbf{R}^n.$$

Thus, c is a minimizer of f on $\mathbf{R}^n$, and since $c \in X$, c minimizes f on X as well.

1.b.2) $a \neq 0$. In this case (4.1.12) reads

$$a^T(x - c) > 0 \Rightarrow f(x) > f(c),$$

so that replacing the ellipsoid E with the half-ellipsoid

$$\widehat{E} = \{x \in E \mid a^T x \leq a^T c\}$$

we ensure the inclusion $X_* \subset \widehat{E}$. Indeed, $X_* \subset E$ by assumption, and when passing from E to $\widehat{E}$, we cut off all points of E where $a^T x > a^T c$ and, consequently, where $f(x) > f(c)$; since $c \in X$, no one of these points can belong to the set X_* of minimizers of f on X.

2) We have seen that as a result of operations described in 1.a-b) we either terminate with an exact minimizer of f on X, or obtain a "half-ellipsoid"

$$\widehat{E} = \{x \in E \mid a^T x \leq a^T c\} \qquad [a \neq 0]$$

containing X_*. It remains to use the following simple geometric fact:

(*) *Let $E = \{x = c + Bu \mid u^T u \leq 1\}$ ($\mathrm{Det}B \neq 0$) be an n-dimensional ellipsoid and $\widehat{E} = \{x \in E \mid a^T x \leq a^T c\}$ ($a \neq 0$) be a "half" of E. If $n > 1$, then $\widehat{E}$ is contained in the ellipsoid*

$$\begin{array}{c} E^+ = \{x = c^+ + B^+ u \mid u^T u \leq 1\}, \\ c^+ = c - \frac{1}{n+1} Bp, \\ B^+ = B\left(\frac{n}{\sqrt{n^2-1}}(I_n - pp^T) + \frac{n}{n+1} pp^T\right) = \frac{n}{\sqrt{n^2-1}} B + \left(\frac{n}{n+1} - \frac{n}{\sqrt{n^2-1}}\right)(Bp)p^T, \\ p = \frac{B^T a}{\sqrt{a^T B B^T a}} \end{array} \qquad (4.1.15)$$

and if $n = 1$, then the set $\widehat{E}$ is contained in the ellipsoid (which now is just a segment)

$$\begin{array}{c} E^+ = \{x \mid c^+ B^+ u \mid |u| \leq 1\}, \\ c^+ = c - \frac{1}{2} \frac{Ba}{|Ba|}, \\ B_+ = \frac{1}{2} B. \end{array}$$

In all cases, the n-dimensional volume $\mathrm{Vol}(E^+)$ of the ellipsoid E^+ is less than the one of E:

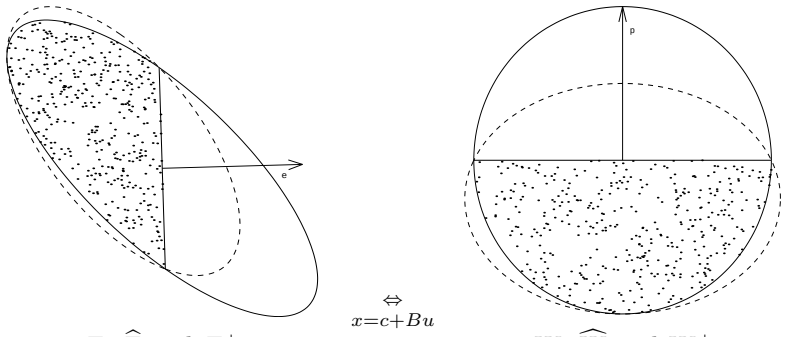
$$\mathrm{Vol}(E^+) = \left(\frac{n}{\sqrt{n^2-1}}\right)^{n-1} \frac{n}{n+1} \mathrm{Vol}(E) \leq \exp\{-1/(2n)\}\mathrm{Vol}(E) \qquad (4.1.16)$$

(in the case of $n = 1$, $\left(\frac{n}{\sqrt{n^2-1}}\right)^{n-1} = 1$).

(*) says that there exists (and can be explicitly specified) an ellipsoid $E^+ \supset \widehat{E}$ with the volume constant times less than the one of E. Since E^+ covers $\widehat{E}$, and the latter set, as we have seen, covers X_*, E^+ covers X_*. Now we can iterate the above construction, thus obtaining a sequence of ellipsoids $E, E^+, (E^+)^+, ...$ with volumes going to 0 at a linear rate (depending on the dimension n only) which "collapses" to the set X_* of optimal solutions of our problem – exactly as in the usual bisection!

Note that (*) is just an exercise in elementary calculus. Indeed, the ellipsoid E is given as an image of the unit Euclidean ball $W = \{u \mid u^T u \leq 1\}$ under the one-to-one affine mapping $u \mapsto c + Bu$; the half-ellipsoid $\widehat{E}$ is then the image, under the same mapping, of the half-ball

$$\widehat{W} = \{u \in W \mid p^T u \leq 0\}$$



E, $\widehat{E}$ and E^+ W, $\widehat{W}$ and W^+

Figure 5.1.

p being the unit vector from (4.1.15); indeed, if $x = c + Bu$, then $a^T x \le a^T c$ if and only if $a^T Bu \le 0$, or, which is the same, if and only if $p^T u \le 0$. Now, instead of covering $\widehat{E}$ by a small in volume ellipsoid E^+, we may cover by a small ellipsoid W^+ the half-ball $\widehat{W}$ and then take E^+ to be the image of W^+ under our affine mapping:

$$E^+ = \{x = c + Bu \mid u \in W^+\}. \tag{4.1.17}$$

Indeed, if W^+ contains $\widehat{W}$, then the image of W^+ under our affine mapping $u \mapsto c + Bu$ contains the image of $\widehat{W}$, i.e., contains $\widehat{E}$. And since the ratio of volumes of two bodies remain invariant under affine mapping (passing from a body to its image under an affine mapping $u \mapsto c + Bu$, we just multiply the volume by $|\mathrm{Det}B|$), we have

$$\frac{\mathrm{Vol}(E^+)}{\mathrm{Vol}(E)} = \frac{\mathrm{Vol}(W^+)}{\mathrm{Vol}(W)}.$$

Thus, the problem of finding a "small" ellipsoid E^+ containing the half-ellipsoid $\widehat{E}$ can be reduced to the one of finding a "small" ellipsoid W^+ containing the half-ball $\widehat{W}$, as shown on Fig. 5.1. Now, the problem of finding a small ellipsoid containing $\widehat{W}$ is very simple: our "geometric data" are invariant with respect to rotations around the p-axis, so that we may look for W^+ possessing the same rotational symmetry. Such an ellipsoid W^+ is given by just 3 parameters: its center should belong to our symmetry axis, i.e., should be $-hp$ for certain h, one of the half-axes of the ellipsoid (let its length be μ) should be directed along p, and the remaining $n-1$ half-axes should be of the same length λ and be orthogonal to p. For our 3 parameters h, μ, λ we have 2 equations expressing the fact that the boundary of W^+ should pass through the "South pole" $-p$ of W and trough the "equator" $\{u \mid u^T u = 1, p^T u = 0\}$; indeed, W^+ should contain $\widehat{W}$ and thus – both the pole and the equator, and since we are looking for W^+ with the smallest possible volume, both the pole and the equator should be on the boundary of W^+. Using our 2 equations to express μ and λ via h, we end up with a single "free" parameter h, and the volume of W^+ (i.e., $\mathrm{const}(n)\mu\lambda^{n-1}$) becomes an explicit function of h; minimizing this function in h, we find the "optimal" ellipsoid W^+, check that it indeed contains $\widehat{W}$ (i.e., that our geometric intuition was correct) and then convert W^+ into E^+ according to (4.1.17), thus coming to the explicit formulas (4.1.15) – (4.1.16); implementation of the outlined scheme takes from 10 to 30 minutes, depending on how many miscalculations are made...

It should be mentioned that although the indicated scheme is quite straightforward and elementary, the fact that it works is not evident a priori: it might happen that the smallest volume ellipsoid containing a half-ball is just the original ball! This would be the death of our idea – instead of a sequence of ellipsoids collapsing to the solution set X_*, we would get a "stationary" sequence $E, E, E...$ Fortunately, it is not happening, and this is a great favour Nature does to Convex Optimization...

Ellipsoid method: the construction. There is a small problem with implementing our idea of "trapping" the optimal set X_* of (4.1.10) by a "collapsing" sequence of ellipsoids. The only thing we can ensure is that all our ellipsoids contain X_* and that their *volumes* rapidly (at a linear rate) converge to 0. However, the linear sizes of the ellipsoids should not necessarily go to 0 – it may happen that the ellipsoids are shrinking in some directions and are not shrinking (or even become larger) in other directions (look what happens if we apply the construction to minimizing a function of 2 variables which in fact depends only on the first coordinate). Thus, to the moment it is unclear how to build a sequence of points converging to X_*. This difficulty, however, can be easily resolved: as we shall see, we can form this sequence from the best feasible solutions generated so far. Another issue which remains open to the moment is when to terminate the method; as we shall see in a while, this issue also can be settled satisfactory.

The precise description of the Ellipsoid method as applied to (4.1.10) is as follows (in this description, we assume that $n \geq 2$, which of course does not restrict generality):

The Ellipsoid Method.

Initialization. Recall that when formulating (4.1.10) it was assumed that the feasible set X of the problem is contained in the ball $E_0 = \{x \mid \|x\|_2 \leq R\}$ of a given radius R and contains an (unknown) Euclidean ball of a known radius $r > 0$. The ball E_0 will be our initial ellipsoid; thus, we set

$$c_0 = 0,\ B_0 = RI,\ E_0 = \{x = c_0 + B_0 u \mid u^T u \leq 1\};$$

note that $E_0 \supset X$.

We also set

$$\rho_0 = R,\ L_0 = 0.$$

The quantities ρ_t will be the "radii" of the ellipsoids E_t to be built, i.e., the radii of the Euclidean balls of the same volumes as E_t's. The quantities L_t will be our guesses for the variation

$$\mathrm{Var}_R(f) = \max_{x \in E_0} f(x) - \min_{x \in E_0} f(x)$$

of the objective on the initial ellipsoid E_0. We shall use these guesses in the termination test.

Finally, we input the accuracy $\epsilon > 0$ to which we want to solve the problem.

Step t, $t = 1, 2, ...$ At the beginning of step t, we have the previous ellipsoid

$$E_{t-1} = \{x = c_{t-1} + B_{t-1} u \mid u^T u \leq 1\} \qquad [c_{t-1} \in \mathbf{R}^n, B_{t-1} \in \mathbf{M}^{n,n}, \mathrm{Det} B_{t-1} \neq 0]$$

(i.e., have c_{t-1}, B_{t-1}) along with the quantities $L_{t-1} \geq 0$ and

$$\rho_{t-1} = |\mathrm{Det} B_{t-1}|^{1/n}.$$

At step t, we act as follows (cf. the preliminary description of the method):

1) We call the separation oracle $\mathrm{Sep}(X)$, c_{t-1} being the input. It is possible that the oracle reports that $c_{t-1} \notin X$ and provides us with a separator

$$a \neq 0: \quad a^T c_{t-1} \geq \sup_{y \in X} a^T y.$$

In this case we call step t *non-productive*, set

$$a_t = a,\ L_t = L_{t-1}$$

and go to rule 3) below. Otherwise – i.e., when $c_{t-1} \in X$ – we call step t *productive* and go to rule 2).

2) We call the first order oracle $\mathcal{O}(f)$, c_{t-1} being the input, and get the value $f(c_{t-1})$ and a subgradient $a \equiv f'(c_{t-1})$ of f at the point c_{t-1}. It is possible that $a = 0$; in this case

we terminate and claim that c_{t-1} is an optimal solution to (4.1.10). In the case of $a \neq 0$ we set

$$a_t = a,$$

compute the quantity

$$\ell_t = \max_{y \in E_0} [a_t^T y - a_t^T c_{t-1}] = R\|a_t\|_2 - a_t^T c_{t-1},$$

update L by setting

$$L_t = \max\{L_{t-1}, \ell_t\}$$

and go to rule 3).

3) We set

$$\widehat{E}_t = \{x \in E_{t-1} \mid a_t^T x \leq a_t^T c_{t-1}\}$$

(cf. the definition of $\widehat{E}$ in our preliminary description of the method) and define the new ellipsoid

$$E_t = \{x = c_t + B_t u \mid u^T u \leq 1\}$$

by setting (see (4.1.15))

$$\begin{array}{rcl} p_t & = & \frac{B_{t-1}^T a_t}{\sqrt{a_t^T B_{t-1} B_{t-1}^T a_t}} \\ c_t & = & c_{t-1} - \frac{1}{n+1} B_{t-1} p_t, \\ B_t & = & \frac{n}{\sqrt{n^2-1}} B_{t-1} + \left(\frac{n}{n+1} - \frac{n}{\sqrt{n^2-1}}\right)(B_{t-1}p_t)p_t^T. \end{array} \tag{4.1.18}$$

We also set

$$\rho_t = |\mathrm{Det} B_t|^{1/n} = \left(\frac{n}{\sqrt{n^2-1}}\right)^{(n-1)/n} \left(\frac{n}{n+1}\right)^{1/n} \rho_{t-1}$$

(see (4.1.16)) and go to rule 4).

4) [Termination test]. We check whether the inequality

$$\frac{\rho_t}{r} < \frac{\epsilon}{L_t + \epsilon} \tag{4.1.19}$$

is satisfied. If it is the case, we terminate and output, as the result of the solution process, the best (i.e., with the smallest value of f) of the "search points" $c_{\tau-1}$ associated with *productive* steps $\tau \leq t$ (we shall see that these productive steps indeed exist, so that the result of the solution process is well-defined). If (4.1.19) is not satisfied, we go to step $t+1$.

Just to get some feeling how the method works, here is a 2D illustration. The problem is

$$\begin{array}{rcl} & & \min\limits_{-1 \leq x_1, x_2 \leq 1} f(x), \\ f(x) & = & \frac{1}{2}(1.443508244x_1 + 0.623233851x_2 - 7.957418455)^2 \\ & & +5(-0.350974738x_1 + 0.799048618x_2 + 2.877831823)^4, \end{array}$$

the optimal solution is $x_1^* = 1, x_2^* = -1$, and the exact optimal value is 70.030152768...

The values of f at the best (i.e., with the smallest value of the objective) feasible solutions found in course of first t steps of the method, $t = 1, 2, ..., 256$, are shown in the following table:

t	best value	t	best value
1	374.61091739	16	76.838253451
2	216.53084103	...	...
3	146.74723394	32	70.901344815
4	112.42945457	...	...
5	93.84206347	64	70.031633483
6	82.90928589	...	...
7	82.90928589	128	70.030154192
8	82.90928589	...	...
...	...	256	70.030152768

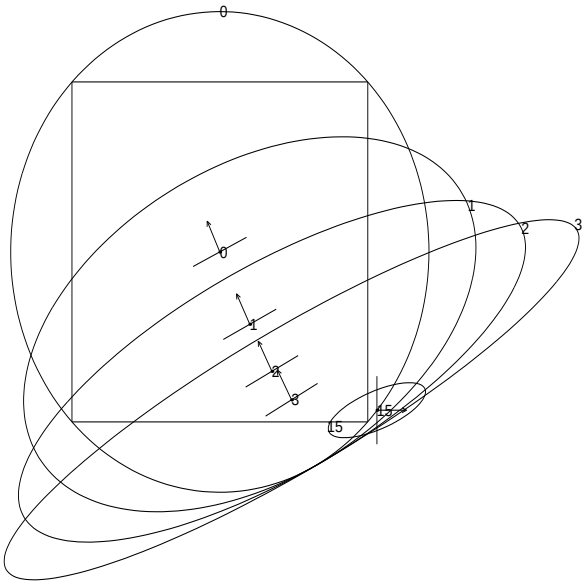


Figure 5.2. Ellipses E_{t-1} and search points c_{t-1}, $t = 1, 2, 3, 4, 16$
Arrows: gradients of the objective $f(x)$
Unmarked segments: tangents to the level lines of $f(x)$

The initial phase of the process looks as shown on Fig. 5.2.

Ellipsoid method: complexity analysis. We are about to establish our key result (which, in particular, immediately implies Theorem 4.1.2):

Theorem 4.1.3 *Let the Ellipsoid method be applied to convex program* (4.1.10) *of dimension $n \geq 2$ such that the feasible set X of the problem contains a Euclidean ball of a given radius $r > 0$ and is contained in the ball $E_0 = \{\|x\|_2 \leq R\}$ of a given radius R. For every input accuracy $\epsilon > 0$, the Ellipsoid method terminates after no more than*

$$N(\epsilon) = \mathrm{Ceil}\left(2n^2\left[\ln\left(\frac{R}{r}\right) + \ln\left(\frac{\epsilon + \mathrm{Var}_R(f)}{\epsilon}\right)\right]\right) + 1 \tag{4.1.20}$$

steps, where

$$\mathrm{Var}_R(f) = \max_{E_0} f - \min_{E_0} f,$$

and $\mathrm{Ceil}(a)$ *is the smallest integer $\geq a$. Moreover, the result $\widehat{x}$ generated by the method is a feasible ϵ-solution to* (4.1.10)*:*

$$\widehat{x} \in X \text{ and } f(x) - \min_X f \leq \epsilon. \tag{4.1.21}$$

Proof. We should prove the following pair of statements:

(i) The method terminates in course of the first $N(\epsilon)$ steps

(ii) The result $\widehat{x}$ is a feasible ϵ-solution to the problem.

1^0. Comparing the preliminary and the final description of the Ellipsoid method and taking into account the initialization rule, we see that if the method does not terminate before step t or terminates at this step according to rule 4), then

$$\begin{array}{lrcl}
(a) & E_0 & \supset & X; \\
(b) & E_\tau & \supset & \widehat{E}_\tau = \left\{x \in E_{\tau-1} \mid a_\tau^T x \leq a_\tau^T c_{\tau-1}\right\},\ \tau = 1, ..., t; \\
(c) & \mathrm{Vol}(E_\tau) & = & \rho_\tau^n \mathrm{Vol}(E_0) = \left(\frac{n}{\sqrt{n^2-1}}\right)^{n-1} \frac{n}{n+1}\mathrm{Vol}(E_{\tau-1}) \\
 & & \leq & \exp\{-1/(2n)\}\mathrm{vol}(E_{\tau-1}),\ \tau = 1, ..., t.
\end{array} \tag{4.1.22}$$

Note that from (c) it follows that

$$\rho_\tau \le \exp\{-\tau/(2n^2)\}R,\ \tau = 1, ..., t. \tag{4.1.23}$$

2^0. We claim that

If the Ellipsoids method terminates at certain step t, then the result $\widehat{x}$ is well-defined and is a feasible ϵ-solution to (4.1.10).

Indeed, there are only two possible reasons for termination. First, it may happen that $c_{t-1} \in X$ and $f'(c_{t-1}) = 0$ (see rule 2)). From our preliminary considerations we know that in this case c_{t-1} is an optimal solution to (4.1.10), which is even more than what we have claimed. Second, it may happen that at step t relation (4.1.19) is satisfied. Let us prove that the claim of 2^0 takes place in this case as well.

2^0.a) Let us set

$$\nu = \frac{\epsilon}{\epsilon + L_t} \in (0, 1].$$

By (4.1.19), we have $\rho_t/r < \nu$, so that there exists ν' such that

$$\frac{\rho_t}{r} < \nu' < \nu \quad [\le 1]. \tag{4.1.24}$$

Let x_* be an optimal solution to (4.1.10), and X^+ be the "ν'-shrinkage" of X to x_*:

$$X^+ = x_* + \nu'(X - x_*) = \{x = (1 - \nu')x_* + \nu' z \mid z \in X\}. \tag{4.1.25}$$

We have

$$\mathrm{Vol}(X^+) = (\nu')^n \mathrm{Vol}(X) \ge \left(\frac{r\nu'}{R}\right)^n \mathrm{Vol}(E_0) \tag{4.1.26}$$

(the last inequality is given by the fact that X contains a Euclidean ball of the radius r), while

$$\mathrm{Vol}(E_t) = \left(\frac{\rho_t}{R}\right)^n \mathrm{Vol}(E_0) \tag{4.1.27}$$

by definition of ρ_t. Comparing (4.1.26), (4.1.27) and taking into account that $\rho_t < r\nu'$ by (4.1.24), we conclude that $\mathrm{Vol}(E_t) < \mathrm{Vol}(X^+)$ and, consequently, X^+ cannot be contained in E_t. Thus, there exists a point y which belongs to X^+:

$$y = (1 - \nu')x_* + \nu' z \qquad [z \in X], \tag{4.1.28}$$

and does *not* belong to E_t.

2^0.b) Since y does not belong to E_t and at the same time belongs to $X \subset E_0$ along with x_* and z (X is convex!), we see that there exists a $\tau \le t$ such that $y \in E_{\tau-1}$ and $y \notin E_\tau$. By (4.1.22.b), every point x from the complement of E_τ in $E_{\tau-1}$ satisfies the relation $a_\tau^T x > a_\tau^T c_{\tau-1}$. Thus, we have

$$a_\tau^T y > a_\tau^T c_{\tau-1} \tag{4.1.29}$$

2^0.c) Observe that the step τ is surely productive. Indeed, otherwise, by construction of the method, a_t would separate X from $c_{\tau-1}$, and (4.1.29) would be impossible (we know that $y \in X$!). Notice that in particular we have just proved that if the method terminates at a step t, then at least one of the steps $1, ..., t$ is productive, so that the result is well-defined.

Since step τ is productive, a_τ is a subgradient of f at $c_{\tau-1}$ (description of the method!), so that

$$f(u) \ge f(c_{\tau-1}) + a_\tau^T(u - c_{\tau-1})$$

for all $u \in X$, and in particular for $u = x_*$. On the other hand, $z \in X \subset E_0$, so that by the definition of ℓ_τ and L_τ we have

$$a_\tau^T(z - c_{\tau-1}) \le \ell_\tau \le L_\tau.$$

Thus,

$$\begin{array}{l} f(x_*) \geq f(c_{\tau-1}) + a_\tau^T(x_* - c_{\tau-1}) \\ L_\tau \geq a_\tau^T(z - c_{\tau-1}) \end{array}$$

Multiplying the first inequality by $(1-\nu')$, the second – by ν' and adding the results, we get

$$\begin{array}{rcl} (1-\nu')f(x_*) + \nu' L_\tau & \geq & (1-\nu')f(c_{\tau-1}) + a_\tau^T([(1-\nu')x_* + \nu' z] - c_{\tau-1}) \\ & = & (1-\nu')f(c_{\tau-1}) + a_\tau^T(y - c_{\tau-1}) \\ & & [\text{see (4.1.28)}] \\ & \geq & (1-\nu')f(c_{\tau-1}) \\ & & [\text{see (4.1.29)}] \end{array}$$

and we come to

$$\begin{array}{rcl} f(c_{\tau-1}) & \leq & f(x_*) + \frac{\nu' L_\tau}{1-\nu'} \\ & \leq & f(x_*) + \frac{\nu' L_t}{1-\nu'} \\ & & [\text{since } L_\tau \leq L_t \text{ in view of } \tau \leq t] \\ & \leq & f(x_*) + \epsilon \\ & & [\text{by definition of } \nu \text{ and since } \nu' < \nu] \\ & = & \mathrm{Opt(C)} + \epsilon. \end{array}$$

We see that *there exists a productive (i.e., with feasible $c_{\tau-1}$) step $\tau \leq t$ such that the corresponding search point $c_{\tau-1}$ is ϵ-optimal.* Since we are in the situation where the result $\widehat{x}$ is the best of the feasible search points generated in course of the first t steps, $\widehat{x}$ is also feasible and ϵ-optimal, as claimed in 2^0.

3^0 It remains to verify that the method does terminate in course of the first $N = N(\epsilon)$ steps. Assume, on the contrary, that it is not the case, and let us lead this assumption to a contradiction.

First, observe that for every productive step t we have

$$c_{t-1} \in X \text{ and } a_t = f'(c_{t-1}),$$

whence, by the definition of a subgradient and the variation $\mathrm{Var}_R(f)$,

$$u \in E_0 \Rightarrow \mathrm{Var}_R(f) \geq f(u) - f(c_{t-1}) \geq a_t^T(u - c_{t-1}),$$

whence

$$\ell_t \equiv \max_{u \in E_0} a_t^T(u - c_{t-1}) \leq \mathrm{Var}_R(f).$$

Looking at the description of the method, we conclude that

$$L_t \leq \mathrm{Var}_R(f) \qquad \forall t. \tag{4.1.30}$$

Since we have assumed that the method does not terminate in course of the first N steps, we have

$$\frac{\rho_N}{r} \geq \frac{\epsilon}{\epsilon + L_N}. \tag{4.1.31}$$

The right hand side in this inequality is $\geq \epsilon/(\epsilon + \mathrm{Var}_R(f))$ by (4.1.30), while the left hand side is $\leq \exp\{-N/(2n^2)\}R$ by (4.1.23). We get

$$\exp\{-N/(2n^2)\}R/r \geq \frac{\epsilon}{\epsilon + \mathrm{Var}_R(f)} \Rightarrow N \leq 2n^2 \left[\ln\left(\frac{R}{r}\right) + \ln\left(\frac{\epsilon + \mathrm{Var}_R(f)}{\epsilon}\right)\right],$$

which is the desired contradiction (see the definition of $N = N(\epsilon)$ in (4.1.20)). ∎

4.1.5 Difficult continuous optimization problems

Real Arithmetic Complexity Theory can borrow from the Combinatorial Complexity Theory techniques for detecting "computationally intractable" problems. Consider the situation as follows: we are given a family $\mathcal{P}$ of optimization programs and want to understand whether the family is computationally tractable. An affirmative answer can be obtained from Theorem 4.1.1; but how could we justify that the family is *in*tractable? A natural course of action here is to demonstrate that certain difficult (NP-complete) combinatorial problem $\mathcal{Q}$ can be reduced to $\mathcal{P}$ in such a way that the possibility to solve $\mathcal{P}$ in polynomial time would imply similar possibility for $\mathcal{Q}$. Assume that the objectives of the instances of $\mathcal{P}$ are polynomially computable, and that we can point out a generic combinatorial problem $\mathcal{Q}$ *known to be* NP-*complete* which can be *reduced* to $\mathcal{P}$ in the following sense:

> *There exists a CCT-polynomial time algorithm $\mathcal{M}$ which, given on input the data vector* $\mathrm{Data}(q)$ *of an instance $(q) \in \mathcal{Q}$, converts it into a triple* $\mathrm{Data}(p[q])$, $\epsilon(q)$, $\mu(q)$ *comprised of the data vector of an instance $(p[q]) \in \mathcal{P}$, positive rational $\epsilon(q)$ and rational $\mu(q)$ such that $(p[q])$ is solvable and*
>
> *— if (q) is unsolvable, then the value of the objective of $(p[q])$ at every $\epsilon(q)$-solution to this problem is $\leq \mu(q) - \epsilon(q)$;*
>
> *— if (q) is solvable, then the value of the objective of $(p[q])$ at every $\epsilon(q)$-solution to this problem is $\geq \mu(q) + \epsilon(q)$.*

We claim that in the case in question we have all reasons to qualify $\mathcal{P}$ as a "computationally intractable" problem. Assume, on contrary, that $\mathcal{P}$ admits a polynomial time solution method $\mathcal{S}$, and let us look what happens if we apply this algorithm to solve $(p[q])$ within accuracy $\epsilon(q)$. Since $(p[q])$ is solvable, the method must produce an $\epsilon(q)$-solution $\widehat{x}$ to $(p[q])$. With additional "polynomial time effort" we may compute the value of the objective of $(p[q])$ at $\widehat{x}$ (recall that the objectives of instances from $\mathcal{P}$ are assumed to be polynomially computable). Now we can compare the resulting value of the objective with $\mu(q)$; by definition of reducibility, if this value is $\leq \mu(q)$, q is unsolvable, otherwise q is solvable. Thus, we get a correct "Real Arithmetic" solvability test for $\mathcal{Q}$. By definition of a Real Arithmetic polynomial time algorithm, the running time of the test is bounded by a polynomial of $s(q) = \mathrm{Size}(p[q])$ and of the quantity

$$d(q) = \mathrm{Digits}((p[q]), \epsilon(q)) = \ln\left(\frac{\mathrm{Size}(p[q]) + \|\mathrm{Data}(p[q])\|_1 + \epsilon^2(q)}{\epsilon(q)}\right).$$

Now note that if $\ell = \mathrm{length}(\mathrm{Data}(q))$, then the total number of bits in $\mathrm{Data}(p[q])$ and in $\epsilon(q)$ is bounded by a polynomial of ℓ (since the transformation $\mathrm{Data}(q) \mapsto (\mathrm{Data}(p[q]), \epsilon(q), \mu(q))$ takes CCT-polynomial time). It follows that both $s(q)$ and $d(q)$ are bounded by polynomials in ℓ, so that our "Real Arithmetic" solvability test for $\mathcal{Q}$ takes polynomial in $\mathrm{length}(\mathrm{Data}(q))$ number of arithmetic operations.

Recall that $\mathcal{Q}$ was assumed to be an NP-complete generic problem, so that it would be "highly improbable" to find a polynomial time solvability test for this problem, while we have managed to build such a test. We conclude that the polynomial solvability of $\mathcal{P}$ is highly improbable as well.

4.2 Interior Point Polynomial Time Methods for LP, CQP and SDP

4.2.1 Motivation

Theorem 4.1.1 states that generic convex programs, under mild computability and boundedness assumptions, are polynomially solvable. This result is extremely important theoretically; however, from the practical viewpoint it is, essentially, no more than "an existence theorem". Indeed, the "universal" complexity bounds coming from Theorem 4.1.2, although polynomial, are not that attractive: by Theorem 4.1.1, when solving problem (4.1.10) with n design variables, the "price" of an accuracy digit (what it costs to reduce current inaccuracy ϵ by factor 2) is $O(n^2)$ calls to the first order and the separation oracles plus $O(n^4)$ arithmetic operations to process the answers of the oracles. Thus, even for simplest

objectives to be minimized over simplest feasible sets, the arithmetic price of an accuracy digit is $O(n^4)$; think how long will it take to solve a problem with, say, 1,000 variables (which is still a "small" size for many applications). The good news about the methods underlying Theorem 4.1.2 is their universality: all they need is a Separation oracle for the feasible set and the possibility to compute the objective and its subgradient at a given point, which is not that much. The bad news about these methods has the same source as the good news: the methods are "oracle-oriented" and capable to use only *local* information on the program they are solving, in contrast to the fact that when solving instances of well-structured programs, like LP, we from the very beginning have in our disposal complete *global* description of the instance. And of course it is ridiculous to use a *complete global* knowledge of the instance just to mimic the *local* in their nature first order and separation oracles. What we would like to have is an optimization technique capable to "utilize efficiently" our global knowledge of the instance and thus allowing to get a solution much faster than it is possible for "nearly blind" oracle-oriented algorithms. The major event in the "recent history" of Convex Optimization, called sometimes "Interior Point revolution", was the invention of these "smart" techniques.

4.2.2 Interior Point methods

The Interior Point revolution was started by the seminal work of N. Karmarkar (1984) where the first interior point method for LP was proposed; in 18 years since then, interior point (IP) polynomial time methods have become an extremely deep and rich theoretically and highly promising computationally area of Convex Optimization. A somehow detailed overview of the history and the recent state of this area is beyond the scope of this course; an interested reader is referred to [17, 19, 13] and references therein. All we intend to do is to give an idea of what are the IP methods, skipping nearly all (sometimes highly instructive and nontrivial) technicalities.

The simplest way to get a proper impression of the (most of) IP methods is to start with a quite traditional *interior penalty* scheme for solving optimization problems.

The Newton method and the Interior penalty scheme

Unconstrained minimization and the Newton method. Seemingly the simplest convex optimization problem is the one of unconstrained minimization of a smooth strongly convex objective:

$$\min_x \{f(x) : x \in \mathbf{R}^n\}; \tag{UC}$$

a "smooth strongly convex" in this context means a 3 times continuously differentiable convex function f such that $f(x) \to \infty$, $\|x\|_2 \to \infty$, and such that the Hessian matrix $f''(x) = \left[\frac{\partial^2 f(x)}{\partial x_i \partial x_j}\right]$ of f is positive definite at every point x. Among numerous techniques for solving (UC), the most remarkable one is the Newton method. In its pure form, the Newton method is extremely transparent and natural: given a current iterate x, we approximate our objective f by its second-order Taylor expansion at the iterate – by the quadratic function

$$f_x(y) = f(x) + (y-x)^T f'(x) + \frac{1}{2}(y-x)^T f''(x)(y-x)$$

– and choose as the next iterate x_+ the minimizer of this quadratic approximation. Thus, the Newton method merely iterates the updating

$$x \mapsto x_+ = x - [f''(x)]^{-1} f'(x). \tag{Nwt}$$

In the case of a (strongly convex) quadratic objective, the approximation coincides with the objective itself, so that the method reaches the exact solution in one step. It is natural to guess (and indeed is true) that in the case when the objective is smooth and strongly convex (although not necessary quadratic) and the current iterate x is close enough to the minimizer x_* of f, the next iterate x_+, although not

being x_* exactly, will be "much closer" to the exact minimizer than x. The precise (and easy) result is that the Newton method converges *locally quadratically*, i.e., that

$$\|x_+ - x_*\|_2 \le C\|x - x_*\|_2^2,$$

provided that $\|x - x_*\|_2 \le r$ with small enough value of $r > 0$ (both this value and C depend on f). Quadratic convergence means essentially that eventually every new step of the process increases by a constant factor the number of accuracy digits in the approximate solution.

When started not "close enough" to the minimizer, the "pure" Newton method (Nwt) can demonstrate weird behaviour (look, e.g., what happens when the method is applied to the univariate function $f(x) = \sqrt{1+x^2}$). The simplest way to overcome this drawback is to pass from the pure Newton method to its *damped* version

$$x \mapsto x_+ = x - \gamma(x)[f''(x)]^{-1} f'(x), \tag{NwtD}$$

where the *stepsize* $\gamma(x) > 0$ is chosen in a way which, on one hand, ensures global convergence of the method and, on the other hand, enforces $\gamma(x) \to 1$ as $x \to x_*$, thus ensuring fast (essentially the same as for the pure Newton method) asymptotic convergence of the process[2)].

Practitioners thought the (properly modified) Newton method to be the fastest, in terms of the iteration count, routine for smooth (not necessarily convex) unconstrained minimization, although sometimes "too heavy" for practical use: the practical drawbacks of the method are both the necessity to invert the Hessian matrix at each step, which is computationally costly in the large-scale case, and especially the necessity to compute this matrix (think how difficult it is to write a code computing 5,050 second order derivatives of a messy function of 100 variables).

Classical interior penalty scheme: the construction. Now consider a constrained convex optimization program. As we remember, one can w.l.o.g. make its objective linear, moving, if necessary, the actual objective to the list of constraints. Thus, let the problem be

$$\min_x \left\{ c^T x : x \in \mathcal{X} \subset \mathbf{R}^n \right\}, \tag{C}$$

where $\mathcal{X}$ is a closed convex set, which we assume to possess a nonempty interior. How could we solve the problem?

Traditionally it was thought that the problems of smooth convex unconstrained minimization are "easy"; thus, a quite natural desire was to reduce the constrained problem (C) to a series of smooth *un*constrained optimization programs. To this end, let us choose somehow a *barrier* (another name – "an interior penalty function") $F(x)$ for the feasible set $\mathcal{X}$ – a function which is well-defined (and is smooth and strongly convex) on the interior of $\mathcal{X}$ and "blows up" as a point from $\operatorname{int}\mathcal{X}$ approaches a boundary point of $\mathcal{X}$:

$$x_i \in \operatorname{int}\mathcal{X},\ x \equiv \lim_{i\to\infty} x_i \in \partial\mathcal{X} \Rightarrow F(x_i) \to \infty,\ i \to \infty,$$

and let us look at the one-parametric family of functions generated by our objective and the barrier:

$$F_t(x) = tc^T x + F(x) : \operatorname{int}\mathcal{X} \to \mathbf{R}.$$

Here the *penalty parameter* t is assumed to be nonnegative.

It is easily seen that under mild regularity assumptions (e.g., in the case of bounded $\mathcal{X}$, which we assume from now on)

- Every function $F_t(\cdot)$ attains its minimum over the interior of $\mathcal{X}$, the minimizer $x_*(t)$ being unique;

[2)] There are many ways to provide the required behaviour of $\gamma(x)$, e.g., to choose $\gamma(x)$ by a linesearch in the direction $e(x) = -[f''(x)]^{-1}f'(x)$ of the Newton step:

$$\gamma(x) = \operatorname*{argmin}_t f(x + te(x)).$$

- The *central path* $x_*(t)$ is a smooth curve, and all its limiting, $t \to \infty$, points belong to the set of optimal solutions of (C).

  This fact is quite clear intuitively. To minimize $F_t(\cdot)$ for large t is the same as to minimize the function $f_\rho(x) = c^T x + \rho F(x)$ for small $\rho = \frac{1}{t}$. When ρ is small, the function f_ρ is very close to $c^T x$ everywhere in $\mathcal{X}$, except a narrow stripe along the boundary of $\mathcal{X}$, the stripe becoming thinner and thinner as $\rho \to 0$. Therefore we have all reasons to believe that the minimizer of F_t for large t (i.e., the minimizer of f_ρ for small ρ) must be close to the set of minimizers of $c^T x$ on $\mathcal{X}$.

We see that the central path $x_*(t)$ is a kind of Ariadne's thread which leads to the solution set of (C). On the other hand, to reach, given a value $t \geq 0$ of the penalty parameter, the point $x_*(t)$ on this path is the same as to minimize a smooth strongly convex function $F_t(\cdot)$ *which attains its minimum at an interior point of* $\mathcal{X}$. The latter problem is "nearly unconstrained one", up to the fact that its objective is not everywhere defined. However, we can easily adapt the methods of unconstrained minimization, including the Newton one, to handle "nearly unconstrained" problems. We see that constrained convex optimization in a sense can be reduced to the "easy" unconstrained one. The conceptually simplest way to make use of this observation would be to choose a "very large" value $\bar{t}$ of the penalty parameter, like $\bar{t} = 10^6$ or $\bar{t} = 10^{10}$, and to run an unconstrained minimization routine, say, the Newton method, on the function $F_{\bar{t}}$, thus getting a good approximate solution to (C) "in one shot". This policy, however, is impractical: since we have no idea where $x_*(\bar{t})$ is, we normally will start our process of minimizing $F_{\bar{t}}$ very far from the minimizer of this function, and thus for a long time will be unable to exploit fast local convergence of the method for unconstrained minimization we have chosen. A smarter way to use our Ariadne's thread is exactly the one used by Theseus: to follow the thread. Assume, e.g., that we know in advance the minimizer of $F_0 \equiv F$, i.e., the point $x_*(0)$[3)]. Thus, we know where the central path starts. Now let us follow this path: at i-th step, standing at a point x_i "close enough" to some point $x_*(t_i)$ of the path, we

• first, increase a bit the current value t_i of the penalty parameter, thus getting a new "target point" $x_*(t_{i+1})$ on the path,

and

• second, approach our new target point $x_*(t_{i+1})$ by running, say, the Newton method, *started at our current iterate* x_i, on the function $F_{t_{i+1}}$, until a new iterate x_{i+1} "close enough" to $x_*(t_{i+1})$ is generated.

As a result of such a step, we restore the initial situation – we again stand at a point which is close to a point on the central path, but this latter point has been moved along the central path towards the optimal set of (C). Iterating this updating and strengthening appropriately our "close enough" requirements as the process goes on, we, same as the central path, approach the optimal set. A conceptual advantage of this "path-following" policy as compared to the "brute force" attempt to reach a target point $x_*(\bar{t})$ with large $\bar{t}$ is that now we have a hope to exploit all the time the strongest feature of our "working horse" (the Newton method) – its fast local convergence. Indeed, assuming that x_i is close to $x_*(t_i)$ and that we do not increase the penalty parameter too rapidly, so that $x_*(t_{i+1})$ is close to $x_*(t_i)$ (recall that the central path is smooth!), we conclude that x_i is close to our new target point $x_*(t_i)$. If all our "close enough" and "not too rapidly" are properly controlled, we may ensure x_i to be in the domain of the quadratic convergence of the Newton method as applied to $F_{t_{i+1}}$, and then it will take a quite small number of steps of the method to recover closeness to our new target point.

Classical interior penalty scheme: the drawbacks. At a qualitative "common sense" level, the interior penalty scheme looks quite attractive and extremely flexible: for the majority of optimization problems treated by the classical optimization, there is a plenty of ways to build a relatively simple barrier meeting all the requirements imposed by the scheme, there is a huge room to play with the policies for increasing the penalty parameter and controlling closeness to the central path, etc. And the theory says that under quite mild and general assumptions on the choice of the numerous "free parameters" of our

[3)] There is no difficulty to ensure thus assumption: given an arbitrary barrier F and an arbitrary starting point $\bar{x} \in \operatorname{int}\mathcal{X}$, we can pass from F to a new barrier $\bar{F} = F(x) - (x - \bar{x})^T F'(\bar{x})$ which attains its minimum exactly at $\bar{x}$, and then use the new barrier $\bar{F}$ instead of our original barrier F; and for the traditional approach we are following for the time being, F has absolutely no advantages as compared to $\bar{F}$.

construction, it still is guaranteed to converge to the optimal set of the problem we have to solve. All looks wonderful, until we realize that the convergence ensured by the theory is completely "unqualified", it is a purely asymptotical phenomenon: we are promised to reach eventually a solution of a whatever accuracy we wish, but how long it will take for a given accuracy – this is the question the "classical" optimization theory, with its "convergence" – "asymptotic linear/superlinear/quadratic convergence" neither posed nor answered. And since our life in this world is finite (moreover, usually more finite than we would like it to be), "asymptotical promises" are perhaps better than nothing, but definitely are not all we would like to know. What is vitally important for us in theory (and to some extent – also in practice) is the issue of *complexity*: given an instance of such and such generic optimization problem and a desired accuracy ϵ, how large is the computational effort (# of arithmetic operations) needed to get an ϵ-solution of the instance? And we would like the answer to be a kind of a polynomial time complexity bound, and not a quantity depending on "unobservable and uncontrollable" properties of the instance, like the "level of regularity" of the boundary of $\mathcal{X}$ at the (unknown!) optimal solution of the instance.

It turns out that the intuitively nice classical theory we have outlined is unable to say a single word on the complexity issues (it is how it should be: a reasoning in purely qualitative terms like "smooth", "strongly convex", etc., definitely cannot yield a quantitative result...) Moreover, from the complexity viewpoint just the very philosophy of the classical convex optimization turns out to be wrong:

• As far as the complexity is concerned, for nearly all "black box represented" classes of unconstrained convex optimization problems (those where all we know is that the objective is called $f(x)$, is (strongly) convex and 2 (3,4,5...) times continuously differentiable, and can be computed, along with its derivatives up to order ... at every given point), there is no such phenomenon as "local quadratic convergence", the Newton method (which uses the second derivatives) has no advantages as compared to the methods which use only the first order derivatives, etc.;

• The very idea to reduce "black-box-represented" constrained convex problems to unconstrained ones – from the complexity viewpoint, the unconstrained problems are not easier than the constrained ones...

4.2.3 But...

Luckily, the pessimistic analysis of the classical interior penalty scheme is not the "final truth". It turned out that what prevents this scheme to yield a polynomial time method is not the structure of the scheme, but the huge amount of freedom it allows for its elements (too much freedom is another word for anarchy...). After some order is added, the scheme becomes a polynomial time one! Specifically, it was understood that

1. There is a (completely non-traditional) class of "good" (self-concordant[4)]) barriers. Every barrier F of this type is associated with a "self-concordance parameter" $\theta(F)$, which is a real ≥ 1;
2. Whenever a barrier F underlying the interior penalty scheme is self-concordant, one can specify the notion of "closeness to the central path" and the policy for updating the penalty parameter in such a way that a *single* Newton step
$$x_i \mapsto x_{i+1} = x_i - [\nabla^2 F_{t_{i+1}}(x_i)]^{-1} \nabla F_{t_{i+1}}(x_i) \tag{4.2.1}$$
suffices to update a "close to $x_*(t_i)$" iterate x_i into a new iterate x_{i+1} which is close, in the same sense, to $x_*(t_{i+1})$. All "close to the central path" points belong to int $\mathcal{X}$, so that the scheme keeps all the iterates strictly feasible.
3. The penalty updating policy mentioned in the previous item is quite simple:
$$t_i \mapsto t_{i+1} = \left(1 + \frac{0.1}{\sqrt{\theta(F)}}\right) t_i;$$

[4)] We do not intend to explain here what is a "self-concordant barrier"; for our purposes it suffices to say that this is a three times continuously differentiable convex barrier F satisfying a pair of specific differential inequalities linking the first, the second and the third directional derivatives of F.

in particular, it does not "slow down" as t_i grows and ensures linear, with the ratio $\left(1+\frac{0.1}{\sqrt{\theta(F)}}\right)$, growth of the penalty. This is vitally important due to the following fact:

4. The inaccuracy of a point x, which is close to some point $x_*(t)$ of the central path, as an approximate solution to (C) is inverse proportional to t:

$$c^T x - \min_{y\in\mathcal{X}} c^T y \le \frac{2\theta(F)}{t}.$$

It follows that

> (!) *After we have managed once to get close to the central path – have built a point x_0 which is close to a point $x(t_0)$, $t_0>0$, on the path, every $O(\sqrt{\theta(F)})$ steps of the scheme improve the quality of approximate solutions generated by the scheme by an absolute constant factor. In particular, it takes no more than*
>
> $$O(1)\sqrt{\theta(F)}\ln\left(2+\frac{\theta(F)}{t_0\epsilon}\right)$$
>
> *steps to generate a strictly feasible ϵ-solution to* (C).

Note that with our simple penalty updating policy all needed to perform a step of the interior penalty scheme is to compute the gradient and the Hessian of the underlying barrier at a single point and to invert the resulting Hessian.

Items 3, 4 say that essentially all we need to derive from the just listed general results a polynomial time method for a generic convex optimization problem is to be able to equip every instance of the problem with a "good" barrier in such a way that both the parameter of self-concordance of the barrier $\theta(F)$ and the arithmetic cost at which we can compute the gradient and the Hessian of this barrier at a given point are polynomial in the size of the instance[5)]. And it turns out that we can meet the latter requirement for all interesting "well-structured" generic convex programs, in particular, for Linear, Conic Quadratic, and Semidefinite Programming. Moreover, "the heroes" of our course – LP, CQP and SDP – are especially nice application fields of the general theory of interior point polynomial time methods; in these particular applications, the theory can be simplified, on one hand, and strengthened, on another.

4.3 Interior point methods for LP, CQP, and SDP: building blocks

We are about to explain what the interior point methods for LP, CQP, SDP look like.

4.3.1 Canonical cones and canonical barriers

We will be interested in a generic conic problem

$$\min_x \left\{c^T x : \mathcal{A}x - B \in \mathbf{K}\right\} \tag{CP}$$

associated with a cone $\mathbf{K}$ given as a direct product of m "basic" cones, each of them being either a second-order, or a semidefinite cone:

$$\mathbf{K} = \mathbf{S}^{k_1}_+ \times ... \times \mathbf{S}^{k_p}_+ \times \mathbf{L}^{k_{p+1}} \times ... \times \mathbf{L}^{k_m} \subset E = \mathbf{S}^{k_1} \times ... \times \mathbf{S}^{k_p} \times \mathbf{R}^{k_{p+1}} \times ... \times \mathbf{R}^{k_m}. \tag{Cone}$$

[5)] Another requirement is to be able once get close to a point $x_*(t_0)$ on the central path with a not "disastrously small" value of t_0 – we should initialize somehow our path-following method! It turns out that such an initialization is a minor problem – it can be carried out via the same path-following technique, provided we are given in advance a strictly feasible solution to our problem.

Of course, the generic problem in question covers LP (no Lorentz factors, all semidefinite factors are of dimension 1), CQP (no semidefinite factors) and SDP (no Lorentz factors).

Now, we shall equip the semidefinite and the Lorentz cones with "canonical barriers":

• The canonical barrier for a semidefinite cone $\mathbf{S}^n_+$ is

$$S_k(X) = -\ln \mathrm{Det}(X) : \mathrm{int}\, \mathbf{S}^k_+ \to \mathbf{R};$$

the *parameter* of this barrier, *by definition*, is $\theta(S_k) = k$ [6)].

• the canonical barrier for a Lorentz cone $\mathbf{L}^k = \{x \in \mathbf{R}^k \mid x_k \geq \sqrt{x_1^2 + ... + x_{k-1}^2}\}$ is

$$L_k(x) = -\ln(x_k^2 - x_1^2 - ... - x_{k-1}^2) = -\ln(x^T J_k x), \quad J_k = \begin{pmatrix} -I_{k-1} & \\ & 1 \end{pmatrix};$$

the parameter of this barrier is $\theta(L_k) = 2$.

• The canonical barrier K for the cone $\mathbf{K}$ given by (Cone), *by definition*, is the direct sum of the canonical barriers of the factors:

$$K(X) = S_{k_1}(X_1) + ... + S_{k_p}(X_p) + L_{k_{p+1}}(X_{p+1}) + ... + L_{k_m}(X_m), \quad X_i \in \begin{cases} \mathrm{int}\, \mathbf{S}^{k_i}_+, & i \leq p \\ \mathrm{int}\, \mathbf{L}^{k_i}, & p < i \leq m \end{cases};$$

from now on, we use upper case Latin letters, like X, Y, Z, to denote elements of the space E; for such an element X, X_i denotes the projection of X onto i-th factor in the direct product representation of E as shown in (Cone).

The parameter of the barrier K, again by definition, is the sum of parameters of the basic barriers involved:

$$\theta(K) = \theta(S_{k_1}) + ... + \theta(S_{k_p}) + \theta(L_{k_{p+1}}) + ... + \theta(L_{k_m}) = \sum_{i=1}^{p} k_i + 2(m-p).$$

Recall that all direct factors in the direct product representation (Cone) of our "universe" E are Euclidean spaces; the matrix factors $\mathbf{S}^{k_i}$ are endowed with the Frobenius inner product

$$\langle X_i, Y_i \rangle_{\mathbf{S}^{k_i}} = \mathrm{Tr}(X_i Y_i),$$

while the "arithmetic factors" $\mathbf{R}^{k_i}$ are endowed with the usual inner product

$$\langle X_i, Y_i \rangle_{\mathbf{R}^{k_i}} = X_i^T Y_i;$$

E itself will be regarded as a Euclidean space endowed with the direct sum of inner products on the factors:

$$\langle X, Y \rangle_E = \sum_{i=1}^{p} \mathrm{Tr}(X_i Y_i) + \sum_{i=p+1}^{m} X_i^T Y_i.$$

It is clearly seen that our basic barriers, same as their direct sum K, indeed are barriers for the corresponding cones: they are C^∞-smooth on the interiors of their domains, blow up to ∞ along every sequence of points from these interiors converging to a boundary point of the corresponding domain and are strongly convex. To verify the latter property, it makes sense to compute explicitly the first and the second directional derivatives of these barriers (we need the corresponding formulae in any case); to simplify notation, we write down the derivatives of the basic functions S_k, L_k at a point x from their

[6)] The barrier S^k, same as the canonical barrier L_k for the Lorentz cone $\mathbf{L}^k$, indeed are self-concordant (whatever it means), and the parameters they are assigned here *by definition* are exactly their parameters of self-concordance.

domain along a direction h (you should remember that in the case of S_k both the point and the direction, in spite of their lower-case denotation, are $k \times k$ symmetric matrices):

$$\begin{array}{|rcl|}
\hline
DS_k(x)[h] \equiv \frac{d}{dt}\Big|_{t=0} S_k(x+th) & = & -\mathrm{Tr}(x^{-1}h) = -\langle x^{-1}, h\rangle_{\mathbf{S}^k}, \\
\text{i.e.} & & \\
\nabla S_k(x) & = & -x^{-1}; \\
D^2S_k(x)[h,h] \equiv \frac{d^2}{dt^2}\Big|_{t=0} S_k(x+th) & = & \mathrm{Tr}(x^{-1}hx^{-1}h) = \langle x^{-1}hx^{-1}, h\rangle_{\mathbf{S}^k}, \\
\text{i.e.} & & \\
[\nabla^2 S_k(x)]h & = & x^{-1}hx^{-1}; \\
\hline
DL_k(x)[h] \equiv \frac{d}{dt}\Big|_{t=0} L_k(x+th) & = & -2\frac{h^TJ_kx}{x^TJ_kx}, \\
\text{i.e.} & & \\
\nabla L_k(x) & = & -\frac{2}{x^TJ_kx}J_kx; \\
D^2L_k(x)[h,h] \equiv \frac{d^2}{dt^2}\Big|_{t=0} L_k(x+th) & = & 4\frac{[h^TJ_kx]^2}{[x^TJ_kx]^2} - 2\frac{h^TJ_kh}{x^TJx}, \\
\text{i.e.} & & \\
\nabla^2 L_k(x) & = & \frac{4}{[x^TJ_kx]^2}J_kxx^TJ_k - \frac{2}{x^TJ_kx}J_k. \\
\hline
\end{array} \tag{4.3.1}$$

From the expression for $D^2S_k(x)[h,h]$ we see that

$$D^2S_k(x)[h,h] = \mathrm{Tr}(x^{-1}hx^{-1}h) = \mathrm{Tr}([x^{-1/2}hx^{-1/2}]^2),$$

so that $D^2S_k(x)[h,h]$ is positive whenever $h \neq 0$. It is not difficult to prove that the same is true for $D^2L_k(x)[h,h]$. Thus, the canonical barriers for semidefinite and Lorentz cones are strongly convex, and so is their direct sum $K(\cdot)$.

It makes sense to illustrate relatively general concepts and results to follow by how they look in a particular case when $\mathbf{K}$ is the semidefinite cone $\mathbf{S}^k_+$; we shall refer to this situation as to the "SDP case". The essence of the matter in our general case is exactly the same as in this particular one, but "straightforward computations" which are easy in the SDP case become nearly impossible in the general case; and we have no possibility to explain here how it is possible (it is!) to get the desired results with minimum amount of computations.

Due to the role played by the SDP case in our exposition, we use for this case special notation, along with the just introduced "general" one. Specifically, we denote the standard – the Frobenius – inner product on $E = \mathbf{S}^k$ as $\langle\cdot,\cdot\rangle_F$, although feel free, if necessary, to use our "general" notation $\langle\cdot,\cdot\rangle_E$ as well; the associated norm is denoted by $\|\cdot\|_2$, so that $\|X\|_2 = \sqrt{\mathrm{Tr}(X^2)}$, X being a symmetric matrix.

4.3.2 Elementary properties of canonical barriers

Let us establish a number of simple and useful properties of canonical barriers.

Proposition 4.3.1 *A canonical barrier, let it be denoted F (F can be either S_k, or L_k, or the direct sum K of several copies of these "elementary" barriers), possesses the following properties:*

(i) *F is logarithmically homogeneous, the parameter of logarithmic homogeneity being $-\theta(F)$, i.e., the following identity holds:*

$$t > 0,\ x \in \mathrm{Dom}\, F \Rightarrow F(tx) = F(x) - \theta(F)\ln t.$$

- In the SDP case, i.e., when $F = S_k = -\ln\mathrm{Det}(x)$ and x is $k \times k$ positive definite matrix, (i) claims that

$$-\ln\mathrm{Det}(tx) = -\ln\mathrm{Det}(x) - k\ln t,$$

which of course is true.

(ii) *Consequently, the following two equalities hold identically in* $x \in \mathrm{Dom}\, F$*:*

$$\begin{array}{ll}(a) & \langle \nabla F(x), x\rangle = -\theta(F);\\ (b) & [\nabla^2 F(x)]x = -\nabla F(x).\end{array}$$

• In the SDP case, $\nabla F(x) = \nabla S_k(x) = -x^{-1}$ and $[\nabla^2 F(x)]h = \nabla^2 S_k(x)h = x^{-1}hx^{-1}$ (see (4.3.1)). Here (a) becomes the identity $\langle x^{-1}, x\rangle_F \equiv \mathrm{Tr}(x^{-1}x) = k$, and (b) kindly informs us that $x^{-1}xx^{-1} = x^{-1}$.

(iii) *Consequently, k-th differential* $D^k F(x)$ *of* F*,* $k \geq 1$*, is homogeneous, of degree* $-k$*, in* $x \in \mathrm{Dom}\, F$*:*

$$\begin{array}{l}\forall (x \in \mathrm{Dom}\, F, t > 0, h_1, ..., h_k):\\ D^k F(tx)[h_1, ..., h_k] \equiv \left.\frac{\partial^k F(tx + s_1h_1 + ... + s_kh_k)}{\partial s_1 \partial s_2 ... \partial s_k}\right|_{s_1 = ... = s_k = 0} = t^{-k} D^k F(x)[h_1, ..., h_k].\end{array} \tag{4.3.2}$$

Proof. (i): it is immediately seen that S_k and L_k are logarithmically homogeneous with parameters of logarithmic homogeneity $-\theta(S_k)$, $-\theta(L_k)$, respectively; and of course the property of logarithmic homogeneity is stable with respect to taking direct sums of functions: if $\mathrm{Dom}\,\Phi(u)$ and $\mathrm{Dom}\,\Psi(v)$ are closed w.r.t. the operation of multiplying a vector by a positive scalar, and both Φ and Ψ are logarithmically homogeneous with parameters α, β, respectively, then the function $\Phi(u) + \Psi(v)$ is logarithmically homogeneous with the parameter $\alpha + \beta$.

(ii): To get (ii.a), it suffices to differentiate the identity

$$F(tx) = F(x) - \theta(F)\ln t$$

in t at $t = 1$:

$$F(tx) = F(x) - \theta(F)\ln t \Rightarrow \langle \nabla F(tx), x\rangle = \frac{d}{dt}F(tx) = -\theta(F)t^{-2},$$

and it remains to set $t = 1$ in the concluding identity.

Similarly, to get (ii.b), it suffices to differentiate the identity

$$\langle \nabla F(x + th), x + th\rangle = -\theta(F)$$

(which is just (ii.a)) in t at $t = 0$, thus arriving at

$$\langle [\nabla^2 F(x)]h, x\rangle + \langle \nabla F(x), h\rangle = 0;$$

since $\langle [\nabla^2 F(x)]h, x\rangle = \langle [\nabla^2 F(x)]x, h\rangle$ (symmetry of partial derivatives!) and since the resulting equality

$$\langle [\nabla^2 F(x)]x, h\rangle + \langle \nabla F(x), h\rangle = 0$$

holds true identically in h, we come to $[\nabla^2 F(x)]x = -\nabla F(x)$.

(iii): Differentiating k times the identity

$$F(tx) = F(x) - \theta \ln t$$

in x, we get

$$t^k D^k F(tx)[h_1, ..., h_k] = D^k F(x)[h_1, ..., h_k].$$ ∎

An especially nice specific feature of the barriers S_k, L_k and K is their *self-duality*:

Proposition 4.3.2 *A canonical barrier, let it be denoted* F *(*F *can be either* S_k*, or* L_k*, or the direct sum* K *of several copies of these "elementary" barriers), possesses the following property: for every* $x \in \mathrm{Dom}\, F$*,* $-\nabla F(x)$ *belongs to* $\mathrm{Dom}\, F$ *as well, and the mapping* $x \mapsto -\nabla F(x) : \mathrm{Dom}\, F \to \mathrm{Dom}\, F$ *is self-inverse:*

$$-\nabla F(-\nabla F(x)) = x \quad \forall x \in \mathrm{Dom}\, F. \tag{4.3.3}$$

Besides this, the mapping $x \mapsto -\nabla F(x)$ *is homogeneous of degree -1:*

$$t > 0, x \in \mathrm{int}\,\mathrm{dom} F \Rightarrow -\nabla F(tx) = -t^{-1}\nabla F(x). \tag{4.3.4}$$

• In the SDP case, i.e., when $F = S_k$ and x is $k \times k$ semidefinite matrix, $\nabla F(x) = \nabla S_k(x) = -x^{-1}$, see (4.3.1), so that the above statements merely say that the mapping $x \mapsto x^{-1}$ is a self-inverse one-to-one mapping of the interior of the semidefinite cone onto itself, and that $-(tx)^{-1} = -t^{-1}x^{-1}$, both claims being trivially true.

4.4 Primal-dual pair of problems and primal-dual central path

4.4.1 The problem(s)

It makes sense to consider simultaneously the "problem of interest" (CP) and its conic dual; since $\mathbf{K}$ is a direct product of self-dual cones, this dual is a conic problem *on the same cone* $\mathbf{K}$. As we remember from Lecture 1, the primal-dual pair associated with (CP) is

$$\begin{array}{ll} \min\limits_x \left\{c^T x : \mathcal{A}x - B \in \mathbf{K}\right\} & \text{(CP)} \\ \max\limits_S \left\{\langle B, S\rangle_E : \mathcal{A}^* S = c,\ S \in \mathbf{K}\right\} & \text{(CD)} \end{array}$$

Assuming that the linear mapping $x \mapsto \mathcal{A}x$ is an embedding (i.e., that $\mathrm{Ker}\mathcal{A} = \{0\}$ – this is Assumption **A** from Lecture 1), we can write down our primal-dual pair in a symmetric geometric form (Lecture 1, Section 1.6.1):

$$\begin{array}{ll} \min\limits_X \left\{\langle C, X\rangle_E : X \in (\mathcal{L} - B) \cap \mathbf{K}\right\} & \text{(P)} \\ \max\limits_S \left\{\langle B, S\rangle_E : S \in (\mathcal{L}^\perp + C) \cap \mathbf{K}\right\} & \text{(D)} \end{array}$$

where $\mathcal{L}$ is a linear subspace in E (the image space of the linear mapping $x \mapsto \mathcal{A}x$), $\mathcal{L}^\perp$ is the orthogonal complement to $\mathcal{L}$ in E, and $C \in E$ satisfies $\mathcal{A}^* C = c$, i.e., $\langle C, \mathcal{A}x\rangle_E \equiv c^T x$.

To simplify things, *from now on we assume that both problems* (CP) *and* (CD) *are strictly feasible*. In terms of (P) and (D) this assumption means that both the primal feasible plane $\mathcal{L} - B$ and the dual feasible plane $\mathcal{L}^\perp + C$ intersect the interior of the cone $\mathbf{K}$.

Remark 4.4.1 *By the Conic Duality Theorem (Lecture 1), both* (CP) *and* (D) *are solvable with equal optimal values:*

$$\mathrm{Opt(CP)} = \mathrm{Opt(D)}$$

(recall that we have assumed strict primal-dual feasibility). Since (P) *is equivalent to* (CP), (P) *is solvable as well, and the optimal value of* (P) *differs from the one of* (P) *by* $\langle C, B\rangle_E$ [7]. *It follows that the optimal values of* (P) *and* (D) *are linked by the relation*

$$\mathrm{Opt(P)} - \mathrm{Opt(D)} + \langle C, B\rangle_E = 0. \tag{4.4.1}$$

4.4.2 The central path(s)

The canonical barrier K of $\mathbf{K}$ induces a barrier for the feasible set $\mathcal{X} = \{x \mid \mathcal{A}x - B \in \mathbf{K}\}$ of the problem (CP) written down in the form of (C), i.e., as

$$\min_x \left\{c^T x : x \in \mathcal{X}\right\};$$

this barrier is

$$\widehat{K}(x) = K(\mathcal{A}x - B) : \mathrm{int}\, X \to \mathbf{R} \tag{4.4.2}$$

[7] Indeed, the values of the respective objectives $c^T x$ and $\langle C, \mathcal{A}x - B\rangle_E$ at the corresponding to each other feasible solutions x of (CP) and $X = \mathcal{A}x - B$ of (P) differ from each other by exactly $\langle C, B\rangle_E$:

$$c^T x - \langle C, X\rangle_E = c^T x - \langle C, \mathcal{A}x - B\rangle_E = \underbrace{c^T x - \langle \mathcal{A}^* C, x\rangle_E}_{=0 \text{ due to } \mathcal{A}^* C = c} + \langle C, B\rangle_E.$$

and is indeed a barrier. Now we can apply the interior penalty scheme to trace the central path $x_*(t)$ associated with the resulting barrier; with some effort it can be derived from the primal-dual strict feasibility that this central path is well-defined (i.e., that the minimizer of

$$\widehat{K}_t(x) = tc^T x + \widehat{K}(x)$$

on $\operatorname{int} X$ exists for every $t \geq 0$ and is unique)[8)]. What is important for us for the moment, is the central path itself, not how to trace it. Moreover, it is highly instructive to pass from the central path $x_*(t)$ in the space of design variables to its image

$$X_*(t) = \mathcal{A}x_*(t) - B$$

in E. The resulting curve has a name – it is called *the primal central path of the primal-dual pair* (P), (D); by its origin, it is a curve comprised of *strictly feasible* solutions of (P) (since it is the same – to say that x belongs to the (interior of) the set $\mathcal{X}$ and to say that $X = \mathcal{A}x - B$ is a (strictly) feasible solution of (P)). A simple and very useful observation is that the primal central path can be defined solely in terms of (P), (D) and thus is a "geometric entity" – it is independent of a particular parameterization of the primal feasible plane $\mathcal{L} - B$ by the design vector x:

(*) *A point $X_*(t)$ of the primal central path is the minimizer of the aggregate*

$$P_t(X) = t\langle C, X\rangle_E + K(X)$$

on the set $(\mathcal{L} - B) \cap \operatorname{int} \mathbf{K}$ of strictly feasible solutions of (P).

This observation is just a tautology: $x_*(t)$ is the minimizer on $\operatorname{int} \mathcal{X}$ of the aggregate

$$\widehat{K}_t(x) \equiv tc^T x + \widehat{K}(x) = t\langle C, \mathcal{A}x\rangle_E + K(\mathcal{A}x - B) = P_t(\mathcal{A}x - B) + t\langle C, B\rangle_E;$$

we see that the function $\widehat{P}_t(x) = P_t(\mathcal{A}x - B)$ of $x \in \operatorname{int} \mathcal{X}$ differs from the function $\widehat{K}_t(x)$ by a constant (depending on t) and has therefore the same minimizer $x_*(t)$ as the function $\widehat{K}_t(x)$. Now, when x runs through $\operatorname{int} \mathcal{X}$, the point $X = \mathcal{A}x - B$ runs exactly through the set of strictly feasible solutions of (P), so that the minimizer X_* of P_t on the latter set and the minimizer $x_*(t)$ of the function $\widehat{P}_t(x) = P_t(\mathcal{A}x - B)$ on $\operatorname{int} \mathcal{X}$ are linked by the relation $X_* = \mathcal{A}x_*(t) - B$.

The "analytic translation" of the above observation is as follows:

(*′) *A point $X_*(t)$ of the primal central path is exactly the strictly feasible solution X to* (P) *such that the vector $tC + \nabla K(X) \in E$ is orthogonal to $\mathcal{L}$ (i.e., belongs to $\mathcal{L}^\perp$).*

Indeed, we know that $X_*(t)$ is the unique minimizer of the smooth convex function $P_t(X) = t\langle C, X\rangle_E + K(X)$ on the intersection of the primal feasible plane $\mathcal{L} - B$ and the interior of the cone $\mathbf{K}$; a necessary and sufficient condition for a point X of this intersection to minimize P_t over the intersection is that ∇P_t must be orthogonal to $\mathcal{L}$.

• In the SDP case, a point $X_*(t)$, $t > 0$, of the primal central path is uniquely defined by the following two requirements: (1) $X_*(t) \succ 0$ should be feasible for (P), and (2) the $k \times k$ matrix

$$tC - X_*^{-1}(t) = tC + \nabla S_k(X_*(t))$$

(see (4.3.1)) should belong to $\mathcal{L}^\perp$, i.e., should be orthogonal, w.r.t. the Frobenius inner product, to every matrix of the form $\mathcal{A}x$.

[8)] In Section 4.2.1, there was no problem with the existence of the central path, since there $\mathcal{X}$ was assumed to be bounded; in our now context, $\mathcal{X}$ not necessarily is bounded.

The dual problem (D) is in no sense "worse" than the primal problem (P) and thus also possesses the central path, now called *the dual central path* $S_*(t)$, $t \geq 0$, of the primal-dual pair (P), (D). Similarly to (*), (*′), the dual central path can be characterized as follows:

(**′) *A point $S_*(t)$, $t \geq 0$, of the dual central path is the unique minimizer of the aggregate*

$$D_t(S) = -t\langle B, S\rangle_E + K(S)$$

on the set of strictly feasible solutions of (D) [9]. *$S_*(t)$ is exactly the strictly feasible solution S to* (D) *such that the vector $-tB + \nabla F(S)$ is orthogonal to $\mathcal{L}^\perp$ (i.e., belongs to $\mathcal{L}$).*

• In the SDP case, a point $S_*(t)$, $t > 0$, of the dual central path is uniquely defined by the following two requirements: (1) $S_*(t) \succ 0$ should be feasible for (D), and (2) the $k \times k$ matrix

$$-tB - S_*^{-1}(t) = -tB + \nabla S_k(S_*(t))$$

(see (4.3.1)) should belong to $\mathcal{L}$, i.e., should be representable in the form $\mathcal{A}x$ for some x.

From Proposition 4.3.2 we can derive a wonderful connection between the primal and the dual central paths:

Theorem 4.4.1 *For $t > 0$, the primal and the dual central paths $X_*(t)$, $S_*(t)$ of a (strictly feasible) primal-dual pair* (P), (D) *are linked by the relations*

$$\begin{array}{l} S_*(t) = -t^{-1}\nabla K(X_*(t)) \\ X_*(t) = -t^{-1}\nabla K(S_*(t)) \end{array} \tag{4.4.3}$$

Proof. By (*′), the vector $tC + \nabla K(X_*(t))$ belongs to $\mathcal{L}^\perp$, so that the vector $S = -t^{-1}\nabla K(X_*(t))$ belongs to the dual feasible plane $\mathcal{L}^\perp + C$. On the other hand, by Proposition 4.4.3 the vector $-\nabla K(X_*(t))$ belongs to $\mathrm{Dom}\, K$, i.e., to the interior of $\mathbf{K}$; since $\mathbf{K}$ is a cone and $t > 0$, the vector $S = -t^{-1}\nabla F(X_*(t))$ belongs to the interior of $\mathbf{K}$ as well. Thus, S is a strictly feasible solution of (D). Now let us compute the gradient of the aggregate D_t at the point S:

$$\begin{array}{rcl} \nabla D_t(S) & = & -tB + \nabla K(-t^{-1}\nabla K(X_*(t))) \\ & = & -tB + t\nabla K(-\nabla K(X_*(t))) \\ & & \qquad \text{[we have used (4.3.4)]} \\ & = & -tB - tX_*(t) \\ & & \qquad \text{[we have used (4.3.3)]} \\ & = & -t(B + X_*(t)) \\ & \in & \mathcal{L} \\ & & \text{[since } X_*(t) \text{ is primal feasible]} \end{array}$$

Thus, S is strictly feasible for (D) and $\nabla D_t(S) \in \mathcal{L}$. But by (**′) these properties *characterize* $S_*(t)$; thus, $S_*(t) = S \equiv -t^{-1}\nabla K(X_*(t))$. This relation, in view of Proposition 4.3.2, implies that $X_*(t) = -t^{-1}\nabla K(S_*(t))$. Another way to get the latter relation from the one $S_*(t) = -t^{-1}\nabla K(X_*(t))$ is just to refer to the primal-dual symmetry. ∎

In fact, the connection between the primal and the dual central paths stated by Theorem 4.4.1 can be used to characterize both the paths:

Theorem 4.4.2 *Let* (P), (D) *be a strictly feasible primal-dual pair.*

For every $t > 0$, there exists a unique strictly feasible solution X of (P) *such that $-t^{-1}\nabla K(X)$ is a feasible solution to* (D), *and this solution X is exactly $X_*(t)$.*

Similarly, for every $t > 0$, there exists a unique strictly feasible solution S of (D) *such that $-t^{-1}\nabla K(S)$ is a feasible solution of* (P), *and this solution S is exactly $S_*(t)$.*

[9] Note the slight asymmetry between the definitions of the primal aggregate P_t and the dual aggregate D_t: in the former, the linear term is $t\langle C, X\rangle_E$, while in the latter it is $-t\langle B, S\rangle_E$. This asymmetry is in complete accordance with the fact that we write (P) as a minimization, and (D) – as a maximization problem; to write (D) in *exactly* the same form as (P), we were supposed to replace B with $-B$, thus getting the formula for D_t completely similar to the one for P_t.

Proof. By primal-dual symmetry, it suffices to prove the first claim. We already know (Theorem 4.4.1) that $X = X_*(t)$ is a strictly feasible solution of (P) such that $-t^{-1}\nabla K(X)$ is feasible for (D); all we need to prove is that $X_*(t)$ is the only point with these properties, which is immediate: if X is a strictly feasible solution of (P) such that $-t^{-1}\nabla K(X)$ is dual feasible, then $-t^{-1}\nabla K(X) \in \mathcal{L}^\perp + C$, or, which is the same, $\nabla K(X) \in \mathcal{L}^\perp - tC$, or, which again is the same, $\nabla P_t(X) = tC + \nabla K(X) \in \mathcal{L}^\perp$. And we already know from $(*')$ that the latter property, taken together with the strict primal feasibility, is characteristic for $X_*(t)$. ∎

On the central path

As we have seen, the primal and the dual central paths are intrinsically linked one to another, and it makes sense to think of them as of a unique entity – the *primal-dual* central path of the primal-dual pair (P), (D). The primal-dual central path is just a curve $(X_*(t), S_*(t))$ in $E \times E$ such that the projection of the curve on the primal space is the primal central path, and the projection of it on the dual space is the dual central path.
To save words, from now on we refer to the primal-dual central path simply as to the *central path*.

The central path possesses a number of extremely nice properties; let us list some of them.

Characterization of the central path. By Theorem 4.4.2, the points $(X_*(t), S_*(t))$ of the central path possess the following properties:

(CentralPath):

1. [Primal feasibility] *The point $X_*(t)$ is strictly primal feasible.*
2. [Dual feasibility] *The point $S_*(t)$ is dual feasible.*
3. ["Augmented complementary slackness"] The points $X_*(t)$ and $S_*(t)$ are linked by the relation
$$S_*(t) = -t^{-1}\nabla K(X_*(t)) \quad [\Leftrightarrow X_*(t) = -t^{-1}\nabla K(S_*(t))].$$
• In the SDP case, $\nabla K(U) = \nabla S_k(U) = -U^{-1}$ (see (4.3.1)), and the augmented complementary slackness relation takes the nice form
$$X_*(t)S_*(t) = t^{-1}I, \tag{4.4.4}$$
where I, as usual, is the unit matrix.

In fact, the indicated properties *fully characterize* the central path: whenever two points X, S possess the properties 1) - 3) with respect to some $t > 0$, X is nothing but $X_*(t)$, and S is nothing but $S_*(t)$ (this again is said by Theorem 4.4.2).

Duality gap along the central path. Recall that for an arbitrary primal-dual feasible pair (X, S) of the (strictly feasible!) primal-dual pair of problems (P), (D), the duality gap
$$\text{DualityGap}(X, S) \equiv [\langle C, X\rangle_E - \text{Opt(P)}] + [\text{Opt(D)} - \langle B, S\rangle_E] = \langle C, X\rangle_E - \langle B, S\rangle_E + \langle C, B\rangle_E$$
(see (4.4.1)) which measures the "total inaccuracy" of X, S as approximate solutions of the respective problems, can be written down equivalently as $\langle S, X\rangle_E$ (see statement (!) in Section 1.7). Now, what is the duality gap along the central path? The answer is immediate:
$$\begin{array}{rcl}\text{DualityGap}(X_*(t), S_*(t)) & = & \langle S_*(t), X_*(t)\rangle_E \\ & = & \langle -t^{-1}\nabla K(X_*(t)), X_*(t)\rangle_E \\ & & \qquad \text{[see (4.4.3)]} \\ & = & t^{-1}\theta(K) \\ & & \text{[see Proposition 4.3.1.(ii)]}\end{array}$$

We have arrived at a wonderful result[10]:

Proposition 4.4.1 *Under assumption of primal-dual strict feasibility, the duality gap along the central path is inverse proportional to the penalty parameter, the proportionality coefficient being the parameter of the canonical barrier K:*

$$\mathrm{DualityGap}(X_*(t), S_*(t)) = \frac{\theta(K)}{t}.$$

In particular, both $X_*(t)$ and $S_*(t)$ are strictly feasible $\left(\frac{\theta(K)}{t}\right)$-approximate solutions to their respective problems:

$$\begin{array}{rcl} \langle C, X_*(t)\rangle_E - \mathrm{Opt}(P) & \leq & \frac{\theta(K)}{t}, \\ \mathrm{Opt}(D) - \langle B, S_*(t)\rangle_E & \leq & \frac{\theta(K)}{t}. \end{array}$$

- In the SDP case, $\mathbf{K} = \mathbf{S}^k_+$ and $\theta(K) = \theta(S_k) = k$.

We see that

> *All we need in order to get "quickly" good primal and dual approximate solutions, is to trace fast the central path; if we were interested to solve only one of the problems (P), (D), it would be sufficient to trace fast the associated – primal or dual – component of this path. The quality guarantees we get in such a process depend – in a completely universal fashion! – solely on the value t of the penalty parameter we have managed to achieve and on the value of the parameter of the canonical barrier K and are completely independent of other elements of the data.*

Near the central path

The conclusion we have just made is a bit too optimistic: well, our life when moving along the central path would be just fine (at the very least, we would know how good are the solutions we already have), but how could we move *exactly* along the path? Among the relations (CentralPath.1-3) defining the path the first two are "simple" – just linear, but the third is in fact a system of *nonlinear* equations, and we have no hope to satisfy these equations exactly. Thus, we arrive at the crucial question which, a bit informally, sounds as follows:

> *How close (and in what sense close) should we be to the path in order for our life to be essentially as nice as if we were exactly on the path?*

There are several ways to answer this question; we will present the simplest one.

A distance to the central path. Our canonical barrier $K(\cdot)$ is a strongly convex smooth function on $\mathrm{int}\,\mathbf{K}$; in particular, its Hessian matrix $\nabla^2 K(Y)$, taken at a point $Y \in \mathrm{int}\,\mathbf{K}$, is positive definite. We can use *the inverse* of this matrix to measure the distances between points of E, thus arriving at the norm

$$\|H\|_Y = \sqrt{\langle [\nabla^2 K(Y)]^{-1} H, H\rangle_E}.$$

It turns out that

> *A good measure of proximity of a strictly feasible primal-dual pair $Z = (X, S)$ to a point $Z_*(t) = (X_*(t), S_*(t))$ from the primal-dual central path is the quantity*
>
> $$\mathrm{dist}(Z, Z_*(t)) \equiv \|tS + \nabla K(X)\|_X \equiv \sqrt{\langle [\nabla^2 K(X)]^{-1}(tS + \nabla K(X)), tS + \nabla K(X)\rangle_E}$$

[10] Which, among other, much more important consequences, explains the name "augmented complementary slackness" of the property 1^0.3): at the primal-dual pair of *optimal* solutions X^*, S^* the duality gap should be zero: $\langle S^*, X^*\rangle_E = 0$. Property 1^0.3, as we just have seen, implies that the duality gap at a primal-dual pair $(X_*(t), S_*(t))$ from the central path, although nonzero, is "controllable" – $\frac{\theta(K)}{t}$ – and becomes small as t grows.

Although written in a non-symmetric w.r.t. X, S form, this quantity is in fact symmetric in X, S: it turns out that

$$\|tS + \nabla K(X)\|_X = \|tX + \nabla K(S)\|_S \tag{4.4.5}$$

for all $t > 0$ and $S, X \in \operatorname{int} \mathbf{K}$.

Observe that $\operatorname{dist}(Z, Z_*(t)) \geq 0$, and $\operatorname{dist}(Z, Z_*(t)) = 0$ if and only if $S = -t^{-1}\nabla K(X)$, which, for a strictly primal-dual feasible pair $Z = (X, S)$, means that $Z = Z_*(t)$ (see the characterization of the primal-dual central path); thus, $\operatorname{dist}(Z, Z_*(t))$ indeed can be viewed as a kind of distance from Z to $Z_*(t)$.

In the SDP case X, S are $k \times k$ symmetric matrices, and

$$\begin{array}{rcll} \operatorname{dist}^2(Z, Z_*(t)) & = & \|tS + \nabla S_k(X)\|_X^2 = \langle [\nabla^2 S_k(X)]^{-1}(tS + \nabla S_k(X)), tS + \nabla S_k(X)\rangle_F & \\ & = & \operatorname{Tr}\left(X(tS - X^{-1})X(tS - X^{-1})\right) & \\ & & & [\text{see (4.3.1)}] \\ & = & \operatorname{Tr}([tX^{1/2}SX^{1/2} - I]^2), & \end{array}$$

so that

$$\operatorname{dist}^2(Z, Z_*(t)) = \operatorname{Tr}\left(X(tS - X^{-1})X(tS - X^{-1})\right) = \|tX^{1/2}SX^{1/2} - I\|_2^2. \tag{4.4.6}$$

Besides this,

$$\begin{array}{rcl} \|tX^{1/2}SX^{1/2} - I\|_2^2 & = & \operatorname{Tr}\left([tX^{1/2}SX^{1/2} - I]^2\right) \\ & = & \operatorname{Tr}\left(t^2X^{1/2}SX^{1/2}X^{1/2}SX^{1/2} - 2tX^{1/2}SX^{1/2} + I\right) \\ & = & \operatorname{Tr}(t^2X^{1/2}SXSX^{1/2}) - 2t\operatorname{Tr}(X^{1/2}SX^{1/2}) + \operatorname{Tr}(I) \\ & = & \operatorname{Tr}(t^2XSXS - 2tXS + I) \\ & = & \operatorname{Tr}(t^2SXSX - 2tSX + I) \\ & = & \operatorname{Tr}(t^2S^{1/2}XS^{1/2}S^{1/2}XS^{1/2} - 2tS^{1/2}XS^{1/2} + I) \\ & = & \operatorname{Tr}([tS^{1/2}XS^{1/2} - I]^2), \end{array}$$

i.e., (4.4.5) indeed is true.

In a moderate $\operatorname{dist}(\cdot, Z_*(\cdot))$-neighbourhood of the central path. It turns out that in such a neighbourhood all is essentially as fine as at the central path itself:

A. *Whenever $Z = (X, S)$ is a pair of primal-dual strictly feasible solutions to* (P), (D) *such that*

$$\operatorname{dist}(Z, Z_*(t)) \leq 1, \tag{Close}$$

Z is "essentially as good as $Z_*(t)$", namely, the duality gap at (X, S) is essentially as small as at the point $Z_*(t)$:

$$\operatorname{DualityGap}(X, S) = \langle S, X\rangle_E \leq 2\operatorname{DualityGap}(Z_*(t)) = \frac{2\theta(K)}{t}. \tag{4.4.7}$$

Let us check **A** in the SDP case. Let (t, X, S) satisfy the premise of **A**. The duality gap at the pair (X, S) of strictly primal-dual feasible solutions is

$$\operatorname{DualityGap}(X, S) = \langle X, S\rangle_F = \operatorname{Tr}(XS),$$

while by (4.4.6) the relation $\operatorname{dist}((S, X), Z_*(t)) \leq 1$ means that

$$\|tX^{1/2}SX^{1/2} - I\|_2 \leq 1,$$

whence

$$\|X^{1/2}SX^{1/2} - t^{-1}I\|_2 \leq \frac{1}{t}.$$

Denoting by δ the vector of eigenvalues of the symmetric matrix $X^{1/2}SX^{1/2}$, we conclude that $\sum_{i=1}^{k}(\delta_i - t^{-1})^2 \le t^{-2}$, whence

$$\begin{aligned}\text{DualityGap}(X,S) &= \text{Tr}(XS) = \text{Tr}(X^{1/2}SX^{1/2}) = \sum_{i=1}^{k}\delta_i \\ &\le kt^{-1} + \sum_{i=1}^{k}|\delta_i - t^{-1}| \le kt^{-1} + \sqrt{k}\sqrt{\sum_{i=1}^{k}(\delta_i - t^{-1})^2} \\ &\le kt^{-1} + \sqrt{k}t^{-1},\end{aligned}$$

and (4.4.7) follows.

It follows from **A** that

For our purposes, it is essentially the same – to move along the primal-dual central path, or to trace this path, staying in its "time-space" neighbourhood

$$\mathcal{N}_\kappa = \{(t,X,S) \mid X \in \mathcal{L} - B, S \in \mathcal{L}^\perp + C, t > 0, \text{dist}((X,S),(X_*(t),S_*(t))) \le \kappa\} \quad (4.4.8)$$

with certain $\kappa \le 1$.

Most of the interior point methods for LP, CQP, and SDP, including those most powerful in practice, solve the primal-dual pair (P), (D) by tracing the central path[11)], although not all of them keep the iterates in $\mathcal{N}_{O(1)}$; some of the methods work in much wider neighbourhoods of the central path, in order to avoid slowing down when passing "highly curved" segments of the path. At the level of ideas, these "long step path following methods" essentially do not differ from the "short step" ones – those keeping the iterates in $\mathcal{N}_{O(1)}$; this is why in the analysis part of our forthcoming presentation we restrict ourselves with the short-step methods. It should be added that as far as the theoretical efficiency estimates are concerned, the short-step methods yield the best known so far complexity bounds for LP, CQP and SDP, and are essentially better than the long-step methods (although in practice the long-step methods usually outperform their short-step counterparts).

4.5 Tracing the central path

4.5.1 The path-following scheme

Assume we are solving a strictly feasible primal-dual pair of problems (P), (D) and intend to trace the associated central path. Essentially all we need is a mechanism for updating a current iterate $(\bar{t}, \bar{X}, \bar{S})$ such that $\bar{t} > 0$, $\bar{X}$ is strictly primal feasible, $\bar{S}$ is strictly dual feasible, and $(\bar{X}, \bar{S})$ is "good", in certain precise sense, approximation of the point $Z_*(\bar{t}) = (X_*(\bar{t}), S_*(\bar{t}))$ on the central path, into a new iterate (t_+, X_+, S_+) with similar properties and a larger value $t_+ > \bar{t}$ of the penalty parameter. Given such an updating and iterating it, we indeed shall trace the central path, with all the benefits (see above) coming from the latter fact[12)] How could we construct the required updating? Recalling the description of the central path, we see that our question is:

[11)] There exist also *potential reduction* interior point methods which do not take explicit care of tracing the central path; an example is the very first IP method for LP – the method of Karmarkar. The potential reduction IP methods are beyond the scope of our course, which is not a big loss for a practically oriented reader, since, as a practical tool, these methods are thought of to be obsolete.

[12)] Of course, besides knowing how to trace the central path, we should also know how to initialize this process – how to come close to the path to be able to start its tracing. There are different techniques to resolve this "initialization difficulty", and basically all of them achieve the goal by using the same path-tracing technique, now applied to an appropriate auxiliary problem where the "initialization difficulty" does not arise at all. Thus, at the level of ideas the initialization techniques do not add something essentially new, which allows us to skip in our presentation all initialization-related issues.

Given a triple $(\bar{t}, \bar{X}, \bar{S})$ which satisfies the relations

$$\begin{array}{rcl} X & \in & \mathcal{L} - B, \\ S & \in & \mathcal{L}^\perp + C \end{array} \tag{4.5.1}$$

(which is in fact a system of linear equations) *and approximately satisfies the system of nonlinear equations*

$$G_t(X, S) \equiv S + t^{-1}\nabla K(X) = 0, \tag{4.5.2}$$

update it into a new triple (t_+, X_+, S_+) with the same properties and $t_+ > \bar{t}$.

Since the left hand side $G(\cdot)$ in our system of nonlinear equations is smooth around $(\bar{t}, \bar{X}, \bar{S})$ (recall that $\bar{X}$ was assumed to be strictly primal feasible), the most natural, from the viewpoint of Computational Mathematics, way to achieve our target is as follows:

1. We choose somehow a desired new value $t_+ > \bar{t}$ of the penalty parameter;
2. We linearize the left hand side $G_{t_+}(X, S)$ of the system of nonlinear equations (4.5.2) at the point $(\bar{X}, \bar{S})$, and replace (4.5.2) with the linearized system of equations

$$G_{t_+}(\bar{X}, \bar{S}) + \frac{\partial G_{t_+}(\bar{X}, \bar{S})}{\partial X}(X - \bar{X}) + \frac{\partial G_{t_+}(\bar{X}, \bar{S})}{\partial S}(S - \bar{S}) = 0 \tag{4.5.3}$$

3. We define the corrections ΔX, ΔS from the requirement that the updated pair $X_+ = \bar{X} + \Delta X$, $S_+ = \bar{S} + \Delta S$ must satisfy (4.5.1) and the linearized version (4.5.3) of (4.5.2). In other words, the corrections should solve the system

$$\begin{array}{c} \Delta X \in \mathcal{L}, \\ \Delta S \in \mathcal{L}^\perp, \\ G_{t_+}(\bar{X}, \bar{S}) + \frac{\partial G_{t_+}(\bar{X}, \bar{S})}{\partial X}\Delta X + \frac{\partial G_{t_+}(\bar{X}, \bar{S})}{\partial S}\Delta S = 0 \end{array} \tag{4.5.4}$$

4. Finally, we define X_+ and S_+ as

$$\begin{array}{rcl} X_+ & = & \bar{X} + \Delta X, \\ S_+ & = & \bar{S} + \Delta S. \end{array} \tag{4.5.5}$$

The primal-dual IP methods we are describing basically fit the outlined scheme, up to the following two important points:

- If the current iterate $(\bar{X}, \bar{S})$ is not enough close to $Z_*(\bar{t})$, and/or if the desired improvement $t_+ - \bar{t}$ is too large, the corrections given by the outlined scheme may be too large; as a result, the updating (4.5.5) as it is may be inappropriate, e.g., X_+, or S_+, or both, may be kicked out of the cone $\mathbf{K}$. (Why not: linearized system (4.5.3) approximates well the "true" system (4.5.2) only locally, and we have no reasons to trust in corrections coming from the linearized system, when these corrections are large.)

  There is a standard way to overcome the outlined difficulty – to use the corrections in a damped fashion, namely, to replace the updating (4.5.5) with

$$\begin{array}{rcl} X_+ & = & \bar{X} + \alpha\Delta X, \\ S_+ & = & \bar{S} + \beta\Delta S, \end{array} \tag{4.5.6}$$

  and to choose the *stepsizes* $\alpha > 0, \beta > 0$ from additional "safety" considerations, like ensuring the updated pair (X_+, S_+) to reside in the interior of $\mathbf{K}$, or enforcing it to stay in a desired neighbourhood of the central path, or whatever else. In IP methods, the solution $(\Delta X, \Delta S)$ of (4.5.4) plays the role of *search direction* (and this is how it is called), and the actual corrections are proportional to the search ones rather than to be exactly the same. In this sense the situation is completely similar to the one with the Newton method from Section 4.2.2 (which is natural: the latter method is *exactly* the linearization method for solving the Fermat equation $\nabla f(x) = 0$).

- The "augmented complementary slackness" system (4.5.2) can be written down in many different forms which are equivalent to each other in the sense that they share a common solution set. E.g., we have the same reasons to express the augmented complementary slackness requirement by the nonlinear system (4.5.2) as to express it by the system

$$\widehat{G}_t(X,S) \equiv X + t^{-1}\nabla K(S) = 0,$$

not speaking about other possibilities. And although all systems of nonlinear equations

$$H_t(X,S) = 0$$

expressing the augmented complementary slackness are "equivalent" in the sense that they share a common solution set, their linearizations are different and thus – lead to different search directions and finally to different path-following methods. Choosing appropriate (in general even varying from iteration to iteration) analytic representation of the augmented complementary slackness requirement, one can gain a lot in the performance of the resulting path-following method, and the IP machinery facilitates this flexibility (see "SDP case examples" below).

4.5.2 Speed of path-tracing

In the LP-CQP-SDP situation, the speed at which the best, from the theoretical viewpoint, path-following methods manage to trace the path, is inverse proportional to the *square root* of the parameter $\theta(K)$ of the underlying canonical barrier. It means the following. Started at a point (t^0, X^0, S^0) from the neighbourhood $\mathcal{N}_{0.1}$ of the central path, the method after $O(1)\sqrt{\theta(K)}$ steps reaches the point $(t^1 = 2t^0, X^1, S^1)$ from the same neighbourhood, after the same $O(1)\sqrt{\theta(K)}$ steps more reaches the point $(t^2 = 2^2t^0, X^2, S^2)$ from the neighbourhood, and so on – it takes the method a *fixed number* $O(1)\sqrt{\theta(K)}$ steps to increase by factor 2 the current value of the penalty parameter, staying all the time in $\mathcal{N}_{0.1}$. By (4.4.7) it means that every $O(1)\sqrt{\theta(K)}$ steps of the method reduce the (upper bound on the) inaccuracy of current approximate solutions by factor 2, or, which is the same, add a fixed number of accuracy digits to these solutions. Thus, "the cost of an accuracy digit" for the (best) path-following methods is $O(1)\sqrt{\theta(K)}$ steps. To realize what this indeed mean, we should, of course, know how "heavy" a step is – what is its arithmetic cost. Well, the arithmetic cost of a step for the "cheapest among the fastest" IP methods as applied to (CP) is *as if* all operations carried out at a step were those required by

1. Assembling, given a point $X \in \operatorname{int}\mathbf{K}$, the symmetric $n \times n$ matrix ($n = \dim x$)

$$\mathcal{H} = \mathcal{A}^*[\nabla^2 K(X)]\mathcal{A};$$

2. Subsequent Choleski factorization of the matrix $\mathcal{H}$ (which, due to its origin, is symmetric positive definite and thus admits Choleski decomposition $\mathcal{H} = DD^T$ with lower triangular D).

Looking at (Cone), (CP) and (4.3.1), we immediately conclude that the arithmetic cost of assembling and factorizing $\mathcal{H}$ is polynomial in the size $\dim \mathrm{Data}(\cdot)$ of the data defining (CP), and that the parameter $\theta(K)$ also is polynomial in this size. Thus, the cost of an accuracy digit for the methods in question is polynomial in the size of the data, as is required from polynomial time methods[13)]. Explicit complexity bounds for $\mathcal{LP}_b$, $\mathcal{CQP}_b$, $\mathcal{SDP}_b$ are given in Sections 4.6.1, 4.6.2, 4.6.3, respectively.

4.5.3 The primal and the dual path-following methods

The simplest way to implement the path-following scheme from Section 4.5.1 is to linearize the augmented complementary slackness equations (4.5.2) as they are, ignoring the option to rewrite these equations

[13)] Strictly speaking, the outlined complexity considerations are applicable to the "highway" phase of the solution process, after we once have reached the neighbourhood $\mathcal{N}_{0.1}$ of the central path. However, the results of our considerations remain unchanged after the initialization expenses are taken into account, see Section 4.6.

equivalently before linearization. Let us look at the resulting method in more details. Linearizing (4.5.2) at a current iterate $\bar{X}$, $\bar{S}$, we get the vector equation

$$t_+(\bar{S}+\Delta S)+\nabla K(\bar{X})+[\nabla^2 K(\bar{X})]\Delta X=0,$$

where t_+ is the target value of the penalty parameter. The system (4.5.4) now becomes

$$\begin{array}{ll}(a) & \Delta X\in\mathcal{L}\\ & \Updownarrow\\ (a') & \Delta X=\mathcal{A}\Delta x\quad[\Delta x\in\mathbf{R}^n]\\ (b) & \Delta S\in\mathcal{L}^\perp\\ & \Updownarrow\\ (b') & \mathcal{A}^*\Delta S=0\\ (c) & t_+[\bar{S}+\Delta S]+\nabla K(\bar{X})+[\nabla^2 K(\bar{X})]\Delta X=0;\end{array}\qquad(4.5.7)$$

the unknowns here are $\Delta X,\Delta S$ and Δx. To process the system, we eliminate ΔX via (a') and multiply both sides of (c) by $\mathcal{A}^*$, thus getting the equation

$$\underbrace{\mathcal{A}^*[\nabla^2 K(\bar{X})]\mathcal{A}}_{\mathcal{H}}\Delta x+[t_+\mathcal{A}^*[\bar{S}+\Delta S]+\mathcal{A}^*\nabla K(\bar{X})]=0.\qquad(4.5.8)$$

Note that $\mathcal{A}^*[\bar{S}+\Delta S]=c$ is the objective of (CP) (indeed, $\bar{S}\in\mathcal{L}^\perp+C$, i.e., $\mathcal{A}^*\bar{S}=c$, while $\mathcal{A}^*\Delta S=0$ by (b')). Consequently, (4.5.8) becomes the *primal Newton system*

$$\mathcal{H}\Delta x=-[t_+c+\mathcal{A}^*\nabla K(\bar{X})].\qquad(4.5.9)$$

Solving this system (which is possible – it is easily seen that the $n\times n$ matrix $\mathcal{H}$ is positive definite), we get Δx and then set

$$\begin{array}{rcl}\Delta X & = & \mathcal{A}\Delta x,\\ \Delta S & = & -t_+^{-1}[\nabla K(\bar{X})+[\nabla^2 K(\bar{X})\Delta X]-\bar{S},\end{array}\qquad(4.5.10)$$

thus getting a solution to (4.5.7). Restricting ourselves with the stepsizes $\alpha=\beta=1$ (see (4.5.6)), we come to the "closed form" description of the method:

$$\begin{array}{ll}(a) & t\mapsto t_+>t\\ (b) & x\mapsto x_+=x+\underbrace{\left(-[\mathcal{A}^*(\nabla^2 K(X))\mathcal{A}]^{-1}[t_+c+\mathcal{A}^*\nabla K(X)]\right)}_{\Delta x},\\ (c) & S\mapsto S_+=-t_+^{-1}[\nabla K(X)+[\nabla^2 K(X)]\mathcal{A}\Delta x],\end{array}\qquad(4.5.11)$$

where x is the current iterate in the space $\mathbf{R}^n$ of design variables and $X=\mathcal{A}x-B$ is its image in the space E.

The resulting scheme admits a quite natural explanation. Consider the function

$$F(x)=K(\mathcal{A}x-B);$$

you can immediately verify that this function is a barrier for the feasible set of (CP). Let also

$$F_t(x)=tc^Tx+F(x)$$

be the associated barrier-generated family of penalized objectives. Relation (4.5.11.b) says that the iterates in the space of design variables are updated according to

$$x\mapsto x_+=x-[\nabla^2 F_{t_+}(x)]^{-1}\nabla F_{t_+}(x),$$

i.e., the process in the space of design variables is exactly the process (4.2.1) from Section 4.2.3.

Note that (4.5.11) is, essentially, a purely primal process (this is where the name of the method comes from). Indeed, the dual iterates S,S_+ just do not appear in formulas for x_+,X_+, and in fact the dual solutions are no more than "shadows" of the primal ones.

Remark 4.5.1 *When constructing the primal path-following method, we have started with the augmented slackness equations in form* (4.5.2)*. Needless to say, we could start our developments with the same conditions written down in the "swapped" form*

$$X + t^{-1}\nabla K(S) = 0$$

as well, thus coming to what is called "dual path-following method". Of course, as applied to a given pair (P), (D)*, the dual path-following method differs from the primal one. However, the constructions and results related to the dual path-following method require no special care – they can be obtained from their "primal counterparts" just by swapping "primal" and "dual" entities.*

The complexity analysis of the primal path-following method can be summarized in the following

Theorem 4.5.1 *Let $0 < \chi \leq \kappa \leq 0.1$. Assume that we are given a starting point (t_0, x_0, S_0) such that $t_0 > 0$ and the point*

$$(X_0 = \mathcal{A}x_0 - B, S_0)$$

is κ-close to $Z_*(t_0)$:

$$\mathrm{dist}((X_0, S_0), Z_*(t_0)) \leq \kappa.$$

Starting with (t_0, x_0, X_0, S_0), let us iterate process (4.5.11) *equipped with the penalty updating policy*

$$t_+ = \left(1 + \frac{\chi}{\sqrt{\theta(K)}}\right) t \tag{4.5.12}$$

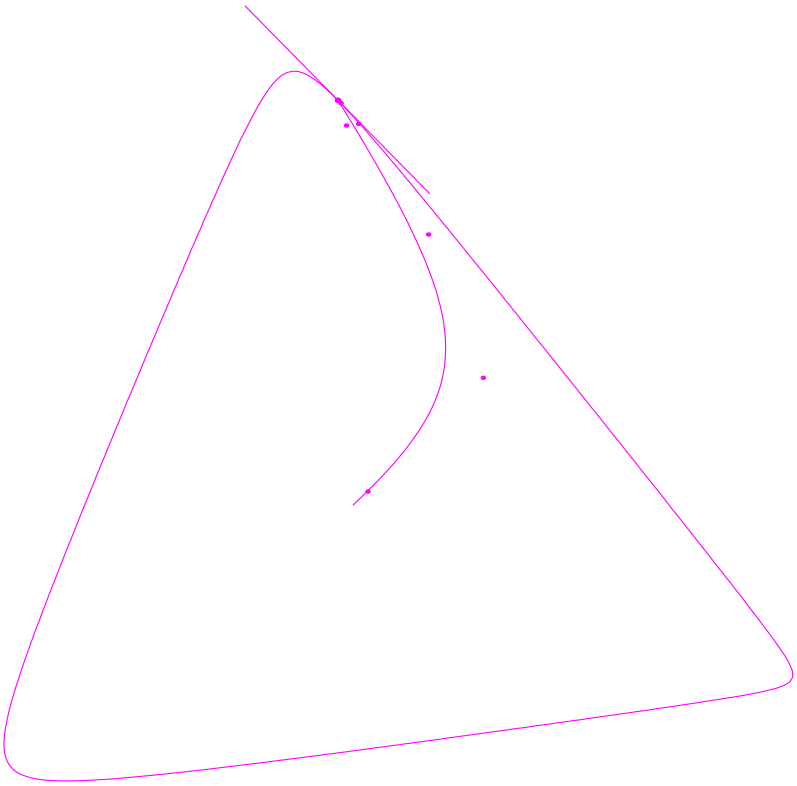
i.e., let us build the iterates (t_i, x_i, X_i, S_i) according to

$$\begin{aligned} t_i &= \left(1 + \tfrac{\chi}{\sqrt{\theta(K)}}\right) t_{i-1}, \\ x_i &= x_{i-1} - \underbrace{[\mathcal{A}^*(\nabla^2 K(X_{i-1}))\mathcal{A}]^{-1}[t_i c + \mathcal{A}^*\nabla K(X_{i-1})]}_{\Delta x_i}, \\ X_i &= \mathcal{A}x_i - B, \\ S_i &= -t_i^{-1}[\nabla K(X_{i-1}) + [\nabla^2 K(X_{i-1})]\mathcal{A}\Delta x_i] \end{aligned}$$

The resulting process is well-defined and generates strictly primal-dual feasible pairs (X_i, S_i) such that (t_i, X_i, S_i) stay in the neighbourhood $\mathcal{N}_\kappa$ of the primal-dual central path.

The theorem says that with properly chosen κ, χ (e.g., $\kappa = \chi = 0.1$) we can, getting once close to the primal-dual central path, trace it by the primal path-following method, keeping the iterates in $\mathcal{N}_\kappa$-neighbourhood of the path and increasing the penalty parameter by an absolute constant factor every $O(\sqrt{\theta(K)})$ steps – exactly as it was claimed in sections 4.2.3, 4.5.2. This fact is extremely important theoretically; in particular, it underlies the polynomial time complexity bounds for LP, CQP and SDP from Section 4.6 below. As a practical tool, the primal and the dual path-following methods, at least in their short-step form presented above, are not that attractive. The computational power of the methods can be improved by passing to appropriate large-step versions of the algorithms, but even these versions are thought of to be inferior as compared to "true" primal-dual path-following methods (those which "indeed work with both (P) and (D)", see below). There are, however, cases when the primal or the dual path-following scheme seems to be unavoidable; these are, essentially, the situations where the pair (P), (D) is "highly asymmetric", e.g., (P) and (D) have different by order of magnitudes design dimensions $\dim \mathcal{L}$, $\dim \mathcal{L}^\perp$. Here it becomes too expensive computationally to treat (P), (D) in a "nearly symmetric way", and it is better to focus solely on the problem with smaller design dimension.

To get an impression of how the primal path-following method works, here is a picture:



What you see is the 2D feasible set of a toy SDP ($\mathbf{K} = \mathbf{S}^3_+$). "Continuous curve" is the primal central path; dots are iterates x_i of the algorithm. We cannot draw the dual solutions, since they "live" in 4-dimensional space ($\dim \mathcal{L}^\perp = \dim \mathbf{S}^3 - \dim \mathcal{L} = 6 - 2 = 4$)

Here are the corresponding numbers:

Itr#	Objective	Duality Gap	Itr#	Objective	Duality Gap
1	-0.100000	2.96	7	-1.359870	8.4e-4
2	-0.906963	0.51	8	-1.360259	2.1e-4
3	-1.212689	0.19	9	-1.360374	5.3e-5
4	-1.301082	6.9e-2	10	-1.360397	1.4e-5
5	-1.349584	2.1e-2	11	-1.360404	3.8e-6
6	-1.356463	4.7e-3	12	-1.360406	9.5e-7

4.5.4 The SDP case

In what follows, we specialize the primal-dual path-following scheme in the SDP case and carry out its complexity analysis.

The path-following scheme in SDP

Let us look at the outlined scheme in the SDP case. Here the system of nonlinear equations (4.5.2) becomes (see (4.3.1))

$$G_t(X, S) \equiv S - t^{-1}X^{-1} = 0, \tag{4.5.13}$$

X, S being positive definite $k \times k$ symmetric matrices.

Recall that our generic scheme of a path-following IP method suggests, given a current triple $(\bar{t}, \bar{X}, \bar{S})$ with positive $\bar{t}$ and strictly primal, respectively, dual feasible $\bar{X}$ and $\bar{S}$, to update the this triple into a new triple (t_+, X_+, S_+) of the same type as follows:

(i) First, we somehow rewrite the system (4.5.13) as an equivalent system

$$\bar{G}_t(X,S)=0; \tag{4.5.14}$$

(ii) Second, we choose somehow a new value $t_+>\bar{t}$ of the penalty parameter and linearize system (4.5.14) (with t set to t_+) at the point $(\bar{X},\bar{S})$, thus coming to the system of linear equations

$$\frac{\partial\bar{G}_{t_+}(\bar{X},\bar{S})}{\partial X}\Delta X+\frac{\partial\bar{G}_{t_+}(\bar{X},\bar{S})}{\partial S}\Delta S=-\bar{G}_{t_+}(\bar{X},\bar{S}), \tag{4.5.15}$$

for the "corrections" $(\Delta X,\Delta S)$;

We add to (4.5.15) the system of linear equations on ΔX, ΔS expressing the requirement that a shift of $(\bar{X},\bar{S})$ in the direction $(\Delta X,\Delta S)$ should preserve the validity of the linear constraints in (P), (D), i.e., the equations saying that $\Delta X\in\mathcal{L}$, $\Delta S\in\mathcal{L}^\perp$. These linear equations can be written down as

$$\begin{array}{rcll}\Delta X&=&\mathcal{A}\Delta x&[\Leftrightarrow\Delta X\in\mathcal{L}]\\ \mathcal{A}^*\Delta S&=&0&[\Leftrightarrow\Delta S\in\mathcal{L}^\perp]\end{array} \tag{4.5.16}$$

(iii) We solve the system of linear equations (4.5.15), (4.5.16), thus obtaining a primal-dual search direction $(\Delta X,\Delta S)$, and update current iterates according to

$$X_+=\bar{X}+\alpha\Delta x,\quad S_+=\bar{S}+\beta\Delta S$$

where the primal and the dual stepsizes α,β are given by certain "side requirements".

The major "degree of freedom" of the construction comes from (i) – from how we construct the system (4.5.14). A very popular way to handle (i), the way which indeed leads to *primal-dual* methods, starts from rewriting (4.5.13) in a form symmetric w.r.t. X and S. To this end we first observe that (4.5.13) is equivalent to every one of the following two matrix equations:

$$XS=t^{-1}I;\quad SX=t^{-1}I.$$

Adding these equations, we get a "symmetric" w.r.t. X,S matrix equation

$$XS+SX=2t^{-1}I, \tag{4.5.17}$$

which, by its origin, is a consequence of (4.5.13). On a closest inspection, it turns out that (4.5.17), regarded as a matrix equation with *positive definite* symmetric matrices, is equivalent to (4.5.13). It is possible to use in the role of (4.5.14) the matrix equation (4.5.17) as it is; this policy leads to the so called AHO (Alizadeh-Overton-Haeberly) search direction and the "$XS+SX$" primal-dual path-following method.

It is also possible to use a "scaled" version of (4.5.17). Namely, let us choose somehow a positive definite *scaling matrix* Q and observe that our original matrix equation (4.5.13) says that $S=t^{-1}X^{-1}$, which is exactly the same as to say that $Q^{-1}SQ^{-1}=t^{-1}(QXQ)^{-1}$; the latter, in turn, is equivalent to every one of the matrix equations

$$QXSQ^{-1}=t^{-1}I;\quad Q^{-1}SXQ=t^{-1}I;$$

Adding these equations, we get the scaled version of (4.5.17):

$$QXSQ^{-1}+Q^{-1}SXQ=2t^{-1}I, \tag{4.5.18}$$

which, same as (4.5.17) itself, is equivalent to (4.5.13).

With (4.5.18) playing the role of (4.5.14), we get a quite flexible scheme with a huge freedom for choosing the scaling matrix Q, which in particular can be varied from iteration to iteration. As we shall see in a while, this freedom reflects the intrinsic (and extremely important in the interior-point context) symmetries of the semidefinite cone.

Analysis of the path-following methods based on search directions coming from (4.5.18) ("Zhang's family of search directions") simplifies a lot when at every iteration we choose its own scaling matrix and ensure that the matrices

$$\widetilde{S} = Q^{-1}\bar{S}Q^{-1},\ \widehat{X} = Q\bar{X}Q$$

commute ($\bar{X}, \bar{S}$ are the iterates to be updated); we call such a policy a "commutative scaling". Popular commutative scalings are:

1. $Q = \bar{S}^{1/2}$ ($\widetilde{S} = I, \widehat{X} = \bar{S}^{1/2}\bar{X}\bar{S}^{1/2}$) (the "$XS$" method);
2. $Q = \bar{X}^{-1/2}$ ($\widetilde{S} = \bar{X}^{1/2}\bar{S}\bar{X}^{1/2}$, $\widehat{X} = I$) (the "SX" method);
3. Q is such that $\widetilde{S} = \widehat{X}$ (the NT (Nesterov-Todd) method, extremely attractive and deep)

   If $\bar{X}$ and $\bar{S}$ were just positive reals, the formula for Q would be simple: $Q = \left(\frac{\bar{S}}{\bar{X}}\right)^{1/4}$. In the matrix case this simple formula becomes a bit more complicated (to make our life easier, below we write X instead of $\bar{X}$ and S instead of $\bar{S}$):

   $$Q = P^{1/2}, \quad P = X^{-1/2}(X^{1/2}SX^{1/2})^{-1/2}X^{1/2}S.$$

   We should verify that (a) P is symmetric positive definite, so that Q is well-defined, and that (b) $Q^{-1}SQ^{-1} = QXQ$.
   (a): Let us first verify that P is symmetric:

$$\begin{array}{c} P\ ?=?\ P^T \\ \Updownarrow \\ X^{-1/2}(X^{1/2}SX^{1/2})^{-1/2}X^{1/2}S\ ?=?\ SX^{1/2}(X^{1/2}SX^{1/2})^{-1/2}X^{-1/2} \\ \Updownarrow \\ \left(X^{-1/2}(X^{1/2}SX^{1/2})^{-1/2}X^{1/2}S\right)\left(X^{1/2}(X^{1/2}SX^{1/2})^{1/2}X^{-1/2}S^{-1}\right)\ ?=?\ I \\ \Updownarrow \\ X^{-1/2}(X^{1/2}SX^{1/2})^{-1/2}(X^{1/2}SX^{1/2})(X^{1/2}SX^{1/2})^{1/2}X^{-1/2}S^{-1}\ ?=?\ I \\ \Updownarrow \\ X^{-1/2}(X^{1/2}SX^{1/2})X^{-1/2}S^{-1}\ ?=?\ I \end{array}$$

   and the concluding $?=?$ indeed is $=$.
   Now let us verify that P is positive definite. Recall that the spectrum of the product of two square matrices, symmetric or not, remains unchanged when swapping the factors. Therefore, denoting $\sigma(A)$ the spectrum of A, we have

$$\begin{array}{rcl} \sigma(P) & = & \sigma\left(X^{-1/2}(X^{1/2}SX^{1/2})^{-1/2}X^{1/2}S\right) \\ & = & \sigma\left((X^{1/2}SX^{1/2})^{-1/2}X^{1/2}SX^{-1/2}\right) \\ & = & \sigma\left((X^{1/2}SX^{1/2})^{-1/2}(X^{1/2}SX^{1/2})X^{-1}\right) \\ & = & \sigma\left((X^{1/2}SX^{1/2})^{1/2}X^{-1}\right) \\ & = & \sigma\left(X^{-1/2}(X^{1/2}SX^{1/2})^{1/2}X^{-1/2}\right), \end{array}$$

   and the argument of the concluding $\sigma(\cdot)$ clearly is a positive definite symmetric matrix. Thus, the spectrum of *symmetric* matrix P is positive, i.e., P is positive definite.
   (b): To verify that $QXQ = Q^{-1}SQ^{-1}$, i.e., that $P^{1/2}XP^{1/2} = P^{-1/2}SP^{-1/2}$, is the same as to verify that $PXP = S$. The latter equality is given by the following computation:

$$\begin{array}{rcl} PXP & = & \left(X^{-1/2}(X^{1/2}SX^{1/2})^{-1/2}X^{1/2}S\right)X\left(X^{-1/2}(X^{1/2}SX^{1/2})^{-1/2}X^{1/2}S\right) \\ & = & X^{-1/2}(X^{1/2}SX^{1/2})^{-1/2}(X^{1/2}SX^{1/2})(X^{1/2}SX^{1/2})^{-1/2}X^{1/2}S \\ & = & X^{-1/2}X^{1/2}S \\ & = & S. \end{array}$$

You should not think that Nesterov and Todd guessed the formula for this scaling matrix. They did much more: they have developed an extremely deep theory (covering the general LP-CQP-SDP case, not just the SDP one!) which, among other things, guarantees that the desired scaling matrix exists (and even is unique). After the existence is established, it becomes much easier (although still not that easy) to find an explicit formula for Q.

Complexity analysis

We are about to carry out the complexity analysis of the primal-dual path-following methods based on "commutative" Zhang's scalings. This analysis, although not that difficult, is more technical than whatever else in our course, and a non-interested reader may skip it without any harm.

Scalings. We already have mentioned what a *scaling* of $\mathbf{S}^k_+$ is: this is the linear one-to-one transformation of $\mathbf{S}^k$ given by the formula

$$H \mapsto QHQ^T, \tag{Scl}$$

where Q is a nonsingular *scaling matrix*. It is immediately seen that (Scl) is a symmetry of the semidefinite cone $\mathbf{S}^k_+$ – it maps the cone onto itself. This family of symmetries is quite rich: for every pair of points A, B from the interior of the cone, there exists a scaling which maps A onto B, e.g., the scaling

$$H \mapsto (\underbrace{B^{1/2}A^{-1/2}}_{Q})H(\underbrace{A^{-1/2}B^{1/2}}_{Q^T}).$$

Essentially, this is exactly the existence of that rich family of symmetries of the underlying cones which makes SDP (same as LP and CQP, where the cones also are "perfectly symmetric") especially well suited for IP methods.

In what follows we will be interested in scalings associated with *positive definite* scaling matrices. The scaling given by such a matrix Q (X,S,...) will be denoted by $\mathcal{Q}$ (resp., $\mathcal{X}$,$\mathcal{S}$,...):

$$\mathcal{Q}[H] = QHQ.$$

Given a problem of interest (CP) (where $\mathbf{K} = \mathbf{S}^k_+$) and a scaling matrix $Q \succ 0$, we can *scale* the problem, i.e., pass from it to the problem

$$\min_x \left\{c^T x : \mathcal{Q}\left[\mathcal{A}x - B\right] \succeq 0\right\} \tag{$\mathcal{Q}$CP}$$

which, of course, is equivalent to (CP) (since $\mathcal{Q}[H]$ is positive semidefinite iff H is so). In terms of "geometric reformulation" (P) of (CP), this transformation is nothing but the substitution of variables

$$QXQ = Y \Leftrightarrow X = Q^{-1}YQ^{-1};$$

with respect to Y-variables, (P) is the problem

$$\min_Y \left\{\mathrm{Tr}(C[Q^{-1}YQ^{-1}]) : Y \in \mathcal{Q}(\mathcal{L}) - \mathcal{Q}[B],\ Y \succeq 0\right\},$$

i.e., the problem

$$\begin{array}{c}\min\limits_Y \left\{\mathrm{Tr}(\widetilde{C}Y) : Y \in \widehat{\mathcal{L}} - \widehat{B},\ Y \succeq 0\right\} \\ \left[\widetilde{C} = Q^{-1}CQ^{-1}, \widehat{B} = QBQ, \widehat{\mathcal{L}} = \mathrm{Im}(\mathcal{Q}\mathcal{A}) = \mathcal{Q}(\mathcal{L})\right]\end{array} \tag{$\widehat{\mathrm{P}}$}$$

The problem dual to ($\widehat{\mathrm{P}}$) is

$$\max_Z \left\{\mathrm{Tr}(\widehat{B}Z) : Z \in \widehat{\mathcal{L}}^\perp + \widehat{C},\ Z \succeq 0\right\}. \tag{$\widetilde{\mathrm{D}}$}$$

It is immediate to realize what is $\widehat{\mathcal{L}}^\perp$:

$$\langle Z, QXQ\rangle_F = \mathrm{Tr}(ZQXQ) = \mathrm{Tr}(QZQX) = \langle QZQ, X\rangle_F;$$

thus, Z is orthogonal to every matrix from $\widehat{\mathcal{L}}$, i.e., to every matrix of the form QXQ with $X \in \mathcal{L}$ iff the matrix QZQ is orthogonal to every matrix from $\mathcal{L}$, i.e., iff $QZQ \in \mathcal{L}^\perp$. It follows that

$$\widehat{\mathcal{L}}^\perp = \mathcal{Q}^{-1}(\mathcal{L}^\perp).$$

Thus, when acting on the primal-dual pair (P), (D) of SDP's, a scaling, given by a matrix $Q \succ 0$, converts it into another primal-dual pair of problems, and this new pair is as follows:

• The "primal" geometric data – the subspace $\mathcal{L}$ and the primal shift B (which has a part-time job to be the dual objective as well) – are replaced with their images under the mapping $\mathcal{Q}$;

• The "dual" geometric data – the subspace $\mathcal{L}^\perp$ and the dual shift C (it is the primal objective as well) – are replaced with their images under the mapping $\mathcal{Q}^{-1}$ *inverse* to $\mathcal{Q}$; this inverse mapping again is a scaling, the scaling matrix being Q^{-1}.

We see that it makes sense to speak about *primal-dual scaling* which acts on both the primal and the dual variables and maps a primal variable X onto QXQ, and a dual variable S onto $Q^{-1}SQ^{-1}$. Formally speaking, the primal-dual scaling associated with a matrix $Q \succ 0$ is the linear transformation $(X, S) \mapsto (QXQ, Q^{-1}SQ^{-1})$ of the direct product of two copies of $\mathbf{S}^k$ (the "primal" and the "dual" ones). A primal-dual scaling acts naturally on different entities associated with a primal-dual pair (P), (S), in particular, at:

- the pair (P), (D) itself – it is converted into another primal-dual pair of problems $(\widehat{\mathrm{P}})$, $(\widetilde{\mathrm{D}})$;
- a primal-dual feasible pair (X, S) of solutions to (P), (D) – it is converted to the pair $(\widehat{X} = QXQ, \widetilde{S} = Q^{-1}SQ^{-1})$, which, as it is immediately seen, is a pair of feasible solutions to $(\widehat{\mathrm{P}})$, $(\widetilde{\mathrm{D}})$. Note that the primal-dual scaling preserves strict feasibility and the duality gap:

  $$\mathrm{DualityGap}_{\mathrm{P,D}}(X, S) = \mathrm{Tr}(XS) = \mathrm{Tr}(QXSQ^{-1}) = \mathrm{Tr}(\widehat{X}\widetilde{S}) = \mathrm{DualityGap}_{\widehat{\mathrm{P}},\widetilde{\mathrm{D}}}(\widehat{X}, \widetilde{S});$$

- the primal-dual central path $(X_*(\cdot), S_*(\cdot))$ of (P), (D); it is converted into the curve $(\widehat{X}_*(t) = QX_*(t)Q, \widetilde{S}_*(t) = Q^{-1}S_*(t)Q^{-1})$, which is nothing but the primal-dual central path $\overline{Z}(t)$ of the primal-dual pair $(\widehat{\mathrm{P}})$, $(\widetilde{\mathrm{D}})$.

  The latter fact can be easily derived from the characterization of the primal-dual central path; a more instructive derivation is based on the fact that our "hero" – the barrier $S_k(\cdot)$ – is "semi-invariant" w.r.t. scaling:

  $$S_k(\mathcal{Q}(X)) = -\ln\mathrm{Det}(QXQ) = -\ln\mathrm{Det}(X) - 2\ln\mathrm{Det}(Q) = S_k(X) + \mathrm{const}(Q).$$

  Now, a point on the primal central path of the problem $(\widehat{\mathrm{P}})$ associated with penalty parameter t, let this point be temporarily denoted by $Y(t)$, is the unique minimizer of the aggregate

  $$S_k^t(Y) = t\langle Q^{-1}CQ^{-1}, Y\rangle_F + S_k(Y) \equiv t\mathrm{Tr}(Q^{-1}CQ^{-1}Y) + S_k(Y)$$

  over the set of strictly feasible solutions of $(\widehat{\mathrm{P}})$. The latter set is exactly the image of the set of strictly feasible solutions of (P) under the transformation $\mathcal{Q}$, so that $Y(t)$ is the image, under the same transformation, of the point, let it be called $X(t)$, which minimizes the aggregate

  $$S_k^t(QXQ) = t\mathrm{Tr}((Q^{-1}CQ^{-1})(QXQ)) + S_k(QXQ) = t\mathrm{Tr}(CX) + S_k(X) + \mathrm{const}(Q)$$

  over the set of strictly feasible solutions to (P). We see that $X(t)$ is exactly the point $X_*(t)$ on the primal central path associated with problem (P). Thus, the point $Y(t)$ of the primal central path associated with $(\widehat{\mathrm{P}})$ is nothing but $\widehat{X}_*(t) = QX_*(t)Q$. Similarly, the point of the central path associated with the problem $(\widetilde{\mathrm{D}})$ is exactly $\widetilde{S}_*(t) = Q^{-1}S_*(t)Q^{-1}$.

- the neighbourhood $\mathcal{N}_\kappa$ of the primal-dual central path $Z(\cdot)$ associated with the pair of problems (P), (D) (see (4.4.8)). As you can guess, the image of $\mathcal{N}_\kappa$ is exactly the neighbourhood $\overline{\mathcal{N}}_\kappa$, given by (4.4.8), of the primal-dual central path $\overline{Z}(\cdot)$ of $(\widehat{\mathrm{P}})$, $(\widetilde{\mathrm{D}})$.

The latter fact is immediate: for a pair (X,S) of strictly feasible primal and dual solutions to (P), (D) and a $t>0$ we have (see (4.4.6)):

$$\begin{array}{rcl}\mathrm{dist}^2((\widehat{X},\widetilde{S}),\overline{Z}_*(t)) & = & \mathrm{Tr}\left([QXQ](tQ^{-1}SQ^{-1}-[QXQ]^{-1})[QXQ](tQ^{-1}SQ^{-1}-[QXQ]^{-1})\right)\\ & = & \mathrm{Tr}\left(QX(tS-X^{-1})X(tS-X^{-1})Q^{-1}\right)\\ & = & \mathrm{Tr}\left(X(tS-X^{-1})X(tS-X^{-1})\right)\\ & = & \mathrm{dist}^2((X,S),Z_*(t)).\end{array}$$

Primal-dual short-step path-following methods based on commutative scalings. Path-following methods we are about to consider trace the primal-dual central path of (P), (D), staying in $\mathcal{N}_\kappa$-neighbourhood of the path; here $\kappa\le 0.1$ is fixed. The path is traced by iterating the following updating:

(U): Given a current pair of strictly feasible primal and dual solutions $(\bar X,\bar S)$ such that the triple

$$\left(\bar t=\frac{k}{\mathrm{Tr}(\bar X\bar S)},\bar X,\bar S\right) \tag{4.5.19}$$

belongs to $\mathcal{N}_\kappa$, i.e. (see (4.4.6))

$$\|\bar t\bar X^{1/2}\bar S\bar X^{1/2}-I\|_2\le\kappa, \tag{4.5.20}$$

we

1. Choose the new value t_+ of the penalty parameter according to
$$t_+=\left(1-\frac{\chi}{\sqrt{k}}\right)^{-1}\bar t, \tag{4.5.21}$$
where $\chi\in(0,1)$ is a parameter of the method;
2. Choose somehow the scaling matrix $Q\succ 0$ such that the matrices $\widehat{X}=Q\bar XQ$ and $\widetilde{S}=Q^{-1}\bar SQ^{-1}$ commute with each other;
3. Linearize the equation
$$QXSQ^{-1}+Q^{-1}SXQ=\frac{2}{t_+}I$$
at the point $(\bar X,\bar S)$, thus coming to the equation
$$Q[\Delta XS+X\Delta S]Q^{-1}+Q^{-1}[\Delta SX+S\Delta X]Q=\frac{2}{t_+}I-[Q\bar X\bar SQ^{-1}+Q^{-1}\bar S\bar XQ]; \tag{4.5.22}$$
4. Add to (4.5.22) the linear equations
$$\begin{array}{rcl}\Delta X & \in & \mathcal{L},\\ \Delta S & \in & \mathcal{L}^\perp;\end{array} \tag{4.5.23}$$
5. Solve system (4.5.22), (4.5.23), thus getting "primal-dual search direction" $(\Delta X,\Delta S)$;
6. Update current primal-dual solutions $(\bar X,\bar S)$ into a new pair (X_+,S_+) according to
$$X_+=\bar X+\Delta X,\quad S_+=\bar S+\Delta S.$$

We already have explained the ideas underlying (U), up to the fact that in our previous explanations we dealt with three "independent" entities $\bar t$ (current value of the penalty parameter), $\bar X$, $\bar S$ (current primal and dual solutions), while in (U) $\bar t$ is a function of $\bar X,\bar S$:

$$\bar t=\frac{k}{\mathrm{Tr}(\bar X\bar S)}. \tag{4.5.24}$$

The reason for establishing this dependence is very simple: if (t, X, S) were on the primal-dual central path: $XS = t^{-1}I$, then, taking traces, we indeed would get $t = \frac{k}{\mathrm{Tr}(XS)}$. Thus, (4.5.24) is a reasonable way to reduce the number of "independent entities" we deal with.

Note also that (U) is a "pure Newton scheme" – here the primal and the dual stepsizes are equal to 1 (cf. (4.5.6)).

The major element of the complexity analysis of path-following polynomial time methods for SDP is as follows:

Theorem 4.5.2 *Let the parameters κ, χ of* (U) *satisfy the relations*

$$0 < \chi \leq \kappa \leq 0.1. \tag{4.5.25}$$

Let, further, $(\bar{X}, \bar{S})$ be a pair of strictly feasible primal and dual solutions to (P), (D) *such that the triple* (4.5.19) *satisfies* (4.5.20)*. Then the updated pair (X_+, S_+) is well-defined (i.e., system* (4.5.22), (4.5.23) *is solvable with a unique solution), X_+, S_+ are strictly feasible solutions to* (P), (D)*, respectively,*

$$t_+ = \frac{k}{\mathrm{Tr}(X_+S_+)}$$

and the triple (t_+, X_+, S_+) belongs to $\mathcal{N}_\kappa$.

The theorem says that with properly chosen κ, χ (say, $\kappa = \chi = 0.1$), updating (U) converts a close to the primal-dual central path, in the sense of (4.5.20), strictly primal-dual feasible iterate $(\bar{X}, \bar{S})$ into a new strictly primal-dual feasible iterate with the same closeness-to-the-path property and larger, by factor $(1 + O(1)k^{-1/2})$, value of the penalty parameter. Thus, after we get close to the path – reach its 0.1-neighbourhood $\mathcal{N}_{0.1}$ – we are able to trace this path, staying in $\mathcal{N}_{0.1}$ and increasing the penalty parameter by absolute constant factor in $O(\sqrt{k}) = O(\sqrt{\theta(K)})$ steps, exactly as announced in Section 4.5.2.

Proof of Theorem 4.5.2. 1^0. Observe, first (this observation is crucial!) that it suffices to prove our Theorem in the particular case when $\bar{X}, \bar{S}$ commute with each other and $Q = I$. Indeed, it is immediately seen that the updating (U) can be represented as follows:

1. We first scale by Q the "input data" of (U) – the primal-dual pair of problems (P), (D) and the strictly feasible pair $\bar{X}, \bar{S}$ of primal and dual solutions to these problems, as explained in sect. "Scaling". Note that the resulting entities – a pair of primal-dual problems and a strictly feasible pair of primal-dual solutions to these problems – are linked with each other exactly in the same fashion as the original entities, due to scaling invariance of the duality gap and the neighbourhood $\mathcal{N}_\kappa$. In addition, the scaled primal and dual solutions commute;
2. We apply to the "scaled input data" yielded by the previous step the updating $(\widehat{\mathrm{U}})$ completely similar to (U), but using the unit matrix in the role of Q;
3. We "scale back" the result of the previous step, i.e., subject this result to the scaling associated with Q^{-1}, thus obtaining the updated iterate (X^+, S^+).

Given that the second step of this procedure preserves primal-dual strict feasibility, w.r.t. the scaled primal-dual pair of problems, of the iterate and keeps the iterate in the κ-neighbourhood $\mathcal{N}_\kappa$ of the corresponding central path, we could use once again the "scaling invariance" reasoning to assert that the result (X^+, S^+) of (U) is well-defined, is strictly feasible for (P), (D) and is close to the original central path, as claimed in the Theorem. Thus, all we need is to justify the above "Given", and this is exactly the same as to prove the theorem in the particular case of $Q = I$ and commuting $\bar{X}$, $\bar{S}$. In the rest of the proof we assume that $Q = I$ and that the matrices $\bar{X}, \bar{S}$ commute with each other. Due to the latter property, $\bar{X}, \bar{S}$ are diagonal in a properly chosen orthonormal basis; representing all matrices from $\mathbf{S}^k$ in this basis, we can reduce the situation to the case when $\bar{X}$ and $\bar{S}$ are diagonal. Thus, we may (and do) assume in the sequel that $\bar{X}$ and $\bar{S}$ are diagonal, with diagonal entries x_i,s_i, $i = 1, ..., k$, respectively, and that $Q = I$. Finally, to simplify notation, we write t, X, S instead of $\bar{t}$, $\bar{X}$, $\bar{S}$, respectively.

2^0. Our situation and goals now are as follows. We are given orthogonal to each other affine planes $\mathcal{L}-B$, $\mathcal{L}^\perp+C$ in $\mathbf{S}^k$ and two positive definite diagonal matrices $X=\mathrm{Diag}(\{x_i\})\in\mathcal{L}-B$, $S=\mathrm{Diag}(\{s_i\})\in\mathcal{L}^\perp+C$. We set

$$\mu=\frac{1}{t}=\frac{\mathrm{Tr}(XS)}{k}$$

and know that

$$\|tX^{1/2}SX^{1/2}-I\|_2\le\kappa.$$

We further set

$$\mu_+=\frac{1}{t_+}=(1-\chi k^{-1/2})\mu \tag{4.5.26}$$

and consider the system of equations w.r.t. unknown symmetric matrices $\Delta X,\Delta S$:

$$\begin{array}{ll}(a) & \Delta X\in\mathcal{L}\\ (b) & \Delta S\in\mathcal{L}^\perp\\ (c) & \Delta XS+X\Delta S+\Delta SX+S\Delta X=2\mu_+I-2XS\end{array} \tag{4.5.27}$$

We should prove that the system has a unique solution such that the matrices

$$X_+=X+\Delta X,\ S_+=S+\Delta S$$

are

(i) positive definite,

(ii) belong, respectively, to $\mathcal{L}-B$, $\mathcal{L}^\perp+C$ and satisfy the relation

$$\mathrm{Tr}(X_+S_+)=\mu_+k; \tag{4.5.28}$$

(iii) satisfy the relation

$$\Omega\equiv\|\mu_+^{-1}X_+^{1/2}S_+X_+^{1/2}-I\|_2\le\kappa. \tag{4.5.29}$$

Observe that the situation can be reduced to the one with $\mu=1$. Indeed, let us pass from the matrices $X,S,\Delta X,\Delta S,X_+,S_+$ to $X,S'=\mu^{-1}S,\Delta X,\Delta S'=\mu^{-1}\Delta S,X_+,S'_+=\mu^{-1}S_+$. Now the "we are given" part of our situation becomes as follows: we are given two diagonal positive definite matrices X,S' such that $X\in\mathcal{L}-B$, $S'\in\mathcal{L}^\perp+C'$, $C'=\mu^{-1}C$,

$$\mathrm{Tr}(XS')=k\times 1$$

and

$$\|X^{1/2}S'X^{1/2}-I\|_2=\|\mu^{-1}X^{1/2}SX^{1/2}-I\|_2\le\kappa.$$

The "we should prove" part becomes: to verify that the system of equations

$$\begin{array}{ll}(a) & \Delta X\in\mathcal{L}\\ (b) & \Delta S'\in\mathcal{L}^\perp\\ (c) & \Delta XS'+X\Delta S'+\Delta S'X+S'\Delta X=2(1-\chi k^{-1/2})I-2XS'\end{array}$$

has a unique solution and that the matrices $X_+=X+\Delta X$, $S'_+=S'+\Delta S'_+$ are positive definite, are contained in $\mathcal{L}-B$, respectively, $\mathcal{L}^\perp+C'$ and satisfy the relations

$$\mathrm{Tr}(X_+S'_+)=\frac{\mu_+}{\mu}=1-\chi k^{-1/2}$$

and

$$\|(1-\chi k^{-1/2})^{-1}X_+^{1/2}S'_+X_+^{1/2}-I\|_2\le\kappa.$$

Thus, the general situation indeed can be reduced to the one with $\mu=1$, $\mu_+=1-\chi k^{-1/2}$, and we loose nothing assuming, in addition to what was already postulated, that

$$\mu\equiv t^{-1}\equiv\frac{\mathrm{Tr}(XS)}{k}=1,\quad \mu_+=1-\chi k^{-1/2},$$

whence

$$[\mathrm{Tr}(XS) =] \quad \sum_{i=1}^{k} x_i s_i = k \tag{4.5.30}$$

and

$$[\|tX^{1/2}SX^{1/2} - I\|_2^2 \equiv] \quad \sum_{i=1}^{n} (x_i s_i - 1)^2 \leq \kappa^2. \tag{4.5.31}$$

3^0. We start with proving that (4.5.27) indeed has a unique solution. It is convenient to pass in (4.5.27) from the unknowns ΔX, ΔS to the unknowns

$$\begin{array}{rcl} \delta X = X^{-1/2}\Delta X X^{-1/2} & \Leftrightarrow & \Delta X = X^{1/2}\delta X X^{1/2}, \\ \delta S = X^{1/2}\Delta S X^{1/2} & \Leftrightarrow & \Delta S = X^{-1/2}\delta S X^{-1/2}. \end{array} \tag{4.5.32}$$

With respect to the new unknowns, (4.5.27) becomes

$$\begin{array}{ll} (a) & X^{1/2}\delta X X^{1/2} \in \mathcal{L}, \\ (b) & X^{-1/2}\delta S X^{-1/2} \in \mathcal{L}^\perp, \\ (c) & X^{1/2}\delta X X^{1/2} S + X^{1/2}\delta S X^{-1/2} + X^{-1/2}\delta S X^{1/2} + S X^{1/2}\delta X X^{1/2} = 2\mu_+ I - 2XS \\ & \Updownarrow \\ (d) & L(\delta X, \delta S) \equiv \left[\underbrace{\sqrt{x_i x_j}(s_i + s_j)}_{\phi_{ij}}(\delta X)_{ij} + \underbrace{\left(\sqrt{\frac{x_i}{x_j}} + \sqrt{\frac{x_j}{x_i}} \right)}_{\psi_{ij}}(\delta S)_{ij} \right]_{i,j=1}^{k} = 2\left[(\mu_+ - x_i s_i)\delta_{ij}\right]_{i,j=1}^{k}, \end{array} \tag{4.5.33}$$

where $\delta_{ij} = \begin{cases} 0, & i \neq j \\ 1, & i = j \end{cases}$ are the Kronecker symbols.

We first claim that (4.5.33), regarded as a system with unknown *symmetric* matrices δX, δS has a unique solution. Observe that (4.5.33) is a system with $2\dim \mathbf{S}^k \equiv 2N$ scalar unknowns and $2N$ scalar linear equations. Indeed, (4.5.33.a) is a system of $N' \equiv N - \dim \mathcal{L}$ linear equations, (4.5.33.b) is a system of $N'' = N - \dim \mathcal{L}^\perp = \dim \mathcal{L}$ linear equations, and (4.5.33.c) has N equations, so that the total # of linear equations in our system is $N' + N'' + N = (N - \dim \mathcal{L}) + \dim \mathcal{L} + N = 2N$. Now, to verify that the *square* system of linear equations (4.5.33) has exactly one solution, it suffices to prove that the homogeneous system

$$X^{1/2}\delta X X^{1/2} \in \mathcal{L},\ X^{-1/2}\delta S X^{-1/2} \in \mathcal{L}^\perp,\ L(\delta X, \delta S) = 0$$

has only trivial solution. Let $(\delta X, \delta S)$ be a solution to the homogeneous system. Relation $L(\delta X, \Delta S) = 0$ means that

$$(\delta X)_{ij} = -\frac{\psi_{ij}}{\phi_{ij}}(\delta S)_{ij}, \tag{4.5.34}$$

whence

$$\mathrm{Tr}(\delta X \delta S) = -\sum_{i,j} \frac{\psi_{ij}}{\phi_{ij}}(\Delta S)_{ij}^2. \tag{4.5.35}$$

Representing $\delta X, \delta S$ via $\Delta X, \Delta S$ according to (4.5.32), we get

$$\mathrm{Tr}(\delta X \delta S) = \mathrm{Tr}(X^{-1/2}\Delta X X^{-1/2} X^{1/2}\Delta S X^{1/2}) = \mathrm{Tr}(X^{-1/2}\Delta X \Delta S X^{1/2}) = \mathrm{Tr}(\Delta X \Delta S),$$

and the latter quantity is 0 due to $\Delta X = X^{1/2}\delta X X^{1/2} \in \mathcal{L}$ and $\Delta S = X^{-1/2}\delta S X^{-1/2} \in \mathcal{L}^\perp$. Thus, the left hand side in (4.5.35) is 0; since $\phi_{ij} > 0$, $\psi_{ij} > 0$, (4.5.35) implies that $\delta S = 0$. But then $\delta X = 0$ in view of (4.5.34). Thus, the homogeneous version of (4.5.33) has the trivial solution only, so that (4.5.33) is solvable with a unique solution.

4^0. Let $\delta X, \delta S$ be the unique solution to (4.5.33), and let ΔX, ΔS be linked to $\delta X, \delta S$ according to (4.5.32). Our local goal is to bound from above the Frobenius norms of δX and δS.

From (4.5.33.c) it follows (cf. derivation of (4.5.35)) that

$$\begin{array}{lll} (a) & (\delta X)_{ij} & = -\frac{\psi_{ij}}{\phi_{ij}}(\delta S)_{ij} + 2\frac{\mu_+ - x_i s_i}{\phi_{ii}}\delta_{ij}, \quad i,j = 1,...,k; \\ (b) & (\delta S)_{ij} & = -\frac{\phi_{ij}}{\psi_{ij}}(\delta X)_{ij} + 2\frac{\mu_+ - x_i s_i}{\psi_{ii}}\delta_{ij}, \quad i,j = 1,...,k. \end{array} \tag{4.5.36}$$

Same as in the concluding part of 3^0, relations (4.5.33.$a-b$) imply that

$$\mathrm{Tr}(\Delta X \Delta S) = \mathrm{Tr}(\delta X \delta S) = \sum_{i,j} (\delta X)_{ij}(\delta S)_{ij} = 0. \tag{4.5.37}$$

Multiplying (4.5.36.a) by $(\delta S)_{ij}$ and taking sum over i, j, we get, in view of (4.5.37), the relation

$$\sum_{i,j} \frac{\psi_{ij}}{\phi_{ij}} (\delta S)_{ij}^2 = 2 \sum_i \frac{\mu_+ - x_i s_i}{\phi_{ii}} (\delta S)_{ii}; \tag{4.5.38}$$

by "symmetric" reasoning, we get

$$\sum_{i,j} \frac{\phi_{ij}}{\psi_{ij}} (\delta X)_{ij}^2 = 2 \sum_i \frac{\mu_+ - x_i s_i}{\psi_{ii}} (\delta X)_{ii}. \tag{4.5.39}$$

Now let

$$\theta_i = x_i s_i, \tag{4.5.40}$$

so that in view of (4.5.30) and (4.5.31) one has

$$\begin{array}{ll} (a) & \sum_i \theta_i = k, \\ (b) & \sum_i (\theta_i - 1)^2 \le \kappa^2. \end{array} \tag{4.5.41}$$

Observe that

$$\phi_{ij} = \sqrt{x_i x_j}(s_i + s_j) = \sqrt{x_i x_j}\left(\frac{\theta_i}{x_i} + \frac{\theta_j}{x_j}\right) = \theta_j \sqrt{\frac{x_i}{x_j}} + \theta_i \sqrt{\frac{x_j}{x_i}}.$$

Thus,

$$\begin{array}{lll} \phi_{ij} & = & \theta_j \sqrt{\frac{x_i}{x_j}} + \theta_i \sqrt{\frac{x_j}{x_i}}, \\ \psi_{ij} & = & \sqrt{\frac{x_i}{x_j}} + \sqrt{\frac{x_j}{x_i}}; \end{array} \tag{4.5.42}$$

since $1 - \kappa \le \theta_i \le 1 + \kappa$ by (4.5.41.b), we get

$$1 - \kappa \le \frac{\phi_{ij}}{\psi_{ij}} \le 1 + \kappa. \tag{4.5.43}$$

By the geometric-arithmetic mean inequality we have $\psi_{ij} \ge 2$, whence in view of (4.5.43)

$$\phi_{ij} \ge (1 - \kappa)\psi_{ij} \ge 2(1 - \kappa) \quad \forall i, j. \tag{4.5.44}$$

We now have

$$\begin{aligned}
(1-\kappa)\sum_{i,j}(\delta X)_{ij}^2 &\le \sum_{i,j}\tfrac{\phi_{ij}}{\psi_{ij}}(\delta X)_{ij}^2 \\
&\qquad\qquad \text{[see (4.5.43)]} \\
&\le 2\sum_i \tfrac{\mu_+ - x_i s_i}{\psi_{ii}}(\delta X)_{ii} \\
&\qquad\qquad \text{[see (4.5.39)]} \\
&\le 2\sqrt{\sum_i(\mu_+ - x_i s_i)^2}\sqrt{\sum_i \psi_{ij}^{-2}(\delta X)_{ii}^2} \\
&\le \sqrt{\sum_i((1-\theta_i)^2 - 2\chi k^{-1/2}(1-\theta_i) + \chi^2 k^{-1})}\sqrt{\sum_{i,j}(\delta X)_{ij}^2} \\
&\qquad\qquad \text{[see (4.5.44)]} \\
&\le \sqrt{\chi^2 + \sum_i(1-\theta_i)^2}\sqrt{\sum_{i,j}(\delta X)_{ij}^2} \\
&\qquad\qquad \text{[since } \textstyle\sum_i(1-\theta_i) = 0 \text{ by (4.5.41.}a\text{)]} \\
&\le \sqrt{\chi^2+\kappa^2}\sqrt{\sum_{i,j}(\delta X)_{ij}^2} \\
&\qquad\qquad \text{[see (4.5.41.}b\text{)]}
\end{aligned}$$

and from the resulting inequality it follows that

$$\|\delta X\|_2 \le \rho \equiv \frac{\sqrt{\chi^2+\kappa^2}}{1-\kappa}. \tag{4.5.45}$$

Similarly,

$$\begin{aligned}
(1+\kappa)^{-1}\sum_{i,j}(\delta S)_{ij}^2 &\le \sum_{i,j}\tfrac{\psi_{ij}}{\phi_{ij}}(\delta S)_{ij}^2 \\
&\qquad\qquad \text{[see (4.5.43)]} \\
&\le 2\sum_i \tfrac{\mu_+ - x_i s_i}{\phi_{ii}}(\delta S)_{ii} \\
&\qquad\qquad \text{[see (4.5.38)]} \\
&\le 2\sqrt{\sum_i(\mu_+ - x_i s_i)^2}\sqrt{\sum_i \phi_{ij}^{-2}(\delta S)_{ii}^2} \\
&\le (1-\kappa)^{-1}\sqrt{\sum_i(\mu_+ - \theta_i)^2}\sqrt{\sum_{i,j}(\delta S)_{ij}^2} \\
&\qquad\qquad \text{[see (4.5.44)]} \\
&\le (1-\kappa)^{-1}\sqrt{\chi^2+\kappa^2}\sqrt{\sum_{i,j}(\delta S)_{ij}^2} \\
&\qquad\qquad \text{[same as above]}
\end{aligned}$$

and from the resulting inequality it follows that

$$\|\delta S\|_2 \le \frac{(1+\kappa)\sqrt{\chi^2+\kappa^2}}{1-\kappa} = (1+\kappa)\rho. \tag{4.5.46}$$

5^0. We are ready to prove 2^0.(i-ii). We have

$$X_+ = X + \Delta X = X^{1/2}(I + \delta X)X^{1/2},$$

and the matrix $I + \delta X$ is positive definite due to (4.5.45) (indeed, the right hand side in (4.5.45) is $\rho \le 1$, whence the Frobenius norm (and therefore - the maximum of modulae of eigenvalues) of δX is less than 1). Note that by the just indicated reasons $I + \delta X \preceq (1+\rho)I$, whence

$$X_+ \preceq (1+\rho)X. \tag{4.5.47}$$

Similarly, the matrix

$$S_+ = S + \Delta S = X^{-1/2}(X^{1/2}SX^{1/2} + \delta S)X^{-1/2}$$

is positive definite. Indeed, the eigenvalues of the matrix $X^{1/2}SX^{1/2}$ are $\geq \min_i \theta_i \geq 1-\kappa$, while the modulae of eigenvalues of δS, by (4.5.46), do not exceed $\frac{(1+\kappa)\sqrt{\chi^2+\kappa^2}}{1-\kappa} < 1-\kappa$. Thus, the matrix $X^{1/2}SX^{1/2}+\delta S$ is positive definite, whence S_+ also is so. We have proved 2^0.(i).

2^0.(ii) is easy to verify. First, by (4.5.33), we have $\Delta X \in \mathcal{L}$, $\Delta S \in \mathcal{L}^\perp$, and since $X \in \mathcal{L}-B$, $S \in \mathcal{L}^\perp + C$, we have $X_+ \in \mathcal{L}-B$, $S_+ \in \mathcal{L}^\perp + C$. Second, we have

$$\begin{array}{rcl}\mathrm{Tr}(X_+S_+) & = & \mathrm{Tr}(XS+X\Delta S+\Delta XS+\Delta X\Delta S)\\ & = & \mathrm{Tr}(XS+X\Delta S+\Delta XS)\\ & & [\text{since } \mathrm{Tr}(\Delta X\Delta S)=0 \text{ due to } \Delta X\in\mathcal{L},\ \Delta S\in\mathcal{L}^\perp]\\ & = & \mu_+ k\\ & & [\text{take the trace of both sides in (4.5.27.}c)]\end{array}$$

2^0.(ii) is proved.

6^0. It remains to verify 2^0.(iii). We should bound from above the quantity

$$\Omega = \|\mu_+^{-1}X_+^{1/2}S_+X_+^{1/2}-I\|_2 = \|X_+^{1/2}(\mu_+^{-1}S_+ - X_+^{-1})X_+^{1/2}\|_2,$$

and our plan is first to bound from above the "close" quantity

$$\begin{array}{rcl}\widehat{\Omega} = \|X^{1/2}(\mu_+^{-1}S_+ - X_+^{-1})X^{1/2}\|_2 & = & \mu_+^{-1}\|Z\|_2,\\ Z & = & X^{1/2}(S_+ - \mu_+X_+^{-1})X^{1/2},\end{array} \tag{4.5.48}$$

and then to bound Ω in terms of $\widehat{\Omega}$.

6^0.1. *Bounding* $\widehat{\Omega}$. We have

$$\begin{array}{rcl}Z & = & X^{1/2}(S_+-\mu_+X_+^{-1})X^{1/2}\\ & = & X^{1/2}(S+\Delta S)X^{1/2}-\mu_+X^{1/2}[X+\Delta X]^{-1}X^{1/2}\\ & = & XS+\delta S-\mu_+X^{1/2}[X^{1/2}(I+\delta X)X^{1/2}]^{-1}X^{1/2}\\ & & [\text{see (4.5.32)}]\\ & = & XS+\delta S-\mu_+(I+\delta X)^{-1}\\ & = & XS+\delta S-\mu_+(I-\delta X)-\mu_+[(I+\delta X)^{-1}-I+\delta X]\\ & = & \underbrace{XS+\delta S+\delta X-\mu_+I}_{Z^1}+\underbrace{(\mu_+-1)\delta X}_{Z^2}+\underbrace{\mu_+[I-\delta X-(I+\delta X)^{-1}]}_{Z^3},\end{array}$$

so that

$$\|Z\|_2 \leq \|Z^1\|_2+\|Z^2\|_2+\|Z^3\|_2. \tag{4.5.49}$$

We are about to bound separately all 3 terms in the right hand side of the latter inequality.

Bounding $\|Z^2\|_2$*:* We have

$$\|Z^2\|_2 = |\mu_+-1|\|\delta X\|_2 \leq \chi k^{-1/2}\rho \tag{4.5.50}$$

(see (4.5.45) and take into account that $\mu_+-1=-\chi k^{-1/2}$).

Bounding $\|Z^3\|_2$*:* Let λ_i be the eigenvalues of δX. We have

$$\begin{array}{rcl}\|Z^3\|_2 & = & \|\mu_+[(I+\delta X)^{-1}-I+\delta X]\|_2\\ & \leq & \|(I+\delta X)^{-1}-I+\delta X\|_2\\ & & [\text{since } |\mu_+|\leq 1]\\ & = & \sqrt{\sum_i\left(\frac{1}{1+\lambda_i}-1+\lambda_i\right)^2}\\ & & [\text{pass to the orthonormal eigenbasis of } \delta X]\\ & = & \sqrt{\sum_i \frac{\lambda_i^4}{(1+\lambda_i)^2}}\\ & \leq & \sqrt{\sum_i \frac{\rho^2\lambda_i^2}{(1-\rho)^2}}\\ & & [\text{see (4.5.45) and note that } \sum_i\lambda_i^2=\|\delta X\|_2^2\leq\rho^2]\\ & \leq & \frac{\rho^2}{1-\rho}\end{array} \tag{4.5.51}$$

Bounding $\|Z^1\|_2$*:* This is a bit more involving. We have

$$\begin{array}{rcl} Z^1_{ij} & = & (XS)_{ij} + (\delta S)_{ij} + (\delta X)_{ij} - \mu_+\delta_{ij} \\ & = & (\delta X)_{ij} + (\delta S)_{ij} + (x_i s_i - \mu_+)\delta_{ij} \\ & = & (\delta X)_{ij}\left[1 - \frac{\phi_{ij}}{\psi_{ij}}\right] + \left[2\frac{\mu_+ - x_i s_i}{\psi_{ii}} + x_i s_i - \mu_+\right]\delta_{ij} \\ & & \qquad\qquad [\text{we have used (4.5.36.}b\text{)}] \\ & = & (\delta X)_{ij}\left[1 - \frac{\phi_{ij}}{\psi_{ij}}\right] \\ & & \qquad\qquad [\text{since } \psi_{ii} = 2, \text{ see (4.5.42)}] \end{array}$$

whence, in view of (4.5.43),

$$|Z^1_{ij}| \le \left|1 - \frac{1}{1-\kappa}\right| |(\delta X)_{ij}| = \frac{\kappa}{1-\kappa}|(\delta X)_{ij}|,$$

so that

$$\|Z^1\|_2 \le \frac{\kappa}{1-\kappa}\|\delta X\|_2 \le \frac{\kappa}{1-\kappa}\rho \tag{4.5.52}$$

(the concluding inequality is given by (4.5.45)).

Assembling (4.5.50), (4.5.51), (4.5.52) and (4.5.49), we come to

$$\|Z\|_2 \le \rho\left[\frac{\chi}{\sqrt{k}} + \frac{\rho}{1-\rho} + \frac{\kappa}{1-\kappa}\right],$$

whence, by (4.5.48),

$$\widehat{\Omega} \le \frac{\rho}{1 - \chi k^{-1/2}}\left[\frac{\chi}{\sqrt{k}} + \frac{\rho}{1-\rho} + \frac{\kappa}{1-\kappa}\right]. \tag{4.5.53}$$

6^0.2. *Bounding* Ω. We have

$$\begin{array}{rcl} \Omega^2 & = & \|\mu_+^{-1}X_+^{1/2}S_+X_+^{1/2} - I\|_2^2 \\ & = & \|X_+^{1/2}\underbrace{[\mu_+^{-1}S_+ - X_+^{-1}]}_{\Theta=\Theta^T}X_+^{1/2}\|_2^2 \\ & = & \mathrm{Tr}\left(X_+^{1/2}\Theta X_+\Theta X_+^{1/2}\right) \\ & \le & (1+\rho)\mathrm{Tr}\left(X_+^{1/2}\Theta X\Theta X_+^{1/2}\right) \\ & & \qquad\qquad [\text{see (4.5.47)}] \\ & = & (1+\rho)\mathrm{Tr}\left(X_+^{1/2}\Theta X^{1/2}X^{1/2}\Theta X_+^{1/2}\right) \\ & = & (1+\rho)\mathrm{Tr}\left(X^{1/2}\Theta X_+^{1/2}X_+^{1/2}\Theta X^{1/2}\right) \\ & = & (1+\rho)\mathrm{Tr}\left(X^{1/2}\Theta X_+\Theta X^{1/2}\right) \\ & \le & (1+\rho)^2\mathrm{Tr}\left(X^{1/2}\Theta X\Theta X^{1/2}\right) \\ & & \qquad\qquad [\text{the same (4.5.47)}] \\ & = & (1+\rho)^2\|X^{1/2}\Theta X^{1/2}\|_2^2 \\ & = & (1+\rho)^2\|X^{1/2}[\mu_+^{-1}S_+ - X_+^{-1}]X^{1/2}\|_2^2 \\ & = & (1+\rho)^2\widehat{\Omega}^2 \\ & & \qquad\qquad [\text{see (4.5.48)}] \end{array}$$

so that

$$\begin{array}{c} \Omega \le (1+\rho)\widehat{\Omega} = \frac{\rho(1+\rho)}{1-\chi k^{-1/2}}\left[\frac{\chi}{\sqrt{k}} + \frac{\rho}{1-\rho} + \frac{\kappa}{1-\kappa}\right], \\ \rho = \frac{\sqrt{\chi^2+\kappa^2}}{1-\kappa}. \end{array} \tag{4.5.54}$$

(see (4.5.53) and (4.5.45)).

It is immediately seen that if $0 < \chi \le \kappa \le 0.1$, the right hand side in the resulting bound for Ω is $\le \kappa$, as required in 2^0.(iii). ∎

Remark 4.5.2 We have carried out the complexity analysis for a large group of primal-dual path-following methods for SDP (i.e., for the case of $\mathbf{K} = \mathbf{S}^k_+$). In fact, the constructions and the analysis we have presented can be word by word extended to the case when $\mathbf{K}$ is a direct product of semidefinite cones – you just should bear in mind that all symmetric matrices we deal with, like the primal and the dual solutions X, S, the scaling matrices Q, the primal-dual search directions ΔX, ΔS, etc., are block-diagonal with common block-diagonal structure. In particular, our constructions and analysis work for the case of LP – this is the case when $\mathbf{K}$ is a direct product of one-dimensional semidefinite cones. Note that in the case of LP Zhang's family of primal-dual search directions reduces to a single direction: since now X, S, Q are diagonal matrices, the scaling (4.5.17) $\mapsto$ (4.5.18) does not vary the equations of augmented complementary slackness.
The recipe to translate all we have presented for the case of SDP to the case of LP is very simple: in the above text, you should assume all matrices like X, S,... to be diagonal and look what the operations with these matrices required by the description of the method do with their diagonals. By the way, one of the very first approaches to the design and the analysis of IP methods for SDP was exactly opposite: you take an IP scheme for LP, replace in its description the words "nonnegative vectors" with "positive semidefinite diagonal matrices" and then erase the adjective "diagonal".

4.6 Complexity bounds for LP, CQP, SDP

In what follows we list the best known so far complexity bounds for LP, CQP and SDP. These bounds are yielded by IP methods and, essentially, say that the *Newton complexity* of finding ϵ-solution to an instance – the total # of steps of a "good" IP algorithm before an ϵ-solution is found – is $O(1)\sqrt{\theta(K)}\ln\frac{1}{\epsilon}$. This is what should be expected in view of discussion in Section 4.5.2; note, however, that the complexity bounds to follow take into account the necessity to "reach the highway" – to come close to the central path before tracing it, while in Section 4.5.2 we were focusing on how fast could we reduce the duality gap after the central path ("the highway") is reached.

Along with complexity bounds expressed in terms of the Newton complexity, we present the bounds on the number of operations of Real Arithmetic required to build an ϵ-solution. Note that these latter bounds typically are conservative – when deriving them, we assume the data of an instance "completely unstructured", which is usually not the case (cf. Warning in Section 4.5.2); exploiting structure of the data, one usually can reduce significantly computational effort per step of an IP method and consequently – the arithmetic cost of ϵ-solution.

4.6.1 Complexity of $\mathcal{LP}_b$

Family of problems:

Problem instance: a program

$$\min_x \left\{c^T x : a_i^T x \le b_i,\ i = 1, ..., m;\ \|x\|_2 \le R\right\} \quad [x \in \mathbf{R}^n]; \tag{p}$$

Data:

$$\begin{aligned}&\mathrm{Data}(p) = [m; n; c; a_1, b_1; ...; a_m, b_m; R],\\ &\mathrm{Size}(p) \equiv \dim \mathrm{Data}(p) = (m+1)(n+1) + 2.\end{aligned}$$

ϵ-solution: an $x \in \mathbf{R}^n$ such that

$$\begin{aligned}\|x\|_\infty &\le R,\\ a_i^T x &\le b_i + \epsilon,\ i = 1, ..., m,\\ c^T x &\le \mathrm{Opt}(p) + \epsilon\end{aligned}$$

(as always, the optimal value of an infeasible problem is $+\infty$).

Newton complexity of ϵ-solution: [14)]

$$\mathrm{Compl}^{\mathrm{Nwt}}(p,\epsilon) = O(1)\sqrt{m+n}\mathrm{Digits}(p,\epsilon),$$

where

$$\mathrm{Digits}(p,\epsilon) = \ln\left(\frac{\mathrm{Size}(p) + \|\mathrm{Data}(p)\|_1 + \epsilon^2}{\epsilon}\right)$$

is the number of accuracy digits in ϵ-solution, see Section 4.1.2.

Arithmetic complexity of ϵ-solution:

$$\mathrm{Compl}(p,\epsilon) = O(1)(m+n)^{3/2}n^2\mathrm{Digits}(p,\epsilon).$$

4.6.2 Complexity of $\mathcal{CQP}_b$

Family of problems:

Problem instance: a program

$$\min_x \left\{c^T x : \|A_i x + b_i\|_2 \le c_i^T x + d_i,\ i = 1,...,m;\ \|x\|_2 \le R\right\} \quad \left[\begin{array}{c} x \in \mathbf{R}^n \\ b_i \in \mathbf{R}^{k_i} \end{array}\right] \qquad (p)$$

Data:

$$\mathrm{Data}(P) = [m; n; k_1, ..., k_m; c; A_1, b_1, c_1, d_1; ...; A_m, b_m, c_m, d_m; R],$$
$$\mathrm{Size}(p) \equiv \dim \mathrm{Data}(p) = (m + \sum_{i=1}^{m} k_i)(n+1) + m + n + 3.$$

ϵ-solution: an $x \in \mathbf{R}^n$ such that

$$\begin{array}{rcl} \|x\|_2 & \le & R, \\ \|A_i x + b_i\|_2 & \le & c_i^T x + d_i + \epsilon,\ i = 1,...,m, \\ c^T x & \le & \mathrm{Opt}(p) + \epsilon. \end{array}$$

Newton complexity of ϵ-solution:

$$\mathrm{Compl}^{\mathrm{Nwt}}(p,\epsilon) = O(1)\sqrt{m+1}\mathrm{Digits}(p,\epsilon).$$

Arithmetic complexity of ϵ-solution:

$$\mathrm{Compl}(p,\epsilon) = O(1)(m+1)^{1/2}n(n^2 + m + \sum_{i=0}^{m} k_i^2)\mathrm{Digits}(p,\epsilon).$$

4.6.3 Complexity of $\mathcal{SDP}_b$

Family of problems:

[14)] In what follows, the precise meaning of a statement "the Newton/arithmetic complexity of finding ϵ-solution of an instance (p) does not exceed N" is as follows: as applied to the input $(\mathrm{Data}(p), \epsilon)$, the method underlying our bound terminates in no more than N steps (respectively, N arithmetic operations) and outputs either a vector, which is an ϵ-solution to the instance, or the correct conclusion "(p) is infeasible".

Problem instance: a program

$$\min_x \left\{ c^T x : A_0 + \sum_{j=1}^n x_j A_j \succeq 0,\ \|x\|_2 \le R \right\} \quad [x \in \mathbf{R}^n], \tag{p}$$

where A_j, $j = 0, 1, ..., n$, are symmetric block-diagonal matrices with m diagonal blocks $A_j^{(i)}$ of sizes $k_i \times k_i$, $i = 1, ..., m$.

Data:

$$\begin{aligned} &\mathrm{Data}(p) = [m; n; k_1, ...k_m; c; A_0^{(1)}, ..., A_0^{(m)}; ...; A_n^{(1)}, ..., A_n^{(m)}; R], \\ &\mathrm{Size}(p) \equiv \dim \mathrm{Data}(P) = \left(\sum_{i=1}^m \tfrac{k_i(k_i+1)}{2} \right)(n+1) + m + n + 3. \end{aligned}$$

ϵ-solution: an x such that

$$\begin{aligned} \|x\|_2 &\le R, \\ A_0 + \sum_{j=1}^n x_j A_j &\succeq -\epsilon I, \\ c^T x &\le \mathrm{Opt}(p) + \epsilon. \end{aligned}$$

Newton complexity of ϵ-solution:

$$\mathrm{Compl}^{\mathrm{Nwt}}(p, \epsilon) = O(1)(1 + \sum_{i=1}^m k_i)^{1/2} \mathrm{Digits}(p, \epsilon).$$

Arithmetic complexity of ϵ-solution:

$$\mathrm{Compl}(p, \epsilon) = O(1)(1 + \sum_{i=1}^m k_i)^{1/2} n(n^2 + n \sum_{i=1}^m k_i^2 + \sum_{i=1}^m k_i^3) \mathrm{Digits}(p, \epsilon).$$

4.7 Concluding remarks

We have discussed IP methods for LP, CQP and SDP as "mathematical animals", with emphasis on the ideas underlying the algorithms and on the theoretical complexity bounds ensured by the methods. Now it is time to say a couple of words on software implementations of IP algorithms and on practical performance of the resulting codes.

As far as the performance of recent IP software is concerned, the situation heavily depends on whether we are speaking about codes for LP, or those for CQP and SDP.

• There exists extremely powerful commercial IP software for LP, capable to handle reliably really large-scale LP's and quite competitive with the best Simplex-type codes for Linear Programming. E.g., one of the best modern LP solvers – CPLEX – allows user to choose between a Simplex-type and IP modes of execution, and in many cases the second option reduces the running time by orders of magnitudes. With a state-of-the-art computer, CPLEX is capable to solve routinely real-world LP's with tens and hundreds thousands of variables and constraints; in the case of favourable structured constraint matrices, the numbers of variables and constraints can become as large as few millions.

• There already exists a very powerful commercial software for CQP – `MOSEK` (Erling Andersen, `http://www.mosek.com`). I would say that as far as LP (and even mixed integer programming) are concerned, `MOSEK` compares favourable to `CPLEX`, and it allows to solve really large CQP's of favourable structure.

• For the time being, IP software for SDP's is not as well-polished, reliable and powerful as the LP one. I would say that the codes available for the moment are capable to solve SDP's with no more than 1,000 – 1,500 design variables.

There are two groups of reasons making the power of SDP software available for the moment that inferior as compared to the capabilities of interior point LP and CQP solvers – the "historical" and the "intrinsic" ones. The "historical" aspect is simple: the development of IP software for LP, on one hand, and for SDP, on the other, has started, respectively, in the mid-eighties and the mid-nineties; for the time being (2002), this is definitely a difference. Well, being too young is the only shortcoming which for sure passes away... Unfortunately, there are intrinsic problems with IP algorithms for large-scale (many thousands of variables) SDP's. Recall that the influence of the size of an SDP/CQP program on the complexity of its solving by an IP method is twofold:

– first, the size affects the Newton complexity of the process. Theoretically, the number of steps required to reduce the duality gap by a constant factor, say, factor 2, is proportional to $\sqrt{\theta(K)}$ ($\theta(K)$ is twice the total # of conic quadratic inequalities for CQP and the total row size of LMI's for SDP). Thus, we could expect an unpleasant growth of the iteration count with $\theta(K)$. Fortunately, the iteration count for good IP methods usually is much less than the one given by the worst-case complexity analysis and is typically about few tens, independently of $\theta(K)$.

– second, the larger is the instance, the larger is the system of linear equations one should solve to generate new primal (or primal-dual) search direction, and, consequently, the larger is the computational effort per step (this effort is dominated by the necessity to assemble and to solve the linear system). Now, the system to be solved depends, of course, on what is the IP method we are speaking about, but it newer is simpler (and for most of the methods, is not more complicated as well) than the system (4.5.8) arising in the primal path-following method:

$$\underbrace{\mathcal{A}^*[\nabla^2 K(\bar{X})]\mathcal{A}}_{\mathcal{H}}\Delta x = \underbrace{-[t_+c + \mathcal{A}^*\nabla K(\bar{X})]}_{h}. \qquad \text{(Nwt)}$$

The size n of this system is exactly the design dimension of problem (CP).

In order to process (Nwt), one should assemble the system (compute $\mathcal{H}$ and h) and then solve it. Whatever is the cost of assembling (Nwt), you should be able to store the resulting matrix $\mathcal{H}$ in memory and to factorize the matrix in order to get the solution. *Both these problems – storing and factorizing $\mathcal{H}$ – become prohibitively expensive when $\mathcal{H}$ is a large dense*[15)] *matrix.* (Think how happy you will be with the necessity to store $\frac{5000\times 5001}{2} = 12,502,500$ reals representing a dense 5000×5000 symmetric matrix $\mathcal{H}$ and with the necessity to perform $\approx \frac{5000^3}{6} \approx 2.08 \times 10^{10}$ arithmetic operations to find its Choleski factor).

The necessity to assemble and to solve large-scale systems of linear equations is intrinsic for IP methods as applied to large-scale optimization programs, and in this respect there is no difference between LP and CQP, on one hand, and SDP, on the other hand. The difference is in how difficult is to handle these large-scale linear systems. In real life LP's-CQP's-SDP's, the structure of the data allows to assemble (Nwt) at a cost negligibly small as compared to the cost of factorizing $\mathcal{H}$, which is a good news. Another good news is that in typical real world LP's, and to some extent for real-world CQP's, $\mathcal{H}$ turns out to be "very well-structured", which reduces dramatically the expenses required by factorizing the matrix and storing the Choleski factor. All practical IP solvers for LP and CQP utilize these favourable properties of real life problems, and this is where their ability to solve problems with tens/hundreds thousands of variables and constraints comes from. Spoil the structure of the problem – and an IP method will be unable to solve an LP with just few thousands of variables. Now, in contrast to real life LP's and CQP's, real life SDP's typically result in dense matrices $\mathcal{H}$, and this is where severe limitations on the sizes of "tractable in practice" SDP's come from. In this respect, real life CQP's are somewhere in-between LP's and SDP's, so that the sizes of "tractable in practice" CQP's could be significantly larger than in the case of SDP's.

It should be mentioned that assembling matrices of the linear systems we are interested in and solving these systems by the standard Linear Algebra techniques is not the only possible way to implement an IP method. Another option is to solve these linear systems by iterative methods. With this approach, all we need to solve a system like (Nwt) is a possibility to multiply a given vector by the matrix of the system, and this does *not* require assembling and storing in memory the matrix itself. E.g., to multiply a

[15)] I.e., with $O(n^2)$ nonzero entries.

vector Δx by $\mathcal{H}$, we can use the multiplicative representation of $\mathcal{H}$ as presented in (Nwt). *Theoretically*, the outlined iterative schemes, as applied to real life SDP's, allow to reduce by orders of magnitudes the arithmetic cost of building search directions and to avoid the necessity to assemble and store huge dense matrices, which is an extremely attractive opportunity. The difficulty, however, is that the iterative schemes are much more affected by rounding errors that the usual Linear Algebra techniques; as a result, *for the time being* "iterative-Linear-Algebra-based" implementation of IP methods is no more than a challenging goal.

Although the sizes of SDP's which can be solved with the existing codes are not that impressive as those of LP's, the possibilities offered to a practitioner by SDP IP methods could hardly be overestimated. Just ten years ago we could not even dream of solving an SDP with more than few tens of variables, while today we can solve routinely 20-25 times larger SDP's, and we have all reasons to believe in further significant progress in this direction.

4.8 Exercises: Around the Ellipsoid method

There are two natural ways to define an ellipsoid W in $\mathbf{R}^n$. The first is to represent W as the set defined by a convex quadratic constraint, namely, as

$$W = \{x \in \mathbf{R}^n \mid (x-c)^T A(x-c) \leq 1\} \tag{4.8.1}$$

A being a symmetric positive definite $n \times n$ matrix and c being a point in $\mathbf{R}^n$ (the center of the ellipsoid).

The second way is to represent W as the image of the unit Euclidean ball under an affine invertible mapping, i.e., as

$$W = \{x = Bu + c \mid u^T u \leq 1\}, \tag{4.8.2}$$

where B is an $n \times n$ nonsingular matrix and c is a point from $\mathbf{R}^n$.

Exercise 4.1 *Prove that the above definitions are equivalent: if $W \subset \mathbf{R}^n$ is given by (4.8.1), then W can be represented by (4.8.2) with B chosen according to*

$$A = (B^{-1})^T B^{-1}$$

(e.g., with B chosen as $A^{-1/2}$). Vice versa, if W is represented by (4.8.2), then W can be represented by (4.8.1), where one should set

$$A = (B^{-1})^T B^{-1}.$$

Note that the (positive definite symmetric) matrix A involved into (4.8.1) is uniquely defined by W (why?); in contrast to this, a nonsingular matrix B involved into (4.8.2) is defined by W up to a right orthogonal factor: the matrices B and B' define the same ellipsoid if and only if $B' = BU$ with an orthogonal $n \times n$ matrix U (why?)

From the second description of an ellipsoid it immediately follows that

if

$$W = \{x = Bu + c \mid u \in \mathbf{R}^n, u^T u \leq 1\}$$

is an ellipsoid and

$$x \mapsto p + B'x$$

is an invertible affine transformation of $\mathbf{R}^n$ (so that B' is a nonsingular $n \times n$ matrix), then the image of W under the transformation also is an ellipsoid.

Indeed, the image is nothing but

$$W' = \{x = B'Bu + (p + B'c) \mid u \in \mathbf{R}^n, u^T u \leq 1\},$$

the matrix $B'B$ being nonsingular along with B and B'. It is also worthy of note that

for any ellipsoid

$$W = \{x = Bu + c \mid u \in \mathbf{R}^n, u^T u \leq 1\}$$

there exists an invertible affine transformation of $\mathbf{R}^n$*, e.g., the transformation*

$$x \mapsto B^{-1}x - B^{-1}c,$$

which transforms the ellipsoid exactly into the unit Euclidean ball

$$V = \{u \in \mathbf{R}^n \mid u^T u \leq 1\}.$$

In what follows we mainly focus on various volume-related issues; to avoid complicated constant factors, it is convenient to take, as the volume unit, the volume of the unit Euclidean ball V in $\mathbf{R}^n$ rather than the volume of the unit cube. The volume of a body[16)] Q measured in this unit, i.e., the ratio

$$\frac{\mathrm{Vol}_n(Q)}{\mathrm{Vol}_n(V)},$$

Vol_n being the usual Lebesque volume in $\mathbf{R}^n$, will be denoted $\mathrm{vol}_n(Q)$ (we omit the subscript n if the value of n is clear from the context).

Exercise 4.2 *Prove that if* W *is an ellipsoid in* $\mathbf{R}^n$ *given by (4.8.2), then*

$$\mathrm{vol}(W) = |\mathrm{Det}B|, \tag{4.8.3}$$

and if W *is given by (4.8.1), then*

$$\mathrm{vol}(W) = |\mathrm{Det}A|^{-1/2}. \tag{4.8.4}$$

Our local goal is to prove the following statement:

Let Q *be a convex body in* $\mathbf{R}^n$ *(i.e., a closed and bounded convex set with a nonempty interior). Then there exist ellipsoids containing* Q*, and among these ellipsoids there is one with the smallest volume. This ellipsoid is unique; it is called the outer extremal ellipsoid associated with* Q*. Similarly, there exist ellipsoids contained in* Q*, and among these ellipsoids there is one with the largest volume. This ellipsoid is unique; it is called the inner extremal ellipsoid associated with* Q*.*

In fact we are not too interested in the uniqueness of the extremal ellipsoids (and you may try to prove the uniqueness yourself); what actually is of interest is the existence and some important properties of the extremal ellipsoids.

Exercise 4.3 *Prove that if* Q *is a closed and bounded convex body in* $\mathbf{R}^n$*, then there exist ellipsoids containing* Q *and among these ellipsoids there is (at least) one with the smallest volume.*

Exercise 4.4 *Prove that if* Q *is a closed and bounded convex body in* $\mathbf{R}^n$*, then there exist ellipsoids contained in* Q *and among these ellipsoids there is (at least) one with the largest volume.*

Note that extremal ellipsoids associated with a closed and bounded convex body Q "accompany Q under affine transformations": if $x \mapsto Ax + b$ is an invertible affine transformation and Q' is the image of Q under this transformation, then the image W' of an extremal outer ellipsoid W associated with Q (note the article: we has not proved the uniqueness!) is an extremal outer ellipsoid associated with Q', and similarly for (an) extremal inner ellipsoid.

The indicated property is, of course, an immediate consequence of the facts that affine images of ellipsoids are again ellipsoids and that the ratio of volumes remains invariant under an affine transformation of the space.

In what follows we focus on outer extremal ellipsoids. Useful information can be obtained from investigating these ellipsoids for "simple parts" of an Euclidean ball.

[16)] in what follows "body" means a set with a nonempty interior

Exercise 4.5 *Let $n > 1$,*

$$V = \{x \in \mathbf{R}^n \mid |x|_2 \equiv \left(\sum_{i=1}^n x_i^2\right)^{1/2} \leq 1\}$$

be the unit Euclidean ball, let e be a unit vector in $\mathbf{R}^n$ and let

$$V_\alpha = \{x \in V \mid e^T x \geq \alpha\},\ \alpha \in [-1, 1]$$

(V_α is what is called a "spherical hat").

Prove that if

$$-\frac{1}{n} < \alpha < 1,$$

then the set V_α can be covered by an ellipsoid W of the volume

$$\mathrm{vol}_n(W) \leq \{\frac{n^2}{n^2-1}\}^{n/2}\sqrt{\frac{n-1}{n+1}}(1-\alpha^2)^{(n-1)/2}(1-\alpha) < 1 = \mathrm{vol}_n(V);$$

W is defined as

$$W = \{x = \frac{1+n\alpha}{n+1}e + Bu \mid u^T u \leq 1\},$$

where

$$B = \left\{(1-\alpha^2)\frac{n^2}{n^2-1}\right\}^{1/2}\left(I - \beta e e^T\right), \quad \beta = 1 - \sqrt{\frac{(1-\alpha)(n-1)}{(1+\alpha)(n+1)}}$$

In fact the ellipsoid given by the latter exercise is the extremal outer ellipsoid associated with V_α.

Looking at the result stated by the latter exercise, one may make a number of useful conclusions.

1. When $\alpha = 0$, i.e., when the spherical hat V_α is a half-ball, we have

$$\mathrm{vol}_n(W) = \left\{1 + \frac{1}{n^2-1}\right\}^{n/2}\sqrt{1 - \frac{2}{n-1}} \leq$$

$$\leq \left\{\exp\{1/(n^2-1)\}\right\}^{n/2}\exp\{-1/(n-1)\} =$$

$$= \exp\{-\frac{n+2}{2(n^2-1)}\} < \exp\{-\frac{1}{2n-2}\} = \exp\{-\frac{1}{2n-2}\}\mathrm{vol}_n(V);$$

thus, for the case of $\alpha = 0$ (and, of course, for the case of $\alpha > 0$) we may cover V_α by an ellipsoid with the volume $1 - O(1/n)$ times less than that one of V. In fact the same conclusion (with another absolute constant factor $O(1)$) holds true when α is negative (so that the spherical hat is greater than half-ball), but "not too negative", say, when $\alpha \geq -\frac{1}{2n}$.

2. In order to cover V_α by an ellipsoid of absolute constant times less volume than that one of V we need α to be positive of order $O(n^{-1/2})$ or greater. In this case, "small" covering of V_α is already given by the Euclidean ball of the radius $\sqrt{1-\alpha^2}$ centered at the point αe (which, anyhow, is *not* the optimal covering presented in exercise 4.5).

Exercise 4.6 *Let V be the unit Euclidean ball in $\mathbf{R}^n$, e be a unit vector and let $\alpha \in (0, 1)$. Consider the "symmetric spherical stripe"*

$$V^\alpha = \{x \in V \mid -\alpha \leq e^T x \leq \alpha\}.$$

Prove that if $0 < \alpha < 1/\sqrt{n}$ then V^α can be covered by an ellipsoid W with the volume

$$\mathrm{vol}_n(W) \leq \alpha\sqrt{n}\left\{\frac{n(1-\alpha^2)}{n-1}\right\}^{(n-1)/2} < 1 = \mathrm{vol}_n(V).$$

Find an explicit representation of the ellipsoid.

We see that in order to cover a symmetric spherical stripe of the unit Euclidean ball V by an ellipsoid of volume less than that one of V, it suffices to have the "half-thickness" α of the stripe to be $< 1/\sqrt{n}$, which again fits our observation (exercise 1.9) that basically all volume of the unit n-dimensional Euclidean ball is concentrated in the $O(1/\sqrt{n})$ neighbourhood of its "equator" - the cross-section of the ball and a hyperplane passing through the center of the ball. A useful exercise is to realize when a non-symmetric spherical stripe

$$V^{\alpha,\beta} = \{x \in V \mid -\alpha \leq e^T x \leq \beta\}$$

of the (centered at the origin) unit Euclidean ball V can be covered by an ellipsoid of volume less than that one of V.

The results of exercises 4.5 and 4.6 imply a number of important geometrical consequences.

Exercise 4.7 *Prove the following theorem of Fritz John:*

Let Q be a closed and bounded convex body in $\mathbf{R}^n$. Then

(i) *Q can be covered by an ellipsoid W in such a way that the concentric to W n times smaller ellipsoid*

$$W' = (1 - \frac{1}{n})c + \frac{1}{n}W$$

(c is the center of W) is contained in Q. One can choose as W the extremal outer ellipsoid associated with Q.

(ii) *If, in addition, Q is central-symmetric with respect to certain point c, then the above result can be improved: Q can be covered by an ellipsoid W centered at c in such a way that the concentric to W $\sqrt{n}$ times smaller ellipsoid*

$$W'' = (1 - \frac{1}{\sqrt{n}})c + \frac{1}{\sqrt{n}}W$$

is contained in Q.

Note that the constants n and $\sqrt{n}$ in the Fritz John Theorem are sharp; an extremal example for (i) is a simplex, and for (ii) - a cube.

Here are several nice geometrical consequences of the Fritz John Theorem:

Let Q be a closed and bounded convex body in $\mathbf{R}^n$. Then

1. *There exist a pair of concentric homothetic with respect to their common center parallelotopes p, P with homothety coefficient equal to $n^{-3/2}$ such that $p \subset Q \subset P$; in other words, there exists an invertible affine transformation of the space such that the image Q' of Q under this transformation satisfies the inclusions*

$$\{x \in \mathbf{R}^n \mid \|x\|_\infty \leq \frac{1}{n^{3/2}}\} \subset Q' \subset \{x \in \mathbf{R}^n \mid \|x\|_\infty \leq 1\};$$

here

$$\|x\|_\infty = \max_{1 \leq i \leq n} |x_i|$$

is the uniform norm of x.

Indeed, from the Fritz John Theorem it follows that there exists an invertible affine transformation resulting in

$$\{x \mid |x|_2 \leq 1/n\} \subset Q' \subset \{x \mid |x|_2 \leq 1\},$$

Q' being the image of Q under the transformation (it suffices to transform the outer extremal ellipsoid associated with Q into the unit Euclidean ball centered at the origin). It remains to note that the smaller Euclidean ball in the above chain of inclusions contains the cube $\{x \mid \|x\|_\infty \leq n^{-3/2}\}$ and the larger one is contained in the unit cube.

2. *If Q is central-symmetric, then the parallelotopes mentioned in 1. can be chosen to have the same center, and the homothety coefficient can be improved to $1/n$; in other words, there*

exists an invertible affine transformation of the space which makes the image Q' of Q central symmetric with respect to the origin and ensures the inclusions

$$\{x \mid \|x\|_\infty \le \frac{1}{n}\} \subset Q' \subset \{x \mid \|x\|_\infty \le 1\}.$$

The statement is given by the reasoning completely similar to that one used for 1., up to the fact that now we should refer to item (ii) of the Fritz John Theorem.

3. *Any norm $\|\cdot\|$ on $\mathbf{R}^n$ can be approximated, within factor $\sqrt{n}$, by a Euclidean norm: given $\|\cdot\|$, one can find a Euclidean norm*

$$|x|_A = (x^T A x)^{1/2},$$

A being a symmetric positive definite $n \times n$ matrix, in such a way that

$$\frac{1}{\sqrt{n}}|x|_A \le \|x\| \le |x|_A$$

for any $x \in \mathbf{R}^n$.

Indeed, let $\mathcal{B} = \{x \mid \|x\| \le 1\}$ be the unit ball with respect to the norm $\|\cdot\|$; this is a closed and bounded convex body, which is central symmetric with respect to the origin. By item (ii) of the Fritz John Theorem, there exists a centered at the origin ellipsoid

$$W = \{x \mid x^T A x \le n\}$$

(A is an $n \times n$ symmetric positive definite matrix) which contains $\mathcal{B}$, while the ellipsoid

$$\{x \mid x^T A x \le 1\}$$

is contained in $\mathcal{B}$; this latter inclusion means exactly that

$$|x|_A \le 1 \Rightarrow x \in \mathcal{B} \Leftrightarrow \|x\| \le 1,$$

i.e., means that $|x|_A \ge \|x\|$. The inclusion $\mathcal{B} \subset W$, by similar reasons, implies that $\|x\| \ge n^{-1/2}|x|_A$.

Remark 4.8.1 The third of the indicated consequences says that any norm on $\mathbf{R}^n$ can be approximated, within constant factor $\sqrt{n}$, by an appropriately chosen Euclidean norm. It turns out that the quality of approximation can be done much better, if we would be satisfied by approximating the norm not at the whole space, but at a properly chosen subspace. Namely, there exists a marvelous and important theorem of Dvoretski which is as follows:

there exists a function $m(n,\varepsilon)$ of positive integer n and positive real ε with the following properties: first,

$$\lim_{n\to\infty} m(n,\varepsilon) = +\infty$$

and,

second, whenever $\|\cdot\|$ is a norm on $\mathbf{R}^n$, one can indicate a $m(n,\varepsilon)$-dimensional subspace $E \subset \mathbf{R}^n$ and a Euclidean norm $|\cdot|_A$ on $\mathbf{R}^n$ such that $|\cdot|_A$ approximates $\|\cdot\|$ on E within factor $1+\varepsilon$:

$$(1-\varepsilon)|x|_A \le \|x\| \le (1+\varepsilon)|x|_A, \quad x \in E.$$

In other words, the Euclidean norm is "marked by God": for any given integer k an arbitrary normed linear space contains an "almost Euclidean" k-dimensional subspace, provided that the dimension of the space is large enough.

Lecture 5

Simple methods for extremely large-scale problems

5.1 Motivation

The polynomial time Interior Point methods, same as all other polynomial time methods for Convex Programming known so far, have a not that pleasant common feature: the arithmetic cost C of an iteration in such a method grows *nonlinearly* with the design dimension n of the problem, unless the latter possesses a very favourable structure. E.g., in IP methods, an iteration requires solving a system of linear equations with (at least) n unknowns. To solve this auxiliary problem, it costs *at least* $O(n^2)$ operations (with the traditional Linear Algebra – even $O(n^3)$ operations), except for the cases when the matrix of the system is very sparse and, moreover, possesses a well-structured sparsity pattern. The latter indeed is the case when solving most of LPs of real-world origin, but nearly *never* is the case for, e.g., SDPs. For other known polynomial time methods, the situation is similar – the arithmetic cost of an iteration, even in the case of extremely simple objectives and feasible sets, is at least $O(n^2)$. With n of order of tens and hundreds of thousands, the computational effort of $O(n^2)$, not speaking about $O(n^3)$, operations per iteration becomes prohibitively large – basically, you never will finish the very first iteration of your method... On the other hand, the design dimensions of tens and hundreds of thousands is exactly what is met in many applications, like SDP relaxations of combinatorial problems involving large graphs or Structural Design (especially for 3D structures). As another important application of this type, consider 3D Medical Imaging problem arising in Positron Emission Tomography.

Positron Emission Tomography (PET) is a powerful, non-invasive, medical diagnostic imaging technique for measuring the metabolic activity of cells in the human body. It has been in clinical use since the early 1990s. PET imaging is unique in that it shows the *chemical functioning* of organs and tissues, while other imaging techniques - such as X-ray, computerized tomography (CT) and magnetic resonance imaging (MRI) - show *anatomic structures*.

A PET scan involves the use of a *radioactive tracer* – a fluid with a small amount of a radioactive material which has the property of emitting positrons. When the tracer is administered to a patient, either by injection or inhalation of gas, it distributes within the body. For a properly chosen tracer, this distribution "concentrates" in desired locations, e.g., in the areas of high metabolic activity where cancer tumors can be expected.

The radioactive component of the tracer disintegrates, emitting positrons. Such a positron nearly immediately annihilates with a near-by electron, giving rise to two photons flying at the speed of light off the point of annihilation in nearly opposite directions along a line with a completely random orientation (i.e., uniformly distributed in space). They penetrate the surrounding tissue and are registered outside the patient by a PET scanner consisting of circular arrays (*rings*) of gamma radiation detectors. Since the two gamma rays are emitted simultaneously and travel in almost exactly opposite directions, we can

say a lot on the location of their source: when a pair of opposing detectors register high-energy photons within a short ($\sim 10^{-8}$sec) timing window ("a coincidence event"), we know that the photons came from a disintegration act, and that the act took place on the line ("line of response" (LOR)) linking the detectors. The measured data set is the collection of numbers of coincidences counted by different pairs of detectors ("bins"), and the problem is to recover from these measurements the 3D density of the tracer.

The mathematical model of the process, after appropriate discretization, is

$$y = P\lambda + \xi,$$

where

- $\lambda \geq 0$ is the vector representing the (discretized) density of the tracer; the entries of λ are indexed by *voxels* – small cubes into which we partition the field of view, and λ_j is the mean density of the tracer in voxel j. Typically, the number n of voxels is in the range from 3×10^5 to 3×10^6, depending on the resolution of the discretization grid;
- y are the measurements; the entries in y are indexed by *bins* – pairs of detectors, and y_i is the number of coincidences counted by i-th pair of detectors. Typically, the dimension m of y – the total number of bins – is millions (at least 3×10^6);
- P is the *projection matrix*; its entries p_{ij} are the probabilities for a LOR originating in voxel j to be registered by bin i. These probabilities are readily given by the geometry of the scanner;
- ξ is the measurement noise coming mainly from the fact that all physical processes underlying PET are random. The standard statistical model for PET implies that y_i, $i = 1, ..., m$, are independent Poisson random variables with the expectations $(P\lambda)_i$.

The problem we are interested in is to recover tracer's density λ given measurements y. As far as the quality of the result is concerned, the most attractive reconstruction scheme is given by the standard in Statistics *Likelihood Ratio* maximization: denoting $p(\cdot|\lambda)$ the density of the probability distribution of the measurements, coming from λ, w.r.t. certain dominating distribution, the estimate of the unknown true value λ_* of λ is

$$\widehat{\lambda} = \operatorname*{argmin}_{\lambda \geq 0} p(y|\lambda),$$

where y is the vector of measurements.

For the aforementioned Poisson model of PET, building the Maximum Likelihood estimate is equivalent to solving the optimization problem

$$\min_{\lambda} \left\{ \sum_{j=1}^n \lambda_j p_j - \sum_{i=1}^m y_i \ln(\sum_{j=1}^n \lambda_j p_{ij}) : \lambda \geq 0 \right\} \quad \left[p_j = \sum_i p_{ij} \right]. \tag{PET}$$

This is a nicely structured convex program (by the way, polynomially reducible to CQP and even LP). The only difficulty – and a severe one – is in huge sizes of the problem: as it was already explained, the number n of decision variables is at least $300,000$, while the number m of log-terms in the objective is in the range from 3×10^6 to 25×10^6.

At the present level of our knowledge, the design dimension n of order of tens and hundreds of thousands rules out the possibility to solve a *nonlinear* convex program, even a well-structured one, by polynomial time methods because of at least quadratic in n "blowing up" the arithmetic cost of an iteration. When n is really large, all we can use are simple methods with linear in n cost of an iteration. As a byproduct of this restriction, *we cannot utilize anymore our knowledge of the analytic structure of the problem*, since all known for the time being ways of doing so are too expensive, provided that n is large. As a result, we are enforced to restrict ourselves with *black-box-oriented* optimization techniques – those which use solely the possibility to compute the values and the (sub)gradients of the objective and the constraints at a point. In Convex Optimization, two types of "cheap" black-box-oriented optimization techniques are known:

- techniques for *unconstrained* minimization of *smooth* convex functions (Gradient Descent, Conjugate Gradients, quasi-Newton methods with restricted memory, etc.);
- subgradient-type techniques for *nonsmooth* convex programs, including constrained ones.

Since the majority of applications are constrained, we restrict our exposition to the techniques of the second type. We start with investigating of what, *in principle*, can be expected of black-box-oriented optimization techniques.

5.2 Information-based complexity of Convex Programming

Black-box-oriented methods and Information-based complexity. Consider a Convex Programming program in the form

$$\min_x \{f(x) : x \in X\}, \tag{CP}$$

where X is a convex compact set in $\mathbf{R}^n$ and the objective f is a continuous convex function on $\mathbf{R}^n$. Let us fix a family $\mathcal{P}(X)$ of convex programs (CP) with X common for all programs from the family, so that such a program can be identified with the corresponding objective, and the family itself is nothing but certain family of convex functions on $\mathbf{R}^n$. We intend to explain what is the *Information-based complexity* of $\mathcal{P}(X)$ – informally, complexity of the family w.r.t. "black-box-oriented" methods. We start with defining such a method as a routine $\mathcal{B}$ as follows:

1. When starting to solve (CP), $\mathcal{B}$ is given an accuracy $\epsilon > 0$ to which the problem should be solved and knows that the problem belongs to a given family $\mathcal{P}(X)$. However, $\mathcal{B}$ does *not* know what is the particular problem it deals with.
2. In course of solving the problem, $\mathcal{B}$ has an access to the First Order oracle for f. This oracle is capable, given on input a point $x \in \mathbf{R}^n$, to report on output what is the value $f(x)$ and a subgradient $f'(x)$ of f at x.

   $\mathcal{B}$ generates somehow a sequence of *search points* $x_1, x_2, ...$ and calls the First Order oracle to get the values and the subgradients of f at these points. The rules for building x_t can be arbitrary, *except for the fact that they should be casual:* x_t can depend only on the information $f(x_1), f'(x_1), ..., f(x_{t-1}), f'(x_{t-1})$ on f accumulated by $\mathcal{B}$ at the first $t-1$ steps.
3. After certain number $T = T_{\mathcal{B}}(f, \epsilon)$ of calls to the oracle, $\mathcal{B}$ terminates and outputs the result $z_{\mathcal{B}}(f, \epsilon)$. This result again should depend solely on the information on f accumulated by $\mathcal{B}$ at the T search steps, and must be an ϵ-solution to (CP), i.e.,

$$z_{\mathcal{B}}(f, \epsilon) \in X \ \& \ f(z_{\mathcal{B}}(f, \epsilon)) - \min_X f \leq \epsilon.$$

We measure the complexity of $\mathcal{P}(X)$ w.r.t. a solution method $\mathcal{B}$ by the function

$$\mathrm{Compl}_{\mathcal{B}}(\epsilon) = \max_{f \in \mathcal{P}(X)} T_{\mathcal{B}}(f, \epsilon)$$

– by the minimal number of steps in which $\mathcal{B}$ is capable to solve within accuracy ϵ every instance of $\mathcal{P}(X)$. Finally, the *Information-based complexity* of the family $\mathcal{P}(X)$ of problems is defined as

$$\mathrm{Compl}(\epsilon) = \min_{\mathcal{B}} \mathrm{Compl}_{\mathcal{B}}(\epsilon),$$

the minimum being taken over all solution methods. Thus, the relation $\mathrm{Compl}(\epsilon) = N$ means, first, that there exists a solution method $\mathcal{B}$ capable to solve within accuracy ϵ every instance of $\mathcal{P}(X)$ in no more than N calls to the First Order oracle, and, second, that for every solution method $\mathcal{B}$ there exists an instance of $\mathcal{P}(X)$ such that $\mathcal{B}$ solves the instance within the accuracy ϵ in at least N steps.

Note that as far as black-box-oriented optimization methods are concerned, the information-based complexity $\mathrm{Compl}(\epsilon)$ of a family $\mathcal{P}(X)$ is a *lower* bound on "actual" computational effort, whatever it means, sufficient to find ϵ-solution to every instance of the family.

Main results on Information-based complexity of Convex Programming can be summarized as follows. Let X be a solid in $\mathbf{R}^n$ (a convex compact set with a nonempty interior), and let $\mathcal{P}(X)$ be the family of all convex functions on $\mathbf{R}^n$ normalized by the condition

$$\max_X f - \min_X f \le 1. \tag{5.2.1}$$

For this family,

I. Complexity of finding high-accuracy solutions in fixed dimension is independent of the geometry of X. Specifically,

$$\begin{array}{ll} \forall(\epsilon \le \epsilon(X)): & O(1)n\ln\left(2+\frac{1}{\epsilon}\right) \le \mathrm{Compl}(\epsilon); \\ \forall(\epsilon > 0): & \mathrm{Compl}(\epsilon) \le O(1)n\ln\left(2+\frac{1}{\epsilon}\right), \end{array} \tag{5.2.2}$$

where

- $O(1)$ are appropriately chosen positive absolute constants,
- $\epsilon(X)$ depends on the geometry of X, but never is less than $\frac{1}{n^2}$, where n is the dimension of X.

II. Complexity of finding solutions of fixed accuracy in high dimensions does depend on the geometry of X. Here are 3 typical results:

(a) Let X be an n-dimensional box: $X = \{x \in \mathbf{R}^n : \|x\|_\infty \le 1\}$. Then

$$\epsilon \le \frac{1}{2} \Rightarrow O(1)n\ln(\frac{1}{\epsilon}) \le \mathrm{Compl}(\epsilon) \le O(1)n\ln(\frac{1}{\epsilon}). \tag{5.2.3}$$

The bounds remain intact when the family is shrunk to include convex Lipschitz continuous, with constant 1/2 w.r.t. the $\|\cdot\|_\infty$-norm, objectives only.

(b) Let X be an n-dimensional ball: $X = \{x \in \mathbf{R}^n : \|x\|_2 \le 1\}$. Then

$$n \ge \frac{1}{\epsilon^2} \Rightarrow \frac{O(1)}{\epsilon^2} \le \mathrm{Compl}(\epsilon) \le \frac{O(1)}{\epsilon^2}. \tag{5.2.4}$$

The bounds remain intact when the family is shrunk to include convex Lipschitz continuous, with constant 1/2 w.r.t. the $\|\cdot\|_2$-norm, objectives only.

(c) Let X be an n-dimensional hyperoctahedron: $X = \{x \in \mathbf{R}^n : \|x\|_1 \le 1\}$. Then

$$n \ge \frac{1}{\epsilon^2} \Rightarrow \frac{O(1)}{\epsilon^2} \le \mathrm{Compl}(\epsilon) \le \frac{O(\ln n)}{\epsilon^2} \tag{5.2.5}$$

(in fact, $O(1)$ in the lower bound can be replaced with $O(\ln n)$, provided that $n >> \frac{1}{\epsilon^2}$). The bounds remain intact when the family is shrunk to include convex Lipschitz continuous, with constant 1/2 w.r.t. the $\|\cdot\|_1$-norm, objectives only.

Since we are interested in extremely large-scale problems, the moral which we can extract from the outlined results is as follows:

• I is discouraging: it says that we have no hope to guarantee high accuracy, like $\epsilon = 10^{-6}$, when solving large-scale problems with black-box-oriented methods; indeed, with $O(n)$ steps per accuracy digit and *at least* $O(n)$ operations per step (this many operations are required already to input a search point to the oracle), the arithmetic cost per accuracy digit is at least $O(n^2)$, which is prohibitively large for really large n.

• II is partly discouraging, partly encouraging. A bad news reported by II is that when X is a box, which is the most typical situation in applications, we have no hope to solve extremely large-scale problems, in a reasonable time, to *guaranteed*, even low, accuracy, since the required number of steps should be at least of order of n. A good news reported by II is that *there exist situations where the complexity of minimizing a convex function to a fixed accuracy is independent, or nearly independent,*

of the design dimension. Of course, the dependence of the complexity bounds in (5.2.4) and (5.2.5) on ϵ is very bad and has nothing in common with being polynomial in $\ln(1/\epsilon)$; however, this drawback is tolerable when we do not intend to get high accuracy. Another drawback is that there are not that many applications where the feasible set is a ball or a hyperoctahedron. Note, however, that in fact we can save the most important for us *upper* complexity bounds in (5.2.4) and (5.2.5) when requiring from X to be a *subset* of a ball, respectively, of a hyperoctahedron, rather than to be the entire ball/hyperoctahedron. This extension is not costless: we should simultaneously strengthen the normalization condition (5.2.1). Specifically, we shall see that

B. The upper complexity bound in (5.2.4) remains valid when $X \subset \{x : \|x\|_2 \leq 1\}$ and

$$\mathcal{P}(X) = \{f : f \text{ is convex and } |f(x) - f(y)| \leq \|x - y\|_2 \ \forall x, y \in X\};$$

S. The upper complexity bound in (5.2.5) remains valid when $X \subset \{x : \|x\|_1 \leq 1\}$ and

$$\mathcal{P}(X) = \{f : f \text{ is convex and } |f(x) - f(y)| \leq \|x - y\|_1 \ \forall x, y \in X\}.$$

Note that the "ball-like" case mentioned in B seems to be rather artificial: the Euclidean norm associated with this case is a very natural mathematical entity, but this is all we can say in its favour. For example, the normalization of the objective in B is that the Lipschitz constant of f w.r.t. $\|\cdot\|_2$ is ≤ 1, or, which is the same, that the vector of the first order partial derivatives of f should, at every point, be of $\|\cdot\|_2$-norm not exceeding 1. In order words, "typical" magnitudes of the partial derivatives of f should become smaller and smaller as the number of variables grows; what could be the reasons for such a strange behaviour? In contrast to this, the normalization condition imposed on f in S is that the Lipschitz constant of f w.r.t. $\|\cdot\|_1$ is ≤ 1, or, which is the same, that the $\|\cdot\|_\infty$-norm of the vector of partial derivatives of f is ≤ 1. In other words, the normalization is that the magnitudes of the first order partial derivatives of f should be ≤ 1, and this normalization is "dimension-independent". Of course, in B we deal with minimization over subsets of the unit ball, while in S we deal with minimization over the subsets of the unit hyperoctahedron, which is much smaller than the unit ball. However, there do exist problems in reality where we should minimize over the standard simplex

$$\Delta_n(1) = \{x \in \mathbf{R}^n : x \geq 0, \sum_i x_i = 1\},$$

which indeed is a subset of the unit hyperoctahedron. For example, it turns out that the PET Image Reconstruction problem (PET) is in fact the problem of minimization over the standard simplex. Indeed, the optimality condition for (PET) reads

$$\lambda_j \left(p_j - \sum_i y_i \frac{p_{ij}}{\sum_\ell p_{i\ell}\lambda_\ell} \right) = 0, \ j = 1, ..., n;$$

summing up these equalities, we get

$$\sum_j p_j \lambda_j = B \equiv \sum_i y_i.$$

It follows that the optimal solution to (PET) remains unchanged when we add to the nonnegativity constraints $\lambda_j \geq 0$ also the constraint $\sum_j p_j \lambda_j = B$. Passing to the new variables $x_j = B^{-1} p_j \lambda_j$, we further convert (PET) to the equivalent form

$$\min_x \left\{ f(x) \equiv -\sum_i y_i \ln(\sum_j q_{ij} x_j) : x \in \Delta_n \right\}, \quad \left[q_{ij} = \frac{B p_{ij}}{p_j} \right] \tag{PET$'$}$$

which is a problem of minimizing a convex function over the standard simplex.

Intermediate conclusion. The discussion above says that this perhaps is a good idea to look for simple convex minimization techniques which, as applied to convex programs (CP) with feasible sets of appropriate geometry, exhibit *dimension-independent* (or nearly dimension-independent) and nearly optimal information-based complexity. We are about to present a family of techniques of this type.

5.3 Methods with Euclidean geometry: Subgradient Descent and Bundle-Level

5.3.1 The simplest of the cheapest – Subgradient Descent

The *Subgradient Descent* method (SD) (N. Shor, 1967) is aimed at solving a convex program

$$f_* = \min_{x\in X} f(x) \tag{5.3.1}$$

where X is a convex compact set in $\mathbf{R}^n$ and f is a Lipschitz continuous on X convex function, is the recurrence

$$x_{t+1} = \Pi_X(x_t - \gamma_t f'(x_t)) \quad [x_1 \in X] \tag{5.3.2}$$

where

- $\gamma_t > 0$ are *stepsizes*
- $\Pi_X(x) = \operatorname{argmin}_{y\in X} \|x-y\|_2^2$ is the standard *projector* on X,
- $f'(x)$ is a *subgradient* of f at x:

$$f(y) \geq f(x) + (y-x)^T f'(x) \quad \forall y \in X.$$

Note: We always assume that $\operatorname{int} X \neq \emptyset$ and that the subgradients $f'(x)$ reported by the First Order oracle at points $x \in X$ satisfy the requirement

$$f'(x) \in \operatorname{cl}\{f'(y) : y \in \operatorname{int} X\}.$$

With this assumption, for every norm $\|\cdot\|$ on $\mathbf{R}^n$ and for every $x \in X$ one has

$$\|f'(x)\|_* \equiv \max_{\xi:\|\xi\|\leq 1} \xi^T f'(x) \leq L_{\|\cdot\|}(f) \equiv \sup_{\substack{x\neq y,\\ x,y\in X}} \frac{|f(x)-f(y)|}{\|x-y\|}, \tag{5.3.3}$$

where

$$\|\xi\|_* = \max_{x:\|x\|\leq 1} \xi^T x \tag{5.3.4}$$

is the norm *conjugate* to the norm $\|\cdot\|$.

When, why and how SD converges?

We start with a simple geometric fact:

Proposition 5.3.1 *Let $X \subset \mathbf{R}^n$ be a closed convex set and $x \in \mathbf{R}^n$. Then the vector $e = x - \Pi_X(x)$ forms an acute angle with every vector of the form $y - \Pi_X(x)$, $y \in X$:*

$$(x-\Pi_X(x))^T(y-\Pi_X(x)) \leq 0 \quad \forall y \in X. \tag{5.3.5}$$

In particular,

$$y \in X \Rightarrow \|y-\Pi_X(x)\|_2^2 \leq \|y-x\|_2^2 - \|x-\Pi_X(x)\|_2^2. \tag{5.3.6}$$

Proof. Indeed, when $y \in X$ and $0 \le t \le 1$, one has

$$\phi(t) = \| \underbrace{[\Pi_X(x) + t(y - \Pi_X(x))]}_{y_t \in X} - x\|_2^2 \ge \|\Pi_X(x) - x\|_2^2 = \phi(0),$$

whence

$$0 \le \phi'(0) = 2(\Pi_X(x) - x)^T (y - \Pi_X(x)).$$

Consequently,

$$\begin{array}{rcl} \|y - x\|_2^2 & = & \|y - \Pi_X(x)\|_2^2 + \|\Pi_X(x) - x\|_2^2 \\ & & +2(y - \Pi_X(x))^T(\Pi_X(x) - x) \\ & \ge & \|y - \Pi_X(x)\|_2^2 + \|\Pi_X(x) - x\|_2^2. \end{array}$$

■

Corollary 5.3.1 *Let problem* (5.3.1) *be solved by SD. Then for every $u \in X$ one has*

$$\gamma_t (x_t - u)^T f'(x_t) \le \underbrace{\frac{1}{2}\|x_t - u\|_2^2}_{d_t} - \underbrace{\frac{1}{2}\|x_{t+1} - u\|_2^2}_{d_{t+1}} + \frac{1}{2}\gamma_t^2 \|f'(x_t)\|_2^2. \tag{5.3.7}$$

Proof. Indeed, by Proposition 5.3.1 we have

$$d_{t+1} \le \frac{1}{2}\|[x_t - u] - \gamma_t f'(x_t)\|_2^2 = d_t - \gamma_t (x_t - u)^T f'(x_t) + \frac{1}{2}\gamma_t^2 \|f'(x_t)\|_2^2.$$

■

Summing up inequalities (5.3.1) over $t = T_0, T_0 + 1, ..., T$, we get

$$\sum_{t=T_0}^{T} \gamma_t (f(x_t) - f(u)) \le \underbrace{d_{T_0} - d_{T+1}}_{\le \Theta} + \sum_{t=T_0}^{T} \frac{1}{2}\gamma_t^2 \|f'(x_t)\|_2^2, \tag{5.3.8}$$

where

$$\Theta = \max_{x,y \in X} \frac{1}{2}\|x - y\|_2^2 \tag{5.3.9}$$

Setting $u = x_* \equiv \operatorname{argmin}_X f$, we arrive at the bound

$$\boxed{\forall (T, T_0, T \ge T_0 \ge 1) : \epsilon_T \equiv \min_{t \le T} f(x_t) - f_* \le \frac{\Theta + \frac{1}{2}\sum\limits_{t=T_0}^{T} \gamma_t^2 \|f'(x_t)\|_2^2}{\sum\limits_{t=T_0}^{T} \gamma_t}} \tag{5.3.10}$$

Relation (5.3.10) allows to arrive at various convergence results.

Example 1: "Divergent Series".

Proposition 5.3.2 *Let $\gamma_t \to 0$ as $t \to \infty$, while $\sum_t \gamma_t = \infty$. Then*

$$\lim_{T \to \infty} \epsilon_T = 0.$$

Proof. Set $T_0 = 1$ and note that

$$\frac{\sum\limits_{t=1}^{T} \gamma_t^2 \|f'(x_t)\|_2^2}{\sum\limits_{t=1}^{T} \gamma_t} \le L^2_{\|\cdot\|_2}(f) \frac{\sum\limits_{t=1}^{T} \gamma_t^2}{\sum\limits_{t=1}^{T} \gamma_t} \to 0, \ T \to \infty.$$

■

Example 2: "Optimal stepsizes".

Proposition 5.3.3 *With the stepsizes*

$$\gamma_t = \frac{\sqrt{2\Theta}}{\|f'(x_t)\|\sqrt{t}} \tag{5.3.11}$$

one has

$$\boxed{\epsilon_T \equiv \min_{t\le T} f(x_t) - f_* \le O(1)\frac{L_{\|\cdot\|_2}(f)\max\limits_{x,y\in X}\|x-y\|_2}{\sqrt{T}},\ T\ge 1} \tag{5.3.12}$$

Proof. Setting $T_0 = \lfloor T/2\rfloor$, we get

$$\epsilon_T \le \frac{\Theta + \Theta\sum\limits_{t=T_0}^{T}\frac{1}{t}}{\sum\limits_{t=T_0}^{T}\frac{\sqrt{\Theta}}{\sqrt{t}\|f'(x_t)\|_2}} \le \frac{\Theta + \Theta\sum\limits_{t=T_0}^{T}\frac{1}{t}}{\sum\limits_{t=T_0}^{T}\frac{\sqrt{\Theta}}{\sqrt{t}L_{\|\cdot\|_2}(f)}} \le L_{\|\cdot\|_2}(f)\sqrt{\Theta}\frac{1+O(1)}{O(1)\sqrt{T}} = O(1)\frac{L_{\|\cdot\|_2}(f)\sqrt{\Theta}}{\sqrt{T}}. \qquad \blacksquare$$

Recalling the definition (5.3.9) of Θ, we can rewrite the efficiency estimate (5.3.12) as

$$\epsilon_T \equiv \min_{1\le t\le T} f(x_t) - f_* \le O(1)\frac{\overbrace{L_{\|\cdot\|_2}(f)\max\limits_{x,y\in X}\|x-y\|_2}^{\mathrm{Var}_{\|\cdot\|_2,X}(f)}}{\sqrt{T}} \tag{5.3.13}$$

We have arrived at efficiency estimate which is *dimension-independent*, provided that the "$\|\cdot\|_2$-variation" of the objective on the feasible domain

$$\mathrm{Var}_{\|\cdot\|_2,X}(f) = L_{\|\cdot\|_2}(f)\max_{x,y\in X}\|x-y\|_2$$

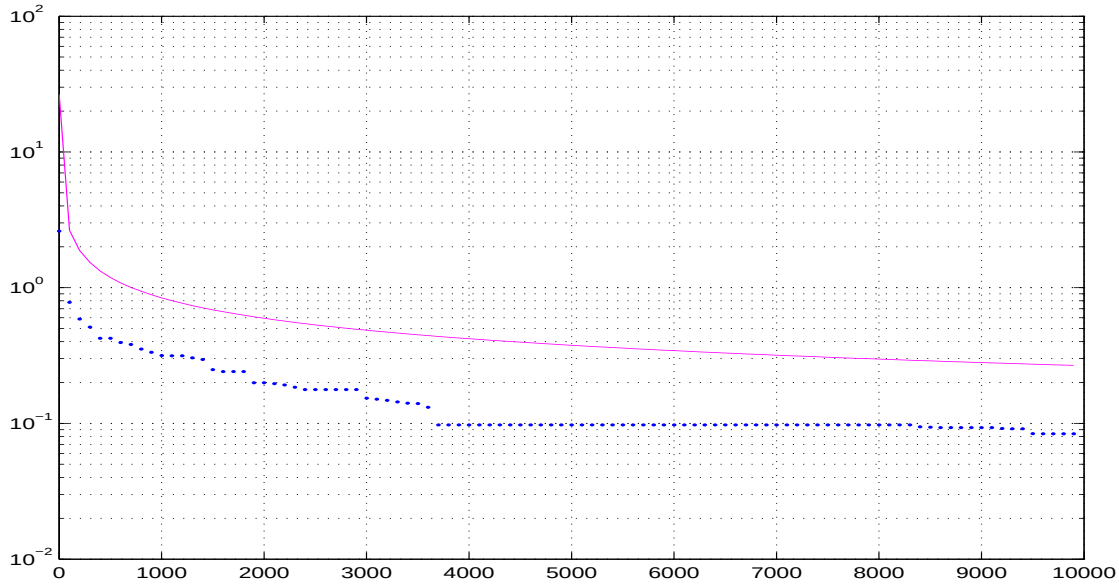
is fixed. Moreover, when X is a Euclidean ball in $\mathbf{R}^n$, this efficiency estimate "is as good as an efficiency estimate of a black-box-oriented method can be", provided that the dimension is large:

$$n \ge \left(\frac{\mathrm{Var}_{\|\cdot\|_2,X}(f)}{\epsilon}\right)^2$$

(see (5.2.4)). Note, however, that our "dimension independent" efficiency estimate

- is pretty slow

• is indeed dimension-independent only for problems with "Euclidean geometry" – those with moderate $\|\cdot\|_2$-variation. As a matter of fact, in applications problems of this type are pretty rare.



SD as applied to $\min_{\|x\|_2\le 1} \|Ax-b\|_1$, $A: 50\times 50$

[red: efficiency estimate; blue: actual error]

5.3.2 From SD to Bundle-Level: Adding memory

An evident drawback of SD is that all information on the objective accumulated so far is "summarized" in the current iterate, and this "summary" is very incomplete. With better usage of past information, one arrives at *bundle methods* which outperform SD significantly in practice, while preserving the most attractive theoretical property of SD – dimension-independent and optimal, in favourable circumstances, rate of convergence.

Bundle-Level method: Description

As applied to problem (5.3.1), BL works as follows.

1. At the beginning of step t of BL, we have in our disposal
   - the first-order information $\{f(x_\tau), f'(x_\tau)\}_{1\le\tau<t}$ on f along the previous search points $x_\tau \in X$, $\tau < t$;
   - current iterate $x_t \in X$.
2. At step t we
   (a) compute $f(x_t), f'(x_t)$; this information, along with the past first-order information on f, provides is with the current *model of the objective*
   $$f_t(x) = \max_{\tau\le t}[f(x_\tau) + (x-x_\tau)^T f'(x_\tau)]$$
   This model underestimates the objective and is exact at the points $x_1, ..., x_t$;
   (b) define the *best found so far value of the objective* $f^t = \min_{\tau\le t} f(x_\tau)$

(c) define the current *lower bound* f_t on f_* by solving the auxiliary problem

$$f_t = \min_{x\in X} f_t(x) \qquad (\mathrm{LP}_t)$$

Note that the current *gap* $\Delta_t = f^t - f_t$ is an upper bound on the inaccuracy of the best found so far approximate solution to the problem;

(d) compute the current *level* $\ell_t = f_t + \lambda\Delta_t$ ($\lambda \in (0,1)$ is a parameter)

(e) build a new search point by solving the auxiliary problem

$$x_{t+1} = \operatorname*{argmin}_x\{\|x - x_t\|_2^2 : x \in X, f_t(x) \le \ell_t\} \qquad (\mathrm{QP}_t)$$

and loop to step $t+1$.

Why and how BL converges?

Preliminary observations: A. The models $f_t(x) = \max_{\tau\le t}[f(x_\tau) + (x - x_\tau)^T f'(x_\tau)]$ grow with t and underestimate f, while the best found so far values of the objective decrease with t and overestimate f_*. Thus,

$$\begin{array}{c} f_1 \le f_2 \le f_3 \le ... \le f_* \\ f^1 \ge f^2 \ge f^3 \le ... \ge f_* \\ \Delta_1 \ge \Delta_2 \ge ... \ge 0 \end{array}$$

B. Let us say that a group of subsequent iterations $J = \{s, s+1, ..., r\}$ form a *segment*, if $\Delta_r \ge (1-\lambda)\Delta_s$. We claim that

(!) *If $J = \{s, s+1, ..., r\}$ is a segment, then*
(i) *All the sets $L_t = \{x \in X : f_t(x) \le \ell_t\}$, $t \in J$, have a point in common, specifically, (any) minimizer u of $f_r(\cdot)$ over X;*
(ii) *For $t \in J$, one has $\|x_t - x_{t+1}\|_2 \ge \frac{(1-\lambda)\Delta_r}{L_{\|\cdot\|_2(f)}}$.*

Indeed, (i): for $t \in J$ we have

$$\begin{array}{rcl} f_t(u) & \le & f_r(u) = f_r = f^r - \Delta_r \le f^t - \Delta_r \le f^t - (1-\lambda)\Delta_s \\ & \le & f^t - (1-\lambda)\Delta_t = \ell_t. \end{array}$$

(ii): We have $f_t(x_t) = f(x_t) \ge f^t$, and $f_t(x_{t+1}) \le \ell_t = f^t - (1-\lambda)\Delta_t$. Thus, when passing from x_t to x_{t+1}, t-th model decreases by at least $(1-\lambda)\Delta_t \ge (1-\lambda)\Delta_r$. It remains to note that $f_t(\cdot)$ is Lipschitz continuous w.r.t. $\|\cdot\|_2$ with constant $L_{\|\cdot\|_2}(f)$. □

C. **Main observation:**

(!!) *The cardinality of a segment $J = \{s, s+1, ..., r\}$ of iterations can be bounded as follows:*

$$\mathrm{Card}(J) \le \frac{\mathrm{Var}^2_{\|\cdot\|_2,X}(f)}{(1-\lambda)^2\Delta_r^2}. \qquad (5.3.14)$$

Indeed, when $t \in J$, the sets $L_t = \{x \in X : f_t(x) \le \ell_t\}$ have a point u in common, and x_{t+1} is the projection of x_t onto L_t. It follows that

$$\begin{array}{rcll} & \|x_{t+1} - u\|_2^2 \le \|x_t - u\|_2^2 - \|x_t - x_{t+1}\|_2^2 \ \forall t \in J & & \\ \Rightarrow & \sum_{t\in J} \|x_t - x_{t+1}\|_2^2 \le \|x_s - u\|_2^2 \le \max_{x,y\in X} \|x - y\|_2^2 & & \\ \Rightarrow & \mathrm{Card}(J) \le \dfrac{\max_{x,y\in X} \|x - y\|_2^2}{\min_{t\in J} \|x_t - x_{t+1}\|_2^2} & & \\ \Rightarrow & \mathrm{Card}(J) \le \dfrac{L^2_{\|\cdot\|_2}(f) \max_{x,y\in X} \|x - y\|_2^2}{(1-\lambda)^2\Delta_r^2} & & \text{[by (!.ii)]} \end{array}$$

We have arrived at the following

Theorem 5.3.1 *For every ϵ, $0 < \epsilon < \Delta_1$, the number N of steps of BL before a gap $\le \epsilon$ is obtained* (i.e., before an ϵ-solution is found) *does not exceed the bound*

$$N(\epsilon) = \frac{\mathrm{Var}^2_{\|\cdot\|_2,X}(f)}{\lambda(1-\lambda)^2(2-\lambda)\epsilon^2}. \tag{5.3.15}$$

Proof. Assume that N is such that $\Delta_N > \epsilon$, and let us bound N from above.
1^0. Let us split the set of iterations $I = \{1, ..., N\}$ into segments $J_1, ..., J_m$ as follows:

- J_1 is the maximal segment which ends with iteration N:

$$J_1 = \{t : t \le N, (1-\lambda)\Delta_t \le \Delta_N\}$$

- J_1 is certain group of subsequent iterations $\{s_1, s_1+1, ..., N\}$. If J_1 differs from I, that is, if $s_1 > 1$, we define J_2 as the maximal segment which ends with iteration $s_1 - 1$:

$$J_2 = \{t : t \le s_1 - 1, (1-\lambda)\Delta_t \le \Delta_{s_1-1}\} = \{s_2, s_2+1, ..., s_1-1\}$$

- If $J_1 \cup J_2$ differs from I, that is, if $s_2 > 1$, we define J_3 as the maximal segment which ends with iteration $s_2 - 1$:

$$J_3 = \{t : t \le s_2 - 1, (1-\lambda)\Delta_t \le \Delta_{s_2-1}\} = \{s_3, s_3+1, ..., s_2-1\}$$

and so on.

2^0. As a result of 1^0, I will be partitioned "from the end to the beginning" into segments of iterations J_1, J_2,...,J_m. Let d_ℓ be the gap corresponding to the last iteration from J_ℓ. By maximality of segments J_ℓ, we have

$$\begin{array}{rcl} d_1 & \ge & \Delta_N > \epsilon \\ d_{\ell+1} & > & (1-\lambda)^{-1} d_\ell,\ \ell = 1, 2, ..., m-1 \end{array}$$

whence

$$d_\ell > \epsilon(1-\lambda)^{-(\ell-1)}.$$

We now have

$$\begin{array}{rcl} N & = & \sum\limits_{\ell=1}^{m} \mathrm{Card}(J_\ell) \le \sum\limits_{\ell=1}^{m} \frac{\mathrm{Var}^2_{\|\cdot\|_2,X}(f)}{(1-\lambda)^2 d_\ell^2} \le \frac{\mathrm{Var}^2_{\|\cdot\|_2,X}(f)}{(1-\lambda)^2} \sum\limits_{\ell=1}^{m} (1-\lambda)^{2(\ell-1)}\epsilon^{-2} \\ & \le & \frac{\mathrm{Var}^2_{\|\cdot\|_2,X}(f)}{(1-\lambda)^2\epsilon^2} \sum\limits_{\ell=1}^{\infty} (1-\lambda)^{2(\ell-1)} = \frac{\mathrm{Var}^2_{\|\cdot\|_2,X}(f)}{(1-\lambda)^2[1-(1-\lambda)^2]\epsilon^2} = N(\epsilon). \end{array}$$

■

Comments. We have seen that the Bundle-Level method shares the dimension-independent (and optimal in a "favourable geometry" large-scale case) theoretical complexity bound:

> **Provable fact:** *For every $\epsilon > 0$, the number of steps before an ϵ-solution to convex program $\min\limits_{x\in X} f(x)$ is found, does not exceed $O(1)\left(\frac{\mathrm{Var}_{\|\cdot\|_2,X}(f)}{\epsilon}\right)^2$.*

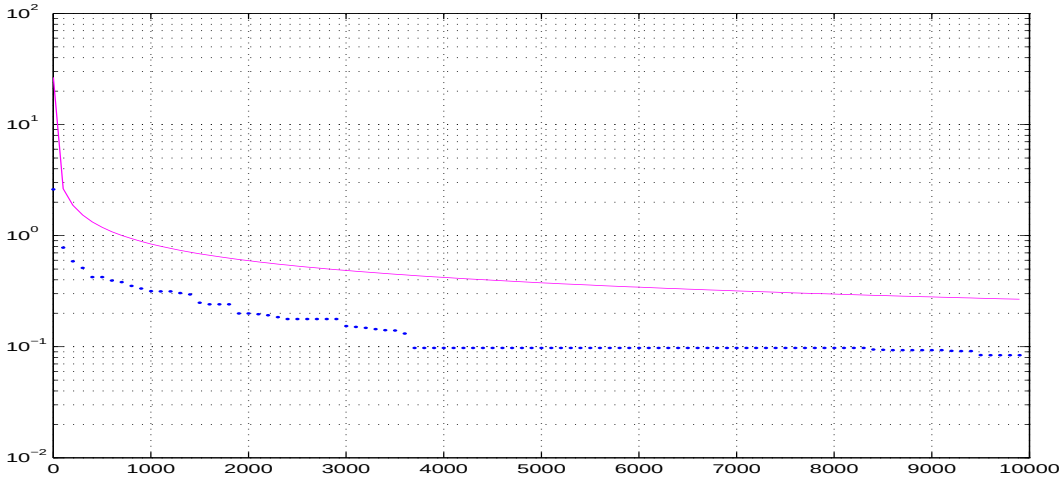
At the same time, there exists quite convincing experimental evidence that the Bundle-Level method obeys the optimal in fixed dimension "polynomial time" complexity bound:

> **Experimental fact:** *For every $\epsilon \in (0, \mathrm{Var}_X(f) \equiv \max\limits_X f - \min\limits_X f)$, the number of steps of BL before an ϵ-solution to convex program $\min\limits_{x\in X} f(x)$ with $X \subset \mathbf{R}^n$ is found, does not exceed $n \ln\left(\frac{\mathrm{Var}_X(f)}{\epsilon}\right) + 1$*

or, equivalently,

> *When solving convex program with n variables by BL, every n steps add new accuracy digit.*

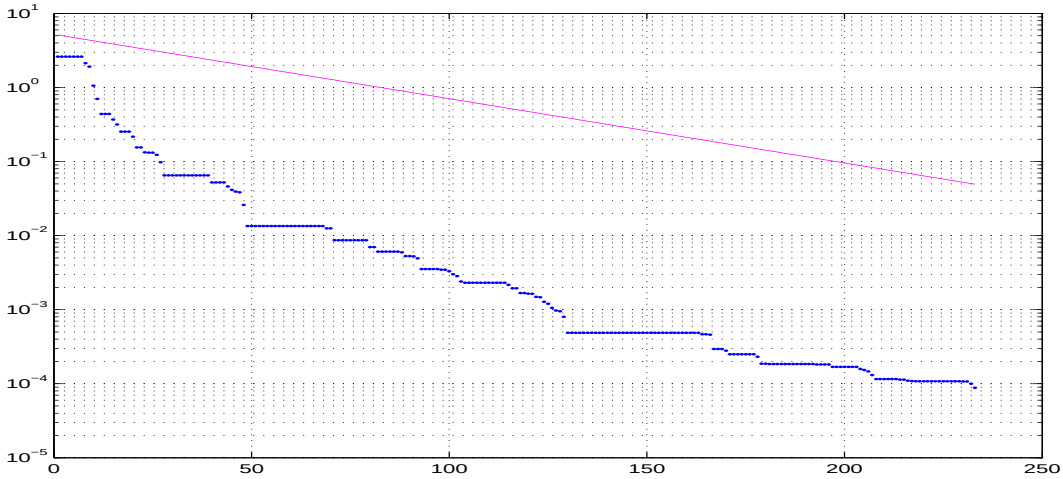
Illustration: Consider a randomly generated problem $\min\limits_{x:\|x\|_2\le 1} f(x) \equiv \|Ax-b\|_1$, $\dim x = 50$ ($f(0) = 2.61$, $f_* = 0$). Here is what happens with SD and BL:



SD, accuracy vs. iteration count

(blue: errors; red: efficiency estimate $3\frac{\mathrm{Var}_{\|\cdot\|_2,X}(f)}{\sqrt{t}}$)

$\epsilon_{10000} = 0.084$



BL, accuracy vs. iteration count

(blue: errors; red: efficiency estimate $e^{-\frac{t}{n}}\mathrm{Var}_X(f)$)

$\epsilon_{233} < 1.e-4$

Bundle-Level vs. Subgradient Descent

5.3.3 Restricted Memory Bundle-Level

In BL, the number of linear constraints in the auxiliary problems

$$f_t = \min_{x\in X} f_t(x) \qquad (\mathrm{LP}_t)$$
$$x_{t+1} = \operatorname*{argmin}_x \left\{\|x_t - x\|_2^2 : x\in X, f_t(x)\le \ell_t\right\} \qquad (\mathrm{QP}_t)$$

is equal to the size t of the current *bundle* – the collection of affine forms $g_\tau(x) = f(x_\tau) + (x-x_\tau)^T f'(x_\tau)$ participating in the model $f_t(\cdot)$. Thus, the complexity of an iteration in BL grows with the iteration

number. In order to suppress this phenomenon, one needs a mechanism for *shrinking* the bundle (and thus – simplifying the models of f).

Simple bundle-shrinking policy

The simplest way of shrinking the bundle is to initialize d as Δ_1 and to run plain BL until an iteration t with $\Delta_t \le d/2$ is met. At such an iteration, we

• shrink the current bundle, keeping in it the minimum number of the forms g_τ sufficient to ensure that

$$f_t \equiv \min_{x\in X} \max 1 \le \tau \le t g_\tau(x) = \min_{x\in X} \max_{\text{selected } \tau} g_\tau(x)$$

(this number is at most n), and

• reset d as Δ_t,

and proceed with plain BL until the gap is again reduced by factor 2, etc.

Computational experience demonstrates that the outlined approach does not slow BL down, while keeping the size of the bundle below the level of about $2n$.

Truncated Proximal Bundle-Level

In *Truncated Proximal Bundle-Level* method, the size of bundle is kept below *a given* desired level m (which, independently of the dimension of the problem of interest, can be as small as 1 or 2).

Phases of TPBL. Execution of TPBL is split into *phases*. Phase s is associated with

1. *prox-center* $c_s \in X$,
2. s-th *upper bound* f^s on f_*, which is the best value of the objective observed before the phase begins,
3. • s-th *lower bound* f_s on f_*, which is the best lower bound on f_* observed before the phase begins.
   Bounds f^s and f_s define
   - s-th *optimality gap* $\Delta_s = f^s - f_s$;
   - s-th *level* $\ell_s = f_s + \lambda\Delta_s$, where $\lambda \in (0,1)$ is parameter of the method.
4. current *model* $\widetilde{f}^s(\cdot) \le f(\cdot)$ of $f(\cdot)$, which is the maximum of $\le m$ affine forms.

To initialize the first phase, we choose $c_1 \in X$, compute $f(c_1), f'(c_1)$ and set

$$\widetilde{f}^1(x) = f(c_1) + (x-c_1)^T f'(c_1), \quad f^1 = f(c_1), \quad f_1 = \min_{x\in X} \widetilde{f}^1(x).$$

Steps of a phase. At the beginning of step $t = 1, 2, ...$ of phase s, we have in our disposal

- upper bound $f^{s,t-1} \le f^s$ on f_*, which is the best found so far value of the objective,
- lower bound $f_{s,t-1} \ge f_s$ on f_*,
- model $\widetilde{f}^{s,t-1}(\cdot) \le f(\cdot)$ of the objective which is the maximum of $\le m$ affine forms,
- • iterate $x_t \in X$ and set $H_{t-1} = \{x : \alpha_{t-1}^T x \ge \beta_{t-1}\}$ such that

$$\begin{array}{rl} x \in X, f(x) \le \ell_s \Rightarrow x \in H_{t-1} & (a_t) \\ x_t = \operatorname{argmin}_x \left\{\|x - c_s\|_2^2 : x \in H_{t-1}\right\} & (b_t) \end{array} \tag{5.3.16}$$

To initialize the first step of phase s, we set

$$f^{s,0} = f^s,\ f_{s,0} = f_s,\ \widetilde{f}^{s,0}(\cdot) = \widetilde{f}^s(\cdot),\ \alpha_0 = 0, \beta_0 = 0\ [\Rightarrow H_0 = \mathbf{R}^n]$$

thus ensuring (5.3.16.a_1), and set $x_1 = c_s$, thus ensuring (5.3.16.b_1).

Step t of phase s is as follows. Given

- bounds $f^{s,t-1} \geq f_*$, $f_{s,t-1} \leq f_*$,
- model $\widetilde{f}^{s,t-1}(\cdot) \leq f(\cdot)$,
- x_t and $H_{t-1} = \{x : \alpha_{t-1}^T x \geq \beta_{t-1}\}$ such that

$$\begin{array}{cc} x \in X, f(x) \leq \ell_s \Rightarrow x \in H_{t-1} & (a_t) \\ x_t = \operatorname*{argmin}_x \left\{\|x - c_s\|_2^2 : x \in H_{t-1}\right\} & (b_t) \end{array}$$

we

1. compute $f(x_t), f'(x_t)$ and set $g_t(x) = f(x_t) + (x - x_t)^T f'(x_t)$;
2. define $\widetilde{f}^{s,t}(\cdot)$ as the maximum of $g_t(\cdot)$ and affine forms associated with $\widetilde{f}^{s,t-1}$ (dropping, if necessary, one of the latter forms to make $\widetilde{f}^{t,s}$ the maximum of at most m forms). If $f(x_t) \leq \ell_s + 0.5(f^s - \ell_s)$ ("significant progress in the upper bound"), we terminate phase s and set

$$f^{s+1} = f^{s,t}, \quad f_{s+1} = f_{s,t-1}, \quad \widetilde{f}^{s+1}(\cdot) = \widetilde{f}^{s,t}(\cdot),$$

   otherwise we proceed as follows:

3. Compute $f_t = \min\limits_x \left\{\widetilde{f}^{s,t}(x) : x \in H_{t-1} \cap X\right\}$. Since $f(x) \geq \ell_s$ in $X \backslash H_{t-1}$, we have $f_* \geq \min[\ell_s, f_t]$, so that

$$f_{s,t} \equiv \max\{f_{s,t-1}, \min[\ell_s, f_t]\} \leq f_*.$$

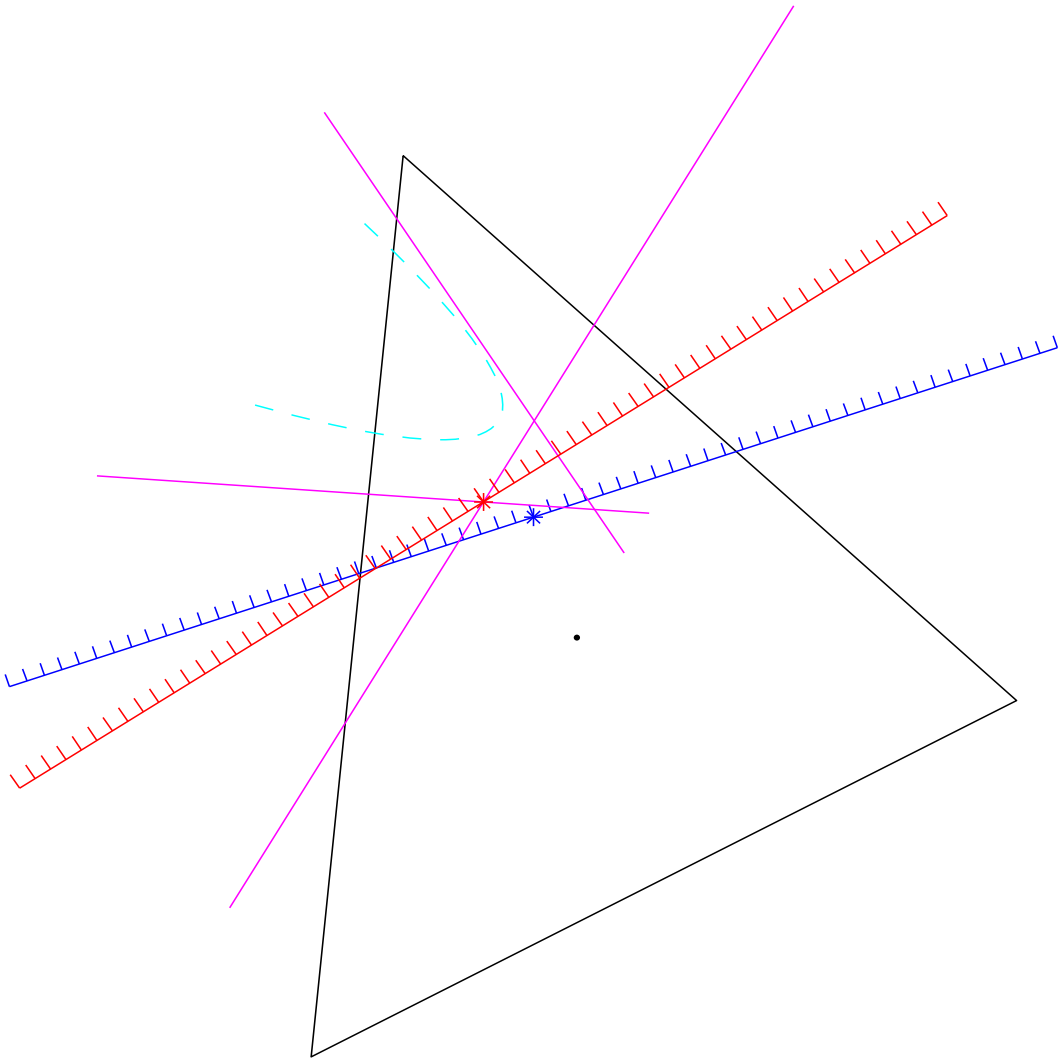
   If $f_{s,t} \geq \ell_s - 0.5(\ell_s - f_s)$ ("significant progress in the lower bound"), we terminate phase s and set

$$f^{s+1} = f^{s,t}, \quad f_{s+1} = f_{s,t}, \quad \widetilde{f}^{s+1}(\cdot) = \widetilde{f}^{s,t}(\cdot)$$

   otherwise we set

$$\begin{array}{rcl} x_{t+1} & = & \operatorname*{argmin}_x \left\{\|x - c_s\|_2^2 : x \in X \cap H_{t-1}, \widetilde{f}^{s,t}(x) \leq \ell_s\right\} \\ H_t & = & \{x : (x_{t+1} - c_s)^T (x - x_{t+1}) \geq 0\} \end{array}$$

   and loop to step $t+1$ of phase s.



Step of TPBL
Black: X
Blue: dot – prox-center; star – x_t; half-plane – H_{t-1}
Cyan: true level set $f(\cdot) \le \ell_s$
Magenta: model level set $\widetilde{f}^{s,t}(\cdot) \le \ell_s$
Red: star: x_{t+1}, half-space – H_t.

Note: If phase s is not terminated at step t, then, by construction,

$$\begin{array}{rcll} x_{t+1} & = & \operatorname{argmin}_x \left\{ \|x - c_s\|_2^2 : x \in X \cap H_{t-1}, \widetilde{f}^{s,t}(x) \le \ell_s \right\} & (a) \\ H_t & = & \{x : (x_{t+1} - c_s)^T (x - x_{t+1}) \ge 0\} & (b) \end{array} \tag{5.3.17}$$

It follows that *when passing to step $t+1$, we have ensured the relations*

$$\begin{array}{rl} x \in X, f(x) \le \ell_s \Rightarrow x \in H_t & (a_{t+1}) \\ x_{t+1} = \operatorname{argmin}_x \left\{ \|x - c_s\|_2^2 : x \in X \cap H_t, \widetilde{f}^{s,t}(x) \le \ell \right\} & (b_{t+1}) \end{array}$$

Indeed, x_{t+1} is the minimizer of $\omega_s(x) \equiv \frac{1}{2}\|x - c_s\|_2^2$ on the set

$$Y_t = X \cap H_{t-1} \cap \{x : \widetilde{f}^{t,s}(x) \le \ell_s\}$$

whence

$$\overbrace{[\omega_s'(x_{t+1})]}^{x_{t+1}-c_s}{}^T(x-x_{t+1})\geq 0\ \forall x\in Y_t$$
$$\Downarrow$$
$$Y_t\subset H_t=\{x:[\omega_s'(x_{t+1})]^T(x-x_{t+1})\geq 0\}\quad(*)$$

Thus,

$$(x\in X, f(x)\leq \ell_s)\underbrace{\Rightarrow}_{(a_t)}(x\in X\cap H_{t-1}, f(x)\leq \ell_s)$$
$$\Rightarrow (x\in X\cap H_{t-1}, \widetilde{f}^{s,t}(x)\leq \ell_s)\underbrace{\Rightarrow}_{(*)} x\in H_t$$

as required in (a_{t+1}). (b_{t+1}) readily follows from the definition of H_t.

Convergence Analysis of TPBL

Preliminary observations: A. *When passing from phase s to phase $s+1$, the optimality gap is decreased at least by the factor*

$$\theta(\lambda)=\frac{\min[1+\lambda,2-\lambda]}{2}.$$

Indeed, phase s can be terminated at step t due to significant progress either in the upper bound on f_*: $f^{s+1}=f^{s,t}\leq \ell_s+\frac{1}{2}(f^s-\ell_s)$

$$\Rightarrow \Delta_{s+1}=f^{s+1}-f_{s+1}\leq \frac{1}{2}\ell_s+\frac{1}{2}f^s-f_s=\frac{1+\lambda}{2}\Delta_s$$

or in the lower bound: $f_{s+1}=f_{s,t}\geq \ell_s-0.5(\ell_s-f_s)$

$$\Rightarrow \Delta_{s+1}=f^{s+1}-f_{s+1}\leq f^s-\frac{1}{2}f_s-\frac{1}{2}\ell_s=\frac{2-\lambda}{2}\Delta_s$$

B. *Let x_t, x_{t+1} be two subsequent search points of phase s. Then*

$$\|x_t-x_{t+1}\|_2>\frac{(1-\lambda)\Delta_s}{2L_{\|\cdot\|_2}(f)}.$$

Indeed, we have $f(x_t)=g_t(x_t)=\widetilde{f}^{s,t}(x_t)\geq \ell_s+\frac{1}{2}(f^s-\ell_s)$, since otherwise phase s would be terminated at step t. At the same time, $g_t(x_{t+1}))\leq m\widetilde{f}^{s,t}(x_{t+1})\leq \ell_s$. Thus, passing from x_t to x_{t+1}, we decrease Lipschitz continuous, with constant $L_{\|\cdot\|_2}(f)$ w.r.t. $\|\cdot\|_2$, function $g_t(\cdot)$ by at least $\frac{1}{2}(f^s-\ell_s)=\frac{1-\lambda}{2}\Delta_s$.

Main observation: *The number of steps at phase s does not exceed*

$$N_s=\frac{4V^2_{\|\cdot\|_2,X}(f)}{(1-\lambda)^2\Delta_s^2}+1. \tag{5.3.18}$$

Indeed, let the number of steps of the phase be $> N$. By construction, $x_{t+1}\in H_{t-1}$ and x_t is the minimizer of $\omega_s(x)=\frac{1}{2}\|x-c_s\|_2^2$ on H_{t-1}, whence

$$\begin{aligned}1\leq t\leq N\Rightarrow \omega_s(x_{t+1}) &= \omega_s(x_t)+\underbrace{(x_{t+1}-x_t)^T\omega_s'(x_t)}_{\geq 0}+\tfrac{1}{2}\|x_t-x_{t+1}\|_2^2\\ &\geq \omega_s(x_t)+\tfrac{1}{2}\|x_t-x_{t+1}\|_2^2.\end{aligned}$$

It follows that

$$\sum_{t=1}^N\underbrace{\frac{1}{2}\|x_t-x_{t+1}\|_2^2}_{\geq\frac{(1-\lambda)^2\Delta_s^2}{8L^2_{\|\cdot\|_2}(f)}}\leq\frac{1}{2}\max_{x,y\in X}\|y-x\|_2^2,$$

whence

$$N \leq \frac{4V^2_{\|\cdot\|_2,X}(f)}{(1-\lambda)^2\Delta_s^2}.$$ □

Same as in the case of BL, (5.3.18) combines with the relation $\Delta_{s+1} \leq \theta(\lambda)\Delta_s$ to yield the following

Theorem 5.3.2 *For every ϵ, $0 < \epsilon < \Delta_1$, the total number of TPBL steps before a gap $\leq \epsilon$ is obtained* (i.e., before an ϵ-solution is found) *does not exceed the bound*

$$N(\epsilon) = c(\lambda)\frac{\mathrm{Var}^2_{\|\cdot\|_2,X}(f)}{\epsilon^2}.$$

Theorem says that when passing from BL to TPBL, we, essentially, preserve the efficiency estimate, while allowing for full control on bundle cardinality, and, consequently, on the complexity of the auxiliary problems.

5.4 The Bundle-Mirror scheme

Subgradient Descent method and its bundle versions are "intrinsically adjusted" to problems with Euclidean geometry; this is where the role of the $\|\cdot\|_2$-variation of the objective

$$\mathrm{Var}_{\|\cdot\|_2,X}(f) = L_{\|\cdot\|_2}(f)\max_{x,x'\in X}\|x-x'\|_2$$

in the efficiency estimate

$$\min_{t\leq T} f(x_t) - f_* \leq O(1)\frac{\mathrm{Var}_{\|\cdot\|_2,X}(f)}{\sqrt{T}}$$

comes from. The *Mirror Descent* scheme extends SD and its bundle versions onto problems with "nice *non-Euclidean* geometry".

5.4.1 Mirror Descent – Building Blocks

Building block #1: Distance-Generating Function. A SD step

$$x \mapsto x_+ = \Pi_X(x - \gamma f'(x))$$

can be viewed as follows: given an iterate $x \in X$, we

1) Form a "local distance" term

$$\left[\frac{1}{2}\|y-x\|_2^2 \equiv\right] \qquad \omega_x(y) = \omega(y) - \omega(x) - \langle\nabla\omega(x), y-x\rangle$$

where

$$\omega(u) = \frac{1}{2}\|u\|_2^2 \tag{2}$$

is a specific "distance-generating function";

2) Augment the linear model of f, built at x:

$$f_x(y) = f(x) + \langle f'(x), y-x\rangle$$

by $\frac{1}{\gamma}$ times the distance term, thus getting the "augmented model"

$$g_x^\gamma(y) = f_x(y) + \frac{1}{\gamma}\omega_x(y) = \frac{1}{\gamma}\left(\langle\gamma f'(x) - \nabla\omega(x), y\rangle + \omega(y)\right) + c(x,\gamma)$$

3) Minimize the augmented model over $y \in X$, thus getting the new iterate x_+:

$$\begin{array}{rcl}\operatorname*{argmin}_{y\in X} g_x^\gamma(y) & = & \operatorname*{argmin}_{y\in X}\left(\langle \gamma f'(x) - \nabla\omega(x), y\rangle + \omega(y)\right)\\ & = & \operatorname*{argmin}_{y\in X}\left(-\langle [x-\gamma f'(x)], y\rangle + \frac{1}{2}\langle y, y\rangle\right)\\ & = & \operatorname*{argmin}_{y\in X}\left[\frac{1}{2}\|y-[x-\gamma f'(x)]\|_2^2 - \frac{1}{2}\|x-\gamma f'(x)\|_2^2\right]\\ & = & \Pi_X(x-\gamma f'(x)).\end{array}$$

Thus,

Subgradient Descent step

$$x \mapsto x_+ = \Pi_X(x-\gamma f'(x))$$

is the step

$$x \mapsto x_+ = \operatorname*{argmin}_{y\in X}\left[\langle \gamma f'(x) - \nabla\omega(x), y\rangle + \omega(y)\right] \tag{5.4.1}$$

associated with the specific distance-generating function

$$\omega(u) = \frac{1}{2}u^T u \tag{5.4.2}$$

Note that the distance-generating function (5.4.2) is continuously differentiable and strongly convex on X, the latter meaning that

$$\langle \nabla\omega(u) - \nabla\omega(v), u-v\rangle \geq \alpha\|u-v\|_2^2 \ \forall u,v\in X \qquad [\alpha>0] \tag{5.4.3}$$

(from ow on, $\langle\cdot,\cdot\rangle$ is the standard inner product in $\mathbf{R}^n \supset X$). Indeed, in the case of (5.4.2) we have

$$\langle \nabla\omega(u) - \nabla\omega(v), u-v\rangle = \langle u-v, u-v\rangle = \|u-v\|_2^2 .eqno\square$$

Building block #2: the potential. Convergence analysis of SD was based on the inequality

$$\begin{array}{l}\forall u\in X:\\ \gamma\underbrace{\langle f'(x), x-u\rangle}_{\geq f(x)-f(u)} - \frac{1}{2}\|\gamma f'(x)\|_2^2 \leq \underbrace{\frac{1}{2}\|x-u\|_2^2 - \frac{1}{2}\|x_+-u\|_2^2}\\ \qquad = [\frac{1}{2}x^Tx - x^Tu] - [\frac{1}{2}x_+^Tx_+ - x_+^Tu]\\ \qquad = [\langle\nabla\omega(x), x-u\rangle - \omega(x)] - [\langle\nabla\omega(x_+), x_+-u\rangle - \omega(x_+)]\\ \qquad \left[\omega(u) = \frac{1}{2}u^Tu\right]\end{array} \tag{5.4.4}$$

ensured by SD step. This inequality states that $H_u(x) = \langle\nabla\omega(x), x-u\rangle - \omega(x)$ is a kind of "potential" for SD: when u is such that $f(x) > f(u)$, a SD step reduces this potential at least by $\gamma[f(x)-f(u)] - O(\gamma^2)$.

Now let us make the following

Observation: *When $\omega(\cdot)$ is continuously differentiable and strongly convex on X:*

$$\langle \nabla\omega(u) - \nabla\omega(v), u-v\rangle \geq \alpha\|u-v\|^2 \ \forall u,v\in X \qquad [\alpha>0]$$

step (5.4.1) *ensures inequality similar to* (5.4.4):

$$\begin{array}{c}\gamma\langle f'(x), x-u\rangle \leq H_u(x) - H_u(x_+) + \frac{1}{2\alpha}\gamma^2\|f'(x)\|_*^2\\ \left[\|\xi\|_* = \max_u\{\langle\xi,u\rangle : \|u\|\leq 1\}\right]\end{array} \tag{5.4.5}$$

Indeed, optimality condition for (5.4.1) reads

$$\langle \gamma f'(x) - \nabla\omega(x) + \nabla\omega(x_+), u-x_+\rangle \geq 0 \ \forall u\in X,$$

whence

$$\begin{array}{rcl}
\gamma\langle f'(x), u - x_+\rangle \geq & \underbrace{\langle \nabla\omega(x) - \nabla\omega(x_+), u - x_+\rangle}_{H_u(x_+) - H_u(x) - [\omega(x_+) - \omega(x) - \langle \nabla\omega(x), x_+ - x\rangle]} & \forall u \in X \\
& \Downarrow & \\
\gamma\langle f'(x), x_+ - u\rangle & \leq & H_u(x) - H_u(x_+) \\
& & + [\omega(x) - \langle \nabla\omega(x), x - x_+\rangle - \omega(x_+)] \\
& & \Downarrow \\
\gamma\langle f'(x), x - u\rangle & \leq & H_u(x) - H_u(x_+) \\
& & + \underbrace{[\omega(x) - \langle \nabla\omega(x), x - x_+\rangle - \omega(x_+)]}_{\leq -\frac{\alpha}{2}\|x - x_+\|^2} \\
& & + \underbrace{\gamma\langle f'(x), x - x_+\rangle}_{\leq \gamma\|f'(x)\|_*\|x - x_+\|} \\
& & \Downarrow \\
\forall u \in X: \ \gamma\langle f'(x), x - u\rangle & \leq & H_u(x) - H_u(x_+) \\
& & + \max\limits_r \left\{\gamma\|f'(x)\|_* r - \frac{\alpha}{2} r^2\right\} \\
& = & H_u(x) - H_u(x_+) + \frac{1}{2\alpha}\gamma^2\|f'(x)\|_*^2
\end{array}$$

as stated in (5.4.5).
Note: With $\omega(u) = \frac{1}{2}\|u\|_2^2$, $\|\cdot\| = \|\cdot\|_2$ one has $\alpha = 1$, $\|\cdot\|_* \equiv \|\cdot\|_2$, and (5.4.5) becomes (5.4.4).

5.4.2 Non-Euclidean SD – Mirror Descent

Same as before, we focus on convex problem (5.3.7) with convex compact domain $X \subset E = \mathbf{R}^n$ and Lipschitz continuous on this domain convex objective f.

The Setup for MD as applied to (5.3.7) is given by

1. continuously differentiable strongly convex function $\omega(u)$ on X
2. a norm $\|\cdot\|$ on E.

$\omega(\cdot)$ and $\|\cdot\|$ define two important parameters:

- *modulus of strong convexity* of ω w.r.t $\|\cdot\|$:
$$\alpha = \max\left\{\alpha' > 0 : \langle \nabla\omega(u) - \nabla\omega(v), u - v\rangle \geq \alpha'\|u - v\|^2 \, \forall u, v \in X\right\} \tag{5.4.6}$$
- *ω-size* of X
$$\Theta = \max_{u,v \in X} [\omega(u) - \omega(v) - \langle \nabla\omega(v), u - v\rangle].$$

Note: With "Ball setup" $\omega(u) = \frac{1}{2}\langle u, u\rangle$, $\|u\| \equiv \|u\|_2 = \sqrt{\langle u, u\rangle}$ one has $\alpha = 1$, $\Theta = \frac{1}{2}\max\limits_{u,v \in X}\|u - v\|_2^2$.

MD: the construction

As applied to (5.3.7), MD generates search points x_t according to the recurrence

$$x_{t+1} = \operatorname*{argmin}_{y \in X} [\langle \gamma_t f'(x_t) - \nabla\omega(x_t), y\rangle + \omega(y)] \tag{5.4.7}$$

where $\gamma_t > 0$ are stepsizes.
Note:

- With Ball setup, (MD) becomes exactly the SD recurrence $x_{t+1} = \Pi_X(x_t - \gamma_t f'(x_t))$
- In order for (MD) to be practical, a step should be easy to implement. This means that X and $\omega(\cdot)$ should fit each other in the sense that auxiliary problems
$$\min_{y \in X} [\langle \zeta, y\rangle + \omega(y)]$$
should be easy to solve.

Why and how MD converges?

By (5.4.5), a MD step ensures the inequality

$$\forall u \in \gamma_t\langle f'(x_t), x_t - u\rangle \le H_u(x_t) - H_u(x_{t+1}) + \tfrac{1}{2\alpha}\gamma_t^2\|f'(x_t)\|_*^2$$
$$[H_u(x) = \langle\nabla\omega(x), x - u\rangle - \omega(x)]$$

Summing up these inequalities, we conclude that for positive integers $T_0 \le T$ one has

$$\sum_{t=T_0}^{T}\gamma_t\langle f'(x_t), x_t - u\rangle \le H_u(x_{T_0}) - H_u(x_T) + \frac{1}{2\alpha}\sum_{t=T_0}^{T}\gamma_t^2\|f'(x_t)\|_*^2. \tag{5.4.8}$$

Lemma 5.4.1 *One has $H_u(x) - H_u(y) \le \Theta$ for all $x, y, u \in X$.*

Proof. Indeed,

$$\begin{aligned} H_u(x) - H_u(y) &= \langle\nabla\omega(x), x - u\rangle - \omega(x) - \langle\nabla\omega(y), y - u\rangle + \omega(y) \\ &= \underbrace{[\omega(u) - \omega(x) - \langle\nabla\omega(x), u - x\rangle]}_{\le\Theta} \\ &\quad + \underbrace{[\omega(y) + \langle\nabla\omega(y), u - y\rangle - \omega(u)]}_{\le 0} \end{aligned}$$

■

Applying Lemma, we conclude from (5.4.8) that

$$\sum_{t=T_0}^{T}\gamma_t\langle f'(x_t), x_t - u\rangle \le \Theta + \frac{1}{2\alpha}\sum_{t=T_0}^{T}\gamma_t^2\|f'(x_t)\|_*^2. \tag{5.4.9}$$

For MD, relation (5.4.9) plays the same crucial role that the inequality (5.3.8) played for SD. Same as in the latter case, it implies that for all positive integers $T \ge T_0$ it holds

$$\epsilon_T \equiv \min_{t\le T} f(x_t) - f_* \le \frac{\Theta + \frac{1}{2\alpha}\sum\limits_{t=T_0}^{T}\gamma_t^2\|f'(x_t)\|_*^2}{\sum\limits_{t=T_0}^{T}\gamma_t} \tag{5.4.10}$$

and we arrive at the following two results (cf. Propositions 5.3.2, 5.3.3):

Proposition 5.4.1 ["Divergent series"] *Whenever $0 < \gamma_t \to 0$ as $t \to \infty$ in such a way that $\sum\limits_t \gamma_t = \infty$, one has $\epsilon_T \to 0$ as $T \to \infty$.*

Proposition 5.4.2 [Optimal stepsize policy] *With stepsizes*

$$\gamma_t = \frac{\sqrt{\Theta\alpha}}{\|f'(x_t)\|_*\sqrt{t}}, \tag{5.4.11}$$

one has

$$\epsilon_T \equiv \min_{t\le T} f(x_t) - f_* \le O(1)\frac{\sqrt{\Theta}L_{\|\cdot\|}(f)}{\sqrt{\alpha}\sqrt{T}} \tag{5.4.12}$$

where $L_{\|\cdot\|}(f)$ is the Lipschitz constant of f w.r.t. the norm $\|\cdot\|$.

Standard Setups and associated efficiency estimates

To get SD as a particular case of MD, one uses
Ball Setup: $\omega(u) = \frac{1}{2}\|u\|_2^2$, $\|\cdot\| = \|\cdot\|_2$. For this setup, one has

$$\alpha = 1,\ \Theta = \frac{1}{2}\max_{x,y\in X}\|x-y\|_2^2,$$

and the associated efficiency estimate (5.4.12) becomes

$$\epsilon_T \le O(1)\frac{L_{\|\cdot\|_2}(f)\max\limits_{x,y\in X}\|x-y\|_2}{\sqrt{T}} \tag{5.4.13}$$

(cf. (5.3.13)).

There are at least two important setups more:
Simplex setup: X is a closed convex subset of the simplex $\Delta_n^+(R) = \{x \in E = \mathbf{R}^n : x \ge 0, \sum\limits_i x_i \le R\}$, $\|\cdot\| = \|\cdot\|_1$,

$$\omega(x) = \sum_i (R^{-1}x_i + n^{-1}\delta)\ln(R^{-1}x_i + n^{-1}\delta) \qquad [\delta = 1.\mathrm{e}-16].$$

Spectahedron Setup: X is a closed convex subset of the *spectahedron* $\Xi_n(R) = \{x \in E = \mathbf{S}^n : x \succeq 0, \mathrm{Tr}(x) \le R\}$, $\|x\| = |x|_1 \equiv \|\lambda(x)\|_1$,

$$\omega(x) = \sum_i (R^{-1}\lambda_i(x) + n^{-1}\delta)\ln(R^{-1}\lambda_i(x) + n^{-1}\delta).$$

For these setups, one has (see Appendix to Lecture 5)

$$\alpha = O(1)R^{-2},\ \Theta \le O(1)\ln n. \tag{5.4.14}$$

and the associated efficiency estimate (5.4.12) becomes nearly-dimension-independent bound

$$\epsilon_T \equiv \min_{t\le T} f(x_t) - f_* \le O(1)\frac{\sqrt{\ln n}L_{\|\cdot\|}(f)R}{\sqrt{T}}. \tag{5.4.15}$$

Discussion: MD vs. SD. Let us compare the convergence properties of MD with Simplex setup and SD (i.e., MD with Ball setup).

The efficiency estimate for the MD with Simplex Setup reads

$$\epsilon_T[\ \begin{matrix}\text{Simplex}\\ \text{setup}\end{matrix}\ ] = \min_{t\le T} f(x_t) - f_* \le O(1)\frac{\ln^{1/2}(n)\overbrace{\max\limits_{x,y\in X}\|x-y\|_1 L_{\|\cdot\|_1}(f)}^{\mathrm{Var}_{\|\cdot\|_1,X}(f)}}{\sqrt{T}} \tag{S}$$

while for SD the efficiency estimate is

$$\epsilon_T[\ \begin{matrix}\text{Ball}\\ \text{setup}\end{matrix}\ ] = \min_{t\le T} f(x_t) - f_* \le O(1)\frac{\overbrace{\max\limits_{x,y\in X}\|x-y\|_2 L_{\|\cdot\|_2}(f)}^{\mathrm{Var}_{\|\cdot\|_2,X}(f)}}{\sqrt{T}} \tag{B}$$

The ratio of the estimates is

$$\chi = \frac{\epsilon_T[\ \begin{matrix}\text{Simplex}\\ \text{setup}\end{matrix}\ ]}{\epsilon_T[\ \begin{matrix}\text{Ball}\\ \text{setup}\end{matrix}\ ]} = O(\sqrt{\ln n})\cdot \underbrace{\frac{\max\limits_{x,y\in X}\|x-y\|_1}{\max\limits_{x,y\in X}\|x-y\|_2}}_{A}\cdot \underbrace{\frac{L_{\|\cdot\|_1}(f)}{L_{\|\cdot\|_2}(f)}}_{B}$$

Observe that

- the factor $O(\sqrt{\ln n})$ is "against" Simplex setup; however, in practice this factor is just a moderate absolute constant.
- the ratio $\frac{\|u\|_1}{\|u\|_2}$ is always ≥ 1 and, depending on x, can be as large as $\sqrt{n}$. It follows that
  - factor A is always ≥ 1 (i.e., is "against" Simplex setup) and can be as large as $\sqrt{n}$
  - factor B is always ≤ 1 (i.e., is "in favour" of Simplex setup) and can be as small as $\frac{1}{\sqrt{n}}$. The actual value of B is
$$\frac{L_{\|\cdot\|_1}(f)}{L_{\|\cdot\|_2}(f)} = \frac{\max\limits_{x\in X}\|f'(x)\|_\infty}{\max\limits_{x\in X}\|f'(x)\|_2}$$
    and it depends on the "geometry" of f. For example,
    - In the case when all first order partial derivatives of f in X are of the same order ("f is nearly equally sensitive to all variables"), we have
$$B = O\left(\frac{\|(a,...,a)^T\|_\infty}{\|(a,...,a)^T\|_2}\right) = O(n^{-1/2})$$
    - In the case when just $O(1)$ first order derivatives of f'in X are of the same order, and the remaining derivatives are negligible small ("f is sensitive to just $O(1)$ variables"), we have
$$B = O\left(\frac{\|(a,0,...,0)^T\|_\infty}{\|(a,0,...,0)^T\|_2}\right) = O(1)$$

It follows that the resulting performance ratio χ depends on the geometry of X and f.
• *Extreme example I: X is the standard Euclidean ball.* In this case, $A = \sqrt{n}$, and since $B \geq \frac{1}{\sqrt{n}}$, we have $\chi \geq 1$ – method with Ball setup (i.e., the classical SD) outperforms the method with Simplex setup by factor which varies from $O(\sqrt{\ln n})$ (f is nearly equally sensitive to all variables) to $O(\sqrt{n\ln n})$ (f is sensitive to just $O(1)$ variables).
• *Extreme example II: X is the standard simplex* $\Delta_n(R) = \{x \in \mathbf{R}^n : x \geq 0, \sum\limits_i x_i = 1\}$. In this case, $A = O(1)$, and since $B \leq 1$ and $O(\sqrt{\ln n})$ in practice a moderate absolute constant, we have $\chi \leq O(1)$ – method with Simplex setup outperforms the classical SD by factor which varies from $O\left(\sqrt{\frac{n}{\ln n}}\right)$ (f is nearly equally sensitive to all variables) to $O\left(\sqrt{\frac{1}{\ln n}}\right)$ (f is sensitive to just $O(1)$ variables).

The conclusion is that *there is no once for ever "optimal" MD; which version of the method to choose, it depends on the geometry of the problem. Flexibility of MD as compared to SD to adjust MD, to some extent, to the geometry of the problem of interest.*

Optimality of standard setups

Ball setup and optimization over the ball. As we remember, in the case of the ball setup the number of steps to solve problem (5.3.1) within accuracy ϵ by SD does not exceed

$$N(\epsilon) = O(1)\left(\frac{D_{\|\cdot\|_2}(X)L_{\|\cdot\|_2}(f)}{\epsilon}\right)^2, \tag{5.4.16}$$

where $D_{\|\cdot\|_2}$ is the $\|\cdot\|_2$-diameter of X and $L_{\|\cdot\|_2}(f)$ is the Lipschitz constant of f w.r.t. $\|\cdot\|_2$.

On the other hand, let $L > 0$, and let $\mathcal{P}_{\|\cdot\|_2,L}(X)$ be the family of all convex problems (5.3.1) with convex Lipschitz continuous, with constant L w.r.t. $\|\cdot\|_2$, objectives. It is known that *if X is an n-dimensional Euclidean ball and $n \geq \frac{D^2_{\|\cdot\|_2}(X)L^2}{\epsilon^2}$, then the information-based complexity of the family $\mathcal{P}_{\|\cdot\|_2,L}(X)$ is at least $O(1)\frac{D^2_{\|\cdot\|_2}(X)L^2}{\epsilon^2}$* (cf. (5.2.4)). Comparing this result with (5.4.16), we conclude that

If X is an n-dimensional Euclidean ball, then the complexity of the family $\mathcal{P}_{\|\cdot\|_2,L}(X)$ w.r.t. the MD algorithm with the ball setup in the "large-scale case" $n \geq \frac{D^2_{\|\cdot\|_2}(X)L^2}{\epsilon^2}$ coincides (within a factor depending solely on θ, λ) with the information-based complexity of the family.

Simplex setup and minimization over the simplex. As we remember (cf. (5.4.15)), in the case of the simplex setup the number of steps to solve problem (5.3.1) within accuracy ϵ by MD does not exceed

$$N(\epsilon) = O(1)\ln n \left(\frac{L_{\|\cdot\|_1}(f) \max\limits_{x,y\in X} \|x-y\|_1}{\epsilon} \right)^2, \tag{5.4.17}$$

provided $0 \in X$. On the other hand, let $L > 0$, and let $\mathcal{P}_{\|\cdot\|_1,L}(X)$ be the family of all convex problems (CP) with convex Lipschitz continuous, with constant L w.r.t. $\|\cdot\|_1$, objectives. It is known that *if X is the n-dimensional simplex $\Delta_n(R)$ (or the full-dimensional simplex $\Delta_n^+(R)$) and $n \geq \frac{L^2R^2}{\epsilon^2}$, then the information-based complexity of the family $\mathcal{P}_{\|\cdot\|_1,L}(X)$ is at least $O(1)\frac{L^2R^2}{\epsilon^2}$* (cf. (5.2.5)). Comparing this result with (5.4.17), we conclude that

If X is the n-dimensional simplex $\Delta_n(R)$ (or the full-dimensional simplex $\Delta_n^+(R)$), then the complexity of the family $\mathcal{P}_{\|\cdot\|_1,L}(X)$ w.r.t. the MD algorithm with the simplex setup in the "large-scale case" $n \geq \frac{L^2R^2}{\epsilon^2}$ coincides, within a factor of order of $\ln n$, with the information-based complexity of the family.

Spectahedron setup and large-scale semidefinite optimization. All the conclusions we have made when speaking about the case of the simplex setup and $X = \Delta_n^+(R)$ (or $X = \Delta_n(R)$) remain valid in the case of the spectahedron setup and X defined as the set of all *block-diagonal* matrices of a given block-diagonal structure contained in $\Xi_n^+(R) = \{x \in \mathbf{S}^n : x \succeq 0, \mathrm{Tr}(x) \leq R\}$ (or contained in $\Xi_n(R) = \{x \in \Xi_n^+(R) : \mathrm{Tr}(x) = R\}$).

We see that *with every one of our standard setups, the MD algorithm under appropriate conditions possesses dimension independent (or nearly dimension independent) complexity bound and, moreover, is nearly optimal in the sense of Information-based complexity theory, provided that the dimension is large.*

Why the standard setups? "The contribution" of $\omega(\cdot)$ to the performance estimate (5.4.12) is in the factor $\Theta = \frac{\Omega}{\kappa}$; the less it is, the better. In principle, given X and $\|\cdot\|$, we could play with $\omega(\cdot)$ to minimize Θ. The standard setups are given by a kind of such optimization for the cases when X is the ball and $\|\cdot\| = \|\cdot\|_2$ ("the ball case"), when X is the simplex and $\|\cdot\| = \|\cdot\|_1$ ("the simplex case"), and when X is the spectahedron and $\|\cdot\| = |\cdot|_1$ ("the spectahedron case"), respectively. We did not try to solve the arising variational problems exactly; however, it can be proved in all three cases that the value of Θ we have reached (i.e., $O(1)$ in the ball case and $O(\ln n)$ in the simplex and the spectahedron cases) cannot be reduced by more than an absolute constant factor. Note that in the simplex case the (regularized) entropy is not the only reasonable choice; similar complexity results can be obtained for, say, $\omega(x) = \sum\limits_i x_i^{p(n)}$ or $\omega(x) = \|x\|^2_{p(n)}$ with $p(n) = 1 + O\left(\frac{1}{\ln n}\right)$.

Application example: Positron Emission Tomography Image Reconstruction. The Maximum Likelihood estimate of tracer's density in PET is

$$\lambda_* = \mathrm{argmax}_{\lambda\geq 0} \left\{ \sum_{j=1}^n p_j\lambda_j - \sum_{i=1}^m y_i \ln(\sum_{j=1}^n p_{ij}\lambda_j) \right\}$$
$$\left[y_i \geq 0 \text{ are observations}, p_{ij} \geq 0, p_j = \sum_i p_{ij} \right]$$

For this convex optimization program, the KKT optimality conditions read

$$\lambda_j \left(p_j - \sum_i y_i \frac{p_{ij}}{\sum_\ell p_{i\ell}\lambda_\ell} \right) = 0 \quad \forall j,$$

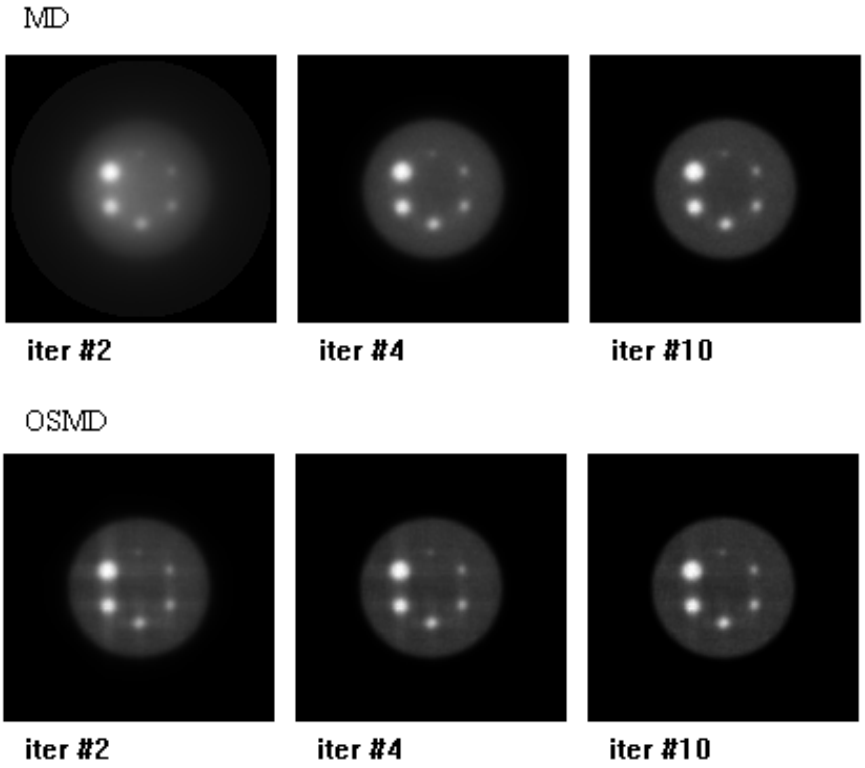
whence, taking sum over j,

$$\sum_j p_j \lambda_j = B \equiv \sum_i y_i.$$

Thus, in fact (PET) is the problem of minimizing over a simplex. Passing to the variables $x_j = p_j B^{-1} \lambda_j$, we end up with the problem

$$\min_x \left\{ f(x) = -\sum_i y_i \ln(\sum_j q_{ij} x_j) : x \in \Delta_n \right\} \qquad \text{(PET)}$$
$$\left[q_{ij} = B p_{ij} p_j^{-1} \right]$$

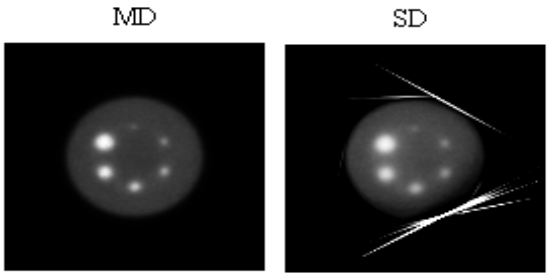
Illustration: "Hot Spheres" phantom ($n = 515,871$).



Itr	1	2	3	4	5	6	7	8	9	10
$f(x_t)$	-4.295	-4.767	-5.079	-5.189	-5.168	-5.230	-5.181	-5.227	-5.189	-5.225

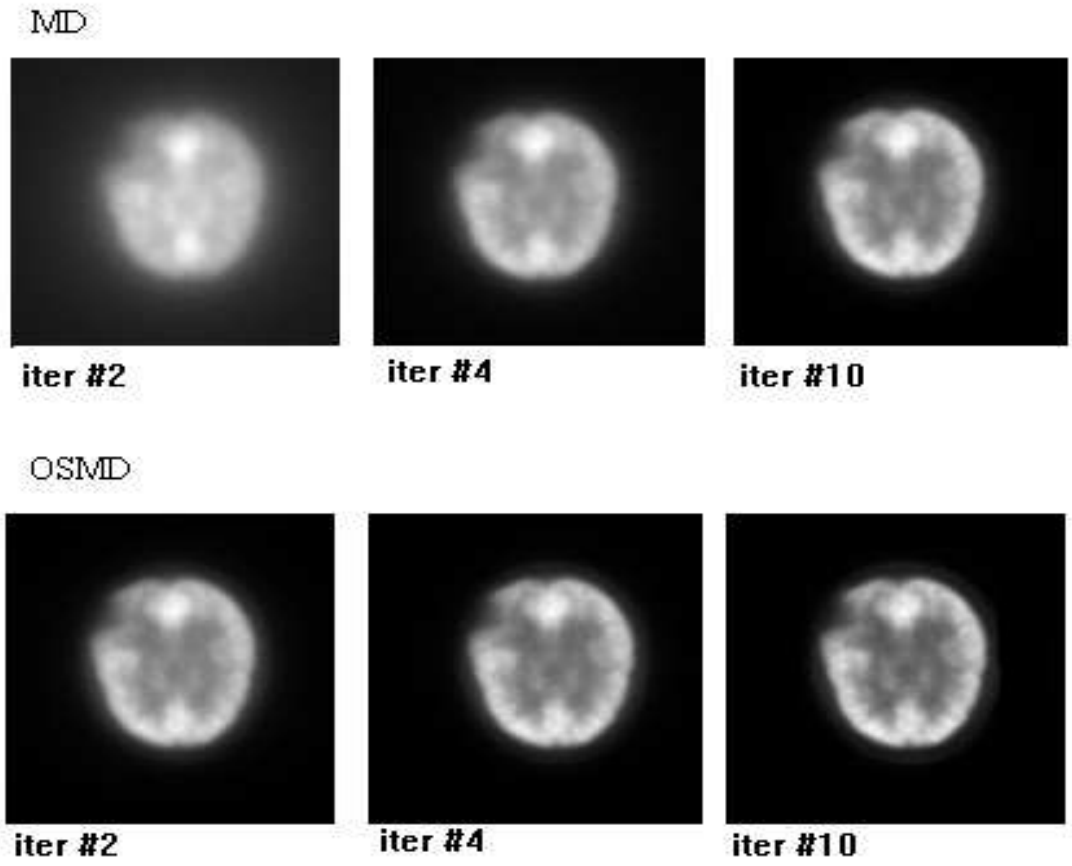
$[f_* \geq -5.283]$

Simplex setup. Progress in accuracy in 10 iterations by factor 21.4



Simplex setup (left) vs. Ball setup (right). Progress in accuracy 21.4 vs. 5.26

Illustration: Brain clinical data ($n = 2,763,635$)



Itr	1	2	3	4	5	6	7	8	9	10
$f(x_t)$	-1.463	-1.848	-2.001	-2.012	-2.015	-2.015	-2.016	-2.016	-2.016	-2.016

$[f_* \geq -2.050]$

Simplex setup. Progress in accuracy in 10 iterations by factor 17.5

5.4.3 Mirror-Level Algorithm

Same as SD, the general Mirror Descent admits a version with memory – Mirror Level (ML) algorithm. The setup for ML is similar to the one for MD and is given by a strongly convex C^1 function $\omega(\cdot)$ on X

and a norm $\|\cdot\|$ on E.

A step of ML. At step t of ML, we

1. compute $f(x_t), f'(x_t)$ and build the current model of f

   $$f_t(x) = \max_{\tau\le t}[f(x_\tau) + \langle f'(x_\tau), x - x_\tau\rangle]$$

   which underestimates the objective and is exact at the points $x_1, ..., x_t$;
2. define the *best found so far value of the objective* $f^t = \min_{\tau\le t} f(x_\tau)$
3. define the current *lower bound* f_t on f_* by solving the auxiliary problem

   $$f_t = \min_{x\in X} f_t(x)$$

   The current *gap* $\Delta_t = f^t - f_t$ is an upper bound on the inaccuracy of the best found so far approximate solution;
4. compute the current *level* $\ell_t = f_t + \lambda\Delta_t$ ($\lambda \in (0,1)$ is a parameter)
5. We solve the optimization problem

   $$d_t = \min_x\{\|x - x_t\| : x \in X, f_t(x) \le \ell_t\}$$

   find e_t, $\|e_t\|_* = 1$, such that

   $$\langle e_t, x_t - x\rangle \ge d_t \quad \forall(x \in X, f_t(x) \le \ell_t),$$

   set

   $$x_{t+1} = \operatorname*{argmin}_{x\in X}\big[\langle \underbrace{\alpha d_t}_{\gamma_t} e_t - \nabla\omega(x_t), x\rangle - \omega(x)\big]$$

   and loop to step $t+1$.

   Observe that with Ball setup,

   - $\gamma_t = \|x_t - z_t\|_2$, where $z_t = \Pi_{L_t}(x_t)$ and

     $$L_t = \{x \in X : f_t(x) \le \ell_t\};$$

   - consequently, $e_t = \frac{x_t - z_t}{\gamma_t}$;
   - consequently,

     $$x_{t+1} = \Pi_X(x_t - \gamma_t e_t) = \Pi_X(x_t - [x_t - z_t]) = \Pi_X(z_t) = z_t = \Pi_{L_t}(x_t),$$

     i.e., the method becomes exactly the BL algorithm.

Why and how ML converges?

Recall that convergence analysis of BL was based on the following fact:

Let $J = \{s, s+1, ..., r\}$ be a segment of iterations of BL, that is, let

$$\Delta_r \ge (1-\lambda)\Delta_s.$$

Then the cardinality of J can be bounded from above as

$$\mathrm{Card}(J) \le \frac{\left(\max\limits_{x,y\in X}\|x - y\|_2 L_{\|\cdot\|_2}(f)\right)^2}{(1-\lambda)^2\Delta_r^2}.$$

Similar fact for ML reads:

(!) *Let $J = \{s, s+1, ..., r\}$ be a segment of iterations of ML, that is, let*

$$\Delta_r \geq (1-\lambda)\Delta_s.$$

Then the cardinality of J can be bounded from above as

$$\mathrm{Card}(J) \leq \frac{(\Theta/\alpha)L^2_{\|\cdot\|}(f)}{(1-\lambda)^2\Delta_r^2}. \tag{5.4.18}$$

From (!), exactly as in the case of BL, one derives

Theorem 5.4.1 *For every ϵ, $0 < \epsilon < \Delta_1$, the number N of steps of ML before a gap $\leq \epsilon$ is obtained* (i.e., before an ϵ-solution is found) *does not exceed the bound*

$$N(\epsilon) = \frac{2(\Theta/\alpha)L^2_{\|\cdot\|}(f)}{\lambda(1-\lambda)^2(2-\lambda)\epsilon^2}.$$

In particular, for Simplex/Spectahedron setup one has

$$N(\epsilon) = O(\ln n)\frac{\left(\max\limits_{x,y\in X}\|x-y\|L_{\|\cdot\|}(f)\right)^2}{\lambda(1-\lambda)^2(2-\lambda)\epsilon^2}.$$

All we need is to verify (!), and here is the verification: Same as in the case of BL, we observe that

1. For t running through a segment of iterations J, the level sets $L_t = \{x \in X : f_t(x) \leq \ell_t\}$ have a point in common, namely, $v \in \operatorname{Argmin}\limits_{x\in X} f_r(x)$;
2. For $t \in J$, the distances $d_t = \min\limits_{x\in L_t}\|x_t - x\|$ are not too small:

$$d_t \geq \frac{(1-\lambda)\Delta_r}{L_{\|\cdot\|}(f)}.$$

When deriving (5.4.5), we have seen that when $\xi \in E$, the updating $x \mapsto x_+ = \mathrm{argmin}_{y\in X}\left[\langle \xi - \nabla\omega(x), y\rangle + \omega(y)\right]$ ensures that

$$\forall u \in X : \langle \xi, x-u\rangle \leq H_u(x) - H_u(x_+) + \frac{1}{2\alpha}\|\xi\|_*^2.$$

Applying this relation to $x = x_t$, $\xi = \gamma_t e_t$, $u = v$, we get

$$\underbrace{\langle \gamma_t e_t, x_t - v\rangle}_{\geq \gamma_t d_t = \frac{\gamma_t^2}{\alpha}} \leq H_v(x_t) - H_v(x_{t+1}) + \frac{1}{2\alpha}\gamma_t^2,$$

whence $H_v(x_t) - H_v(x_{t+1}) \geq \frac{1}{2\alpha}\gamma_t^2 = \frac{\alpha}{2}d_t^2$. Thus,

$$\underbrace{\frac{\alpha}{2}\sum_{t=s}^{r} d_t^2}_{\geq \frac{\alpha(1-\lambda)^2\Delta_r^2}{2L^2_{\|\cdot\|}(f)}\mathrm{Card}(J)} \leq H_v(x_s) - H_v(x_{r+1}) \leq \Theta,$$

and (5.4.18) follows. □

5.4.4 NERML – Non-Euclidean Restricted Memory Level algorithm

The algorithm we are about to present is in the same relation to Mirror Level as Truncated Proximal Bundle-Level is to Bundle-Level - in both cases, we want to get full control on bundle cardinality (and thus – on the complexity of auxiliary problems) while not losing in the theoretical efficiency estimate. Specifically, NERML is a version of ML where the bundle size is kept below a given desired level m. The construction of NERML is very similar to the one of TPBL.

The setup for NERML, same as those for MD and ML, is given by a continuously differentiable and strongly convex on X function $\omega(\cdot)$ and a norm $\|\cdot\|$ on the Euclidean space E where X lives.

Execution of NERML is split into *phases*. Phase s is associated with

1. prox-center $c_s \in X$
2. s-th upper bound f^s on f_*, which is the best value of the objective observed before the phase begins,
3. s-th lower bound f_s on f_*, which is the best lower bound on f_* observed before the phase begins.

   f^s and f_s define
   - s-th optimality gap $\Delta_s = f^s - f_s$;
   - s-th level $\ell_s = f_s + \lambda\Delta_s$, where $\lambda \in (0,1)$ is parameter of the method;
   - s-th local distance
   $$\omega_s(x) = \omega(x) - \langle \nabla\omega(c_s), x\rangle - \omega(c_s)$$
4. current model $\widetilde{f}^s(\cdot) \le f(\cdot)$ of $f(\cdot)$, which is the maximum of $\le m$ affine forms.

To initialize the first phase, we choose $c_1 \in X$, compute $f(c_1), f'(c_1)$ and set

$$\widetilde{f}^1(x) = f(c_1) + \langle f'(c_1), x - c_1\rangle, \quad f^1 = f(c_1), \quad f_1 = \min_{x\in X} \widetilde{f}^1(x).$$

Steps of a phase. At the beginning of step $t = 1, 2, ...$ of phase s, we have

- upper bound $f^{s,t-1} \le f^s$ on f_*, which is the best found so far value of the objective,
- lower bound $f_{s,t-1} \ge f_s$ on f_*,
- model $\widetilde{f}^{s,t-1}(\cdot) \le f(\cdot)$ of the objective which is the maximum of $\le m$ affine forms
- iterate $x_t \in X$ and set $H_{t-1} = \{x : \langle \alpha_{t-1}, x\rangle \ge \beta_{t-1}\}$ such that

$$\begin{array}{cc} x \in X, f(x) \le \ell_s \Rightarrow x \in H_{t-1} & (a_t) \\ x_t = \operatorname{argmin}_x \{\omega_s(x) : x \in H_{t-1} \cap X\} & (b_t) \end{array} \tag{5.4.19}$$

To initialize the first step of phase s, we set

$$f^{s,0} = f^s,\ f_{s,0} = f_s,\ \widetilde{f}^{s,0}(\cdot) = \widetilde{f}^s(\cdot),\ \alpha_0 = 0, \beta_0 = 0\ [\Rightarrow H_0 = E]$$

thus ensuring (5.4.19.a_1), and set $x_1 = c_s$, thus ensuring (5.4.19.b_1).

Step t phase s. Given

- bounds $f^{s,t-1} \geq f_*$, $f_{s,t-1} \leq f_*$,
- model $\widetilde{f}^{s,t-1}(\cdot) \leq f(\cdot)$,
- x_t and $H_{t-1} = \{x : \langle \alpha_{t-1}, x\rangle \geq \beta_{t-1}\}$ such that

$$\begin{array}{rl} x \in X, f(x) \leq \ell_s \Rightarrow x \in H_{t-1} & (a_t) \\ x_t = \operatorname*{argmin}_x \{\omega_s(x) : x \in H_{t-1} \cap X\} & (b_t) \end{array}$$

we act as follows:

1. compute $f(x_t), f'(x_t)$ and set $g_t(x) = f(x_t) + \langle f'(x_t), x - x_t\rangle$;
2. Define $\widetilde{f}^{s,t}(\cdot)$ as the maximum of $g_t(\cdot)$ and affine forms associated with $\widetilde{f}^{s,t-1}$ (dropping, if necessary, one of the latter forms to make $\widetilde{f}^{s,t}$ the maximum of at most m forms).
   If $f(x_t) \leq \ell_s + 0.5(f^s - \ell_s)$ ("significant progress in the upper bound"), we terminate phase s and set
   $$f^{s+1} = f^{s,t}, \quad f_{s+1} = f_{s,t-1}, \quad \widetilde{f}^{s+1}(\cdot) = \widetilde{f}^{s,t}(\cdot),$$
   otherwise
3. Compute $f_t = \min_x \left\{\widetilde{f}^{s,t}(x) : x \in H_{t-1} \cap X\right\}$. Since $f(x) \geq \ell_s$ in $X \backslash H_{t-1}$, we have $f_* \geq \min[\ell_s, f_t]$, so that
   $$f_{s,t} \equiv \max\{f_{s,t-1}, \min[\ell_s, f_t]\} \leq f_*.$$
   If $f_{s,t} \geq \ell_s - 0.5(\ell_s - f_s)$ ("significant progress in the lower bound"), we terminate phase s and set
   $$f^{s+1} = f^{s,t}, \quad f_{s+1} = f_{s,t}, \quad \widetilde{f}^{s+1}(\cdot) = \widetilde{f}^{s,t}(\cdot)$$
   otherwise we set
   $$\begin{array}{rcl} x_{t+1} & = & \operatorname*{argmin}_x \left\{\omega_s(x) : x \in X \cap H_{t-1}, \widetilde{f}^{s,t}(x) \leq \ell_s\right\} \\ H_t & = & \{x : \langle \nabla \omega_s(x_{t+1}), x - x_{t+1}\rangle \geq 0\} \end{array}$$
   and loop to step $t+1$ of phase s.

Note: When passing to step $t+1$, we ensure the relations

$$\begin{array}{rl} x \in X, f(x) \leq \ell_s \Rightarrow x \in H_t & (a_{t+1}) \\ x_{t+1} = \operatorname*{argmin}_x \left\{\omega_s(x) : x \in X \cap H_t, \widetilde{f}^{s,t}(x) \leq \ell\right\} & (b_{t+1}) \end{array}$$

Indeed, x_{t+1} is the minimizer of $\omega_s(x)$ on the set

$$Y_t = X \cap H_{t-1} \cap \{x : \widetilde{f}^{t,s}(x) \leq \ell_s\}$$

whence

$$\begin{array}{c} \langle \nabla \omega_s(x), x - x_{t+1}\rangle \geq 0 \ \forall x \in Y_t \\ \Downarrow \\ Y_t \subset H_t = \{x : \langle \nabla \omega_s(x_{t+1}), x - x_{t+1}\rangle \geq 0\} \quad (*) \end{array}$$

Thus,

$$\begin{array}{l} (x \in X, f(x) \leq \ell_s) \underbrace{\Rightarrow}_{(a_t)} (x \in X \cap H_{t-1}, f(x) \leq \ell_s) \\ \Rightarrow (x \in X \cap H_{t-1}, \widetilde{f}^{s,t}(x) \leq \ell_s) \underbrace{\Rightarrow}_{(*)} x \in H_t \end{array}$$

as required in (a_{t+1}). (b_{t+1}) readily follows from the definition of H_t.

Convergence of NERML

Recall that the efficiency estimate for TPBL was a nearly straightforward consequence of the following fact:

(*) *The number of steps of TPBL at a phase s does not exceed*

$$N_s = \frac{4\left(\max\limits_{x,y} \|x-y\|_2 L_{\|\cdot\|_2}(f)\right)^2}{(1-\lambda)^2\Delta_s^2} + 1.$$

For NERML, a similar fact is valid:

(!) *The number of steps of NERML at a phase s does not exceed*

$$N_s = \frac{8(\Theta/\alpha)L^2_{\|\cdot\|}(f)}{(1-\lambda)^2\Delta_s^2} + 1.$$

The same reasoning as in the case of TPBL, with (!) playing the role of (*), yields

Theorem 5.4.2 *For every ϵ, $0 < \epsilon < \Delta_1$, the total number of NERML steps before a gap $\le \epsilon$ is obtained* (i.e., before an ϵ-solution is found) *does not exceed the bound*

$$N(\epsilon) = c(\lambda)\frac{(\Theta/\alpha)L^2_{\|\cdot\|}(f)}{\epsilon^2}.$$

All we need is to verify (!), and here is the verification:

Assume that phase s was not terminated in course of N steps. By construction, for $1 \le t \le N$ we have

$$\begin{array}{rcl} x_{t+1} \in H_{t-1}\cap X \ \&\ x_t = \operatorname{argmin}\limits_x\{\omega_s(x) : x \in H_{t-1}\cap X\} & & \\ \Downarrow & & \\ \omega_s(x_{t+1}) & \ge & \omega_s(x_t) + \underbrace{\langle \nabla\omega(x_t), x_{t+1}-x_t\rangle}_{\ge 0} + \frac{\alpha}{2}\|x_{t+1}-x_t\|^2 \\ & \ge & \omega_s(x_t) + \frac{\alpha}{2}\|x_{t+1}-x_t\|^2 \end{array} \qquad (5.4.20)$$

Further, when passing from x_t to

$$x_{t+1} = \operatorname*{argmin}_x\left\{\omega_s(x) : x \in H_{t-1}\cap X, \widetilde{f}^{s,t}(x) \le \ell_s\right\},$$

the function $g_t(x) \equiv f(x_t) + \langle f'(x_t), x - x_t\rangle \le \widetilde{f}^{s,t}$ varies from the value $f(x_t) \ge f^{s,t}$ to a value $\le \ell_s$ and thus decreases by at least $0.5(1-\lambda)\Delta_s$ (otherwise phase s would be terminated at step t due to significant progress in the best found so far value of the objective). Since $g_t(\cdot)$ is Lipschitz continuous, with constant $L_{\|\cdot\|}(f)$ w.r.t. $\|\cdot\|$, we conclude that

$$0.5(1-\lambda)\Delta_s \le \|x_t - x_{t+1}\|L_{\|\cdot\|}(f) \Rightarrow \|x_t - x_{t+1}\| \ge \frac{0.5(1-\lambda)\Delta_s}{L_{\|\cdot\|}(f)}.$$

Applying the resulting inequality in (5.4.20), we arrive at

$$\omega_s(x_{t+1}) \ge \omega_s(x_t) + \frac{\alpha(1-\lambda)^2}{8L^2_{\|\cdot\|}(f)}\Delta_s^2,\ 1 \le t \le N. \qquad (5.4.21)$$

On the other hand, the function $\omega_s(x) = \omega(x) - \langle\nabla\omega(c_s), x - c_S\rangle + \omega(c_s)$ on the set X varies between 0 and Θ, and (!) follows from (5.4.21). □

5.5 Implementation issues and illustrations

5.5.1 Implementing SD and MD

Recall that our Implementability assumption on the setup of MD and related methods is that one can easily solve auxiliary problems of the form

$$\min_{x\in X}\left\{\langle\xi,x\rangle+\omega(x)\right\}. \tag{5.5.1}$$

Under this assumption, there is no difficulty to implement SD and MD – both methods require at every step solving a single auxiliary problem of the outlined type.

Implementing NERML

The situation with methods with memory is more complicated, since here auxiliary problems are of a more general form. For the sake of definiteness, let us restrict ourselves with the case of NERML (a reader can think of how the subsequent recommendations should be modified in the case of the ML algorithm). Note that in NERML we have to handle two types of auxiliary problems:

$$\begin{array}{ll}(a) & \min\limits_{x}\left\{\max\limits_{i}\phi(x):\langle\alpha,x\rangle\le 0, x\in X\right\}\\ (b) & \min\limits_{x}\left\{\langle a,x\rangle+\omega(x):\max\limits_{j}\psi_j(x)\le 0, x\in X\right\}\end{array} \tag{5.5.2}$$

where $\phi(\cdot)$ and $\psi(\cdot)$ are maxima of given affine functions. Problems (a) arise when updating current lower bounds on the optimal value, while problems (b) are responsible for updating the iterates. In principle, there are to ways to solve these problems.

A. Straightforward approach. Depending on the structure and the sizes of X and ω, we can solve the auxiliary problems "as they are" by dedicated high-performance optimization techniques. For example, when X is given by a list of linear constraints and the ball setup is used, the auxiliary problems are those of minimizing linear (or convex quadratic) objectives under linear inequality constraints, and to solve these problems, one could use well-developed pivoting or interior point methods fir convex quadratic optimization under linear constraints. This is how all standard BL/TPBL methods are implemented.

B. Exploiting duality. A severe disadvantage of the straightforward approach is that when the problem of interest is an extremely large-scale one (which is *the* major case we are interested in), so are the auxiliary problems, which can make their solution by traditional routines too time-consuming. An attractive alternative is offered by Lagrange Duality. Specifically, it is easily seen that in the NERML context all auxiliary problems to be solved are strictly feasible in the sense that their feasible sets intersect the relative interior of X. this, combined with the fact that the functional constraints $\phi(x)\le 0$ in these problems are linear, combines with the standard Lagrange duality to imply that the optimal values in (a), (b) are equal to those in their (solvable!) Lagrange dual problems

$$\begin{array}{ll}(a_*) & \max\limits_{\lambda\ge 0,\mu_i\ge 0,\sum_i\mu_i=1} H(\lambda,\mu),\quad H(\lambda,\mu)=\min\limits_{x\in X}\left[\sum\limits_i\mu_i\phi(x)+\lambda(\langle\alpha,x\rangle-\beta)\right]\\ (b_*) & \max\limits_{\lambda_j\ge 0} F(\lambda),\quad F(\lambda)=\min\limits_{x\in X}\left[\langle a,x\rangle+\omega(x)+\sum\limits_j\lambda_j\psi_j(x)\right]\end{array} \tag{5.5.3}$$

Besides this, since $\omega(\cdot)$ is strongly convex, a high-accuracy solution $\bar\lambda$ to (5.5.3.b) implies a high-accuracy solution

$$\bar x=\operatorname*{argmin}_{x\in X}\left[\langle a,x\rangle+\omega(x)+\sum_j\bar\lambda_j\psi_j(x)\right].$$

Now observe that *by our initial assumption on the easiness of (5.5.1), both $H(\lambda,\mu)$ and $F(\lambda)$ are easily computable, along with their supergradients, at any given point.* For example, to compute $H(\lambda,\mu)$ and

$H'_\lambda(\lambda,\mu)$, $H'_\mu(\mu,\lambda)$, one should compute the point $\bar{x} = \mathrm{argmin}_{x\in X}\left[\sum_i \mu_i\phi(x) + \lambda(\langle\alpha, x\rangle - \beta)\right]$ (which is assumed to be easy) and to set

$$H(\lambda,\mu) = \sum_i \mu_i\phi(\bar{x}) + \lambda(\langle\alpha,\bar{x}x\rangle - \beta),\ H'_\lambda(\lambda,\mu) = \phi(\bar{x}),\ H'_m u(\lambda,\mu) = (\langle\alpha,\bar{x}\rangle - \beta).$$

Now, since *the design dimension of the dual problems is under full our control and thus can be enforced to be reasonably small*, we can solve the dual problems rapidly by *low dimensional* black-box oriented convex optimization techniques (Ellipsoid method, Bundle-Level with straightforward implementation, etc.). thus getting the required optimal values of the "actual" auxiliary problems and a high-accuracy solution to the second of them (for the first of them, only the optimal value is needed).

When the Implementability assumption is satisfied?

The bottom line of the above considerations is that essentially what we need to make the outlined methods practical is the Implementability assumption. The latter takes place, e.g., in the case of

- *Ball setup and simple X* (ball, box, positive part of ball, standard simplex,...),
- *Simplex setup and simple X* (the simplexes $\Delta_n^+(R)$, $\Delta_n(R)$, the intersection of $\Delta_n(R)$ and a box,...)
- *Spectahedron setup with X* comprised of block-diagonal matrices with diagonal blocks of size $O(1)$.

It is easily seen that in all these cases, problem (5.5.1) can be solved within accuracy ϵ in $O(1)n\ln n\ln(1/\epsilon)$ arithmetic operations, that is, it takes $O(1)n\ln n$ a.o. to solve (5.5.1) within machine precision. We are about to support this claim.

Ball setup. Here problem (5.5.1) becomes

$$\min_{x\in X}\left[\frac{1}{2}x^Tx - p^Tx\right],$$

or, which is the same,

$$\min_{s\in X}\left[\frac{1}{2}\|x-p\|_2^2\right].$$

We see that to solve (5.5.1) is the same as to *project* on X - to find the point in X which is as close as possible, in the usual $\|\cdot\|_2$-norm, to a given point p. This problem is easy to solve for several simple solids X, e.g.,

- a ball $\{x : \|x-a\|_2 \le r\}$,
- a box $\{x : a \le x \le b\}$,
- the simplex $\Delta_n(R) = \{x : x \ge 0, \sum_i x_i = R\}$.

In the first two cases, it takes $O(n)$ operations to compute the solution – it is given by evident explicit formulas. In the third case, to project is a bit more involving: you can easily demonstrate that the projection is given by the relations $x_i = x_i(\lambda_*)$, where $x_i(\lambda) = \max[0, p_i - \lambda]$ and λ_* is the unique root of the equation

$$\sum_i x_i(\lambda) = R.$$

The left hand side of this equation is nonincreasing and continuous in λ and, as it is immediately seen, its value varies from something $\ge R$ when $\lambda = \max_i p_i - R$ to 0 when $\lambda = \max_i p_i$. It follows that one can easily approximate λ_* by Bisection, and that it takes a moderate absolute constant of bisection steps to compute λ_* (and thus – the projection) within the machine precision. The arithmetic cost of a Bisection step clearly is $O(n)$, and the overall arithmetic complexity of finding the projection becomes $O(n)$.

Simplex setup. Let us restrict ourselves with the two simplest cases:
S.A: X is the standard simplex $\Delta_n(R) = \{x \in \mathbf{R}^n : x \geq 0, \sum_i x_i = R\}$;
S.B: X is the standard full-dimensional simplex $\Delta_n^+(R) = \{x \in \mathbf{R}^n : x \geq 0, \sum_i x_i \leq R\}$.

By evident scaling arguments, we lose nothing when setting $R = 1$.
Case S.A. When $X = \Delta_n(1)$, the problem (5.5.1) becomes

$$\min\left\{\sum_i (x_i + \sigma)\ln(x_i + \sigma) - p^T x : x \geq 0, \sum_i x_i = 1\right\} \qquad [\sigma = \delta n^{-1}] \tag{5.5.4}$$

The solution to this optimization problem, as it is immediately seen, is given by $x_i = x_i(\lambda_*)$, where

$$x_i(\lambda) = \max[\exp\{\widehat{p}_i - \lambda\} - \sigma, 0] \qquad [\widehat{p}_i = p_i - \max_j p_j] \tag{5.5.5}$$

and λ_* is the solution to the equation

$$\sum_i x_i(\lambda) = 1.$$

Here again the left hand side of the equation is nonincreasing and continuous in λ and, as it is immediately seen, its value varies from something which is ≥ 1 when $\lambda = -\sigma$ to something which is < 1 when $\lambda = \ln n$, so that we again can compute λ_* (and thus – $x(\lambda_*)$) within machine precision, in a realistic range of values of n, in a moderate absolute constant of bisection steps. As a result, the arithmetic cost of solving (5.5.4) is again $O(n)$.

Note that "numerically speaking", we should not bother about Bisection at all. Indeed, let us set δ to something really small, say, $\delta =$ 1.e-16. Then $\sigma = \delta n^{-1} <<$ 1.e-16, while (at least some of) $x_i(\lambda_*)$ should be of order of $1/n$ (since their sum should be 1). It follows that with actual (i.e., finite precision) computations, the quantity σ in (5.5.5) is negligible. Omitting σ in (5.5.4) (i.e., replacing in (5.5.1) the regularized entropy by the usual one), we can explicitly write down the solution x_* to (5.5.4):

$$x_i = \frac{\exp\{-\widehat{p}_i\}}{\sum_j \exp\{-\widehat{p}_j\}},\ i = 1, ..., n.$$

Case S.B. The case of $X = \Delta_n^+(1)$ is very close to the one of $X = \Delta_n(1)$. The only difference is that now we first should check whether

$$\sum_i \max\left[\exp\{-1 - p_i\} - \delta n^{-1}, 0\right] \leq 1;$$

if it is the case, then the optimal solution to (5.5.1) is given by

$$x_i = \max\left[\exp\{-1 - p_i\} - \delta n^{-1}, 0\right],\ i = 1, ..., n,$$

otherwise the optimal solution to (5.5.1) is exactly the optimal solution to (5.5.4).

Spectahedron setup. Consider the case of the spectahedron setup, and assume that either
Sp.A: X is comprised of all block-diagonal matrices, of a given block-diagonal structure, belonging to $\Xi_n(R)$,
or
Sp.B: X is comprised of all block-diagonal matrices, of a given block-diagonal structure, belonging to $\Xi_n^+(R)$.
As above, we lose nothing by assuming $R = 1$.
Case Sp.A. Here the problem (5.5.1) becomes

$$\min_{x \in X}\{\mathrm{Tr}((x + \sigma I_n)\ln(x + \sigma I_n)) + \mathrm{Tr}(px)\} \qquad [\sigma = \delta n^{-1}]$$

We lose nothing by assuming that p is a symmetric block-diagonal matrix of the same block-diagonal structure as the one of matrices from X. Let $p = U\pi U^T$ be the singular value decomposition of p with orthogonal U and diagonal π of the same block-diagonal structure as that one of p. Passing from x to the new matrix variable ξ according to $x = U\xi U^T$, we convert our problem to the problem

$$\min_{\xi\in X}\{\mathrm{Tr}((\xi+\sigma I_n)\ln(\xi+\sigma I_n)) + \mathrm{Tr}(\pi\xi)\} \tag{5.5.6}$$

We claim that the unique (due to strong convexity of ω) optimal solution ξ^* to the latter problem is a diagonal matrix. Indeed, for every diagonal matrix D with diagonal entries ± 1 and for every feasible solution ξ to our problem, the matrix $D\xi D$ clearly is again a feasible solution with the same value of the objective (recall that π is diagonal). It follows that the optimal set $\{\xi^*\}$ of our problem should be invariant w.r.t. the aforementioned transformations $\xi \mapsto D\xi D$, which is possible if and only if ξ^* is a diagonal matrix. Thus, when solving (5.5.6), we may from the very beginning restrict ourselves with diagonal ξ, and with this restriction the problem becomes

$$\min_{\xi\in\mathbf{R}^n}\left\{\sum_i(\xi_i+\sigma)\ln(\xi_i+\sigma) + \pi^T\xi : \xi\geq 0, \sum_i \xi_i = 1\right\}, \tag{5.5.7}$$

which is exactly the problem we have considered in the case of the simplex setup and $X = \Delta_n$. We see that the only elaboration in the case of the spectahedron setup as compared to the simplex one is in the necessity to find the singular value decomposition of p. The latter task is easy, provided that the diagonal blocks in the matrices in question are of small sizes. Note that this favourable situation does occur in several important applications, e.g., in Structural Design.

Case Sp.B. This case is completely similar to the previous one; the only difference is that the role of (5.5.6) is now played by the problem

$$\min_{\xi\in\mathbf{R}^n}\left\{\sum_i(\xi_i+\sigma)\ln(\xi_i+\sigma) + \pi^T\xi : \xi\geq 0, \sum_i \xi_i \leq 1\right\},$$

which we have already considered when speaking about the simplex setup.

Updating prox-centers

The complexity results stated in Theorem 5.4.2 are absolutely independent of how we update the prox-centers, so that in this respect we, *in principle*, are completely free. The common sense, however, says that the most natural policy here is to use as the prox-center at every stage the best (with the smallest value of f) solution among those we have at our disposal at the beginning of the stage.

5.5.2 Illustration: PET Image Reconstruction problem

To get an impression of the practical performance of the NERML algorithm, let us look at numerical results related to the 2D version of the PET Image Reconstruction problem.

The model. We process simulated measurements as if they were registered by a ring of 360 detectors, the inner radius of the ring being 1 (Fig. 5.1). The field of view is a concentric circle of the radius 0.9, and it is covered by the 129×129 rectangular grid. The grid partitions the field of view into $10,471$ pixels, and we act as if tracer's density was constant in every pixel. Thus, the design dimension of the problem (PET$'$) we are interested to solve is "just" $n = 10471$.

The number of bins (i.e., number m of log-terms in the objective of (PET$'$)) is 39784, while the number of nonzeros among q_{ij} is 3,746,832.

The true image is "black and white" (the density in every pixel is either 1 or 0). The measurement time (which is responsible for the level of noise in the measurements) is mimicked as follows: we model the measurements according to the Poisson model *as if* during the period of measurements the expected number of positrons emitted by a single pixel with unit density was a given number M.

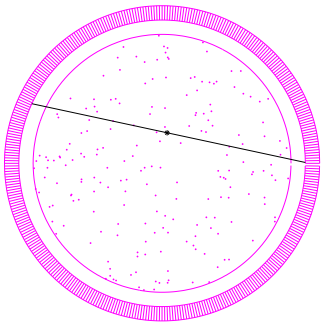


Figure 5.1. Ring with 360 detectors, field of view and a line of response

The algorithm we are using to solve (PET′) is the plain NERML method with the simplex setup. The parameters λ, θ of the algorithm were chosen as

$$\lambda = 0.95, \quad \theta = 0.5.$$

The approximate solution reported by the algorithm at a step is the best found so far search point (the one with the best value of the objective we have seen to the moment).

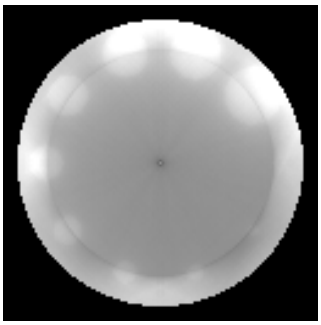
The results of two sample runs we are about to present are not that bad.

Experiment 1: Noiseless measurements. The evolution of the best, in terms of the objective, solutions x^t found in course of the first t calls to the oracle is displayed at Fig. 5.2 (on pictures, brighter areas correspond to higher density of the tracer). The numbers are as follows. With the noiseless measurements, we know in advance the optimal value in (PET′) – it is easily seen that without noises, the true image (which in our simulated experiment we do know) is an optimal solution. In our problem, this optimal value equals to 2.8167; the best value of the objective found in 111 oracle calls is 2.8171 (optimality gap 4.e-4). The progress in accuracy is plotted on Fig. 5.3. We have built totally 111 search points, and the entire computation took $18'51''$ on a 350 MHz Pentium II laptop with 96 MB RAM.

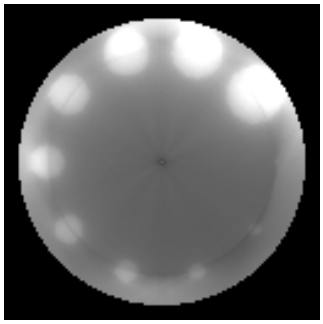
Experiment 2: Noisy measurements (40 LOR's per pixel with unit density, totally 63,092 LOR's registered). The pictures are presented at Fig. 5.4. Here are the numbers. With noisy measurements, we have no a priori knowledge of the true optimal value in (PET′); in simulated experiments, a kind of orientation is given by the value of the objective at the true image (which is hopefully a close to f_* *upper* bound on f_*). In our experiment, this bound equals to -0.8827. The best value of the objective found in 115 oracle calls is -0.8976 (which is less that the objective at the true image in fact, the algorithm went below the value of f at the true image already after 35 oracle calls). The upper bound on the optimality gap at termination is 9.7e-4. The progress in accuracy is plotted on Fig. 5.5. We have built totally 115 search points; the entire computation took $20'41''$.

True image: 10 “hot spots”
$f = 2.817$

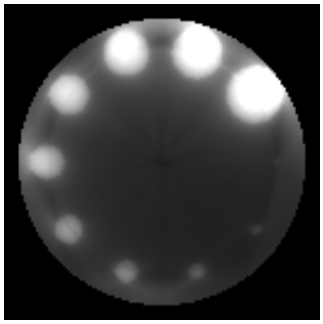
$x^1 = n^{-1}(1, ..., 1)^T$
$f = 3.247$



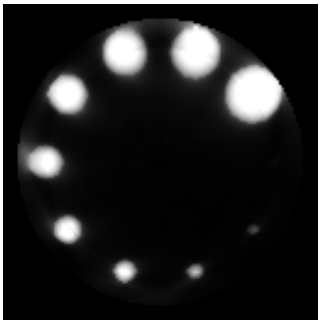
x^2 – some traces of 8 spots
$f = 3.185$



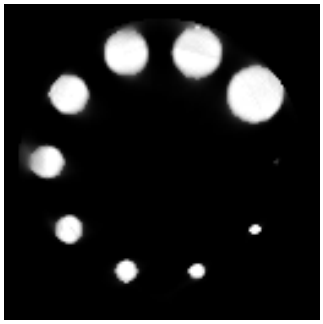
x^3 – traces of 8 spots
$f = 3.126$



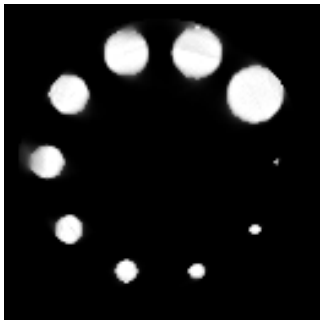
x^5 – some trace of 9-th spot
$f = 3.016$



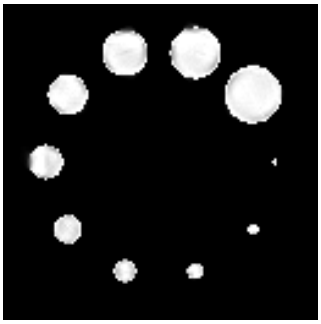
x^8 – 10-th spot still missing...
$f = 2.869$



x^{24} – trace of 10-th spot
$f = 2.828$



x^{27} – all 10 spots in place
$f = 2.823$



x^{31} – that is it...
$f = 2.818$

Figure 5.2. Reconstruction from noiseless measurements

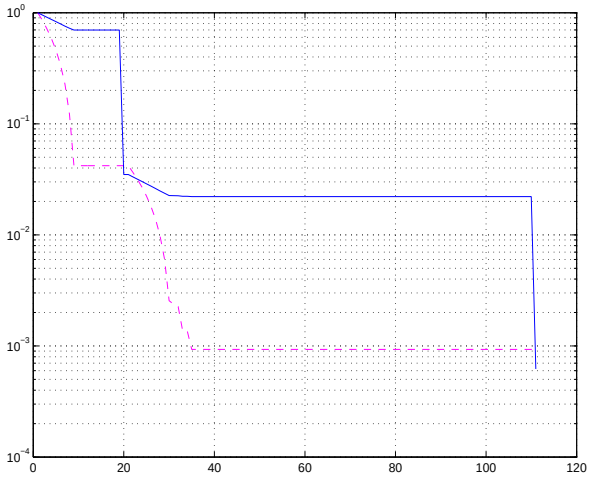


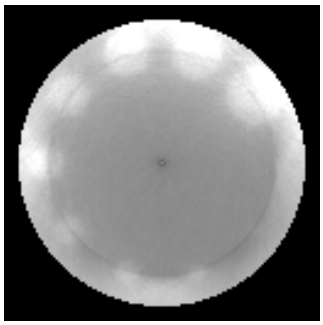
Figure 5.3. Progress in accuracy, noiseless measurements.

solid line: Relative gap $\frac{\text{Gap}(t)}{\text{Gap}(1)}$ vs. step number t; $\text{Gap}(t)$ is the difference between the best found so far value $f(x^t)$ of f and the current lower bound on f_*.
In 111 steps, the gap was reduced by factor > 1600

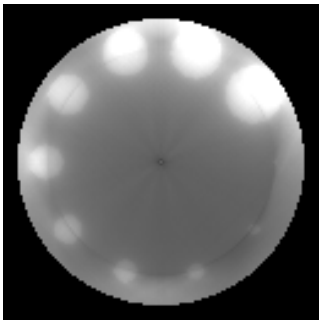
dashed line: Progress in accuracy $\frac{f(x^t)-f_*}{f(x^1)-f_*}$ vs. step number t
In 111 steps, the accuracy was improved by factor > 1080

True image: 10 "hot spots"
$f = -0.883$

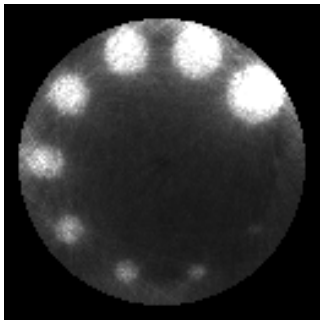
$x^1 = n^{-1}(1, ..., 1)^T$
$f = -0.452$



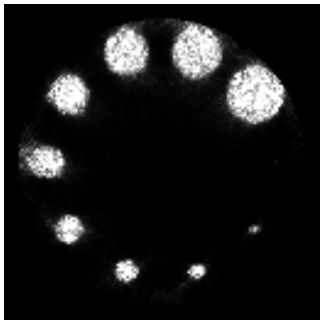
x^2 – light traces of 5 spots
$f = -0.520$



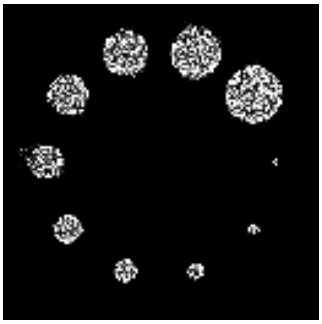
x^3 – traces of 8 spots
$f = -0.585$



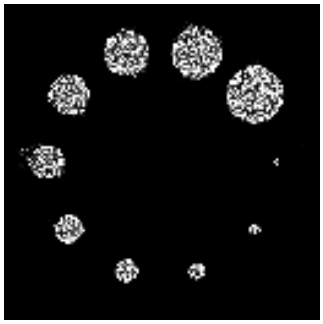
x^5 – 8 spots in place
$f = -0.707$



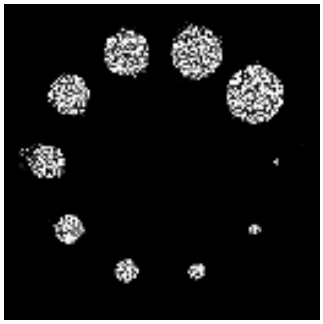
x^8 – 10th spot still missing...
$f = -0.865$



x^{12} – all 10 spots in place
$f = -0.872$



x^{35} – all 10 spots in place
$f = -0.886$



x^{43} – ...
$f = -0.896$

Figure 5.4. Reconstruction from noisy measurements

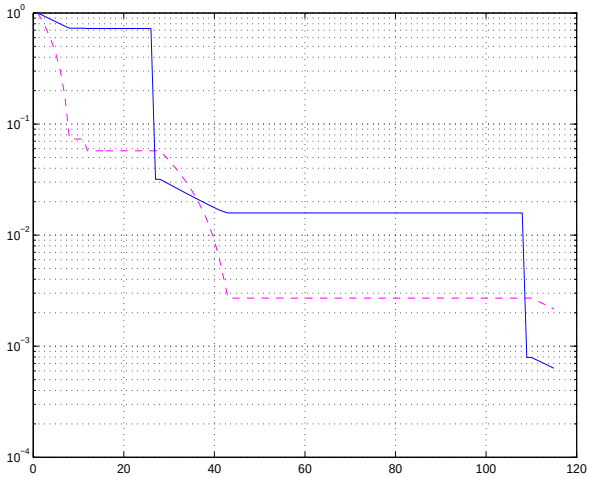


Figure 5.5. Progress in accuracy, noisy measurements.

solid line: Relative gap $\frac{\mathrm{Gap}(t)}{\mathrm{Gap}(1)}$ vs. step number t
In 115 steps, the gap was reduced by factor 1580

dashed line: Progress in accuracy $\frac{f(x^t)-\underline{f}}{f(x^1)-\underline{f}}$ vs. step number t
($\underline{f}$ is the last lower bound on f_* built in the run)
In 115 steps, the accuracy was improved by factor > 460

5.6 Appendix: strong convexity of $\omega(\cdot)$ for standard setups

The case of the ball setup is trivial.

The case of the simplex setup: W.l.o.g., we may assume that $R = 1$. For a C^2 function $\omega(\cdot)$, a sufficient condition for (5.4.6) is the relation

$$h^T\omega''(x)h \geq \alpha\|h\|^2 \quad \forall(x, h : x, x+h \in X) \tag{5.6.1}$$

(why?) For the simplex setup, we have

$$\begin{array}{rcl}\|h\|_1^2 & = & \left[\sum\limits_i |h_i|\right]^2 = \left[\sum\limits_i \frac{|h_i|}{\sqrt{x_i+\delta n^{-1}}}\sqrt{x_i+\delta n^{-1}}\right]^2 \\ & \leq & \left[\sum\limits_i (x_i+\delta n^{-1})\right]\left[\sum\limits_i \frac{h_i^2}{x_i+\delta n^{-1}}\right] \leq (1+\delta)h^T\omega''(x)h,\end{array}$$

and (5.6.1) indeed is satisfied with $\alpha = (1+\delta)^{-1}$.

To prove (5.4.14), note that for all $x, y \in \Delta_n^+ \supset X$, setting $\bar{x} = x + \delta n^{-1}(1, ..., 1)^T$, $\bar{y} = y + \delta n^{-1}(1, ..., 1)^T$, one has

$$\begin{array}{rcl}
& \omega(y) - \omega(x) - (y-x)^T\nabla\omega(x) & \\
= & \sum\limits_i [\bar{y}_i \ln(\bar{y}_i) - \bar{x}_i\ln(\bar{x}_i) - (\bar{y}_i - \bar{x}_i)(1+\ln(\bar{x}_i))] \leq -\sum\limits_i(\bar{y}_i - \bar{x}_i) + \sum\limits_i \bar{y}_i\ln(\frac{\bar{y}_i}{\bar{x}_i}) & \\
\leq & 1 + \delta + \sum\limits_i \bar{y}_i\ln(\frac{\bar{y}_i}{\bar{x}_i}) & \\
& & [\text{since } \bar{y}_i \geq 0 \text{ and } \sum\limits_i \bar{x}_i \leq 1+\delta] \\
\leq & 1 + \delta + \sum\limits_i \bar{y}_i\ln(\frac{\bar{y}_i}{\delta n^{-1}}) & \\
& & [\text{since } \bar{x}_i \geq \delta n^{-1}] \\
\leq & 1 + \delta + \max\limits_z \left\{\sum\limits_i z_i\ln(nz_i/\delta) : z \geq 0, \sum\limits_i z_i \leq 1+\delta\right\} & \\
& & [\text{since } \sum\limits_i \bar{y}_i \leq 1+\delta] \\
= & (1+\delta)\left[1 + \ln\left(\frac{n(1+\delta)}{\delta}\right)\right] &
\end{array}$$

and (5.4.14) follows.

The case of the spectahedron setup: We again assume w.l.o.g. that $R = 1$ and again intend to use the sufficient condition (5.6.1) for strong convexity, but now it is a bit more involving. First of all, let us compute the second derivative of the regularized matrix entropy

$$\omega(x) = \mathrm{Tr}((x+\sigma I_n)\ln(x+\sigma I_n)) : \Xi_n^+(1) \to \mathbf{R} \qquad [\sigma = \delta m^{-1}]$$

Setting $y[x] = x + \sigma I_n$,

$$f(z) = z\ln z$$

(z is complex variable restricted to belong to the open right hand half-plane, and $\ln z$ is the principal branch of the logarithm in this half-plane), in a neighbourhood of a given point $\bar{x} \in \Xi_n^+(1)$ we have, by Cauchy's integral formula,

$$Y(x) \equiv y[x]\ln(y[x]) = \frac{1}{2\pi i}\oint\limits_\gamma f(z)(zI_n - y[x])^{-1}dz, \tag{5.6.2}$$

where γ is a closed contour in the right half-plane with all the eigenvalues of $y[\bar{x}]$ inside the contour. Consequently,

$$\begin{array}{rcl} DY(x)[h] & = & \frac{1}{2\pi i}\oint\limits_\gamma f(z)(zI_n - y[x])^{-1}h(zI_n - y[x])^{-1}dz, \\ D^2Y(x)[h,h] & = & \frac{1}{\pi i}\oint\limits_\gamma f(z)(zI_n - y[x])^{-1}h(zI_n - y[x])^{-1}h(zI_n - y[x])^{-1}dz, \end{array} \tag{5.6.3}$$

whence

$$D^2\omega(\bar{x})[h,h] = \mathrm{Tr}\left(\frac{1}{\pi i}\oint_\gamma f(z)(zI_n - y[x])^{-1}h(zI_n - y[x])^{-1}h(zI_n - y[x])^{-1}dz\right).$$

Passing to the eigenbasis of $y[\bar{x}]$, we may assume that $y[\bar{x}]$ is diagonal with positive diagonal entries $\mu_1 \le \mu_2 \le ... \le \mu_n$. In this case the formula above reads

$$D^2\omega(\bar{x})[h,h] = \frac{1}{\pi i}\sum_{p,q=1}^n \oint_\gamma h_{pq}^2 \frac{f(z)}{(z-\mu_p)^2(z-\mu_q)}dz. \tag{5.6.4}$$

Computing the residuals of the integrands at their poles, we get

$$D^2\omega(\bar{x})[h,h] = \sum_{p,q=1}^n \frac{\ln(\mu_p) - \ln(\mu_q)}{\mu_p - \mu_q}h_{pq}^2, \tag{5.6.5}$$

where, by convention, the expression $\frac{\ln(\mu_p)-\ln(\mu_q)}{\mu_p-\mu_q}$ with $\mu_p = \mu_q$ is assigned the value $\frac{1}{\mu_p}$. Since $\ln(\cdot)$ is concave, we have $\frac{\ln(\mu_p)-\ln(\mu_q)}{\mu_p-\mu_q} \ge \frac{1}{\max[\mu_p,\mu_q]}$, so that

$$D^2\omega(\bar{x})[h,h] \quad \ge \quad \sum_{p,q=1}^n \frac{1}{\max[\mu_p,\mu_q]}h_{pq}^2 = \sum_{p=1}^n \frac{1}{\mu_p}\left[h_{pp}^2 + 2\sum_{q=1}^{p-1} h_{pq}^2\right]. \tag{5.6.6}$$

It follows that

$$\begin{aligned}
\left(\sum_{p=1}^n \sqrt{h_{pp}^2 + 2\sum_{q=1}^{p-1} h_{pq}^2}\right)^2 &= \left(\sum_{p=1}^n \frac{\sqrt{h_{pp}^2+2\sum_{q=1}^{p-1} h_{pq}^2}}{\sqrt{\mu_p}}\sqrt{\mu_p}\right)^2 \\
&\le \left(\sum_{p=1}^n \frac{h_{pp}^2+2\sum_{q=1}^{p-1} h_{pq}^2}{\mu_p}\right)\left(\sum_{p=1}^n \mu_p\right) \\
&\le D^2\omega(\bar{x})[h,h]\mathrm{Tr}(y[\bar{x}]) \qquad \text{[see (5.6.6)]} \\
&\le (1+\delta)D^2\omega(\bar{x})[h,h].
\end{aligned} \tag{5.6.7}$$

Note that we have

$$h = \sum_{p=1}^n h^p, \quad (h^p)_{rs} = \begin{cases} 0, & r \ne p \ \&\ s \ne p \\ h_{rp}, & s = p \\ h_{ps}, & r = p \end{cases}.$$

Every matrix h^p is of the form $h_{pp}e_pe_p^T + r_pe_p^T + e_pr_p^T$, where $r_p = (h_{1p}, ..., h_{p-1,p}, 0, ..., 0)^T$ and e_p are the standard basic orths. From this representation it is immediately seen that

$$|h^p|_1 = \sqrt{h_{pp}^2 + 4\|r_p\|_2^2} \le \sqrt{2}\sqrt{h_{pp}^2 + 2\sum_{q=1}^{p-1} h_{pq}^2},$$

whence

$$|h|_1 \le \sum_{p=1}^n |h^p|_1 \le \sqrt{2}\sum_{p=1}^n \sqrt{h_{pp}^2 + 2\sum_{q=1}^{p-1} h_{pq}^2}.$$

Combining this relation with (5.6.7), we get

$$D^2\omega(\bar{x})[h,h] \geq \frac{1}{2(1+\delta)}|h|_1^2,$$

so that (5.6.1) is satisfied with $\alpha = 0.5(1+\delta)^{-1}$.

Now let us bound Θ. Let $x, y \in \Xi_n^+(1)$, let $\bar{x} = x + \sigma I_n$, $\bar{y} = y + \sigma I_n$, $\sigma = \delta n^{-1}$, and let ξ_p, η_p be the eigenvalues of $\bar{x}$ and $\bar{y}$, respectively. For a contour γ in the open right half-plane such that all ξ_p are inside γ we have (cf. (5.6.2) – (5.6.3)):

$$D\omega(x)[h] = \mathrm{Tr}\left(\frac{1}{2\pi i}\oint_\gamma f(z)(zI_n - \bar{x})^{-1}h(zI_n - \bar{x})^{-1}dz\right).$$

We lose nothing by assuming that $\bar{x}$ is a diagonal matrix; in this case, the latter equality implies that

$$D\omega(x)[h] = \sum_{p=1}^n \frac{1}{2\pi i}\oint_\gamma f(z)(z-\xi_p)^{-2}h_{pp}dz,$$

whence, computing the residuals of the integrands,

$$D\omega(x)[h] = \sum_p (1 + \ln(\xi_p))h_{pp}.$$

It follows that

$$\begin{aligned}
&\omega(y) - \omega(x) - D\omega(x)[y-x] \\
&\quad = \omega(y) - \sum_p \xi_p \ln\xi_p - \sum_p (1+\ln(\xi_p))(\bar{y}_{pp} - \xi_p) \\
&\quad = \omega(y) - \sum_p \bar{y}_{pp}\ln(\xi_p) + \sum_p (\xi_p - \bar{y}_{pp}) \\
&\quad \leq 1 + \delta + \omega(y) - \sum_p \bar{y}_{pp}\ln(\xi_p) \\
&\qquad\qquad [\text{since } \mathrm{Tr}(\bar{x}) \leq 1+\delta.\mathrm{Tr}(\bar{Y}) \geq 0] \\
&\quad = 1 + \delta + \sum_p \eta_p \ln(\eta_p) + \sum_p \bar{y}_{pp}\ln(1/\xi_p) \\
&\quad \leq 1 + \delta + (1+\delta)\ln(1+\delta) + \ln(1/\sigma)\sum_p \bar{y}_{pp} \\
&\qquad\qquad [\text{since } \eta \geq 0 \text{ and } \sum_p \eta_p = \mathrm{Tr}(\bar{y}) \leq 1+\delta \text{ and } 1/\xi_p \leq 1/\sigma] \\
&\quad = 1 + \delta + (1+\delta)\ln(1+\delta) + (1+\delta)\ln(n/\delta) = (1+\delta)\left[1 + \ln\left(\tfrac{n(1+\delta)}{\delta}\right)\right].
\end{aligned}$$

The resulting inequality implies (5.4.14).

Bibliography

[1] Ben-Tal, A., and Nemirovski, A., "Robust Solutions of Uncertain Linear Programs" – *OR Letters* **v.** 25 (1999), 1–13.

[2] Ben-Tal, A., Margalit, T., and Nemirovski, A., "The Ordered Subsets Mirror Descent Optimization Method with Applications to Tomography" – *SIAM Journal on Optimization* **v.** 12 (2001), 79–108.

[3] Ben-Tal, A., and Nemirovski, A., *Lectures on Modern Convex Optimization: Analysis, Algorithms, Engineering Applications* – MPS-SIAM Series on Optimization, SIAM, Philadelphia, 2001.

[4] Ben-Tal, A., Nemirovski, A., and Roos, C. (2001), "Robust solutions of uncertain quadratic and conic-quadratic problems" – to appear in *SIAM J. on Optimization*.

[5] Boyd, S., El Ghaoui, L., Feron, E., and Balakrishnan, V., *Linear Matrix Inequalities in System and Control Theory* – volume 15 of *Studies in Applied Mathematics*, SIAM, Philadelphia, 1994.

[6] Vanderberghe, L., Boyd, S., El Gamal, A.,"Optimizing dominant time constant in RC circuits", *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* **v.** 17 (1998), 110–125.

[7] Goemans, M.X., and Williamson, D.P., "Improved approximation algorithms for Maximum Cut and Satisfiability problems using semidefinite programming" – *Journal of ACM* **42** (1995), 1115–1145.

[8] Grotschel, M., Lovasz, L., and Schrijver, A., *The Ellipsoid Method and Combinatorial Optimization*, Springer, Heidelberg, 1988.

[9] Kiwiel, K., "Proximal level bundle method for convex nondifferentable optimization, saddle point problems and variational inequalities", *Mathematical Programming Series B* **v.** 69 (1995), 89-109.

[10] Lemarechal, C., Nemirovski, A., and Nesterov, Yu., "New variants of bundle methods", *Mathematical Programming Series B* **v.** 69 (1995), 111-148.

[11] Lobo, M.S., Vanderbeghe, L., Boyd, S., and Lebret, H., "Second-Order Cone Programming" – *Linear Algebra and Applications* **v.** 284 (1998), 193–228.

[12] Nemirovski, A. "Polynomial time methods in Convex Programming" – in: J. Renegar, M. Shub and S. Smale, Eds., *The Mathematics of Numerical Analysis*, 1995 AMS-SIAM Summer Seminar on Mathematics in Applied Mathematics, July 17 – August 11, 1995, Park City, Utah. – Lectures in Applied Mathematics, **v.** 32 (1996), AMS, Providence, 543–589.

[13] Nesterov, Yu., and Nemirovski, A. *Interior point polynomial time methods in Convex Programming*, SIAM, Philadelphia, 1994.

[14] Nesterov, Yu. "Squared functional systems and optimization problems", in: H. Frenk, K. Roos, T. Terlaky, S. Zhang, Eds., *High Performance Optimization*, Kluwer Academic Publishers, 2000, 405–440.

[15] Nesterov, Yu. "Semidefinite relaxation and non-convex quadratic optimization" – *Optimization Methods and Software* **v.** 12 (1997), 1–20.

[16] Nesterov, Yu. "Nonconvex quadratic optimization via conic relaxation", in: R. Saigal, H. Wolkowicz, L. Vandenberghe, Eds. *Handbook on Semidefinite Programming*, Kluwer Academis Publishers, 2000, 363-387.

[17] Roos. C., Terlaky, T., and Vial, J.-P. *Theory and Algorithms for Linear Optimization: An Interior Point Approach*, J. Wiley & Sons, 1997.

[18] Wu S.-P., Boyd, S., and Vandenberghe, L. (1997), "FIR Filter Design via Spectral Factorization and Convex Optimization" – Biswa Datta, Ed., *Applied and Computational Control, Signal and Circuits*, Birkhauser, 1997, 51–81.

[19] Ye, Y. *Interior Point Algorithms: Theory and Analysis*, J. Wiley & Sons, 1997.

Appendix A

Prerequisites from Linear Algebra and Analysis

Regarded as mathematical entities, the objective and the constraints in a Mathematical Programming problem are functions of several real variables; therefore before entering the Optimization Theory and Methods, we need to recall several basic notions and facts about the spaces $\mathbf{R}^n$ where these functions live, same as about the functions themselves. The reader is supposed to know most of the facts to follow, so he/she should not be surprised by a "cooking book" style which we intend to use in this Lecture.

A.1 Space $\mathbf{R}^n$: algebraic structure

Basically all events and constructions to be considered will take place in the *space* $\mathbf{R}^n$ *of n-dimensional real vectors*. This space can be described as follows.

A.1.1 A point in $\mathbf{R}^n$

A point in $\mathbf{R}^n$ (called also an *n-dimensional vector*) is an ordered collection $x = (x_1, ..., x_n)$ of n reals, called the *coordinates*, or *components*, or *entries* of vector x; the space $\mathbf{R}^n$ itself is the set of all collections of this type.

A.1.2 Linear operations

$\mathbf{R}^n$ is equipped with two *basic operations*:

- *Addition of vectors.* This operation takes on input two vectors $x = (x_1, ..., x_n)$ and $y = (y_1, ..., y_n)$ and produces from them a new vector

$$x + y = (x_1 + y_1, ..., x_n + y_n)$$

  with entries which are sums of the corresponding entries in x and in y.

- *Multiplication of vectors by reals.* This operation takes on input a real λ and an n-dimensional vector $x = (x_1, ..., x_n)$ and produces from them a new vector

$$\lambda x = (\lambda x_1, ..., \lambda x_n)$$

  with entries which are λ times the entries of x.

The as far as addition and multiplication by reals are concerned, the arithmetic of $\mathbf{R}^n$ inherits most of the common rules of Real Arithmetic, like $x + y = y + x$, $(x + y) + z = x + (y + z)$, $(\lambda + \mu)(x + y) = \lambda x + \mu x + \lambda y + \mu y$, $\lambda(\mu x) = (\lambda\mu)x$, etc.

A.1.3 Linear subspaces

Linear subspaces in $\mathbf{R}^n$ are, by definition, nonempty subsets of $\mathbf{R}^n$ which are closed with respect to addition of vectors and multiplication of vectors by reals:

$$L \subset \mathbf{R}^n \text{ is a linear subspace } \Leftrightarrow \left\{ \begin{array}{c} L \neq \emptyset; \\ x, y \in L \Rightarrow x + y \in L; \\ x \in L, \lambda \in \mathbf{R} \Rightarrow \lambda x \in L. \end{array} \right.$$

A.1.3.A. Examples of linear subspaces:

1. The entire $\mathbf{R}^n$;
2. The *trivial* subspace containing the single zero vector $0 = (0, ..., 0)$ [1)]; (this vector/point is called also the origin)
3. The set $\{x \in \mathbf{R}^n : x_1 = 0\}$ of all vectors x with the first coordinate equal to zero.

   The latter example admits a natural extension:
4. The set of all solutions to a *homogeneous* (i.e., with zero right hand side) system of linear equations

$$\left\{ x \in \mathbf{R}^n : \begin{array}{rcl} a_{11}x_1 + ... + a_{1n}x_n & = & 0 \\ a_{21}x_1 + ... + a_{2n}x_n & = & 0 \\ ... & & \\ a_{m1}x_1 + ... + a_{mn}x_n & = & 0 \end{array} \right\} \tag{A.1.1}$$

   always is a linear subspace in $\mathbf{R}^n$. This example is "generic", that is, *every* linear subspace in $\mathbf{R}^n$ is the solution set of a (finite) system of homogeneous linear equations, see Proposition A.3.6 below.
5. *Linear span of a set of vectors.* Given a nonempty set X of vectors, one can form a linear subspace $\mathrm{Lin}(X)$, called the *linear span* of X; this subspace consists of all vectors x which can be represented as *linear combinations* $\sum_{i=1}^{N} \lambda_i x_i$ of vectors from X (in $\sum_{i=1}^{N} \lambda_i x_i$, N is an arbitrary positive integer, λ_i are reals and x_i belong to X). Note that

   > $\mathrm{Lin}(X)$ *is the smallest linear subspace which contains* X*: if* L *is a linear subspace such that* $L \supset X$*, then* $L \supset L(X)$ (why?).

   The "linear span" example also is generic:

   > *Every linear subspace in* $\mathbf{R}^n$ *is the linear span of an appropriately chosen finite set of vectors from* $\mathbf{R}^n$*.*

   (see Theorem A.1.2.(i) below).

A.1.3.B. Sums and intersections of linear subspaces. Let $\{L_\alpha\}_{\alpha \in I}$ be a family (finite or infinite) of linear subspaces of $\mathbf{R}^n$. From this family, one can build two sets:

1. *The sum* $\sum_\alpha L_\alpha$ of the subspaces L_α which consists of all vectors which can be represented as finite sums of vectors taken each from its own subspace of the family;
2. *The intersection* $\bigcap_\alpha L_\alpha$ of the subspaces from the family.

[1)] Pay attention to the notation: we use the same symbol 0 to denote the real zero and the n-dimensional vector with all coordinates equal to zero; these two zeros are not the same, and one should understand from the context (it always is very easy) which zero is meant.

Theorem A.1.1 *Let $\{L_\alpha\}_{\alpha\in I}$ be a family of linear subspaces of $\mathbf{R}^n$. Then*

(i) *The sum $\sum\limits_\alpha L_\alpha$ of the subspaces from the family is itself a linear subspace of $\mathbf{R}^n$; it is the smallest of those subspaces of $\mathbf{R}^n$ which contain every subspace from the family;*

(ii) *The intersection $\bigcap\limits_\alpha L_\alpha$ of the subspaces from the family is itself a linear subspace of $\mathbf{R}^n$; it is the largest of those subspaces of $\mathbf{R}^n$ which are contained in every subspace from the family.*

A.1.4 Linear independence, bases, dimensions

A collection $X = \{x^1, ..., x^N\}$ of vectors from $\mathbf{R}^n$ is called *linearly independent*, if no nontrivial (i.e., with at least one nonzero coefficient) linear combination of vectors from X is zero.

> *Example of linearly independent set:* the collection of n standard basic orths $e_1 = (1, 0, ..., 0)$, $e_2 = (0, 1, 0, ..., 0)$, ..., $e_n = (0, ..., 0, 1)$.
>
> *Examples of linearly dependent sets:* (1) $X = \{0\}$; (2) $X = \{e_1, e_1\}$; (3) $X = \{e_1, e_2, e_1+e_2\}$.

A collection of vectors $f^1, ..., f^m$ is called a *basis* in $\mathbf{R}^n$, if

1. The collection is linearly independent;
2. Every vector from $\mathbf{R}^n$ is a linear combination of vectors from the collection (i.e., $\mathrm{Lin}\{f^1, ..., f^m\} = \mathbf{R}^n$).

> *Example of a basis:* The collection of standard basic orths $e_1, ..., e_n$ is a basis in $\mathbf{R}^n$.
>
> *Examples of non-bases:* (1) The collection $\{e_2, ..., e_n\}$. This collection is linearly independent, but not every vector is a linear combination of the vectors from the collection; (2) The collection $\{e_1, e_1, e_2, ..., e_n\}$. Every vector is a linear combination of vectors form the collection, but the collection is not linearly independent.

Besides the bases of the entire $\mathbf{R}^n$, one can speak about the bases of linear subspaces:

A collection $\{f^1, ..., f^m\}$ of vectors is called a *basis of a linear subspace L*, if

1. The collection is linearly independent,
2. $L = \mathrm{Lin}\{f^1, ..., f^m\}$, i.e., all vectors f^i belong to L, and every vector from L is a linear combination of the vectors $f^1, ..., f^m$.

In order to avoid trivial remarks, it makes sense to agree once for ever that

> *An empty set of vectors is linearly independent, and an empty linear combination of vectors $\sum\limits_{i\in\emptyset} \lambda_i x_i$ equals to zero.*

With this convention, the trivial linear subspace $L = \{0\}$ also has a basis, specifically, an empty set of vectors.

Theorem A.1.2 (i) *Let L be a linear subspace of $\mathbf{R}^n$. Then L admits a (finite) basis, and all bases of L are comprised of the same number of vectors; this number is called the dimension of L and is denoted by $\dim(L)$.*

> We have seen that $\mathbf{R}^n$ admits a basis comprised of n elements (the standard basic orths). From (i) it follows that *every* basis of $\mathbf{R}^n$ contains exactly n vectors, and the dimension of $\mathbf{R}^n$ is n.

(ii) *The larger is a linear subspace of $\mathbf{R}^n$, the larger is its dimension: if $L \subset L'$ are linear subspaces of $\mathbf{R}^n$, then $\dim(L) \le \dim(L')$, and the equality takes place if and only if $L = L'$.*

> We have seen that the dimension of $\mathbf{R}^n$ is n; according to the above convention, the trivial linear subspace $\{0\}$ of $\mathbf{R}^n$ admits an empty basis, so that its dimension is 0. Since $\{0\} \subset L \subset \mathbf{R}^n$ for every linear subspace L of $\mathbf{R}^n$, it follows from (ii) that the dimension of a linear subspace in $\mathbf{R}^n$ is an integer between 0 and n.

(iii) *Let L be a linear subspace in $\mathbf{R}^n$. Then*

(iii.1) *Every linearly independent subset of vectors from L can be extended to a basis of L;*

(iii.2) *From every spanning subset X for L – i.e., a set X such that $\mathrm{Lin}(X) = L$ – one can extract a basis of L.*

It follows from (iii) that

– every linearly independent subset of L contains at most $\dim(L)$ vectors, and if it contains exactly $\dim(L)$ vectors, it is a basis of L;

– every spanning set for L contains at least $\dim(L)$ vectors, and if it contains exactly $\dim(L)$ vectors, it is a basis of L.

(iv) *Let L be a linear subspace in $\mathbf{R}^n$, and $f^1, ..., f^m$ be a basis in L. Then every vector $x \in L$ admits exactly one representation*

$$x = \sum_{i=1}^{m} \lambda_i(x) f^i$$

as a linear combination of vectors from the basis, and the mapping

$$x \mapsto (\lambda_1(x), ..., \lambda_m(x)) : L \to \mathbf{R}^m$$

is a one-to-one mapping of L onto $\mathbf{R}^m$ which is linear, i.e. for every $i = 1, ..., m$ one has

$$\begin{array}{rcl} \lambda_i(x+y) & = & \lambda_i(x) + \lambda_i(y) \quad \forall (x, y \in L); \\ \lambda_i(\nu x) & = & \nu \lambda_i(x) \quad \forall (x \in L, \nu \in \mathbf{R}). \end{array} \tag{A.1.2}$$

The reals $\lambda_i(x)$, $i = 1, ..., m$, are called the coordinates of $x \in L$ in the basis $f^1, ..., f^m$.

E.g., the coordinates of a vector $x \in \mathbf{R}^n$ in the *standard basis* $e_1, ..., e_n$ of $\mathbf{R}^n$ – the one comprised of the standard basic orths – are exactly the entries of x.

(v) [Dimension formula] *Let L_1, L_2 be linear subspaces of $\mathbf{R}^n$. Then*

$$\dim(L_1 \cap L_2) + \dim(L_1 + L_2) = \dim(L_1) + \dim(L_2).$$

A.1.5 Linear mappings and matrices

A function $\mathcal{A}(x)$ (another name – *mapping*) defined on $\mathbf{R}^n$ and taking values in $\mathbf{R}^m$ is called *linear*, if it preserves linear operations:

$$\mathcal{A}(x+y) = \mathcal{A}(x) + \mathcal{A}(y) \quad \forall (x, y \in \mathbf{R}^n); \quad \mathcal{A}(\lambda x) = \lambda \mathcal{A}(x) \quad \forall (x \in \mathbf{R}^n, \lambda \in \mathbf{R}).$$

It is immediately seen that a linear mapping from $\mathbf{R}^n$ to $\mathbf{R}^m$ can be represented as multiplication by an $m \times n$ matrix:

$$\mathcal{A}(x) = Ax,$$

and this matrix is uniquely defined by the mapping: the columns A_j of A are just the images of the standard basic orths e_j under the mapping $\mathcal{A}$:

$$A_j = \mathcal{A}(e_j).$$

Linear mappings from $\mathbf{R}^n$ into $\mathbf{R}^m$ can be added to each other:

$$(\mathcal{A} + \mathcal{B})(x) = \mathcal{A}(x) + \mathcal{B}(x)$$

and multiplied by reals:

$$(\lambda \mathcal{A})(x) = \lambda \mathcal{A}(x),$$

and the results of these operations again are linear mappings from $\mathbf{R}^n$ to $\mathbf{R}^m$. The addition of linear mappings and multiplication of these mappings by reals correspond to the same operations with the

matrices representing the mappings: adding/multiplying by reals mappings, we add, respectively, multiply by reals the corresponding matrices.

Given two linear mappings $\mathcal{A}(x) : \mathbf{R}^n \to \mathbf{R}^m$ and $\mathcal{B}(y) : \mathbf{R}^m \to \mathbf{R}^k$, we can build their superposition

$$\mathcal{C}(x) \equiv \mathcal{B}(\mathcal{A}(x)) : \mathbf{R}^n \to \mathbf{R}^k,$$

which is again a linear mapping, now from $\mathbf{R}^n$ to $\mathbf{R}^k$. In the language of matrices representing the mappings, the superposition corresponds to matrix multiplication: the $k \times n$ matrix C representing the mapping $\mathcal{C}$ is the product of the matrices representing $\mathcal{A}$ and $\mathcal{B}$:

$$\mathcal{A}(x) = Ax,\ \mathcal{B}(y) = By \Rightarrow \mathcal{C}(x) \equiv \mathcal{B}(\mathcal{A}(x)) = B \cdot (Ax) = (BA)x.$$

Important convention. When speaking about adding n-dimensional vectors and multiplying them by reals, it is absolutely unimportant whether we treat the vectors as the column ones, or the row ones, or write down the entries in rectangular tables, or something else. However, when matrix operations (matrix-vector multiplication, transposition, etc.) become involved, it is important whether we treat our vectors as columns, as rows, or as something else. For the sake of definiteness, *from now on we treat all vectors as column ones*, independently of how we refer to them in the text. For example, when saying for the first time what a vector is, we wrote $x = (x_1, ..., x_n)$, which might suggest that we were speaking about row vectors. We stress that it is not the case, and the only reason for using the notation $x = (x_1, ..., x_n)$ instead of the "correct" one $x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$ is to save space and to avoid ugly formulas like $f(\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix})$ when speaking about functions with vector arguments. After we have agreed that *there is no such thing as a row vector in this Lecture course*, we can use (and do use) without any harm whatever notation we want.

Exercise A.1 *1. Mark in the list below those subsets of $\mathbf{R}^n$ which are linear subspaces, find out their dimensions and point out their bases:*

(a) $\mathbf{R}^n$

(b) $\{0\}$

(c) $\emptyset$

(d) $\{x \in \mathbf{R}^n : \sum_{i=1}^n ix_i = 0\}$

(e) $\{x \in \mathbf{R}^n : \sum_{i=1}^n ix_i^2 = 0\}$

(f) $\{x \in \mathbf{R}^n : \sum_{i=1}^n ix_i = 1\}$

(g) $\{x \in \mathbf{R}^n : \sum_{i=1}^n ix_i^2 = 1\}$

2. It is known that L is a subspace of $\mathbf{R}^n$ with exactly one basis. What is L?

3. Consider the space $\mathbf{R}^{m\times n}$ of $m \times n$ matrices with real entries. As far as linear operations – addition of matrices and multiplication of matrices by reals – are concerned, this space can be treated as certain $\mathbf{R}^N$.

(a) Find the dimension of $\mathbf{R}^{m\times n}$ and point out a basis in this space

(b) In the space $\mathbf{R}^{n\times n}$ of square $n\times n$ matrices, there are two interesting subsets: the set $\mathbf{S}^n$ of symmetric matrices $\{A=[A_{ij}] : A_{ij}=A_{ij}\}$ and the set $\mathbf{J}^n$ of skew-symmetric matrices $\{A=[A_{ij}] : A_{ij}=-A_{ji}\}$.

i. Verify that both $\mathbf{S}^n$ and $\mathbf{J}^n$ are linear subspaces of $\mathbf{R}^{n\times n}$

ii. Find the dimension and point out a basis in $\mathbf{S}^n$

iii. Find the dimension and point out a basis in $\mathbf{J}^n$

iv. What is the sum of $\mathbf{S}^n$ and $\mathbf{J}^n$? What is the intersection of $\mathbf{S}^n$ and $\mathbf{J}^n$?

A.2 Space $\mathbf{R}^n$: Euclidean structure

So far, we were interested solely in the algebraic structure of $\mathbf{R}^n$, or, which is the same, in the properties of the *linear* operations (addition of vectors and multiplication of vectors by scalars) the space is endowed with. Now let us consider another structure on $\mathbf{R}^n$ – the *standard Euclidean structure* – which allows to speak about distances, angles, convergence, etc., and thus makes the space $\mathbf{R}^n$ a much richer mathematical entity.

A.2.1 Euclidean structure

The standard Euclidean structure on $\mathbf{R}^n$ is given by the *standard inner product* – an operation which takes on input two vectors x, y and produces from them a real, specifically, the real

$$\langle x,y\rangle \equiv x^T y = \sum_{i=1}^n x_i y_i$$

The basic properties of the inner product are as follows:

1. [bi-linearity]: The real-valued function $\langle x,y\rangle$ of two vector arguments $x,y\in\mathbf{R}^n$ is linear with respect to every one of the arguments, the other argument being fixed:

$$\begin{array}{rcll}\langle \lambda u+\mu v, y\rangle &=& \lambda\langle u,y\rangle+\mu\langle v,y\rangle & \forall(u,v,y\in\mathbf{R}^n,\lambda,\mu\in\mathbf{R})\\ \langle x,\lambda u+\mu v\rangle &=& \lambda\langle x,u\rangle+\mu\langle x,v\rangle & \forall(x,u,v\in\mathbf{R}^n,\lambda,\mu\in\mathbf{R})\end{array}$$

2. [symmetry]: The function $\langle x,y\rangle$ is symmetric:

$$\langle x,y\rangle = \langle y,x\rangle \quad \forall(x,y\in\mathbf{R}^n).$$

3. [positive definiteness]: The quantity $\langle x,x\rangle$ always is nonnegative, and it is zero if and only if x is zero.

Remark A.2.1 The outlined 3 properties – bi-linearity, symmetry and positive definiteness – form a definition of an Euclidean inner product, and there are infinitely many different from each other ways to satisfy these properties; in other words, there are infinitely many different Euclidean inner products on $\mathbf{R}^n$. The standard inner product $\langle x,y\rangle = x^T y$ is just a particular case of this general notion. Although in the sequel we normally work with the standard inner product, the reader should remember that the facts we are about to recall are valid for all Euclidean inner products, and not only for the standard one.

The notion of an inner product underlies a number of purely algebraic constructions, in particular, those of *inner product representation of linear forms* and of *orthogonal complement*.

A.2.2 Inner product representation of linear forms on $\mathbf{R}^n$

A *linear form* on $\mathbf{R}^n$ is a real-valued function $f(x)$ on $\mathbf{R}^n$ which is additive ($f(x+y)=f(x)+f(y)$) and homogeneous ($f(\lambda x)=\lambda f(x)$)

Example of linear form: $f(x)=\sum_{i=1}^{n} ix_i$

Examples of non-linear functions: (1) $f(x)=x_1+1$; (2) $f(x)=x_1^2-x_2^2$; (3) $f(x)=\sin(x_1)$.

When adding/multiplying by reals linear forms, we again get linear forms (scientifically speaking: "linear forms on $\mathbf{R}^n$ form a linear space"). *Euclidean structure allows to identify linear forms on $\mathbf{R}^n$ with vectors from $\mathbf{R}^n$*:

Theorem A.2.1 *Let $\langle\cdot,\cdot\rangle$ be a Euclidean inner product on $\mathbf{R}^n$.*

(i) *Let $f(x)$ be a linear form on $\mathbf{R}^n$. Then there exists a uniquely defined vector $f\in\mathbf{R}^n$ such that the form is just the inner product with f:*

$$f(x)=\langle f,x\rangle \quad \forall x$$

(ii) *Vice versa, every vector $f\in\mathbf{R}^n$ defines, via the formula*

$$f(x)\equiv\langle f,x\rangle,$$

a linear form on $\mathbf{R}^n$;

(iii) *The above one-to-one correspondence between the linear forms and vectors on $\mathbf{R}^n$ is linear: adding linear forms (or multiplying a linear form by a real), we add (respectively, multiply by the real) the vector(s) representing the form(s).*

A.2.3 Orthogonal complement

An Euclidean structure allows to associate with a linear subspace $L\subset\mathbf{R}^n$ another linear subspace $L^\perp$ – the *orthogonal complement* (or the *annulator*) of L; by definition, $L^\perp$ consists of all vectors which are orthogonal to every vector from L:

$$L^\perp=\{f:\langle f,x\rangle=0 \quad \forall x\in L\}.$$

Theorem A.2.2 (i) *Twice taken, orthogonal complement recovers the original subspace: whenever L is a linear subspace of $\mathbf{R}^n$, one has*

$$(L^\perp)^\perp=L;$$

(ii) *The larger is a linear subspace L, the smaller is its orthogonal complement: if $L_1\subset L_2$ are linear subspaces of $\mathbf{R}^n$, then $L_1^\perp\supset L_2^\perp$*

(iii) *The intersection of a subspace and its orthogonal complement is trivial, and the sum of these subspaces is the entire $\mathbf{R}^n$:*

$$L\cap L^\perp=\{0\},\quad L+L^\perp=\mathbf{R}^n.$$

Remark A.2.2 From Theorem A.2.2.(iii) and the Dimension formula (Theorem A.1.2.(v)) it follows, first, that for every subspace L in $\mathbf{R}^n$ one has

$$\dim(L)+\dim(L^\perp)=n.$$

Second, every vector $x\in\mathbf{R}^n$ admits a unique decomposition

$$x=x_L+x_{L^\perp}$$

into a sum of two vectors: the first of them, x_L, belongs to L, and the second, $x_{L^\perp}$, belongs to $L^\perp$. This decomposition is called the *orthogonal decomposition* of x *taken with respect to* $L, L^\perp$; x_L is called the *orthogonal projection* of x onto L, and $x_{L^\perp}$ – the orthogonal projection of x onto the orthogonal complement of L. Both projections depend on x linearly, for example,

$$(x+y)_L=x_L+y_L,\quad (\lambda x)_L=\lambda x_L.$$

The mapping $x\mapsto x_L$ is called the *orthogonal projector* onto L.

A.2.4 Orthonormal bases

A collection of vectors $f^1, ..., f^m$ is called *orthonormal* w.r.t. Euclidean inner product $\langle \cdot, \cdot \rangle$, if distinct vector from the collection are orthogonal to each other:

$$i \neq j \Rightarrow \langle f^i, f^j \rangle = 0$$

and inner product of every vector f^i with itself is unit:

$$\langle f^i, f^i \rangle = 1, \ i = 1, ..., m.$$

Theorem A.2.3 (i) *An orthonormal collection $f^1, ..., f^m$ always is linearly independent and is therefore a basis of its linear span $L = \mathrm{Lin}(f^1, ..., f^m)$* (such a basis in a linear subspace is called *orthonormal*). *The coordinates of a vector $x \in L$ w.r.t. an orthonormal basis $f^1, ..., f^m$ of L are given by explicit formulas:*

$$x = \sum_{i=1}^{m} \lambda_i(x) f^i \Leftrightarrow \lambda_i(x) = \langle x, f^i \rangle.$$

Example of an orthonormal basis in $\mathbf{R}^n$: The standard basis $\{e_1, ..., e_n\}$ is orthonormal *with respect to the standard inner product* $\langle x, y \rangle = x^T y$ on $\mathbf{R}^n$ (but is not orthonormal w.r.t. other Euclidean inner products on $\mathbf{R}^n$).

Proof of (i): Taking inner product of both sides in the equality

$$x = \sum_j \lambda_j(x) f^j$$

with f^i, we get

$$\begin{array}{rcll} \langle x, f_i \rangle & = & \langle \sum\limits_j \lambda_j(x) f^j, f^i \rangle & \\ & = & \sum\limits_j \lambda_j(x) \langle f^j, f^i \rangle & \text{[bilinearity of inner product]} \\ & = & \lambda_i(x) & \text{[orthonormality of } \{f^i\}] \end{array}$$

Similar computation demonstrates that if 0 is represented as a linear combination of f^i with certain coefficients λ_i, then $\lambda_i = \langle 0, f^i \rangle = 0$, i.e., all the coefficients are zero; this means that an orthonormal system is linearly independent.

(ii) *If $f^1, ..., f^m$ is an orthonormal basis in a linear subspace L, then the inner product of two vectors $x, y \in L$ in the coordinates $\lambda_i(\cdot)$ w.r.t. this basis is given by the standard formula*

$$\langle x, y \rangle = \sum_{i=1}^{m} \lambda_i(x) \lambda_i(y).$$

Proof:

$$\begin{array}{rcll} x & = & \sum\limits_i \lambda_i(x) f^i, \ y = \sum\limits_i \lambda_i(y) f^i & \\ \Rightarrow \langle x, y \rangle & = & \langle \sum\limits_i \lambda_i(x) f^i, \sum\limits_i \lambda_i(y) f^i \rangle & \\ & = & \sum\limits_{i,j} \lambda_i(x) \lambda_j(y) \langle f^i, f^j \rangle & \text{[bilinearity of inner product]} \\ & = & \sum\limits_i \lambda_i(x) \lambda_i(y) & \text{[orthonormality of } \{f^i\}] \end{array}$$

(iii) *Every linear subspace L of $\mathbf{R}^n$ admits an orthonormal basis; moreover, every orthonormal system $f^1, ..., f^m$ of vectors from L can be extended to an orthonormal basis in L.*

Important corollary: *All Euclidean spaces of the same dimension are "the same". Specifically, if L is an m-dimensional space in a space $\mathbf{R}^n$ equipped with an Euclidean inner product $\langle\cdot,\cdot\rangle$, then there exists a one-to-one mapping $x \mapsto A(x)$ of L onto $\mathbf{R}^m$ such that*

- *The mapping preserves linear operations:*

$$A(x+y) = A(x) + A(y) \quad \forall(x, y \in L); A(\lambda x) = \lambda A(x) \quad \forall(x \in L, \lambda \in \mathbf{R});$$

- *The mapping converts the $\langle\cdot,\cdot\rangle$ inner product on L into the standard inner product on $\mathbf{R}^m$:*

$$\langle x, y\rangle = (A(x))^T A(y) \quad \forall x, y \in L.$$

Indeed, by (iii) L admits an orthonormal basis $f^1, ..., f^m$; using (ii), one can immediately check that the mapping

$$x \mapsto A(x) = (\lambda_1(x), ..., \lambda_m(x))$$

which maps $x \in L$ into the m-dimensional vector comprised of the coordinates of x in the basis $f^1, ..., f^m$, meets all the requirements.

Proof of (iii) is given by important by its own right *Gram-Schmidt orthogonalization process* as follows. We start with an arbitrary basis $h^1, ..., h^m$ in L and step by step convert it into an orthonormal basis $f^1, ..., f^m$. At the beginning of a step t of the construction, we already have an orthonormal collection $f^1, ..., f^{t-1}$ such that $\mathrm{Lin}\{f^1, ..., f^{t-1}\} = \mathrm{Lin}\{h^1, ..., h^{t-1}\}$. At a step t we

1. Build the vector
$$g^t = h^t - \sum_{j=1}^{t-1} \langle h^t, f^j\rangle f^j.$$
It is easily seen (check it!) that
   (a) One has
$$\mathrm{Lin}\{f^1, ..., f^{t-1}, g^t\} = \mathrm{Lin}\{h^1, ..., h^t\}; \tag{A.2.1}$$
   (b) $g^t \neq 0$ (derive this fact from (A.2.1) and the linear independence of the collection $h^1, ..., h^m$);
   (c) g^t is orthogonal to $f^1, ..., f^{t-1}$
2. Since $g^t \neq 0$, the quantity $\langle g^t, g^t\rangle$ is positive (positive definiteness of the inner product), so that the vector
$$f^t = \frac{1}{\sqrt{\langle g^t, g^t\rangle}} g^t$$
is well defined. It is immediately seen (check it!) that the collection $f^1, ..., f^t$ is orthonormal and
$$\mathrm{Lin}\{f^1, ..., f^t\} = \mathrm{Lin}\{f^1, ..., f^{t-1}, g^t\} = \mathrm{Lin}\{h^1, ..., h^t\}.$$
Step t of the orthogonalization process is completed.

After m steps of the optimization process, we end up with an orthonormal system $f^1, ..., f^m$ of vectors from L such that

$$\mathrm{Lin}\{f^1, ..., f^m\} = \mathrm{Lin}\{h^1, ..., h^m\} = L,$$

so that $f^1, ..., f^m$ is an orthonormal basis in L.

The construction can be easily modified (do it!) to extend a given orthonormal system of vectors from L to an orthonormal basis of L.

Exercise A.2 *1. What is the orthogonal complement (w.r.t. the standard inner product) of the subspace* $\{x \in \mathbf{R}^n : \sum_{i=1}^n x_i = 0\}$ *in* $\mathbf{R}^n$*?*

2. Find an orthonormal basis (w.r.t. the standard inner product) in the linear subspace $\{x \in \mathbf{R}^n : x_1 = 0\}$ *of* $\mathbf{R}^n$

3. Let L *be a linear subspace of* $\mathbf{R}^n$*, and* $f^1, ..., f^m$ *be an orthonormal basis in* L*. Prove that for every* $x \in \mathbf{R}^n$*, the orhoprojection* x_L *of* x *onto* L *is given by the formula*

$$x_L = \sum_{i=1}^m (x^T f^i) f^i.$$

4. Let L_1, L_2 *be linear subspaces in* $\mathbf{R}^n$*. Verify the formulas*

$$(L_1 + L_2)^\perp = L_1^\perp \cap L_2^\perp; \quad (L_1 \cap L_2)^\perp = L_1^\perp + L_2^\perp.$$

5. Consider the space of $m \times n$ *matrices* $\mathbf{R}^{m\times n}$*, and let us equip it with the "standard inner product" (called the Frobenius inner product)*

$$\langle A, B\rangle = \sum_{i,j} A_{ij} B_{ij}$$

(as if we were treating $m\times n$ *matrices as* mn*-dimensional vectors, writing the entries of the matrices column by column, and then taking the standard inner product of the resulting long vectors).*

(a) Verify that in terms of matrix multiplication the Frobenius inner product can be written as

$$\langle A, B\rangle = \mathrm{Tr}(AB^T),$$

where $\mathrm{Tr}(C)$ *is the trace (the sum of diagonal elements) of a square matrix* C*.*

(b) Build an orthonormal basis in the linear subspace $\mathbf{S}^n$ *of symmetric* $n \times n$ *matrices*

(c) What is the orthogonal complement of the subspace $\mathbf{S}^n$ *of symmetric* $n \times n$ *matrices in the space* $\mathbf{R}^{n\times n}$ *of square* $n \times n$ *matrices?*

(d) Find the orthogonal decomposition, w.r.t. $\mathbf{S}^2$*, of the matrix* $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$

A.3 Affine subspaces in $\mathbf{R}^n$

Many of events to come will take place not in the entire $\mathbf{R}^n$, but in its *affine subspaces* which, geometrically, are planes of different dimensions in $\mathbf{R}^n$. Let us become acquainted with these subspaces.

A.3.1 Affine subspaces and affine hulls

Definition of an affine subspace. Geometrically, a linear subspace L of $\mathbf{R}^n$ is a special plane – the one passing through the origin of the space (i.e., containing the zero vector). To get an arbitrary plane M, it suffices to subject an appropriate special plane L to a translation – to add to all points from L a fixed *shifting vector* a. This geometric intuition leads to the following

Definition A.3.1 [Affine subspace] *An affine subspace (a plane) in* $\mathbf{R}^n$ *is a set of the form*

$$M = a + L = \{y = a + x \mid x \in L\}, \tag{A.3.1}$$

where L *is a linear subspace in* $\mathbf{R}^n$ *and* a *is a vector from* $\mathbf{R}^n$ [2)].

[2)] according to our convention on arithmetic of sets, I was supposed to write in (A.3.1) $\{a\} + L$ instead of $a + L$ – we did not define arithmetic sum of a vector and a set. Usually people ignore this difference and omit the brackets when writing down singleton sets in similar expressions: we shall write $a + L$ instead of $\{a\} + L$, $\mathbf{R}d$ instead of $\mathbf{R}\{d\}$, etc.

E.g., shifting the linear subspace L comprised of vectors with zero first entry by a vector $a = (a_1, ..., a_n)$, we get the set $M = a + L$ of all vectors x with $x_1 = a_1$; according to our terminology, this is an affine subspace.

Immediate question about the notion of an affine subspace is: what are the "degrees of freedom" in decomposition (A.3.1) – how "strict" M determines a and L? The answer is as follows:

Proposition A.3.1 *The linear subspace L in decomposition (A.3.1) is uniquely defined by M and is the set of all differences of the vectors from M:*

$$L = M - M = \{x - y \mid x, y \in M\}. \tag{A.3.2}$$

The shifting vector a is not uniquely defined by M and can be chosen as an arbitrary vector from M.

A.3.2 Intersections of affine subspaces, affine combinations and affine hulls

An immediate conclusion of Proposition A.3.1 is as follows:

Corollary A.3.1 *Let $\{M_\alpha\}$ be an arbitrary family of affine subspaces in $\mathbf{R}^n$, and assume that the set $M = \cap_\alpha M_\alpha$ is nonempty. Then M_α is an affine subspace.*

From Corollary A.3.1 it immediately follows that for every nonempty subset Y of $\mathbf{R}^n$ there exists the smallest affine subspace containing Y – the intersection of all affine subspaces containing Y. This smallest affine subspace containing Y is called the *affine hull* of Y (notation: $\mathrm{Aff}(Y)$).

All this resembles a lot the story about linear spans. Can we further extend this analogy and to get a description of the affine hull $\mathrm{Aff}(Y)$ in terms of elements of Y similar to the one of the linear span ("linear span of X is the set of all linear combinations of vectors from X")? Sure we can!

Let us choose somehow a point $y_0 \in Y$, and consider the set

$$X = Y - y_0.$$

All affine subspaces containing Y should contain also y_0 and therefore, by Proposition A.3.1, can be represented as $M = y_0 + L$, L being a linear subspace. It is absolutely evident that an affine subspace $M = y_0 + L$ contains Y if and only if the subspace L contains X, and that the larger is L, the larger is M:

$$L \subset L' \Rightarrow M = y_0 + L \subset M' = y_0 + L'.$$

Thus, to find the smallest among *affine subspaces containing Y*, it suffices to find the smallest among the *linear subspaces containing X* and to translate the latter space by y_0:

$$\mathrm{Aff}(Y) = y_0 + \mathrm{Lin}(X) = y_0 + \mathrm{Lin}(Y - y_0). \tag{A.3.3}$$

Now, we know what is $\mathrm{Lin}(Y - y_0)$ – this is a set of all linear combinations of vectors from $Y - y_0$, so that a generic element of $\mathrm{Lin}(Y - y_0)$ is

$$x = \sum_{i=1}^{k} \mu_i (y_i - y_0) \quad [k \text{ may depend of } x]$$

with $y_i \in Y$ and real coefficients μ_i. It follows that the generic element of $\mathrm{Aff}(Y)$ is

$$y = y_0 + \sum_{i=1}^{k} \mu_i (y_i - y_0) = \sum_{i=0}^{k} \lambda_i y_i,$$

where

$$\lambda_0 = 1 - \sum_i \mu_i, \ \lambda_i = \mu_i, \ i \geq 1.$$

We see that a generic element of $\mathrm{Aff}(Y)$ is a linear combination of vectors from Y. Note, however, that the coefficients λ_i in this combination are not completely arbitrary: their sum is equal to 1. Linear combinations of this type – with the unit sum of coefficients – have a special name – they are called affine combinations.

We have seen that every vector from $\mathrm{Aff}(Y)$ is an affine combination of vectors of Y. Whether the inverse is true, i.e., whether $\mathrm{Aff}(Y)$ contains all affine combinations of vectors from Y? The answer is positive. Indeed, if

$$y = \sum_{i=1}^{k} \lambda_i y_i$$

is an affine combination of vectors from Y, then, using the equality $\sum_i \lambda_i = 1$, we can write it also as

$$y = y_0 + \sum_{i=1}^{k} \lambda_i (y_i - y_0),$$

y_0 being the "marked" vector we used in our previous reasoning, and the vector of this form, as we already know, belongs to $\mathrm{Aff}(Y)$. Thus, we come to the following

Proposition A.3.2 [Structure of affine hull]

$$\mathrm{Aff}(Y) = \{\textit{the set of all affine combinations of vectors from } Y\}.$$

When Y itself is an affine subspace, it, of course, coincides with its affine hull, and the above Proposition leads to the following

Corollary A.3.2 *An affine subspace M is closed with respect to taking affine combinations of its members – every combination of this type is a vector from M. Vice versa, a nonempty set which is closed with respect to taking affine combinations of its members is an affine subspace.*

A.3.3 Affinely spanning sets, affinely independent sets, affine dimension

Affine subspaces are closely related to linear subspaces, and the basic notions associated with linear subspaces have natural and useful affine analogies. Here we introduce these notions and discuss their basic properties.

Affinely spanning sets. Let $M = a + L$ be an affine subspace. We say that a subset Y of M is *affinely spanning* for M (we say also that Y spans M affinely, or that M is affinely spanned by Y), if $M = \mathrm{Aff}(Y)$, or, which is the same due to Proposition A.3.2, if every point of M is an affine combination of points from Y. An immediate consequence of the reasoning of the previous Section is as follows:

Proposition A.3.3 *Let $M = a + L$ be an affine subspace and Y be a subset of M, and let $y_0 \in Y$. The set Y affinely spans M – $M = \mathrm{Aff}(Y)$ – if and only if the set*

$$X = Y - y_0$$

spans the linear subspace L: $L = \mathrm{Lin}(X)$.

Affinely independent sets. A linearly independent set $x_1, ..., x_k$ is a set such that no nontrivial linear combination of $x_1, ..., x_k$ equals to zero. An equivalent definition is given by Theorem A.1.2.(iv): $x_1, ..., x_k$ are linearly independent, if the coefficients in a linear combination

$$x = \sum_{i=1}^{k} \lambda_i x_i$$

are uniquely defined by the value x of the combination. This equivalent form reflects the essence of the matter – what we indeed need, is the uniqueness of the coefficients in expansions. Accordingly, this equivalent form is the prototype for the notion of an affinely independent set: we want to introduce this notion in such a way that the coefficients λ_i in an *affine* combination

$$y = \sum_{i=0}^{k} \lambda_i y_i$$

of "affinely independent" set of vectors $y_0, ..., y_k$ would be uniquely defined by y. *Non*-uniqueness would mean that

$$\sum_{i=0}^{k} \lambda_i y_i = \sum_{i=0}^{k} \lambda'_i y_i$$

for two different collections of coefficients λ_i and λ'_i with unit sums of coefficients; if it is the case, then

$$\sum_{i=0}^{m} (\lambda_i - \lambda'_i) y_i = 0,$$

so that y_i's are linearly dependent and, moreover, there exists a nontrivial zero combination of then with *zero sum of coefficients* (since $\sum_i (\lambda_i - \lambda'_i) = \sum_i \lambda_i - \sum_i \lambda'_i = 1 - 1 = 0$). Our reasoning can be inverted – if there exists a nontrivial linear combination of y_i's with zero sum of coefficients which is zero, then the coefficients in the representation of a vector as an affine combination of y_i's are not uniquely defined. Thus, in order to get uniqueness we should for sure forbid relations

$$\sum_{i=0}^{k} \mu_i y_i = 0$$

with nontrivial zero sum coefficients μ_i. Thus, we have motivated the following

Definition A.3.2 [Affinely independent set] *A collection $y_0, ..., y_k$ of n-dimensional vectors is called affinely independent, if no nontrivial linear combination of the vectors with zero sum of coefficients is zero:*

$$\sum_{i=1}^{k} \lambda_i y_i = 0, \ \sum_{i=0}^{k} \lambda_i = 0 \Rightarrow \lambda_0 = \lambda_1 = ... = \lambda_k = 0.$$

With this definition, we get the result completely similar to the one of Theorem A.1.2.(iv):

Corollary A.3.3 *Let $y_0, ..., y_k$ be affinely independent. Then the coefficients λ_i in an affine combination*

$$y = \sum_{i=0}^{k} \lambda_i y_i \quad [\sum_i \lambda_i = 1]$$

of the vectors $y_0, ..., y_k$ are uniquely defined by the value y of the combination.

Verification of affine independence of a collection can be immediately reduced to verification of linear independence of closely related collection:

Proposition A.3.4 *$k+1$ vectors $y_0, ..., y_k$ are affinely independent if and only if the k vectors $(y_1 - y_0), (y_2 - y_0), ..., (y_k - y_0)$ are linearly independent.*

From the latter Proposition it follows, e.g., that the collection $0, e_1, ..., e_n$ comprised of the origin and the standard basic orths is affinely independent. Note that this collection is linearly dependent (as every collection containing zero). You should definitely know the difference between the two notions of independence we deal with: linear independence means that no nontrivial linear combination of the vectors can be zero, while affine independence means that no nontrivial linear combination *from certain restricted class of them* (with zero sum of coefficients) can be zero. Therefore, there are more affinely independent sets than the linearly independent ones: a linearly independent set is for sure affinely independent, but not vice versa.

Affine bases and affine dimension. Propositions A.3.2 and A.3.3 reduce the notions of affine spanning/affine independent sets to the notions of spanning/linearly independent ones. Combined with Theorem A.1.2, they result in the following analogies of the latter two statements:

Proposition A.3.5 [Affine dimension] *Let $M = a + L$ be an affine subspace in $\mathbf{R}^n$. Then the following two quantities are finite integers which are equal to each other:*

(i) *minimal # of elements in the subsets of M which affinely span M;*

(ii) *maximal # of elements in affine independent subsets of M.*

The common value of these two integers is by 1 more than the dimension $\dim L$ of L.

By definition, the *affine dimension* of an affine subspace $M = a + L$ is the dimension $\dim L$ of L. Thus, if M is of affine dimension k, then the minimal cardinality of sets affinely spanning M, same as the maximal cardinality of affine independent subsets of M, is $k + 1$.

Theorem A.3.1 [Affine bases] *Let $M = a + L$ be an affine subspace in $\mathbf{R}^n$.*

A. *Let $Y \subset M$. The following three properties of X are equivalent:*

(i) *Y is an affine independent set which affinely spans M;*

(ii) *Y is affine independent and contains $1 + \dim L$ elements;*

(iii) *Y affinely spans M and contains $1 + \dim L$ elements.*

A subset Y of M possessing the indicated equivalent to each other properties is called an affine basis of M. Affine bases in M are exactly the collections $y_0, ..., y_{\dim L}$ such that $y_0 \in M$ and $(y_1 - y_0), ..., (y_{\dim L} - y_0)$ is a basis in L.

B. *Every affinely independent collection of vectors of M either itself is an affine basis of M, or can be extended to such a basis by adding new vectors. In particular, there exists affine basis of M.*

C. *Given a set Y which affinely spans M, you can always extract from this set an affine basis of M.*

We already know that the standard basic orths $e_1, ..., e_n$ form a basis of the entire space $\mathbf{R}^n$. And what about affine bases in $\mathbf{R}^n$? According to Theorem A.3.1.**A**, you can choose as such a basis a collection $e_0, e_0 + e_1, ..., e_0 + e_n$, e_0 being an arbitrary vector.

Barycentric coordinates. Let M be an affine subspace, and let $y_0, ..., y_k$ be an affine basis of M. Since the basis, by definition, affinely spans M, every vector y from M is an affine combination of the vectors of the basis:

$$y = \sum_{i=0}^{k} \lambda_i y_i \quad [\sum_{i=0}^{k} \lambda_i = 1],$$

and since the vectors of the affine basis are affinely independent, the coefficients of this combination are uniquely defined by y (Corollary A.3.3). These coefficients are called *barycentric coordinates* of y with respect to the affine basis in question. In contrast to the usual coordinates with respect to a (linear) basis, the barycentric coordinates could not be quite arbitrary: their sum should be equal to 1.

A.3.4 Dual description of linear subspaces and affine subspaces

To the moment we have introduced the notions of linear subspace and affine subspace and have presented a scheme of generating these entities: to get, e.g., a linear subspace, you start from an arbitrary nonempty set $X \subset \mathbf{R}^n$ and add to it all linear combinations of the vectors from X. When replacing linear combinations with the affine ones, you get a way to generate affine subspaces.

The just indicated way of generating linear subspaces/affine subspaces resembles the approach of a worker building a house: he starts with the base and then adds to it new elements until the house is ready. There exists, anyhow, an approach of an artist creating a sculpture: he takes something large and then deletes extra parts of it. Is there something like "artist's way" to represent linear subspaces and affine subspaces? The answer is positive and very instructive.

Affine subspaces and systems of linear equations

Let L be a linear subspace. According to Theorem A.2.2.(i), it is an orthogonal complement – namely, the orthogonal complement to the linear subspace $L^\perp$. Now let $a_1, ..., a_m$ be a finite spanning set in $L^\perp$. A vector x which is orthogonal to $a_1, ..., a_m$ is orthogonal to the entire $L^\perp$ (since every vector from $L^\perp$ is a linear combination of $a_1, ..., a_m$ and the inner product is bilinear); and of course vice versa, a vector orthogonal to the entire $L^\perp$ is orthogonal to $a_1, ..., a_m$. We see that

$$L = (L^\perp)^\perp = \{x \mid a_i^T x = 0,\ i = 1, ..., k\}. \tag{A.3.4}$$

Thus, we get a very important, although simple,

Proposition A.3.6 ["Outer" description of a linear subspace] *Every linear subspace L in $\mathbf{R}^n$ is a set of solutions to a homogeneous linear system of equations*

$$a_i^T x = 0,\ i = 1, ..., m, \tag{A.3.5}$$

given by properly chosen m and vectors $a_1, ..., a_m$.

Proposition A.3.6 is an "if and only if" statement: as we remember from Example A.1.3.A.4, solution set to a homogeneous system of linear equations with n variables always is a linear subspace in $\mathbf{R}^n$.

From Proposition A.3.6 and the facts we know about the dimension we can easily derive several important consequences:

- Systems (A.3.5) which define a given linear subspace L are exactly the systems given by the vectors $a_1, ..., a_m$ which span $L^\perp$ [3)]
- The smallest possible number m of equations in (A.3.5) is the dimension of $L^\perp$, i.e., by Remark A.2.2, is $\operatorname{codim} L \equiv n - \dim L$ [4)]

Now, an affine subspace M is, by definition, a translation of a linear subspace: $M = a + L$. As we know, vectors x from L are exactly the solutions of certain *homogeneous* system of linear equations

$$a_i^T x = 0,\ i = 1, ..., m.$$

It is absolutely clear that adding to these vectors a fixed vector a, we get exactly the set of solutions to the *inhomogeneous* solvable linear system

$$a_i^T x = b_i \equiv a_i^T a,\ i = 1, ..., m.$$

Vice versa, the set of solutions to a *solvable* system of linear equations

$$a_i^T x = b_i,\ i = 1, ..., m,$$

with n variables is the sum of a particular solution to the system and the solution set to the corresponding homogeneous system (the latter set, as we already know, is a linear subspace in $\mathbf{R}^n$), i.e., is an affine subspace. Thus, we get the following

Proposition A.3.7 ["Outer" description of an affine subspace]
Every affine subspace $M = a + L$ in $\mathbf{R}^n$ is a set of solutions to a solvable linear system of equations

$$a_i^T x = b_i,\ i = 1, ..., m, \tag{A.3.6}$$

given by properly chosen m and vectors $a_1, ..., a_m$.

Vice versa, the set of all solutions to a solvable system of linear equations with n variables is an affine subspace in $\mathbf{R}^n$.

The linear subspace L associated with M is exactly the set of solutions of the homogeneous (with the right hand side set to 0) version of system (A.3.6).

We see, in particular, that an affine subspace always is closed.

[3)] the reasoning which led us to Proposition A.3.6 says that $[a_1, ..., a_m$ span $L^\perp] \Rightarrow$ [(A.3.5) defines L]; now we claim that the inverse also is true

[4)] to make this statement true also in the extreme case when $L = \mathbf{R}^n$ (i.e., when $\operatorname{codim} L = 0$), we from now on make a convention that an *empty* set of equations or inequalities defines, as the solution set, the entire space

Comment. The "outer" description of a linear subspace/affine subspace – the "artist's" one – is in many cases much more useful than the "inner" description via linear/affine combinations (the "worker's" one). E.g., with the outer description it is very easy to check whether a given vector belongs or does not belong to a given linear subspace/affine subspace, which is not that easy with the inner one[5)]. In fact both descriptions are "complementary" to each other and perfectly well work in parallel: what is difficult to see with one of them, is clear with another. The idea of using "inner" and "outer" descriptions of the entities we meet with – linear subspaces, affine subspaces, convex sets, optimization problems – the general idea of *duality* – is, I would say, the main driving force of Convex Analysis and Optimization, and in the sequel we would all the time meet with different implementations of this fundamental idea.

A.3.5 Structure of the simplest affine subspaces

This small subsection deals mainly with terminology. According to their dimension, affine subspaces in $\mathbf{R}^n$ are named as follows:

- Subspaces of dimension 0 are translations of the only 0-dimensional linear subspace $\{0\}$, i.e., are singleton sets – vectors from $\mathbf{R}^n$. These subspaces are called *points*; a point is a solution to a square system of linear equations with nonsingular matrix.

- Subspaces of dimension 1 (lines). These subspaces are translations of one-dimensional linear subspaces of $\mathbf{R}^n$. A one-dimensional linear subspace has a single-element basis given by a nonzero vector d and is comprised of all multiples of this vector. Consequently, line is a set of the form

  $$\{y = a + td \mid t \in \mathbf{R}\}$$

  given by a pair of vectors a (the origin of the line) and d (the direction of the line), $d \neq 0$. The origin of the line and its direction are not uniquely defined by the line; you can choose as origin any point on the line and multiply a particular direction by nonzero reals.

  In the barycentric coordinates a line is described as follows:

  $$l = \{\lambda_0 y_0 + \lambda_1 y_1 \mid \lambda_0 + \lambda_1 = 1\} = \{\lambda y_0 + (1-\lambda) y_1 \mid \lambda \in \mathbf{R}\},$$

  where y_0, y_1 is an affine basis of l; you can choose as such a basis any pair of distinct points on the line.

  The "outer" description a line is as follows: it is the set of solutions to a linear system with n variables and $n-1$ linearly independent equations.

- Subspaces of dimension > 2 and $< n-1$ have no special names; sometimes they are called affine planes of such and such dimension.

- Affine subspaces of dimension $n-1$, due to important role they play in Convex Analysis, have a special name – they are called *hyperplanes*. The outer description of a hyperplane is that a hyperplane is the solution set of a *single* linear equation

  $$a^T x = b$$

  with nontrivial left hand side ($a \neq 0$). In other words, a hyperplane is the level set $a(x) = \text{const}$ of a nonconstant linear form $a(x) = a^T x$.

- The "largest possible" affine subspace – the one of dimension n – is unique and is the entire $\mathbf{R}^n$. This subspace is given by an empty system of linear equations.

[5)] in principle it is not difficult to certify that a given point belongs to, say, a linear subspace given as the linear span of some set – it suffices to point out a representation of the point as a linear combination of vectors from the set. But how could you certify that the point does *not* belong to the subspace?

A.4 Space $\mathbf{R}^n$: metric structure and topology

Euclidean structure on the space $\mathbf{R}^n$ gives rise to a number of extremely important *metric* notions – distances, convergence, etc. For the sake of definiteness, we associate these notions with the standard inner product $x^T y$.

A.4.1 Euclidean norm and distances

By positive definiteness, the quantity $x^T x$ always is nonnegative, so that the quantity

$$|x| \equiv \|x\|_2 = \sqrt{x^T x} = \sqrt{x_1^2 + x_2^2 + ... + x_n^2}$$

is well-defined; this quantity is called the (standard) *Euclidean norm* of vector x (or simply the norm of x) and is treated as the distance from the origin to x. The distance between two arbitrary points $x, y \in \mathbf{R}^n$ is, by definition, the norm $|x - y|$ of the difference $x - y$. The notions we have defined satisfy all basic requirements on the general notions of a norm and distance, specifically:

1. *Positivity of norm: The norm of a vector always is nonnegative; it is zero if and only is the vector is zero*:
$$|x| \geq 0 \quad \forall x; \quad |x| = 0 \Leftrightarrow x = 0.$$
2. *Homogeneity of norm: When a vector is multiplied by a real, its norm is multiplied by the absolute value of the real*:
$$|\lambda x| = |\lambda| \cdot |x| \quad \forall (x \in \mathbf{R}^n, \lambda \in \mathbf{R}).$$
3. *Triangle inequality: Norm of the sum of two vectors is $\leq$ the sum of their norms*:
$$|x + y| \leq |x| + |y| \quad \forall (x, y \in \mathbf{R}^n).$$

   In contrast to the properties of positivity and homogeneity, which are absolutely evident, the Triangle inequality is not trivial and definitely requires a proof. The proof goes through a fact which is extremely important by its own right – the *Cauchy Inequality*, which perhaps is the most frequently used inequality in Mathematics:

   Theorem A.4.1 [Cauchy's Inequality] *The absolute value of the inner product of two vectors does not exceed the product of their norms:*

$$|x^T y| \leq |x||y| \quad \forall (x, y \in \mathbf{R}^n)$$

   and is equal to the product of the norms if and only if one of the vectors is proportional to the other one:

$$|x^T y| = |x||y| \Leftrightarrow \{\exists \alpha : x = \alpha y \text{ or } \exists \beta : y = \beta x\}$$

   Proof is immediate: we may assume that both x and y are nonzero (otherwise the Cauchy inequality clearly is equality, and one of the vectors is constant times (specifically, zero times) the other one, as announced in Theorem). Assuming $x, y \neq 0$, consider the function

$$f(\lambda) = (x - \lambda y)^T (x - \lambda y) = x^T x - 2\lambda x^T y + \lambda^2 y^T y.$$

   By positive definiteness of the inner product, this function – which is a second order polynomial – is nonnegative on the entire axis, whence the discriminant of the polynomial

$$(x^T y)^2 - (x^T x)(y^T y)$$

   is nonpositive:

$$(x^T y)^2 \leq (x^T x)(y^T y).$$

Taking square roots of both sides, we arrive at the Cauchy Inequality. We also see that the inequality is equality if and only if the discriminant of the second order polynomial $f(\lambda)$ is zero, i.e., if and only if the polynomial has a (multiple) real root; but due to positive definiteness of inner product, $f(\cdot)$ has a root λ if and only if $x = \lambda y$, which proves the second part of Theorem. ∎

From Cauchy's Inequality to the Triangle Inequality: Let $x, y \in \mathbf{R}^n$. Then

$$\begin{array}{rcll} |x+y|^2 & = & (x+y)^T(x+y) & \text{[definition of norm]} \\ & = & x^Tx + y^Ty + 2x^Ty & \text{[opening parentheses]} \\ & \leq & \underbrace{x^Tx}_{|x|^2} + \underbrace{y^Ty}_{|y|^2} + 2|x||y| & \text{[Cauchy's Inequality]} \\ & = & (|x| + |y|)^2 & \\ \Rightarrow |x+y| & \leq & |x| + |y| & \blacksquare \end{array}$$

The properties of norm (i.e., of the distance to the origin) we have established induce properties of the distances between pairs of arbitrary points in $\mathbf{R}^n$, specifically:

1. *Positivity of distances: The distance $|x-y|$ between two points is positive, except for the case when the points coincide ($x = y$), when the distance between x and y is zero;*

2. *Symmetry of distances: The distance from x to y is the same as the distance from y to x:*

$$|x - y| = |y - x|;$$

3. *Triangle inequality for distances:* For every three points x, y, z, the distance from x to z does not exceed the sum of distances between x and y and between y and z:

$$|z - x| \leq |y - x| + |z - y| \quad \forall (x, y, z \in \mathbf{R}^n)$$

A.4.2 Convergence

Equipped with distances, we can define the fundamental notion of *convergence of a sequence of vectors*. Specifically, we say that *a sequence $x^1, x^2, ...$ of vectors from $\mathbf{R}^n$ converges to a vector $\bar{x}$*, or, equivalently, that *$\bar{x}$ is the limit of the sequence $\{x^i\}$* (notation: $\bar{x} = \lim\limits_{i\to\infty} x^i$), if the distances from $\bar{x}$ to x^i go to 0 as $i \to \infty$:

$$\bar{x} = \lim_{i\to\infty} x^i \Leftrightarrow |\bar{x} - x^i| \to 0, i \to \infty,$$

or, which is the same, for every $\epsilon > 0$ there exists $i = i(\epsilon)$ such that the distance between every point x^i, $i \geq i(\epsilon)$, and $\bar{x}$ does not exceed ϵ:

$$\left\{|\bar{x} - x^i| \to 0, i \to \infty\right\} \Leftrightarrow \left\{\forall \epsilon > 0 \exists i(\epsilon) : i \geq i(\epsilon) \Rightarrow |\bar{x} - x^i| \leq \epsilon\right\}.$$

Exercise A.3 *Verify the following facts:*

1. *$\bar{x} = \lim\limits_{i\to\infty} x^i$ if and only if for every $j = 1, ..., n$ the coordinates # j of the vectors x^i converge, as $i \to \infty$, to the coordinate # j of the vector $\bar{x}$;*

2. *If a sequence converges, its limit is uniquely defined;*

3. *Convergence is compatible with linear operations:*
   — *if $x^i \to x$ and $y^i \to y$ as $i \to \infty$, then $x^i + y^i \to x + y$ as $i \to \infty$;*
   — *if $x^i \to x$ and $\lambda_i \to \lambda$ as $i \to \infty$, then $\lambda_i x^i \to \lambda x$ as $i \to \infty$.*

A.4.3 Closed and open sets

After we have in our disposal distance and convergence, we can speak about *closed* and *open* sets:

- A set $X \subset \mathbf{R}^n$ is called *closed*, if it contains limits of all converging sequences of elements of X:

$$\left\{x^i \in X, x = \lim_{i\to\infty} x^i\right\} \Rightarrow x \in X$$

- A set $X \subset \mathbf{R}^n$ is called *open*, if whenever x belongs to X, all points close enough to x also belong to X:

$$\forall (x \in X)\exists(\delta > 0) : |x' - x| < \delta \Rightarrow x' \in X.$$

An open set containing a point x is called a *neighbourhood* of x.

Examples of closed sets: (1) $\mathbf{R}^n$; (2) $\emptyset$; (3) the sequence $x^i = (i, 0, ..., 0)$, $i = 1, 2, 3, ...$; (4) $\{x \in \mathbf{R}^n : \sum_{i=1}^{n} a_{ij}x_j = 0, i = 1, ..., m\}$ (in other words: a linear subspace in $\mathbf{R}^n$ always is closed, see Proposition A.3.6);(5) $\{x \in \mathbf{R}^n : \sum_{i=1}^{n} a_{ij}x_j = b_i, i = 1, ..., m\}$ (in other words: an affine subset of $\mathbf{R}^n$ always is closed, see Proposition A.3.7);; (6) Any finite subset of $\mathbf{R}^n$

Examples of non-closed sets: (1) $\mathbf{R}^n \backslash \{0\}$; (2) the sequence $x^i = (1/i, 0, ..., 0)$, $i = 1, 2, 3, ...$; (3) $\{x \in \mathbf{R}^n : x_j > 0, j = 1, ..., n\}$; (4) $\{x \in \mathbf{R}^n : \sum_{i=1}^{n} x_j > 5\}$.

Examples of open sets: (1) $\mathbf{R}^n$; (2) $\emptyset$; (3) $\{x \in \mathbf{R}^n : \sum_{j=1}^{n} a_{ij}x_j > b_j, i = 1, ..., m\}$; (4) complement of a finite set.

Examples of non-open sets: (1) A nonempty finite set; (2) the sequence $x^i = (1/i, 0, ..., 0)$, $i = 1, 2, 3, ...$, and the sequence $x^i = (i, 0, 0, ..., 0)$, $i = 1, 2, 3, ...$; (3) $\{x \in \mathbf{R}^n : x_j \geq 0, j = 1, ..., n\}$; (4) $\{x \in \mathbf{R}^n : \sum_{i=1}^{n} x_j \geq 5\}$.

Exercise A.4 *Mark in the list to follows those sets which are closed and those which are open:*

1. *All vectors with integer coordinates*
2. *All vectors with rational coordinates*
3. *All vectors with positive coordinates*
4. *All vectors with nonnegative coordinates*
5. $\{x : |x| < 1\}$;
6. $\{x : |x| = 1\}$;
7. $\{x : |x| \leq 1\}$;
8. $\{x : |x| \geq 1\}$:
9. $\{x : |x| > 1\}$;
10. $\{x : 1 < |x| \leq 2\}$.

Verify the following facts

1. *A set $X \subset \mathbf{R}^n$ is closed if and only if its complement $\bar{X} = \mathbf{R}^n \backslash X$ is open;*
2. *Intersection of every family (finite or infinite) of closed sets is closed. Union of every family (finite of infinite) of open sets is open.*
3. *Union of finitely many closed sets is closed. Intersection of finitely many open sets is open.*

A.4.4 Local compactness of $\mathbf{R}^n$

A fundamental fact about convergence in $\mathbf{R}^n$, which in certain sense is characteristic for this series of spaces, is the following

Theorem A.4.2 *From every bounded sequence $\{x^i\}_{i=1}^{\infty}$ of points from $\mathbf{R}^n$ one can extract a converging subsequence $\{x^{i_j}\}_{j=1}^{\infty}$. Equivalently: A closed and bounded subset X of $\mathbf{R}^n$ is compact, i.e., a set possessing the following two equivalent to each other properties:*

(i) *From every sequence of elements of X one can extract a subsequence which converges to certain point of X;*

(ii) *From every open covering of X (i.e., a family $\{U_\alpha\}_{\alpha\in A}$ of open sets such that $X \subset \bigcup\limits_{\alpha\in A} U_\alpha$) one can extract a finite sub-covering, i.e., a finite subset of indices $\alpha_1, ..., \alpha_N$ such that $X \subset \bigcup\limits_{i=1}^{N} U_{\alpha_i}$.*

A.5 Continuous functions on $\mathbf{R}^n$

A.5.1 Continuity of a function

Let $X \subset \mathbf{R}^n$ and $f(x) : X \to \mathbf{R}^m$ be a function (another name – mapping) defined on X and taking values in $\mathbf{R}^m$.

1. f is called *continuous at a point* $\bar{x} \in X$, if for every sequence x^i of points of X converging to $\bar{x}$ the sequence $f(x^i)$ converges to $f(\bar{x})$. Equivalent definition:

   $f : X \to \mathbf{R}^m$ is continuous at $\bar{x} \in X$, if for every $\epsilon > 0$ there exists $\delta > 0$ such that

   $$x \in X, |x - \bar{x}| < \delta \Rightarrow |f(x) - f(\bar{x})| < \epsilon.$$

2. f is called *continuous on* X, if f is continuous at every point from X. Equivalent definition: f preserves convergence: whenever a sequence of points $x^i \in X$ converges to a point $x \in X$, the sequence $f(x^i)$ converges to $f(x)$.

   Examples of continuous mappings:

   1. An *affine* mapping

      $$f(x) = \begin{bmatrix} \sum\limits_{j=1}^{m} A_{1j}x_j + b_1 \\ \vdots \\ \sum\limits_{j=1}^{m} A_{mj}x_j + b_m \end{bmatrix} \equiv Ax + b : \mathbf{R}^n \to \mathbf{R}^m$$

      is continuous on the entire $\mathbf{R}^n$ (and thus – on every subset of $\mathbf{R}^n$) (check it!).
   2. The norm $|x|$ is a continuous on $\mathbf{R}^n$ (and thus – on every subset of $\mathbf{R}^n$) real-valued function (check it!).

Exercise A.5 • *Consider the function*

$$f(x_1, x_2) = \begin{cases} \frac{x_1^2 - x_2^2}{x_1^2 + x_2^2}, & (x_1, x_2) \neq 0 \\ 0, & x_1 = x_2 = 0 \end{cases} : \mathbf{R}^2 \to \mathbf{R}.$$

Check whether this function is continuous on the following sets:

1. $\mathbf{R}^2$;
2. $\mathbf{R}^2 \backslash \{0\}$;

3. $\{x \in \mathbf{R}^2 : x_1 = 0\}$;
4. $\{x \in \mathbf{R}^2 : x_2 = 0\}$;
5. $\{x \in \mathbf{R}^2 : x_1 + x_2 = 0\}$;
6. $\{x \in \mathbf{R}^2 : x_1 - x_2 = 0\}$;
7. $\{x \in \mathbf{R}^2 : |x_1 - x_2| \leq x_1^4 + x_2^4\}$;

- *Let $f : \mathbf{R}^n \to \mathbf{R}^m$ be a continuous mapping. Mark those of the following statements which always are true:*

1. *If U is an open set in $\mathbf{R}^m$, then so is the set $f^{-1}(U) = \{x : f(x) \in U\}$;*
2. *If U is an open set in $\mathbf{R}^n$, then so is the set $f(U) = \{f(x) : x \in U\}$;*
3. *If F is a closed set in $\mathbf{R}^m$, then so is the set $f^{-1}(F) = \{x : f(x) \in F\}$;*
4. *If F is an closed set in $\mathbf{R}^n$, then so is the set $f(F) = \{f(x) : x \in F\}$.*

A.5.2 Elementary continuity-preserving operations

All "elementary" operations with mappings preserve continuity. Specifically,

Theorem A.5.1 *Let X be a subset in $\mathbf{R}^n$.*

(i) [stability of continuity w.r.t. linear operations] *If $f_1(x), f_2(x)$ are continuous functions on X taking values in $\mathbf{R}^m$ and $\lambda_1(x)$, $\lambda_2(x)$ are continuous real-valued functions on X, then the function*

$$f(x) = \lambda_1(x)f_1(x) + \lambda_2(x)f_2(x) : X \to \mathbf{R}^m$$

is continuous on X;

(ii) [stability of continuity w.r.t. superposition] *Let*

- *$X \subset \mathbf{R}^n$, $Y \subset \mathbf{R}^m$;*
- *$f : X \to \mathbf{R}^m$ be a continuous mapping such that $f(x) \in Y$ for every $x \in X$;*
- *$g : Y \to \mathbf{R}^k$ be a continuous mapping.*

Then the composite mapping

$$h(x) = g(f(x)) : X \to \mathbf{R}^k$$

is continuous on X.

A.5.3 Basic properties of continuous functions on $\mathbf{R}^n$

The basic properties of continuous functions on $\mathbf{R}^n$ can be summarized as follows:

Theorem A.5.2 *Let X be a nonempty closed and bounded subset of $\mathbf{R}^n$.*

(i) *If a mapping $f : X \to \mathbf{R}^m$ is continuous on X, it is bounded on X: there exists $C < \infty$ such that $|f(x)| \leq C$ for all $x \in X$.*

> Proof. Assume, on the contrary to what should be proved, that f is unbounded, so that for every i there exists a point $x^i \in X$ such that $|f(x^i)| > i$. By Theorem A.4.2, we can extract from the sequence $\{x^i\}$ a subsequence $\{x^{i_j}\}_{j=1}^{\infty}$ which converges to a point $\bar{x} \in X$. The real-valued function $g(x) = |f(x)|$ is continuous (as the superposition of two continuous mappings, see Theorem A.5.1.(ii)) and therefore its values at the points x^{i_j} should converge, as $j \to \infty$, to its value at $\bar{x}$; on the other hand, $g(x^{i_j}) \geq i_j \to \infty$ as $j \to \infty$, and we get the desired contradiction.

(ii) *If a mapping $f : X \to \mathbf{R}^m$ is continuous on X, it is uniformly continuous: for every $\epsilon > 0$ there exists $\delta > 0$ such that*

$$x, y \in X, |x - y| < \delta \Rightarrow |f(x) - f(y)| < \epsilon.$$

Proof. Assume, on the contrary to what should be proved, that there exists $\epsilon > 0$ such that for every $\delta > 0$ one can find a pair of points x, y in X such that $|x - y| < \delta$ and $|f(x) - f(y)| \geq \epsilon$. In particular, for every $i = 1, 2, ...$ we can find two points x^i, y^i in X such that $|x^i - y^i| \leq 1/i$ and $|f(x^i) - f(y^i)| \geq \epsilon$. By Theorem A.4.2, we can extract from the sequence $\{x^i\}$ a subsequence $\{x^{i_j}\}_{j=1}^{\infty}$ which converges to certain point $\bar{x} \in X$. Since $|y^{i_j} - x^{i_j}| \leq 1/i_j \to 0$ as $j \to \infty$, the sequence $\{y^{i_j}\}_{j=1}^{\infty}$ converges to the same point $\bar{x}$ as the sequence $\{x^{i_j}\}_{j=1}^{\infty}$ (why?) Since f is continuous, we have

$$\lim_{j\to\infty} f(y^{i_j}) = f(\bar{x}) = \lim_{j\to\infty} f(x^{i_j}),$$

whence $\lim\limits_{j\to\infty} (f(x^{i_j}) - f(y^{i_j})) = 0$, which contradicts the fact that $|f(x^{i_j}) - f(y^{i_j})| \geq \epsilon > 0$ for all j.

(iii) *Let f be a real-valued continuous function on X. The f attains its minimum on X:*

$$\operatorname*{Argmin}_X f \equiv \{x \in X : f(x) = \inf_{y\in X} f(y)\} \neq \emptyset,$$

same as f attains its maximum at certain points of X:

$$\operatorname*{Argmax}_X f \equiv \{x \in X : f(x) = \sup_{y\in X} f(y)\} \neq \emptyset.$$

Proof: Let us prove that f attains its maximum on X (the proof for minimum is completely similar). Since f is bounded on X by (i), the quantity

$$f^* = \sup_{x\in X} f(x)$$

is finite; of course, we can find a sequence $\{x^i\}$ of points from X such that $f^* = \lim\limits_{i\to\infty} f(x^i)$. By Theorem A.4.2, we can extract from the sequence $\{x^i\}$ a subsequence $\{x^{i_j}\}_{j=1}^{\infty}$ which converges to certain point $\bar{x} \in X$. Since f is continuous on X, we have

$$f(\bar{x}) = \lim_{j\to\infty} f(x^{i_j}) = \lim_{i\to\infty} f(x^i) = f^*,$$

so that the maximum of f on X indeed is achieved (e.g., at the point $\bar{x}$).

Exercise A.6 *Prove that in general no one of the three statements in Theorem A.5.2 remains valid when X is closed, but not bounded, same as when X is bounded, but not closed.*

A.6 Differentiable functions on $\mathbf{R}^n$

A.6.1 The derivative

The reader definitely is familiar with the notion of derivative of a real-valued function $f(x)$ of real variable x:

$$f'(x) = \lim_{\Delta x\to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}$$

This definition does not work when we pass from functions of single real variable to functions of several real variables, or, which is the same, to functions with vector arguments. Indeed, in this case the shift in the argument Δx should be a vector, and we do not know what does it mean to *divide* by a vector...

A proper way to extend the notion of the derivative to real- and vector-valued functions of vector argument is to realize what in fact is the meaning of the derivative in the univariate case. What $f'(x)$ says to us is *how to approximate f in a neighbourhood of x by a linear function*. Specifically, *if* $f'(x)$

exists, then the linear function $f'(x)\Delta x$ of Δx approximates the change $f(x+\Delta x)-f(x)$ in f up to a remainder which is of highest order as compared with Δx as $\Delta x \to 0$:

$$|f(x+\Delta x)-f(x)-f'(x)\Delta x| \le \bar{o}(|\Delta x|) \text{ as } \Delta x \to 0.$$

In the above formula, we meet with the notation $\bar{o}(|\Delta x|)$, and here is the explanation of this notation:

> *$\bar{o}(|\Delta x|)$ is a common name of all functions $\phi(\Delta x)$ of Δx which are well-defined in a neighbourhood of the point $\Delta x = 0$ on the axis, vanish at the point $\Delta x = 0$ and are such that*
>
> $$\frac{\phi(\Delta x)}{|\Delta x|} \to 0 \text{ as } \Delta x \to 0.$$
>
> For example,
>
> 1. $(\Delta x)^2 = \bar{o}(|\Delta x|)$, $\Delta x \to 0$,
> 2. $|\Delta x|^{1.01} = \bar{o}(|\Delta x|)$, $\Delta x \to 0$,
> 3. $\sin^2(\Delta x) = \bar{o}(|\Delta x|)$, $\Delta x \to 0$,
> 4. $\Delta x \neq \bar{o}(|\Delta x|)$, $\Delta x \to 0$.
>
> Later on we shall meet with the notation "$\bar{o}(|\Delta x|^k)$ as $\Delta x \to 0$", where k is a positive integer. The definition is completely similar to the one for the case of $k = 1$:
>
> *$\bar{o}(|\Delta x|^k)$ is a common name of all functions $\phi(\Delta x)$ of Δx which are well-defined in a neighbourhood of the point $\Delta x = 0$ on the axis, vanish at the point $\Delta x = 0$ and are such that*
>
> $$\frac{\phi(\Delta x)}{|\Delta x|^k} \to 0 \text{ as } \Delta x \to 0.$$

Note that if $f(\cdot)$ is a function defined in a neighbourhood of a point x on the axis, then there perhaps are many linear functions $a\Delta x$ of Δx which well approximate $f(x+\Delta x)-f(x)$, in the sense that the remainder in the approximation

$$f(x+\Delta x)-f(x)-a\Delta x$$

tends to 0 as $\Delta x \to 0$; among these approximations, however, there exists *at most one* which approximates $f(x+\Delta x)-f(x)$ "very well" – so that the remainder is $\bar{o}(|\Delta x|)$, and not merely tends to 0 as $\Delta x \to 0$. Indeed, if

$$f(x+\Delta x)-f(x)-a\Delta x = \bar{o}(|\Delta x|),$$

then, dividing both sides by Δx, we get

$$\frac{f(x+\Delta x)-f(x)}{\Delta x} - a = \frac{\bar{o}(|\Delta x|)}{\Delta x};$$

by definition of $\bar{o}(\cdot)$, the right hand side in this equality tends to 0 as $\Delta x \to 0$, whence

$$a = \lim_{\Delta x \to 0} \frac{f(x+\Delta x)-f(x)}{\Delta x} = f'(x).$$

Thus, *if* a linear function $a\Delta x$ of Δx approximates the change $f(x+\Delta x)-f(x)$ in f up to the remainder which is $\bar{o}(|\Delta x|)$ as $\Delta x \to 0$, *then* a is the derivative of f at x. You can easily verify that the inverse statement also is true: *if* the derivative of f at x exists, *then* the linear function $f'(x)\Delta x$ of Δx approximates the change $f(x+\Delta x)-f(x)$ in f up to the remainder which is $\bar{o}(|\Delta x|)$ as $\Delta x \to 0$.

The advantage of the "$\bar{o}(|\Delta x|)$"-definition of derivative is that it can be naturally extended onto vector-valued functions of vector arguments (you should just replace "axis" with $\mathbf{R}^n$ in the definition of $\bar{o}$) and enlightens the *essence* of the notion of derivative: when it exists, this is exactly *the linear function of Δx which approximates the change $f(x+\Delta x)-f(x)$ in f up to a remainder which is $\bar{o}(|\Delta x|)$*. The precise definition is as follows:

Definition A.6.1 [Frechet differentiability] *Let f be a function which is well-defined in a neighbourhood of a point $x \in \mathbf{R}^n$ and takes values in $\mathbf{R}^m$. We say that f is differentiable at x, if there exists a linear function $Df(x)[\Delta x]$ of $\Delta x \in \mathbf{R}^n$ taking values in $\mathbf{R}^m$ which approximates the change $f(x+\Delta x)-f(x)$ in f up to a remainder which is $\bar{o}(|\Delta x|)$:*

$$|f(x+\Delta x)-f(x)-Df(x)[\Delta x]| \le \bar{o}(|\Delta x|). \tag{A.6.1}$$

Equivalently: a function f which is well-defined in a neighbourhood of a point $x \in \mathbf{R}^n$ and takes values in $\mathbf{R}^m$ is called differentiable at x, if there exists a linear function $Df(x)[\Delta x]$ of $\Delta x \in \mathbf{R}^n$ taking values in $\mathbf{R}^m$ such that for every $\epsilon > 0$ there exists $\delta > 0$ satisfying the relation

$$|\Delta x| \le \delta \Rightarrow |f(x+\Delta x)-f(x)-Df(x)[\Delta x]| \le \epsilon |\Delta x|.$$

A.6.2 Derivative and directional derivatives

We have defined what does it mean that a function $f : \mathbf{R}^n \to \mathbf{R}^m$ is differentiable at a point x, but did not say yet what is the derivative. The reader could guess that the derivative is exactly the "linear function $Df(x)[\Delta x]$ of $\Delta x \in \mathbf{R}^n$ taking values in $\mathbf{R}^m$ which approximates the change $f(x+\Delta x)-f(x)$ in f up to a remainder which is $\le \bar{o}(|\Delta x|)$" participating in the definition of differentiability. The guess is correct, but we cannot merely call the entity participating in the definition the derivative – why do we know that this entity is unique? Perhaps there are many different linear functions of Δx approximating the change in f up to a remainder which is $\bar{o}(|\Delta x|)$. In fact there is no more than a single linear function with this property due to the following observation:

> *Let f be differentiable at x, and $Df(x)[\Delta x]$ be a linear function participating in the definition of differentiability. Then*
>
> $$\forall \Delta x \in \mathbf{R}^n : \quad Df(x)[\Delta x] = \lim_{t \to +0} \frac{f(x+t\Delta x)-f(x)}{t}. \tag{A.6.2}$$
>
> *In particular, the derivative $Df(x)[\cdot]$ is uniquely defined by f and x.*
>
> **Proof.** We have
>
> $$\begin{array}{ll}
> |f(x+t\Delta x)-f(x)-Df(x)[t\Delta x]| \le \bar{o}(|t\Delta x|) & \\
> \Downarrow & \\
> |\frac{f(x+t\Delta x)-f(x)}{t} - \frac{Df(x)[t\Delta x]}{t}| \le \frac{\bar{o}(|t\Delta x|)}{t} & \\
> \Updownarrow & \\
> |\frac{f(x+t\Delta x)-f(x)}{t} - Df(x)[\Delta x]| \le \frac{\bar{o}(|t\Delta x|)}{t} & [\text{since } Df(x)[\cdot] \text{ is linear}] \\
> \Downarrow & \\
> Df(x)[\Delta x] = \lim_{t \to +0} \frac{f(x+t\Delta x)-f(x)}{t} & \left[\begin{array}{l}\text{passing to limit as } t \to +0;\\ \text{note that } \frac{\bar{o}(|t\Delta x|)}{t} \to 0, t \to +0\end{array}\right]
> \end{array}$$

Pay attention to important remarks as follows:

1. The right hand side limit in (A.6.2) is an important entity called the *directional derivative of f taken at x along (a direction) Δx*; note that this quantity is defined in the "purely univariate" fashion – by dividing the change in f by the magnitude of a shift in a direction Δx and passing to limit as the magnitude of the shift approaches 0. Relation (A.6.2) says that the derivative, if exists, is, at every Δx, nothing that the directional derivative of f taken at x along Δx. Note, however, that differentiability is much more than the existence of directional derivatives along all directions Δx; differentiability requires also *the directional derivatives to be "well-organized" – to depend linearly on the direction Δx*. It is easily seen that just existence of directional derivatives does not imply their "good organization": for example, the Euclidean norm

$$f(x) = |x|$$

at $x = 0$ possesses directional derivatives along all directions:

$$\lim_{t\to+0}\frac{f(0+t\Delta x)-f(0)}{t} = |\Delta x|;$$

these derivatives, however, depend *non-linearly* on Δx, so that the Euclidean norm is *not* differentiable at the origin (although is differentiable everywhere outside the origin, but this is another story).

2. It should be stressed that the derivative, if exists, is what it is: *a linear function of $\Delta x \in \mathbf{R}^n$ taking values in $\mathbf{R}^m$*. As we shall see in a while, we can *represent* this function by something "tractable", like a vector or a matrix, and can understand how to compute such a representation; however, an intelligent reader should bear in mind that *a* representation is not exactly the same as *the* represented entity. Sometimes the difference between derivatives and the entities which represent them is reflected in the terminology: what we call the *derivative*, is also called the *differential*, while the word "derivative" is reserved for the vector/matrix representing the differential.

A.6.3 Representations of the derivative

indexderivatives!representation ofBy definition, the derivative of a mapping $f : \mathbf{R}^n \to \mathbf{R}^m$ at a point x is a linear function $Df(x)[\Delta x]$ taking values in $\mathbf{R}^m$. How could we represent such a function?

Case of $m = 1$ – the gradient. Let us start with real-valued functions (i.e., with the case of $m = 1$); in this case the derivative is a *linear* real-valued function on $\mathbf{R}^n$. As we remember, the standard Euclidean structure on $\mathbf{R}^n$ allows to represent every linear function on $\mathbf{R}^n$ as the inner product of the argument with certain fixed vector. In particular, the derivative $Df(x)[\Delta x]$ of a scalar function can be represented as

$$Df(x)[\Delta x] = [\text{vector}]^T\Delta x;$$

what is denoted "vector" in this relation, is called the *gradient* of f at x and is denoted by $\nabla f(x)$:

$$Df(x)[\Delta x] = (\nabla f(x))^T\Delta x. \tag{A.6.3}$$

How to compute the gradient? The answer is given by (A.6.2). Indeed, let us look what (A.6.3) and (A.6.2) say when Δx is the i-th standard basic orth. According to (A.6.3), $Df(x)[e_i]$ is the i-th coordinate of the vector $\nabla f(x)$; according to (A.6.2),

$$\left.\begin{array}{rl} Df(x)[e_i] &= \lim\limits_{t\to+0}\frac{f(x+te_i)-f(x)}{t},\\ Df(x)[e_i] = -Df(x)[-e_i] &= -\lim\limits_{t\to+0}\frac{f(x-te_i)-f(x)}{t} = \lim\limits_{t\to-0}\frac{f(x+te_i)-f(x)}{t}\end{array}\right\} \Rightarrow Df(x)[e_i] = \frac{\partial f(x)}{\partial x_i}.$$

Thus,

If a real-valued function f is differentiable at x, then the first order partial derivatives of f at x exist, and the gradient of f at x is just the vector with the coordinates which are the first order partial derivatives of f taken at x:

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1}\\ \vdots \\ \frac{\partial f(x)}{\partial x_n}\end{bmatrix}.$$

The derivative of f, taken at x, is the linear function of Δx given by

$$Df(x)[\Delta x] = (\nabla f(x))^T\Delta x = \sum_{i=1}^n \frac{\partial f(x)}{\partial x_i}(\Delta x)_i.$$

General case – the Jacobian. Now let $f : \mathbf{R}^n \to \mathbf{R}^m$ with $m \geq 1$. In this case, $Df(x)[\Delta x]$, regarded as a function of Δx, is a linear mapping from $\mathbf{R}^n$ to $\mathbf{R}^m$; as we remember, the standard way to represent a linear mapping from $\mathbf{R}^n$ to $\mathbf{R}^m$ is to represent it as the multiplication by $m \times n$ matrix:

$$Df(x)[\Delta x] = [m \times n \text{ matrix}] \cdot \Delta x. \tag{A.6.4}$$

What is denoted by "matrix" in (A.6.4), is called *the Jacobian* of f at x and is denoted by $f'(x)$. How to compute the entries of the Jacobian? Here again the answer is readily given by (A.6.2). Indeed, on one hand, we have

$$Df(x)[\Delta x] = f'(x)\Delta x, \tag{A.6.5}$$

whence

$$[Df(x)[e_j]]_i = (f'(x))_{ij},\ i = 1, ..., m, j = 1, ..., n.$$

On the other hand, denoting

$$f(x) = \begin{bmatrix} f_1(x) \\ \vdots \\ f_m(x) \end{bmatrix},$$

the same computation as in the case of gradient demonstrates that

$$[Df(x)[e_j]]_i = \frac{\partial f_i(x)}{\partial x_j}$$

and we arrive at the following conclusion:

> *If a vector-valued function $f(x) = (f_1(x), ..., f_m(x))$ is differentiable at x, then the first order partial derivatives of all f_i at x exist, and the Jacobian of f at x is just the $m \times n$ matrix with the entries $[\frac{\partial f_i(x)}{\partial x_j}]_{i,j}$ (so that the rows in the Jacobian are $[\nabla f_1(x)]^T$,..., $[\nabla f_m(x)]^T$. The derivative of f, taken at x, is the linear vector-valued function of Δx given by*
>
> $$Df(x)[\Delta x] = f'(x)\Delta x = \begin{bmatrix} [\nabla f_1(x)]^T \Delta x \\ \vdots \\ [\nabla f_m(x)]^T \Delta x \end{bmatrix}.$$

Remark A.6.1 Note that for a real-valued function f we have defined both the gradient $\nabla f(x)$ and the Jacobian $f'(x)$. These two entities are "nearly the same", but not exactly the same: the Jacobian is a vector-row, and the gradient is a vector-column linked by the relation

$$f'(x) = (\nabla f(x))^T.$$

Of course, both these representations of the derivative of f yield the same linear approximation of the change in f:

$$Df(x)[\Delta x] = (\nabla f(x))^T \Delta x = f'(x)\Delta x.$$

A.6.4 Existence of the derivative

We have seen that the existence of the derivative of f at a point implies the existence of the first order partial derivatives of the (components $f_1, ..., f_m$ of) f. The inverse statement is not exactly true – the existence of all first order partial derivatives $\frac{\partial f_i(x)}{\partial x_j}$ not necessarily implies the existence of the derivative; we need a bit more:

Theorem A.6.1 [Sufficient condition for differentiability] *Assume that*

1. *The mapping $f = (f_1, ..., f_m) : \mathbf{R}^n \to \mathbf{R}^m$ is well-defined in a neighbourhood U of a point $x_0 \in \mathbf{R}^n$,*
2. *The first order partial derivatives of the components f_i of f exist everywhere in U, and*
3. *The first order partial derivatives of the components f_i of f are continuous at the point x_0.*

Then f is differentiable at the point x_0.

A.6.5 Calculus of derivatives

The calculus of derivatives is given by the following result:

Theorem A.6.2 (i) [Differentiability and linear operations] *Let $f_1(x)$, $f_2(x)$ be mappings defined in a neighbourhood of $x_0 \in \mathbf{R}^n$ and taking values in $\mathbf{R}^m$, and $\lambda_1(x), \lambda_2(x)$ be real-valued functions defined in a neighbourhood of x_0. Assume that $f_1, f_2, \lambda_1, \lambda_2$ are differentiable at x_0. Then so is the function $f(x) = \lambda_1(x)f_1(x) + \lambda_2(x)f_2(x)$, with the derivative at x_0 given by*

$$\begin{array}{rcl} Df(x_0)[\Delta x] & = & [D\lambda_1(x_0)[\Delta x]]f_1(x_0) + \lambda_1(x_0)Df_1(x_0)[\Delta x] \\ & & +[D\lambda_2(x_0)[\Delta x]]f_2(x_0) + \lambda_2(x_0)Df_2(x_0)[\Delta x] \\ & & \Downarrow \\ f'(x_0) & = & f_1(x_0)[\nabla\lambda_1(x_0)]^T + \lambda_1(x_0)f_1'(x_0) \\ & & +f_2(x_0)[\nabla\lambda_2(x_0)]^T + \lambda_2(x_0)f_2'(x_0). \end{array}$$

(ii) [chain rule] *Let a mapping $f : \mathbf{R}^n \to \mathbf{R}^m$ be differentiable at x_0, and a mapping $g : \mathbf{R}^m \to \mathbf{R}^n$ be differentiable at $y_0 = f(x_0)$. Then the superposition $h(x) = g(f(x))$ is differentiable at x_0, with the derivative at x_0 given by*

$$\begin{array}{c} Dh(x_0)[\Delta x] = Dg(y_0)[Df(x_0)[\Delta x]] \\ \Downarrow \\ h'(x_0) = g'(y_0)f'(x_0) \end{array}$$

If the outer function g is real-valued, then the latter formula implies that

$$\nabla h(x_0) = [f'(x_0)]^T \nabla g(y_0)$$

(recall that for a real-valued function ϕ, $\phi' = (\nabla\phi)^T$).

A.6.6 Computing the derivative

Representations of the derivative via first order partial derivatives normally allow to compute it by the standard Calculus rules, in a completely mechanical fashion, not thinking at all of *what* we are computing. The examples to follow (especially the third of them) demonstrate that it often makes sense to bear in mind *what* is the derivative; this sometimes yield the result much faster than blind implementing Calculus rules.

Example 1: The gradient of an affine function. An *affine* function

$$f(x) = a + \sum_{i=1}^{n} g_i x_i \equiv a + g^T x : \mathbf{R}^n \to \mathbf{R}$$

is differentiable at every point (Theorem A.6.1) and its gradient, of course, equals g:

$$\begin{array}{rcll} (\nabla f(x))^T \Delta x & = & \lim\limits_{t \to +0} t^{-1} \left[f(x + t\Delta x) - f(x) \right] & [(A.6.2)] \\ & = & \lim\limits_{t \to +0} t^{-1} [t g^T \Delta x] & [\text{arithmetics}] \end{array}$$

and we arrive at

$$\boxed{\nabla(a + g^T x) = g}$$

Example 2: The gradient of a quadratic form. For the time being, let us define a homogeneous quadratic form on $\mathbf{R}^n$ as a function

$$f(x) = \sum_{i,j} A_{ij} x_i x_j = x^T A x,$$

where A is an $n \times n$ matrix. Note that the matrices A and A^T define the same quadratic form, and therefore the *symmetric* matrix $B = \frac{1}{2}(A + A^T)$ also produces the same quadratic form as A and A^T. It follows that we always may assume (and do assume from now on) that the matrix A producing the quadratic form in question is symmetric.

A quadratic form is a simple polynomial and as such is differentiable at every point (Theorem A.6.1). What is the gradient of f at a point x? Here is the computation:

$$\begin{array}{rcll}
(\nabla f(x))^T \Delta x & = & Df(x)[\Delta x] & \\
& = & \lim\limits_{t \to +0} \left[(x + t\Delta x)^T A(x + t\Delta x) - x^T A x\right] & \\
& & & [(A.6.2)] \\
& = & \lim\limits_{t \to +0} \left[x^T A x + t(\Delta x)^T A x + t x^T A \Delta x + t^2 (\Delta x)^T A \Delta x - x^T A x\right] & \\
& & & [\text{opening parentheses}] \\
& = & \lim\limits_{t \to +0} t^{-1} \left[2t(Ax)^T \Delta x + t^2 (\Delta x)^T A \Delta x\right] & \\
& & & [\text{arithmetics + symmetry of } A] \\
& = & 2(Ax)^T \Delta x &
\end{array}$$

We conclude that

$$\boxed{\nabla(x^T A x) = 2Ax}$$

(recall that $A = A^T$).

Example 3: The derivative of the log-det barrier. Let us compute the derivative of the *log-det barrier* (playing an extremely important role in modern optimization)

$$F(X) = \ln \mathrm{Det}(X);$$

here X is an $n \times n$ matrix (or, if you prefer, n^2-dimensional vector). Note that $F(X)$ is well-defined and differentiable in a neighbourhood of every point $\bar{X}$ with positive determinant (indeed, $\mathrm{Det}(X)$ is a polynomial of the entries of X and thus – is everywhere continuous and differentiable with continuous partial derivatives, while the function $\ln(t)$ is continuous and differentiable on the positive ray; by Theorems A.5.1.(ii), A.6.2.(ii), F is differentiable at every X such that $\mathrm{Det}(X) > 0$). The reader is kindly asked to try to find the derivative of F by the standard techniques; if the result will not be obtained in, say, 30 minutes, please look at the 8-line computation to follow (in this computation, $\mathrm{Det}(\bar{X}) > 0$, and $G(X) = \mathrm{Det}(X)$):

$$\begin{array}{rcll}
 & DF(\bar{X})[\Delta X] & & \\
= & D\ln(G(\bar{X}))[DG(\bar{X})[\Delta X]] & & [\text{chain rule}] \\
= & G^{-1}(\bar{X}) DG(\bar{X})[\Delta X] & & [\ln'(t) = t^{-1}] \\
= & \mathrm{Det}^{-1}(\bar{X}) \lim\limits_{t \to +0} t^{-1} \left[\mathrm{Det}(\bar{X} + t\Delta X) - \mathrm{Det}(\bar{X})\right] & & [\text{definition of } G \text{ and (A.6.2)}] \\
= & \mathrm{Det}^{-1}(\bar{X}) \lim\limits_{t \to +0} t^{-1} \left[\mathrm{Det}(\bar{X}(I + t\bar{X}^{-1}\Delta X)) - \mathrm{Det}(\bar{X})\right] & & \\
= & \mathrm{Det}^{-1}(\bar{X}) \lim\limits_{t \to +0} t^{-1} \left[\mathrm{Det}(\bar{X})(\mathrm{Det}(I + t\bar{X}^{-1}\Delta X) - 1)\right] & & [\mathrm{Det}(AB) = \mathrm{Det}(A)\mathrm{Det}(B)] \\
= & \lim\limits_{t \to +0} t^{-1} \left[\mathrm{Det}(I + t\bar{X}^{-1}\Delta X) - 1\right] & & \\
= & \mathrm{Tr}(\bar{X}^{-1}\Delta X) = \sum\limits_{i,j} [\bar{X}^{-1}]_{ji} (\Delta X)_{ij} & &
\end{array}$$

where the concluding equality

$$\lim_{t \to +0} t^{-1}[\mathrm{Det}(I + tA) - 1] = \mathrm{Tr}(A) \equiv \sum_i A_{ii} \tag{A.6.6}$$

is immediately given by recalling what is the determinant of $I + tA$: this is a polynomial of t which is the sum of products, taken along all diagonals of a $n \times n$ matrix and assigned certain signs, of the entries of $I + tA$. At every one of these diagonals, except for the main one, there are at least two cells with the entries

proportional to t, so that the corresponding products do not contribute to the constant and the linear in t terms in $\mathrm{Det}(I+tA)$ and thus do not affect the limit in (A.6.6). The only product which does contribute to the linear and the constant terms in $\mathrm{Det}(I+tA)$ is the product $(1+tA_{11})(1+tA_{22})...(1+tA_{nn})$ coming from the main diagonal; it is clear that in this product the constant term is 1, and the linear in t term is $t(A_{11}+...+A_{nn})$, and (A.6.6) follows.

A.6.7 Higher order derivatives

Let $f:\mathbf{R}^n\to\mathbf{R}^m$ be a mapping which is well-defined and differentiable at every point x from an open set U. The Jacobian of this mapping $J(x)$ is a mapping from $\mathbf{R}^n$ to the space $\mathbf{R}^{m\times n}$ matrices, i.e., is a mapping taking values in certain $\mathbf{R}^M$ ($M=mn$). The derivative of this mapping, if it exists, is called the *second derivative* of f; it again is a mapping from $\mathbf{R}^n$ to certain $\mathbf{R}^M$ and as such can be differentiable, and so on, so that we can speak about the second, the third, ... derivatives of a vector-valued function of vector argument. A *sufficient* condition for the existence of k derivatives of f in U is that f is C^k in U, i.e., that all partial derivatives of f of orders $\le k$ exist and are continuous everywhere in U (cf. Theorem A.6.1).

We have explained what does it mean that f has k derivatives in U; note, however, that according to the definition, highest order derivatives at a point x are just long vectors; say, the second order derivative of a scalar function f of 2 variables is the Jacobian of the mapping $x\mapsto f'(x):\mathbf{R}^2\to\mathbf{R}^2$, i.e., a mapping from $\mathbf{R}^2$ to $\mathbf{R}^{2\times 2}=\mathbf{R}^4$; the third order derivative of f is therefore the Jacobian of a mapping from $\mathbf{R}^2$ to $\mathbf{R}^4$, i.e., a mapping from $\mathbf{R}^2$ to $\mathbf{R}^{4\times 2}=\mathbf{R}^8$, and so on. The question which should be addressed now is: *What is a natural and transparent way to represent the highest order derivatives?*

The answer is as follows:

> (*) *Let $f:\mathbf{R}^n\to\mathbf{R}^m$ be C^k on an open set $U\subset\mathbf{R}^n$. The derivative of order $\ell\le k$ of f, taken at a point $x\in U$, can be naturally identified with a function*
>
> $$D^\ell f(x)[\Delta x^1,\Delta x^2,...,\Delta x^\ell]$$
>
> *of ℓ vector arguments $\Delta x^i\in\mathbf{R}^n$, $i=1,...,\ell$, and taking values in $\mathbf{R}^m$. This function is linear in every one of the arguments Δx^i, the other arguments being fixed, and is symmetric with respect to permutation of arguments $\Delta x^1,...,\Delta x^\ell$.*
>
> *In terms of f, the quantity $D^\ell f(x)[\Delta x^1,\Delta x^2,...,\Delta x^\ell]$ (full name: "the ℓ-th derivative (or differential) of f taken at a point x along the directions $\Delta x^1,...,\Delta x^\ell$") is given by*
>
> $$D^\ell f(x)[\Delta x^1,\Delta x^2,...,\Delta x^\ell]=\frac{\partial^\ell}{\partial t_\ell\partial t_{\ell-1}...\partial t_1}\Big|_{t_1=...=t_\ell=0}f(x+t_1\Delta x^1+t_2\Delta x^2+...+t_\ell\Delta x^\ell). \tag{A.6.7}$$

The explanation to our claims is as follows. Let $f:\mathbf{R}^n\to\mathbf{R}^m$ be C^k on an open set $U\subset\mathbf{R}^n$.

1. When $\ell=1$, (*) says to us that the first order derivative of f, taken at x, is a linear function $Df(x)[\Delta x^1]$ of $\Delta x^1\in\mathbf{R}^n$, taking values in $\mathbf{R}^m$, and that the value of this function at every Δx^1 is given by the relation
$$Df(x)[\Delta x^1]=\frac{\partial}{\partial t_1}\Big|_{t_1=0}f(x+t_1\Delta x^1) \tag{A.6.8}$$
(cf. (A.6.2)), which is in complete accordance with what we already know about the derivative.

2. To understand what is the second derivative, let us take the first derivative $Df(x)[\Delta x^1]$, *let us temporarily fix somehow the argument* Δx^1 and treat the derivative as a function of x. As a function of x, Δx^1 being fixed, the quantity $Df(x)[\Delta x^1]$ is again a mapping which maps U into $\mathbf{R}^m$ and is differentiable by Theorem A.6.1 (provided, of course, that $k\ge 2$). The derivative of this mapping is certain linear function of $\Delta x\equiv\Delta x^2\in\mathbf{R}^n$, depending on x as on a parameter; and of course it depends on Δx^1 as on a parameter as well. Thus, the derivative of $Df(x)[\Delta x^1]$ in x is certain function
$$D^2f(x)[\Delta x^1,\Delta x^2]$$

of $x \in U$ and $\Delta x^1, \Delta x^2 \in \mathbf{R}^n$ and taking values in $\mathbf{R}^m$. What we know about this function is that it is linear in Δx^2. In fact, it is also linear in Δx^1, since it is the derivative in x of certain function (namely, of $Df(x)[\Delta x^1]$) *linearly depending on the parameter* Δx^1, so that the derivative of the function *in* x is linear in the parameter Δx^1 as well (differentiation is a linear operation with respect to a function we are differentiating: summing up functions and multiplying them by real constants, we sum up, respectively, multiply by the same constants, the derivatives). Thus, $D^2f(x)[\Delta x^1, \Delta x^2]$ is linear in Δx^1 when x and Δx^2 are fixed, and is linear in Δx^2 when x and Δx^1 are fixed. Moreover, we have

$$\begin{array}{rcll} D^2f(x)[\Delta x^1, \Delta x^2] & = & \frac{\partial}{\partial t_2}\Big|_{t_2=0} Df(x+t_2\Delta x^2)[\Delta x^1] & \text{[cf. (A.6.8)]} \\ & = & \frac{\partial}{\partial t_2}\Big|_{t_2=0}\frac{\partial}{\partial t_1}\Big|_{t_1=0} f(x+t_2\Delta x^2+t_1\Delta x^1) & \text{[by (A.6.8)]} \\ & = & \left.\frac{\partial^2}{\partial t_2 \partial t_1}\right|_{t_1=t_2=0} f(x+t_1\Delta x^1+t_2\Delta x^2) & \end{array} \tag{A.6.9}$$

as claimed in (A.6.7) for $\ell = 2$. The only piece of information about the second derivative which is contained in $(*)$ and is not justified yet is that $D^2f(x)[\Delta x^1, \Delta x^2]$ is symmetric in $\Delta x^1, \Delta x^2$; but this fact is readily given by the representation (A.6.7), since, as they prove in Calculus, if a function ϕ possesses *continuous* partial derivatives of orders $\le \ell$ in a neighbourhood of a point, then these derivatives in this neighbourhood are independent of the order in which they are taken; it follows that

$$\begin{array}{rcll} D^2f(x)[\Delta x^1, \Delta x^2] & = & \left.\frac{\partial^2}{\partial t_2 \partial t_1}\right|_{t_1=t_2=0} \underbrace{f(x+t_1\Delta x^1+t_2\Delta x^2)}_{\phi(t_1,t_2)} & \text{[(A.6.9)]} \\ & = & \left.\frac{\partial^2}{\partial t_1 \partial t_2}\right|_{t_1=t_2=0} \phi(t_1,t_2) & \\ & = & \left.\frac{\partial^2}{\partial t_1 \partial t_2}\right|_{t_1=t_2=0} f(x+t_2\Delta x^2+t_1\Delta x^1) & \\ & = & D^2f(x)[\Delta x^2, \Delta x^1] & \text{[the same (A.6.9)]} \end{array}$$

3. Now it is clear how to proceed: to define $D^3f(x)[\Delta x^1, \Delta x^2, \Delta x^3]$, we fix in the second order derivative $D^2f(x)[\Delta x^1, \Delta x^2]$ the arguments $\Delta x^1, \Delta x^2$ and treat it as a function of x only, thus arriving at a mapping which maps U into $\mathbf{R}^m$ and depends on $\Delta x^1, \Delta x^2$ as on parameters (linearly in every one of them). Differentiating the resulting mapping in x, we arrive at a function $D^3f(x)[\Delta x^1, \Delta x^2, \Delta x^3]$ which by construction is linear in every one of the arguments Δx^1, Δx^2, Δx^3 and satisfies (A.6.7); the latter relation, due to the Calculus result on the symmetry of partial derivatives, implies that $D^3f(x)[\Delta x^1, \Delta x^2, \Delta x^3]$ is symmetric in $\Delta x^1, \Delta x^2, \Delta x^3$. After we have in our disposal the third derivative D^3f, we can build from it in the already explained fashion the fourth derivative, and so on, until k-th derivative is defined.

Remark A.6.2 Since $D^\ell f(x)[\Delta x^1, ..., \Delta x^\ell]$ is linear in every one of Δx^i, we can expand the derivative in a multiple sum:

$$\begin{array}{rcl} & & \Delta x^i = \sum\limits_{j=1}^n \Delta x^i_j e_j \\ & & \Downarrow \\ D^\ell f(x)[\Delta x^1, ..., \Delta x^\ell] & = & D^\ell f(x)[\sum\limits_{j_1=1}^n \Delta x^1_{j_1} e_{j_1}, ..., \sum\limits_{j_\ell=1}^n \Delta x^\ell_{j_\ell} e_{j_\ell}] \\ & = & \sum\limits_{1\le j_1,...,j_\ell\le n} D^\ell f(x)[e_{j_1}, ..., e_{j_\ell}]\Delta x^1_{j_1}...\Delta x^\ell_{j_\ell} \end{array} \tag{A.6.10}$$

What is the origin of the coefficients $D^\ell f(x)[e_{j_1}, ..., e_{j_\ell}]$? According to (A.6.7), one has

$$\begin{array}{rcl} D^\ell f(x)[e_{j_1}, ..., e_{j_\ell}] & = & \left.\frac{\partial^\ell}{\partial t_\ell \partial t_{\ell-1}...\partial t_1}\right|_{t_1=...=t_\ell=0} f(x+t_1e_{j_1}+t_2e_{j_2}+...+t_\ell e_{j_\ell}) \\ & = & \frac{\partial^\ell}{\partial x_{j_\ell}\partial x_{j_{\ell-1}}...\partial x_{j_1}} f(x). \end{array}$$

so that the coefficients in (A.6.10) are nothing but the partial derivatives, of order ℓ, of f.

Remark A.6.3 An important particular case of relation (A.6.7) is the one when $\Delta x^1 = \Delta x^2 = ... = \Delta x^\ell$; let us call the common value of these ℓ vectors d. According to (A.6.7), we have

$$D^\ell f(x)[d,d,...,d] = \left.\frac{\partial^\ell}{\partial t_\ell \partial t_{\ell-1}...\partial t_1}\right|_{t_1=...=t_\ell=0} f(x+t_1 d + t_2 d + ... + t_\ell d).$$

This relation can be interpreted as follows: consider the function

$$\phi(t) = f(x+td)$$

of a real variable t. Then (check it!)

$$\phi^{(\ell)}(0) = \left.\frac{\partial^\ell}{\partial t_\ell \partial t_{\ell-1}...\partial t_1}\right|_{t_1=...=t_\ell=0} f(x+t_1 d + t_2 d + ... + t_\ell d) = D^\ell f(x)[d,...,d].$$

In other words, $D^\ell f(x)[d,...,d]$ is what is called *ℓ-th directional derivative of f taken at x along the direction d*; to define this quantity, we pass from function f of several variables to the univariate function $\phi(t) = f(x+td)$ – restrict f onto the line passing through x and directed by d – and then take the "usual" derivative of order ℓ of the resulting function of single real variable t at the point $t = 0$ (which corresponds to the point x of our line).

Representation of higher order derivatives. k-th order derivative $D^k f(x)[\cdot,...,\cdot]$ of a C^k function $f : \mathbf{R}^n \to \mathbf{R}m$ is what it is – it is a symmetric k-linear mapping on $\mathbf{R}^n$ taking values in $\mathbf{R}^m$ and depending on x as on a parameter. Choosing somehow coordinates in $\mathbf{R}^n$, we can represent such a mapping in the form

$$D^k f(x)[\Delta x_1,...,\Delta x_k] = \sum_{1\le i_1,...,i_k\le n} \frac{\partial^k f(x)}{\partial x_{i_k}\partial x_{i_{k-1}}...\partial x_{i_1}} (\Delta x_1)_{i_1}...(\Delta x_k)_{i_k}.$$

We may say that the derivative can be represented by k-index collection of m-dimensional vectors $\frac{\partial^k f(x)}{\partial x_{i_k}\partial x_{i_{k-1}}...\partial x_{i_1}}$. This collection, however, is a difficult-to-handle entity, so that such a representation does not help. There is, however, a case when the collection becomes an entity we know to handle; this is the case of the second-order derivative of a scalar function ($k = 2, m = 1$). In this case, the collection in question is just a symmetric matrix $H(x) = \left[\frac{\partial^2 f(x)}{\partial x_i \partial x_j}\right]_{1\le i,j\le n}$. This matrix is called the *Hessian* of f at x. Note that

$$D^2 f(x)[\Delta x_1, \Delta x_2] = \Delta x_1^T H(x) \Delta x_2.$$

A.6.8 Calculus of C^k mappings

The calculus of C^k mappings can be summarized as follows:

Theorem A.6.3 (i) *Let U be an open set in $\mathbf{R}^n$, $f_1(\cdot), f_2(\cdot) : \mathbf{R}^n \to \mathbf{R}^m$ be C^k in U, and let real-valued functions $\lambda_1(\cdot), \lambda_2(\cdot)$ be C^k in U. Then the function*

$$f(x) = \lambda_1(x) f_1(x) + \lambda_2(x) f_2(x)$$

is C^k in U.

(ii) *Let U be an open set in $\mathbf{R}^n$, V be an open set in $\mathbf{R}^m$, let a mapping $f : \mathbf{R}^n \to \mathbf{R}^m$ be C^k in U and such that $f(x) \in V$ for $x \in U$, and, finally, let a mapping $g : \mathbf{R}^m \to \mathbf{R}^p$ be C^k in V. Then the superposition*

$$h(x) = g(f(x))$$

is C^k in U.

Remark A.6.4 For higher order derivatives, in contrast to the first order ones, there is no simple "chain rule" for computing the derivative of superposition. For example, the second-order derivative of the superposition $h(x) = g(f(x))$ of two C^2-mappings is given by the formula

$$Dh(x)[\Delta x^1, \Delta x^2] = Dg(f(x))[D^2f(x)[\Delta x^1, \Delta x^2]] + D^2g(x)[Df(x)[\Delta x^1], Df(x)[\Delta x^2]]$$

(check it!). We see that both the first- and the second-order derivatives of f and g contribute to the second-order derivative of the superposition h.

The only case when there does exist a simple formula for high order derivatives of a superposition is the case when the inner function is affine: if $f(x) = Ax + b$ and $h(x) = g(f(x)) = g(Ax + b)$ with a C^ℓ mapping g, then

$$D^\ell h(x)[\Delta x^1, ..., \Delta x^\ell] = D^\ell g(Ax + b)[A\Delta x^1, ..., A\Delta x^\ell]. \tag{A.6.11}$$

A.6.9 Examples of higher-order derivatives

Example 1: Second-order derivative of an affine function $f(x) = a + b^T x$ is, of course, identically zero. Indeed, as we have seen,

$$Df(x)[\Delta x^1] = b^T \Delta x^1$$

is independent of x, and therefore the derivative of $Df(x)[\Delta x^1]$ in x, which should give us the second derivative $D^2f(x)[\Delta x^1, \Delta x^2]$, is zero. Clearly, the third, the fourth, etc., derivatives of an affine function are zero as well.

Example 2: Second-order derivative of a homogeneous quadratic form $f(x) = x^T Ax$ (A is a symmetric $n \times n$ matrix). As we have seen,

$$Df(x)[\Delta x^1] = 2x^T A\Delta x^1.$$

Differentiating in x, we get

$$D^2f(x)[\Delta x^1, \Delta x^2] = \lim_{t \to +0} t^{-1} \left[2(x + t\Delta x^2)^T A\Delta x^1 - 2x^T A\Delta x^1\right] = 2(\Delta x^2)^T A\Delta x^1,$$

so that

$$\boxed{D^2f(x)[\Delta x^1, \Delta x^2] = 2(\Delta x^2)^T A\Delta x^1}$$

Note that the second derivative of a quadratic form is independent of x; consequently, the third, the fourth, etc., derivatives of a quadratic form are identically zero.

Example 3: Second-order derivative of the log-det barrier $F(X) = \ln \mathrm{Det}(X)$. As we have seen, this function of an $n \times n$ matrix is well-defined and differentiable on the set U of matrices with positive determinant (which is an open set in the space $\mathbf{R}^{n \times n}$ of $n \times n$ matrices). In fact, this function is C^∞ in U. Let us compute its second-order derivative. As we remember,

$$DF(X)[\Delta X^1] = \mathrm{Tr}(X^{-1}\Delta X^1). \tag{A.6.12}$$

To differentiate the right hand side in X, let us first find the derivative of the mapping $G(X) = X^{-1}$ which is defined on the open set of non-degenerate $n \times n$ matrices. We have

$$\begin{array}{rcl} DG(X)[\Delta X] & = & \lim\limits_{t\to+0} t^{-1}\left[(X+t\Delta X)^{-1} - X^{-1}\right] \\ & = & \lim\limits_{t\to+0} t^{-1}\left[(X(I+tX^{-1}\Delta X))^{-1} - X^{-1}\right] \\ & = & \lim\limits_{t\to+0} t^{-1}\big[(I+t\underbrace{X^{-1}\Delta X}_{Y})^{-1}X^{-1} - X^{-1}\big] \\ & = & \left[\lim\limits_{t\to+0} t^{-1}\left[(I+tY)^{-1} - I\right]\right]X^{-1} \\ & = & \left[\lim\limits_{t\to+0} t^{-1}\left[I-(I+tY)\right](I+tY)^{-1}\right]X^{-1} \\ & = & \left[\lim\limits_{t\to+0}\left[-Y(I+tY)^{-1}\right]\right]X^{-1} \\ & = & -YX^{-1} \\ & = & -X^{-1}\Delta X X^{-1} \end{array}$$

and we arrive at the important by its own right relation

$$\boxed{D(X^{-1})[\Delta X] = -X^{-1}\Delta X X^{-1}, \quad [X \in \mathbf{R}^{n\times n}, \mathrm{Det}(X) \neq 0]}$$

which is the "matrix extension" of the standard relation $(x^{-1})' = -x^{-2}$, $x \in \mathbf{R}$.

Now we are ready to compute the second derivative of the log-det barrier:

$$\begin{array}{rcl} & & F(X) = \ln \mathrm{Det}(X) \\ & & \Downarrow \\ & & DF(X)[\Delta X^1] = \mathrm{Tr}(X^{-1}\Delta X^1) \\ & & \Downarrow \\ D^2F(X)[\Delta X^1, \Delta X^2] & = & \lim\limits_{t\to+0} t^{-1}\left[\mathrm{Tr}((X+t\Delta X^2)^{-1}\Delta X^1) - \mathrm{Tr}(X^{-1}\Delta X^1)\right] \\ & = & \lim\limits_{t\to+0} \mathrm{Tr}\left(t^{-1}(X+t\Delta X^2)^{-1}\Delta X^1 - X^{-1}\Delta X^1\right) \\ & = & \lim\limits_{t\to+0} \mathrm{Tr}\left(\left[t^{-1}(X+t\Delta X^2)^{-1} - X^{-1}\right]\Delta X^1\right) \\ & = & \mathrm{Tr}\left([-X^{-1}\Delta X^2 X^{-1}]\Delta X^1\right), \end{array}$$

and we arrive at the formula

$$\boxed{D^2F(X)[\Delta X^1, \Delta X^2] = -\mathrm{Tr}(X^{-1}\Delta X^2 X^{-1}\Delta X^1) \quad [X \in \mathbf{R}^{n\times n}, \mathrm{Det}(X) > 0]}$$

Since $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$ (check it!) for all matrices A, B such that the product AB makes sense and is square, the right hand side in the above formula is symmetric in ΔX^1, ΔX^2, as it should be for the second derivative of a C^2 function.

A.6.10 Taylor expansion

Assume that $f : \mathbf{R}^n \to \mathbf{R}^m$ is C^k in a neighbourhood U of a point $\bar{x}$. The *Taylor expansion of order k* of f, built at the point $\bar{x}$, is the function

$$\begin{array}{rcl} F_k(x) & = & f(\bar{x}) + \frac{1}{1!}Df(\bar{x})[x-\bar{x}] + \frac{1}{2!}D^2f(\bar{x})[x-\bar{x}, x-\bar{x}] \\ & & + \frac{1}{3!}D^2f(\bar{x})[x-\bar{x}, x-\bar{x}, x-\bar{x}] + ... + \frac{1}{k!}D^kf(\bar{x})[\underbrace{x-\bar{x}, ..., x-\bar{x}}_{k \text{ times}}] \end{array} \qquad (A.6.13)$$

We are already acquainted with the Taylor expansion of order 1

$$F_1(x) = f(\bar{x}) + Df(\bar{x})[x-\bar{x}]$$

– this is the affine function of x which approximates "very well" $f(x)$ in a neighbourhood of $\bar{x}$, namely, within approximation error $\bar{o}(|x-\bar{x}|)$. Similar fact is true for Taylor expansions of higher order:

Theorem A.6.4 *Let $f : \mathbf{R}^n \to \mathbf{R}^m$ be C^k in a neighbourhood of $\bar{x}$, and let $F_k(x)$ be the Taylor expansion of f at $\bar{x}$ of degree k. Then*

(i) *$F_k(x)$ is a vector-valued polynomial of full degree $\leq k$ (i.e., every one of the coordinates of the vector $F_k(x)$ is a polynomial of $x_1, ..., x_n$, and the sum of powers of x_i's in every term of this polynomial does not exceed k);*

(ii) *$F_k(x)$ approximates $f(x)$ in a neighbourhood of $\bar{x}$ up to a remainder which is $\bar{o}(|x-\bar{x}|^k)$ as $x \to \bar{x}$:*

For every $\epsilon > 0$, there exists $\delta > 0$ such that

$$|x - \bar{x}| \leq \delta \Rightarrow |F_k(x) - f(x)| \leq \epsilon |x - \bar{x}|^k.$$

$F_k(\cdot)$ is the unique polynomial with components of full degree $\leq k$ which approximates f up to a remainder which is $\bar{o}(|x - \bar{x}|^k)$.

(iii) *The value and the derivatives of F_k of orders $1, 2, ..., k$, taken at $\bar{x}$, are the same as the value and the corresponding derivatives of f taken at the same point.*

As stated in Theorem, $F_k(x)$ approximates $f(x)$ for x close to $\bar{x}$ up to a remainder which is $\bar{o}(|x - \bar{x}|^k)$. In many cases, it is not enough to know that the reminder is "$\bar{o}(|x - \bar{x}|^k)$" — we need an explicit bound on this remainder. The standard bound of this type is as follows:

Theorem A.6.5 *Let k be a positive integer, and let $f : \mathbf{R}^n \to \mathbf{R}^m$ be C^{k+1} in a ball $B_r = B_r(\bar{x}) = \{x \in \mathbf{R}^n : |x - \bar{x}| < r\}$ of a radius $r > 0$ centered at a point $\bar{x}$. Assume that the directional derivatives of order $k+1$, taken at every point of B_r along every unit direction, do not exceed certain $L < \infty$:*

$$|D^{k+1} f(x)[d, ..., d]| \leq L \quad \forall (x \in B_r) \forall (d, |d| = 1).$$

Then for the Taylor expansion F_k of order k of f taken at $\bar{x}$ one has

$$|f(x) - F_k(x)| \leq \frac{L|x - \bar{x}|^{k+1}}{(k+1)!} \quad \forall (x \in B_r).$$

Thus, in a neighbourhood of $\bar{x}$ the remainder of the k-th order Taylor expansion, taken at $\bar{x}$, is of order of $L|x - \bar{x}|^{k+1}$, where L is the maximal (over all unit directions and all points from the neighbourhood) magnitude of the directional derivatives of order $k+1$ of f.

A.7 Symmetric matrices

A.7.1 Spaces of matrices

Let $\mathbf{S}^m$ be the space of symmetric $m \times m$ matrices, and $\mathbf{M}^{m,n}$ be the space of rectangular $m \times n$ matrices with real entries. From the viewpoint of their linear structure (i.e., the operations of addition and multiplication by reals) $\mathbf{S}^m$ is just the arithmetic linear space $\mathbf{R}^{m(m+1)/2}$ of dimension $\frac{m(m+1)}{2}$: by arranging the elements of a symmetric $m \times m$ matrix X in a single column, say, in the row-by-row order, you get a usual m^2-dimensional column vector; multiplication of a matrix by a real and addition of matrices correspond to the same operations with the "representing vector(s)". When X runs through $\mathbf{S}^m$, the vector representing X runs through $m(m+1)/2$-dimensional subspace of $\mathbf{R}^{m^2}$ consisting of vectors satisfying the "symmetry condition" – the coordinates coming from symmetric to each other pairs of entries in X are equal to each other. Similarly, $\mathbf{M}^{m,n}$ as a linear space is just $\mathbf{R}^{mn}$, and it is natural to equip $\mathbf{M}^{m,n}$ with the inner product defined as the usual inner product of the vectors representing the matrices:

$$\langle X, Y \rangle = \sum_{i=1}^{m} \sum_{j=1}^{n} X_{ij} Y_{ij} = \mathrm{Tr}(X^T Y).$$

Here Tr stands for the *trace* – the sum of diagonal elements of a (square) matrix. With this inner product (called the *Frobenius inner product*), $\mathbf{M}^{m,n}$ becomes a legitimate Euclidean space, and we may use in

connection with this space all notions based upon the Euclidean structure, e.g., the (Frobenius) norm of a matrix

$$\|X\|_2 = \sqrt{\langle X, X\rangle} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}^2} = \sqrt{\mathrm{Tr}(X^T X)}$$

and likewise the notions of orthogonality, orthogonal complement of a linear subspace, etc. The same applies to the space $\mathbf{S}^m$ equipped with the Frobenius inner product; of course, the Frobenius inner product of symmetric matrices can be written without the transposition sign:

$$\langle X, Y\rangle = \mathrm{Tr}(XY),\ X, Y \in \mathbf{S}^m.$$

A.7.2 Main facts on symmetric matrices

Let us focus on the space $\mathbf{S}^m$ of symmetric matrices. The most important property of these matrices is as follows:

Theorem A.7.1 [Eigenvalue decomposition] *$n \times n$ matrix A is symmetric if and only if it admits an orthonormal system of eigenvectors: there exist orthonormal basis $\{e_1, ..., e_n\}$ such that*

$$Ae_i = \lambda_i e_i,\ i = 1, ..., n, \tag{A.7.1}$$

for reals λ_i.

In connection with Theorem A.7.1, it is worthy to recall the following notions and facts:

A.7.2.A. Eigenvectors and eigenvalues. An *eigenvector* of an $n \times n$ matrix A is a *nonzero* vector e (real or complex) such that $Ae = \lambda e$ for (real or complex) scalar λ; this scalar is called the *eigenvalue* of A *corresponding to the eigenvector* e.

Eigenvalues of A are exactly the roots of the *characteristic polynomial*

$$\pi(z) = \mathrm{Det}(zI - A) = z^n + b_1 z^{n-1} + b_2 z^{n-2} + ... + b_n$$

of A.

Theorem A.7.1 states, in particular, that for a symmetric matrix A, all eigenvalues are real, and the corresponding eigenvectors can be chosen to be real and to form an orthonormal basis in $\mathbf{R}^n$.

A.7.2.B. Eigenvalue decomposition of a symmetric matrix. Theorem A.7.1 admits equivalent reformulation as follows (check the equivalence!):

Theorem A.7.2 *An $n \times n$ matrix A is symmetric if and only if it can be represented in the form*

$$A = U\Lambda U^T, \tag{A.7.2}$$

where

- *U is an orthogonal matrix: $U^{-1} = U^T$ (or, which is the same, $U^T U = I$, or, which is the same, $UU^T = I$, or, which is the same, the columns of U form an orthonormal basis in $\mathbf{R}^n$, or, which is the same, the columns of U form an orthonormal basis in $\mathbf{R}^n$).*
- *Λ is the diagonal matrix with the diagonal entries $\lambda_1, ..., \lambda_n$.*

Representation (A.7.2) with orthogonal U and diagonal Λ is called the *eigenvalue decomposition* of A. In such a representation,

- The columns of U form an orthonormal system of eigenvectors of A;
- The diagonal entries in Λ are the eigenvalues of A corresponding to these eigenvectors.

A.7.2.C. Vector of eigenvalues. When speaking about eigenvalues $\lambda_i(A)$ of a symmetric $n \times n$ matrix A, we always arrange them in the non-ascending order:

$$\lambda_1(A) \geq \lambda_2(A) \geq ... \geq \lambda_n(A);$$

$\lambda(A) \in \mathbf{R}^n$ denotes the vector of eigenvalues of A taken in the above order.

A.7.2.D. Freedom in eigenvalue decomposition. Part of the data Λ, U in the eigenvalue decomposition (A.7.2) is uniquely defined by A, while the other data admit certain "freedom". Specifically, the sequence $\lambda_1, ..., \lambda_n$ of eigenvalues of A (i.e., diagonal entries of Λ) is exactly the sequence of roots of the characteristic polynomial of A (every root is repeated according to its multiplicity) and thus is uniquely defined by A (provided that we arrange the entries of the sequence in the non-ascending order). The columns of U are not uniquely defined by A. What is uniquely defined, are the *linear spans* $E(\lambda)$ of the columns of U corresponding to all eigenvalues equal to certain λ; such a linear span is nothing but the *spectral subspace* $\{x : Ax = \lambda x\}$ of A corresponding to the eigenvalue λ. There are as many spectral subspaces as many different eigenvalues; spectral subspaces corresponding to different eigenvalues of symmetric matrix are orthogonal to each other, and their sum is the entire space. When building an orthogonal matrix U in the spectral decomposition, one chooses an orthonormal eigenbasis in the spectral subspace corresponding to the largest eigenvalue and makes the vectors of this basis the first columns in U, then chooses an orthonormal basis in the spectral subspace corresponding to the second largest eigenvalue and makes the vector from this basis the next columns of U, and so on.

A.7.2.E. "Simultaneous" decomposition of commuting symmetric matrices. Let $A_1, ..., A_k$ be $n \times n$ symmetric matrices. It turns out that *the matrices commute with each other ($A_iA_j = A_jA_i$ for all i, j) if and only if they can be "simultaneously diagonalized", i.e., there exist a single orthogonal matrix U and diagonal matrices Λ_1,...,Λ_k such that*

$$A_i = U\Lambda_i U^T,\ i = 1, ..., k.$$

You are welcome to prove this statement by yourself; to simplify your task, here are two simple and important by their own right statements which help to reach your target:

> **A.7.2.E.1:** *Let λ be a real and A, B be two commuting $n \times n$ matrices. Then the spectral subspace $E = \{x : Ax = \lambda x\}$ of A corresponding to λ is invariant for B (i.e., $Be \in E$ for every $e \in E$).*
>
> **A.7.2.E.2:** *If A is an $n \times n$ matrix and L is an invariant subspace of A (i.e., L is a linear subspace such that $Ae \in L$ whenever $e \in L$), then the orthogonal complement $L^\perp$ of L is invariant for the matrix A^T. In particular, if A is symmetric and L is invariant subspace of A, then $L^\perp$ is invariant subspace of A as well.*

A.7.3 Variational characterization of eigenvalues

Theorem A.7.3 [VCE – Variational Characterization of Eigenvalues] *Let A be a symmetric matrix. Then*

$$\lambda_\ell(A) = \min_{E \in \mathcal{E}_\ell} \max_{x \in E, x^Tx=1} x^TAx,\ \ell = 1, ..., n, \tag{A.7.3}$$

where $\mathcal{E}_\ell$ is the family of all linear subspaces in $\mathbf{R}^n$ of the dimension $n - \ell + 1$.

VCE says that to get the largest eigenvalue $\lambda_1(A)$, you should maximize the quadratic form x^TAx over the unit sphere $S = \{x \in \mathbf{R}^n : x^Tx = 1\}$; the maximum is exactly $\lambda_1(A)$. To get the second largest eigenvalue $\lambda_2(A)$, you should act as follows: you choose a linear subspace E of dimension $n - 1$ and maximize the quadratic form x^TAx over the cross-section of S by this subspace; the maximum value of the form depends on E, and you minimize this maximum over linear subspaces E of the dimension

$n-1$; the result is exactly $\lambda_2(A)$. To get $\lambda_3(A)$, you replace in the latter construction subspaces of the dimension $n-1$ by those of the dimension $n-2$, and so on. In particular, the smallest eigenvalue $\lambda_n(A)$ is just the minimum, over all linear subspaces E of the dimension $n-n+1=1$, i.e., over all lines passing through the origin, of the quantities x^TAx, where $x\in E$ is unit ($x^Tx=1$); in other words, $\lambda_n(A)$ is just the minimum of the quadratic form x^TAx over the unit sphere S.

Proof of the VCE is pretty easy. Let $e_1,...,e_n$ be an orthonormal eigenbasis of A: $Ae_\ell=\lambda_\ell(A)e_\ell$. For $1\le\ell\le n$, let $F_\ell=\mathrm{Lin}\{e_1,...,e_\ell\}$, $G_\ell=\mathrm{Lin}\{e_\ell,e_{\ell+1},...,e_n\}$. Finally, for $x\in\mathbf{R}^n$ let $\xi(x)$ be the vector of coordinates of x in the orthonormal basis $e_1,...,e_n$. Note that

$$x^Tx=\xi^T(x)\xi(x),$$

since $\{e_1,...,e_n\}$ is an orthonormal basis, and that

$$\begin{aligned} x^TAx &= x^TA\sum_i\xi_i(x)e_i)=x^T\sum_i\lambda_i(A)\xi_i(x)e_i= \\ \sum_i\lambda_i(A)\xi_i(x)&\underbrace{(x^Te_i)}_{\xi_i(x)} \\ &= \sum_i\lambda_i(A)\xi_i^2(x). \end{aligned} \tag{A.7.4}$$

Now, given ℓ, $1\le\ell\le n$, let us set $E=G_\ell$; note that E is a linear subspace of the dimension $n-\ell+1$. In view of (A.7.4), the maximum of the quadratic form x^TAx over the intersection of our E with the unit sphere is

$$\max\left\{\sum_{i=\ell}^n\lambda_i(A)\xi_i^2:\sum_{i=\ell}^n\xi_i^2=1\right\},$$

and the latter quantity clearly equals to $\max\limits_{\ell\le i\le n}\lambda_i(A)=\lambda_\ell(A)$. Thus, for appropriately chosen $E\in\mathcal{E}_\ell$, the inner maximum in the right hand side of (A.7.3) equals to $\lambda_\ell(A)$, whence the right hand side of (A.7.3) is $\le\lambda_\ell(A)$. It remains to prove the opposite inequality. To this end, consider a linear subspace E of the dimension $n-\ell+1$ and observe that it has nontrivial intersection with the linear subspace F_ℓ of the dimension ℓ (indeed, $\dim E+\dim F_\ell=(n-\ell+1)+\ell>n$, so that $\dim(E\cap F)>0$ by the Dimension formula). It follows that there exists a unit vector y belonging to both E and F_ℓ. Since y is a unit vector from F_ℓ, we have $y=\sum\limits_{i=1}^\ell\eta_ie_i$ with $\sum\limits_{i=1}^\ell\eta_i^2=1$, whence, by (A.7.4),

$$y^TAy=\sum_{i=1}^\ell\lambda_i(A)\eta_i^2\ge\min_{1\le i\le\ell}\lambda_i(A)=\lambda_\ell(A).$$

Since y is in E, we conclude that

$$\max_{x\in E:x^Tx=1}x^TAx\ge y^TAy\ge\lambda_\ell(A).$$

Since E is an arbitrary subspace form $\mathcal{E}_\ell$, we conclude that the right hand side in (A.7.3) is $\ge\lambda_\ell(A)$. ∎

A simple and useful byproduct of our reasoning is the relation (A.7.4):

Corollary A.7.1 *For a symmetric matrix A, the quadratic form x^TAx is weighted sum of squares of the coordinates $\xi_i(x)$ of x taken with respect to an orthonormal eigenbasis of A; the weights in this sum are exactly the eigenvalues of A:*

$$x^TAx=\sum_i\lambda_i(A)\xi_i^2(x).$$

Corollaries of the VCE

VCE admits a number of extremely important corollaries as follows:

A.7.3.A. Eigenvalue characterization of positive (semi)definite matrices. Recall that a matrix A is called positive definite (notation: $A \succ 0$), if it is symmetric and the quadratic form x^TAx is positive outside the origin; A is called positive semidefinite (notation: $A \succeq 0$), if A is symmetric and the quadratic form x^TAx is nonnegative everywhere. VCE provides us with the following eigenvalue characterization of positive (semi)definite matrices:

Proposition A.7.1 *: A symmetric matrix A is positive semidefinite if and only if its eigenvalues are nonnegative; A is positive definite if and only if all eigenvalues of A are positive*

Indeed, A is positive definite, if and only if the minimum value of x^TAx over the unit sphere is positive, and is positive semidefinite, if and only if this minimum value is nonnegative; it remains to note that by VCE, the minimum value of x^TAx over the unit sphere is exactly the minimum eigenvalue of A.

A.7.3.B. $\succeq$-Monotonicity of the vector of eigenvalues. Let us write $A \succeq B$ ($A \succ B$) to express that A, B are symmetric matrices of the same size such that $A - B$ is positive semidefinite (respectively, positive definite).

Proposition A.7.2 *If $A \succeq B$, then $\lambda(A) \geq \lambda(B)$, and if $A \succ B$, then $\lambda(A) > \lambda(B)$.*

Indeed, when $A \succeq B$, then, of course,

$$\max_{x \in E: x^Tx=1} x^TAx \geq \max_{x \in E: x^Tx=1} x^TBx$$

for every linear subspace E, whence

$$\lambda_\ell(A) = \min_{E \in \mathcal{E}_\ell} \max_{x \in E: x^Tx=1} x^TAx \geq \min_{E \in \mathcal{E}_\ell} \max_{x \in E: x^Tx=1} x^TBx = \lambda_\ell(B),\ \ell = 1, ..., n,$$

i.e., $\lambda(A) \geq \lambda(B)$. The case of $A \succ B$ can be considered similarly.

A.7.3.C. Eigenvalue Interlacement Theorem. We shall formulate this extremely important theorem as follows:

Theorem A.7.4 [Eigenvalue Interlacement Theorem] *Let A be a symmetric $n \times n$ matrix and $\bar{A}$ be the angular $(n-k) \times (n-k)$ submatrix of A. Then, for every $\ell \leq n-k$, the ℓ-th eigenvalue of $\bar{A}$ separates the ℓ-th and the $(\ell+k)$-th eigenvalues of A:*

$$\lambda_\ell(A) \succeq \lambda_\ell(\bar{A}) \succeq \lambda_{\ell+k}(A). \tag{A.7.5}$$

Indeed, by VCE, $\lambda_\ell(\bar{A}) = \min_{E \in \bar{\mathcal{E}}_\ell} \max_{x \in E: x^Tx=1} x^TAx$, where $\bar{\mathcal{E}}_\ell$ is the family of all linear subspaces of the dimension $n-k-\ell+1$ contained in the linear subspace $\{x \in \mathbf{R}^n : x_{n-k+1} = x_{n-k+2} = ... = x_n = 0\}$. Since $\bar{\mathcal{E}}_\ell \subset \mathcal{E}_{\ell+k}$, we have

$$\lambda_\ell(\bar{A}) = \min_{E \in \bar{\mathcal{E}}_\ell} \max_{x \in E: x^Tx=1} x^TAx \geq \min_{E \in \mathcal{E}_{\ell+k}} \max_{x \in E: x^Tx=1} x^TAx = \lambda_{\ell+k}(A).$$

We have proved the left inequality in (A.7.5). Applying this inequality to the matrix $-A$, we get

$$-\lambda_\ell(\bar{A}) = \lambda_{n-k-\ell}(-\bar{A}) \geq \lambda_{n-\ell}(-A) = -\lambda_\ell(A),$$

or, which is the same, $\lambda_\ell(\bar{A}) \leq \lambda_\ell(A)$, which is the first inequality in (A.7.5).

A.7.4 Positive semidefinite matrices and the semidefinite cone

A.7.4.A. Positive semidefinite matrices. Recall that an $n \times n$ matrix A is called *positive semidefinite* (notation: $A \succeq 0$), if A is symmetric and produces nonnegative quadratic form:

$$A \succeq 0 \Leftrightarrow \{A = A^T \quad \text{and} \quad x^T Ax \geq 0 \quad \forall x\}.$$

A is called positive *definite* (notation: $A \succ 0$), if it is positive semidefinite and the corresponding quadratic form is positive outside the origin:

$$A \succ 0 \Leftrightarrow \{A = A^T \quad \text{and} \quad x^T Ax > 00 \quad \forall x \neq 0\}.$$

It makes sense to list a number of equivalent definitions of a positive semidefinite matrix:

Theorem A.7.5 *Let A be a symmetric $n \times n$ matrix. Then the following properties of A are equivalent to each other:*

(i) $A \succeq 0$
(ii) $\lambda(A) \geq 0$
(iii) $A = D^T D$ *for certain rectangular matrix* D
(iv) $A = \Delta^T \Delta$ *for certain upper triangular* $n \times n$ *matrix* Δ
(v) $A = B^2$ *for certain symmetric matrix* B;
(vi) $A = B^2$ *for certain* $B \succeq 0$.

The following properties of a symmetric matrix A also are equivalent to each other:

(i′) $A \succ 0$
(ii′) $\lambda(A) > 0$
(iii′) $A = D^T D$ *for certain rectangular matrix* D *of rank* n
(iv′) $A = \Delta^T \Delta$ *for certain nondegenerate upper triangular* $n \times n$ *matrix* Δ
(v′) $A = B^2$ *for certain nondegenerate symmetric matrix* B;
(vi′) $A = B^2$ *for certain* $B \succ 0$.

Proof. (i)⇔(ii): this equivalence is stated by Proposition A.7.1.

(ii)⇔(vi): Let $A = U\Lambda U^T$ be the eigenvalue decomposition of A, so that U is orthogonal and Λ is diagonal with nonnegative diagonal entries $\lambda_i(A)$ (we are in the situation of (ii) !). Let $\Lambda^{1/2}$ be the diagonal matrix with the diagonal entries $\lambda_i^{1/2}(A)$; note that $(\Lambda^{1/2})^2 = \Lambda$. The matrix $B = U\Lambda^{1/2}U^T$ is symmetric with nonnegative eigenvalues $\lambda_i^{1/2}(A)$, so that $B \succeq 0$ by Proposition A.7.1, and

$$B^2 = U\Lambda^{1/2}\underbrace{U^T U}_{I}\Lambda^{1/2}U^T = U(\Lambda^{1/2})^2 U^T = U\Lambda U^T = A,$$

as required in (vi).

(vi)⇒(v): evident.

(v)⇒(iv): Let $A = B^2$ with certain symmetric B, and let b_i be i-th column of B. Applying the Gram-Schmidt orthogonalization process (see proof of Theorem A.2.3.(iii)), we can find an orthonormal system of vectors $u_1, ..., u_n$ and lower triangular matrix L such that $b_i = \sum_{j=1}^{i} L_{ij}u_j$, or, which is the same, $B^T = LU$, where U is the orthogonal matrix with the rows $u_1^T, ..., u_n^T$. We now have $A = B^2 = B^T(B^T)^T = LUU^T L^T = LL^T$. We see that $A = \Delta^T\Delta$, where the matrix $\Delta = L^T$ is upper triangular.

(iv)⇒(iii): evident.

(iii)⇒(i): If $A = D^T D$, then $x^T Ax = (Dx)^T(Dx) \geq 0$ for all x.

We have proved the equivalence of the properties (i) – (vi). Slightly modifying the reasoning (do it yourself!), one can prove the equivalence of the properties (i′) – (vi′). ∎

Remark A.7.1 (i) [Checking positive semidefiniteness] Given an $n \times n$ symmetric matrix A, one can check whether it is positive semidefinite by a purely algebraic finite algorithm (the so called *Lagrange diagonalization of a quadratic form*) which requires at most $O(n^3)$ arithmetic operations. Positive definiteness

of a matrix can be checked also by the *Choleski factorization algorithm* which finds the decomposition in (iv′), if it exists, in approximately $\frac{1}{6}n^3$ arithmetic operations.

There exists another useful algebraic criterion (Sylvester's criterion) for positive semidefiniteness of a matrix; according to this criterion, a symmetric matrix A is positive definite if and only if its angular minors are positive, and A is positive semidefinite if and only if all its principal minors are nonnegative. For example, a symmetric 2×2 matrix $A = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$ is positive semidefinite if and only if $a \geq 0$, $c \geq 0$ and $\mathrm{Det}(A) \equiv ac - b^2 \geq 0$.

(ii) [Square root of a positive semidefinite matrix] By the first chain of equivalences in Theorem A.7.5, a symmetric matrix A is $\succeq 0$ if and only if A is the square of a positive semidefinite matrix B. The latter matrix is uniquely defined by $A \succeq 0$ and is called the *square root* of A (notation: $A^{1/2}$).

A.7.4.B. The semidefinite cone. When adding symmetric matrices and multiplying them by reals, we add, respectively multiply by reals, the corresponding quadratic forms. It follows that

> **A.7.4.B.1:** *The sum of positive semidefinite matrices and a product of a positive semidefinite matrix and a nonnegative real is positive semidefinite*,

or, which is the same (see Section B.1.4),

> **A.7.4.B.2:** *$n \times n$ positive semidefinite matrices form a cone $\mathbf{S}^n_+$ in the Euclidean space $\mathbf{S}^n$ of symmetric $n \times n$ matrices, the Euclidean structure being given by the Frobenius inner product $\langle A, B\rangle = \mathrm{Tr}(AB) = \sum\limits_{i,j} A_{ij}B_{ij}$.*

The cone $\mathbf{S}^n_+$ is called the *semidefinite* cone of size n. It is immediately seen that the semidefinite cone $\mathbf{S}^n_+$ is "good" (see Lecture 5), specifically,

- $\mathbf{S}^n_+$ is closed: the limit of a converging sequence of positive semidefinite matrices is positive semidefinite;
- $\mathbf{S}^n_+$ is pointed: the only $n \times n$ matrix A such that both A and $-A$ are positive semidefinite is the zero $n \times n$ matrix;
- $\mathbf{S}^n_+$ possesses a nonempty interior which is comprised of positive definite matrices.

Note that the relation $A \succeq B$ means exactly that $A - B \in \mathbf{S}^n_+$, while $A \succ B$ is equivalent to $A - B \in \mathrm{int}\,\mathbf{S}^n_+$. The "matrix inequalities" $A \succeq B$ ($A \succ B$) match the standard properties of the usual scalar inequalities, e.g.:

$A \succeq A$	[reflexivity]
$A \succeq B$, $B \succeq A \Rightarrow A = B$	[antisymmetry]
$A \succeq B$, $B \succeq C \Rightarrow A \succeq C$	[transitivity]
$A \succeq B$, $C \succeq D \Rightarrow A + C \succeq B + D$	[compatibility with linear operations, I]
$A \succeq B$, $\lambda \geq 0 \Rightarrow \lambda A \succeq \lambda B$	[compatibility with linear operations, II]
$A_i \succeq B_i, A_i \to A, B_i \to B$ as $i \to \infty \Rightarrow A \succeq B$	[closedness]

with evident modifications when $\succeq$ is replaced with $\succ$, or

$$A \succeq B, C \succ D \Rightarrow A + C \succ B + D,$$

etc. Along with these standard properties of inequalities, the inequality $\succeq$ possesses a nice additional property:

> **A.7.4.B.3:** *In a valid $\succeq$-inequality*
>
> $$A \succeq B$$
>
> *one can multiply both sides from the left and by the right by a (rectangular) matrix and its transpose:*
>
> $$A, B \in \mathbf{S}^n, \quad A \succeq B, \quad V \in \mathbf{M}^{n,m}$$
> $$\Downarrow$$
> $$V^T A V \succeq V^T B V$$

Indeed, we should prove that if $A - B \succeq 0$, then also $V^T(A-B)V \succeq 0$, which is immediate – the quadratic form $y^T[V^T(A-B)V]y = (Vy)^T(A-B)(Vy)$ of y is nonnegative along with the quadratic form $x^T(A-B)x$ of x.

An important additional property of the semidefinite cone is its *self-duality*:

Theorem A.7.6 *A symmetric matrix Y has nonnegative Frobenius inner products with all positive semidefinite matrices if and only if Y itself is positive semidefinite.*

Proof. "if" part: Assume that $Y \succeq 0$, and let us prove that then $\mathrm{Tr}(YX) \geq 0$ for every $X \succeq 0$. Indeed, the eigenvalue decomposition of Y can be written as

$$Y = \sum_{i=1}^{n} \lambda_i(Y) e_i e_i^T,$$

where e_i are the orthonormal eigenvectors of Y. We now have

$$\begin{array}{rcl} \mathrm{Tr}(YX) & = & \mathrm{Tr}((\sum\limits_{i=1}^{n} \lambda_i(Y) e_i e_i^T)X) = \sum\limits_{i=1}^{n} \lambda_i(Y)\mathrm{Tr}(e_i e_i^T X) \\ & = & \sum\limits_{i=1}^{n} \lambda_i(Y)\mathrm{Tr}(e_i^T X e_i), \end{array} \tag{A.7.6}$$

where the concluding equality is given by the following well-known property of the trace:

A.7.4.B.4: *Whenever matrices A, B are such that the product AB makes sense and is a square matrix, one has*

$$\mathrm{Tr}(AB) = \mathrm{Tr}(BA).$$

Indeed, we should verify that if $A \in \mathbf{M}^{p,q}$ and $B \in \mathbf{M}^{q,p}$, then $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$. The left hand side quantity in our hypothetic equality is $\sum\limits_{i=1}^{p}\sum\limits_{j=1}^{q} A_{ij}B_{ji}$, and the right hand side quantity is $\sum\limits_{j=1}^{q}\sum\limits_{i=1}^{p} B_{ji}A_{ij}$; they indeed are equal.

Looking at the concluding quantity in (A.7.6), we see that it indeed is nonnegative whenever $X \succeq 0$ (since $Y \succeq 0$ and thus $\lambda_i(Y) \geq 0$ by P.7.5).

"only if" part: We are given Y such that $\mathrm{Tr}(YX) \geq 0$ for all matrices $X \succeq 0$, and we should prove that $Y \succeq 0$. This is immediate: for every vector x, the matrix $X = xx^T$ is positive semidefinite (Theorem A.7.5.(iii)), so that $0 \leq \mathrm{Tr}(Yxx^T) = \mathrm{Tr}(x^TYx) = x^TYx$. Since the resulting inequality $x^TYx \geq 0$ is valid for every x, we have $Y \succeq 0$. ∎

Appendix B

Convex sets in $\mathbf{R}^n$

B.1 Definition and basic properties

B.1.1 A convex set

In the school geometry a figure is called convex if it contains, along with every pair of its points x, y, also the entire segment $[x, y]$ linking the points. This is exactly the definition of a convex set in the multidimensional case; all we need is to say what does it mean "the segment $[x, y]$ linking the points $x, y \in \mathbf{R}^n$". This is said by the following

Definition B.1.1 [Convex set]
1) *Let x, y be two points in $\mathbf{R}^n$. The set*

$$[x, y] = \{z = \lambda x + (1 - \lambda)y \mid 0 \leq \lambda \leq 1\}$$

is called a segment with the endpoints x, y.
2) *A subset M of $\mathbf{R}^n$ is called convex, if it contains, along with every pair of its points x, y, also the entire segment $[x, y]$:*

$$x, y \in M,\ 0 \leq \lambda \leq 1 \Rightarrow \lambda x + (1 - \lambda)y \in M.$$

Note that by this definition an empty set is convex (by convention, or better to say, by the exact sense of the definition: for the empty set, you cannot present a counterexample to show that it is not convex).

B.1.2 Examples of convex sets

B.1.2.A. Affine subspaces

Example B.1.1 *A linear/affine subspace of $\mathbf{R}^n$ is convex.*

Convexity of affine subspaces immediately follows from the possibility to represent these sets as solution sets of systems of linear equations (Proposition A.3.7), due to the following simple and important fact:

Proposition B.1.1 *The solution set of an arbitrary (possibly, infinite) system*

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

of linear inequalities with n unknowns x – the set

$$S = \{x \in \mathbf{R}^n \mid a_\alpha^T x \leq b_\alpha,\ \alpha \in \mathcal{A}\}$$

is convex.

In particular, the solution set of a finite system

$$Ax \leq b$$

of m inequalities with n variables (A is $m \times n$ matrix) is convex; a set of this latter type is called polyhedral.

Exercise B.1 *Prove Proposition B.1.1.*

Remark B.1.1 *Note that every set given by Proposition B.1.1 is not only convex, but also closed (why?). In fact, from Separation Theorem (Theorem B.2.5 below) it follows that*

> *Every closed convex set in $\mathbf{R}^n$ is the solution set of a (perhaps, infinite) system of nonstrict linear inequalities.*

B.1.2.B. Unit balls of norms

Let $\|\cdot\|$ be a norm on $\mathbf{R}^n$ i.e., a real-valued function on $\mathbf{R}^n$ satisfying the three characteristic properties of a norm (Section A.4.1), specifically:

A. [positivity] $\|x\| \geq 0$ for all $x \in \mathbf{R}^n$; $\|x\| = 0$ is and only if $x = 0$;

B. [homogeneity] For $x \in \mathbf{R}^n$ and $\lambda \in \mathbf{R}$, one has

$$\|\lambda x\| = |\lambda| \|x\|;$$

C. [triangle inequality] For all $x, y \in \mathbf{R}^n$ one has

$$\|x + y\| \leq \|x\| + \|y\|.$$

Example B.1.2 *The unit ball of the norm $\|\cdot\|$ – the set*

$$\{x \in E \mid \|x\| \leq 1\},$$

same as every other $\|\cdot\|$-ball

$$\{x \mid \|x - a\| \leq r\}$$

($a \in \mathbf{R}^n$ and $r \geq 0$ are fixed) is convex.

In particular, Euclidean balls ($\|\cdot\|$-balls associated with the standard Euclidean norm $\|x\|_2 = \sqrt{x^T x}$) are convex.

The standard examples of norms on $\mathbf{R}^n$ are the ℓ_p-norms

$$\|x\|_p = \begin{cases} \left(\sum_{i=1}^n |x_i|^p \right)^{1/p}, & 1 \leq p < \infty \\ \max_{1 \leq i \leq n} |x_i|, & p = \infty \end{cases}.$$

These indeed are norms (which is not clear in advance). When $p = 2$, we get the usual Euclidean norm; of course, you know how the Euclidean ball looks. When $p = 1$, we get

$$\|x\|_1 = \sum_{i=1}^n |x_i|,$$

and the unit ball is the *hyperoctahedron*

$$V = \{x \in \mathbf{R}^n \mid \sum_{i=1}^n |x_i| \leq 1\}$$

When $p = \infty$, we get

$$\|x\|_\infty = \max_{1\le i\le n} |x_i|,$$

and the unit ball is the *hypercube*

$$V = \{x \in \mathbf{R}^n \mid -1 \le x_i \le 1, 1 \le i \le n\}.$$

Exercise B.2 † *Prove that unit balls of norms on $\mathbf{R}^n$ are exactly the same as convex sets V in $\mathbf{R}^n$ satisfying the following three properties:*

1. *V is symmetric w.r.t. the origin: $x \in V \Rightarrow -x \in V$;*
2. *V is bounded and closed;*
3. *V contains a neighbourhood of the origin.*

A set V satisfying the outlined properties is the unit ball of the norm

$$\|x\| = \inf\left\{t \ge 0 : t^{-1}x \in V\right\}.$$

Hint: You could find useful to verify and to exploit the following facts:

1. A norm $\|\cdot\|$ on $\mathbf{R}^n$ is Lipschitz continuous with respect to the standard Euclidean distance: there exists $C_{\|\cdot\|} < \infty$ such that $|\|x\| - \|y\|| \le C_{\|\cdot\|}\|x - y\|_2$ for all x, y
2. Vice versa, the Euclidean norm is Lipschitz continuous with respect to a given norm $\|\cdot\|$: there exists $c_{\|\cdot\|} < \infty$ such that $|\|x\|_2 - \|y\|_2| \le c_{\|\cdot\|}\|x - y\|$ for all x, y

B.1.2.C. Ellipsoid

Example B.1.3 [Ellipsoid] *Let Q be a $n \times n$ matrix which is symmetric ($Q = Q^T$) and positive definite ($x^TQx \ge 0$, with $\ge$ being $=$ if and only if $x = 0$). Then, for every nonnegative r, the Q-ellipsoid of radius r centered at a – the set*

$$\{x \mid (x - a)^TQ(x - a) \le r^2\}$$

is convex.

To see that an ellipsoid $\{x : (x - a)^TQ(x - a) \le r^2\}$ is convex, note that since Q is positive definite, the matrix $Q^{1/2}$ is well-defined and positive definite. Now, if $\|\cdot\|$ is a norm on $\mathbf{R}^n$ and P is a nonsingular $n \times n$ matrix, the function $\|Px\|$ is a norm along with $\|\cdot\|$ (why?). Thus, the function $\|x\|_Q \equiv \sqrt{x^TQx} = \|Q^{1/2}x\|_2$ is a norm along with $\|\cdot\|_2$, and the ellipsoid in question clearly is just $\|\cdot\|_Q$-ball of radius r centered at a.

B.1.2.C. Neighbourhood of a convex set

Example B.1.4 *Let M be a convex set in $\mathbf{R}^n$, and let $\epsilon > 0$. Then, for every norm $\|\cdot\|$ on $\mathbf{R}^n$, the ϵ-neighbourhood of M, i.e., the set*

$$M_\epsilon = \{y \in \mathbf{R}^n \mid \mathrm{dist}_{\|\cdot\|}(y, M) \equiv \inf_{x\in M} \|y - x\| \le \epsilon\}$$

is convex.

Exercise B.3 *Justify the statement of Example B.1.4.*

B.1.3 Inner description of convex sets: Convex combinations and convex hull

B.1.3.A. Convex combinations

Recall the notion of *linear combination* y of vectors $y_1, ..., y_m$ – this is a vector represented as

$$y = \sum_{i=1}^{m} \lambda_i y_i,$$

where λ_i are real coefficients. Specifying this definition, we have come to the notion of an *affine combination* - this is a linear combination with the sum of coefficients equal to one. The last notion in this genre is the one of *convex combination*.

Definition B.1.2 *A convex combination of vectors $y_1, ..., y_m$ is their affine combination with nonnegative coefficients, or, which is the same, a linear combination*

$$y = \sum_{i=1}^{m} \lambda_i y_i$$

with nonnegative coefficients with unit sum:

$$\lambda_i \geq 0, \ \sum_{i=1}^{m} \lambda_i = 1.$$

The following statement resembles those in Corollary A.3.2:

Proposition B.1.2 *A set M in $\mathbf{R}^n$ is convex if and only if it is closed with respect to taking all convex combinations of its elements, i.e., if and only if every convex combination of vectors from M again is a vector from M.*

Exercise B.4 *Prove Proposition B.1.2.*
Hint: Assuming $\lambda_1, ..., \lambda_m > 0$, one has

$$\sum_{i=1}^{m} \lambda_i y_i = \lambda_1 y_1 + (\lambda_2 + \lambda_3 + ... + \lambda_m) \sum_{i=2}^{m} \mu_i y_i, \quad \mu_i = \frac{\lambda_i}{\lambda_2 + \lambda_3 + ... + \lambda_m}.$$

B.1.3.B. Convex hull

Same as the property to be linear/affine subspace, the property to be convex is preserved by taking intersections (why?):

Proposition B.1.3 *Let $\{M_\alpha\}_\alpha$ be an arbitrary family of convex subsets of $\mathbf{R}^n$. Then the intersection*

$$M = \cap_\alpha M_\alpha$$

is convex.

As an immediate consequence, we come to the notion of *convex hull* $\mathrm{Conv}(M)$ of a nonempty subset in $\mathbf{R}^n$ (cf. the notions of linear/affine hull):

Corollary B.1.1 [Convex hull]
Let M be a nonempty subset in $\mathbf{R}^n$. Then among all convex sets containing M (these sets do exist, e.g., $\mathbf{R}^n$ itself) there exists the smallest one, namely, the intersection of all convex sets containing M. This set is called the convex hull of M [notation: $\mathrm{Conv}(M)$].

The linear span of M is the set of all linear combinations of vectors from M, the affine hull is the set of all affine combinations of vectors from M. As you guess,

Proposition B.1.4 [Convex hull via convex combinations] *For a nonempty $M \subset \mathbf{R}^n$:*

$$\mathrm{Conv}(M) = \{\textit{the set of all convex combinations of vectors from } M\}.$$

Exercise B.5 *Prove Proposition B.1.4.*

B.1.3.C. Simplex

The convex hull of $m+1$ affinely independent points $y_0, ..., y_m$ (Section A.3.3) is called *m-dimensional simplex with the vertices $y_0, .., y_m$*. By results of Section A.3.3, every point x of an m-dimensional simplex with vertices $y_0, ..., y_m$ admits exactly one representation as a convex combination of the vertices; the corresponding coefficients form the unique solution to the system of linear equations

$$\sum_{i=0}^{m} \lambda_i x_i = x, \sum_{i=0}^{m} \lambda_i = 1.$$

This system is solvable if and only if $x \in M = \mathrm{Aff}(\{y_0, .., y_m\})$, and the components of the solution (the barycentric coordinates of x) are affine functions of $x \in \mathrm{Aff}(M)$; the simplex itself is comprised of points from M with nonnegative barycentric coordinates.

B.1.4 Cones

A nonempty subset M of $\mathbf{R}^n$ is called *conic*, if it contains, along with every point $x \in M$, the entire ray $\mathbf{R}x = \{tx \mid t \geq 0\}$ spanned by the point:

$$x \in M \Rightarrow tx \in M \ \forall t \geq 0.$$

A convex conic set is called a *cone*.

Proposition B.1.5 *A nonempty subset M of $\mathbf{R}^n$ is a cone if and only if it possesses the following pair of properties:*

- *is conic: $x \in M, t \geq 0 \Rightarrow tx \in M$;*
- *contains sums of its elements: $x, y \in M \Rightarrow x + y \in M$.*

Exercise B.6 *Prove Proposition B.1.5.*

As an immediate consequence, we get that a cone is closed with respect to taking linear combinations *with nonnegative coefficients* of the elements, and vice versa – a nonempty set closed with respect to taking these combinations is a cone.

Example B.1.5 *The solution set of an arbitrary (possibly, infinite) system*

$$a_\alpha^T x \leq 0, \ \alpha \in \mathcal{A}$$

of homogeneous linear inequalities with n unknowns x – the set

$$K = \{x \mid a_\alpha^T x \leq 0 \ \forall \alpha \in \mathcal{A}\}$$

– is a cone.

In particular, the solution set to a homogeneous finite system of m homogeneous linear inequalities

$$Ax \leq 0$$

(A is $m \times n$ matrix) is a cone; a cone of this latter type is called polyhedral.

Note that the cones given by systems of linear homogeneous nonstrict inequalities necessarily are closed. From Separation Theorem B.2.5 it follows that, vice versa, every closed convex cone is the solution set to such a system, so that Example B.1.5 is the generic example of a closed convex cone.

> Cones form a very important family of convex sets, and one can develop theory of cones absolutely similar (and in a sense, equivalent) to that one of all convex sets. E.g., introducing the notion of *conic combination* of vectors $x_1, ..., x_k$ as a linear combination of the vectors with nonnegative coefficients, you can easily prove the following statements completely similar to those for general convex sets, with conic combination playing the role of convex one:

- A set is a cone if and only if it is nonempty and is closed with respect to taking all conic combinations of its elements;
- Intersection of a family of cones is again a cone; in particular, for every nonempty set $M \subset \mathbf{R}^n$ there exists the smallest cone containing M – its conic!hull $\mathrm{Cone}\,(M)$, and this conic hull is comprised of all conic combinations of vectors from M.

In particular, the conic hull of a nonempty finite set $M = \{u_1, ..., u_N\}$ of vectors in $\mathbf{R}^n$ is the cone

$$\mathrm{Cone}\,(M) = \{\sum_{i=1}^{N} \lambda_i u_i \mid \lambda_i \geq 0,\, i = 1, ..., N\}.$$

B.1.5 "Calculus" of convex sets

Proposition B.1.6 *The following operations preserve convexity of sets:*

1. *Intersection: if M_α, $\alpha \in \mathcal{A}$, are convex sets, so is the set $\bigcap_\alpha M_\alpha$.*
2. *Direct product: if $M_1 \subset \mathbf{R}^{n_1}$ and $M_2 \subset \mathbf{R}^{n_2}$ are convex sets, so is the set*

$$M_1 \times M_2 = \{y = (y_1, y_2) \in \mathbf{R}^{n_1} \times \mathbf{R}^{n_2} = \mathbf{R}^{n_1+n_2} : y_1 \in M_1, y_2 \in M_2\}.$$

3. *Arithmetic summation and multiplication by reals: if $M_1, ..., M_k$ are convex sets in $\mathbf{R}^n$ and $\lambda_1, ..., \lambda_k$ are arbitrary reals, then the set*

$$\lambda_1 M_1 + ... + \lambda_k M_k = \{\sum_{i=1}^{k} \lambda_i x_i \mid x_i \in M_i,\, i = 1, ..., k\}$$

is convex.

4. *Taking the image under affine mapping: if $M \subset \mathbf{R}^n$ is convex and $x \mapsto \mathcal{A}(x) \equiv Ax + b$ is an affine mapping from $\mathbf{R}^n$ into $\mathbf{R}^m$ (A is $m \times n$ matrix, b is m-dimensional vector), then the set*

$$\mathcal{A}(M) = \{y = \mathcal{A}(x) \equiv Ax + a \mid x \in M\}$$

is a convex set in $\mathbf{R}^m$;

5. *Taking the inverse image under affine mapping: if $M \subset \mathbf{R}^n$ is convex and $y \mapsto Ay + b$ is an affine mapping from $\mathbf{R}^m$ to $\mathbf{R}^n$ (A is $n \times m$ matrix, b is n-dimensional vector), then the set*

$$\mathcal{A}^{-1}(M) = \{y \in \mathbf{R}^m \mid \mathcal{A}(y) \in M\}$$

is a convex set in $\mathbf{R}^m$.

Exercise B.7 *Prove Proposition B.1.6.*

B.1.6 Topological properties of convex sets

Convex sets and closely related objects - convex functions - play the central role in Optimization. To play this role properly, the convexity alone is insufficient; we need convexity plus closedness.

B.1.6.A. The closure

It is clear from definition of a closed set (Section A.4.3) that the intersection of a family of closed sets in $\mathbf{R}^n$ is also closed. From this fact it, as always, follows that for every subset M of $\mathbf{R}^n$ there exists the smallest closed set containing M; this set is called the *closure* of M and is denoted $\mathrm{cl}\, M$. In Analysis they prove the following inner description of the closure of a set in a metric space (and, in particular, in $\mathbf{R}^n$):

The closure of a set $M \subset \mathbf{R}^n$ is exactly the set comprised of the limits of all converging sequences of elements of M.

With this fact in mind, it is easy to prove that, e.g., the closure of the open Euclidean ball

$$\{x \mid |x-a| < r\} \quad [r > 0]$$

is the closed ball $\{x \mid \|x-a\|_2 \le r\}$. Another useful application example is the closure of a set

$$M = \{x \mid a_\alpha^T x < b_\alpha,\ \alpha \in \mathcal{A}\}$$

given by strict linear inequalities: *if such a set is nonempty*, then its closure is given by the nonstrict versions of the same inequalities:

$$\mathrm{cl}\, M = \{x \mid a_\alpha^T x \le b_\alpha,\ \alpha \in \mathcal{A}\}.$$

Nonemptiness of M in the latter example is essential: the set M given by two strict inequalities

$$x < 0, \quad -x < 0$$

in $\mathbf{R}$ clearly is empty, so that its closure also is empty; in contrast to this, applying formally the above rule, we would get wrong answer

$$\mathrm{cl}\, M = \{x \mid x \le 0, x \ge 0\} = \{0\}.$$

B.1.6.B. The interior

Let $M \subset \mathbf{R}^n$. We say that a point $x \in M$ is an *interior* point of M, if some neighbourhood of the point is contained in M, i.e., there exists centered at x ball of positive radius which belongs to M:

$$\exists r > 0 \quad B_r(x) \equiv \{y \mid \|y-x\|_2 \le r\} \subset M.$$

The set of all interior points of M is called the *interior* of M [notation: $\mathrm{int}\, M$].

E.g.,

- The interior of an open set is the set itself;
- The interior of the closed ball $\{x \mid \|x-a\|_2 \le r\}$ is the open ball $\{x \mid \|x-a\|_2 < r\}$ (why?)
- The interior of a polyhedral set $\{x \mid Ax \le b\}$ with matrix A not containing zero rows is the set $\{x \mid Ax < b\}$ (why?)

  The latter statement is not, generally speaking, valid for sets of solutions of infinite systems of linear inequalities. E.g., the system of inequalities

  $$x \le \frac{1}{n},\ n = 1, 2, ...$$

  in $\mathbf{R}$ has, as a solution set, the nonpositive ray $\mathbf{R}_- = \{x \le 0\}$; the interior of this ray is the negative ray $\{x < 0\}$. At the same time, strict versions of our inequalities

  $$x < \frac{1}{n},\ n = 1, 2, ...$$

  define the same nonpositive ray, not the negative one.

It is also easily seen (this fact is valid for arbitrary metric spaces, not for $\mathbf{R}^n$ only), that

- the interior of an arbitrary set is open

The interior of a set is, of course, contained in the set, which, in turn, is contained in its closure:

$$\operatorname{int} M \subset M \subset \operatorname{cl} M. \tag{B.1.1}$$

The complement of the interior in the closure – the set

$$\partial M = \operatorname{cl} M \backslash \operatorname{int} M$$

– is called the *boundary* of M, and the points of the boundary are called *boundary points* of M (Warning: these points not necessarily belong to M, since M can be less than $\operatorname{cl} M$; in fact, all boundary points belong to M if and only if $M = \operatorname{cl} M$, i.e., if and only if M is closed).

The boundary of a set clearly is closed (as the intersection of two closed sets $\operatorname{cl} M$ and $\mathbf{R}^n \backslash \operatorname{int} M$; the latter set is closed as a complement to an open set). From the definition of the boundary,

$$M \subset \operatorname{int} M \cup \partial M \quad [= \operatorname{cl} M],$$

so that a point from M is either an interior, or a boundary point of M.

B.1.6.C. The relative interior

Many of the constructions in Optimization possess nice properties in the interior of the set the construction is related to and may lose these nice properties at the boundary points of the set; this is why in many cases we are especially interested in interior points of sets and want the set of these points to be "enough massive". What to do if it is not the case – e.g., there are no interior points at all (look at a segment in the plane)? It turns out that in these cases we can use a good surrogate of the "normal" interior – the *relative interior* defined as follows.

Definition B.1.3 [Relative interior] *Let $M \subset \mathbf{R}^n$. We say that a point $x \in M$ is relative interior for M, if M contains the intersection of a small enough ball centered at x with* $\operatorname{Aff}(M)$*:*

$$\exists r > 0 \quad B_r(x) \cap \operatorname{Aff}(M) \equiv \{y \mid y \in \operatorname{Aff}(M), \|y - x\|_2 \le r\} \subset M.$$

The set of all relative interior points of M is called its relative interior [*notation:* $\operatorname{ri} M$].

E.g. the relative interior of a singleton is the singleton itself (since a point in the 0-dimensional space is the same as a ball of a positive radius); more generally, the relative interior of an affine subspace is the set itself. The interior of a segment $[x, y]$ $(x \neq y)$ in $\mathbf{R}^n$ is empty whenever $n > 1$; in contrast to this, the relative interior is nonempty independently of n and is the interval (x, y) – the segment with deleted endpoints. Geometrically speaking, the relative interior is the interior we get when regard M as a subset of its affine hull (the latter, geometrically, is nothing but $\mathbf{R}^k$, k being the affine dimension of $\operatorname{Aff}(M)$).

Exercise B.8 *Prove that the relative interior of a simplex with vertices $y_0, ..., y_m$ is exactly the set* $\{x = \sum_{i=0}^{m} \lambda_i y_i : \lambda_i > 0, \sum_{i=0}^{m} \lambda_i = 1\}$.

We can play with the notion of the relative interior in basically the same way as with the one of interior, namely:

- since $\operatorname{Aff}(M)$, as every affine subspace, is closed and contains M, it contains also the smallest closed sets containing M, i.e., $\operatorname{cl} M$. Therefore we have the following analogies of inclusions (B.1.1):

$$\operatorname{ri} M \subset M \subset \operatorname{cl} M \quad [\subset \operatorname{Aff}(M)]; \tag{B.1.2}$$

- we can define the *relative boundary* $\partial_{\operatorname{ri}} M = \operatorname{cl} M \backslash \operatorname{ri} M$ which is a closed set contained in $\operatorname{Aff}(M)$, and, as for the "actual" interior and boundary, we have

$$\operatorname{ri} M \subset M \subset \operatorname{cl} M = \operatorname{ri} M + \partial_{\operatorname{ri}} M.$$

Of course, if $\operatorname{Aff}(M) = \mathbf{R}^n$, then the relative interior becomes the usual interior, and similarly for boundary; this for sure is the case when $\operatorname{int} M \neq \emptyset$ (since then M contains a ball B, and therefore the affine hull of M is the entire $\mathbf{R}^n$, which is the affine hull of B).

B.1.6.D. Nice topological properties of a convex set

An arbitrary set M in $\mathbf{R}^n$ may possess very pathological topology: both inclusions in the chain

$$\mathrm{ri}\, M \subset M \subset \mathrm{cl}\, M$$

can be very "non-tight". E.g., let M be the set of rational numbers in the segment $[0,1] \subset \mathbf{R}$. Then $\mathrm{ri}\, M = \mathrm{int}\, M = \emptyset$ – since every neighbourhood of every rational real contains irrational reals – while $\mathrm{cl}\, M = [0,1]$. Thus, $\mathrm{ri}\, M$ is "incomparably smaller" than M, $\mathrm{cl}\, M$ is "incomparably larger", and M is contained in its relative boundary (by the way, what is this relative boundary?).

The following proposition demonstrates that the topology of a *convex* set M is much better than it might be for an arbitrary set.

Theorem B.1.1 *Let M be a convex set in $\mathbf{R}^n$. Then*

(i) *The interior* $\mathrm{int}\, M$*, the closure* $\mathrm{cl}\, M$ *and the relative interior* $\mathrm{ri}\, M$ *are convex;*

(ii) *If M is nonempty, then the relative interior* $\mathrm{ri}\, M$ *of M is nonempty*

(iii) *The closure of M is the same as the closure of its relative interior:*

$$\mathrm{cl}\, M = \mathrm{cl}\, \mathrm{ri}\, M$$

(in particular, every point of $\mathrm{cl}\, M$ *is the limit of a sequence of points from* $\mathrm{ri}\, M$*)*

(iv) *The relative interior remains unchanged when we replace M with its closure:*

$$\mathrm{ri}\, M = \mathrm{ri}\, \mathrm{cl}\, M.$$

Proof. (i): prove yourself!

(ii): Let M be a nonempty convex set, and let us prove that $\mathrm{ri}\, M \neq \emptyset$. By translation, we may assume that $0 \in M$. Further, we may assume that the linear span of M is the entire $\mathbf{R}^n$. Indeed, as far as linear operations and the Euclidean structure are concerned, the linear span L of M, as every other linear subspace in $\mathbf{R}^n$, is equivalent to certain $\mathbf{R}^k$; since the notion of relative interior deals only with linear and Euclidean structures, we lose nothing thinking of $\mathrm{Lin}(M)$ as of $\mathbf{R}^k$ and taking it as our universe instead of the original universe $\mathbf{R}^n$. Thus, in the rest of the proof of (ii) we assume that $0 \in M$ and $\mathrm{Lin}(M) = \mathbf{R}^n$; what we should prove is that the interior of M (which in the case in question is the same as relative interior) is nonempty. Note that since $0 \in M$, we have $\mathrm{Aff}(M) = \mathrm{Lin}(M) = \mathbf{R}^n$.

Since $\mathrm{Lin}(M) = \mathbf{R}^n$, we can find in M n linearly independent vectors $a_1, .., a_n$. Let also $a_0 = 0$. The $n+1$ vectors $a_0, ..., a_n$ belong to M, and *since M is convex*, the convex hull of these vectors also belongs to M. This convex hull is the set

$$\Delta = \{x = \sum_{i=0}^{n} \lambda_i a_i : \lambda \geq 0, \sum_i \lambda_i = 1\} = \{x = \sum_{i=1}^{n} \mu_i a_i : \mu \geq 0, \sum_{i=1}^{n} \mu_i \leq 1\}.$$

We see that Δ is the image of the *standard full-dimensional simplex*

$$\{\mu \in \mathbf{R}^n : \mu \geq 0, \sum_{i=1}^{n} \mu_i \leq 1\}$$

under linear transformation $\mu \mapsto A\mu$, where A is the matrix with the columns $a_1, ..., a_n$. The standard simplex clearly has a nonempty interior (comprised of all vectors $\mu > 0$ with $\sum_i \mu_i < 1$); since A is nonsingular (due to linear independence of $a_1, ..., a_n$), multiplication by A maps open sets onto open ones, so that Δ has a nonempty interior. Since $\Delta \subset M$, the interior of M is nonempty. □

(iii): We should prove that the closure of $\mathrm{ri}\, M$ is exactly the same that the closure of M. In fact we shall prove even more:

Lemma B.1.1 *Let $x \in \mathrm{ri}\, M$ and $y \in \mathrm{cl}\, M$. Then all points from the half-segment $[x, y)$,*

$$[x, y) = \{z = (1-\lambda)x + \lambda y \mid 0 \leq \lambda < 1\}$$

belong to the relative interior of M.

Proof of the Lemma. Let $\mathrm{Aff}(M) = a + L$, L being linear subspace; then

$$M \subset \mathrm{Aff}(M) = x + L.$$

Let B be the unit ball in L:

$$B = \{h \in L \mid \|h\|_2 \leq 1\}.$$

Since $x \in \mathrm{ri}\, M$, there exists positive radius r such that

$$x + rB \subset M. \tag{B.1.3}$$

Now let $\lambda \in [0, 1)$, and let $z = (1 - \lambda)x + \lambda y$. Since $y \in \mathrm{cl}\, M$, we have $y = \lim_{i\to\infty} y_i$ for certain sequence of points from M. Setting $z_i = (1 - \lambda)x + \lambda y_i$, we get $z_i \to z$ as $i \to \infty$. Now, from (B.1.3) and the convexity of M is follows that the sets $Z_i = \{u = (1 - \lambda)x' + \lambda y_i : x' \in x + rB\}$ are contained in M; clearly, Z_i is exactly the set $z_i + r'B$, where $r' = (1 - \lambda)r > 0$. Thus, z is the limit of sequence z_i, and r'-neighbourhood (in $\mathrm{Aff}(M)$) of every one of the points z_i belongs to M. For every $r'' < r'$ and for all i such that z_i is close enough to z, the r'-neighbourhood of z_i contains the r''-neighbourhood of z; thus, a neighbourhood (in $\mathrm{Aff}(M)$) of z belongs to M, whence $z \in \mathrm{ri}\, M$. □

A useful byproduct of Lemma B.1.1 is as follows:

Corollary B.1.2 *Let M be a convex set. Then every convex combination*

$$\sum_i \lambda_i x_i$$

of points $x_i \in \mathrm{cl}\, M$ where at least one term with positive coefficient corresponds to $x_i \in \mathrm{ri}\, M$ is in fact a point from $\mathrm{ri}\, M$.

(iv): The statement is evidently true when M is empty, so assume that M is nonempty. The inclusion $\mathrm{ri}\, M \subset \mathrm{ri}\,\mathrm{cl}\, M$ is evident, and all we need is to prove the inverse inclusion. Thus, let $z \in \mathrm{ri}\,\mathrm{cl}\, M$, and let us prove that $z \in \mathrm{ri}\, M$. Let $x \in \mathrm{ri}\, M$ (we already know that the latter set is nonempty). Consider the segment $[x, z]$; since z is in the relative interior of $\mathrm{cl}\, M$, we can extend a little bit this segment through the point z, not leaving $\mathrm{cl}\, M$, i.e., there exists $y \in \mathrm{cl}\, M$ such that $z \in [x, y)$. We are done, since by Lemma B.1.1 from $z \in [x, y)$, with $x \in \mathrm{ri}\, M$, $y \in \mathrm{cl}\, M$, it follows that $z \in \mathrm{ri}\, M$. ■

We see from the proof of Theorem B.1.1 that to get a closure of a (nonempty) convex set, it suffices to subject it to the "radial" closure, i.e., to take a point $x \in \mathrm{ri}\, M$, take all rays in $\mathrm{Aff}(M)$ starting at x and look at the intersection of such a ray l with M; such an intersection will be a convex set on the line which contains a one-sided neighbourhood of x, i.e., is either a segment $[x, y_l]$, or the entire ray l, or a half-interval $[x, y_l)$. In the first two cases we should not do anything; in the third we should add y to M. After all rays are looked through and all "missed" endpoints y_l are added to M, we get the closure of M. To understand what is the role of convexity here, look at the *nonconvex* set of rational numbers from $[0, 1]$; the interior (≡ relative interior) of this "highly percolated" set is empty, the closure is $[0, 1]$, and there is no way to restore the closure in terms of the interior.

B.2 Main theorems on convex sets

B.2.1 Caratheodory Theorem

Let us call the *affine dimension* (or simple *dimension* of a nonempty set $M \subset \mathbf{R}^n$ (notation: $\dim M$) the affine dimension of $\mathrm{Aff}(M)$.

Theorem B.2.1 [Caratheodory] *Let $M \subset \mathbf{R}^n$, and let $\dim \mathrm{Conv} M = m$. Then every point $x \in \mathrm{Conv} M$ is a convex combination of at most $m + 1$ points from M.*

Proof. Let $x \in \mathrm{Conv}M$. By Proposition B.1.4 on the structure of convex hull, x is convex combination of certain points $x_1, ..., x_N$ from M:

$$x = \sum_{i=1}^{N} \lambda_i x_i, \quad [\lambda_i \geq 0, \sum_{i=1}^{N} \lambda_i = 1].$$

Let us choose among all these representations of x as a convex combination of points from M the one with the smallest possible N, and let it be the above combination. I claim that $N \leq m+1$ (this claim leads to the desired statement). Indeed, if $N > m+1$, then the system of $m+1$ homogeneous equations

$$\begin{aligned} \sum_{i=1}^{N} \mu_i x_i &= 0 \\ \sum_{i=1}^{N} \mu_i &= 0 \end{aligned}$$

with N unknowns $\mu_1, ..., \mu_N$ has a nontrivial solution $\delta_1, ..., \delta_N$:

$$\sum_{i=1}^{N} \delta_i x_i = 0, \sum_{i=1}^{N} \delta_i = 0, (\delta_1, ..., \delta_N) \neq 0.$$

It follows that, for every real t,

$$(*) \quad \sum_{i=1}^{N} [\lambda_i + t\delta_i] x_i = x.$$

What is to the left, is an affine combination of x_i's. When $t = 0$, this is a convex combination - all coefficients are nonnegative. When t is large, this is *not* a convex combination, since some of δ_i's are negative (indeed, not all of them are zero, and the sum of δ_i's is 0). There exists, of course, the largest t for which the combination (*) has nonnegative coefficients, namely

$$t^* = \min_{i:\delta_i<0} \frac{\lambda_i}{|\delta_i|}.$$

For this value of t, the combination (*) is with nonnegative coefficients, and at least one of the coefficients is zero; thus, we have represented x as a convex combination of less than N points from M, which contradicts the definition of N. ∎

B.2.2 Radon Theorem

Theorem B.2.2 [Radon] *Let S be a set of at least $n+2$ points $x_1, ..., x_N$ in $\mathbf{R}^n$. Then one can split the set into two nonempty subsets S_1 and S_2 with intersecting convex hulls: there exists partitioning $I \cup J = \{1, ..., N\}$, $I \cap J = \emptyset$, of the index set $\{1, ..., N\}$ into two nonempty sets I and J and convex combinations of the points $\{x_i, i \in I\}$, $\{x_j, j \in J\}$ which coincide with each other, i.e., there exist α_i, $i \in I$, and β_j, $j \in J$, such that*

$$\sum_{i \in I} \alpha_i x_i = \sum_{j \in J} \beta_j x_j; \quad \sum_i \alpha_i = \sum_j \beta_j = 1; \quad \alpha_i, \beta_j \geq 0.$$

Proof. Since $N > n+1$, the homogeneous system of $n+1$ scalar equations with N unknowns $\mu_1, ..., \mu_N$

$$\begin{aligned} \sum_{i=1}^{N} \mu_i x_i &= 0 \\ \sum_{i=1}^{N} \mu_i &= 0 \end{aligned}$$

has a nontrivial solution $\lambda_1, ..., \lambda_N$:

$$\sum_{i=1}^{N} \mu_i x_i = 0, \sum_{i=1}^{N} \lambda_i = 0, [(\lambda_1, ..., \lambda_N) \neq 0].$$

Let $I = \{i \mid \lambda_i \geq 0\}$, $J = \{i \mid \lambda_i < 0\}$; then I and J are nonempty and form a partitioning of $\{1, ..., N\}$. We have

$$a \equiv \sum_{i \in I} \lambda_i = \sum_{j \in J} (-\lambda_j) > 0$$

(since the sum of all λ's is zero and not all λ's are zero). Setting

$$\alpha_i = \frac{\lambda_i}{a}, i \in I, \ \beta_j = \frac{-\lambda_j}{a}, j \in J,$$

we get

$$\alpha_i \geq 0, \beta_j \geq 0, \sum_{i \in I} \alpha_i = 1, \sum_{j \in J} \beta_j = 1,$$

and

$$[\sum_{i \in I} \alpha_i x_i] - [\sum_{j \in J} \beta_j x_j] = a^{-1} \left([\sum_{i \in I} \lambda_i x_i] - [\sum_{j \in J} (-\lambda_j) x_j] \right) = a^{-1} \sum_{i=1}^{N} \lambda_i x_i = 0. \quad \blacksquare$$

B.2.3 Helley Theorem

Theorem B.2.3 [Helley, I] *Let $\mathcal{F}$ be a finite family of convex sets in $\mathbf{R}^n$. Assume that every $n+1$ sets from the family have a point in common. The all the sets have a point in common.*

Proof. Let us prove the statement by induction on the number N of sets in the family. The case of $N \leq n+1$ is evident. Now assume that the statement holds true for all families with certain number $N \geq n+1$ of sets, and let $S_1, ..., S_N, S_{N+1}$ be a family of $N+1$ convex sets which satisfies the premise of the Helley Theorem; we should prove that the intersection of the sets $S_1, ..., S_N, S_{N+1}$ is nonempty.

Deleting from our $N+1$-set family the set S_i, we get N-set family which satisfies the premise of the Helley Theorem and thus, by the inductive hypothesis, the intersection of its members is nonempty:

$$(\forall i \leq N+1): \ T^i = S_1 \cap S_2 \cap ... \cap S_{i-1} \cap S_{i+1} \cap ... \cap S_{N+1} \neq \emptyset.$$

Let us choose a point x_i in the (nonempty) set T^i. We get $N+1 \geq n+2$ points from $\mathbf{R}^n$. By Radon's Theorem, we can partition the index set $\{1, ..., N+1\}$ into two nonempty subsets I and J in such a way that certain convex combination x of the points x_i, $i \in I$, is a convex combination of the points x_j, $j \in J$, as well. Let us verify that x belongs to all the sets $S_1, ..., S_{N+1}$, which will complete the proof. Indeed, let i^* be an index from our index set; let us prove that $x \in S_{i^*}$. We have either $i^* \in I$, or $i^* \in J$. In the first case all the sets T^j, $j \in J$, are contained in S_{i^*} (since S_{i^*} participates in all intersections which give T^i with $i \neq i^*$). Consequently, all the points x_j, $j \in J$, belong to S_{i^*}, and therefore x, which is a convex combination of these points, also belongs to S_{i^*} (all our sets are convex!), as required. In the second case similar reasoning says that all the points x_i, $i \in I$, belong to S_{i^*}, and therefore x, which is a convex combination of these points, belongs to S_{i^*}. $\blacksquare$

Exercise B.9 *Let $S_1, ..., S_N$ be a family of N convex sets in $\mathbf{R}^n$, and let m be the affine dimension of $\mathrm{Aff}(S_1 \cup ... \cup S_N)$. Assume that every $m+1$ sets from the family have a point in common. Prove that all sets from the family have a point in common.*

In the aforementioned version of the Helley Theorem we dealt with finite families of convex sets. To extend the statement to the case of infinite families, we need to strengthen slightly the assumption. The resulting statement is as follows:

Theorem B.2.4 [Helley, II] *Let $\mathcal{F}$ be an arbitrary family of convex sets in $\mathbf{R}^n$. Assume that*

(a) every $n+1$ sets from the family have a point in common,

and

(b) every set in the family is closed, and the intersection of the sets from certain finite subfamily of the family is bounded (e.g., one of the sets in the family is bounded).

Then all the sets from the family have a point in common.

Proof. By the previous theorem, all finite subfamilies of $\mathcal{F}$ have nonempty intersections, and these intersections are convex (since intersection of a family of convex sets is convex, Theorem B.1.3); in view of (a) these intersections are also closed. Adding to $\mathcal{F}$ all intersections of finite subfamilies of $\mathcal{F}$, we get a larger family $\mathcal{F}'$ comprised of closed convex sets, and a finite subfamily of this larger family again has a nonempty intersection. Besides this, from (b) it follows that this new family contains a bounded set Q. Since all the sets are closed, the family of sets

$$\{Q \cap Q' \mid Q' \in \mathcal{F}\}$$

is a *nested family of compact sets* (i.e., a family of compact sets with nonempty intersection of sets from every finite subfamily); by the well-known Analysis theorem such a family has a nonempty intersection[1)]. ■

B.2.4 Homogeneous Farkas Lemma

Let $a_1, ..., a_N$ be vectors from $\mathbf{R}^n$, and let a be another vector. Here we address the question: when a belongs to the cone spanned by the vectors $a_1, ..., a_N$, i.e., when a can be represented as a linear combination of a_i with *nonnegative* coefficients? A *necessary* condition is evident: if

$$a = \sum_{i=1}^{n} \lambda_i a_i \quad [\lambda_i \geq 0,\, i = 1, ..., N]$$

then every vector h which has nonnegative inner products with all a_i should also have nonnegative inner product with a:

$$a = \sum_i \lambda_i a_i \;\&\; \lambda_i \geq 0 \,\forall i \;\&\; h^T a_i \geq 0 \,\forall i \Rightarrow h^T a \geq 0.$$

The Homogeneous Farkas Lemma says that this evident necessary condition is also sufficient:

Lemma B.2.1 [Homogeneous Farkas Lemma] *Let $a, a_1, ..., a_N$ be vectors from $\mathbf{R}^n$. The vector a is a conic combination of the vectors a_i (linear combination with nonnegative coefficients) if and only if every vector h satisfying $h^T a_i \geq 0$, $i = 1, ..., N$, satisfies also $h^T a \geq 0$.*

Proof. The necessity – the "only if" part of the statement – was proved before the Farkas Lemma was formulated. Let us prove the "if" part of the Lemma. Thus, assume that every vector h satisfying $h^T a_i \geq 0 \;\forall i$ satisfies also $h^T a \geq 0$, and let us prove that a is a conic combination of the vectors a_i.

There is nothing to prove when $a = 0$ – the zero vector of course is a conic combination of the vectors a_i. Thus, from now on we assume that $a \neq 0$.

[1)] here is the proof of this Analysis theorem: assume, on contrary, that the compact sets Q_α, $\alpha \in \mathcal{A}$, have empty intersection. Choose a set Q_{α^*} from the family; for every $x \in Q_{\alpha^*}$ there is a set Q^x in the family which does not contain x - otherwise x would be a common point of all our sets. Since Q^x is closed, there is an open ball V_x centered at x which does not intersect Q^x. The balls V_x, $x \in Q_{\alpha^*}$, form an open covering of the compact set Q_{α^*}, and therefore there exists a finite subcovering $V_{x_1}, ..., V_{x_N}$ of Q_{α^*} by the balls from the covering. Since Q^{x_i} does not intersect V_{x_i}, we conclude that the intersection of the finite subfamily $Q_{\alpha^*}, Q^{x_1}, ..., Q^{x_N}$ is empty, which is a contradiction

1^0. Let

$$\Pi = \{h \mid a^T h = -1\},$$

and let

$$A_i = \{h \in \Pi \mid a_i^T h \geq 0\}.$$

Π is a hyperplane in $\mathbf{R}^n$, and every A_i is a polyhedral set contained in this hyperplane and is therefore convex.

2^0. What we know is that the intersection of all the sets A_i, $i = 1, ..., N$, is empty (since a vector h from the intersection would have nonnegative inner products with all a_i and the inner product -1 with a, and we are given that no such h exists). Let us choose the smallest, in the number of elements, of those sub-families of the family of sets $A_1, ..., A_N$ which still have empty intersection of their members; without loss of generality we may assume that this is the family $A_1, ..., A_k$. Thus, the intersection of all k sets $A_1, ..., A_k$ is empty, but the intersection of every $k-1$ sets from the family $A_1, ..., A_k$ is nonempty.

3^0. We claim that

(A) $a \in \mathrm{Lin}(\{a_1, ..., a_k\})$;

(B) The vectors $a_1, ..., a_k$ are linearly independent.

(A) is easy: assuming that $a \notin E = \mathrm{Lin}(\{a_1, ..., a_k\})$, we conclude that the orthogonal projection f of the vector a onto the orthogonal complement $E^\perp$ of E is nonzero. The inner product of f and a is the same as $f^T f$, is.e., is positive, while $f^T a_i = 0$, $i = 1, ..., k$. Taking $h = -(f^T f)^{-1} f$, we see that $h^T a = -1$ and $h^T a_i = 0$, $i = 1, ..., k$. In other words, h belongs to every set A_i, $i = 1, ..., k$, by definition of these sets, and therefore the intersection of the sets $A_1, ..., A_k$ is nonempty, which is a contradiction.

(B) is given by the Helley Theorem I. Indeed, assume that $a_1, ..., a_k$ are linearly dependent, and let us lead this assumption to a contradiction. Since $a_1, ..., a_k$ are linearly dependent, the dimension of $E = \mathrm{Lin}(\{a_1, ..., a_k\})$ is certain $m < k$. We already know from A. that $a \in E$. Now let $A_i' = A_i \cap E$. We claim that every $k-1$ of the sets A_i' have a nonempty intersection, while all k these sets have empty intersection. The second claim is evident – since the sets $A_1, ..., A_k$ have empty intersection, the same is the case with their parts A_i'. The first claim also is easily supported: let us take $k-1$ of the dashed sets, say, $A_1', ..., A_{k-1}'$. By construction, the intersection of $A_1, ..., A_{k-1}$ is nonempty; let h be a vector from this intersection, i.e., a vector with nonnegative inner products with $a_1, ..., a_{k-1}$ and the product -1 with a. When replacing h with its orthogonal projection h' on E, we do not vary all these inner products, since these are products with vectors from E; thus, h' also is a common point of $A_1, ..., A_{k-1}$, and since this is a point from E, it is a common point of the dashed sets $A_1', ..., A_{k-1}'$ as well.

Now we can complete the proof of (B): the sets $A_1', ..., A_k'$ are convex sets belonging to the *hyperplane* $\Pi' = \Pi \cap E = \{h \in E \mid a^T h = -1\}$ (Π' indeed is a hyperplane in E, since $0 \neq a \in E$) *in the m-dimensional linear subspace E*. Π' is an affine subspace of the affine dimension $\ell = \dim E - 1 = m - 1 < k - 1$ (recall that we are in the situation when $m = \dim E < k$), and every $\ell + 1 \leq k - 1$ subsets from the family A_1',...,A_k' have a nonempty intersection. From the Helley Theorem I (see Exercise B.9) it follows that all the sets $A_1', ..., A_k'$ have a point in common, which, as we know, is not the case. The contradiction we have got proves that $a_1, ..., a_k$ are linearly independent.

4^0. With (A) and (B) in our disposal, we can easily complete the proof of the "if" part of the Farkas Lemma. Specifically, by (A), we have

$$a = \sum_{i=1}^{k} \lambda_i a_i$$

with some real coefficients λ_i, and all we need is to prove that these coefficients are nonnegative. Assume, on the contrary, that, say, $\lambda_1 < 0$. Let us extend the (linearly independent in view of (B)) system of vectors $a_1, ..., a_k$ by vectors $f_1, ..., f_{n-k}$ to a basis in $\mathbf{R}^n$, and let $\xi_i(x)$ be the coordinates of a vector x in this basis. The function $\xi_1(x)$ is a linear form of x and therefore is the inner product with certain vector:

$$\xi_1(x) = f^T x \quad \forall x.$$

Now we have

$$f^T a = \xi_1(a) = \lambda_1 < 0$$

and

$$f^T a_i = \begin{cases} 1, & i = 1 \\ 0, & i = 2, ..., k \end{cases}$$

so that $f^T a_i \geq 0$, $i = 1, ..., k$. We conclude that a proper normalization of f – namely, the vector $|\lambda_1|^{-1} f$ – belongs to $A_1, ..., A_k$, which is the desired contradiction – by construction, this intersection is empty. ∎

B.2.5 Separation Theorem

B.2.5.A. Separation: definition

Recall that a *hyperplane* M in $\mathbf{R}^n$ is, by definition, an affine subspace of the dimension $n - 1$. By Proposition A.3.7, hyperplanes are exactly the same as level sets of nontrivial linear forms:

$$\begin{array}{c} M \subset \mathbf{R}^n \text{ is a hyperplane} \\ \Updownarrow \\ \exists a \in \mathbf{R}^n, b \in \mathbf{R}, a \neq 0 : \quad M = \{x \in \mathbf{R}^n \mid a^T x = b\} \end{array}$$

We can, consequently, associate with the hyperplane (or, better to say, with the associated linear form a; this form is defined uniquely, up to multiplication by a nonzero real) the following sets:

- "upper" and "lower" open half-spaces $M^{++} = \{x \in \mathbf{R}^n \mid a^T x > b\}$, $M^{--} = \{x \in \mathbf{R}^n \mid a^T x < b\}$;

  these sets clearly are convex, and since a linear form is continuous, and the sets are given by strict inequalities on the value of a continuous function, they indeed are open.

  Note that since a is uniquely defined by M, up to multiplication by a nonzero real, these open half-spaces are uniquely defined by the hyperplane, up to swapping the "upper" and the "lower" ones (which half-space is "upper", it depends on the particular choice of a);

- "upper" and "lower" closed half-spaces $M^{+} = \{x \in \mathbf{R}^n \mid a^T x \geq b\}$, $M^{-} = \{x \in \mathbf{R}^n \mid a^T x \leq b\}$;

  these are also convex sets, now closed (since they are given by non-strict inequalities on the value of a continuous function). It is easily seen that the closed upper/lower half-space is the closure of the corresponding open half-space, and M itself is the boundary (i.e., the complement of the interior to the closure) of all four half-spaces.

It is clear that our half-spaces and M itself partition $\mathbf{R}^n$:

$$\mathbf{R}^n = M^{--} \cup M \cup M^{++}$$

(partitioning by disjoint sets),

$$\mathbf{R}^n = M^{-} \cup M^{+}$$

(M is the intersection of the right hand side sets).

Now we define the basic notion of *separation* of two convex sets T and S by a hyperplane.

Definition B.2.1 [separation] *Let S, T be two nonempty convex sets in $\mathbf{R}^n$.*

- *A hyperplane*
$$M = \{x \in \mathbf{R}^n \mid a^T x = b\} \quad [a \neq 0]$$
is said to separate S and T, if, first,
$$S \subset \{x : a^T x \leq b\}, \quad T \subset \{x : a^T x \geq b\}$$
(i.e., S and T belong to the opposite closed half-spaces into which M splits $\mathbf{R}^n$), and, second, at least one of the sets S, T is not contained in M itself:
$$S \cup T \not\subset M.$$
The separation is called strong, if there exist b', b'', $b' < b < b''$, such that
$$S \subset \{x : a^T x \leq b'\}, \quad T \subset \{x : a^T x \geq b''\}.$$

- *A linear form $a \neq 0$ is said to separate (strongly separate) S and T, if for properly chosen b the hyperplane $\{x : a^T x = b\}$ separates (strongly separates) S and T.*
- *We say that S and T can be (strongly) separated, if there exists a hyperplane which (strongly) separates S and T.*

E.g.,

- the hyperplane $\{x : a^T x \equiv x_2 - x_1 = 1\}$ in $\mathbf{R}^2$ strongly separates convex polyhedral sets $T = \{x \in \mathbf{R}^2 : 0 \leq x_1 \leq 1, 3 \leq x_2 \leq 5\}$ and $S = \{x \in \mathbf{R}^2 : x_2 = 0; x_1 \geq -1\}$;
- the hyperplane $\{x : a^T x \equiv x = 1\}$ in $\mathbf{R}^1$ separates (but not strongly separates) the convex sets $S = \{x \leq 1\}$ and $T = \{x \geq 1\}$;
- the hyperplane $\{x : a^T x \equiv x_1 = 0\}$ in $\mathbf{R}^2$ separates (but not strongly separates) the sets $S = \{x \in \mathbf{R}^2 :, x_1 < 0, x_2 \geq -1/x_1\}$ and $T = \{x \in \mathbf{R}^2 : x_1 > 0, x_2 > 1/x_1\}$;
- the hyperplane $\{x : a^T x \equiv x_2 - x_1 = 1\}$ in $\mathbf{R}^2$ does *not* separate the convex sets $S = \{x \in \mathbf{R}^2 : x_2 \geq 1\}$ and $T = \{x \in \mathbf{R}^2 : x_2 = 0\}$;
- the hyperplane $\{x : a^T x \equiv x_2 = 0\}$ in $\mathbf{R}^2$ does not separate the sets $S = \{x \in \mathbf{R}^2 : x_2 = 0, x_1 \leq -1\}$ and $T = \{x \in \mathbf{R}^2 : x_2 = 0, x_1 \geq 1\}$.

The following Exercise presents an equivalent description of separation:

Exercise B.10 *Let S, T be nonempty convex sets in $\mathbf{R}^n$. Prove that a linear form a separates S and T if and only if*
$$\sup_{x \in S} a^T x \leq \inf_{y \in T} a^T y$$
and
$$\inf_{x \in S} a^T x < \sup_{y \in T} a^T y.$$
This separation is strong if and only if
$$\sup_{x \in S} a^T x < \inf_{y \in T} a^T y.$$

Exercise B.11 *Whether the sets $S = \{x \in \mathbf{R}^2 : x_1 > 0, x_2 \geq 1/x_1\}$ and $T = \{x \in \mathbf{R}^2 : x_1 < 0, x_2 \geq -1/x_1\}$ can be separated? Whether they can be strongly separated?*

B.2.5.B. Separation Theorem

Theorem B.2.5 [Separation Theorem] *Let S and T be nonempty convex sets in $\mathbf{R}^n$.*

(i) *S and T can be separated if and only if their relative interiors do not intersect:* $\mathrm{ri}\, S \cap \mathrm{ri}\, T = \emptyset$.

(ii) *S and T can be strongly separated if and only if the sets are at a positive distance from each other:*

$$\mathrm{dist}(S,T) \equiv \inf\{\|x-y\|_2 : x \in S, y \in T\} > 0.$$

In particular, if S, T are closed nonempty non-intersecting convex sets and one of these sets is compact, S and T can be strongly separated.

Proof takes several steps.

(i), Necessity. Assume that S, T can be separated, so that for certain $a \neq 0$ we have

$$\inf_{x\in S} a^T x \leq \inf_{y\in T} a^T y; \quad \inf_{x\in S} a^T x < \sup_{y\in T} a^T y. \tag{B.2.1}$$

We should lead to a contradiction the assumption that $\mathrm{ri}\, S$ and $\mathrm{ri}\, T$ have in common certain point $\bar{x}$. Assume that it is the case; then from the first inequality in (B.2.1) it is clear that $\bar{x}$ maximizes the linear function $f(x) = a^T x$ on S and simultaneously minimizes this function on T. Now, we have the following simple and important

> **Lemma B.2.2** *A linear function $f(x) = a^T x$ can attain its maximum/minimum over a convex set Q at a point $x \in \mathrm{ri}\, Q$ if and only if the function is constant on Q.*
>
> **Proof.** "if" part is evident. To prove the "only if" part, let $\bar{x} \in \mathrm{ri}\, Q$ be, say, a minimizer of f over Q and y be an arbitrary point of Q; we should prove that $f(\bar{x}) = f(y)$. There is nothing to prove if $y = \bar{x}$, so let us assume that $y \neq \bar{x}$. Since $\bar{x} \in \mathrm{ri}\, Q$, the segment $[y, \bar{x}]$, which is contained in M, can be extended a little bit through the point $\bar{x}$, not leaving M (since $\bar{x} \in \mathrm{ri}\, Q$), so that there exists $z \in Q$ such that $\bar{x} \in [y, z)$, i.e., $\bar{x} = (1-\lambda)y + \lambda z$ with certain $\lambda \in (0, 1]$; since $y \neq \bar{x}$, we have in fact $\lambda \in (0, 1)$. Since f is linear, we have
>
> $$f(\bar{x}) = (1-\lambda)f(y) + \lambda f(z);$$
>
> since $f(\bar{x}) \leq \min\{f(y), f(z)\}$ and $0 < \lambda < 1$, this relation can be satisfied only when $f(\bar{x}) = f(y) = f(z)$. □

By Lemma B.2.2, $f(x) = f(\bar{x})$ on S and on T, so that $f(\cdot)$ is constant on $S \cup T$, which yields the desired contradiction with the second inequality in (B.2.1). □

(i), Sufficiency. The proof of sufficiency part of the Separation Theorem is much more instructive. There are several ways to prove it, and I choose the one which goes via the Homogeneous Farkas Lemma B.2.1, which is extremely important in its own right.

(i), Sufficiency, Step 1: Separation of a convex polytope and a point outside the polytope. Let us start with seemingly very particular case of the Separation Theorem – the one where S is the convex full points $x_1, ..., x_N$, and T is a singleton $T = \{x\}$ which does not belong to S. We intend to prove that in this case there exists a linear form which separates x and S; in fact we shall prove even the existence of strong separation.

Let us associate with n-dimensional vectors $x_1, ..., x_N, x$ the $(n+1)$-dimensional vectors $a = \begin{pmatrix} x \\ 1 \end{pmatrix}$ and $a_i = \begin{pmatrix} x_i \\ 1 \end{pmatrix}$, $i = 1, ..., N$. I claim that a does not belong to the conic hull of $a_1, ..., a_N$. Indeed, if a would be representable as a linear combination of $a_1, ..., a_N$ with nonnegative coefficients, then, looking at the last, $(n+1)$-st, coordinates in such a representation, we would conclude that the sum of coefficients

should be 1, so that the representation, actually, represents x as a convex combination of $x_1, ..., x_N$, which was assumed to be impossible.

Since a does not belong to the conic hull of $a_1, ..., a_N$, by the Homogeneous Farkas Lemma (Lemma B.2.1) there exists a vector $h = \begin{pmatrix} f \\ \alpha \end{pmatrix} \in \mathbf{R}^{n+1}$ which "separates" a and $a_1, ..., a_N$ in the sense that

$$h^T a > 0, \ h^T a_i \leq 0, \ i = 1, ..., N,$$

whence, of course,

$$h^T a > \max_i h^T a_i.$$

Since the components in all the inner products $h^T a, h^T a_i$ coming from the $(n+1)$-st coordinates are equal to each other , we conclude that the n-dimensional component f of h separates x and $x_1, ..., x_N$:

$$f^T x > \max_i f^T x_i.$$

Since for every convex combination $y = \sum_i \lambda_i x_i$ of the points x_i one clearly has $f^T y \leq \max_i f^T x_i$, we conclude, finally, that

$$f^T x > \max_{y \in \mathrm{Conv}(\{x_1, ..., x_N\})} f^T y,$$

so that f strongly separates $T = \{x\}$ and $S = \mathrm{Conv}(\{x_1, ..., x_N\})$. □

(i), Sufficiency, Step 2: Separation of a convex set and a point outside of the set. Now consider the case when S is an arbitrary nonempty convex set and $T = \{x\}$ is a singleton outside S (the difference with Step 1 is that now S is not assumed to be a polytope).

First of all, without loss of generality we may assume that S contains 0 (if it is not the case, we may subject S and T to translation $S \mapsto p + S$, $T \mapsto p + T$ with $p \in -S$). Let L be the linear span of S. If $x \notin L$, the separation is easy: taking as f the orthogonal to L component of x, we shall get

$$f^T x = f^T f > 0 = \max_{y \in S} f^T y,$$

so that f strongly separates S and $T = \{x\}$.

It remains to consider the case when $x \in L$. Since $S \subset L$, $x \in L$ and $x \notin S$, L is a nonzero linear subspace; w.l.o.g., we can assume that $L = \mathbf{R}^n$.

Let $\Sigma = \{h : \|h\|_2 = 1\}$ be the unit sphere in $L = \mathbf{R}^n$. This is a closed and bounded set in $\mathbf{R}^n$ (boundedness is evident, and closedness follows from the fact that $\|\cdot\|_2$ is continuous). Consequently, Σ is a compact set. Let us prove that there exists $f \in \Sigma$ which separates x and S in the sense that

$$f^T x \geq \sup_{y \in S} f^T y. \tag{B.2.2}$$

Assume, on the contrary, that no such f exists, and let us lead this assumption to a contradiction. Under our assumption for every $h \in \Sigma$ there exists $y_h \in S$ such that

$$h^T y_h > h^T x.$$

Since the inequality is strict, it immediately follows that there exists a neighbourhood U_h of the vector h such that

$$(h')^T y_h > (h')^T x \quad \forall h' \in U_h. \tag{B.2.3}$$

The family of open sets $\{U_h\}_{h \in \Sigma}$ covers Σ; since Σ is compact, we can find a finite subfamily $U_{h_1}, ..., U_{h_N}$ of the family which still covers Σ. Let us take the corresponding points $y_1 = y_{h_1}, y_2 = y_{h_2}, ..., y_N = y_{h_N}$ and the polytope $S' = \mathrm{Conv}(\{y_1, ..., y_N\})$ spanned by the points. Due to the origin of y_i, all of them are points from S; since S is convex, the polytope S' is contained in S and, consequently, does not contain x. By Step 1, x can be strongly separated from S': there exists a such that

$$a^T x > \sup_{y \in S'} a^T y. \tag{B.2.4}$$

By normalization, we may also assume that $\|a\|_2 = 1$, so that $a \in \Sigma$. Now we get a contradiction: since $a \in \Sigma$ and $U_{h_1}, ..., U_{h_N}$ form a covering of Σ, a belongs to certain U_{h_i}. By construction of U_{h_i} (see (B.2.3)), we have

$$a^T y_i \equiv a^T y_{h_i} > a^T x,$$

which contradicts (B.2.4) – recall that $y_i \in S'$.

The contradiction we get proves that there exists $f \in \Sigma$ satisfying (B.2.2). We claim that f separates S and $\{x\}$; in view of (B.2.2), all we need to verify our claim is to show that the linear form $f(y) = f^T y$ is non-constant on $S \cup T$, which is evident: we are in the situation when $0 \in S$ and $L \equiv \mathrm{Lin}(S) = \mathbf{R}^n$ and $f \neq 0$, so that $f(y)$ is non-constant already on S. □

Mathematically oriented reader should take into account that the simple-looking reasoning underlying Step 2 in fact brings us into a completely new world. Indeed, the considerations at Step 1 and in the proof of Homogeneous Farkas Lemma are "pure arithmetic" – we never used things like convergence, compactness, etc., and used rational arithmetic only – no square roots, etc. It means that the Homogeneous Farkas Lemma and the result stated a Step 1 remain valid if we, e.g., replace our universe $\mathbf{R}^n$ with the space $\mathbf{Q}^n$ of n-dimensional rational vectors (those with rational coordinates; of course, the multiplication by reals in this space should be restricted to multiplication by rationals). The "rational" Farkas Lemma or the possibility to separate a rational vector from a "rational" polytope by a rational linear form, which is the "rational" version of the result of Step 1, definitely are of interest (e.g., for Integer Programming). In contrast to these "purely arithmetic" considerations, at Step 2 we used compactness – something heavily exploiting the fact that our universe is $\mathbf{R}^n$ and not, say, $\mathbf{Q}^n$ (in the latter space bounded and closed sets not necessary are compact). Note also that we could not avoid things like compactness arguments at Step 2, since the very fact we are proving is true in $\mathbf{R}^n$ but not in $\mathbf{Q}^n$. Indeed, consider the "rational plane" – the universe comprised of all 2-dimensional vectors with rational entries, and let S be the half-plane in this rational plane given by the linear inequality

$$x_1 + \alpha x_2 \leq 0,$$

where α is irrational. S clearly is a "convex set" in $\mathbf{Q}^2$; it is immediately seen that a point outside this set cannot be separated from S by a rational linear form.

(i), Sufficiency, Step 3: Separation of two nonempty and non-intersecting convex sets. Now we are ready to prove that two nonempty and non-intersecting convex sets S and T can be separated. To this end consider the arithmetic difference

$$\Delta = S - T = \{x - y \mid x \in S, y \in T\}.$$

By Proposition B.1.6.3, Δ is convex (and, of course, nonempty) set; since S and T do not intersect, Δ does not contain 0. By Step 2, we can separate Δ and $\{0\}$: there exists $f \neq 0$ such that

$$f^T 0 = 0 \geq \sup_{z \in \Delta} f^T z \ \& \ f^T 0 > \inf_{z \in \Delta} f^T z.$$

In other words,

$$0 \geq \sup_{x \in S, y \in T} [f^T x - f^T y] \ \& \ 0 > \inf_{x \in S, y \in T} [f^T x - f^T y],$$

which clearly means that f separates S and T. ■

(i), Sufficiency, Step 4: Separation of nonempty convex sets with non-intersecting relative interiors. Now we are able to complete the proof of the "if" part of the Separation Theorem. Let S and T be two nonempty convex sets with non-intersecting relative interiors; we should prove that S and T can be properly separated. This is immediate: as we know from Theorem B.1.1, the sets $S' = \mathrm{ri}\, S$ and $T' = \mathrm{ri}\, T$ are nonempty and convex; since we are given that they do not intersect, they can be separated by Step 3: there exists f such that

$$\inf_{x \in T'} f^T x \geq \sup_{y \in S'} f^T x \ \& \ \sup_{x \in T'} f^T x > \inf_{y \in S'} f^T x. \tag{B.2.5}$$

It is immediately seen that in fact f separates S and T. Indeed, the quantities in the left and the right hand sides of the first inequality in (B.2.5) clearly remain unchanged when we replace S' with $\mathrm{cl}\, S'$ and T' with $\mathrm{cl}\, T'$; by Theorem B.1.1, $\mathrm{cl}\, S' = \mathrm{cl}\, S \supset S$ and $\mathrm{cl}\, T' = \mathrm{cl}\, T \supset T$, and we get $\inf_{x\in T} f^T x = \inf_{x\in T'} f^T x$, and similarly $\sup_{y\in S} f^T y = \sup_{y\in S'} f^T y$. Thus, we get from (B.2.5)

$$\inf_{x\in T} f^T x \geq \sup_{y\in S} f^T y.$$

It remains to note that $T' \subset T$, $S' \subset S$, so that the second inequality in (B.2.5) implies that

$$\sup_{x\in T} f^T x > \inf_{y\in S} f^T x. \quad \square$$

(ii), Necessity: prove yourself.

(ii), Sufficiency: Assuming that $\rho \equiv \inf\{\|x-y\|_2 : x \in S, y \in T\} > 0$, consider the sets $S' = \{x : \inf_{y\in S} \|x-y\|_2 \leq \rho\}$. Note that S' is convex along with S (Example B.1.4) and that $S' \cap T = \emptyset$ (why?) By (i), S' and T can be separated, and if f is a linear form which separates S' and T, then the same form strongly separates S and T (why?). The "in particular" part of (ii) readily follows from the just proved statement due to the fact that if two closed nonempty sets in $\mathbf{R}^n$ do not intersect and one of them is compact, then the sets are at positive distance from each other (why?). ∎

Exercise B.12 *Derive the statement in Remark B.1.1 from the Separation Theorem.*

Exercise B.13 *Implement the following alternative approach to the proof of Separation Theorem:*

1. *Prove that if x is a point in $\mathbf{R}^n$ and S is a nonempty closed convex set in $\mathbf{R}^n$, then the problem*

$$\min_y \{\|x-y\|_2 : y \in S\}$$

has a unique optimal solution $\bar{x}$.

2. *In the situation of 1), prove that if $x \notin S$, then the linear form $e = x - \bar{x}$ strongly separates $\{x\}$ and S:*

$$\max_{y\in S} e^T y = e^T \bar{x} = e^T x - e^T e < e^T x,$$

thus getting a direct proof of the possibility to separate strongly a nonempty closed convex set and a point outside this set.

3. *Derive from 2) the Separation Theorem.*

B.2.5.C. Supporting hyperplanes

By the Separation Theorem, a closed and nonempty convex set M is the intersection of all closed half-spaces containing M. Among these half-spaces, the most interesting are the "extreme" ones – those with the boundary hyperplane touching M. The notion makes sense for an arbitrary (not necessary closed) convex set, but we shall use it for closed sets only, and include the requirement of closedness in the definition:

Definition B.2.2 [Supporting plane] *Let M be a convex closed set in $\mathbf{R}^n$, and let x be a point from the relative boundary of M. A hyperplane*

$$\Pi = \{y \mid a^T y = a^T x\} \quad [a \neq 0]$$

is called supporting to M at x, if it separates M and $\{x\}$, i.e., if

$$a^T x \geq \sup_{y\in M} a^T y \quad \& \quad a^T x > \inf_{y\in M} a^T y. \tag{B.2.6}$$

Note that since x is a point from the relative boundary of M and therefore belongs to $\mathrm{cl}\, M = M$, the first inequality in (B.2.6) in fact is equality. Thus, an equivalent definition of a supporting plane is as follows:

> *Let M be a closed convex set and x be a relative boundary point of M. The hyperplane $\{y \mid a^T y = a^T x\}$ is called supporting to M at x, if the linear form $a(y) = a^T y$ attains its maximum on M at the point x and is nonconstant on M.*

E.g., the hyperplane $\{x_1 = 1\}$ in $\mathbf{R}^n$ clearly is supporting to the unit Euclidean ball $\{x \mid |x| \le 1\}$ at the point $x = e_1 = (1, 0, ..., 0)$.

The most important property of a supporting plane is its existence:

Proposition B.2.1 [Existence of supporting hyperplane] *Let M be a convex closed set in $\mathbf{R}^n$ and x be a point from the relative boundary of M. Then*

(i) *There exists at least one hyperplane which is supporting to M at x;*

(ii) *If Π is supporting to M at x, then the intersection $M \cap \Pi$ is of affine dimension less than the one of M (recall that the affine dimension of a set is, by definition, the affine dimension of the affine hull of the set).*

Proof. (i) is easy: if x is a point from the relative boundary of M, then it is outside the relative interior of M and therefore $\{x\}$ and $\mathrm{ri}\, M$ can be separated by the Separation Theorem; the separating hyperplane is exactly the desired supporting to M at x hyperplane.

To prove (ii), note that if $\Pi = \{y \mid a^T y = a^T x\}$ is supporting to M at $x \in \partial_{\mathrm{ri}} M$, then the set $M' = M \cap \Pi$ is a nonempty (it contains x) convex set, and the linear form $a^T y$ is constant on M' and therefore (why?) on $\mathrm{Aff}(M')$. At the same time, the form is nonconstant on M by definition of a supporting plane. Thus, $\mathrm{Aff}(M')$ is a proper (less than the entire $\mathrm{Aff}(M)$) subset of $\mathrm{Aff}(M)$, and therefore the affine dimension of $\mathrm{Aff}(M')$ (i.e., the affine dimension of M') is less than the affine dimension of $\mathrm{Aff}(M)$ (i.e., than the affine dimension of M). [2]. ∎

B.2.6 Polar of a convex set and Milutin-Dubovitski Lemma

B.2.6.A. Polar of a convex set

Let M be a nonempty convex set in $\mathbf{R}^n$. The *polar* $\mathrm{Polar}\,(M)$ of M is the set of all linear forms which do not exceed 1 on M, i.e., the set of all vectors a such that $a^T x \le 1$ for all $x \in M$:

$$\mathrm{Polar}\,(M) = \{a : a^T x \le 1 \forall x \in M\}.$$

For example, $\mathrm{Polar}\,(\mathbf{R}^n) = \{0\}$, $\mathrm{Polar}\,(\{0\}) = \mathbf{R}^n$; if L is a liner subspace in $\mathbf{R}^n$, then $\mathrm{Polar}\,(L) = L^\perp$ (why?).

The following properties of the polar are evident:

1. $0 \in \mathrm{Polar}\,(M)$;
2. $\mathrm{Polar}\,(M)$ is convex;
3. $\mathrm{Polar}\,(M)$ is closed.

It turns out that these properties characterize polars:

Proposition B.2.2 *Every closed convex set M containing the origin is polar, specifically, it is polar of its polar:*

$$\begin{array}{c} M \text{ is closed and convex, } 0 \in M \\ \Updownarrow \\ M = \mathrm{Polar}\,(\mathrm{Polar}\,(M)) \end{array}$$

[2] In the latter reasoning we used the following fact: if $P \subset Q$ are two affine subspaces, then the affine dimension of P is $\le$ the one of Q, with $\le$ being $=$ if and only if $P = Q$. Please prove this fact

Proof. All we need is to prove that if M is closed and convex and $0 \in M$, then $M = \mathrm{Polar}\,(\mathrm{Polar}\,(M))$. By definition,

$$y \in \mathrm{Polar}\,(M), x \in M \Rightarrow y^T x \leq 1,$$

so that $M \subset \mathrm{Polar}\,(\mathrm{Polar}\,(M))$. To prove that this inclusion is in fact equality, assume, on the contrary, that there exists $\bar{x} \in \mathrm{Polar}\,(\mathrm{Polar}\,(M))\backslash M$. Since M is nonempty, convex and closed and $\bar{x} \notin M$, the point $\bar{x}$ can be strongly separated from M (Separation Theorem, (ii)). Thus, for appropriate b one has

$$b^T \bar{x} > \sup_{x \in M} b^T x.$$

Since $0 \in M$, the left hand side quantity in this inequality is positive; passing from b to a proportional vector $a = \lambda b$ with appropriately chosen positive λ, we may ensure that

$$a^T \bar{x} > 1 \geq \sup_{x \in M} a^T x.$$

This is the desired contradiction, since the relation $1 \geq \sup_{x \in M} a^T x$ implies that $a \in \mathrm{Polar}\,(M)$, so that the relation $a^T \bar{x} > 1$ contradicts the assumption that $\bar{x} \in \mathrm{Polar}\,(\mathrm{Polar}\,(M))$. ∎

Exercise B.14 *Let M be a convex set containing the origin, and let M' be the polar of M. Prove the following facts:*

1. $\mathrm{Polar}\,(M) = \mathrm{Polar}\,(\mathrm{cl}\,M)$*;*
2. *M is bounded if and only if $0 \in \mathrm{int}\,M'$;*
3. *$\mathrm{int}\,M \neq \emptyset$ if and only if the only vector $h \in M'$ such that $\pm h \in M'$ is the zero vector;*
4. *M is a closed cone of and only if M' is a closed cone. If M is a cone (not necessarily closed), then*

$$M' = \{a : a^T x \leq 0 \forall x \in M\}. \tag{B.2.7}$$

B.2.6.B. Dual cone

Let $M \subset \mathbf{R}^n$ be a cone. By Exercise B.14.4, the polar M' of M is a closed cone given by (B.2.7). The set $M_* = -M'$ (which also is a closed cone), that is, the set

$$M_* = \{a : a^T x \geq 0 \forall x \in M\}$$

of all vectors which have nonnegative inner products with all vectors from M, is called the cone *dual* to M. By Proposition B.2.2 and Exercise B.14.4, the family of closed cones in $\mathbf{R}^n$ is closed with respect to passing to a dual cone, and the duality is symmetric: for a closed cone M, M_* also is a closed cone, and $(M_*)_* = M$.

Exercise B.15 *Let M be a closed cone in $\mathbf{R}^n$, and M_* be its dual cone. Prove that*

1. *M is pointed (i.e., does not contain lines) if and only M_* has a nonempty interior. Derive from this fact that M is a closed pointed cone with a nonempty interior if and only if the dual cone has the same properties.*
2. *Prove that $a \in \mathrm{int}\,M_*$ if and only if $a^T x > 0$ for all nonzero vectors $x \in M$.*

B.2.6.C. Dubovitski-Milutin Lemma

Let $M_1, ..., M_k$ be cones (not necessarily closed), and M be their intersection; of course, M also is a cone. How to compute the cone dual to M?

Proposition B.2.3 *Let $M_1, ..., M_k$ be cones. The cone M' dual to the intersection M of the cones M_1,...,M_k contains the arithmetic sum $\widetilde{M}$ of the cones M'_1,...,M'_k dual to M_1,...,M_k. If all the cones $M_1, ..., M_k$ are closed, then M' is equal to $\mathrm{cl}\,\widetilde{M}$. In particular, for closed cones M_1,...,M_k, M' coincides with $\widetilde{M}$ if and only if the latter set is closed.*

Proof. Whenever $a_i \in M'_i$ and $x \in M$, we have $a_i^T x \geq 0$, $i = 1, ..., k$, whence $(a_1 + ... + a_k)^T x \geq 0$. Since the latter relation is valid for all $x \in M$, we conclude that $a_1 + ... + a_k \in M'$. Thus, $\widetilde{M} \subseteq M'$.

Now assume that the cones $M_1, ..., M_k$ are closed, and let us prove that $M = \mathrm{cl}\,\widetilde{M}$. Since M' is closed and we have seen that $\widetilde{M} \in M'$, all we should prove is that if $a \in M'$, then $a \in \widehat{M} = \mathrm{cl}\,\widetilde{M}$ as well. Assume, on the contrary, that $a \in M' \backslash \widehat{M}$. Since the set $\widetilde{M}$ clearly is a cone, its closure $\widehat{M}$ is a closed cone; by assumption, a does not belong to this closed cone and therefore, by Separation Theorem (ii), a can be strongly separated from $\widehat{M}$ and therefore – from $\widetilde{M} \subset \widehat{M}$. Thus, for some x one has

$$a^T x < \inf_{b \in \widetilde{M}} b^T x = \inf_{a_i \in M'_i, i=1,...,k} (a_1 + ... + a_k)^T x = \sum_{i=1}^{k} \inf_{a_i \in M'_i} a_i^T x. \tag{B.2.8}$$

From the resulting inequality it follows that $\inf\limits_{a_i \in M'_i} a_i^T x > -\infty$; since M'_i is a cone, the latter is possible if and only if $\inf\limits_{a_i \in M'_i} a_i^T x = 0$, i.e., if and only if for every i one has $x \in \mathrm{Polar}\,(M'_i) = M_i$ (recall that the cones M_i are closed). Thus, $x \in M_i$ for all i, and the concluding quantity in (B.2.8) is 0. We see that $x \in M = \cup_i M_i$, and that (B.2.8) reduces to $a^T x < 0$. This contradicts the inclusion $a \in M'$. ∎

Note that in general $\widetilde{M}$ can be non-closed even when all the cones $M_1, ..., M_k$ are closed. Indeed, take $k = 2$, and let M'_1 be the ice-cream cone $\{(x, y, z) \in \mathbf{R}^3 : z \geq \sqrt{x^2 + y^2}\}$. The orthogonal projection of this cone on a 2D plane Π tangent to M_1 at a nonzero point of the boundary of M_1 clearly is non-closed: it is an interior of an appropriate half-plane with the boundary passing through the origin plus the origin itself (that is, in appropriate coordinates (u, v) on the plane the projection is the non-closed cone $K = \{(u, v) : u > 0\} \cup \{(0, 0)\}$). We conclude that the sum of M'_1 and the ray M'_2 emanating from the origin and orthogonal to P (this ray is a closed cone) is, in appropriate coordinates, the set $K^+ = \{(u, v, w) \in \mathbf{R}^3 : u > 0, w \geq 0\} \cup \{(0, 0, w) : w \geq 0\}$, which is a non-closed cone.

Dubovistki-Milutin Lemma presents a simple sufficient condition for $\widetilde{M}$ to be closed and thus to coincide with M':

Proposition B.2.4 [Dubovistki-Milutin Lemma] *Let $M_1, ..., M_k$ be cones such that M_k is closed and the set $M_k \cap \mathrm{int}\, M_1 \cap \mathrm{int}\, M_2 \cap ... \cap \mathrm{int}\, M_{k-1}$ is nonempty, and let $M = M_1 \cap ... \cap M_k$. Let also M'_i be the cones dual to M_i. Then*

(i) $\mathrm{cl}\, M = \bigcap\limits_{i=1}^{k} \mathrm{cl}\, M_i$;

(ii) *the cone $\widetilde{M} = M'_1 + ... + M'_k$ is closed, and thus coincides with the cone M' dual to $\mathrm{cl}\, M$ (or, which is the same by Exercise B.14.1, with the cone dual to M). In other words, every linear form which is nonnegative on M can be represented as a sum of k linear forms which are nonnegative on the respective cones M_1,...,M_k.*

Proof. (i): We should prove that under the premise of the Dobovitski-Milutin Lemma, $\mathrm{cl}\, M = \bigcap\limits_i \mathrm{cl}\, M_i$. The right hand side here contains M and is closed, so that all we should prove is that every point x in $\bigcap\limits_{i=1}^{k} \mathrm{cl}\, M_i$ is the limit of an appropriate sequence $x_t \in M$. By premise of the Lemma, there exists a point $\bar{x} \in M_k \cap \mathrm{int}\, M_1 \cap \mathrm{int}\, M_2 \cap ... \cap \mathrm{int}\, M_{k-1}$; setting $x_t = t^{-1}\bar{x} + (1 - t^{-1})x$, we get a sequence converging to x as $t \to \infty$; at the same time, $x_t \in M_k$ (since x, $\bar{x}$ are in $\mathrm{cl}\, M_k$, and the latter set is closed) and $x_t \in M_i$ for every $i < k$ (by Lemma B.1.1; note that for $i < k$ one has $\bar{x} \in \mathrm{int}\, M_i$ and $x \in \mathrm{cl}\, M_i$). Thus, every point $x \in \bigcap\limits_{i=1}^{k} \mathrm{cl}\, M_i$ is the limit of a sequence from M. □

(ii): Under the premise of the Lemma, when replacing the cones $M_1, ..., M_k$ with their closures, we do not vary the polars M'_i of the cones (and thus do not vary $\widetilde{M}$) and replace the intersection of the sets

$M_1, ..., M_k$ with its closure (by (i)), thus not varying the polar of the intersection. And of course when replacing the cones $M_1, ..., M_k$ with their closures, we preserve the premise of Lemma. Thus, we lose nothing when assuming, in addition to the premise of Lemma, that the cones $M_1, ..., M_k$ are closed. To prove the lemma for closed cones M_1,...,M_k, we use induction in $k \geq 2$.

Base $k = 2$: Let a sequence $\{f_t + g_t\}_{t=1}^{\infty}$ with $f_t \in M_1'$ and $g_t \in M_2'$ converge to certain h; we should prove that $h = f + g$ for appropriate $f \in M_1'$ and $g \in M_2'$. To achieve our goal, it suffices to verify that for an appropriate subsequence t_j of indices there exists $f \equiv \lim\limits_{j\to\infty} f_{t_j}$. Indeed, if this is the case, then $g = \lim\limits_{j\to\infty} g_{t_j}$ also exists (since $f_t + g_t \to h$ as $t \to \infty$ and $f + g = h$; besides this, $f \in M_1'$ and $g \in M_2'$, since both the cones in question are closed. In order to verify the existence of the desired subsequence, it suffices to lead to a contradiction the assumption that $\|f_t\|_2 \to \infty$ as $t \to \infty$. Let the latter assumption be true. Passing to a subsequence, we may assume that the unit vectors $\phi_t = f_t/\|f_t\|_2$ have a limit ϕ as $t \to \infty$; since M_1' is a closed cone, ϕ is a unit vector from M_1'. Now, since $f_t + g_t \to h$ as $t \to \infty$, we have $\phi = \lim\limits_{t\to\infty} f_t/\|f_t\|_2 = -\lim\limits_{t\to\infty} g_t/\|f_t\|_2$ (recall that $\|f_t\|_2 \to \infty$ as $t \to \infty$, whence $h/\|f_t\|_2 \to 0$ as $t \to \infty$). We see that the vector $-\phi$ belongs to M_2'. Now, by assumption M_2 intersects the interior of the cone M_1; let $\bar{x}$ be a point in this intersection. We have $\phi^T\bar{x} \geq 0$ (since $\bar{x} \in M_1$ and $\phi \in M_1'$) and $\phi^T\bar{x} \leq 0$ (since $-\phi \in M_2'$ and $\bar{x} \in M_2$). We conclude that $\phi^T\bar{x} = 0$, which is contradicts the facts that $0 \neq \phi \in M_1'$ and $\bar{x} \in \operatorname{int} M_1$ (see Exercise B.15.2). □

Inductive step: Assume that the statement we are proving is valid in the case of $k-1 \geq 2$ cones, and let M_1,...,M_k be k cones satisfying the premise of the Dubovitski-Milutin Lemma. By this premise, the come $M^1 = M_1 \cup ... \cup M_{k-1}$ has a nonempty interior, and M_k intersects this interior. Applying to the pair of cones M, M_k the already proved 2-cone version of the Lemma, we see that the set $(M^1)' + M_k'$ is closed; here $(M^1)'$ is the cone dual to M^1. Further, the cones $M_1, ..., M_{k-1}$ satisfy the premise of the $(k-1)$-cone version of the Lemma; by inductive hypothesis, the set $M_1' + ... + M_{k-1}'$ is closed and therefore, by Proposition B.2.3, equals to $(M^1)'$. Thus, $M_1' + ... + M_k' = (M^1)' + M_k'$, and we have seen that the latter set is closed. ■

B.2.7 Extreme points and Krein-Milman Theorem

Supporting planes are useful tool to prove existence of *extreme points* of convex sets. Geometrically, an extreme point of a convex set M is a point in M which cannot be obtained as a convex combination of other points of the set; and the importance of the notion comes from the fact (which we shall prove in the mean time) that the set of all extreme points of a "good enough" convex set M is the "shortest worker's instruction for building the set" – this is the smallest set of points for which M is the convex hull.

B.2.7.A. Extreme points: definition

The exact definition of an extreme point is as follows:

Definition B.2.3 [extreme points] *Let M be a nonempty convex set in $\mathbf{R}^n$. A point $x \in M$ is called an extreme point of M, if there is no nontrivial (of positive length) segment $[u, v] \in M$ for which x is an interior point, i.e., if the relation*

$$x = \lambda u + (1 - \lambda)v$$

with $\lambda \in (0, 1)$ and $u, v \in M$ is valid if and only if

$$u = v = x.$$

E.g., the extreme points of a segment are exactly its endpoints; the extreme points of a triangle are its vertices; the extreme points of a (closed) circle on the 2-dimensional plane are the points of the circumference.

An equivalent definitions of an extreme point is as follows:

Exercise B.16 *Let M be a convex set and let $x \in M$. Prove that*

1. *x is extreme if and only if the only vector h such that $x \pm h \in M$ is the zero vector;*
2. *x is extreme if and only if the set $M \backslash \{x\}$ is convex.*

B.2.7.B. Krein-Milman Theorem

It is clear that a convex set M not necessarily possesses extreme points; as an example you may take the open unit ball in $\mathbf{R}^n$. This example is not interesting – the set in question is not closed; when replacing it with its closure, we get a set (the closed unit ball) with plenty of extreme points – these are all points of the boundary. There are, however, *closed* convex sets which do not possess extreme points – e.g., a line or an affine subspace of larger dimension. A nice fact is that the absence of extreme points in a closed convex set M always has the standard reason – the set contains a line. Thus, a closed and nonempty convex set M which does not contain lines for sure possesses extreme points. And if M is nonempty convex compact set, it possesses a quite representative set of extreme points – their convex hull is the entire M.

Theorem B.2.6 *Let M be a closed and nonempty convex set in $\mathbf{R}^n$. Then*
(i) *The set $\mathrm{Ext}(M)$ of extreme points of M is nonempty if and only if M does not contain lines;*
(ii) *If M is bounded, then M is the convex hull of its extreme points:*

$$M = \mathrm{Conv}(\mathrm{Ext}(M)),$$

so that every point of M is a convex combination of the points of $\mathrm{Ext}(M)$.

Part (ii) of this theorem is the finite-dimensional version of the famous *Krein-Milman Theorem*.
Proof. Let us start with (i). The "only if" part is easy, due to the following simple

Lemma B.2.3 *Let M be a closed convex set in $\mathbf{R}^n$. Assume that for some $\bar{x} \in M$ and $h \in \mathbf{R}^n$ M contains the ray*

$$\{\bar{x} + th \mid t \geq 0\}$$

starting at $\bar{x}$ with the direction h. Then M contains also all parallel rays starting at the points of M:

$$(\forall x \in M) : \{x + th \mid t \geq 0\} \subset M.$$

In particular, if M contains certain line, then it contains also all parallel lines passing through the points of M.

Comment. For a closed convex set M, the set of all directions h such that $x + th \in M$ for some x and all $t \geq 0$ (i.e., by Lemma – such that $x + th \in M$ for all $x \in M$ and all $t \geq 0$) is called *the recessive cone of M* [notation: $\mathrm{Rec}(M)$]. With Lemma B.2.3 it is immediately seen (prove it!) that $\mathrm{Rec}(M)$ indeed is a closed cone, and that

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

Directions from $\mathrm{Rec}(M)$ are called recessive for M.

Proof of the lemma is immediate: if $x \in M$ and $\bar{x} + th \in M$ for all $t \geq 0$, then, due to convexity, for any fixed $\tau \geq 0$ we have

$$\epsilon(\bar{x} + \frac{\tau}{\epsilon}h) + (1 - \epsilon)x \in M$$

for all $\epsilon \in (0, 1)$. As $\epsilon \to +0$, the left hand side tends to $x + \tau h$, and since M is closed, $x + \tau h \in M$ for every $\tau \geq 0$. □

Exercise B.17 *Let M be a closed nonempty convex set. Prove that* $\mathrm{Rec}(M) \neq \{0\}$ *if and only if M is unbounded.*

Lemma B.2.3, of course, resolves all our problems with the "only if" part. Indeed, here we should prove that if M possesses extreme points, then M does not contain lines, or, which is the same, that if M contains lines, then it has no extreme points. But the latter statement is immediate: if M contains a line, then, by Lemma, there is a line in M passing through every given point of M, so that no point can be extreme. □

Now let us prove the "if" part of (i). Thus, from now on we assume that M does not contain lines; our goal is to prove that then M possesses extreme points. Let us start with the following

Lemma B.2.4 *Let Q be a nonempty closed convex set, $\bar{x}$ be a relative boundary point of Q and Π be a hyperplane supporting to Q at $\bar{x}$. Then all extreme points of the nonempty closed convex set $\Pi \cap Q$ are extreme points of Q.*

Proof of the Lemma. First, the set $\Pi \cap Q$ is closed and convex (as an intersection of two sets with these properties); it is nonempty, since it contains $\bar{x}$ (Π contains $\bar{x}$ due to the definition of a supporting plane, and Q contains $\bar{x}$ due to the closedness of Q). Second, let a be the linear form associated with Π:

$$\Pi = \{y \mid a^T y = a^T \bar{x}\},$$

so that

$$\inf_{x \in Q} a^T x < \sup_{x \in Q} a^T x = a^T \bar{x} \tag{B.2.9}$$

(see Proposition B.2.1). Assume that y is an extreme point of $\Pi \cap Q$; what we should do is to prove that y is an extreme point of Q, or, which is the same, to prove that

$$y = \lambda u + (1-\lambda) v$$

for some $u, v \in Q$ and $\lambda \in (0,1)$ is possible only if $y = u = v$. To this end it suffices to demonstrate that under the above assumptions $u, v \in \Pi \cap Q$ (or, which is the same, to prove that $u, v \in \Pi$, since the points are known to belong to Q); indeed, we know that y is an extreme point of $\Pi \cap Q$, so that the relation $y = \lambda u + (1-\lambda) v$ with $\lambda \in (0,1)$ *and* $u, v \in \Pi \cap Q$ does imply $y = u = v$.

To prove that $u, v \in \Pi$, note that since $y \in \Pi$ we have

$$a^T y = a^T \bar{x} \geq \max\{a^T u, a^T v\}$$

(the concluding inequality follows from (B.2.9)). On the other hand,

$$a^T y = \lambda a^T u + (1-\lambda) a^T v;$$

combining these observations and taking into account that $\lambda \in (0,1)$, we conclude that

$$a^T y = a^T u = a^T v.$$

But these equalities imply that $u, v \in \Pi$. □

Equipped with the Lemma, we can easily prove (i) by induction on the dimension of the convex set M (recall that this is nothing but the affine dimension of the affine span of M, i.e., the linear dimension of the linear subspace L such that $\mathrm{Aff}(M) = a + L$).

There is nothing to do if the dimension of M is zero, i.e., if M is a point – then, of course, $M = \mathrm{Ext}(M)$. Now assume that we already have proved the nonemptiness of $\mathrm{Ext}(T)$ for all nonempty closed and not containing lines convex sets T of certain dimension k, and let us prove that the same statement is valid for the sets of dimension $k+1$. Let M be a closed convex nonempty and not containing lines set of dimension $k+1$. Since M does not contain lines and is of positive dimension, it differs from $\mathrm{Aff}(M)$

and therefore it possesses a relative boundary point $\bar{x}$ [3)]. According to Proposition B.2.1, there exists a hyperplane $\Pi = \{x \mid a^T x = a^T \bar{x}\}$ which supports M at $\bar{x}$:

$$\inf_{x \in M} a^T x < \max_{x \in M} a^T x = a^T \bar{x}.$$

By the same Proposition, the set $T = \Pi \cap M$ (which is closed, convex and nonempty) is of affine dimension less than the one of M, i.e., of the dimension $\leq k$. T clearly does not contain lines (since even the larger set M does not contain lines). By the inductive hypothesis, T possesses extreme points, and by Lemma B.2.4 all these points are extreme also for M. The inductive step is completed, and (i) is proved. □

Now let us prove (ii). Thus, let M be nonempty, convex, closed and bounded; we should prove that

$$M = \mathrm{Conv}(\mathrm{Ext}(M)).$$

What is immediately seen is that the right hand side set is contained in the left hand side one. Thus, all we need is to prove that every $x \in M$ is a convex combination of points from $\mathrm{Ext}(M)$. Here we again use induction on the affine dimension of M. The case of 0-dimensional set M (i.e., a point) is trivial. Assume that the statement in question is valid for all k-dimensional convex closed and bounded sets, and let M be a convex closed and bounded set of dimension $k+1$. Let $x \in M$; to represent x as a convex combination of points from $\mathrm{Ext}(M)$, let us pass through x an arbitrary line $\ell = \{x + \lambda h \mid \lambda \in \mathbf{R}\}$ $(h \neq 0)$ in the affine span $\mathrm{Aff}(M)$ of M. Moving along this line from x in each of the two possible directions, we eventually leave M (since M is bounded); as it was explained in the proof of (i), it means that there exist nonnegative λ_+ and λ_- such that the points

$$\bar{x}_\pm = x + \lambda_\pm h$$

both belong to the relative boundary of M. Let us verify that $\bar{x}_\pm$ are convex combinations of the extreme points of M (this will complete the proof, since x clearly is a convex combination of the two points $\bar{x}_\pm$). Indeed, M admits supporting at $\bar{x}_+$ hyperplane Π; as it was explained in the proof of (i), the set $\Pi \cap M$ (which clearly is convex, closed and bounded) is of affine dimension less than that one of M; by the inductive hypothesis, the point $\bar{x}_+$ of this set is a convex combination of extreme points of the set, and by Lemma B.2.4 all these extreme points are extreme points of M as well. Thus, $\bar{x}_+$ is a convex combination of extreme points of M. Similar reasoning is valid for $\bar{x}_-$. ■

B.2.8 Structure of polyhedral sets

B.2.8.A. Main result

By definition, a polyhedral set M is the set of all solutions to a finite system of nonstrict linear inequalities:

$$M = \{x \in \mathbf{R}^n \mid Ax \leq b\}, \tag{B.2.10}$$

where A is a matrix of the column size n and certain row size m and b is m-dimensional vector. This is an "outer" description of a polyhedral set. We are about to establish an important result on the equivalent "inner" representation of a polyhedral set.

[3)] Indeed, there exists $z \in \mathrm{Aff}(M) \backslash M$, so that the points

$$x_\lambda = x + \lambda(z - x)$$

(x is an arbitrary fixed point of M) do not belong to M for some $\lambda \geq 1$, while $x_0 = x$ belongs to M. The set of those $\lambda \geq 0$ for which $x_\lambda \in M$ is therefore nonempty and bounded from above; this set clearly is closed (since M is closed). Thus, there exists the largest $\lambda = \lambda^*$ for which $x_\lambda \in M$. We claim that x_{λ^*} is a relative boundary point of M. Indeed, by construction this is a point from M. If it would be a point from the relative interior of M, then all the points x_λ with close to λ^* and greater than λ^* values of λ would also belong to M, which contradicts the origin of λ^*

Consider the following construction. Let us take two finite nonempty set of vectors V ("vertices") and R ("rays") and build the set

$$M(V,R) = \mathrm{Conv}(V) + \mathrm{Cone}\,(R) = \{\sum_{v\in V}\lambda_v v + \sum_{r\in R}\mu_r r \mid \lambda_v \geq 0, \mu_r \geq 0, \sum_v \lambda_v = 1\}.$$

Thus, we take all vectors which can be represented as sums of convex combinations of the points from V and conic combinations of the points from R. The set $M(V,R)$ clearly is convex (as the arithmetic sum of two convex sets $\mathrm{Conv}(V)$ and $\mathrm{Cone}\,(R)$). The promised inner description polyhedral sets is as follows:

Theorem B.2.7 [Inner description of a polyhedral set] *The sets of the form $M(V,R)$ are exactly the nonempty polyhedral sets: $M(V,R)$ is polyhedral, and every nonempty polyhedral set M is $M(V,R)$ for properly chosen V and R.*

The polytopes $M(V,\{0\}) = \mathrm{Conv}(V)$ are exactly the nonempty and bounded polyhedral sets. The sets of the type $M(\{0\},R)$ are exactly the polyhedral cones (sets given by finitely many nonstrict homogeneous linear inequalities).

Remark B.2.1 In addition to the results of the Theorem, it can be proved that in the representation of a nonempty polyhedral set M as $M = \mathrm{Conv}(V) + \mathrm{Cone}\,(R)$

– the "conic" part $\mathrm{Conv}(R)$ (not the set R itself!) is uniquely defined by M and is the recessive cone of M (see Comment to Lemma B.2.3);

– if M does not contain lines, then V can be chosen as the set of all extreme points of M.

Postponing temporary the proof of Theorem B.2.7, let us explain why this theorem is that important – why it is so nice to know both inner and outer descriptions of a polyhedral set.

Consider a number of natural questions:

- A. Is it true that the inverse image of a polyhedral set $M \subset \mathbf{R}^n$ under an affine mapping $y \mapsto \mathcal{P}(y) = Py + p : \mathbf{R}^m \to \mathbf{R}^n$, i.e., the set
$$\mathcal{P}^{-1}(M) = \{y \in \mathbf{R}^m \mid Py + p \in M\}$$
is polyhedral?
- B. Is it true that the image of a polyhedral set $M \subset \mathbf{R}^n$ under an affine mapping $x \mapsto y = \mathcal{P}(x) = Px + p : \mathbf{R}^n \to \mathbf{R}^m$ – the set
$$\mathcal{P}(M) = \{Px + p \mid x \in M\}$$
is polyhedral?
- C. Is it true that the intersection of two polyhedral sets is again a polyhedral set?
- D. Is it true that the arithmetic sum of two polyhedral sets is again a polyhedral set?

The answers to all these questions are positive; what is very instructive is how these positive answers are obtained.

It is very easy to answer affirmatively to A, starting from the original – outer – definition of a polyhedral set: if $M = \{x \mid Ax \leq b\}$, then, of course,

$$\mathcal{P}^{-1}(M) = \{y \mid A(Py + p) \leq b\} = \{y \mid (AP)y \leq b - Ap\}$$

and therefore $\mathcal{P}^{-1}(M)$ is a polyhedral set.

An attempt to answer affirmatively to B via the same definition fails – there is not seen an easy way to update the linear inequalities defining a polyhedral set into those defining its image, and it is absolutely unclear why the image indeed is given by finitely many linear inequalities. Note, however, that there is no difficulty to answer affirmatively to B with the inner description of a nonempty polyhedral set: if $M = M(V,R)$, then, evidently,

$$\mathcal{P}(M) = M(\mathcal{P}(V), PR),$$

where $PR = \{Pr \mid r \in R\}$ is the image of R under the action of the homogeneous part of $\mathcal{P}$.

Similarly, positive answer to C becomes evident, when we use the outer description of a polyhedral set: taking intersection of the solution sets to two systems of nonstrict linear inequalities, we, of course, again get the solution set to a system of this type – you simply should put together all inequalities from the original two systems. And it is very unclear how to answer positively to D with the outer definition of a polyhedral set – what happens with inequalities when we add the solution sets? In contrast to this, the inner description gives the answer immediately:

$$\begin{array}{rcl} M(V,R)+M(V',R') & = & \mathrm{Conv}(V)+\mathrm{Cone}\,(R)+\mathrm{Conv}(V')+\mathrm{Cone}\,(R') \\ & = & [\mathrm{Conv}(V)+\mathrm{Conv}(V')]+[\mathrm{Cone}\,(R)+\mathrm{Cone}\,(R')] \\ & = & \mathrm{Conv}(V+V')+\mathrm{Cone}\,(R\cup R') \\ & = & M(V+V',R\cup R'). \end{array}$$

Note that in this computation we used two rules which should be justified: $\mathrm{Conv}(V)+\mathrm{Conv}(V') = \mathrm{Conv}(V+V')$ and $\mathrm{Cone}\,(R)+\mathrm{Cone}\,(R') = \mathrm{Cone}\,(R\cup R')$. The second is evident from the definition of the conic hull, and only the first needs simple reasoning. To prove it, note that $\mathrm{Conv}(V)+\mathrm{Conv}(V')$ is a convex set which contains $V+V'$ and therefore contains $\mathrm{Conv}(V+V')$. The inverse inclusion is proved as follows: if

$$x=\sum_i \lambda_i v_i,\ y=\sum_j \lambda'_j v'_j$$

are convex combinations of points from V, resp., V', then, as it is immediately seen (please check!),

$$x+y=\sum_{i,j}\lambda_i\lambda'_j(v_i+v'_j)$$

and the right hand side is a convex combination of points from $V+V'$.

We see that it is extremely useful to keep in mind both descriptions of polyhedral sets – what is difficult to see with one of them, is absolutely clear with another.

As a seemingly "more important" application of the developed theory, let us look at Linear Programming.

B.2.8.B. Theory of Linear Programming

A general Linear Programming program is the problem of maximizing a linear objective function over a polyhedral set:

$$\text{(P)}\quad c^Tx\to\max \mid\ x\in M=\{x\in\mathbf{R}^n \mid\ Ax\le b\};$$

here c is a given n-dimensional vector – the objective, A is a given $m\times n$ *constraint matrix* and $b\in\mathbf{R}^m$ is the right hand side vector. Note that (P) is called "Linear Programming program in the canonical form"; there are other equivalent forms of the problem.

B.2.8.B.1. Solvability of a Linear Programming program. According to the Linear Programming terminology which you for sure know, (P) is called

- feasible, if it admits a feasible solution, i.e., the system $Ax\le b$ is solvable, and infeasible otherwise;
- bounded, if it is feasible and the objective is above bounded on the feasible set, and unbounded, if it is feasible, but the objective is not bounded from above on the feasible set;
- solvable, if it is feasible and the optimal solution exists – the objective attains its maximum on the feasible set.

If the program is bounded, then the upper bound of the values of the objective on the feasible set is a real; this real is called the optimal value of the program and is denoted by c^*. It is convenient to assign optimal value to unbounded and infeasible programs as well – for an unbounded program it, by definition, is $+\infty$, and for an infeasible one it is $-\infty$.

Note that our terminology is aimed to deal with maximization programs; if the program is to minimize the objective, the terminology is updated in the natural way: when defining bounded/unbounded programs, we should speak about below boundedness rather than about the above boundedness of the objective, etc. E.g., the optimal value of an unbounded minimization program is $-\infty$, and of an infeasible one it is $+\infty$. This terminology is consistent with the usual way of converting a minimization problem into an equivalent maximization one by replacing the original objective c with $-c$: the properties of feasibility – boundedness – solvability remain unchanged, and the optimal value in all cases changes its sign.

I have said that you for sure know the above terminology; this is not exactly true, since you definitely have heard and used the words "infeasible LP program", "unbounded LP program", but hardly used the words "bounded LP program" – only the "solvable" one. This indeed is true, although absolutely unclear in advance – a bounded LP program always is solvable. With the tools we have now we can immediately prove this fundamental for Linear Programming fact.

Theorem B.2.8 (i) *A Linear Programming program is solvable if and only if it is bounded.*

(ii) *If the program is solvable and the feasible set of the program does not contain lines, then at least one of the optimal solutions is an extreme point of the feasible set.*

Proof. (i): The "only if" part of the statement is tautological: the definition of solvability includes boundedness. What we should prove is the "if" part – that a bounded program is solvable. This is immediately given by the inner description of the feasible set M of the program: this is a polyhedral set, so that being nonempty (as it is for a bounded program), it can be represented as

$$M(V,R) = \mathrm{Conv}(V) + \mathrm{Cone}\,(R)$$

for some nonempty finite sets V and R. I claim first of all that since (P) is bounded, the inner product of c with every vector from R is nonpositive. Indeed, otherwise there would be $r \in R$ with $c^T r > 0$; since $M(V,R)$ clearly contains with every its point x the entire ray $\{x + tr \mid t \geq 0\}$, and the objective evidently is unbounded on this ray, it would be above unbounded on M, which is not the case.

Now let us choose in the *finite and nonempty* set V the point, let it be called v^*, which maximizes the objective on V. I claim that v^* is an optimal solution to (P), so that (P) is solvable. The justification of the claim is immediate: v^* clearly belongs to M; now, a generic point of $M = M(V,R)$ is

$$x = \sum_{v\in V} \lambda_v v + \sum_{r\in R} \mu_r r$$

with nonnegative λ_v and μ_r and with $\sum_v \lambda_v = 1$, so that

$$\begin{array}{rcll} c^T x & = & \sum_v \lambda_v c^T v + \sum_r \mu_r c^T r & \\ & \leq & \sum_v \lambda_v c^T v & [\text{since } \mu_r \geq 0 \text{ and } c^T r \leq 0,\ r \in R] \\ & \leq & \sum_v \lambda_v c^T v^* & [\text{since } \lambda_v \geq 0 \text{ and } c^T v \leq c^T v^*] \\ & = & c^T v^* & [\text{since } \sum_v \lambda_v = 1]\ \square \end{array}$$

(ii): if the feasible set of (P), let it be called M, does not contain lines, it, being convex and closed (as a polyhedral set) possesses extreme points. It follows that (ii) is valid in the trivial case when the objective of (ii) is constant on the entire feasible set, since then every extreme point of M can be taken as the desired optimal solution. The case when the objective is nonconstant on M can be immediately reduced to the aforementioned trivial case: if x^* is an optimal solution to (P) and the linear form $c^T x$ is nonconstant on M, then the hyperplane $\Pi = \{x \mid c^T x = c^*\}$ is supporting to M at x^*; the set $\Pi \cap M$ is closed, convex, nonempty and does not contain lines, therefore it possesses an extreme point x^{**} which, on one hand, clearly is an optimal solution to (P), and on another hand is an extreme point of M by Lemma B.2.4. ∎

B.2.8.C. Structure of a polyhedral set: proofs

B.2.8.C.1. Extreme points of a polyhedral set. Consider a polyhedral set

$$K = \{x \in \mathbf{R}^n \mid Ax \leq b\},$$

A being a $m \times n$ matrix and b being a vector from $\mathbf{R}^m$. What are the extreme points of K? The answer is given by the following

Theorem B.2.9 [Extreme points of polyhedral set]
Let $x \in K$. The vector x is an extreme point of K if and only if some n linearly independent (i.e., with linearly independent vectors of coefficients) inequalities of the system $Ax \leq b$ are equalities at x.

Proof. Let a_i^T, $i = 1, ..., m$, be the rows of A.

The "only if" part: let x be an extreme point of K, and let I be the set of those indices i for which $a_i^T x = b_i$; we should prove that the set F of vectors $\{a_i \mid i \in I\}$ contains n linearly independent vectors, or, which is the same, that $\mathrm{Lin}(F) = \mathbf{R}^n$. Assume that it is not the case; then the orthogonal complement to F contains a nonzero vector h (since the dimension of $F^\perp$ is equal to $n - \dim \mathrm{Lin}(F)$ and is therefore positive). Consider the segment $\Delta_\epsilon = [x - \epsilon h, x + \epsilon h]$, $\epsilon > 0$ being the parameter of our construction. Since h is orthogonal to the "active" vectors a_i – those with $i \in I$, all points y of this segment satisfy the relations $a_i^T y = a_i^T x = b_i$. Now, if i is a "nonactive" index – one with $a_i^T x < b_i$ – then $a_i^T y \leq b_i$ for all $y \in \Delta_\epsilon$, provided that ϵ is small enough. Since there are finitely many nonactive indices, we can choose $\epsilon > 0$ in such a way that all $y \in \Delta_\epsilon$ will satisfy all "nonactive" inequalities $a_i^T x \leq b_i$, $i \notin I$. Since $y \in \Delta_\epsilon$ satisfies, as we have seen, also all "active" inequalities, we conclude that with the above choice of ϵ we get $\Delta_\epsilon \subset K$, which is a contradiction: $\epsilon > 0$ and $h \neq 0$, so that Δ_ϵ is a nontrivial segment with the midpoint x, and no such segment can be contained in K, since x is an extreme point of K. □

To prove the "if" part, assume that $x \in K$ is such that among the inequalities $a_i^T x \leq b_i$ which are equalities at x there are n linearly independent, say, those with indices $1, ..., n$, and let us prove that x is an extreme point of K. This is immediate: assuming that x is not an extreme point, we would get the existence of a nonzero vector h such that $x \pm h \in K$. In other words, for $i = 1, ..., n$ we would have $b_i \pm a_i^T h \equiv a_i^T (x \pm h) \leq b_i$, which is possible only if $a_i^T h = 0$, $i = 1, ..., n$. But the only vector which is orthogonal to n linearly independent vectors in $\mathbf{R}^n$ is the zero vector (why?), and we get $h = 0$, which was assumed not to be the case. ■.

Corollary B.2.1 *The set of extreme points of a polyhedral set is finite.*

Indeed, according the above Theorem, every extreme point of a polyhedral set $K = \{x \in \mathbf{R}^n \mid Ax \leq b\}$ satisfies the equality version of certain n-inequality subsystem of the original system, the matrix of the subsystem being nonsingular. Due to the latter fact, an extreme point is uniquely defined by the corresponding subsystem, so that the number of extreme points does not exceed the number C_m^n of $n \times n$ submatrices of the matrix A and is therefore finite. ■

Note that C_m^n is nothing but an upper (ant typically very conservative) bound on the number of extreme points of a polyhedral set given by m inequalities in $\mathbf{R}^n$: some $n \times n$ submatrices of A can be singular and, what is more important, the majority of the nonsingular ones normally produce "candidates" which do not satisfy the remaining inequalities.

> **Remark B.2.2** The result of Theorem B.2.9 is very important, in particular, for the theory of the Simplex method – the traditional computational tool of Linear Programming. When applied to the LP program in the standard form
>
> $$c^T x \to \min \mid Px = p,\ x \geq 0 \quad [x \in \mathbf{R}^n],$$
>
> with $k \times n$ matrix P, the result of Theorem B.2.9 is that extreme points of the feasible set are exactly *the basic feasible solutions* of the system $Px = p$, i.e., nonnegative vectors x such that $Px = p$ and the set of columns of P associated with positive entries of x is linearly independent. Since the feasible set of an LP program in the standard form clearly

does not contain lines, among the optimal solutions (if they exist) to an LP program in the standard form at least one is an extreme point of the feasible set (Theorem B.2.8.(ii)). Thus, in principle we could look through the finite set of all extreme points of the feasible set ($\equiv$ through all basic feasible solutions) and to choose the one with the best value of the objective. This receipt allows to find a feasible solution in finitely many arithmetic operations, provided that the program is solvable, and is, basically, what the Simplex method does; this latter method, of course, looks through the basic feasible solutions in a smart way which normally allows to deal with a negligible part of them only.

Another useful consequence of Theorem B.2.9 is that if all the data in an LP program are rational, then every extreme point of the feasible domain of the program is a vector with rational entries. In particular, a solvable standard form LP program with rational data has at least one rational optimal solution.

B.2.8.C.2. Structure of a bounded polyhedral set. Now we are in a position to prove a significant part of Theorem B.2.7 – the one describing *bounded* polyhedral sets.

Theorem B.2.10 [Structure of a bounded polyhedral set] *A bounded and nonempty polyhedral set M in $\mathbf{R}^n$ is a polytope, i.e., is the convex hull of a finite nonempty set:*

$$M = M(V, \{0\}) = \mathrm{Conv}(V);$$

one can choose as V the set of all extreme points of M.

Vice versa – a polytope is a bounded and nonempty polyhedral set.

Proof. The first part of the statement – that a bounded nonempty polyhedral set is a polytope – is readily given by the Krein-Milman Theorem combined with Corollary B.2.1. Indeed, a polyhedral set always is closed (as a set given by nonstrict inequalities involving continuous functions) and convex; if it is also bounded and nonempty, it, by the Krein-Milman Theorem, is the convex hull of the set V of its extreme points; V is finite by Corollary B.2.1. □

Now let us prove the more difficult part of the statement – that a polytope is a bounded polyhedral set. The fact that a convex hull of a finite set is bounded is evident. Thus, all we need is to prove that the convex hull of finitely many points is a polyhedral set.

To make our terminology more brief, let us temporary call the polytopes – convex hulls of nonempty finite sets – V-sets ("V" from "vertex"), and the bounded polyhedral nonempty sets – PB-sets ("P" from "polyhedral", "B" from "bounded"). From the already proved part of the Theorem we know that every PB-set is a V-set as well, and what we should prove is that every V-set M is a PB-set.

Let $M = \mathrm{Conv}(\{v_1, ..., v_N\}$ be a V-set, and let us prove that it is a PB-set. As always, we can assume without loss of generality that the set is full-dimensional[4)]. Thus, we may assume that $\mathrm{int}\, M \neq 0$. By translation, we can also ensure that $0 \in \mathrm{int}\, M$. Now let us look at the polar $M' = \mathrm{Polar}\,(M)$ of M. By Exercise B.14.1, this set is bounded. I claim that this set is also polyhedral, so that M^* is a PB-set. Indeed, a point f belongs to M' if and only if $f^T x \leq 1$ for all x's which are convex combinations of the points $v_1, ..., v_N$, or, which is clearly the same, $f \in M'$ if and only if $f^T v_i \leq 1$, $i = 1, ..., N$. Thus, M' is given by a finite system of nonstrict linear inequalities

$$v_i^T f \leq 1,\ i = 1, ..., N$$

and indeed is polyhedral.

[4)] here is the justification: shifting M, we can assume that M contains 0; replacing $\mathbf{R}^n$ with $L = \mathrm{Lin}(M)$ we come to the situation when the interior of M is nonempty. Given that the result we are proving is valid in this particular case – when the V-set in question possesses a nonempty interior – we are able to conclude that M, as a subset of L, is defined by finitely many nonstrict linear inequalities. Adding to these inequalities the linear equalities defining L – we know from Proposition A.3.7 that a linear subspace is a polyhedral set – we get the desired polyhedral description of M as a subset of $\mathbf{R}^n$.

Now we are done. M' is a PB-set, and therefore, as we already know, is a V-set. Besides this, M' is the polar of a bounded set and therefore 0 is an interior point of M' (Exercise B.14.1). But we just now have proved that the polar to every V-set with 0 in the interior of the set is a PB-set. Thus, the polar to M' – and this is M by Proposition B.2.2 – is a PB-set. ■

B.2.8.C.3. Structure of a general polyhedral set: completing the proof. Now let us prove the general Theorem B.2.7. The proof basically follows the lines of the one of Theorem B.2.10, but with one elaboration: now we cannot use the Krein-Milman Theorem to take upon itself part of our difficulties.

Same as above, to simplify language let us call VR-sets ("V" from "vertex", "R" from rays) the sets of the form $M(V,R)$, and P-sets the nonempty polyhedral sets. We should prove that every P-set is a VR-set, and vice versa. We start with proving that every P-set is a VR-set.

B.2.8.C.3.A. P⇒VR:

P⇒VR, Step 1: reduction to the case when the P-set does not contain lines. Let M be a P-set, so that M is the set of all solutions to a solvable system of linear inequalities:

$$M = \{x \in \mathbf{R}^n \mid Ax \leq b\} \tag{B.2.11}$$

with $m \times n$ matrix A. Such a set may contain lines; if h is the direction of a line in M, then $A(x+th) \leq b$ for some x and all $t \in \mathbf{R}$, which is possible only if $Ah = 0$. Vice versa, if h is from the kernel of A, i.e., if $Ah = 0$, then the line $x + \mathbf{R}^h$ with $x \in M$ clearly is contained in M. Thus, we come to the following fact:

> **Lemma B.2.5** *Nonempty polyhedral set (B.2.11) contains lines if and only if the kernel of A is nontrivial, and the nonzero vectors from the kernel are exactly the directions of lines contained in M: if M contains a line with direction h, then $h \in \mathrm{Ker}A$, and vice versa: if $0 \neq h \in \mathrm{Ker}A$ and $x \in M$, then M contains the entire line $x + \mathbf{R}h$.*

Given a nonempty set (B.2.11), let us denote by L the kernel of A and by $L^\perp$ the orthogonal complement to the kernel, and let M' be the cross-section of M by $L^\perp$:

$$M' = \{x \in L^\perp \mid Ax \leq b\}.$$

The set M' clearly does not contain lines (since the direction of every line in M', on one hand, should belong to $L^\perp$ due to $M' \subset L^\perp$, and on the other hand – should belong to $L = \mathrm{Ker}A$, since a line in $M' \subset M$ is a line in M as well). The set M' is nonempty and, moreover, $M = M' + L$. Indeed, M' contains the orthogonal projections of all points from M onto $L^\perp$ (since to project a point onto $L^\perp$, you should move from this point along certain line with the direction in L, and all these movements, started in M, keep you in M by the Lemma) and therefore is nonempty, first, and is such that $M' + L \supset M$, second. On the other hand, $M' \subset M$ and $M + L = M$ by Lemma B.2.5, whence $M' + L \subset M$. Thus, $M' + L = M$.

Finally, M' is a polyhedral set together with M, since the inclusion $x \in L^\perp$ can be represented by $\dim L$ linear equations (i.e., by $2\dim L$ nonstrict linear inequalities): you should say that x is orthogonal to $\dim L$ somehow chosen vectors $a_1, ..., a_{\dim L}$ forming a basis in L.

The results of our effort are as follows: given an arbitrary P-set M, we have represented is as the sum of a P-set M' not containing lines and a linear subspace L. With this decomposition in mind we see that in order to achieve our current goal – to prove that every P-set is a VR-set – it suffices to prove the same statement for P-sets not containing lines. Indeed, given that $M' = M(V,R')$ and denoting by R' a finite set such that $L = \mathrm{Cone}\,(R')$ (to get R', take the set of $2\dim L$ vectors $\pm a_i$, $i = 1, ..., \dim L$, where $a_1, ..., a_{\dim L}$ is a basis in L), we would obtain

$$\begin{array}{rcl} M & = & M' + L \\ & = & [\mathrm{Conv}(V) + \mathrm{Cone}\,(R)] + \mathrm{Cone}\,(R') \\ & = & \mathrm{Conv}(V) + [\mathrm{Cone}\,(R) + \mathrm{Cone}\,(R')] \\ & = & \mathrm{Conv}(V) + \mathrm{Cone}\,(R \cup R') \\ & = & M(V, R \cup R') \end{array}$$

We see that in order to establish that a P-set is a VR-set it suffices to prove the same statement for the case when the P-set in question does not contain lines.

P⇒VR, Step 2: the P-set does not contain lines. Our situation is as follows: we are given a not containing lines P-set in $\mathbf{R}^n$ and should prove that it is a VR-set. We shall prove this statement by induction on the dimension n of the space. The case of $n = 0$ is trivial. Now assume that the statement in question is valid for $n \le k$, and let us prove that it is valid also for $n = k + 1$. Let M be a not containing lines P-set in $\mathbf{R}^{k+1}$:

$$M = \{x \in \mathbf{R}^{k+1} \mid a_i^T x \le b_i,\ i = 1, ..., m\}. \tag{B.2.12}$$

Without loss of generality we may assume that all a_i are nonzero vectors (since M is nonempty, the inequalities with $a_i = 0$ are valid on the entire $\mathbf{R}^n$, and removing them from the system, we do not vary its solution set). Note that $m > 0$ – otherwise M would contain lines, since $k \ge 0$.

1^0. We may assume that M is unbounded – otherwise the desired result is given already by Theorem B.2.10. By Exercise B.17, there exists a recessive direction $r \ne 0$ of M Thus, M contains the ray $\{x + tr \mid t \ge 0\}$, whence, by Lemma B.2.3, $M + \mathrm{Cone}\,(\{r\}) = M$. □

2^0. For every $i \le m$, where m is the row size of the matrix A from (B.2.12), that is, the number of linear inequalities in the description of M, let us denote by M_i the corresponding "facet" of M – the polyhedral set given by the system of inequalities (B.2.12) with the inequality $a_i^T x \le b_i$ replaced by the equality $a_i^T x = b_i$. Some of these "facets" can be empty; let I be the set of indices i of nonempty M_i's.

When $i \in I$, the set M_i is a nonempty polyhedral set – i.e., a P-set – which does not contain lines (since $M_i \subset M$ and M does not contain lines). Besides this, M_i belongs to the hyperplane $\{a_i^T x = b_i\}$, i.e., actually it is a P-set in $\mathbf{R}^k$. By the inductive hypothesis, we have representations

$$M_i = M(V_i, R_i),\ i \in I,$$

for properly chosen finite nonempty sets V_i and R_i. I claim that

$$M = M(\cup_{i\in I} V_i, \cup_{i\in I} R_i \cup \{r\}), \tag{B.2.13}$$

where r is a recessive direction of M found in 1^0; after the claim will be supported, our induction will be completed.

To prove (B.2.13), note, first of all, that the right hand side of this relation is contained in the left hand side one. Indeed, since $M_i \subset M$ and $V_i \subset M_i$, we have $V_i \subset M$, whence also $V = \cup_i V_i \subset M$; since M is convex, we have

$$\mathrm{Conv}(V) \subset M. \tag{B.2.14}$$

Further, if $r' \in R_i$, then r' is a recessive direction of M_i; since $M_i \subset M$, r' is a recessive direction of M by Lemma B.2.3. Thus, every vector from $\cup_{i\in I} R_i$ is a recessive direction for M, same as r; thus, every vector from $R = \cup_{i\in I} R_i \cup \{r\}$ is a recessive direction of M, whence, again by Lemma B.2.3,

$$M + \mathrm{Cone}\,(R) = M.$$

Combining this relation with (B.2.14), we get $M(V, R) \subset M$, as claimed.

It remains to prove that M is contained in the right hand side of (B.2.13). Let $x \in M$, and let us move from x along the direction $(-r)$, i.e., move along the ray $\{x - tr : t \ge 0\}$. After large enough step along this ray we leave M. (Indeed, otherwise the ray with the direction $-r$ started at x would be contained in M, while the opposite ray for sure is contained in M since r is a recessive direction of M; we would conclude that M contains a line, which is not the case by assumption.) Since the ray $\{x - tr : t \ge 0\}$ eventually leaves M and M is bounded, there exists the largest t, let it be called t^*, such that $x' = x - t^* r$ still belongs to M. It is clear that at x' one of the linear inequalities defining M becomes equality – otherwise we could slightly increase the parameter t^* still staying in M. Thus, $x' \in M_i$ for some $i \in I$. Consequently,

$$x' \in \mathrm{Conv}(V_i) + \mathrm{Cone}\,(R_i),$$

whence $x = x' + t^* r \in \mathrm{Conv}(V_i) + \mathrm{Cone}\,(R_i \cup \{r\}) \subset M(V, R)$, as claimed. □

B.2.8.C.3.B. VR⇒P: We already know that every P-set is a VR-set. Now we shall prove that every VR-set is a P-set, thus completing the proof of Theorem B.2.7. This will be done via the polarity – exactly as in the case of Theorem B.2.10.

Thus, let M be a VR-set:

$$M = M(V, R), \quad V = \{v_1, ..., v_N\},\ R = \{r_1, ..., r_M\};$$

we should prove that it is a P-set. Without loss of generality we may assume that $0 \in M$.

1^0. Let M' be the polar of M. I claim that M' is a P-set. Indeed, $f \in M'$ if and only if $f^T x \le 1$ for every x of the form

$$(\text{convex combination of } v_i) + (\text{conic combination of } r_j),$$

i.e., if and only if $f^T r_j \le 0$ for all j (otherwise $f^T x$ clearly would be unbounded from above on M) and $f^T v_i \le 1$ for all i. Thus,

$$M' = \{f \mid v_i^T f \le 1,\ i = 1, ..., N,\ r_j^T f \le 0,\ j = 1, ..., n\}$$

is a P-set.

2^0. Now we are done: M' is a P-set, and consequently - we already know it – is a VR-set. By 1^0, the polar of a VR-set M' is a P-set; since M is closed and convex and contains the origin, this polar is nothing but M (Proposition B.2.2). Thus, M is a P-set. □

Theorem B.2.7 claims also that the sets of the type $M(V, \{0\})$ are exactly the bounded polyhedral sets – we already know this from Theorem B.2.10 – and that the sets of the type $M(\{0\}, R)$ are exactly the polyhedral cones – i.e., those given by finite systems of homogeneous nonstrict linear inequalities. This latter fact is all which we still should prove. This is easy:

First, let us prove that a polyhedral cone M can be represented as $M(\{0\}, S)$ for some S. Since M is a polyhedral cone, it, as every polyhedral set, can be represented as

$$M = \text{Conv}(V) + \text{Cone}\,(R); \tag{B.2.15}$$

since, by evident reasons, $\text{Conv}(V) \subset \text{Cone}\,(V)$, we get

$$M \subset \text{Cone}\,(V) + \text{Cone}\,(R) = \text{Cone}\,(V \cup R). \tag{B.2.16}$$

On the other hand, since M, being a cone, contains 0, on one hand, and, on the other hand,

$$M + \text{Cone}\,(R) = \text{Conv}(V) + \text{Cone}\,(R) + \text{Cone}\,(R) = \text{Conv}(V) + \text{Cone}\,(R) = M$$

(since $\text{Cone}\,(R) + \text{Cone}\,(R)$ clearly is the same as $\text{Cone}\,(R)$), we get

$$\text{Cone}\,(R) = 0 + \text{Cone}\,(R) \subset M + \text{Cone}\,(R) = M;$$

since $\text{Cone}\,(R) \subset M$ and from (B.2.15) $V \subset M$, the right hand side in (B.2.16) is the conic hull of vectors from M and therefore is a subset of the cone M. Thus, the inclusion in (B.2.16) is in fact equality, and $M = M(\{0\}, V \cup R)$, as required.

It remains to prove that the set of the type $M = M(\{0\}, R)$ – which clearly is a cone – is a polyhedral cone. As every VR-set, M is given by a finite system of inequalities

$$a_i^T x \le b_i,\ i = 1, ..., m,$$

and all we should prove is that the inequalities in the system can be chosen to be homogeneous (with $b_i = 0$). This is immediate: since M is a cone, for every solution x of the above system all vectors tx, $t \ge 0$, also are solutions, which is possible if and only if $b_i \ge 0$ for all i and $a_i^T x \le 0$ for all i and all solutions x to the system. It follows that when "strengthening" the system – replacing in it $b_i \ge 0$ by $b_i = 0$, thus making the system homogeneous – we do not vary the solution set. ■

Appendix C

Convex functions

C.1 Convex functions: first acquaintance

C.1.1 Definition and Examples

Definition C.1.1 [convex function] *A function $f : Q \to \mathbf{R}$ defined on a nonempty subset Q of $\mathbf{R}^n$ and taking real values is called convex, if*

- *the domain Q of the function is convex;*
- *for every $x, y \in Q$ and every $\lambda \in [0, 1]$ one has*

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y). \tag{C.1.1}$$

If the above inequality is strict whenever $x \neq y$ and $0 < \lambda < 1$, f is called strictly convex.

A function f such that $-f$ is convex is called *concave*; the domain Q of a concave function should be convex, and the function itself should satisfy the inequality opposite to (C.1.1):

$$f(\lambda x + (1-\lambda)y) \geq \lambda f(x) + (1-\lambda)f(y), \;\; x, y \in Q, \lambda \in [0, 1].$$

The simplest example of a convex function is an *affine function*

$$f(x) = a^T x + b$$

– the sum of a linear form and a constant. This function clearly is convex on the entire space, and the "convexity inequality" for it is equality. An affine function is both convex and concave; it is easily seen that a function which is both convex and concave on the entire space is affine.

Here are several elementary examples of "nonlinear" convex functions of one variable:

- functions convex on the whole axis:
  x^{2p}, p is a positive integer;
  $\exp\{x\}$;
- functions convex on the nonnegative ray:
  x^p, $1 \leq p$;
  $-x^p$, $0 \leq p \leq 1$;
  $x \ln x$;
- functions convex on the positive ray:
  $1/x^p$, $p > 0$;
  $-\ln x$.

To the moment it is not clear why these functions are convex; in the mean time we shall derive a simple analytic criterion for detecting convexity which immediately demonstrates that the above functions indeed are convex.

A very convenient equivalent definition of a convex function is in terms of its *epigraph*. Given a real-valued function f defined on a nonempty subset Q of $\mathbf{R}^n$, we define its epigraph as the set

$$\mathrm{Epi}(f) = \{(t,x) \in \mathbf{R}^{n+1} : x \in Q, t \geq f(x)\};$$

geometrically, to define the epigraph, you should take the *graph* of the function – the surface $\{t = f(x), x \in Q\}$ in $\mathbf{R}^{n+1}$ – and add to this surface all points which are "above" it. The equivalent, more geometrical, definition of a convex function is given by the following simple statement (prove it!):

Proposition C.1.1 [definition of convexity in terms of the epigraph]
A function f defined on a subset of $\mathbf{R}^n$ is convex if and only if its epigraph is a nonempty convex set in $\mathbf{R}^{n+1}$.

More examples of convex functions: norms. Equipped with Proposition C.1.1, we can extend our initial list of convex functions (several one-dimensional functions and affine ones) by more examples – *norms*. Let $\pi(x)$ be a norm on $\mathbf{R}^n$ (see Section B.1.2.B). To the moment we know three examples of norms – the Euclidean norm $\|x\|_2 = \sqrt{x^T x}$, the 1-norm $\|x\|_1 = \sum\limits_i |x_i|$ and the ∞-norm $\|x\|_\infty = \max\limits_i |x_i|$. It was also claimed (although not proved) that these are three members of an infinite family of norms

$$\|x\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}, \quad 1 \leq p \leq \infty$$

(the right hand side of the latter relation for $p = \infty$ is, by definition, $\max\limits_i |x_i|$).

We are about to prove that every norm is convex:

Proposition C.1.2 *Let $\pi(x)$ be a real-valued function on $\mathbf{R}^n$ which is positively homogeneous of degree 1:*

$$\pi(tx) = t\pi(x) \quad \forall x \in \mathbf{R}^n, t \geq 0.$$

π is convex if and only if it is subadditive:

$$\pi(x+y) \leq \pi(x) + \pi(y) \quad \forall x, y \in \mathbf{R}^n.$$

In particular, a norm (which by definition is positively homogeneous of degree 1 and is subadditive) is convex.

Proof is immediate: the epigraph of a positively homogeneous of degree 1 function π clearly is a conic set: $(t,x) \in \mathrm{Epi}(\pi) \Rightarrow \lambda(t,x) \in \mathrm{Epi}(\pi)$ whenever $\lambda \geq 0$. Now, by Proposition C.1.1 π is convex if and only if $\mathrm{Epi}(\pi)$ is convex. From Proposition 1.7.2 we know that a conic set is convex (i.e., is a cone) if and only if it contains the sum of every two its elements; this latter property is satisfied for the epigraph of a real-valued function if and only if the function is subadditive (evident). ∎

C.1.2 Elementary properties of convex functions

C.1.2.A. Jensen's inequality

The following elementary observation is, I believe, one of the most useful observations in the world:

Proposition C.1.3 [Jensen's inequality] *Let f be convex and Q be the domain of f. Then for every convex combination*

$$\sum_{i=1}^N \lambda_i x_i$$

of points from Q one has

$$f(\sum_{i=1}^N \lambda_i x_i) \le \sum_{i=1}^N \lambda_i f(x_i).$$

The proof is immediate: the points $(f(x_i), x_i)$ clearly belong to the epigraph of f; since f is convex, its epigraph is a convex set, so that the convex combination

$$\sum_{i=1}^N \lambda_i (f(x_i), x_i) = (\sum_{i=1}^N \lambda_i f(x_i), \sum_{i=1}^N \lambda_i x_i)$$

of the points also belongs to $\mathrm{Epi}(f)$. By definition of the epigraph, the latter means exactly that $\sum_{i=1}^N \lambda_i f(x_i) \ge f(\sum_{i=1}^N \lambda_i x_i)$. ∎

Note that the definition of convexity of a function f is exactly the requirement on f to satisfy the Jensen inequality for the case of $N = 2$; we see that to satisfy this inequality for $N = 2$ is the same as to satisfy it for *all* N.

C.1.2.B. Convexity of level sets of a convex function

The following simple observation is also very useful:

Proposition C.1.4 [convexity of level sets] *Let f be a convex function with the domain Q. Then, for every real α, the set*

$$\mathrm{lev}_\alpha(f) = \{x \in Q : f(x) \le \alpha\}$$

– the level set of f – is convex.

The proof takes one line: if $x, y \in \mathrm{lev}_\alpha(f)$ and $\lambda \in [0,1]$, then $f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y) \le \lambda\alpha + (1-\lambda)\alpha = \alpha$, so that $\lambda x + (1-\lambda)y \in \mathrm{lev}_\alpha(f)$.

Note that the convexity of level sets does *not* characterize convex functions; there are nonconvex functions which share this property (e.g., every monotone function on the axis). The "proper" characterization of convex functions in terms of convex sets is given by Proposition C.1.1 – convex functions are exactly the functions with convex epigraphs. Convexity of level sets specify a wider family of functions, the so called *quasiconvex* ones.

C.1.3 What is the value of a convex function outside its domain?

Literally, the question which entitles this subsection is senseless. Nevertheless, when speaking about *convex* functions, it is extremely convenient to think that the function outside its domain also has a value, namely, takes the value $+\infty$; with this convention, we can say that

a convex function f on $\mathbf{R}^n$ is a function taking values in the extended real axis $\mathbf{R} \cup \{+\infty\}$ such that the domain $\mathrm{Dom}\, f$ of the function – the set of those x's where $f(x)$ is finite – is nonempty, and for all $x, y \in \mathbf{R}^n$ and all $\lambda \in [0,1]$ one has

$$f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y). \tag{C.1.2}$$

If the expression in the right hand side involves infinities, it is assigned the value according to the standard and reasonable conventions on what are arithmetic operations in the "extended real axis" $\mathbf{R} \cup \{+\infty\} \cup \{-\infty\}$:

- arithmetic operations with reals are understood in their usual sense;
- the sum of $+\infty$ and a real, same as the sum of $+\infty$ and $+\infty$ is $+\infty$; similarly, the sum of a real and $-\infty$, same as the sum of $-\infty$ and $-\infty$ is $-\infty$. The sum of $+\infty$ and $-\infty$ is undefined;

- the product of a real and $+\infty$ is $+\infty$, 0 or $-\infty$, depending on whether the real is positive, zero or negative, and similarly for the product of a real and $-\infty$. The product of two "infinities" is again infinity, with the usual rule for assigning the sign to the product.

Note that it is not clear in advance that our new definition of a convex function is equivalent to the initial one: initially we included into the definition requirement for the domain to be convex, and now we omit explicit indicating this requirement. In fact, of course, the definitions are equivalent: convexity of Dom f – i.e., the set where f is finite – is an immediate consequence of the "convexity inequality" (C.1.2).

It is convenient to think of a convex function as of something which is defined everywhere, since it saves a lot of words. E.g., with this convention I can write $f + g$ (f and g are convex functions on $\mathbf{R}^n$), and everybody will understand what is meant; without this convention, I am supposed to add to this expression the explanation as follows: "$f+g$ is a function with the domain being the intersection of those of f and g, and in this intersection it is defined as $(f + g)(x) = f(x) + g(x)$".

C.2 How to detect convexity

In an optimization problem

$$f(x) \to \min \mid \ g_j(x) \leq 0,\ j = 1, ..., m$$

convexity of the objective f and the constraints g_i is crucial: it turns out that problems with this property possess nice theoretical properties (e.g., the local *necessary* optimality conditions for these problems are *sufficient for global optimality*); and what is much more important, convex problems can be efficiently (both in theoretical and, to some extent, in the practical meaning of the word) solved numerically, which is not, unfortunately, the case for general nonconvex problems. This is why it is so important to know how one can detect convexity of a given function. This is the issue we are coming to.

The scheme of our investigation is typical for mathematics. Let me start with the example which you know from Analysis. How do you detect continuity of a function? Of course, there is a definition of continuity in terms of ϵ and δ, but it would be an actual disaster if each time we need to prove continuity of a function, we were supposed to write down the proof that "for every positive ϵ there exists positive δ such that ...". In fact we use another approach: we list once for ever a number of standard operations which preserve continuity, like addition, multiplication, taking superpositions, etc., and point out a number of standard examples of continuous functions – like the power function, the exponent, etc. To prove that the operations in the list preserve continuity, same as to prove that the standard functions are continuous, this takes certain effort and indeed is done in $\epsilon - \delta$ terms; but after this effort is once invested, we normally have no difficulties with proving continuity of a given function: it suffices to demonstrate that the function can be obtained, in finitely many steps, from our "raw materials" – the standard functions which are known to be continuous – by applying our machinery – the combination rules which preserve continuity. Normally this demonstration is given by a single word "evident" or even is understood by default.

This is exactly the case with convexity. Here we also should point out the list of operations which preserve convexity and a number of standard convex functions.

C.2.1 Operations preserving convexity of functions

These operations are as follows:

- [stability under taking weighted sums] if f, g are convex functions on $\mathbf{R}^n$, then their linear combination $\lambda f + \mu g$ with *nonnegative* coefficients again is convex, provided that it is finite at least at one point;

  [this is given by straightforward verification of the definition]

- [stability under affine substitutions of the argument] the superposition $f(Ax + b)$ of a convex function f on $\mathbf{R}^n$ and affine mapping $x \mapsto Ax + b$ from $\mathbf{R}^m$ into $\mathbf{R}^n$ is convex, provided that it is finite at least at one point.

[you can prove it directly by verifying the definition or by noting that the epigraph of the superposition, if nonempty, is the inverse image of the epigraph of f under an affine mapping]

- [stability under taking pointwise sup] upper bound $\sup_\alpha f_\alpha(\cdot)$ of every family of convex functions on $\mathbf{R}^n$ is convex, provided that this bound is finite at least at one point.

  [to understand it, note that the epigraph of the upper bound clearly is the intersection of epigraphs of the functions from the family; recall that the intersection of every family of convex sets is convex]

- ["Convex Monotone superposition"] Let $f(x) = (f_1(x), ..., f_k(x))$ be vector-function on $\mathbf{R}^n$ with convex components f_i, and assume that F is a convex function on $\mathbf{R}^k$ which is monotone, i.e., such that $z \le z'$ always implies that $F(z) \le F(z')$. Then the superposition

$$\phi(x) = F(f(x)) = F(f_1(x), ..., f_k(x))$$

  is convex on $\mathbf{R}^n$, provided that it is finite at least at one point.

  Remark C.2.1 *The expression $F(f_1(x), ..., f_k(x))$ makes no evident sense at a point x where some of f_i's are $+\infty$. By definition, we assign the superposition at such a point the value $+\infty$.*

  [To justify the rule, note that if $\lambda \in (0,1)$ and $x, x' \in \operatorname{Dom}\phi$, then $z = f(x), z' = f(x')$ are vectors from $\mathbf{R}^k$ which belong to $\operatorname{Dom} F$, and due to the convexity of the components of f we have

$$f(\lambda x + (1-\lambda)x') \le \lambda z + (1-\lambda)z';$$

  in particular, the left hand side is a vector from $\mathbf{R}^k$ – it has no "infinite entries", and we may further use the monotonicity of F:

$$\phi(\lambda x + (1-\lambda)x') = F(f(\lambda x + (1-\lambda)x')) \le F(\lambda z + (1-\lambda)z')$$

  and now use the convexity of F:

$$F(\lambda z + (1-\lambda)z') \le \lambda F(z) + (1-\lambda)F(z')$$

  to get the required relation

$$\phi(\lambda x + (1-\lambda)x') \le \lambda\phi(x) + (1-\lambda)\phi(x').$$

  ]

Imagine how many extra words would be necessary here if there were no convention on the value of a convex function outside its domain!

Two more rules are as follows:

- [stability under partial minimization] if $f(x,y) : \mathbf{R}^n_x \times \mathbf{R}^m_y$ is convex (as a function of $z = (x,y)$; this is called *joint convexity*) and the function

$$g(x) = \inf_y f(x,y)$$

  is *proper*, i.e., is $> -\infty$ everywhere and is finite at least at one point, then g is convex [this can be proved as follows. We should prove that if $x, x' \in \operatorname{Dom} g$ and $x'' = \lambda x + (1-\lambda)x'$ with $\lambda \in [0,1]$, then $x'' \in \operatorname{Dom} g$ and $g(x'') \le \lambda g(x) + (1-\lambda)g(x')$. Given positive ϵ, we can find y and y' such that $(x,y) \in \operatorname{Dom} f$, $(x',y') \in \operatorname{Dom} f$ and $g(x) + \epsilon \ge f(x,y)$, $g(y') + \epsilon \ge f(x',y')$. Taking weighted sum of these two inequalities, we get

$$\lambda g(x) + (1-\lambda)g(y) + \epsilon \ge \lambda f(x,y) + (1-\lambda)f(x',y') \ge$$

[since f is convex]

$$\geq f(\lambda x + (1-\lambda)x', \lambda y + (1-\lambda)y') = f(x'', \lambda y + (1-\lambda)y')$$

(the last $\geq$ follows from the convexity of f). The concluding quantity in the chain is $\geq g(x'')$, and we get $g(x'') \leq \lambda g(x) + (1-\lambda)g(x') + \epsilon$. In particular, $x'' \in \mathrm{Dom}\, g$ (recall that g is assumed to take only the values from $\mathbf{R}$ and the value $+\infty$). Moreover, since the resulting inequality is valid for all $\epsilon > 0$, we come to $g(x'') \leq \lambda g(x) + (1-\lambda)g(x')$, as required.]

- the "conic transformation" of a convex function f on $\mathbf{R}^n$ – the function $g(y,x) = yf(x/y)$ – is convex in the half-space $y > 0$ in $\mathbf{R}^{n+1}$.

Now we know what are the basic operations preserving convexity. Let us look what are the standard functions these operations can be applied to. A number of examples was already given, but we still do not know why the functions in the examples are convex. The usual way to check convexity of a "simple" – given by a simple formula – function is based on *differential criteria of convexity*. Let us look what are these criteria.

C.2.2 Differential criteria of convexity

From the definition of convexity of a function if immediately follows that convexity is one-dimensional property: a proper (i.e., finite at least at one point) function f on $\mathbf{R}^n$ taking values in $\mathbf{R} \cup \{+\infty\}$ is convex if and only if its restriction on every line, i.e., every function of the type $g(t) = f(x + th)$ on the axis, is either convex, or is identically $+\infty$.

It follows that to detect convexity of a function, it, in principle, suffices to know how to detect convexity of functions of one variable. This latter question can be resolved by the standard Calculus tools. Namely, in the Calculus they prove the following simple

Proposition C.2.1 [Necessary and Sufficient Convexity Condition for smooth functions on the axis] *Let (a,b) be an interval in the axis (we do not exclude the case of $a = -\infty$ and/or $b = +\infty$). Then*

(i) *A differentiable everywhere on (a,b) function f is convex on (a,b) if and only if its derivative f' is monotonically nondecreasing on (a,b);*

(ii) *A twice differentiable everywhere on (a,b) function f is convex on (a,b) if and only if its second derivative f'' is nonnegative everywhere on (a,b).*

With the Proposition, you can immediately verify that the functions listed as examples of convex functions in Section C.1.1 indeed are convex. The only difficulty which you may meet is that some of these functions (e.g., x^p, $p \geq 1$, and $-x^p$, $0 \leq p \leq 1$, were claimed to be convex on the half-interval $[0, +\infty)$, while the Proposition speaks about convexity of functions on intervals. To overcome this difficulty, you may use the following simple

Proposition C.2.2 *Let M be a convex set and f be a function with $\mathrm{Dom}\, f = M$. Assume that f is convex on $\mathrm{ri}\, M$ and is continuous on M, i.e.,*

$$f(x_i) \to f(x),\ i \to \infty,$$

whenever $x_i, x \in M$ and $x_i \to x$ as $i \to \infty$. Then f is convex on M.

Proof of Proposition C.2.1:

(i), necessity. Assume that f is differentiable and convex on (a,b); we should prove that then f' is monotonically nondecreasing. Let $x < y$ be two points of (a,b), and let us prove that $f'(x) \leq f'(y)$. Indeed, let $z \in (x,y)$. We clearly have the following representation of z as a convex combination of x and y:

$$z = \frac{y-z}{y-x}x + \frac{x-z}{y-x}y,$$

whence, from convexity,

$$f(z) \leq \frac{y-z}{y-x}f(x) + \frac{x-z}{y-x}f(y),$$

whence

$$\frac{f(z)-f(x)}{x-z} \leq \frac{f(y)-f(z)}{y-z}.$$

Passing here to limit as $z \to x+0$, we get

$$f'(x) \leq \frac{(f(y)-f(x)}{y-x},$$

and passing in the same inequality to limit as $z \to y-0$, we get

$$f'(y) \geq \frac{(f(y)-f(x)}{y-x},$$

whence $f'(x) \leq f'(y)$, as claimed. ∎

(i), sufficiency. We should prove that if f is differentiable on (a,b) and f' is monotonically nondecreasing on (a,b), then f is convex on (a,b). It suffices to verify that if $x < y$, $x, y \in (a,b)$, and $z = \lambda x + (1-\lambda)y$ with $0 < \lambda < 1$, then

$$f(z) \leq \lambda f(x) + (1-\lambda)f(y),$$

or, which is the same (write $f(z)$ as $\lambda f(z) + (1-\lambda)f(z)$), that

$$\frac{f(z)-f(x)}{\lambda} \leq \frac{f(y)-f(z)}{1-\lambda}.$$

noticing that $z - x = \lambda(y-x)$ and $y - z = (1-\lambda)(y-x)$, we see that the inequality we should prove is equivalent to

$$\frac{f(z)-f(x)}{z-x} \leq \frac{f(y)-f(z)}{y-z}.$$

But in this equivalent form the inequality is evident: by the Lagrange Mean Value Theorem, its left hand side is $f'(\xi)$ with some $\xi \in (x,z)$, while the right hand one is $f'(\eta)$ with some $\eta \in (z,y)$. Since f' is nondecreasing and $\xi \leq z \leq \eta$, we have $f'(\xi) \leq f'(\eta)$, and the left hand side in the inequality we should prove indeed is $\leq$ the right hand one. □

(ii) is immediate consequence of (i), since, as we know from the very beginning of Calculus, a differentiable function – in the case in question, it is f' – is monotonically nondecreasing on an interval if and only if its derivative is nonnegative on this interval. ∎

In fact, for functions of one variable there is a differential criterion of convexity which does not preassume any smoothness (we shall not prove this criterion):

Proposition C.2.3 [convexity criterion for univariate functions]

Let $g : \mathbf{R} \to \mathbf{R} \cup \{+\infty\}$ be a function. Let the domain $\Delta = \{t : g(t) < \infty\}$ of the function be a convex set which is not a singleton, i.e., let it be an interval (a,b) with possibly added one or both endpoints ($-\infty \leq a < b \leq \infty$). g is convex if and only if it satisfies the following 3 requirements:

1) g is continuous on (a,b);

2) g is differentiable everywhere on (a,b), excluding, possibly, a countable set of points, and the derivative $g'(t)$ is nondecreasing on its domain;

3) at each endpoint u of the interval (a,b) which belongs to Δ g is upper semicontinuous:

$$g(u) \geq \limsup_{t \in (a,b), t \to u} g(t).$$

Proof of Proposition C.2.2: Let $x, y \in M$ and $z = \lambda x + (1-\lambda)y$, $\lambda \in [0,1]$, and let us prove that

$$f(z) \le \lambda f(x) + (1-\lambda) f(y).$$

As we know from Theorem B.1.1.(iii), there exist sequences $x_i \in \operatorname{ri} M$ and $y_i \in \operatorname{ri} M$ converging, respectively to x and to y. Then $z_i = \lambda x_i + (1-\lambda) y_i$ converges to z as $i \to \infty$, and since f is convex on $\operatorname{ri} M$, we have

$$f(z_i) \le \lambda f(x_i) + (1-\lambda) f(y_i);$$

passing to limit and taking into account that f is continuous on M and x_i, y_i, z_i converge, as $i \to \infty$, to $x, y, z \in M$, respectively, we obtain the required inequality. ∎

From Propositions C.2.1.(ii) and C.2.2 we get the following convenient *necessary and sufficient* condition for convexity of a *smooth* function of n variables:

Corollary C.2.1 [convexity criterion for smooth functions on $\mathbf{R}^n$]
Let $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ be a function. Assume that the domain Q of f is a convex set with a nonempty interior and that f is

- *continuous on Q*

  and

- *twice differentiable on the interior of Q.*

Then f is convex if and only if its Hessian is positive semidefinite on the interior of Q:

$$h^T f''(x) h \ge 0 \ \ \forall x \in \operatorname{int} Q \ \forall h \in \mathbf{R}^n.$$

Proof. The "only if" part is evident: if f is convex and $x \in Q' = \operatorname{int} Q$, then the function of one variable

$$g(t) = f(x + th)$$

(h is an arbitrary fixed direction in $\mathbf{R}^n$) is convex in certain neighbourhood of the point $t = 0$ on the axis (recall that affine substitutions of argument preserve convexity). Since f is twice differentiable in a neighbourhood of x, g is twice differentiable in a neighbourhood of $t = 0$, so that $g''(0) = h^T f''(x) h \ge 0$ by Proposition C.2.1. □

Now let us prove the "if" part, so that we are given that $h^T f''(x) h \ge 0$ for every $x \in \operatorname{int} Q$ and every $h \in \mathbf{R}^n$, and we should prove that f is convex.

Let us first prove that f is convex on the interior Q' of the domain Q. By Theorem B.1.1, Q' is a convex set. Since, as it was already explained, the convexity of a function on a convex set is one-dimensional fact, all we should prove is that every one-dimensional function

$$g(t) = f(x + t(y-x)), \ \ 0 \le t \le 1$$

(x and y are from Q') is convex on the segment $0 \le t \le 1$. Since f is continuous on $Q \supset Q'$, g is continuous on the segment; and since f is twice continuously differentiable on Q', g is continuously differentiable on $(0,1)$ with the second derivative

$$g''(t) = (y-x)^T f''(x + t(y-x))(y-x) \ge 0.$$

Consequently, g is convex on $[0,1]$ (Propositions C.2.1.(ii) and C.2.2). Thus, f is convex on Q'. It remains to note that f, being convex on Q' and continuous on Q, is convex on Q by Proposition C.2.2. ∎

Applying the combination rules preserving convexity to simple functions which pass the "infinitesimal' convexity tests, we can prove convexity of many complicated functions. Consider, e.g., an *exponential posynomial* – a function

$$f(x) = \sum_{i=1}^{N} c_i \exp\{a_i^T x\}$$

with positive coefficients c_i (this is why the function is called *pos*ynomial). How could we prove that the function is convex? This is immediate:

$\exp\{t\}$ is convex (since its second order derivative is positive and therefore the first derivative is monotone, as required by the infinitesimal convexity test for smooth functions of one variable);

consequently, all functions $\exp\{a_i^T x\}$ are convex (stability of convexity under affine substitutions of argument);

consequently, f is convex (stability of convexity under taking linear combinations with nonnegative coefficients).

And if we were supposed to prove that the maximum of three posynomials is convex? Ok, we could add to our three steps the fourth, which refers to stability of convexity under taking pointwise supremum.

C.3 Gradient inequality

An extremely important property of a convex function is given by the following

Proposition C.3.1 [Gradient inequality] *Let f be a function taking finite values and the value $+\infty$, x be an interior point of the domain of f and Q be a convex set containing x. Assume that*

- *f is convex on Q*

and

- *f is differentiable at x,*

and let $\nabla f(x)$ be the gradient of the function at x. Then the following inequality holds:

$$(\forall y \in Q): \quad f(y) \geq f(x) + (y-x)^T \nabla f(x). \tag{C.3.1}$$

Geometrically: the graph

$$\{(y,t) \in \mathbf{R}^{n+1} : y \in \mathrm{Dom}\, f \cap Q,\ t = f(y)\}$$

of the function f restricted onto the set Q is above the graph

$$\{(y,t) \in \mathbf{R}^{n+1} : t = f(x) + (y-x)^T \nabla f(x)\}$$

of the linear form tangent to f at x.

Proof. Let $y \in Q$. There is nothing to prove if $y \notin \mathrm{Dom}\, f$ (since there the right hand side in the gradient inequality is $+\infty$), same as there is nothing to prove when $y = x$. Thus, we can assume that $y \neq x$ and $y \in \mathrm{Dom}\, f$. Let us set

$$y_\tau = x + \tau(y-x),\ \ 0 < \tau \leq 1,$$

so that $y_1 = y$ and y_τ is an interior point of the segment $[x,y]$ for $0 < \tau < 1$. Now let us use the following extremely simple

> **Lemma C.3.1** *Let x, x', x'' be three distinct points with $x' \in [x, x'']$, and let f be convex and finite on $[x, x'']$. Then*
>
> $$\frac{f(x') - f(x)}{\|x' - x\|_2} \leq \frac{f(x'') - f(x)}{\|x'' - x\|_2}. \tag{C.3.2}$$

Proof of the Lemma. We clearly have

$$x' = x + \lambda(x'' - x), \quad \lambda = \frac{\|x' - x\|_2}{\|x'' - x\|_2} \in (0, 1)$$

or, which is the same,

$$x' = (1 - \lambda)x + \lambda x''.$$

From the convexity inequality

$$f(x') \le (1 - \lambda)f(x) + \lambda f(x''),$$

or, which is the same,

$$f(x') - f(x) \le \lambda(f(x'') - f(x')).$$

Dividing by λ and substituting the value of λ, we come to (C.3.2). □

Applying the Lemma to the triple $x, x' = y_\tau, x'' = y$, we get

$$\frac{f(x + \tau(y - x)) - f(x)}{\tau\|y - x\|_2} \le \frac{f(y) - f(x)}{\|y - x\|_2};$$

as $\tau \to +0$, the left hand side in this inequality, by the definition of the gradient, tends to $\|y - x\|_2^{-1}(y - x)^T\nabla f(x)$, and we get

$$\|y - x\|_2^{-1}(y - x)^T\nabla f(x) \le \|y - x\|_2^{-1}(f(y) - f(x)),$$

or, which is the same,

$$(y - x)^T\nabla f(x) \le f(y) - f(x);$$

this is exactly the inequality (C.3.1). ∎

It is worthy of mentioning that in the case when Q is convex set with a nonempty interior and f is continuous on Q and differentiable on $\operatorname{int} Q$, f is convex on Q if and only if the Gradient inequality (C.3.1) is valid for every pair $x \in \operatorname{int} Q$ and $y \in Q$.

Indeed, the "only if" part, i.e., the implication

$$\textit{convexity of } f \Rightarrow \textit{Gradient inequality for all } x \in \operatorname{int} Q \textit{ and all } y \in Q$$

is given by Proposition C.3.1. To prove the "if" part, i.e., to establish the implication inverse to the above, assume that f satisfies the Gradient inequality for all $x \in \operatorname{int} Q$ and all $y \in Q$, and let us verify that f is convex on Q. It suffices to prove that f is convex on the interior Q' of the set Q (see Proposition C.2.2; recall that by assumption f is continuous on Q and Q is convex). To prove that f is convex on Q', note that Q' is convex (Theorem B.1.1) and that, due to the Gradient inequality, on Q' f is the upper bound of the family of affine (and therefore convex) functions:

$$f(y) = \sup_{x \in Q'} f_x(y), \;\; f_x(y) = f(x) + (y - x)^T\nabla f(x).$$ ∎

C.4 Boundedness and Lipschitz continuity of a convex function

Convex functions possess nice local properties.

Theorem C.4.1 [local boundedness and Lipschitz continuity of convex function]

Let f be a convex function and let K be a closed and bounded set contained in the relative interior of the domain $\operatorname{Dom} f$ *of f. Then f is Lipschitz continuous on K – there exists constant L – the Lipschitz constant of f on K – such that*

$$|f(x) - f(y)| \le L\|x - y\|_2 \quad \forall x, y \in K. \tag{C.4.1}$$

In particular, f is bounded on K.

Remark C.4.1 All three assumptions on K – (1) closedness, (2) boundedness, and (3) $K \subset \mathrm{ri}\,\mathrm{Dom}\, f$ – are essential, as it is seen from the following three examples:

- $f(x) = 1/x$, $\mathrm{Dom}\, F = (0, +\infty)$, $K = (0, 1]$. We have (2), (3) but not (1); f is neither bounded, nor Lipschitz continuous on K.
- $f(x) = x^2$, $\mathrm{Dom}\, f = \mathbf{R}$, $K = \mathbf{R}$. We have (1), (3) and not (2); f is neither bounded nor Lipschitz continuous on K.
- $f(x) = -\sqrt{x}$, $\mathrm{Dom}\, f = [0, +\infty)$, $K = [0, 1]$. We have (1), (2) and not (3); f is not Lipschitz continuous on K [1)], although is bounded. With properly chosen convex function f of two variables and non-polyhedral compact domain (e.g., with $\mathrm{Dom}\, f$ being the unit circle), we could demonstrate also that lack of (3), even in presence of (1) and (2), may cause unboundedness of f at K as well.

Remark C.4.2 *Theorem C.4.1 says that a convex function f is bounded on every compact (i.e., closed and bounded) subset of the relative interior of* $\mathrm{Dom}\, f$. *In fact there is much stronger statement on the below boundedness of f: f is below bounded on any bounded subset of $\mathbf{R}^n$!.*

Proof of Theorem C.4.1. We shall start with the following local version of the Theorem.

Proposition C.4.1 *Let f be a convex function, and let $\bar{x}$ be a point from the relative interior of the domain* $\mathrm{Dom}\, f$ *of f. Then*

(i) *f is bounded at $\bar{x}$: there exists a positive r such that f is bounded in the r-neighbourhood $U_r(\bar{x})$ of $\bar{x}$ in the affine span of* $\mathrm{Dom}\, f$*:*

$$\exists r > 0, C: \quad |f(x)| \le C \ \ \forall x \in U_r(\bar{x}) = \{x \in \mathrm{Aff}(\mathrm{Dom}\, f) : \|x - \bar{x}\|_2 \le r\};$$

(ii) *f is Lipschitz continuous at $\bar{x}$, i.e., there exists a positive ρ and a constant L such that*

$$|f(x) - f(x')| \le L\|x - x'\|_2 \ \forall x, x' \in U_\rho(\bar{x}).$$

Implication "Proposition C.4.1 $\Rightarrow$ Theorem C.4.1" is given by standard Analysis reasoning. All we need is to prove that if K is a bounded and closed (i.e., a compact) subset of $\mathrm{ri}\,\mathrm{Dom}\, f$, then f is Lipschitz continuous on K (the boundedness of f on K is an evident consequence of its Lipschitz continuity on K and boundedness of K). Assume, on contrary, that f is not Lipschitz continuous on K; then for every integer i there exists a pair of points $x_i, y_i \in K$ such that

$$f(x_i) - f(y_i) \ge i\|x_i - y_i\|_2. \tag{C.4.2}$$

Since K is compact, passing to a subsequence we can ensure that $x_i \to x \in K$ and $y_i \to y \in K$. By Proposition C.4.1 the case $x = y$ is impossible – by Proposition f is Lipschitz continuous in a neighbourhood B of $x = y$; since $x_i \to x, y_i \to y$, this neighbourhood should contain all x_i and y_i with large enough indices i; but then, from the Lipschitz continuity of f in B, the ratios $(f(x_i) - f(y_i))/\|x_i - y_i\|_2$ form a bounded sequence, which we know is not the case. Thus, the case $x = y$ is impossible. The case $x \ne y$ is "even less possible" – since, by Proposition, f is continuous on $\mathrm{Dom}\, f$ at both the points x and y (note that Lipschitz continuity at a point clearly implies the usual continuity at it), so that we would have $f(x_i) \to f(x)$ and $f(y_i) \to f(y)$ as $i \to \infty$. Thus, the left hand side in (C.4.2) remains bounded as $i \to \infty$. In the right hand side one factor – i – tends to ∞, and the other one has a nonzero limit $\|x - y\|$, so that the right hand side tends to ∞ as $i \to \infty$; this is the desired contradiction. □

Proof of Proposition C.4.1.

1^0. We start with proving the *above boundedness* of f in a neighbourhood of $\bar{x}$. This is immediate: we know that there exists a neighbourhood $U_{\bar{r}}(\bar{x})$ which is contained in $\mathrm{Dom}\, f$ (since, by assumption, $\bar{x}$ is a relative interior point of $\mathrm{Dom}\, f$). Now, we can find a small simplex Δ of the dimension $m =$

[1)] indeed, we have $\lim\limits_{t \to +0} \frac{f(0) - f(t)}{t} = \lim\limits_{t \to +0} t^{-1/2} = +\infty$, while for a Lipschitz continuous f the ratios $t^{-1}(f(0) - f(t))$ should be bounded

$\dim \mathrm{Aff}(\mathrm{Dom}\, f)$ with the vertices $x_0, ..., x_m$ in $U_{\bar r}(\bar x)$ in such a way that $\bar x$ will be a convex combination of the vectors x_i with *positive* coefficients, even with the coefficients $1/(m+1)$:

$$\bar x = \sum_{i=0}^{m} \frac{1}{m+1} x_i \quad {}^{2)}.$$

We know that $\bar x$ is the point from the relative interior of Δ (Exercise B.8); since Δ spans the same affine subspace as $\mathrm{Dom}\, f$, it means that Δ contains $U_r(\bar x)$ with certain $r > 0$. Now, we have

$$\Delta = \{\sum_{i=0}^{m} \lambda_i x_i : \lambda_i \geq 0, \sum_i \lambda_i = 1\}$$

so that in Δ f is bounded from above by the quantity $\max\limits_{0\leq i\leq m} f(x_i)$ by Jensen's inequality:

$$f(\sum_{i=0}^{m} \lambda_i x_i) \leq \sum_{i=0}^{m} \lambda_i f(x_i) \leq \max_i f(x_i).$$

Consequently, f is bounded from above, by the same quantity, in $U_r(\bar x)$.

2^0. Now let us prove that if f is above bounded, by some C, in $U_r = U_r(\bar x)$, then it in fact is below bounded in this neighbourhood (and, consequently, is bounded in U_r). Indeed, let $x \in U_r$, so that $x \in \mathrm{Aff}(\mathrm{Dom}\, f)$ and $\|x - \bar x\|_2 \leq r$. Setting $x' = \bar x - [x - \bar x] = 2\bar x - x$, we get $x' \in \mathrm{Aff}(\mathrm{Dom}\, f)$ and $\|x' - \bar x\|_2 = \|x - \bar x\|_2 \leq r$, so that $x' \in U_r$. Since $\bar x = \frac{1}{2}[x + x']$, we have

$$2f(\bar x) \leq f(x) + f(x'),$$

whence

$$f(x) \geq 2f(\bar x) - f(x') \geq 2f(\bar x) - C, \;\; x \in U_r(\bar x),$$

and f is indeed below bounded in U_r.

(i) is proved.

3^0. (ii) is an immediate consequence of (i) and Lemma C.3.1. Indeed, let us prove that f is Lipschitz continuous in the neighbourhood $U_{r/2}(\bar x)$, where $r > 0$ is such that f is bounded in $U_r(\bar x)$ (we already know from (i) that the required r does exist). Let $|f| \leq C$ in U_r, and let $x, x' \in U_{r/2}$, $x \neq x'$. Let us extend the segment $[x, x']$ through the point x' until it reaches, at certain point x'', the (relative) boundary of U_r. We have

$$x' \in (x, x''); \quad \|x'' - \bar x\|_2 = r.$$

From (C.3.2) we have

$$f(x') - f(x) \leq \|x' - x\|_2 \frac{f(x'') - f(x)}{\|x'' - x\|_2}.$$

The second factor in the right hand side does not exceed the quantity $(2C)/(r/2) = 4C/r$; indeed, the numerator is, in absolute value, at most $2C$ (since $|f|$ is bounded by C in U_r and both x, x'' belong to U_r), and the denominator is at least $r/2$ (indeed, x is at the distance at most $r/2$ from $\bar x$, and x'' is at the distance exactly r from $\bar x$, so that the distance between x and x'', by the triangle inequality, is at least $r/2$). Thus, we have

$$f(x') - f(x) \leq (4C/r)\|x' - x\|_2, \;\; x, x' \in U_{r/2};$$

[2] to see that the required Δ exists, let us act as follows: first, the case of $\mathrm{Dom}\, f$ being a singleton is evident, so that we can assume that $\mathrm{Dom}\, f$ is a convex set of dimension $m \geq 1$. Without loss of generality, we may assume that $\bar x = 0$, so that $0 \in \mathrm{Dom}\, f$ and therefore $\mathrm{Aff}(\mathrm{Dom}\, f) = \mathrm{Lin}(\mathrm{Dom}\, f)$. By Linear Algebra, we can find m vectors $y_1, ..., y_m$ in $\mathrm{Dom}\, f$ which form a basis in $\mathrm{Lin}(\mathrm{Dom}\, f) = \mathrm{Aff}(\mathrm{Dom}\, f)$. Setting $y_0 = -\sum\limits_{i=1}^{m} y_i$ and taking into account that $0 = \bar x \in \mathrm{ri}\, \mathrm{Dom}\, f$, we can find $\epsilon > 0$ such that the vectors $x_i = \epsilon y_i$, $i = 0, ..., m$, belong to $U_{\bar r}(\bar x)$. By construction, $\bar x = 0 = \frac{1}{m+1} \sum\limits_{i=0}^{m} x_i$.

swapping x and x', we come to

$$f(x) - f(x') \le (4C/r)\|x' - x\|_2,$$

whence

$$|f(x) - f(x')| \le (4C/r)\|x - x'\|_2,\ x, x' \in U_{r/2},$$

as required in (ii). ∎

C.5 Maxima and minima of convex functions

As it was already mentioned, optimization problems involving convex functions possess nice theoretical properties. One of the most important of these properties is given by the following

Theorem C.5.1 ["Unimodality"] *Let f be a convex function on a convex set $Q \subset \mathbf{R}^n$, and let $x^* \in Q \cap \mathrm{Dom}\, f$ be a local minimizer of f on Q:*

$$(\exists r > 0): \quad f(y) \ge f(x^*) \quad \forall y \in Q,\ \|y - x\|_2 < r. \tag{C.5.1}$$

Then x^* is a global minimizer of f on Q:

$$f(y) \ge f(x^*) \quad \forall y \in Q. \tag{C.5.2}$$

Moreover, the set $\operatorname{Argmin}_Q f$ of all local ($\equiv$ global) minimizers of f on Q is convex.

If f is strictly convex (i.e., the convexity inequality $f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda)f(y)$ is strict whenever $x \ne y$ and $\lambda \in (0,1)$), then the above set is either empty or is a singleton.

Proof. 1) Let x^* be a local minimizer of f on Q and $y \in Q$, $y \ne x^*$; we should prove that $f(y) \ge f(x^*)$. There is nothing to prove if $f(y) = +\infty$, so that we may assume that $y \in \mathrm{Dom}\, f$. Note that also $x^* \in \mathrm{Dom}\, f$ for sure – by definition of a local minimizer.

For all $\tau \in (0,1)$ we have, by Lemma C.3.1,

$$\frac{f(x^* + \tau(y - x^*)) - f(x^*)}{\tau\|y - x^*\|_2} \le \frac{f(y) - f(x^*)}{\|y - x^*\|_2}.$$

Since x^* is a local minimizer of f, the left hand side in this inequality is nonnegative for all small enough values of $\tau > 0$. We conclude that the right hand side is nonnegative, i.e., $f(y) \ge f(x^*)$. □

2) To prove convexity of $\operatorname{Argmin}_Q f$, note that $\operatorname{Argmin}_Q f$ is nothing but the level set $\mathrm{lev}_\alpha(f)$ of f associated with the minimal value $\min_Q f$ of f on Q; as a level set of a convex function, this set is convex (Proposition C.1.4).

3) To prove that the set $\operatorname{Argmin}_Q f$ associated with a strictly convex f is, if nonempty, a singleton, note that if there were two distinct minimizers x', x'', then, from strict convexity, we would have

$$f(\frac{1}{2}x' + \frac{1}{2}x'') < \frac{1}{2}[f(x') + f(x'') == \min_Q f,$$

which clearly is impossible - the argument in the left hand side is a point from Q! ∎

Another pleasant fact is that in the case of differentiable convex functions the known from Calculus necessary optimality condition (the Fermat rule) is sufficient for global optimality:

Theorem C.5.2 [Necessary and sufficient optimality condition for a differentiable convex function]

Let f be convex function on convex set $Q \subset \mathbf{R}^n$, and let x^* be an interior point of Q. Assume that f is differentiable at x^*. Then x^* is a minimizer of f on Q if and only if

$$\nabla f(x^*) = 0.$$

Proof. As a *necessary* condition for local optimality, the relation $\nabla f(x^*) = 0$ is known from Calculus; it has nothing in common with convexity. The essence of the matter is, of course, the *sufficiency* of the condition $\nabla f(x^*) = 0$ for *global optimality* of x^* in the case of *convex* f. This sufficiency is readily given by the Gradient inequality (C.3.1): by virtue of this inequality and due to $\nabla f(x^*) = 0$,

$$f(y) \geq f(x^*) + (y - x^*)\nabla f(x^*) = f(x^*)$$

for all $y \in Q$. ∎

A natural question is what happens if x^* in the above statement is not necessarily an interior point of Q. Thus, assume that x^* is an arbitrary point of a convex set Q and that f is convex on Q and differentiable at x^* (the latter means exactly that Dom f contains a neighbourhood of x^* and f possesses the first order derivative at x^*). Under these assumptions, when x^* is a minimizer of f on Q?

The answer is as follows: let

$$T_Q(x^*) = \{h \in \mathbf{R}^n : x^* + th \in Q \quad \forall \text{ small enough } t > 0\}$$

be the *radial cone* of Q at x^*; geometrically, this is the set of all directions leading from x^* inside Q, so that a small enough positive step from x^* along the direction keeps the point in Q. From the convexity of Q it immediately follows that the radial cone indeed is a convex cone (not necessary closed). E.g., when x^* is an interior point of Q, then the radial cone to Q at x^* clearly is the entire $\mathbf{R}^n$. A more interesting example is the radial cone to a polyhedral set

$$Q = \{x : a_i^T x \leq b_i,\ i = 1, ..., m\}; \tag{C.5.3}$$

for $x^* \in Q$ the corresponding radial cone clearly is the polyhedral cone

$$\{h : a_i^T h \leq 0 \quad \forall i : a_i^T x^* = b_i\} \tag{C.5.4}$$

corresponding to the *active* at x^* (i.e., satisfied at the point as equalities rather than as strict inequalities) constraints $a_i^T x \leq b_i$ from the description of Q.

Now, for the functions in question the necessary and sufficient condition for x^* to be a minimizer of f on Q is as follows:

Proposition C.5.1 *Let Q be a convex set, let $x^* \in Q$, and let f be a convex on Q function which is differentiable at x^*. The necessary and sufficient condition for x^* to be a minimizer of f on Q is that the derivative of f taken at x^* along every direction from $T_Q(x^*)$ should be nonnegative:*

$$h^T \nabla f(x^*) \geq 0 \quad \forall h \in T_Q(x^*).$$

Proof is immediate. The necessity is an evident fact which has nothing in common with convexity: assuming that x^* is a local minimizer of f on Q, we note that if there were $h \in T_Q(x^*)$ with $h^T \nabla f(x^*) < 0$, then we would have

$$f(x^* + th) < f(x^*)$$

for all small enough positive t. On the other hand, $x^* + th \in Q$ for all small enough positive t due to $h \in T_Q(x^*)$. Combining these observations, we conclude that in every neighbourhood of x^* there are points from Q with strictly better than the one at x^* values of f; this contradicts the assumption that x^* is a local minimizer of f on Q.

The sufficiency is given by the Gradient Inequality, exactly as in the case when x^* is an interior point of Q. ∎

Proposition C.5.1 says that whenever f is convex on Q and differentiable at $x^* \in Q$, the necessary and sufficient condition for x^* to be a minimizer of f on Q is that the linear form given by the gradient $\nabla f(x^*)$ of f at x^* should be nonnegative at all directions from the radial cone $T_Q(x^*)$. The linear forms nonnegative at all directions from the radial cone also form a cone; it is called the cone *normal* to Q at x^* and is denoted $N_Q(x^*)$. Thus, Proposition says that the necessary and sufficient condition for x^* to minimize f on Q is the inclusion $\nabla f(x^*) \in N_Q(x^*)$. What does this condition actually mean, it depends

on what is the normal cone: whenever we have an explicit description of it, we have an explicit form of the optimality condition.

E.g., when $T_Q(x^*) = \mathbf{R}^n$ (it is the same as to say that x^* is an interior point of Q), then the normal cone is comprised of the linear forms nonnegative at the entire space, i.e., it is the trivial cone $\{0\}$; consequently, for the case in question the optimality condition becomes the Fermat rule $\nabla f(x^*) = 0$, as we already know.

When Q is the polyhedral set (C.5.3), the normal cone is the polyhedral cone (C.5.4); it is comprised of all directions which have nonpositive inner products with all a_i coming from the active, in the aforementioned sense, constraints. The normal cone is comprised of all vectors which have nonnegative inner products with all these directions, i.e., of vectors a such that the inequality $h^T a \geq 0$ is a consequence of the inequalities $h^T a_i \leq 0$, $i \in I(x^*) \equiv \{i : a_i^T x^* = b_i\}$. From the Homogeneous Farkas Lemma we conclude that the normal cone is simply the conic hull of the vectors $-a_i$, $i \in I(x^*)$. Thus, in the case in question (*) reads:

$x^* \in Q$ is a minimizer of f on Q if and only if there exist nonnegative reals λ_i^* associated with "active" (those from $I(x^*)$) values of i such that

$$\nabla f(x^*) + \sum_{i \in I(x^*)} \lambda_i^* a_i = 0.$$

These are the famous *Karush-Kuhn-Tucker* optimality conditions; these conditions are necessary for optimality in an essentially wider situation.

The indicated results demonstrate that the fact that a point $x^* \in \mathrm{Dom}\, f$ is a global minimizer of a convex function f depends only on the local behaviour of f at x^*. This is not the case with maximizers of a convex function. First of all, such a maximizer, if exists, in all nontrivial cases should belong to the boundary of the domain of the function:

Theorem C.5.3 *Let f be convex, and let Q be the domain of f. Assume that f attains its maximum on Q at a point x^* from the relative interior of Q. Then f is constant on Q.*

Proof. Let $y \in Q$; we should prove that $f(y) = f(x^*)$. There is nothing to prove if $y = x^*$, so that we may assume that $y \neq x^*$. Since, by assumption, $x^* \in \mathrm{ri}\, Q$, we can extend the segment $[x^*, y]$ through the endpoint x^*, keeping the left endpoint of the segment in Q; in other words, there exists a point $y' \in Q$ such that x^* is an interior point of the segment $[y', y]$:

$$x^* = \lambda y' + (1 - \lambda) y$$

for certain $\lambda \in (0, 1)$. From the definition of convexity

$$f(x^*) \leq \lambda f(y') + (1 - \lambda) f(y).$$

Since both $f(y')$ and $f(y)$ do not exceed $f(x^*)$ (x^* is a maximizer of f on Q!) and both the weights λ and $1 - \lambda$ are strictly positive, the indicated inequality can be valid only if $f(y') = f(y) = f(x^*)$. ∎

The next theorem gives further information on maxima of convex functions:

Theorem C.5.4 *Let f be a convex function on $\mathbf{R}^n$ and E be a subset of $\mathbf{R}^n$. Then*

$$\sup_{\mathrm{Conv} E} f = \sup_E f. \tag{C.5.5}$$

In particular, if $S \subset \mathbf{R}^n$ is convex and compact set, then the supremum of f on S is equal to the supremum of f on the set of extreme points of S:

$$\sup_S f = \sup_{\mathrm{Ext}(S)} f \tag{C.5.6}$$

Proof. To prove (C.5.5), let $x \in \mathrm{Conv}E$, so that x is a convex combination of points from E (Theorem B.1.4 on the structure of convex hull):

$$x = \sum_i \lambda_i x_i \quad [x_i \in E,\ \lambda_i \geq 0,\ \sum_i \lambda_i = 1].$$

Applying Jensen's inequality (Proposition C.1.3), we get

$$f(x) \leq \sum_i \lambda_i f(x_i) \leq \sum_i \lambda_i \sup_E f = \sup_E f,$$

so that the left hand side in (C.5.5) is $\leq$ the right hand one; the inverse inequality is evident, since $\mathrm{Conv}E \supset E$. □

To derive (C.5.6) from (C.5.5), it suffices to note that from the Krein-Milman Theorem (Theorem B.2.6) for a convex compact set S one has $S = \mathrm{ConvExt}(S)$. ■

The last theorem on maxima of convex functions is as follows:

Theorem C.5.5 *Let f be a convex function such that the domain Q of f is closed and does not contain lines. Then*

(i) *If the set*

$$\operatorname*{Argmax}_Q f \equiv \{x \in Q : f(x) \geq f(y)\, \forall y \in Q\}$$

of global maximizers of f is nonempty, then it intersects the set $\mathrm{Ext}(Q)$ of the extreme points of Q, so that at least one of the maximizers of f is an extreme point of Q;

(ii) *If the set Q is polyhedral and f is above bounded on Q, then the maximum of f on Q is achieved:* $\operatorname*{Argmax}_Q f \neq \emptyset$.

Proof. Let us start with (i). We shall prove this statement by induction on the dimension of Q. The base $\dim Q = 0$, i.e., the case of a singleton Q, is trivial, since here $Q = \mathrm{Ext}Q = \operatorname*{Argmax}_Q f$. Now assume that the statement is valid for the case of $\dim Q \leq p$, and let us prove that it is valid also for the case of $\dim Q = p + 1$. Let us first verify that the set $\operatorname*{Argmax}_Q f$ intersects with the (relative) boundary of Q. Indeed, let $x \in \operatorname*{Argmax}_Q f$. There is nothing to prove if x itself is a relative boundary point of Q; and if x is not a boundary point, then, by Theorem C.5.3, f is constant on Q, so that $\operatorname*{Argmax}_Q f = Q$; and since Q is closed, every relative boundary point of Q (such a point does exist, since Q does not contain lines and is of positive dimension) is a maximizer of f on Q, so that here again $\operatorname*{Argmax}_Q f$ intersects $\partial_{\mathrm{ri}} Q$.

Thus, among the maximizers of f there exists at least one, let it be x, which belongs to the relative boundary of Q. Let H be the hyperplane which supports Q at x (see Section B.2.5), and let $Q' = Q \cap H$. The set Q' is closed and convex (since Q and H are), nonempty (it contains x) and does not contain lines (since Q does not). We have $\max_Q f = f(x) = \max_{Q'} f$ (note that $Q' \subset Q$), whence

$$\emptyset \neq \operatorname*{Argmax}_{Q'} f \subset \operatorname*{Argmax}_Q f.$$

Same as in the proof of the Krein-Milman Theorem (Theorem B.2.6), we have $\dim Q' < \dim Q$. In view of this inequality we can apply to f and Q' our inductive hypothesis to get

$$\mathrm{Ext}(Q') \cap \operatorname*{Argmax}_{Q'} f \neq \emptyset.$$

Since $\mathrm{Ext}(Q') \subset \mathrm{Ext}(Q)$ by Lemma B.2.4 and, as we just have seen, $\operatorname*{Argmax}_{Q'} f \subset \operatorname*{Argmax}_{Q} f$, we conclude that the set $\mathrm{Ext}(Q) \cap \operatorname*{Argmax}_{Q} f$ is not smaller than $\mathrm{Ext}(Q') \cap \operatorname*{Argmax}_{Q'} f$ and is therefore nonempty, as required. □

To prove (ii), let us use the known to us from Lecture 4 results on the structure of a polyhedral convex set:

$$Q = \mathrm{Conv}(V) + \mathrm{Cone}\,(R),$$

where V and R are finite sets. We are about to prove that the upper bound of f on Q is exactly the maximum of f on the finite set V:

$$\forall x \in Q: \quad f(x) \le \max_{v \in V} f(v). \tag{C.5.7}$$

This will mean, in particular, that f attains its maximum on Q – e.g., at the point of V where f attains its maximum on V.

To prove the announced statement, I first claim that if f is above bounded on Q, then every direction $r \in \mathrm{Cone}\,(R)$ is *descent* for f, i.e., is such that every step in this direction taken from every point $x \in Q$ decreases f:

$$f(x + tr) \le f(x) \quad \forall x \in Q \forall t \ge 0. \tag{C.5.8}$$

Indeed, if, on contrary, there were $x \in Q$, $r \in R$ and $t \ge 0$ such that $f(x + tr) > f(x)$, we would have $t > 0$ and, by Lemma C.3.1,

$$f(x + sr) \ge f(x) + \frac{s}{t}(f(x + tr) - f(x)), \ s \ge t.$$

Since $x \in Q$ and $r \in \mathrm{Cone}\,(R)$, $x + sr \in Q$ for all $s \ge 0$, and since f is above bounded on Q, the left hand side in the latter inequality is above bounded, while the right hand one, due to $f(x + tr) > f(x)$, goes to $+\infty$ as $s \to \infty$, which is the desired contradiction.

Now we are done: to prove (C.5.7), note that a generic point $x \in Q$ can be represented as

$$x = \sum_{v \in V} \lambda_v v + r \quad [r \in \mathrm{Cone}\,(R); \sum_v \lambda_v = 1, \lambda_v \ge 0],$$

and we have

$$\begin{array}{rcll} f(x) & = & f(\sum\limits_{v \in V} \lambda_v v + r) & \\ & \le & f(\sum\limits_{v \in V} \lambda_v v) & \text{[by (C.5.8)]} \\ & \le & \sum\limits_{v \in V} \lambda_v f(v) & \text{[Jensen's Inequality]} \\ & \le & \max\limits_{v \in V} f(v) & \blacksquare \end{array}$$

C.6 Subgradients and Legendre transformation

C.6.1 Proper functions and their representation

According to one of two equivalent definitions, a convex function f on $\mathbf{R}^n$ is a function taking values in $\mathbf{R} \cup \{+\infty\}$ such that the epigraph

$$\mathrm{Epi}(f) = \{(t, x) \in \mathbf{R}^{n+1} : t \ge f(x)\}$$

is a nonempty convex set. Thus, there is no essential difference between convex functions and convex sets: convex function generates a convex set – its epigraph – which of course remembers everything about the function. And the only specific property of the epigraph as

a convex set is that it has a recessive direction – namely, $e = (1, 0)$ – such that the intersection of the epigraph with every line directed by h is either empty, or is a closed ray. Whenever a nonempty convex set possesses such a property with respect to certain direction, it can be represented, in properly chosen coordinates, as the epigraph of some convex function. Thus, a convex function is, basically, nothing but a way to look, in the literal meaning of the latter verb, at a convex set.

Now, we know that "actually good" convex sets are closed ones: they possess a lot of important properties (e.g., admit a good outer description) which are not shared by arbitrary convex sets. It means that among convex functions there also are "actually good" ones – those with closed epigraphs. Closedness of the epigraph can be "translated" to the functional language and there becomes a special kind of continuity – *lower semicontinuity:*

Definition C.6.1 [Lower semicontinuity] *Let f be a function (not necessarily convex) defined on $\mathbf{R}^n$ and taking values in $\mathbf{R} \cup \{+\infty\}$. We say that f is lower semicontinuous at a point $\bar{x}$, if for every sequence of points $\{x_i\}$ converging to $\bar{x}$ one has*

$$f(\bar{x}) \leq \lim \inf_{i \to \infty} f(x_i)$$

(here, of course, $\lim\inf$ *of a sequence with all terms equal to $+\infty$ is $+\infty$).*

f is called lower semicontinuous, if it is lower semicontinuous at every point.

A trivial example of a lower semicontinuous function is a continuous one. Note, however, that a semicontinuous function is not obliged to be continuous; what it is obliged, is to make only "jumps down". E.g., the function

$$f(x) = \begin{cases} 0, & x \neq 0 \\ a, & x = 0 \end{cases}$$

is lower semicontinuous if $a \leq 0$ ("jump down at $x = 0$ or no jump at all"), and is <u>not</u> lower semicontinuous if $a > 0$ ("jump up").

The following statement links lower semicontinuity with the geometry of the epigraph:

Proposition C.6.1 *A function f defined on $\mathbf{R}^n$ and taking values from $\mathbf{R} \cup \{+\infty\}$ is lower semicontinuous if and only if its epigraph is closed (e.g., due to its emptiness).*

I shall not prove this statement, same as most of other statements in this Section; the reader definitely is able to restore (very simple) proofs I am skipping.

An immediate consequence of the latter proposition is as follows:

Corollary C.6.1 *The upper bound*

$$f(x) = \sup_{\alpha \in \mathcal{A}} f_\alpha(x)$$

of arbitrary family of lower semicontinuous functions is lower semicontinuous.

[from now till the end of the Section, if the opposite is not explicitly stated, "a function" means "a function defined on the entire $\mathbf{R}^n$ and taking values in $\mathbf{R} \cup \{+\infty\}$"]

Indeed, the epigraph of the upper bound is the intersection of the epigraphs of the functions forming the bound, and the intersection of closed sets always is closed.

Now let us look at *convex* lower semicontinuous functions; according to our general convention, "convex" means "satisfying the convexity inequality and finite at least at one point", or, which is the same, "with convex nonempty epigraph"; and as we just have seen, "lower semicontinuous" means "with closed epigraph". Thus, we are interested in functions with closed convex nonempty epigraphs; to save words, let us call these functions <u>proper</u>.

What we are about to do is to translate to the functional language several constructions and results related to convex sets. In the usual life, a translation (e.g. of poetry) typically results in something less rich than the original; in contrast to this, in mathematics this is a powerful source of new ideas and constructions.

"Outer description" of a proper function. We know that a closed convex set is intersection of closed half-spaces. What does this fact imply when the set is the epigraph of a proper function f? First of all, note that the epigraph is not a completely arbitrary convex set: it has a recessive direction $e = (1, 0)$ – the basic orth of the t-axis in the space of variables $t \in \mathbf{R}, x \in \mathbf{R}^n$ where the epigraph lives. This direction, of course, should be recessive for every closed half-space

$$(*) \quad \Pi = \{(t, x) : \alpha t \geq d^T x - a\} \quad [|\alpha| + |d| > 0]$$

containing $\mathrm{Epi}(f)$ (note that what is written in the right hand side of the latter relation, is one of many universal forms of writing down a general nonstrict linear inequality in the space where the epigraph lives; this is the form the most convenient for us now). Thus, e should be a recessive direction of $\Pi \supset \mathrm{Epi}(f)$; as it is immediately seen, recessivity of e for Π means exactly that $\alpha \geq 0$. Thus, speaking about closed half-spaces containing $\mathrm{Epi}(f)$, we in fact are considering some of the half-spaces (*) with $\alpha \geq 0$.

Now, there are two essentially different possibilities for α to be nonnegative – (A) to be positive, and (B) to be zero. In the case of (B) the boundary hyperplane of Π is "vertical" – it is parallel to e, and in fact it "bounds" only x – Π is comprised of all pairs (t, x) with x belonging to certain half-space in the x-subspace and t being arbitrary real. These "vertical" subspaces will be of no interest for us.

The half-spaces which indeed are of interest for us are the "nonvertical" ones: those given by the case (A), i.e., with $\alpha > 0$. For a non-vertical half-space Π, we always can divide the inequality defining Π by α and to make $\alpha = 1$. Thus, a "nonvertical" candidate to the role of a closed half-space containing $\mathrm{Epi}(f)$ always can be written down as

$$(**) \quad \Pi = \{(t, x) : t \geq d^T x - a\},$$

i.e., *can be represented as the epigraph of an affine function of x.*

Now, when such a candidate indeed is a half-space containing $\mathrm{Epi}(f)$? The answer is clear: it is the case *if and only if the affine function $d^T x - a$ everywhere in $\mathbf{R}^n$ is $\leq f(\cdot)$* – as we shall say, "is an affine minorant of f"; indeed, the smaller is the epigraph, the larger is the function. *If we knew that* $\mathrm{Epi}(f)$ – which definitely is the intersection of all closed half-spaces containing $\mathrm{Epi}(f)$ – *is in fact the intersection of already nonvertical closed half-spaces containing* $\mathrm{Epi}(f)$, or, which is the same, *the intersection of the epigraphs of all affine minorants of f*, we would be able to get a nice and nontrivial result:

> (!) *a proper convex function is the upper bound of affine functions – all its affine minorants.*

(indeed, we already know that it is the same – to say that a function is an upper bound of certain family of functions, and to say that the epigraph of the function is the intersection of the epigraphs of the functions of the family).

(!) indeed is true:

Proposition C.6.2 *A proper convex function f is the upper bound of all its affine minorants. Moreover, at every point $\bar{x} \in \mathrm{ri}\,\mathrm{Dom}\, f$ from the relative interior of the domain f f is even not the upper bound, but simply the maximum of its minorants: there exists an affine function $f_{\bar{x}}(x)$ which is $\leq f(x)$ everywhere in $\mathbf{R}^n$ and is equal to f at $x = \bar{x}$.*

Proof. I. We start with the "Moreover" part of the statement; this is the key to the entire statement. Thus, we are about to prove that if $\bar{x} \in \mathrm{ri}\,\mathrm{Dom}\, f$, then there exists an affine function $f_{\bar{x}}(x)$ which is everywhere $\leq f(x)$, and at $x = \bar{x}$ the inequality becomes an equality.

I.1^0 First of all, we easily can reduce the situation to the one when $\mathrm{Dom}\, f$ is full-dimensional. Indeed, by shifting f we may make the affine span $\mathrm{Aff}(\mathrm{Dom}\, f)$ of the domain of f to be a

linear subspace L in $\mathbf{R}^n$; restricting f onto this linear subspace, we clearly get a proper function on L. If we believe that our statement is true for the case when the domain of f is full-dimensional, we can conclude that there exists an affine function

$$d^T x - a \quad [x \in L]$$

on L ($d \in L$) such that

$$f(x) \geq d^T x - a \quad \forall x \in L; f(\bar{x}) = d^T \bar{x} - a.$$

The affine function we get clearly can be extended, by the same formula, from L on the entire $\mathbf{R}^n$ and is a minorant of f on the entire $\mathbf{R}^n$ – outside of $L \supset \mathrm{Dom}\, f$ f simply is $+\infty$! This minorant on $\mathbf{R}^n$ is exactly what we need.

I.2^0. Now let us prove that our statement is valid when $\mathrm{Dom}\, f$ is full-dimensional, so that $\bar{x}$ is an interior point of the domain of f. Let us look at the point $y = (f(\bar{x}), \bar{x})$. This is a point from the epigraph of f, and I claim that it is a point from the relative boundary of the epigraph. Indeed, if y were a relative interior point of $\mathrm{Epi}(f)$, then, taking $y' = y + e$, we would get a segment $[y', y]$ contained in $\mathrm{Epi}(f)$; since the endpoint y of the segment is assumed to be relative interior for $\mathrm{Epi}(f)$, we could extend this segment a little through this endpoint, not leaving $\mathrm{Epi}(f)$; but this clearly is impossible, since the t-coordinate of the new endpoint would be $< f(\bar{x})$, and the x-component of it still would be $\bar{x}$.

Thus, y is a point from the relative boundary of $\mathrm{Epi}(f)$. Now I claim that y' is an interior point of $\mathrm{Epi}(f)$. This is immediate: we know from Theorem C.4.1 that f is continuous at $\bar{x}$, so that there exists a neighbourhood U of $\bar{x}$ in $\mathrm{Aff}(\mathrm{Dom}\, f) = \mathbf{R}^n$ such that $f(x) \leq f(\bar{x} + 0.5)$ whenever $x \in U$, or, in other words, the set

$$V = \{(t, x) : x \in U, t > f(\bar{x}) + 0.5\}$$

is contained in $\mathrm{Epi}(f)$; but this set clearly contains a neighbourhood of y' in $\mathbf{R}^{n+1}$.

Now let us look at the supporting linear form to $\mathrm{Epi}(f)$ at the point y of the relative boundary of $\mathrm{Epi}(f)$. This form gives us a linear inequality on $\mathbf{R}^{n+1}$ which is satisfied everywhere on $\mathrm{Epi}(f)$ and becomes equality at y; besides this, the inequality is not equality identically on $\mathrm{Epi}(f)$, it is strict somewhere on $\mathrm{Epi}(f)$. Without loss of generality we may assume that the inequality is of the form

$$(+) \quad \alpha t \geq d^T x - a.$$

Now, since our inequality is satisfied at $y' = y + e$ and becomes equality at $(t, x) = y$, α should be ≥ 0; it cannot be 0, since in the latter case the inequality in question would be equality also at $y' \in \mathrm{int}\, \mathrm{Epi}(f)$. But a linear inequality which is satisfied at a convex set and is *equality* at an *interior* point of the set is trivial – coming from the zero linear form (this is exactly the statement that a linear form attaining its minimum on a convex set at a point from the relative interior of the set is constant on the set and on its affine hull).

Thus, inequality $(+)$ which is satisfied on $\mathrm{Epi}(f)$ and becomes equality at y is an inequality with $\alpha > 0$. Let us divide both sides of the inequality by α; we get a new inequality of the form

$$(\&) \quad t \geq d^T x - a$$

(I keep the same notation for the right hand side coefficients – we never will come back to the old coefficients); this inequality is valid on $\mathrm{Epi}(f)$ and is equality at $y = (f(\bar{x}), \bar{x})$. Since the inequality is valid on $\mathrm{Epi}(f)$, it is valid at every pair (t, x) with $x \in \mathrm{Dom}\, f$ and $t = f(x)$:

$$(\#) \quad f(x) \geq d^T x - a \quad \forall x \in \mathrm{Dom}\, f;$$

so that the right hand side is an affine minorant of f on $\mathrm{Dom}\, f$ and therefore – on $\mathbf{R}^n$ ($f = +\infty$ outside $\mathrm{Dom}\, f$!). It remains to note that $(\#)$ is equality at $\bar{x}$, since $(\&)$ is equality at y. $\square$

II. We have proved that if $\mathcal{F}$ if the set of all affine functions which are minorants of f, then the function

$$\bar{f}(x) = \sup_{\phi\in\mathcal{F}} \phi(x)$$

is equal to f on $\operatorname{ri}\operatorname{Dom} f$ (and at x from the latter set in fact sup in the right hand side can be replaced with max); to complete the proof of the Proposition, we should prove that $\bar{f}$ is equal to f also outside $\operatorname{ri}\operatorname{Dom} f$.

II.1^0. Let us first prove that $\bar{f}$ is equal to f outside $\operatorname{cl}\operatorname{Dom} f$, or. which is the same, prove that $\bar{f}(x) = +\infty$ outside $\operatorname{cl}\operatorname{Dom} f$. This is easy: is $\bar{x}$ is a point outside $\operatorname{cl}\operatorname{Dom} f$, it can be strongly separated from $\operatorname{Dom} f$, see Separation Theorem (ii) (Theorem B.2.5). Thus, there exists $z \in \mathbf{R}^n$ such that

$$z^T\bar{x} \geq z^T x + \zeta \quad \forall x \in \operatorname{Dom} f \quad [\zeta > 0]. \tag{C.6.1}$$

Besides this, we already know that there exists at least one affine minorant of f, or, which is the same, there exist a and d such that

$$f(x) \geq d^T x - a \quad \forall x \in \operatorname{Dom} f. \tag{C.6.2}$$

Let us add to (C.6.2) inequality (C.6.1) multiplied by positive weight λ; we get

$$f(x) \geq \phi_\lambda(x) \equiv (d + \lambda z)^T x + [\lambda\zeta - a - \lambda z^T\bar{x}] \quad \forall x \in \operatorname{Dom} f.$$

This inequality clearly says that $\phi_\lambda(\cdot)$ is an affine minorant of f on $\mathbf{R}^n$ for every $\lambda > 0$. The value of this minorant at $x = \bar{x}$ is equal to $d^T\bar{x} - a + \lambda\zeta$ and therefore it goes to $+\infty$ as $\lambda \to +\infty$. We see that the upper bound of affine minorants of f at $\bar{x}$ indeed is $+\infty$, as claimed.

II.2^0. Thus, we know that the upper bound $\bar{f}$ of all affine minorants of f is equal to f everywhere on the relative interior of $\operatorname{Dom} f$ and everywhere outside the closure of $\operatorname{Dom} f$; all we should prove that this equality is also valid at the points of the relative boundary of $\operatorname{Dom} f$. Let $\bar{x}$ be such a point. There is nothing to prove if $\bar{f}(\bar{x}) = +\infty$, since by construction $\bar{f}$ is everywhere $\leq f$. Thus, we should prove that if $\bar{f}(\bar{x}) = c < \infty$, then $f(\bar{x}) = c$. Since $\bar{f} \leq f$ everywhere, to prove that $f(\bar{x}) = c$ is the same as to prove that $f(\bar{x}) \leq c$. This is immediately given by lower semicontinuity of f: let us choose $x' \in \operatorname{ri}\operatorname{Dom} f$ and look what happens along a sequence of points $x_i \in [x', \bar{x})$ converging to $\bar{x}$. All the points of this sequence are relative interior points of $\operatorname{Dom} f$ (Lemma B.1.1), and consequently

$$f(x_i) = \bar{f}(x_i).$$

Now, $x_i = (1 - \lambda_i)\bar{x} + \lambda_i x'$ with $\lambda_i \to +0$ as $i \to \infty$; since $\bar{f}$ clearly is convex (as the upper bound of a family of affine and therefore convex functions), we have

$$\bar{f}(x_i) \leq (1 - \lambda_i)\bar{f}(\bar{x}) + \lambda_i\bar{f}(x').$$

Putting things together, we get

$$f(x_i) \leq (1 - \lambda_i)\bar{f}(\bar{x}) + \lambda_i f(x');$$

as $i \to \infty$, $x_i \to \bar{x}$, and the right hand side in our inequality converges to $\bar{f}(\bar{x}) = c$; since f is lower semicontinuous, we get $f(\bar{x}) \leq c$. ∎

We see why "translation of mathematical facts from one mathematical language to another" – in our case, from the language of convex sets to the language of convex functions – may be fruitful: because we invest a lot into the process rather than run it mechanically.

Closure of a convex function. We got a nice result on the "outer description" of a *proper* convex function: it is the upper bound of a family of affine functions. Note that, vice versa, the upper bound of every family of affine functions is a proper function, provided that this upper bound is finite at least at one point (indeed, as we know from Section C.2.1, upper bound of every family of convex functions is convex, provided that it is finite at least at one point; and Corollary C.6.1 says that upper bound of lower semicontinuous functions (e.g., affine ones – they are even continuous) is lower semicontinuous).

Now, what to do with a convex function which is not lower semicontinuous? The similar question about convex sets – what to do with a convex set which is not closed – can be resolved very simply: we can pass from the set to its closure and thus get a "normal" object which is very "close" to the original one: the "main part" of the original set – its relative interior – remains unchanged, and the "correction" adds to the set something relatively small – the relative boundary. The same approach works for convex functions: if a convex function f is not proper (i.e., its epigraph, being convex and nonempty, is not closed), we can "correct" the function – replace it with a new function with the epigraph being the closure of $\mathrm{Epi}(f)$. To justify this approach, we, of course, should be sure that the closure of the epigraph of a convex function is also an epigraph of such a function. This indeed is the case, and to see it, it suffices to note that a set G in $\mathbf{R}^{n+1}$ is the epigraph of a function taking values in $\mathbf{R} \cup \{+\infty\}$ if and only if the intersection of G with every vertical line $\{x = \text{const}, t \in \mathbf{R}\}$ is either empty, or is a closed ray of the form $\{x = \text{const}, t \geq \bar{t} > -\infty\}$. Now, it is absolutely evident that if G is the closure of the epigraph of a function f, that its intersection with a vertical line is either empty, or is a closed ray, or is the entire line (the last case indeed can take place – look at the closure of the epigraph of the function equal to $-\frac{1}{x}$ for $x > 0$ and $+\infty$ for $x \leq 0$). We see that in order to justify our idea of "proper correction" of a convex function we should prove that if f is convex, then the last of the indicated three cases – the intersection of $\mathrm{cl}\,\mathrm{Epi}(f)$ with a vertical line is the entire line – never occurs. This fact evidently is a corollary of the following simple

Proposition C.6.3 *A convex function is below bounded on every bounded subset of* $\mathbf{R}^n$.

Proof. Without loss of generality we may assume that the domain of the function f is full-dimensional and that 0 is the interior point of the domain. According to Theorem C.4.1, there exists a neighbourhood U of the origin – which can be thought of to be a centered at the origin ball of some radius $r > 0$ – where f is bounded from above by some C. Now, if $R > 0$ is arbitrary and x is an arbitrary point with $|x| \leq R$, then the point

$$y = -\frac{r}{R}x$$

belongs to U, and we have

$$0 = \frac{r}{r+R}x + \frac{R}{r+R}y;$$

since f is convex, we conclude that

$$f(0) \leq \frac{r}{r+R}f(x) + \frac{R}{r+R}f(y) \leq \frac{r}{r+R}f(x) + \frac{R}{r+R}c,$$

and we get the lower bound

$$f(x) \geq \frac{r+R}{r}f(0) - \frac{r}{R}c$$

for the values of f in the centered at 0 ball of radius R. ∎

Thus, we conclude that the closure of the epigraph of a convex function f is the epigraph of certain function, let it be called *the closure* $\mathrm{cl}\, f$ *of* f. Of course, this latter function is convex (its epigraph is convex – it is the closure of a convex set), and since its epigraph is closed, $\mathrm{cl}\, f$ is proper. The following statement gives direct description of $\mathrm{cl}\, f$ in terms of f:

Proposition C.6.4 *Let f be a convex function and $\mathrm{cl}\, f$ be its closure. Then*

(i) *For every x one has*

$$\mathrm{cl}\, f(x) = \lim_{r \to +0} \inf_{x' : \|x' - x\|_2 \le r} f(x').$$

In particular,

$$f(x) \ge \mathrm{cl}\, f(x)$$

for all x, and

$$f(x) = \mathrm{cl}\, f(x)$$

whenever $x \in \mathrm{ri}\,\mathrm{Dom}\, f$, same as whenever $x \notin \mathrm{cl}\,\mathrm{Dom}\, f$.

Thus, the "correction" $f \mapsto \mathrm{cl}\, f$ may vary f only at the points from the relative boundary of $\mathrm{Dom}\, f$,

$$\mathrm{Dom}\, f \subset \mathrm{Dom}\,\mathrm{cl}\, f \subset \mathrm{cl}\,\mathrm{Dom}\, f,$$

whence also

$$\mathrm{ri}\,\mathrm{Dom}\, f = \mathrm{ri}\,\mathrm{Dom}\,\mathrm{cl}\, f.$$

(ii) *The family of affine minorants of* $\mathrm{cl}\, f$ *is exactly the family of affine minorants of f, so that*

$$\mathrm{cl}\, f(x) = \sup\{\phi(x) : \phi \text{ is an affine minorant of } f\},$$

and the sup *in the right hand side can be replaced with* max *whenever $x \in \mathrm{ri}\,\mathrm{Dom}\,\mathrm{cl}\, f = \mathrm{ri}\,\mathrm{Dom}\, f$.*

["so that" comes from the fact that $\mathrm{cl}\, f$ is proper and is therefore the upper bound of its affine minorants]

C.6.2 Subgradients

Let f be a convex function, and let $x \in \mathrm{Dom}\, f$. It may happen that there exists an affine minorant $d^T x - a$ of f which coincides with f at x:

$$f(y) \ge d^T y - a \quad \forall y, \quad f(x) = d^T x - a.$$

From the equality in the latter relation we get $a = d^T x - f(x)$, and substituting this representation of a into the first inequality, we get

$$f(y) \ge f(x) + d^T(y - x) \quad \forall y. \tag{C.6.3}$$

Thus, if f admits an affine minorant which is exact at x, then there exists d which gives rise to inequality (C.6.3). Vice versa, if d is such that (C.6.3) takes place, then the right hand side of (C.6.3), regarded as a function of y, is an affine minorant of f which is exact at x.

Now note that (C.6.3) expresses certain property of a vector d. A vector satisfying, for a given x, this property – i.e., the slope of an exact at x affine minorant of f – is called a *subgradient* of f at x, and the set of all subgradients of f at x is denoted $\partial f(x)$.

Subgradients of convex functions play important role in the theory and numerical methods of Convex Programming – they are quite reasonable surrogates of gradients. The most elementary properties of the subgradients are summarized in the following statement:

Proposition C.6.5 *Let f be a convex function and x be a point from* $\mathrm{Dom}\, f$. *Then*

(i) *$\partial f(x)$ is a closed convex set which for sure is nonempty when $x \in \mathrm{ri}\,\mathrm{Dom}\, f$*

(ii) *If $x \in \mathrm{int}\,\mathrm{Dom}\, f$ and f is differentiable at x, then $\partial f(x)$ is the singleton comprised of the usual gradient of f at x.*

Proof. (i): Closedness and convexity of $\partial f(x)$ are evident – (C.6.3) is an infinite system of nonstrict linear inequalities with respect to d, the inequalities being indexed by $y \in \mathbf{R}^n$. Nonemptiness of $\partial f(x)$ for the case when $x \in \operatorname{ri}\operatorname{Dom} f$ – this is the most important fact about the subgradients – is readily given by our preceding results. Indeed, we should prove that if $x \in \operatorname{ri}\operatorname{Dom} f$, then there exists an affine minorant of f which is exact at x. But this is an immediate consequence of Proposition C.6.4: part (i) of the proposition says that there exists an affine minorant of f which is equal to $\operatorname{cl} f(x)$ at the point x, and part (i) says that $f(x) = \operatorname{cl} f(x)$.

(ii): If $x \in \operatorname{int}\operatorname{Dom} f$ and f is differentiable at x, then $\nabla f(x) \in \partial f(x)$ by the Gradient Inequality. To prove that in the case in question $\nabla f(x)$ is the only subgradient of f at x, note that if $d \in \partial f(x)$, then, by definition,

$$f(y) - f(x) \geq d^T(y - x) \quad \forall y$$

Substituting $y - x = th$, h being a fixed direction and t being > 0, dividing both sides of the resulting inequality by t and passing to limit as $t \to +0$, we get

$$h^T \nabla f(x) \geq h^T d.$$

This inequality should be valid for all h, which is possible if and only if $d = \nabla f(x)$. ∎

Proposition C.6.5 explains why subgradients are good surrogates of gradients: at a point where gradient exists, it is the only subgradient, but, in contrast to the gradient, a subgradient exists basically everywhere (for sure in the relative interior of the domain of the function). E.g., let us look at the simple function

$$f(x) = |x|$$

on the axis. It is, of course, convex (as maximum of two linear forms x and $-x$). Whenever $x \neq 0$, f is differentiable at x with the derivative $+1$ for $x > 0$ and -1 for $x < 0$. At the point $x = 0$ f is not differentiable; nevertheless, it must have subgradients at this point (since 0 is an interior point of the domain of the function). And indeed, it is immediately seen that the subgradients of $|x|$ at $x = 0$ are exactly the reals from the segment $[-1, 1]$. Thus,

$$\partial|x| = \begin{cases} \{-1\}, & x < 0 \\ [-1, 1], & x = 0 \\ \{+1\}, & x > 0 \end{cases}.$$

Note also that if x is a relative boundary point of the domain of a convex function, even a "good" one, the set of subgradients of f at x may be empty, as it is the case with the function

$$f(y) = \begin{cases} -\sqrt{y}, & y \geq 0 \\ +\infty, & y < 0 \end{cases};$$

it is clear that there is no non-vertical supporting line to the epigraph of the function at the point $(0, f(0))$, and, consequently, there is no affine minorant of the function which is exact at $x = 0$.

A significant – and important – part of Convex Analysis deals with *subgradient calculus* – with the rules for computing subgradients of "composite" functions, like sums, superpositions, maxima, etc., given subgradients of the operands. These rules extend onto nonsmooth convex case the standard Calculus rules and are very nice and instructive; the related considerations, however, are beyond our scope.

C.6.3 Legendre transformation

Let f be a convex function. We know that f "basically" is the upper bound of all its affine minorants; this is exactly the case when f is proper, otherwise the corresponding equality

takes place everywhere except, perhaps, some points from the relative boundary of Dom f. Now, when an affine function $d^Tx - a$ is an affine minorant of f? It is the case if and only if

$$f(x) \geq d^Tx - a$$

for all x or, which is the same, if and only if

$$a \geq d^Tx - f(x)$$

for all x. We see that if the slope d of an affine function $d^Tx - a$ is fixed, then in order for the function to be a minorant of f we should have

$$a \geq \sup_{x \in \mathbf{R}^n} [d^Tx - f(x)].$$

The supremum in the right hand side of the latter relation is certain function of d; this function is called the *Legendre transformation* of f and is denoted f^*:

$$f^*(d) = \sup_{x \in \mathbf{R}^n} [d^Tx - f(x)].$$

Geometrically, the Legendre transformation answers the following question: given a slope d of an affine function, i.e., given the hyperplane $t = d^Tx$ in $\mathbf{R}^{n+1}$, what is the minimal "shift down" of the hyperplane which places it below the graph of f?

From the definition of the Legendre transformation it follows that this is a proper function. Indeed, we loose nothing when replacing $\sup_{x \in \mathbf{R}^n} [d^Tx - f(x)]$ by $\sup_{x \in \mathrm{Dom}\, f} [d^Tx - f(x)]$, so that the Legendre transformation is the upper bound of a family of affine functions. Since this bound is finite at least at one point (namely, at every d coming form affine minorant of f; we know that such a minorant exists), it is a convex lower semicontinuous function, as claimed.

The most elementary (and the most fundamental) fact about the Legendre transformation is its symmetry:

Proposition C.6.6 *Let f be a convex function. Then twice taken Legendre transformation of f is the closure* $\mathrm{cl}\, f$ *of f:*

$$(f^*)^* = \mathrm{cl}\, f.$$

In particular, if f is proper, then it is the Legendre transformation of its Legendre transformation (which also is proper).

Proof is immediate. The Legendre transformation of f^* at the point x is, by definition,

$$\sup_{d \in \mathbf{R}^n} [x^Td - f^*(d)] = \sup_{d \in \mathbf{R}^n, a \geq f^*(d)} [d^Tx - a];$$

the second sup here is exactly the supremum of all affine minorants of f (this is the origin of the Legendre transformation: $a \geq f^*(d)$ if and only if the affine form $d^Tx - a$ is a minorant of f). And we already know that the upper bound of all affine minorants of f is the closure of f. ∎

The Legendre transformation is a very powerful tool – this is a "global" transformation, so that *local* properties of f^* correspond to *global* properties of f. E.g.,

- $d = 0$ belongs to the domain of f^* if and only if f is below bounded, and if it is the case, then $f^*(0) = -\inf f$;
- if f is proper, then the subgradients of f^* at $d = 0$ are exactly the minimizers of f on $\mathbf{R}^n$;

- Dom f^* is the entire $\mathbf{R}^n$ if and only if $f(x)$ grows, as $\|x\|_2 \to \infty$, faster than $\|x\|_2$: there exists a function $r(t) \to \infty$, as $t \to \infty$ such that

$$f(x) \geq r(\|x\|_2) \quad \forall x,$$

etc. Thus, whenever we can compute explicitly the Legendre transformation of f, we get a lot of "global" information on f. Unfortunately, the more detailed investigation of the properties of Legendre transformation is beyond our scope; I simply list several simple facts and examples:

- From the definition of Legendre transformation,

$$f(x) + f^*(d) \geq x^T d \quad \forall x, d.$$

Specifying here f and f^*, we get certain inequality, e.g., the following one:
[Young's Inequality] *if p and q are positive reals such that $\frac{1}{p} + \frac{1}{q} = 1$, then*

$$\frac{|x|^p}{p} + \frac{|d|^q}{q} \geq xd \quad \forall x, d \in \mathbf{R}$$

(indeed, as it is immediately seen, the Legendre transformation of the function $|x|^p/p$ is $|d|^q/q$)

Consequences. Very simple-looking Young's inequality gives rise to a very nice and useful *Hölder inequality:*
Let $1 \leq p \leq \infty$ and let q be such $\frac{1}{p} + \frac{1}{q} = 1$ ($p = 1 \Rightarrow q = \infty$, $p = \infty \Rightarrow q = 1$). For every two vectors $x, y \in \mathbf{R}^n$ one has

$$\sum_{i=1}^{n} |x_i y_i| \leq \|x\|_p \|y\|_q \tag{C.6.4}$$

Indeed, there is nothing to prove if p or q is ∞ – if it is the case, the inequality becomes the evident relation

$$\sum_i |x_i y_i| \leq (\max_i |x_i|)(\sum_i |y_i|).$$

Now let $1 < p < \infty$, so that also $1 < q < \infty$. In this case we should prove that

$$\sum_i |x_i y_i| \leq (\sum_i |x_i|^p)^{1/p} (\sum_i |y_i|^q)^{1/q}.$$

There is nothing to prove if one of the factors in the right hand side vanishes; thus, we can assume that $x \neq 0$ and $y \neq 0$. Now, both sides of the inequality are of homogeneity degree 1 with respect to x (when we multiply x by t, both sides are multiplied by $|t|$), and similarly with respect to y. Multiplying x and y by appropriate reals, we can make both factors in the right hand side equal to 1: $\|x\|_p = \|y\|_p = 1$. Now we should prove that under this normalization the left hand side in the inequality is ≤ 1, which is immediately given by the Young inequality:

$$\sum_i |x_i y_i| \leq \sum_i [|x_i|^p/p + |y_i|^q/q] = 1/p + 1/q = 1.$$

Note that the Hölder inequality says that

$$|x^T y| \leq \|x\|_p \|y\|_q; \tag{C.6.5}$$

when $p = q = 2$, we get the Cauchy inequality. Now, inequality (C.6.5) is *exact* in the sense that for every x there exists y with $\|y\|_q = 1$ such that

$$x^T y = \|x\|_p \quad [= \|x\|_p \|y\|_q];$$

it suffices to take

$$y_i = \|x\|_p^{1-p} |x_i|^{p-1} \mathrm{sign}(x_i)$$

(here $x \neq 0$; the case of $x = 0$ is trivial – here y can be an arbitrary vector with $\|y\|_q = 1$).

Combining our observations, we come to an extremely important, although simple, fact:

$$\|x\|_p = \max\{y^T x : \|y\|_q \leq 1\} \quad [\frac{1}{p} + \frac{1}{q} = 1]. \tag{C.6.6}$$

It follows, in particular, that $\|x\|_p$ is convex (as an upper bound of a family of linear forms), whence

$$\|x' + x''\|_p = 2\|\frac{1}{2}x' + \frac{1}{2}x''\|_p \leq 2(\|x'\|_p/2 + \|x''\|_p/2) = \|x'\|_p + \|x''\|_p;$$

this is nothing but the triangle inequality. Thus, $\|x\|_p$ satisfies the triangle inequality; it clearly possesses two other characteristic properties of a norm – positivity and homogeneity. Consequently, $\|\cdot\|_p$ is a norm – the fact that we announced twice and have finally proven now.

- The Legendre transformation of the function

$$f(x) \equiv -a$$

is the function which is equal to a at the origin and is $+\infty$ outside the origin; similarly, the Legendre transformation of an affine function $\bar{d}^T x - a$ is equal to a at $d = \bar{d}$ and is $+\infty$ when $d \neq \bar{d}$;

- The Legendre transformation of the strictly convex quadratic form

$$f(x) = \frac{1}{2} x^T A x$$

(A is positive definite symmetric matrix) is the quadratic form

$$f^*(d) = \frac{1}{2} d^T A^{-1} d$$

- The Legendre transformation of the Euclidean norm

$$f(x) = \|x\|_2$$

is the function which is equal to 0 in the closed unit ball centered at the origin and is $+\infty$ outside the ball.

The latter example is a particular case of the following statement:

Let $\|x\|$ be a norm on $\mathbf{R}^n$, and let

$$\|d\|_* = \sup\{d^T x : \|x\| \leq 1\}$$

be the conjugate to $\|\cdot\|$ norm.

Exercise C.1 *Prove that $\|\cdot\|_*$ is a norm, and that the norm conjugate to $\|\cdot\|_*$ is the original norm $\|\cdot\|$.*

Hint: Observe that the unit ball of $\|\cdot\|_*$ is exactly the polar of the unit ball of $\|\cdot\|$.

The Legendre transformation of $\|x\|$ is the characteristic function of the unit ball of the conjugate norm, i.e., is the function of d equal to 0 when $\|d\|_* \leq 1$ and is $+\infty$ otherwise.

E.g., (C.6.6) says that the norm conjugate to $\|\cdot\|_p$, $1 \leq p \leq \infty$, is $\|\cdot\|_q$, $1/p + 1/q = 1$; consequently, the Legendre transformation of p-norm is the characteristic function of the unit $\|\cdot\_q$-ball.

Appendix D

Convex Programming, Lagrange Duality, Saddle Points

D.1 Mathematical Programming Program

A (constrained) Mathematical Programming program is a problem as follows:

$$\text{(P)} \quad \min\{f(x) : x \in X, \quad g(x) \equiv (g_1(x), ..., g_m(x)) \leq 0, \ h(x) \equiv (h_1(x), ..., h_k(x)) = 0\}. \qquad \text{(D.1.1)}$$

The standard terminology related to (D.1.1) is:

- [domain] X is called the *domain* of the problem
- [objective] f is called the *objective*
- [constraints] g_i, $i = 1, ..., m$, are called the (functional) *inequality constraints*; h_j, $j = 1, ..., k$, are called the *equality constraints*[1)]

In the sequel, if the opposite is not explicitly stated, it always is assumed that the objective and the constraints are well-defined on X.

- [feasible solution] a point $x \in \mathbf{R}^n$ is called a *feasible solution* to (D.1.1), if $x \in X$, $g_i(x) \leq 0$, $i = 1, ..., m$, and $h_j(x) = 0$, $j = 1, ..., k$, i.e., if x satisfies all restrictions imposed by the formulation of the problem
  - [feasible set] the set of all feasible solutions is called the *feasible set* of the problem
  - [feasible problem] a problem with a nonempty feasible set (i.e., the one which admits feasible solutions) is called *feasible* (or consistent)
  - [active constraints] an inequality constraint $g_i(\cdot) \leq 0$ is called *active at a given feasible solution* x, if this constraint is satisfied at the point as an equality rather than strict inequality, i.e., if

    $$g_i(x) = 0.$$

    A equality constraint $h_i(x) = 0$ by definition is active at every feasible solution x.

[1)] rigorously speaking, the constraints are not the functions g_i, h_j, but the relations $g_i(x) \leq 0$, $h_j(x) = 0$; in fact the word "constraints" is used in both these senses, and it is always clear what is meant. E.g., saying that x satisfies the constraints, we mean the relations, and saying that the constraints are differentiable, we mean the functions

- [optimal value] the quantity

$$f^* = \begin{cases} \inf\limits_{x\in X: g(x)\le 0, h(x)=0} f(x), & \text{the problem is feasible} \\ +\infty, & \text{the problem is infeasible} \end{cases}$$

  is called *the optimal value* of the problem

  - [below boundedness] the problem is called *below bounded*, if its optimal value is $> -\infty$, i.e., if the objective is below bounded on the feasible set

- [optimal solution] a point $x \in \mathbf{R}^n$ is called an *optimal solution* to (D.1.1), if x is feasible and $f(x) \le f(x')$ for any other feasible solution, i.e., if

$$x \in \operatorname*{Argmin}_{x'\in X: g(x')\le 0, h(x')=0} f(x')$$

  - [solvable problem] a problem is called *solvable*, if it admits optimal solutions
  - [optimal set] the set of all optimal solutions to a problem is called its *optimal set*

To solve the problem *exactly* means to find its optimal solution or to detect that no optimal solution exists.

D.2 Convex Programming program and Lagrange Duality Theorem

A Mathematical Programming program (P) is called *convex* (or *Convex Programming* program), if

- X is a *convex* subset of $\mathbf{R}^n$
- $f, g_1, ..., g_m$ are *real-valued convex* functions on X,
  and
- there are no equality constraints at all.

Note that instead of saying that there are no equality constraints, we could say that there are constraints of this type, but only *linear* ones; this latter case can be immediately reduced to the one without equality constraints by replacing $\mathbf{R}^n$ with the affine subspace given by the (linear) equality constraints.

D.2.1 Convex Theorem on Alternative

The simplest case of a convex program is, of course, a Linear Programming program – the one where $X = \mathbf{R}^n$ and the objective and all the constraints are linear. We already know what are optimality conditions for this particular case – they are given by the Linear Programming Duality Theorem. How did we get these conditions?

We started with the observation that the fact that a point x^* is an optimal solution can be expressed in terms of solvability/unsolvability of certain systems of inequalities: in our now terms, these systems are

$$x \in G,\ f(x) \le c,\ g_j(x) \le 0,\ j = 1, ..., m \tag{D.2.1}$$

and

$$x \in G,\ f(x) < c,\ g_j(x) \le 0,\ j = 1, ..., m; \tag{D.2.2}$$

here c is a parameter. Optimality of x^* for the problem means exactly that for appropriately chosen c (this choice, of course, is $c = f(x^*)$) the first of these systems is solvable and x^* is its solution, while the second system is unsolvable. Given this trivial observation, we converted the "negative" part of it – the

claim that (D.2.2) is unsolvable – into a positive statement, using the General Theorem on Alternative, and this gave us the LP Duality Theorem.

Now we are going to use the same approach. What we need is a "convex analogy" to the Theorem on Alternative – something like the latter statement, but for the case when the inequalities in question are given by convex functions rather than the linear ones (and, besides it, we have a "convex inclusion" $x \in X$).

It is easy to *guess* the result we need. How did we come to the formulation of the Theorem on Alternative? The question we were interested in was, basically, how to express in an affirmative manner the fact that a system of linear inequalities has no solutions; to this end we observed that if we can combine, in a linear fashion, the inequalities of the system and get an obviously false inequality like $0 \leq -1$, then the system is unsolvable; this condition is certain affirmative statement with respect to the weights with which we are combining the original inequalities.

Now, the scheme of the above reasoning has nothing in common with linearity (and even convexity) of the inequalities in question. Indeed, consider *an arbitrary* inequality system of the type (D.2.2):

$$\text{(I)} \qquad \begin{array}{rcl} f(x) & < & c \\ g_j(x) & \leq & 0,\ j = 1, ..., m \\ x & \in & X; \end{array}$$

all we assume is that X is a nonempty subset in $\mathbf{R}^n$ and $f, g_1, ..., g_m$ are real-valued functions on X. It is absolutely evident that

if there exist nonnegative $\lambda_1, ..., \lambda_m$ such that the inequality

$$f(x) + \sum_{j=1}^{m} \lambda_j g_j(x) < c \tag{D.2.3}$$

has no solutions in X, then (I) *also has no solutions.*

Indeed, a solution to (I) clearly is a solution to (D.2.3) – the latter inequality is nothing but a combination of the inequalities from (I) with the weights 1 (for the first inequality) and λ_j (for the remaining ones).

Now, what does it mean that (D.2.3) has no solutions? A necessary and sufficient condition for this is that the infimum of the left hand side of (D.2.3) in $x \in X$ is $\geq c$. Thus, we come to the following evident

Proposition D.2.1 [Sufficient condition for insolvability of (I)] *Consider a system* (I) *with arbitrary data and assume that the system*

$$\text{(II)} \qquad \begin{array}{rcl} \inf\limits_{x \in X} \left[f(x) + \sum\limits_{j=1}^{m} \lambda_j g_j(x) \right] & \geq & c \\ \lambda_j & \geq & 0,\ j = 1, ..., m \end{array}$$

with unknowns $\lambda_1, ..., \lambda_m$ has a solution. Then (I) *is infeasible.*

Let me stress that this result is completely general; it does not require any assumptions on the entities involved.

The result we have obtained, unfortunately, does not help us: the actual power of the Theorem on Alternative (and the fact used to prove the Linear Programming Duality Theorem) is not the *sufficiency* of the condition of Proposition for infeasibility of (I), but the *necessity* of this condition. Justification of necessity of the condition in question has nothing in common with the evident reasoning which gives the sufficiency. The necessity in the linear case ($X = \mathbf{R}^n$, f, $g_1, ..., g_m$ are linear) can be established via the Homogeneous Farkas Lemma. Now we shall prove the necessity of the condition for the *convex* case, and already here we need some additional, although minor, assumptions; and in the general nonconvex case

the condition in question simply is *not* necessary for infeasibility of (I) [and this is very bad – this is the reason why there exist difficult optimization problems which we do not know how to solve efficiently].

The just presented "preface" explains what we should do; now let us carry out our plan. We start with the aforementioned "minor regularity assumptions".

Definition D.2.1 [Slater Condition] *Let $X \subset \mathbf{R}^n$ and $g_1, ..., g_m$ be real-valued functions on X. We say that these functions satisfy the Slater condition on X, if there exists $x \in X$ such that $g_j(x) < 0$, $j = 1, ..., m$.*

An inequality constrained program

$$\text{(IC)} \qquad \min\{f(x) : g_j(x) \le 0,\ j = 1, ..., m,\ x \in X\}$$

($f, g_1, ..., g_m$ are real-valued functions on X) is called to satisfy the Slater condition, if $g_1, ..., g_m$ satisfy this condition on X.

We are about to establish the following fundamental fact:

Theorem D.2.1 [Convex Theorem on Alternative]
Let $X \subset \mathbf{R}^n$ be convex, let $f, g_1, ..., g_m$ be real-valued convex functions on X, and let $g_1, ..., g_m$ satisfy the Slater condition on X. Then system (I) *is solvable if and only if system* (II) *is unsolvable.*

Proof. The first part of the statement – "if (II) has a solution, then (I) has no solutions" – is given by Proposition D.2.1. What we need is to prove the inverse statement. Thus, let us assume that (I) has no solutions, and let us prove that then (II) has a solution.

Without loss of generality we may assume that X is full-dimensional: $\operatorname{ri} X = \operatorname{int} X$ (indeed, otherwise we could replace our "universe" $\mathbf{R}^n$ with the affine span of X).

1^0. Let us set

$$F(x) = \begin{pmatrix} f(x) \\ g_1(x) \\ ... \\ g_m x) \end{pmatrix}$$

and consider two sets in $\mathbf{R}^{m+1}$:

$$S = \{u = (u_0, ..., u_m) \mid \exists x \in X : F(x) \le u\}$$

and

$$T = \{(u_0, ..., u_m) \mid u_0 < c, u_1 \le 0, u_2 \le 0, ..., u_m \le 0\}.$$

I claim that

- (i) S and T are nonempty convex sets;
- (ii) S and T does not intersect.

Indeed, convexity and nonemptiness of T is evident, same as nonemptiness of S. Convexity of S is an immediate consequence of the fact that X and $f, g_1, ..., g_m$ are convex. Indeed, assuming that $u', u'' \in S$, we conclude that there exist $x', x'' \in X$ such that $F(x') \le u'$ and $F(x'') \le u''$, whence, for every $\lambda \in [0, 1]$.

$$\lambda F(x') + (1 - \lambda)F(x'') \le \lambda u' + (1 - \lambda)u''.$$

The left hand side in this inequality, due to convexity of X and $f, g_1, ..., g_m$, is $\ge F(y)$, $y = \lambda x' + (1-\lambda)x''$. Thus, for the point $v = \lambda u' + (1 - \lambda)u''$ there exists $y \in X$ with $F(y) \le v$, whence $v \in S$. Thus, S is convex.

The fact that $S \cap T = \emptyset$ is an evident equivalent reformulation of the fact that (I) has no solutions.

2^0. Since S and T are nonempty convex sets with empty intersection, by Separation Theorem (Theorem B.2.5) they can be separated by a linear form: there exist $a = (a_0, ..., a_m) \ne 0$ such that

$$\inf_{u \in S} \sum_{j=0}^{m} a_j u_j \ge \sup_{u \in T} \sum_{j=0}^{m} a_j u_j. \tag{D.2.4}$$

3^0. Let us look what can be said about the vector a. I claim that, first,

$$a \geq 0 \tag{D.2.5}$$

and, second,

$$a_0 > 0. \tag{D.2.6}$$

Indeed, to prove (D.2.5) note that if some a_i were negative, then the right hand side in (D.2.4) would be $+\infty$ [2)], which is forbidden by (D.2.4).

Thus, $a \geq 0$; with this in mind, we can immediately compute the right hand side of (D.2.4):

$$\sup_{u\in T} \sum_{j=0}^{m} a_j u_j = \sup_{u_0<c, u_1,\ldots,u_m\leq 0} \sum_{j=0}^{m} a_j u_j = a_0 c.$$

Since for every $x \in X$ the point $F(x)$ belongs to S, the left hand side in (D.2.4) is not less that

$$\inf_{x\in X} \left[a_0 f(x) + \sum_{j=1}^{m} a_j g_j(x) \right];$$

combining our observations, we conclude that (D.2.4) implies the inequality

$$\inf_{x\in X} \left[a_0 f(x) + \sum_{j=1}^{m} a_j g_j(x) \right] \geq a_0 c. \tag{D.2.7}$$

Now let us prove that $a_0 > 0$. This crucial fact is an immediate consequence of the Slater condition. Indeed, let $\bar{x} \in X$ be the point given by this condition, so that $g_j(\bar{x}) < 0$. From (D.2.7) we conclude that

$$a_0 f(\bar{x}) + \sum_{j=0}^{m} a_j g_j(\bar{x}) \geq a_0 c.$$

If a_0 were 0, then the right hand side of this inequality would be 0, while the left one would be the combination $\sum_{j=0}^{m} a_j g_j(\bar{x})$ of *negative* reals $g_j(\bar{x})$ with *nonnegative* coefficients a_j *not all equal to 0* [3)], so that the left hand side is strictly negative, which is the desired contradiction.

4^0. Now we are done: since $a_0 > 0$, we are in our right to divide both sides of (D.2.7) by a_0 and thus get

$$\inf_{x\in X} \left[f_0(x) + \sum_{j=1}^{m} \lambda_j g_j(x) \right] \geq c, \tag{D.2.8}$$

where $\lambda_j = a_j/a_0 \geq 0$. Thus, (II) has a solution. ∎

D.2.2 Lagrange Function and Lagrange Duality

D.2.2.A. Lagrange function

The result of Convex Theorem on Alternative brings to our attention the function

$$\underline{L}(\lambda) = \inf_{x\in X} \left[f_0(x) + \sum_{j=1}^{m} \lambda_j g_j(x) \right], \tag{D.2.9}$$

[2)] look what happens when all coordinates in u, except the ith one, are fixed at values allowed by the description of T and u_i is a large in absolute value negative real

[3)] indeed, from the very beginning we know that $a \neq 0$, so that if $a_0 = 0$, then not all a_j, $j \geq 1$, are zeros

same as the aggregate

$$L(x,\lambda) = f_0(x) + \sum_{j=1}^{m} \lambda_j g_j(x) \tag{D.2.10}$$

from which this function comes. Aggregate (D.2.10) is called the *Lagrange function* of the inequality constrained optimization program

$$\text{(IC)} \qquad \min\{f(x) : g_j(x) \le 0,\ j = 1,...,m,\ x \in X\}.$$

The Lagrange function of an optimization program is a very important entity: most of optimality conditions are expressed in terms of this function. Let us start with translating of what we already know to the language of the Lagrange function.

D.2.2.B. Convex Programming Duality Theorem

Theorem D.2.2 *Consider an arbitrary inequality constrained optimization program* (IC)*. Then*

(i) *The infimum*

$$\underline{L}(\lambda) = \inf_{x \in X} L(x,\lambda)$$

of the Lagrange function in $x \in X$ is, for every $\lambda \ge 0$, a lower bound on the optimal value in (IC)*, so that the optimal value in the optimization program*

$$\text{(IC}^*\text{)} \qquad \sup_{\lambda \ge 0} \underline{L}(\lambda)$$

also is a lower bound for the optimal value in (IC)*;*

(ii) [Convex Duality Theorem] *If* (IC)

- *is convex,*
- *is below bounded*

and

- *satisfies the Slater condition,*

then the optimal value in (IC*) *is attained and is equal to the optimal value in* (IC)*.*

Proof. (i) is nothing but Proposition D.2.1 (why?). It makes sense, however, to repeat here the corresponding one-line reasoning:

> Let $\lambda \ge 0$; in order to prove that
>
> $$\underline{L}(\lambda) \equiv \inf_{x \in X} L(x,\lambda) \le c^* \quad [L(x,\lambda) = f(x) + \sum_{j=1}^{m} \lambda_j g_j(x)],$$
>
> where c^* is the optimal value in (IC), note that if x is feasible for (IC), then evidently $L(x,\lambda) \le f(x)$, so that the infimum of L over $x \in X$ is $\le$ the infimum c^* of f over the feasible set of (IC). □

(ii) is an immediate consequence of the Convex Theorem on Alternative. Indeed, let c^* be the optimal value in (IC). Then the system

$$f(x) < c^*, g_j(x) \le 0,\ j = 1,...,m$$

has no solutions in X, and by the above Theorem the system (II) associated with $c = c^*$ has a solution, i.e., there exists $\lambda^* \ge 0$ such that $\underline{L}(\lambda^*) \ge c^*$. But we know from (i) that the strict inequality here is impossible and, besides this, that $\underline{L}(\lambda) \le c^*$ for every $\lambda \ge 0$. Thus, $\underline{L}(\lambda^*) = c^*$ and λ^* is a maximizer of $\underline{L}$ over $\lambda \ge 0$. ■

C.1.2.C. Dual program

Theorem D.2.2 establishes certain connection between two optimization programs – the "primal" program

$$\text{(IC)} \qquad \min\{f(x) : g_j(x) \le 0,\ j = 1, ..., m,\ x \in X\}$$

and its *Lagrange dual program*

$$\text{(IC}^*\text{)} \qquad \max\left\{\underline{L}(\lambda) \equiv \inf_{x\in X} L(x,\lambda) : \lambda \ge 0\right\}$$

(the variables λ of the dual problem are called the *Lagrange multipliers* of the primal problem). The Theorem says that the optimal value in the dual problem is $\le$ the one in the primal, and under some favourable circumstances (the primal problem is convex below bounded and satisfies the Slater condition) the optimal values in the programs are equal to each other.

In our formulation there is some asymmetry between the primal and the dual programs. In fact both of the programs are related to the Lagrange function in a quite symmetric way. Indeed, consider the program

$$\min_{x\in X} \overline{L}(x), \quad \overline{L}(x) = \sup_{\lambda\ge 0} L(\lambda, x).$$

The objective in this program clearly is $+\infty$ at every point $x \in X$ which is not feasible for (IC) and is $f(x)$ on the feasible set of (IC), so that the program is equivalent to (IC). We see that both the primal and the dual programs come from the Lagrange function: in the primal problem, we minimize over X the result of maximization of $L(x,\lambda)$ in $\lambda \ge 0$, and in the dual program we maximize over $\lambda \ge 0$ the result of minimization of $L(x,\lambda)$ in $x \in X$. This is a particular (and the most important) example of a *zero sum two person game* – the issue we will speak about later.

We have seen that under certain convexity and regularity assumptions the optimal values in (IC) and (IC*) are equal to each. There is also another way to say when these optimal values are equal – this is always the case when the Lagrange function possesses a saddle point, i.e., there exists a pair $x^* \in X, \lambda^* \ge 0$ such that at the pair $L(x,\lambda)$ attains its minimum as a function of $x \in X$ and attains its maximum as a function of $\lambda \ge 0$:

$$L(x,\lambda^*) \ge L(x^*,\lambda^*) \ge L(x^*,\lambda) \quad \forall x \in X, \lambda \ge 0.$$

It can be easily demonstrated (do it by yourself or look at Theorem D.3.1) that

Proposition D.2.2 *(x^*,λ^*) is a saddle point of the Lagrange function L of* (IC) *if and only if x^* is an optimal solution to* (IC)*, λ^* is an optimal solution to* (IC*) *and the optimal values in the indicated problems are equal to each other.*

Our current goal is to extract from what we already know optimality conditions for convex programs.

D.2.3 Optimality Conditions in Convex Programming

D.2.3.A. Saddle point form of optimality conditions

Theorem D.2.3 [Saddle Point formulation of Optimality Conditions in Convex Programming]
Let (IC) *be an optimization program, $L(x,\lambda)$ be its Lagrange function, and let $x^* \in X$. Then*

(i) *A sufficient condition for x^* to be an optimal solution to* (IC) *is the existence of the vector of Lagrange multipliers $\lambda^* \ge 0$ such that (x^*,λ^*) is a saddle point of the Lagrange function $L(x,\lambda)$, i.e., a point where $L(x,\lambda)$ attains its minimum as a function of $x \in X$ and attains its maximum as a function of $\lambda \ge 0$:*

$$L(x,\lambda^*) \ge L(x^*,\lambda^*) \ge L(x^*,\lambda) \quad \forall x \in X, \lambda \ge 0. \tag{D.2.11}$$

(ii) *if the problem* (IC) *is convex and satisfies the Slater condition, then the above condition is necessary for optimality of x^*: if x^* is optimal for* (IC)*, then there exists $\lambda^* \ge 0$ such that (x^*,λ^*) is a saddle point of the Lagrange function.*

Proof. (i): assume that for a given $x^* \in X$ there exists $\lambda^* \geq 0$ such that (D.2.11) is satisfied, and let us prove that then x^* is optimal for (IC). First of all, x^* is feasible: indeed, if $g_j(x^*) > 0$ for some j, then, of course, $\sup_{\lambda \geq 0} L(x^*, \lambda) = +\infty$ (look what happens when all λ's, except λ_j, are fixed, and $\lambda_j \to +\infty$); but $\sup_{\lambda \geq 0} L(x^*, \lambda) = +\infty$ is forbidden by the second inequality in (D.2.11).

Since x^* is feasible, $\sup_{\lambda \geq 0} L(x^*, \lambda) = f(x^*)$, and we conclude from the second inequality in (D.2.11) that $L(x^*, \lambda^*) = f(x^*)$. Now the first inequality in (D.2.11) reads

$$f(x) + \sum_{j=1}^{m} \lambda_j^* g_j(x) \geq f(x^*) \quad \forall x \in X.$$

This inequality immediately implies that x^* is optimal: indeed, if x is feasible for (IC), then the left hand side in the latter inequality is $\leq f(x)$ (recall that $\lambda^* \geq 0$), and the inequality implies that $f(x) \geq f(x^*)$. □

(ii): Assume that (IC) is a convex program, x^* is its optimal solution and the problem satisfies the Slater condition; we should prove that then there exists $\lambda^* \geq 0$ such that (x^*, λ^*) is a saddle point of the Lagrange function, i.e., that (D.2.11) is satisfied. As we know from the Convex Programming Duality Theorem (Theorem D.2.2.(ii)), the dual problem (IC*) has a solution $\lambda^* \geq 0$ and the optimal value of the dual problem is equal to the optimal value in the primal one, i.e., to $f(x^*)$:

$$f(x^*) = \underline{L}(\lambda^*) \equiv \inf_{x \in X} L(x, \lambda^*). \tag{D.2.12}$$

We immediately conclude that

$$\lambda_j^* > 0 \Rightarrow g_j(x^*) = 0$$

(this is called *complementary slackness*: positive Lagrange multipliers can be associated only with active (satisfied at x^* as equalities) constraints. Indeed, from (D.2.12) it for sure follows that

$$f(x^*) \leq L(x^*, \lambda^*) = f(x^*) + \sum_{j=1}^{m} \lambda_j^* g_j(x^*);$$

the terms in the $\sum_j$ in the right hand side are nonpositive (since x^* is feasible for (IC)), and the sum itself is nonnegative due to our inequality; it is possible if and only if all the terms in the sum are zero, and this is exactly the complementary slackness.

From the complementary slackness we immediately conclude that $f(x^*) = L(x^*, \lambda^*)$, so that (D.2.12) results in

$$L(x^*, \lambda^*) = f(x^*) = \inf_{x \in X} L(x, \lambda^*).$$

On the other hand, since x^* is feasible for (IC), we have $L(x^*, \lambda) \leq f(x^*)$ whenever $\lambda \geq 0$. Combining our observations, we conclude that

$$L(x^*, \lambda) \leq L(x^*, \lambda^*) \leq L(x, \lambda^*)$$

for all $x \in X$ and all $\lambda \geq 0$. ■

Note that (i) is valid for an arbitrary inequality constrained optimization program, not necessarily convex. However, in the nonconvex case the *sufficient* condition for optimality given by (i) is extremely far from being necessary and is "almost never" satisfied. In contrast to this, in the convex case the condition in question is not only sufficient, but also "nearly necessary" – it for sure is necessary when (IC) is a convex program satisfying the Slater condition.

We are about to prove a modification of Theorem D.2.3, where we slightly relax the Slater condition.

Theorem D.2.4 *Consider a convex problem* (IC)*, and let x^* be a feasible solution of the problem. Assume that the functions $g_1, ..., g_k$ are affine, while the functions f, $g_{k+1}, ..., g_m$ are differentiable at x. Finally, assume the restricted Slater condition: there exists $\bar{x} \in \mathrm{ri}\, X$ such that $g_i(\bar{x}) \leq 0$ for $i \leq k$ and $g_i(\bar{x}) < 0$ for $i > k$. Then x_* is an optimal solution to* (IC) *if and only if there exists $\lambda^* \geq 0$ such that (x^*, λ^*) is a saddle point of $L(x, \lambda)$ on $X \times \{\lambda \geq 0\}$.*

Proof. The "if" part of the statement is given by Theorem D.2.3.(i). Let us focus on the "only if" part. Thus, assume that x^* is an optimal solution of (IC), and let us prove the existence of required λ_*. As always, we may assume without loss of generality that $\mathrm{int}\, X \neq \emptyset$. Let $I(x^*)$ be the set of indices of the constraints which are active at x^*. Consider the radial cone of X at x^*:

$$M_1 = \{h : \exists t > 0 : x^* + th \in X\}$$

along with the polyhedral cone

$$M_2 = \{h : (\nabla g_j^T(x^*)h \leq 0 \ \forall j \in I(x^*)\}.$$

We claim that

(I): M_2 is a closed cone which has a nonempty intersection with the interior of the convex cone M_1;

(II): the vector $\nabla f(x^*)$ belongs to the cone dual to the cone $M = M_1 \cap M_2$.

Postponing for the time being the proofs, let us derive from (I), (II) the existence of the required vector of Lagrange multipliers. Applying the Dubovitski-Milutin Lemma (Theorem B.2.4), which is legitimate due to (I), (II), we conclude that there exists a representation

$$\nabla f(x^*) = u + v, \quad u \in M_1', \, v \in M_2',$$

where M_i' is the cone dual to the cone M_i. By the Homogeneous Farkas Lemma, we have

$$v = -\sum_{j \in I(x^*)} \lambda_j^* \nabla g_j(x^*),$$

where $\lambda_j^* \geq 0$. Setting $\lambda_j^* = 0$ for $j \notin I(x^*)$, we get a vector $\lambda^* \geq 0$ such that

$$\begin{array}{c} \nabla_x \Big|_{x=x^*} L(x, \lambda^*) = \nabla f(x^*) + \sum_j \lambda_j^* \nabla g_j(x^*) = \nabla f(x^*) - v = u, \\ \lambda_j^* g_j(x^*) = 0, \, j = 1, ..., m. \end{array} \tag{D.2.13}$$

Since the function $L(x, \lambda^*)$ is convex in $x \in X$ and differentiable at x^*, the first relation in (D.2.13) combines with the inclusion $u \in M_1'$ and Proposition C.5.1 to imply that x^* is a minimizer of $L(x, \lambda^*)$ over $x \in X$. The second relation in (D.2.13) is the complementary slackness which, as we remember from the proof of Theorem D.2.3.(ii), combines with the feasibility of x^* to imply that λ^* is a maximizer of $L(x^*, \lambda)$ over $\lambda \geq 0$. Thus, (x^*, λ^*) is a saddle point of the Lagrange function, as claimed.

It remains to verify (I) and (II).

(I): the fact that M_2 is a closed cone is evident (M_2 is a polyhedral cone). The fact that M_1 is a convex cone with a nonempty interior is an immediate consequence of the convexity of X and the relation $\mathrm{int}\, X \neq \emptyset$. By assumption, there exists a point $\bar{x} \in \mathrm{int}\, X$ such that $g_j(\bar{x} \leq 0$ for all j. Since $\bar{x} \in \mathrm{int}\, X$,the vector $h = \bar{x} - x^*$ clearly belongs to $\mathrm{int}\, M_1$; since $g_j(x^*) = 0$, $j \in I(x^*)$, and $g_j(\bar{x}) \leq 0$, from Gradient Inequality it follows that $h^T \nabla g_j(x^*) \leq g_j(\bar{x}) - g_j(x^*) \leq 0$ for $j \in I(x^*)$, so that $h \in M_1$. Thus, h is the intersection of $\mathrm{int}\, M_1$ and M_2, so that this intersection is nonempty. □

(II): Assume, on the contrary to what should be proven, that there exists a vector $d \in M_1 \cap M_2$ such that $d^T \nabla f(x^*) < 0$. Let h be the same vector as in the proof of (I). Since $d^T \nabla f(x^*) < 0$, we can choose $\epsilon > 0$ such that with $d_\epsilon = d + \epsilon h$ one has $d_\epsilon^T \nabla f(x^*) < 0$. Since both d and h belong to M_1, there exists $\delta > 0$ such that $x_t = x^* + t d_\epsilon \in X$ for $0 \leq t \leq \delta$; since $d_\epsilon^T \nabla f(x^*) < 0$, we may further assume that $f(x_t) < f(x^*)$ when $0 < t \leq \delta$. Let us verify that for every $j \leq m$ one has

$(*_j)$: *There exists* $\delta_j > 0$ *such that* $g_j(x_t) \le 0$ *for* $0 \le t \le \delta_j$.

This will yield the desired contradiction, since, setting $t = \min[\delta, \min_j \delta_j]$, we would have $x_t \in X$, $g_j(x_t) \le 0$, $j = 1, ..., m$, $f(x_t) < f(x^*)$, which is impossible, since x^* is an optimal solution of (IC).

To prove $(*_j)$, consider the following three possibilities:

$j \notin I(x^*)$: here $(*_j)$ is evident, since $g_j(x)$ is negative at x^* and is continuous in $x \in X$ at the point x^* (recall that all g_j are assumed even to be differentiable at x^*).

$j \in I(x^*)$ and $j \le k$: For j in question, the function $g_j(x)$ is affine and vanishes at x^*, while $\nabla g_j(x^*)$ has nonpositive inner products with both d (due to $d \in M_2$) and h (due to $g_j(x^*) = 0$, $g_j(x^* + h) = g_j(\bar{x}) \le 0$); it follows that $\nabla g_j(x^*)$ has nonpositive inner product with d_ϵ, and since the function is affine, we arrive at $g_j(x^* + td_\epsilon) \le g_j(x^*) = 0$ for $t \ge 0$.

$j \in I(x^*)$ and $j > k$: In this case, the function $\gamma_j(t) = g_j(x^* + t_\epsilon)$ vanishes at $t = 0$ and is differentiable at $t = 0$ with the derivative $\gamma_j'(0) = (\epsilon h + d)^T \nabla g_j(x^*)$. This derivative is negative, since $d^T \nabla g_j(x^*) \le 0$ due to $d \in M_2$ and $j \in I(x^*)$, while by the Gradient Inequality $h^T \nabla g_j(x^*) \le g_j(x^* + h) - g_j(x^*) = g_j(\bar{x}) - g_j(x^*) \le g_j(\bar{x}) < 0$. Since $\gamma_j(0) = 0$, $\gamma_j'(0) < 0$, $\gamma_j(t)$ is negative for all small enough positive t, as required in $(*_j)$. ∎

D.2.3.B. Karush-Kuhn-Tucker form of optimality conditions

Theorems D.2.3, D.2.4 express, basically, the strongest optimality conditions for a Convex Programming program. These conditions, however, are "implicit" – they are expressed in terms of saddle point of the Lagrange function, and it is unclear how to verify that something is or is not the saddle point of the Lagrange function. Fortunately, the proof of Theorem D.2.4 yields more or less explicit optimality conditions as follows:

Theorem D.2.5 [Karush-Kuhn-Tucker Optimality Conditions in Convex Programming] *Let* (IC) *be a convex program, let* x^* *be its feasible solution, and let the functions* f, g_1,...,g_m *be differentiable at* x^*. *Then*

(i) [Sufficiency] *The Karush-Kuhn-Tucker condition:*

There exist nonnegative Lagrange multipliers λ_j^*, $j = 1, ..., m$, *such that*

$$\lambda_j^* g_j(x^*) = 0,\ j = 1, ..., m \quad \text{[complementary slackness]} \tag{D.2.14}$$

and

$$\nabla f(x^*) + \sum_{j=1}^{m} \lambda_j^* \nabla g_j(x^*) \in T_X^*(x^*) \tag{D.2.15}$$

(that is, $(x - x^*)^T \nabla f(x^*) + \sum_{j=1}^{m} \lambda_j^* \nabla g_j(x^*) \ge 0$ *for all* $x \in X$*)*

is sufficient for x^* *to be optimal solution to* (IC).

(ii) [Necessity and sufficiency] *If, in addition to the premise, the "restricted Slater assumption" holds, that is, there exists* $\bar{x} \in X$ *such that at* $\bar{x}$ *the nonlinear* g_j *are strictly negative, and linear* g_j *are nonpositive, then the Karush-Kuhn-Tucker condition from* (i) *is necessary and sufficient for* x^* *to be optimal solution to* (IC).

Proof. (i) is readily given by Theorem D.2.3.(ii); indeed, it is immediately seen that under the premise of Theorem D.2.5 the Karush-Kuhn-Tucker condition is sufficient for x^*, λ^*) to be a saddle point of the Lagrange function.

(ii) is contained in the proof of Theorem D.2.4. ∎

Note that the optimality conditions stated in Theorem C.5.2 and Proposition C.5.1 are particular cases of the above Theorem corresponding to $m = 0$.

D.3 Saddle Points

D.3.1 Definition and Game Theory interpretation

When speaking about the "saddle point" formulation of optimality conditions in Convex Programming, we touched a very interesting in its own right topic of Saddle Points. This notion is related to the situation as follows. Let $X \subset \mathbf{R}^n$ and $\Lambda \in \mathbf{R}^m$ be two nonempty sets, and let

$$L(x, \lambda) : X \times \Lambda \to \mathbf{R}$$

be a real-valued function of $x \in X$ and $\lambda \in \Lambda$. We say that a point $(x^*, \lambda^*) \in X \times \Lambda$ is a *saddle point* of L on $X \times \Lambda$, if L attains in this point its maximum in $\lambda \in \Lambda$ and attains at the point its minimum in $x \in X$:

$$L(x, \lambda^*) \geq L(x^*, \lambda^*) \geq L(x^*, \lambda) \quad \forall (x, \lambda) \in X \times \Lambda. \tag{D.3.1}$$

The notion of a saddle point admits natural interpretation in *game terms*. Consider what is called a *two person zero sum game* where player I chooses $x \in X$ and player II chooses $\lambda \in \Lambda$; after the players have chosen their decisions, player I pays to player II the sum $L(x, \lambda)$. Of course, I is interested to minimize his payment, while II is interested to maximize his income. What is the natural notion of the equilibrium in such a game – what are the choices (x, λ) of the players I and II such that every one of the players is not interested to vary his choice independently on whether he knows the choice of his opponent? It is immediately seen that the equilibria are exactly the saddle points of the cost function L. Indeed, if (x^*, λ^*) is such a point, than the player I is not interested to pass from x to another choice, given that II keeps his choice λ fixed: the first inequality in (D.3.1) shows that such a choice cannot decrease the payment of I. Similarly, player II is not interested to choose something different from λ^*, given that I keeps his choice x^* – such an action cannot increase the income of II. On the other hand, if (x^*, λ^*) is not a saddle point, then either the player I can decrease his payment passing from x^* to another choice, given that II keeps his choice at λ^* – this is the case when the first inequality in (D.3.1) is violated, or similarly for the player II; thus, equilibria are exactly the saddle points.

The game interpretation of the notion of a saddle point motivates deep insight into the structure of the set of saddle points. Consider the following two situations:

(A) player I makes his choice first, and player II makes his choice already knowing the choice of I;

(B) vice versa, player II chooses first, and I makes his choice already knowing the choice of II.

In the case (A) the reasoning of I is: If I choose some x, then II of course will choose λ which maximizes, for my x, my payment $L(x, \lambda)$, so that I shall pay the sum

$$\overline{L}(x) = \sup_{\lambda \in \Lambda} L(x, \lambda);$$

Consequently, my policy should be to choose x which minimizes my *loss function* L, i.e., the one which solves the optimization problem

$$\text{(I)} \qquad \min_{x \in X} \overline{L}(x);$$

with this policy my anticipated payment will be

$$\inf_{x \in X} \overline{L}(x) = \inf_{x \in X} \sup_{\lambda \in \Lambda} L(x, \lambda).$$

In the case (B), similar reasoning of II enforces him to choose λ maximizing his *profit function*

$$\underline{L}(\lambda) = \inf_{x \in X} L(x, \lambda),$$

i.e., the one which solves the optimization problem

$$\text{(II)} \qquad \max_{\lambda \in \Lambda} \underline{L}(\lambda);$$

with this policy, the anticipated profit of II is

$$\sup_{\lambda\in\Lambda}\underline{L}(\lambda)=\sup_{\lambda\in\Lambda}\inf_{x\in X}L(x,\lambda).$$

Note that these two reasonings relate to two *different* games: the one with priority of II (when making his decision, II already knows the choice of I), and the one with similar priority of I. Therefore we should not, generally speaking, expect that the anticipated loss of I in (A) is equal to the anticipated profit of II in (B). What can be guessed is that the anticipated loss of I in (B) is *less than or equal to* the anticipated profit of II in (A), since the conditions of the game (B) are better for I than those of (A). Thus, we may guess that independently of the structure of the function $L(x,\lambda)$, there is the inequality

$$\sup_{\lambda\in\Lambda}\inf_{x\in X}L(x,\lambda)\le\inf_{x\in X}\sup_{\lambda\in\Lambda}L(x,\lambda). \tag{D.3.2}$$

This inequality indeed is true; which is seen from the following reasoning:

$$\begin{array}{l}\forall y\in X:\quad\inf\limits_{x\in X}L(x,\lambda)\le L(y,\lambda)\Rightarrow\\ \forall y\in X:\sup\limits_{\lambda\in\Lambda}\inf\limits_{x\in X}L(x,\lambda)\le\sup\limits_{\lambda\in\Lambda}L(y,\lambda)\equiv\underline{L}(y);\end{array}$$

consequently, the quantity $\sup\limits_{\lambda\in\Lambda}\inf\limits_{x\in X}L(x,\lambda)$ is a lower bound for the function $\underline{L}(y)$, $y\in X$, and is therefore a lower bound for the infimum of the latter function over $y\in X$, i.e., is a lower bound for $\inf\limits_{y\in X}\sup\limits_{\lambda\in\Lambda}L(y,\lambda)$.

Now let us look what happens when the game in question has a saddle point (x^*,λ^*), so that

$$L(x,\lambda^*)\ge L(x^*,\lambda^*)\ge L(x^*,\lambda)\quad\forall(x,\lambda)\in X\times\Lambda. \tag{D.3.3}$$

I claim that if it is the case, then

(*) *x^* is an optimal solution to* (I), *λ^* is an optimal solution to* (II) *and the optimal values in these two optimization problems are equal to each other* (and are equal to the quantity $L(x^*,\lambda^*)$).

Indeed, from (D.3.3) it follows that

$$\underline{L}(\lambda^*)\ge L(x^*,\lambda^*)\ge\overline{L}(x^*),$$

whence, of course,

$$\sup_{\lambda\in\Lambda}\underline{L}(\lambda)\ge\underline{L}(\lambda^*)\ge L(x^*,\lambda^*)\ge\overline{L}(x^*)\ge\inf_{x\in X}\overline{L}(x).$$

the very first quantity in the latter chain is $\le$ the very last quantity by (D.3.2), which is possible if and only if all the inequalities in the chain are equalities, which is exactly what is said by (A) and (B).

Thus, if (x^*,λ^*) is a saddle point of L, then (*) takes place. We are about to demonstrate that the inverse also is true:

Theorem D.3.1 [Structure of the saddle point set] *Let $L:X\times Y\to\mathbf{R}$ be a function. The set of saddle points of the function is nonempty if and only if the related optimization problems* (I) *and* (II) *are solvable and the optimal values in the problems are equal to each other. If it is the case, then the saddle points of L are exactly all pairs (x^*,λ^*) with x^* being an optimal solution to* (I) *and λ^* being an optimal solution to* (II)*, and the value of the cost function $L(\cdot,\cdot)$ at every one of these points is equal to the common optimal value in* (I) *and* (II)*.*

Proof. We already have established "half" of the theorem: if there are saddle points of L, then their components are optimal solutions to (I), respectively, (II), and the optimal values in these two problems are equal to each other and to the value of L at the saddle point in question. To complete the proof, we should demonstrate that if x^* is an optimal solution to (I), λ^* is an optimal solution to (II) and the

optimal values in the problems are equal to each other, then (x^*, λ^*) is a saddle point of L. This is immediate: we have

$$\begin{array}{rcll} L(x,\lambda^*) & \geq & \underline{L}(\lambda^*) & [\text{ definition of } \underline{L}] \\ & = & \overline{L}(x^*) & [\text{by assumption}] \\ & \geq & L(x^*,\lambda) & [\text{definition of } \overline{L}] \end{array}$$

whence

$$L(x,\lambda^*) \geq L(x^*,\lambda) \quad \forall x \in X, \lambda \in \Lambda;$$

substituting $\lambda = \lambda^*$ in the right hand side of this inequality, we get $L(x,\lambda^*) \geq L(x^*,\lambda^*)$, and substituting $x = x^*$ in the right hand side of our inequality, we get $L(x^*,\lambda^*) \geq L(x^*,\lambda)$; thus, (x^*,λ^*) indeed is a saddle point of L. ∎

D.3.2 Existence of Saddle Points

It is easily seen that a "quite respectable" cost function may have no saddle points, e.g., the function $L(x,\lambda) = (x-\lambda)^2$ on the unit square $[0,1] \times [0,1]$. Indeed, here

$$\underline{L}(x) = \sup_{\lambda \in [0,1]} (x-\lambda)^2 = \max\{x^2, (1-x)^2\},$$

$$\overline{L}(\lambda) = \inf_{x \in [0,1]} (x-\lambda)^2 = 0, \ \lambda \in [0,1],$$

so that the optimal value in (I) is $\frac{1}{4}$, and the optimal value in (II) is 0; according to Theorem D.3.1 it means that L has no saddle points.

On the other hand, there are generic cases when L has a saddle point, e.g., when

$$L(x,\lambda) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) : X \times \mathbf{R}^m_+ \to \mathbf{R}$$

is the Lagrange function of a solvable convex program satisfying the Slater condition. Note that in this case L is convex in x for every $\lambda \in \Lambda \equiv \mathbf{R}^m_+$ and is linear (and therefore concave) in λ for every fixed X. As we shall see in a while, these are the structural properties of L which take upon themselves the "main responsibility" for the fact that in the case in question the saddle points exist. Namely, there exists the following

Theorem D.3.2 [Existence of saddle points of a convex-concave function (Sion-Kakutani)] *Let X and Λ be convex compact sets in $\mathbf{R}^n$ and $\mathbf{R}^m$, respectively, and let*

$$L(x,\lambda) : X \times \Lambda \to \mathbf{R}$$

be a continuous function which is convex in $x \in X$ for every fixed $\lambda \in \Lambda$ and is concave in $\lambda \in \Lambda$ for every fixed $x \in X$. Then L has saddle points on $X \times \Lambda$.

Proof. According to Theorem D.3.1, we should prove that

- (i) Optimization problems (I) and (II) are solvable
- (ii) the optimal values in (I) and (II) are equal to each other.

(i) is valid independently of convexity-concavity of L and is given by the following routine reasoning from the Analysis:

Since X and Λ are compact sets and L is continuous on $X \times \Lambda$, due to the well-known Analysis theorem L is uniformly continuous on $X \times \Lambda$: for every $\epsilon > 0$ there exists $\delta(\epsilon) > 0$ such that

$$|x - x'| + |\lambda - \lambda'| \leq \delta(\epsilon) \Rightarrow |L(x,\lambda) - L(x',\lambda')| \leq \epsilon \quad [4] \tag{D.3.4}$$

[4] for those not too familiar with Analysis, I wish to stress the difference between the usual continuity and the uniform continuity: continuity of L means that given $\epsilon > 0$ *and a point* (x,λ), it is possible to choose $\delta > 0$ such that (D.3.4) is valid; the corresponding δ may depend on (x,λ), not only on ϵ. Uniform continuity means that this positive δ may be chosen as a function of ϵ only. The fact that a continuous on a compact set function automatically is uniformly continuous on the set is one of the most useful features of compact sets

In particular,

$$|x - x'| \leq \delta(\epsilon) \Rightarrow |L(x,\lambda) - L(x'\lambda)| \leq \epsilon,$$

whence, of course, also

$$|x - x'| \leq \delta(\epsilon) \Rightarrow |\overline{L}(x) - \overline{L}(x')| \leq \epsilon,$$

so that the function $\overline{L}$ is continuous on X. Similarly, $\underline{L}$ is continuous on Λ. Taking in account that X and Λ are compact sets, we conclude that the problems (I) and (II) are solvable.

(ii) is the essence of the matter; here, of course, the entire construction heavily exploits convexity-concavity of L.

0^0. To prove (ii), we first establish the following statement, which is important by its own right:

Lemma D.3.1 [Minmax Lemma] *Let X be a convex compact set and $f_0, ..., f_N$ be a collection of $N+1$ convex and continuous functions on X. Then the minmax*

$$\min_{x\in X} \max_{i=0,...,N} f_i(x) \tag{D.3.5}$$

of the collection is equal to the minimum in $x \in X$ of certain convex combination of the functions: there exist nonnegative μ_i, $i = 0, ..., N$, with unit sum such that

$$\min_{x\in X} \max_{i=0,...,N} f_i(x) = \min_{x\in X} \sum_{i=0}^{N} \mu_i f_i(x)$$

Remark D.3.1 Minimum of *every* convex combination of a collection of *arbitrary* functions is $\leq$ the minmax of the collection; this evident fact can be also obtained from (D.3.2) as applied to the function

$$M(x,\mu) = \sum_{i=0}^{N} \mu_i f_i(x)$$

on the direct product of X and the standard simplex

$$\Delta = \{\mu \in \mathbf{R}^{N+1} \mid \mu \geq 0, \sum_i \mu_i = 1\}.$$

The Minmax Lemma says that if f_i are convex and continuous on a convex compact set X, then the indicated inequality is in fact equality; you can easily verify that this is nothing but the claim that the function M possesses a saddle point. Thus, the Minmax Lemma is in fact a particular case of the Sion-Kakutani Theorem; we are about to give a direct proof of this particular case of the Theorem and then to derive the general case from this particular one.

Proof of the Minmax Lemma. Consider the optimization program

$$(S) \quad t \to \min \mid f_0(x) - t \leq 0,\ f_1(x) - t \leq 0, ..., f_N(x) - t \leq 0,\ x \in X.$$

This clearly is a convex program with the optimal value

$$t^* = \min_{x\in X} \max_{i=0,...,N} f_i(x)$$

(note that (t, x) is feasible solution for (S) if and only if $x \in X$ and $t \geq \max_{i=0,...,N} f_i(x)$). The problem clearly satisfies the Slater condition and is solvable (since X is compact set and f_i, $i = 0, ..., N$, are continuous on X; therefore their maximum also is continuous on X and thus attains its minimum on the compact set X). Let (t^*, x^*) be an optimal solution to the problem. According to Theorem D.2.3, there exists $\lambda^* \geq 0$ such that $((t^*, x^*), \lambda^*)$ is a saddle point of the corresponding Lagrange function

$$L(t, x; \lambda) = t + \sum_{i=0}^{N} \lambda_i (f_i(x) - t) = t(1 - \sum_{i=0}^{N} \lambda_i) + \sum_{i=0}^{N} \lambda_i f_i(x),$$

and the value of this function at $((t^*, x^*), \lambda^*)$ is equal to the optimal value in (S), i.e., to t^*.

Now, since $L(t, x; \lambda^*)$ attains its minimum in (t, x) over the set $\{t \in \mathbf{R}, x \in X\}$ at (t^*, x^*), we should have

$$\sum_{i=0}^{N} \lambda_i^* = 1$$

(otherwise the minimum of L in (t, x) would be $-\infty$). Thus,

$$[\min_{x\in X} \max_{i=0,...,N} f_i(x) =] \quad t^* = \min_{t\in\mathbf{R}, x\in X} \left[t \times 0 + \sum_{i=0}^{N} \lambda_i^* f_i(x) \right],$$

so that

$$\min_{x\in X} \max_{i=0,...,N} f_i(x) = \min_{x\in X} \sum_{i=0}^{N} \lambda_i^* f_i(x)$$

with some $\lambda_i^* \geq 0$, $\sum_{i=0}^{N} \lambda_i^* = 1$, as claimed. ∎

From the Minmax Lemma to the Sion-Kakutani Theorem. We should prove that the optimal values in (I) and (II) (which, by (i), are well defined reals) are equal to each other, i.e., that

$$\inf_{x\in X} \sup_{\lambda\in\Lambda} L(x, \lambda) = \sup_{\lambda\in\Lambda} \inf_{x\in X} L(x, \lambda).$$

We know from (D.3.4) that the first of these two quantities is greater than or equal to the second, so that all we need is to prove the inverse inequality. For me it is convenient to assume that the right quantity (the optimal value in (II)) is 0, which, of course, does not restrict generality; and all we need to prove is that the left quantity – the optimal value in (I) – cannot be positive.

1^0. What does it mean that the optimal value in (II) is zero? When it is zero, then the function $\underline{L}(\lambda)$ is nonpositive for every λ, or, which is the same, the convex continuous function of $x \in X$ – the function $L(x, \lambda)$ – has nonpositive minimal value over $x \in X$. Since X is compact, this minimal value is achieved, so that the set

$$X(\lambda) = \{x \in X \mid L(x, \lambda) \leq 0\}$$

is nonempty; and since X is convex and L is convex in $x \in X$, the set $X(\lambda)$ is convex (as a level set of a convex function, Proposition C.1.4). Note also that the set is closed (since X is closed and $L(x, \lambda)$ is continuous in $x \in X$).

2^0. Thus, if the optimal value in (II) is zero, then the set $X(\lambda)$ is a nonempty convex compact set for every $\lambda \in \Lambda$. And what does it mean that the optimal value in (I) is nonpositive? It means exactly that there is a point $x \in X$ where the function $\overline{L}$ is nonpositive, i.e., the point $x \in X$ where $L(x, \lambda) \leq 0$ for all $\lambda \in \Lambda$. In other words, to prove that the optimal value in (I) is nonpositive is the same as to prove that *the sets $X(\lambda)$, $\lambda \in \Lambda$, have a point in common.*

3^0. With the above observations we see that the situation is as follows: we are given a family of closed nonempty convex subsets $X(\lambda)$, $\lambda \in \Lambda$, of a compact set X, and we should prove that these sets have a point in common. To this end, in turn, it suffices to prove that every *finite* number of sets from our family have a point in common (to justify this claim, I can refer to the Helley Theorem II, which gives us much stronger result: to prove that all $X(\lambda)$ have a point in common, it suffices to prove that every $(n + 1)$ sets of this family, n being the affine dimension of X, have a point in common). Let $X(\lambda_0), ..., X(\lambda_N)$ be $N + 1$ sets from our family; we should prove that the sets have a point in common. In other words, let

$$f_i(x) = L(x, \lambda_i),\ i = 0, ..., N;$$

all we should prove is that there exists a point x where all our functions are nonpositive, or, which is the same, that the minmax of our collection of functions – the quantity

$$\alpha \equiv \min_{x\in X} \max_{i=1,...,N} f_i(x)$$

– is nonpositive.

The proof of the inequality $\alpha \leq 0$ is as follows. According to the Minmax Lemma (which can be applied in our situation – since L is convex and continuous in x, all f_i are convex and continuous, and X is compact), α is the minimum in $x \in X$ of certain convex combination $\phi(x) = \sum_{i=0}^{N} \nu_i f_i(x)$ of the functions $f_i(x)$. We have

$$\phi(x) = \sum_{i=0}^{N} \nu_i f_i(x) \equiv \sum_{i=0}^{N} \nu_i L(x, \lambda_i) \leq L(x, \sum_{i=0}^{N} \nu_i \lambda_i)$$

(the last inequality follows from concavity of L in λ; this is the only – and crucial – point where we use this assumption). We see that $\phi(\cdot)$ is majorated by $L(\cdot, \lambda)$ for a properly chosen λ; it follows that the minimum of ϕ in $x \in X$ – and we already know that this minimum is exactly α – is nonpositive (recall that the minimum of L in x is nonpositive for every λ). ∎