INTERIOR POINT

POLYNOMIAL TIME METHODS

IN

CONVEX PROGRAMMING

A. Nemirovski

Spring Semester 1996

Interior Point Polynomial Methods in Convex Programming

Goals. During the last decade the area of interior point polynomial methods (started in 1984 when N. Karmarkar invented his famous algorithm for Linear Programming) became one of the dominating fields, or even *the* dominating field, of theoretical and computational activity in Convex Optimization. The goal of the course is to present a general theory of interior point polynomial algorithms in Convex Programming. The theory allows to explain all known methods of this type and to extend them from the initial area of interior point technique - Linear and Quadratic Programming - onto a wide variety of essentially nonlinear classes of convex programs.

We present in a self-contained manner the basic theory along with its applications to several important classes of convex programs (LP, QP, Quadratically constrained Quadratic programming, Geometrical programming, Eigenvalue problems, etc.)

The course follows the recent book
Yu. Nesterov, A. Nemirovski *Interior-Point Polynomial Algorithms in Convex Programming* SIAM Studies in Applied Mathematics, 1994

Prerequisites for the course are the standard Calculus and the most elementary parts of Convex Analysis.

Duration: one semester, 2 hours weekly

Contents:

I decided to add to the course three Appendices:

Appendices I and II contain two recent papers on the subject; The first of them develops the approach to the design of long-step interior point methods discussed, in its simplest form, in Lecture 8. The second paper is devoted to the particular application – Truss Topology Design – discussed in the exercises to Lecture 5 (Section 5.5). I think the ability to go through these papers is a good indication of mastering the course.

Appendix III is a 4-lecture Minicourse on polynomial time methods in Convex Programming which I wrote for the 1995i Summer AMS Seminar on Mathematics of Numerical Analysis (June-August 1995, Park City, Utah, USA). It can be regarded as a “technicality-free” summary of the main body of the Course a summary which contains also some new details. In fact I think it makes sense to start reading with this summary. Appendix III (same as Appendices I and II) can be read independently of other parts of the text.

About Exercises

The majority of Lectures are accompanied by the "Exercise" sections. In several cases, the exercises are devoted to the lecture where they are placed; sometimes they prepare the reader to the next lecture.

The mark * at the word "Exercise" or at an item of an exercise means that you may use hints given in Appendix "Hints". A hint, in turn, may refer you to the solution of the exercise given in the Appendix "Solutions"; this is denoted by the mark $^+$. Some exercises are marked by $^+$ rather than by *; this refers you directly to the solution of an exercise.

Exercises marked by $^\#$ are closely related to the lecture where they are placed; it would be a good thing to solve such an exercise or at least to become acquainted with its solution (if any is given).

Exercises which I find difficult are marked with $^>$.

The exercises, usually, are not that simple. They in no sense are obligatory, and the reader is not expected to solve all or even the majority of the exercises. Those who would like to work on the solutions should take into account that the order of exercises is important: a problem which could cause serious difficulties as it is becomes much simpler in the context (at least I hope so).

Contents

Chapter 1

Introduction to the Course

What we are about to study in this semester are the theory and the applications of interior point polynomial time methods in Convex Programming. Today, in the introductory lecture, I am not going to prove theorems and present algorithms. My goal is to explain what the course is about, what are the interior point methods and why so many researchers and practitioners are now deeply involved in this new area.

1.1 Some history

The modern theory of polynomial time interior point methods takes its origin in the seminal paper of Narendra Karmarkar published in 1984. Now, after 10 years, there are hundreds of researchers working in the area, and thousands of papers and preprints on the subject. The electronic bibliography on interior point methods collected and maintained by Dr. Eberhard Kranich, although far from being complete, contains now over 1,500 entries. For Optimization Community which covers not so many people, this is a tremendous concentration of effort in a single area, for sure incomparable with all happened in the previous years.

Although to the moment the majority of the papers on interior point methods deal with the theoretical issues, the practical yield also is very remarkable. It suffices to say that the Karmarkar algorithm for Linear Programming was used as the working horse for the US Army logistic planning (i.e., planning of all kinds of supplies) in the Gulf War. Another interior point method for Linear Programming, the so called primal-dual one, forms the nucleus of an extremely efficient and very popular now software package OSL2. Let me present you a citation from G. Dantzig: "At the present time (1990), interior algorithms are in open competition with variants of the simplex methods"[1)]. It means something when new-borned methods can be competitive against an extremely powerful and polished for almost 50 years by thousands of people Simplex method.

Now let me switch from the style of advertisements to the normal one. What actually happened in 1984, was the appearance of a new iterative polynomial-time algorithm for Linear Programming. We already know what does it mean "a polynomial time algorithm for LP" - recall the lecture about the Ellipsoid method and the Khachiyan theorem on polynomial solvability of LP. As we remember, Khachiyan proved in 1979 that Linear Programming is polynomially solvable, namely, that an LP problem with rational coefficients, m inequality constraints and n variables can be solved exactly in $O(n^3(n+m)L)$ arithmetic operations, L being the input length of the problem, i.e., the total binary length of the numerical data specifying the problem instance. The new method of Karmarkar possessed the complexity bound of $O(m^{3/2}n^2L)$ operations. In the standard for the complexity analysis case of more or less "square" problems $m = O(n)$ the former estimate becomes $O(n^4L)$, the latter $O(n^{3.5}L)$. Thus, there was some progress in the complexity. And it can be said for sure that neither this moderate progress, nor remarkable elegance of the new algorithm never could cause the revolution in Optimization. What indeed was a sensation, what inspired extremely intensive activity in the new area and in a few years resulted in significant theoretical and computational progress, was the claim that the new algorithm in

[1] *History of Mathematica Programming*, J.K. Lenstra. A.H.G. Rinnooy Kan, A. Schrijver, Eds. CWI, North-Holland, 1991

real-world computations was by order of magnitudes more efficient than the Simplex method. Let me explain you why this was a sensation. It is known that the Simplex method is not polynomial: there exist bad problem instances where the number of pivotings grows exponentially with the dimension of the instance. Thus, any polynomial time algorithm for LP, the Ellipsoid one, the method of Karmarkar or whatever else, for sure is incomparably better in its worst-case behaviour than the Simplex. But this is the theoretical worst-case behaviour which, as is demonstrated by almost 50-year practice, never occurs in real-world applications; from the practical viewpoint, the Simplex method is an extremely efficient algorithm with fairy low empirical complexity; this is why the method is able to solve very large-scale real world LP problems in reasonable time. In contrast to this, the Ellipsoid method works more or less in accordance with its theoretical worst-case complexity bound, so that in practical computations this "theoretically good" method is by far dominated by the Simplex even on very small problems with tens of variables and constraints. If the method of Karmarkar would also behave itself according to its theoretical complexity bound, it would be only slightly better then the Ellipsoid method and still would be incomparably worse than the Simplex. The point, anyhow, is that actual behaviour of the method of Karmarkar turned out to be much better than it is said by the worst-case theoretical complexity bound. This phenomenon combined with the theoretical advantages of a polynomial time algorithm, not the latter advantages alone, (same as, I believe, not the empirical behaviour of the method alone), inspired an actual revolution in optimization which continues up today and hardly will terminate in the nearest future.

I have said something about the birth of the "interior point science". As it often happens in our field, later it turned out that this was the second birth; the first one was in 1967 in Russia, where Ilya Dikin, then the Ph.D. student of Leonid Kantorovich, invented what is now called the affine scaling algorithm for LP. This algorithm which hardly is theoretically polynomial, is certain simplification of the method of Karmarkar which shares all practical advantages of the basic Karmarkar algorithm; thus, as a computational tool, interior point methods exist at least since 1967. A good question is why this computational tool which is in extreme fashion now was completely overlooked in the West, same as in Russia. I think that this happened due to two reasons: first, Dikin came too early, when there was no interest to iterative procedures for LP - a new-borned iterative procedure, even of a great potential, hardly could overcome as a practical tool perfectly polished Simplex method, and the theoretical complexity issues in these years did not bother optimization people (even today we do not know whether the theoretical complexity of the Dikin algorithm is better than that one of the Simplex; and in 1967 the question itself hardly could occur). Second, the Dikin algorithm appeared in Russia, where there were neither hardware base for Dikin to perform large-scale tests of his algorithm, nor "social demand" for solving large-scale LP problems, so it was almost impossible to realize the practical potential of the new algorithm and to convince people in Russia, not speaking about the West, that this is something which worths attention.

Thus, although the prehistory of the interior point technique for LP started in 1967, the actual history of this subject started only in 1984. It would be impossible to outline numerous significant contributions to the field done since then; it would require mentioning tens, if not hundreds, of authors. There is, anyhow, one contribution which must be indicated explicitly. I mean the second cornerstone of the subject, the paper of James Renegar (1986) where the first *path-following* polynomial time interior point method for LP was developed. The efficiency estimate of this method was better than that one of the method of Karmarkar, namely, $O(n^3L)$ [2] - cubic in the dimension, same as for classical methods of solving systems of linear equations; up to now this is the best known theoretical complexity bound for LP. Besides this remarkable theoretical advantage, the method of Renegar possesses an important advantage in, let me say, the human dimension: the method belongs to a quite classical and well-known in Optimization scheme, in contrast to rather unusual Ellipsoid and Karmarkar algorithms. The paper of Renegar was extremely important for the understanding of the new methods and it, same as a little bit later independent paper of Clovis Gonzaga with close result, brought the area in the position very favourable for future developments.

To the moment I was speaking about interior point methods for Linear Programming, and this reflects the actual history of the subject: not only the first interior point methods vere developed for this case, but till the very last years the main activity, both theoretical and computational, in the field was focused on Linear Programming and the very close to it Linearly constrained Quadratic Programming. To extend the

[2] recall that we are speaking about "almost square" problems with the number of inequalities m being of order of the number of variables n

approach to more general classes of problems, it was actually a challenge: the original constructions and proofs heavily exploited the polyhedral structure of the feasible domain of an LP problem, and in order to pass to the nonlinear case, it required to realize what is the deep intrinsic nature of the methods. This latter problem was solved in a series of papers of Yurii Nesterov in 1988; the ideas of these papers form the basis of the theory the course is devoted to, the theory which now has became a kind of standard for unified explanation and development of polynomial time interior point algorithms for convex problems, both linear and nonlinear. To present this theory and its applications, this is the goal of my course. In the remaining part of this introductory lecture I am going to explain what we are looking for and what will be our general strategy.

1.2 The goal: poynomial time methods

I have declared that the purpose of the theory to be presented is developing of polynomial time algorithms for convex problems. Let me start with explaining what a polynomial time method is. Consider a family of convex problems

$$(p): \qquad \text{minimize } f(x) \text{ s.t. } g_j(x) \leq 0, \ i = 1, ..., m, \ x \in G$$

of a given analytical structure, like the family of LP problems, or Linearly constrained Quadratic problems, or Quadratically constrained Quadratic ones, etc. The only formal assumption on the family is that a problem instance p from it is identified by a finite-dimensional *data vector* $\mathcal{D}(p)$; normally you can understand this vector as the collection of the numeric coefficients in analytical expressions for the objective and the constraints; these expressions themselves are fixed by the description of the family. The dimension of the data vector is called the *size* $l(p)$ of the problem instance. A numerical method for solving problems from the family is a routine which, given on input the data vector, generates a sequence of approximate solutions to the problem in such a way that every of these solutions is obtained in finitely many operations of precise real arithmetic, like the four arithmetic operations, taking square roots, exponents, logarithms and other elementary functions; each operand in an operation is either an entry of the data vector, or the result of one of the preceding operations. We call a numerical method *convergent*, if, for any positive ε and for any problem instance p from the family, the approximate solutions x_i generated by the method, starting with certain $i = i^*(\varepsilon, p)$, are ε-solutions to the problem, i.e., they belong to G and satisfy the relations

$$f(x_i) - f^* \leq \varepsilon, \ g_j(x_i) \leq \varepsilon, \ j = 1, ..., m,$$

(f^* is the optimal value in the problem). We call a method *polynomial*, if it is convergent and the arithmetic cost $C(\varepsilon, p)$ of ε-solution, i.e., the total number of arithmetic operations at the first $i^*(\varepsilon, p)$ steps of the method as applied to p, admits an upper bound as follows:

$$C(\varepsilon, p) \leq \pi(l(p)) \ln\left(\frac{\mathcal{V}(p)}{\varepsilon}\right),$$

where π is certain polynomial independent on the data and $\mathcal{V}(p)$ is certain data-dependent *scale factor*. The ratio $\mathcal{V}(p)/\varepsilon$ can be interpreted as the relative accuracy which corresponds to the absolute accuracy ε, and the quantity $\ln(\frac{\mathcal{V}(p)}{\varepsilon})$ can be thought of as the number of accuracy digits in ε-solution. With this interpretation, the polynomiality of a method means that for this method *the arithmetic cost of an accuracy digit is bounded from above by a polynomial of the problem size, and this polynomial can be thought of as the characteristic of the complexity of the method.*

It is reasonable to compare this approach with the information-based approach we dealt with in the previous course. In the information-based complexity theory the problem was assumed to be represented by an oracle, by a black box, so that a method, starting its work, had no information on the instance; this information was accumulated via sequential calls to the oracle, and the number of these calls sufficient to find an ε-solution was thought of as the complexity of the method; we did not include in this complexity neither the computational effort of the oracle, nor the arithmetic cost of processing the answers of the oracle by the method. In contrast to this, in our now approach the data specifying the problem instance form the input to the method, so that the method from the very beginning possesses complete global information on the problem instance. What the method should do is to transform this input information into ε-solution to the problem, and the complexity of the method (which now might be called *algorithmic* or *combinatorial* complexity) is defined by the arithmetic cost of this transformation. It is clear that our new approach is not as general as the information-based one, since now we can speak only on families of problems of a reasonable analytic structure (otherwise the notion of the data vector becomes senseless). As a compensation, the combinatorial complexity is much more adequate measure of the actual computational effort than the information-based complexity.

After I have outlined what is our final goals, let me give you an idea of how this goal will be achieved. In what follows we will develop methods of two different types: the *path-following* and the *potential reduction* ones; the LP prototypes of these methods are, respectively, the methods of Renegar and Gonzaga, which are path-following routines, and the method of Karmarkar, which is a potential reduction one. In contrast to the actual historical order, we shall start with the quite traditional path-following scheme, since we are unprepared to understand what in fact happens in the methods of the Karmarkar type.

1.3 The path-following scheme

The, let me say, "classical" stage in developing the scheme is summarized in the seminal monograph of Fiacco and McCormic (1967). Assume we intend to solve a convex program

$$(P): \qquad \textit{minimize } f(x) \textit{ s.t. } g_i(x) \leq 0, \ i = 1, ..., m$$

associated with smooth (at least twice continuously defferentiable) convex functions f, g_i on $\mathbf{R}^n$. Let

$$G = \{x \in \mathbf{R}^n \mid g_i(x) \leq 0\}$$

be the feasible domain of the problem; assume for the sake of simplicity that this domain is bounded, and let the constraints $\{g_i\}$ satisfy the Slater condition:

$$\exists x: \ g_i(x) < 0, \ i = 1, ..., m.$$

Under these assumptions the feasible domain G is a solid - a closed and bounded convex set in $\mathbf{R}^n$ with a nonempty interior.

In 60's people believed that it is not difficult to solve *unconstrained* smooth convex problems, and it was very natural to try to reduce the constrained problem (P) to a series of unconstrained problems. To this end it was suggested to associate with the feasible domain G of problem (P) a *barrier* - an interior penalty function $F(x)$, i.e., a smooth convex function F defined on the interior of G and tending to ∞ when we approach from inside the boundary of G:

$$\lim_{i\to\infty} F(x_i) = \infty \textit{ for any sequence } \{x_i \in \operatorname{int} G\} \textit{ with } \lim_{i\to\infty} x_i \in \partial G.$$

It is also reasonble to assume that F is nondegenerate, i.e.,

$$F''(x) > 0, \ x \in \operatorname{int} G$$

(here > 0 stands for "positive definite").

Given such a barrier, one can associate with it and with the objective f of (P) the *barrier-generated family* comprised of the problems

$$(P_t): \qquad \textit{minimize } F_t(x) \equiv tf(x) + F(x).$$

Here the *penalty parameter* t is positive. Of course, x in (P_t) is subject to the "induced" restriction $x \in \operatorname{int} G$, since F_t is outside the latter set.

From our assumptions on G it immediately follows that

a) every of the problems (P_t) has a unique solution $x^*(t)$; this solution is, of course, in the interior of G;

b) the *path* $x^*(t)$ of solutions to (P_t) is a continuous function of $t \in [0, \infty)$, and all its limiting, as $t \to \infty$, points belong to the set of optimal solutions to (P).

It immediately follows that if we are able to *follow the path* $x^*(t)$ along certain sequence $t_i \to \infty$ of values of the penalty parameter, i.e., know how to form "good enough" approximations $x_i \in \operatorname{int} G$ to the points $x^*(t_i)$, say, such that

$$x_i - x^*(t_i) \to 0, \; i \to \infty, \tag{1.1}$$

then we know how to solve (P): b) and (1.1) imply that all limiting points of the sequance of our iterates $\{x_i\}$ belong to the optimal set of (P).

Now, to be able to meet the requirement (1.1) is, basically, the same as to be able to solve to a prescribed accuracy each of the "penalized" problems (P_t). What are our abilities in this respect? (P_t) is a minimization problem with smooth and nondegenerate (i.e., with nonsingular Hessian) objective. Of course, this objective is defined on the proper open convex subset of $\mathbf{R}^n$ rather than on the whole $\mathbf{R}^n$, so that the problem, rigorously speaking, is a constrained one, same as the initial problem (P). The constrained nature of (P_t) is, anyhow, nothing but an illusion: the solution to the problem is unique and belongs to the interior of G, and any converging minimization method of a relaxation type (i.e., monotonically decreasing the value of the objective along the sequence of iterates) started in an interior point of G would automatically keep the iterates away from the boundary of G (since $F_t \to \infty$ together with F as the argument approaches the boundary from inside); thus, qualitatively speaking, the behaviour of the method as applied to (P_t) would be the same as if the objective F_t was defined everywhere. In other words, we have basically the same possibilities to solve (P_t) as if it was an unconstrained problem with smooth and nondegenerate objective. Thus, the outlined path-following scheme indeed achieves our goal - it reduces the constrained problem (P) to a series of in fact unconstrained problems (P_t).

We have outlined what are our abilities to solve to a prescribed accuracy every particular problem (P_t) - to this end we can apply to the problem any relaxation iterative routine for smooth unconstrained minimization, starting the routine from an interior point of G. What we need, anyhow, is to solve not a single problem from the family, but a sequence of these problems associated with certain tending to ∞ sequence of values of the penalty parameter. Of course, in principle we could choose an arbitrary sequence $\{t_i\}$ and solve each of the problems (P_{t_i}) independently, but anybody understands that it is senseless. What makes sense is to use the approximate solution x_i to the "previous" problem (P_{t_i}) as the starting point when solving the "new" problem $(P_{t_{i+1}})$. Since $x^*(t)$, as we just have mentioned, is a continuous function of t, a good approximate solution to the previous problem will be a good initial point for solving the new one, provided that $t_{i+1} - t_i$ is not too large; this latter asumption can be ensured by a proper policy of updating the penalty parameter.

To implement the aforementioned scheme, one should specify its main blocks, namely, to choose somehow:

1) the barrier F;

2) the "working horse" - the unconstrained minimization method for solving the problems (P_t), along with the stopping criterion for the method;

3) the policy for updating the penalty parameter.

The traditional recommendations here were rather diffuse. The qualitative theory insisted on at least C^2-smoothness and nondegeneracy of the barrier, and this was basically all; within this class of barriers, there were no clear theoretical priorities. What people were adviced to do, was

for 1): to choose F as certain "preserving smoothness" aggregate of g_i, e.g.,

$$F(x) = \sum_{i=1}^{m} \left(\frac{1}{-g_i(x)} \right)^{\alpha} \tag{1.2}$$

with some $\alpha > 0$, or

$$F(x) = -\sum_{i=1}^{m} \ln(-g_i(x)), \tag{1.3}$$

or something else of this type; the idea was that the local information on this barrier required by the "working horse" should be easily computed via similar information on the constraints g_i;

for 2): to choose as the "working horse" the Newton method; this recommendation came from computational experience and had no serious theoretical justification;

for 3): qualitatively, updating the penalty at a high rate, we reduce the number of auxiliary unconstrained problems at the cost of elaborating each of the problems (since for large $t_{i+1} - t_i$ a good approximation of $x^*(t_i)$ may be a bad starting point for solving the updated problem; a low rate of updating the penalty simplifies the auxiliary problems and increases the number of the problems to be solved before a prescribed value of the penalty (which corresponds to the required accuracy of solving (P)) is achieved. The traitional theory was unable to offer explicit recommendations on the "balanced" rate resulting in the optimal overall effort, and this question normally was solved on the basis of "computational experience".

What was said looks very natural and is known for more than 30 years. Nevertheless, the classical results on the path-following scheme have nothing in common with polynomial complexity bounds, and not only because in 60's nobody bothered about polynomiality: even after you pose this question, the traditional results do not allow to answer this question affirmatively. The reason is as follows: to perform the complexity analysis of the path-following scheme, one needs not only qualitative information like "the Newton method, as applied to a smooth convex function with nondegenerate Hessian, converges quadratically, provided that the starting point is close enough to the minimizer of the objective", but also quantitive information: what is this "close enough". The results of this latter type also existed and everybody in Optimization knew them, but it did not help much. Indeed, the typical quantitive result on the behaviour of the Newton optimization method was as follows:

let ϕ be a C^2*-continuous convex function defined in the Euclidean ball V of radius R centered at x^* and taking minimum at x^* such that*

$\phi''(x^*)$ is nondegenerate with the spectrum from certain segment segment $[L_0, L_1]$, $0 < L_0 < L_1$;

$\phi''(x)$ is Lipschitz continuous at x^* with certain constant L_3:

$$|\phi''(x) - \phi''(x^*)| \leq L_3 |x - x^*|, \; x \in V.$$

Then there exist

$$\rho = \rho(R, L_0, L_1, L_2) > 0, \;\; c = c(R, L_0, L_1, L_2)$$

such that the Newton iterate

$$x^+ = x - [\phi''(x)]^{-1}\phi'(x)$$

of a point x satisfies the relation

$$|x + -x^*| \leq c|x - x^*|^2, \tag{1.4}$$

provided that

$$|x - x^*| \leq \rho.$$

The functions $\rho(\cdot)$ and $c(\cdot)$ can be written down explicitly, the statement itself can be modified and a little bit strengthen, but it does not matter for us: the point is the structure of traditional results on the Newton method, not the results themselves. These results are *local*: the quantitive description of the convergence properties of the method is given in terms of the parameters responsible for smoothness and nondegeneracy of the objective, and the "constant factor" c in the rate-of-convergence expression (1.4), same as the size ρ of the "domain of quadratic convergence" become worse and worse as the aforementioned parameters of smoothness and nondegeneracy of the objective become worse. This is the structure of the traditional rate-of-convergence results for the Newton method; the structure traditional results on any other standard method for smooth unconstrained optimization is completely similar: these results always involve some data-dependent parameters of smoothness and/or nondegeneracy of the objective, and the quantitive description of the rate of convergence always becomes worse and worse as these parameters become worse.

Now it is easy to realize why the traditional rate-of-convergence results for our candidate "working horses" - the Newton method or something else - do not allow to establish polynomiality of the path-following scheme. As the method goes on, the parameters of smoothness and nondegeneracy of our auxiliary objectives F_t inevitably become worse and worse: if the solution to (P) is on the boundary of G, and this is the only case of interest in constrained minimization, the minimizers $x^*(t)$ of F_t approach the boundary of G as t grows, and the behaviour of F_t in a neighbourhood of $x^*(t)$ becomes less and less regular (indeed, for large t the function F_t goes to ∞ very close to $x^*(t)$. Since the parameters of smoothness/nondegeneracy of F_t become worse and worse as t grows, the auxiliary problems, from the traditional viewpoint, become quantitively more and more complicated, and the progress in accuracy (# of new digits of accuracy per unit computational effort) tends to 0 as the method goes on.

The seminal contribution of Renegar and Gonzaga was in demonstration of the fact that *the above scheme applied to a Linear Programming problem*

$$\textit{minimize } f(x) = c^T x \textit{ s.t. } g_j(x) \equiv a_i^T - b_j \le 0,\ j = 1, ..., m,\ x \in \mathbf{R}^n$$

and to the concrete barrier for the feasible domain G of the problem - to the standard logarithmic barrier

$$F(x) = -\sum_{j=1}^{m} \ln(b_j - a_j^T x)$$

for the polytope G - is polynomial.

More specifically, it was proved that the method

$$t_{i+1} = (1 + \frac{0.001}{\sqrt{m}})t_i;\ \ x_{i+1} = x_i - [\nabla_x^2 F_{t_{i+1}}(x_i)]^{-1}\nabla_x F_{t_{i+1}}(x_i) \tag{1.5}$$

(a single Newton step per each step in the penalty parameter) keeps the iterates in the interior of G, maintains the "closeness relation"

$$F_{t_i}(x_i) - \min F_{t_i} \le 0.01$$

(provided that this relation was satisfied by the initial pair (t_0, x_0)) and ensures linear data-independent rate of convergence

$$f(x_i) - f^* \le 2mt_i^{-1} \le 2mt_0^{-1} \exp\{-O(1)im^{-1/2}\}. \tag{1.6}$$

Thus, in spite of the above discussion, it turned out that for the particular barrier in question the path-following scheme is polynomial - the penalty can be increased at a constant rate $(1 + 0.001m^{-1/2})$ depending only on the size of the problem instance, and each step in the penalty should be accompanied by a *single* Newton step in x. According to (1.6), the absolute inaccuracy is inverse proportional to the penalty parameter, so that to add an extra accuracy digit it suffices to increase the parameter by an absolute constant factor, which, in view of the description of the method, takes $O(\sqrt{m})$ steps. Thus, the *Newton complexity* - the # of Newton steps - of finding an ε-solution is

$$\mathcal{N}(\varepsilon, p) = O(\sqrt{m}) \ln\left(\frac{\mathcal{V}(p)}{\varepsilon}\right), \tag{1.7}$$

and since each Newton step costs, as it is easily seen, $O(mn^2)$ operations, the combinatorial complexity of the method turns out to be polynomial, namely,

$$\mathcal{C}(\varepsilon, p) \le O(m^{1.5}n^2) \ln\left(\frac{\mathcal{V}(p)}{\varepsilon}\right).$$

1.4 What is inside: self-concordance

Needless to say that the proofs of the announced results given by Renegar and Gonzaga were completely non-standard and heavily exploited the specific form of the logarithmic barrier for the polytope. The same can be said about subsequent papers devoted to the Linear Programming case. The key to nonlinear extensions found by Yurii Nesterov was in realizing that *among all various properties of the logarithmic barrier for a polytope, in fact only two are responsible for the polynomiality of the path-following methods associated with this polytope. These properties are expressed by the following pair of differential inequalities:*
[self-concordance]:

$$|\frac{d^3}{dt^3}|_{t=0}F(x+th)| \le 2\left(\frac{d^2t}{dt^2}|_{t=0}F(x+th)\right)^{3/2},\ \forall h\ \forall x \in \operatorname{int} G,$$

[finiteness of the barrier parameter]:

$$\exists\vartheta < \infty:\ |\frac{d}{dt}|_{t=0}F(x+th)| \le \vartheta^{1/2}\left(\frac{d^2t}{dt^2}|_{t=0}F(x+th)\right)^{1/2},\ \forall h\ \forall x \in \operatorname{int} G.$$

The inequality in the second relation in fact is satisfied with $\theta = m$.

I am not going to comment these properties now; this is the goal of the forthcoming lectures. What should be said is that these properties do not refer explicitly to the polyhedral structure of G. Given an *arbitrary* solid G, not necessarily polyhedral, one can try to find for this solid a barrier F with the indicated properties. It turns out that such a *self-concordant* barrier always exists; moreover, in many important cases it can be written down in explicit and "computable" form. And the essense of the theory is that

> *given a self-concordant barrier F for a solid G, one can associate with this barrier interior-point methods for minimizing linear objectives over G in completely the same manner as in the case when G is a polytope and F is the standard logarithmic barrier for G. E.g., to get a path-following method, it suffices to replace in the relations (1.5) the standard logarithmic barrier for a polytope with the given self-concordant barrier for the solid G, and the quantity m with the parameter ϑ of the latter barrier, with similar substitution $m \Leftarrow \vartheta$ in the expression for the Newton complexity of the method.*
>
> In particular, if F is "polynomially computable", so that its gradient and Hessian at a given point can be computed at a polynomial arithmetic cost, then the associated with F path-following method turns out to be polynomial.

Note that in the above claim I spoke about minimizing *linear* objectives only. This does not cause any loss of generality, since, given a general convex problem

$$\textit{minimize } f(u) \textit{ s.t. } g_j(u) \leq 0,\ j = 1, ..., m,\ u \in Q \subset \mathbf{R}^k,$$

you always can pass from it to an equivalent problem

$$\textit{minimize } t \textit{ s.t. } x \equiv (t, u) \in G \equiv \{(t, u) \mid f(u) - t \leq 0,\ g_j(u) \leq 0,\ i = 1, ..., m,\ u \in Q\}$$

of minimizing a linear objective over convex set. Thus, the possibilities to solve convex problems by interior point polynomial time methods are restricted only by our abilities to point out "explicit polynomially computable" self-concordant barriers for the corresponding feasible domains, which normally is not so difficult.

1.5 Structure of the course

I hope now you have certain preliminary impression of what we are going to do. More specifically, our plans are as follows.

1) First of all, we should study the basic properties of self-concordant functions and barriers; these properties underly all our future constructions and proofs. This preliminary part of the course is technical; I hope we shall survive the technicalities which, I think, will take two lectures.

2) As an immediate consequence of our technical effort, we shall find ourselves in a fine position to develop and study path-following interior point methods for convex problems, and this will be the first application of our theory.

3) To extend onto the nonlinear case another group of interior point methods known for LP, the potential reduction ones (like the method of Karmarkar), we start with a specific and very interesting in its own right geometry - *conic formulation of a Convex Programming Problem and Conic Duality*. After developing the corresponding geometrical tools, we would be in a position to develop potential reduction methods for general convex problems.

4) The outlined "general" part of the course is, in a sense, conditional: the typical statements here claim that, given a "good" - self-concordant - barrier for the feasible domain of the problem in question, you should act in such and such way and will obtain such and such polynomial efficiency estimate. As far as applications are concerned, these general schemes should, of course, be accompanied by technique for constructing the required "good" barriers. This technique is developed in the second part of the course. Applying this technique and our general schemes, we shall come to concrete "ready-to-use" interior point polynomial time algorithms for a series of important classes of Convex Programming problems, including, besides Linear Programming, Linearly constrained Quadratic Programming, Quadratically constrained Quadratic Programming, Geometrical Programming, Optimization over the cone of positive semidefinite matrices, etc.

Chapter 2

Self-concordant functions

In this lecture I introduce the main concept of the theory in question - the notion of a *self-concordant function*. The goal is to define a family of smooth convex functions convenient for minimization by the Newton method. Recall that a step of the Newton method as applied to the problem of (unconstrained) minimization of a smooth convex function f is based on the following rule:

in order to find the Newton iterate of a point x compute the second-order Taylor expansion of f at x, find the minimizer $\widehat{x}$ of this expansion and perform a step from x along the direction $\widehat{x} - x$.

What the step should be, it depends on the version of the method: in the pure Newton routine the iterate is exactly $\widehat{x}$; it the relaxation version of the method one minimizes f along the ray $[x, \widehat{x})$, etc.

As it was mentioned in the introductory lecture, the traditional results on the Newton method state, under reasonable smoothness and nondegeneracy assumptions, its local quadratic convergence. These results, as it became clear recently, possess a generic conceptual drawback: the quantitive description of the region of quadratic convergence, same as the convergence itself, is given in terms of the condition number of the Hessian of f at the minimizer and the Lipschitz constant of this Hessian. These quantities, anyhow, are "frame-dependent": they are defined not by f itself, but also by the Euclidean structure in the space of variables. Indeed, we need this structure simply to define the Hessian matrix of f, same, by the way, as to define the gradient of f. When we change the Euclidean structure, the gradient and the Hessian are subject to certain transformation which *does not* remain invariant the quantities like the condition number of the Hessian or its Lipschitz constant. As a result, the traditional description of the behaviour of the method depends not only on the objective itself, but also on an arbitrary choice of the Euclidean structure used in the description, which contradicts the affine-invariant nature of the method (note that no "metric notions" are involved into the formulation of the method). To overcome this drawback, note that the objective itself at any point x induces certain Euclidean structure $\mathcal{E}_x$; to define this structure, let us regard the second order differential

$$D^2 f(x)[h, g] = \frac{\partial^2}{\partial t \partial s} \mid_{t=s=0} f(x + th + sg)$$

of f taken at x along the pair of directions h and g as the inner product of the vectors h and g. Since f is convex, this inner product possesses all required properties (except, possibly, the nondegeneracy requirement "the square of a nonzero vector is strictly positive"; as we shall see, this is a minor difficulty). Of course, this Euclidean structure is *local* - it depends on x. Note that the Hessian of f, taken at x with respect to the Euclidean structure $\mathcal{E}_x$, is fine - this is simply the unit matrix, the matrix with the smallest possible condition number, namely, 1. The traditional results on the Newton method say that what is important for besides this condition number is the Lipschitz constant of the Hessian, or, which is basically the same, the magnitude of the third order derivatives of f. What happens if we relate these latter quantities to the local Euclidean structure defined by f? This is the key to the notion of self-concordance. And the definition is as follows:

Definition 2.0.1 *Let Q be a nonempty open convex set in $\mathbf{R}^n$ and F be a C^3 smooth convex function defined on Q. F is called self-concordant on Q, if it possesses the following two properties:*

[Barrier property] *$F(x_i) \to \infty$ along every sequence $\{x_i \in Q\}$ converging, as $i \to \infty$, to a boundary point of Q;*

[Differential inequality of self-concordance] *F satisfies the differential inequality*

$$|D^3F(x)[h,h,h]| \leq 2\left(D^2F(x)[h,h]\right)^{3/2} \tag{2.1}$$

for all $x \in Q$ and all $h \in \mathbf{R}^n$.

From now on

$$D^kF(x)[h_1,...,h_k] \equiv \frac{\partial^k}{\partial t_1...\partial t_k}|_{t_1=...=t_k=0}F(x+t_1h_1+...+t_kh_k)$$

denotes kth differential of F taken at x along the directions $h_1,...,h_k$.

(2.1) says exactly that if a vector h is of local Euclidean length 1, then the third order derivative of F in the direction h is, in absolute value, at most 2; this is nothing but the aforementioned "Lipschitz continuity", with certain once for ever fixed constant, namely, 2, of the second-order derivative of F with respect to the local Euclidean metric defined by this derivative itself.

You can ask what is so magic in the constant 2. The answer is as follows: both sides of (2.1) should be nad actually are of the same homogeneity degree with respect to h (this is the origin of the exponentual 3/2 in the right hand side). As a consequence, they are of different homogeneity degrees with respect to F. Therefore, given a function F satisfying the inequality

$$|D^3F(x)[h,h,h]| \leq 2\alpha\left(D^2F(x)[h,h]\right)^{3/2},$$

with certain positive α, you always may scale F, namely, multiply it by $\sqrt{\alpha}$, and come to a function satisfying (2.1). We see that the choice of the constant factor in (2.1) is of no actual importance and is nothing but a normalization condition. The indicated choice of this factor is motivated by the desire to make the function $-\ln t$, which plays important role in what follows, to satisfy (2.1) "as it is", without any scaling.

2.1 Examples and elementary combination rules

We start with a pair of examples of self-concordant functions.

Example 2.1.1 *A convex quadratic form*

$$f(x) = x^TAx - 2b^Tx + c$$

on $\mathbf{R}^n$ (and, in particular, a linear form on $\mathbf{R}^n$) is self-concordant on $\mathbf{R}^n$.

This is immediate: the left hand side of (2.1) is identically zero. An single-line verification of the definition justifies also the following example:

Example 2.1.2 *The function $-\ln t$ is self-concordant on the positive ray $\{t \in \mathbf{R} \mid t > 0\}$.*

The number of examples can be easily increased, due to the following extremely simple (and very useful) combination rules:

Proposition 2.1.1 (i) [stability with respect to affine substitutions of argument] *Let F be self-concordant on $Q \subset \mathbf{R}^n$ and $x = Ay + b$ be affine mapping from $\mathbf{R}^k$ to $\mathbf{R}^n$ with the image intersecting Q. Then the inverse image of Q under the mapping, i.e., the set*

$$Q^+ = \{y \in \mathbf{R}^k \mid Ay + b \in Q\}$$

is an open convex subset of $\mathbf{R}^k$, and the composite function

$$F^+(y) = F(Ay + b) : Q^+ \to \mathbf{R}$$

is self-concordant on Q^+.

(ii) [stability with respect to summation and multiplication by reals ≥ 1] *Let F_i be self-concordant functions on the open convex domains $Q_i \subset \mathbf{R}^n$ and $\alpha_i \geq 1$ be reals, $i = 1, ..., m$. Assume that the set $Q = \cap_{i=1}^m Q_i$ is nonempty. Then the function*

$$F(x) = \alpha_1 F_1(x) + ... + \alpha_m F_m(x) : Q \to \mathbf{R}$$

is self-concordant on Q.

(iii) [stability with respect to direct summation] *Let F_i be self-concordant on open convex domains $Q_i \subset \mathbf{R}^{n_i}$, $i = 1, ..., m$. Then the function*

$$F(x_1, ..., x_m) = F_1(x_1) + ... + F_m(x_m) : Q \equiv Q_1 \times ... \times Q_m \to \mathbf{R}$$

is self-concordant on Q.

Proof is given by immediate and absolutely trivial verification of the definition. E.g., let us prove (ii). Since O_i are open convex domains with nonempty intersection Q, Q is an open convex domain, as it should be. Further, F, is, of course, C^3 smooth and convex on Q. To prove the barrier property, note that since F_i are convex, they are below bounded on any bounded subset of Q. It follows that if $\{x_j \in Q\}$ is a sequence converging to a boundary point x of Q, then all the sequences $\{\alpha_i F_i(x_j)\}$, $i = 1, ..., m$, are below bounded, and at least one of them diverges to ∞ (since x belongs to the boundary of at least one of the sets Q_i); consequently, $F(x_j) \to \infty$, as required.

To verify (2.1), add the inequalities

$$\alpha_i |D^3 F_i(x)[h, h, h]| \leq 2\alpha_i \left(D^2 F_i(x)[h, h]\right)^{3/2}$$

($x \in Q$, $h \in \mathbf{R}^n$). The left hand side of the resulting inequality clearly will be $\geq |D^3 F(x)[h, h, h]|$, while the right hand side will be $\leq 2\left(D^2 F(x)[h, h]\right)^{3/2}$, since for nonnegative b_i and $\alpha_i \geq 1$ one has

$$\sum_i \alpha_i b_i^{3/2} \leq (\sum_i \alpha_i b_i)^{3/2}.$$

Thus, F satisfies (2.1). ■

An immediate consequence of our combination rules is the following

Corollary 2.1.1 *Let*

$$G = \{x \in \mathbf{R}^n \mid a_i^T x - b_i \leq 0, \ i = 1, ..., m\}$$

be a convex polyhedron defined by a set of linear inequalities satisfying the Slater condition:

$$\exists x \in \mathbf{R}^n : \ a_i^T x - b_i < 0, \ i = 1, ..., m.$$

Then the standard logarithmic barrier for G given by

$$F(x) = -\sum_{i=1}^m \ln(b_i - a_i^T x)$$

is self-concordant on the interior of G.

Proof. From the Slater condition it follows that

$$\mathrm{int}\, G = \{x \in \mathbf{R}^n \mid a_i^T x - b_i < 0, \ i = 1, ..., m\} = \cap_{i=1}^m G_i, \ G_i = \{x \in \mathbf{R}^n \mid a_i^T x - b_i < 0\}.$$

Since the function $-\ln t$ is self-concordant on the positive half-axis, every of the functions $F_i(x) = -\ln(b_i - a_i^T x)$ is self-concordant on G_i (item (i) of Proposition; note that G_i is the inverse image of the positive half-axis under the affine mapping $x \mapsto b_i - a_i^T x$), whence $F(x) = \sum_i F_i(x)$ is self-concordant on $G = \cap_i G_i$ (item (ii) of Proposition). ■

In spite of its extreme simplicity, the fact stated in Corollary, as we shall see in the mean time, is responsible for 50% of all polynomial time results in Linear Programming.

Now let us come to systematic investigation of properties of self-concordant functions, with the final goal to analyze the behaviour of the Newton method as applied to a function of this type.

2.2 Properties of self-concordant functions

Let Q be an open convex domain in $E = \mathbf{R}^n$ and F be self-concordant on Q. For $x \in Q$ and $h, g \in E$ let us define

$$\langle g, h\rangle_x = D^2F(x)[g,h],\ |h|_x = \langle h,h\rangle_x^{1/2}$$

so that $|\cdot|_x$ is a Euclidean seminorm on E; it is a norm if and only if $D^2F(x)$ is nondegenerate.

Let us establish the basic properties of F.

0. Basic inequality. *For any $x \in Q$ and any triple $h_i \in E$, $i = 1, 2, 3$, one has*

$$|D^3F(x)[h_1,h_2,h_3]| \le 2\prod_{i=1}^{3}|h_i|_x.$$

Comment. This is the result of applying to the symmetric 3-linear form $D^3F(x)[h_1,h_2,h_3]$ and 2-linear positive semidefinite form $D^2F(x)[h_1,h_2]$ the following general fact:

let $A[h_1,...,h_k]$ be a symmetric k-linear form on $\mathbf{R}^n$ and $B[h_1,h_2]$ be a symmetrice positive semidefinite bilinear form such that

$$|A[h,h,...,h]| \le \alpha B^{k/2}[h,h]$$

for certain α and all h. Then

$$|A[h_1,...,h_k]| \le \alpha B^{1/2}[h_1,h_1]B^{1/2}[h_2,h_2]...B^{1/2}[h_k,h_k]$$

for all $h_1,...,h_k$.

The proof of this statement is among the exercises to the lecture.

I. Behaviour in the Dikin ellipsoid *For $x \in Q$ let us define the centered at x open Dikin ellipsoid of radius r as the set*

$$W_r(x) = \{y \in E \mid |y-x|_x < r\},$$

and the closed Dikin ellipsoid as the set

$$\widehat{W}_r(x) = \mathrm{cl}\,W_r(x) = \{y \in E \mid |y-x|_x \le r\}.$$

The open unit Dikin ellipsoid $W_1(x)$ is contained in Q. Within this ellipsoid the Hessians of F are "almost proportional" to $F''(x)$,

$$(1-|h|_x)^2F''(x) \le F''(x+h) \le (1-|h|_x)^{-2}F''(x) \text{ whenever } |h|_x < 1, \tag{2.2}$$

the gradients of F satisfy the following Lipschitz-type condition:

$$|z^T(F'(x+h)-F'(x))| \le \frac{|h|_x}{1-|h|_x}|z|_x \quad \forall z \text{ whenever } |h|_x < 1, \tag{2.3}$$

and we have the following lower and upper bounds on F:

$$F(x)+DF(x)[h]+\rho(-|h|_x) \le F(x+h) \le F(x)+DF(x)[h]+\rho(|h|_x),\ |h|_x < 1. \tag{2.4}$$

where

$$\rho(s) = -\ln(1-s) - s = \frac{s^2}{2}+\frac{s^3}{3}+\frac{s^4}{4}+... \tag{2.5}$$

Lower bound in (2.4) is valid for all h such that $x+h \in Q$, not only for those h with $|h|_x < 1$.

Proof. Let h be such that

$$r \equiv |h|_x < 1 \text{ and } x+h \in Q.$$

Let us prove that relations (2.2), (2.3) and (2.4) are satisfied at this particular h.

1^0. Let us set

$$\phi(t) = D^2F(x+th)[h,h],$$

so that ϕ is continuously differentiable on $[0,1]$. We have

$$0 \leq \phi(t),\ r^2 = \phi(0) < 1,\ |\phi'(t)| = |D^3F(x+th)[h,h,h]| \leq 2\phi^{3/2}(t),$$

whence, for all small enough positive ϵ,

$$0 < \phi_\epsilon(t) \equiv \epsilon + \phi(t),\ \phi_\epsilon(0) < 1,\ |\phi'_\epsilon(t)| \leq 2\phi_\epsilon^{3/2}(t),$$

so that

$$|\frac{d}{dt}\phi_\epsilon^{-1/2}(t)| \leq 1.$$

It follows that

$$\phi_\epsilon^{-1/2}(0) - t \leq \phi_\epsilon^{-1/2}(t) \leq \phi_\epsilon^{-1/2}(0) + t,\ 0 \leq t \leq 1,$$

whence

$$\frac{\phi_\epsilon(0)}{(1+t\phi_\epsilon^{1/2}(0))^2} \leq \phi_\epsilon(t) \leq \frac{\phi_\epsilon(0)}{(1-t\phi_\epsilon^{1/2}(0))^2}.$$

The resulting inequalities hold true for all $t \in [0,1]$ and all $\epsilon > 0$; passing to limit as $\epsilon \to +0$, we come to

$$\frac{r^2}{(1+rt)^2} \leq \phi(t) \equiv D^2F(x+th)[h,h] \leq \frac{r^2}{(1-rt)^2},\ 0 \leq t \leq 1. \tag{2.6}$$

2^0. Two sequential integrations of (2.6) result in

$$F(x) + DF(x)[h] + \int_0^1 \{\int_0^\tau \frac{r^2}{(1+rt)^2}dt\}d\tau \leq F(x+h) \leq$$

$$\leq F(x) + DF(x)[h] + \int_0^1 \{\int_0^\tau \frac{r^2}{(1-rt)^2}dt\}d\tau,$$

which after straightforward computation leads to (2.4) (recall that $r = |h|_x$).

Looking at the presented reasoning, one can immediately see that the restriction $r < 1$ was used only in the derivation of the upper, not the *lower* bound in (2.4); therefore this lower bound is valid for all h such that $x + h \in Q$, as claimed.

3^0. Now let us fix $g \in E$ and set

$$\psi(t) = D^2F(x+th)[g,g],$$

so that ψ a continuously differentiable nonnegative function on $[0,1]$. We have

$$|\psi'(t)| = |D^3F(x+th)[g,g,h]| \leq 2D^2F(x+th)[g,g]\left[D^2F(x+th)[h,h]\right]^{1/2} \tag{2.7}$$

(we have used **0.**). Relation (2.7) means that ψ satisfies the linear differential inequality

$$|\psi'(t)| \leq 2\psi(t)\phi^{1/2}(t) \leq 2\psi(t)\frac{r}{1-rt},\ 0 \leq t \leq 1$$

(the second inequality follows from (2.6) combined with $\psi \geq 0$). It follows that

$$\frac{d}{dt}[(1-rt)^2\psi(t)] \equiv (1-rt)^2[\psi'(t) - 2r(1-rt)^{-1}\psi(t)] \leq 0,\ 0 \leq t \leq 1,$$

and

$$\frac{d}{dt}[(1-rt)^{-2}\psi(t)] \equiv (1-rt)^{-2}[\psi'(t) + 2r(1-rt)^{-1}\psi(t)] \geq 0,\ 0 \leq t \leq 1,$$

whence, respectively,

$$(1-rt)^2\psi(t) \leq \psi(0),\ (1-rt)^{-2}\psi(t) \geq \psi(0),$$

or, recalling what ψ and r are,

$$(1-|h|_xt)^{-2}D^2F(x+th)[g,g] \geq D^2F(x)[g,g] \geq (1-|h|_xt)^2D^2F(x+th)[g,g];$$

since g is arbitrary, we come to (2.2).

4^0. We have proved that (2.2) and (2.4) hold true for any h such that $x+h$ is in the open unit Dikin ellipsoid $W_1(x)$ and $x+h \in Q$. To complete the proof, it remains to demonstrate that the latter "and" is redundant: $x+h \in Q$ whenever $x+h$ belongs to the open unit Dikin ellipsoid $W_1(x)$. To prove the latter statement, assume, on contrary, that $W_1(x)$ is not contained in Q. Then there is a point y in $W_1(x)$ such that the half-segment $[x, y)$ belongs to Q and y itself does not belong to Q. The function F is well-defined on this half-segment; moreover, as we already have seen, at any point $x+h$ of this half-segment (2.4) holds. When $x+h$ runs over the half-segment, the quantities $|h|_x$ are bounded from above by $|y-x|_x$ and are therefore less than 1 and bounded away from 1. It follows from (2.4) that F is bounded on the half-segment, which is the desired contradiction: since y is a boundary point of Q, F should tend to ∞ as a point from $[x, y)$ approaches to y.

5^0. It remains to prove (2.3). To this end let us fix an arbitrary vector z and let us set

$$g(t) = z^T(F'(x+th) - F'(x)).$$

Since the open unit Dikin ellipsoid $W_1(x)$ is contained in Q, the function g is well-defined on the segment $[0, 1]$. We have

$$\begin{array}{rcl} g(0) & = & 0; \\ |g'(t)| & = & |z^T F''(x+th)h| \\ & \leq & \sqrt{z^T F''(x+th)z}\sqrt{h^T F''(x+th)h} \\ & & \text{[we have used Cauchy's inequality]} \\ & \leq & (1-t|h|_x)^{-2}\sqrt{z^T F''(x)z}\sqrt{h^T F''(x)h} \\ & & \text{[we have used (2.2)]} \\ & = & |h|_x(1-t|h|_x)^{-2}\sqrt{z^T F''(x)z}, \end{array}$$

whence

$$|g(1)| \leq \int_0^1 \frac{|h|_x}{(1-t|h|_x)^2} dt \sqrt{z^T F''(x)z} = \frac{|h|_x}{1-|h|_x}\sqrt{z^T F''(x)z},$$

as claimed in (2.3). ■

II. Recessive subspace of a self-concordant function. *For $x \in Q$ consider the subspace $\{h \in E \mid D^2F(x)[h,h] = 0\}$ - the kernel of the Hessian of F at x. This recessive subspace E_F of F is independent of the choice of x and is such that*

$$Q = Q + E_F.$$

In particular, the Hessian of F is nonsingular everywhere if and only if there exists a point where the Hessian of F is nonsingular; this is for sure the case if Q is bounded.

Terminology: we call F *nondegenerate*, if $E_F = \{0\}$, or, which is the same, if the Hessian of F is nonsingular somewhere (and then everywhere) on Q.

Proof of II. To prove that the kernel of the Hessian of F is independent of the point where the Hessian is taken is the same as to prove that if $D^2F(x_0)[h,h] = 0$, then $D^2F(y)[h,h] \equiv 0$ identically in $y \in Q$. To demonstrate this, let us fix $y \in Q$ and consider the function

$$\psi(t) = D^2F(x_0 + t(y-x))[h,h],$$

which is consinuously differentiable on the segment $[0, 1]$. Same as in the item 3^0 of the previous proof, we have

$$|\psi'(t)| = |D^3F(x_0 + t(y-x))[h,h,y-x]| \leq$$

$$\leq 2D^2F(x_0 + t(y-x))[h,h]\left[D^2F(x_0 + t(y-x))[y-x,y-x]\right]^{1/2} \equiv \psi(t)\xi(t)$$

with certain continuous on $[0, 1]$ function ξ. It follows that

$$|\psi'(t)| \leq M\psi(t)$$

with certain constant M, whence $0 \leq \psi(t) \leq \psi(0)\exp\{Mt\}, 0 \leq t \leq 1$ (look at the derivative of the function $\psi(t)\exp\{-Mt\}$). Since $\psi(0) = 0$, we come to $\psi(1) = 0$, i.e., $D^2F(y)[h,h] = 0$, as claimed.

Thus, the kernel of the Hessian of F is independent of the point where the Hessian is taken. If $h \in E_F$ and $x \in Q$, then, of course, $|h|_x = 0$, so that $x + h \in W_1(x)$; from **I.** we know that $W_1(x)$ belongs to Q, so that $x + h \in Q$; thus, $x + E_F \subset Q$ whenever $x \in Q$, as required. ■

Now it is time to introduce a very important concept of *Newton decrement* of a self-concordant function at a point. Let $x \in Q$. The Newton decrement of F at x is defined as

$$\lambda(F, x) = \max\{DF(x)[h] \mid h \in E, |h|_x \leq 1\}.$$

In other words, the Newton decrement is nothing but the conjugate to $|\cdot|_x$ norm of the first-order derivative of F at x. To be more exact, we should note that $|\cdot|_x$ is not necessary a norm: it may be a *semi*norm, i.e., may be zero at certain nonzero vectors; this happens if and only if the recessive subspace E_F of F is nontrivial, or, which is the same, if the Dikin ellipsoid of F is not an actual ellipsoid, but an unbounded set - elliptic cylinder. In this latter case the maximum in the definition of the Newton decrement may (not necessarily should) be $+\infty$. We can immediately realize when this is the case.

III. Continuity of the Newton decrement. *The Newton decrement of F at $x \in Q$ is finite if and only if $DF(x)[h] = 0$ for all $h \in E_F$. If it is the case for certain $x = x_0 \in Q$, then it is also the case for all $x \in Q$, and in this case the Newton decrement is continuous in $x \in Q$ and F is constant along its recessive subspace:*

$$F(x + h) = F(x) \ \ \forall x \in Q \ \ \forall h \in E_F; \tag{2.8}$$

otherwise the Newton decrement is identically $+\infty$.

Proof. It is clear that if there is $h \in E_F$ such that $DF(x)[h] \neq 0$, then $\lambda(F, x) = \infty$, since $|th|_x = 0$ for all real t and, consequently, $DF(x)[u]$ is above unbounded on the set $\{|u|_x \leq 1\}$. Vice versa, assume that $DF(x)[h] = 0$ for all $h \in E_F$, and let us prove that then $\lambda(F, x) < \infty$. There is nothing to prove if $E_F = E$, so that let us assume that $E_F \neq E$. Let $E_F^{\perp}$ be certain subspace of E complementary to E_F: $E_F \cap E_F^{\perp} = \{0\}$, $E_F + E_F^{\perp} = E$, and let π be the projector of E onto $E_F^{\perp}$ parallel to E_F, i.e., if

$$h = h_F + h_F^{\perp}$$

is the (unique) representation of $h \in E$ as the sum of vectors from E_F and $E_F^{\perp}$, then

$$\pi h = h_F^{\perp}.$$

It is clear that

$$|\pi h|_x \equiv |h|_x$$

(since the difference $h - \pi h$ belongs to E_F and therefore is of zero $|\cdot|_x$-seminorm), and since we have assumed that $DF(x)[u]$ is zero for $u \in E_F$, we also have

$$DF(x)[h] = DF(x)[\pi h].$$

Combining these observations, we see that it is possible to replace E in the definition of the Newton decrement by $E_F^{\perp}$:

$$\lambda(F, x) = \max\{DF(x)[h] \mid h \in E_F^{\perp}, |h|_x \leq 1\}. \tag{2.9}$$

Since $|\cdot|_x$ restricted onto $E_F^{\perp}$ is a norm rather than a seminorm, the right hand side of the latter relation is finite, as claimed.

Now let us demonstrate that if $\lambda(F, x)$ is finite at certain point $x_0 \in Q$, then it is also finite at any other point x of Q and is continuous in x. To prove finiteness, as we just have seen, it suffices to demonstrate that $DF(x)[h] = 0$ for any x and any $h \in E_F$. To this end let us fix $x \in Q$ and $h \in E_F$ and consider the function

$$\psi(t) = DF(x_0 + t(x - x_0))[h].$$

This function is continuously differentiable on $[0, 1]$ and is zero at the point $t = 0$ (since $\lambda(F, x_0)$ is assumed finite); besides this,

$$\psi'(t) = D^2F(x_0 + t(x - x_0))[h, x - x_0] = 0$$

(since h belongs to the null space of the positive semidefinite symmetric bilinear form $D^2F(x_0+t(x-x_0))[h_1,h_2]$), so that ψ is constant, namely, 0, and $\psi(1)=0$, as required. As a byproduct of our reasonong, we see that if $\lambda(F,\cdot)$ is finite, then

$$F(x+h)=F(x),\ x\in Q,\ h\in E_F,$$

since the derivative of F at any point from Q in any direction from E_F is zero.

It remains to prove that if $\lambda(F,x)$ is finite at certain (and then, as we just have proved, at any) point, then this is a continuous function of x. This is immediate: we already know that if $\lambda(F,x)$ is finite, it can be defined by relation (2.9), and this relation, by the standard reasons, defines a continuous function of x (since $|\cdot|_x$ restricted onto $E_F^{\perp}$ is a continuously depending on x norm, not a seminorm). ■

The following simple observation clarifies the origin of the Newton decrement and its relation to the Newton method.

IV. Newton Decrement and Newton Iterate. *Given $x\in Q$, consider the second-order Newton expansion of F at x, i.e., the convex quadratic form*

$$N_{F,x}(h)=F(x)+DF(x)[h]+\frac{1}{2}D^2F(x)[h,h]\equiv F(x)+DF(x)[h]+\frac{1}{2}|h|_x^2.$$

This form is below bounded if and only if it attains its minimum on E and if and only if $\lambda(F,x)<\infty$; if it is the case, then for (any) Newton direction e of F at x, i.e., any minimizer of this form, one has

$$D^2F(x)[e,h]\equiv -DF(x)[h],\ h\in E, \tag{2.10}$$

$$|e|_x=\lambda(F,x) \tag{2.11}$$

and

$$N_{F,x}(0)-N_{F,x}(e)=\frac{1}{2}\lambda^2(F,x). \tag{2.12}$$

Thus, the Newton decrement is closely related to the amount by which the Newton iteration

$$x\mapsto x+e$$

decreases F in its second-order expansion.

Proof. This is an immediate consequence of the standard fact of Linear Algebra: a convex quadratic form

$$f_{A,b}(h)=\frac{1}{2}h^TAh+b^Th+c$$

is below bounded if and only if it attains its minimum and if and only if the quantity

$$\lambda=\max\{b^Th\mid h^TAh\le 1\}$$

is finite; if it is the case, then the minimizers y of the form are exactly the vectors such that

$$y^TAh=-b^Th,\ h\in E,$$

for every minimizer y one has

$$y^TAy=\lambda^2$$

and

$$f_{A,b}(0)-\min f_{A,b}=\frac{1}{2}\lambda^2.$$

■

The observation given by **IV.** allows to compute the Newton decrement in the nondegenerate case $E_F=\{0\}$.

IVa. Expressions for the Newton direction and the Newton decrement. *If F is nondegenerate and $x\in Q$, then the Newton direction of F at x is unique and is nothing but*

$$e(F,x)=-[F''(x)]^{-1}F'(x),$$

F' and F'' being the gradient and the Hessian of F with respect to certain Euclidean structure on E, and the Newton decrement is given by

$$\lambda(F,x) = \sqrt{(F'(x))^T[F''(x)]^{-1}F'(x)} = \sqrt{e^T(F,x)F''(x)e(F,x)} = \sqrt{-e^T(F,x)F'(x)}.$$

Proof. This is an immediate consequence of **IV.** (pass from the "coordinateless" differentials to "coordinate" representation in terms of the gradient and the Hessian).

Now comes the main statement about the behaviour of the Newton method as applied to a self-concordant function.

V. Damped Newton Method: relaxation property. *Let $\lambda(F,\cdot)$ be finite on Q. Given $x \in Q$, consider the damped Newton iterate of x*

$$x^+ \equiv x^+(F,x) = x + \frac{1}{1+\lambda(F,x)}e,$$

e being (any) Newton direction of F at x. Then

$$x^+ \in Q$$

and

$$F(x) - F(x^+) \geq \lambda(F,x) - \ln(1+\lambda(F,x)). \tag{2.13}$$

Proof. As we know from **IV.**, $|e|_x = \lambda \equiv \lambda(F,x)$, and therefore $|x^+ - x|_x = \lambda/(1+\lambda) < 1$. Thus, x^+ belongs to the open unit Dikin ellipsoid of F centered at x, and, consequently, to Q (see **I.**). In view of (2.4) we have

$$F(x^+) \leq F(x) + \frac{1}{1+\lambda}DF(x)[e] + \rho((1+\lambda)^{-1}|e|_x) =$$

[see (2.10) - (2.12)]

$$= F(x) - \frac{1}{1+\lambda}D^2F(x)[e,e] + \rho\left(\frac{\lambda}{1+\lambda}\right) = F(x) - \frac{\lambda^2}{1+\lambda} + \rho\left(\frac{\lambda}{1+\lambda}\right) =$$

[see the definition of ρ in (2.4)]

$$= F(x) - \frac{\lambda^2}{1+\lambda} - \ln\left(1 - \frac{\lambda}{1+\lambda}\right) - \frac{\lambda}{1+\lambda} =$$

$$= F(x) - \lambda + \ln(1+\lambda),$$

so that

$$F(x) - F(x^+) \geq \lambda - \ln(1+\lambda),$$

as claimed. ∎

VI. Existence of minimizer, A. *F attains its minimum on Q if and only if it is below bounded on Q; if it is the case, then $\lambda(F,\cdot)$ is finite and, moreover, $\min_{x\in Q}\lambda(F,x) = 0$.*

Proof. Of course, if F attains its minimum on Q, it is below bounded on this set. To prove the inverse statement, assume that F is below bounded on Q, and let us prove that it attains its minimum on Q. First of all, $\lambda(F,\cdot)$ is finite. Indeed, if there would be $x \in Q$ with infinite $\lambda(F,x)$, it would mean that the derivative of F taken at x in certain direction $h \in E_F$ is nonzero. As we know from **II.**, the affine plane $x + E_F$ is contained in Q, and the second order derivative of the restriction of F onto this plane is identically zero, so that the restriction is linear (and nonconstant, since the first order derivative of F at x in certain direction from E_F is nonzero). And a nonconstant linear function $F\mid_{x+E_F}$ is, of course, below unbounded. Now let $Q^\perp$ be the cross-section of Q by the plane $x + E_F^\perp$, where $x \in Q$ is certain fixed point and $E_F^\perp$ is a subspace complementary to E_F. Then $Q^\perp$ is an open convex set in certain $\mathbf{R}^k$ and, in view of **II.**, $Q = Q^\perp + E_F$; in view of **III.** F is constant along any translation of E_F, and we see that it is the same to prove that F attains its minimum on Q and to prove that the restriction of F onto $Q^{\perp}$ attains its minimum on $Q^{\perp}$. This restriction is a self-concordant function on $Q^{\perp}$ (Proposition 2.1.1); of course, it is below bounded on $Q^{\perp}$, and its recessive subspace is trivial. Passing from (Q, F) to $(Q^{\perp}, F|_{Q^{\perp}})$, we see that the statement in question can be reduced to a similar statement for a *nondegenerate* self-concordant below bounded function; to avoid complicated notation, let us assume that F itself is nondegenerate.

Since F is below bounded, the quantity $\inf_{x\in Q}\lambda(F,x)$ is 0; indeed, if it were positive:

$$\lambda(F,x) > \lambda > 0 \ \forall x \in Q,$$

then, according to **V.**, we would have a possibility to pass from any point $x \in Q$ to another point x^{+} with at least by the constant $\lambda - \ln(1+\lambda)$ less value of F, which, of course, is impossible, since F is assumed below bounded. Since $\inf_{x\in Q}\lambda(F,x) = 0$, there exists a point x with $\lambda \equiv \lambda(F,x) \le 1/6$. From (2.4) it follows that

$$F(x+h) \ge F(x) + DF(x)[h] + |h|_x - \ln(1+|h|_x), \ \ |h|_x < 1.$$

Further, in view of (2.10),

$$DF(x)[h] = -D^2F(x)[e,h] \ge -|e|_x|h|_x$$

(we have used the Cauchy inequality), which combined with (2.11) results in

$$DF(x)[h] \ge -\lambda |h|_x,$$

and we come to

$$F(x+h) \ge F(x) - \lambda|h|_x + |h|_x - \ln(1+|h|_x). \tag{2.14}$$

When $0 \le t < 1$, we have

$$f(t) \equiv -\lambda t + t - \ln(1+t) \ge -\lambda t + t - t + \frac{1}{2}t^2 - \frac{1}{3}t^3 + \frac{1}{4}t^4 - ... \ge$$

$$\ge -\lambda t + \frac{1}{2}t^2 - \frac{1}{3}t^3 = t\left[\frac{1}{2}t - \frac{1}{3}t^2 - \lambda\right],$$

and we see that if

$$t(\lambda) = 2(1+3\lambda)\lambda,$$

then $f(t(\lambda)) > 0$ and $t(\lambda) < 1$. From (2.14) we conclude that $F(x+h) > F(x)$ whenever $x+h$ belongs to the boundary of the closed Dikin ellipsoid $\widehat{W}_{t(\lambda)}(x)$ which in the case in question is a compact subset of Q (recall that F is assumed to be nondegenerate). It follows that the minimizer of F over the ellipsoid (which for sure exists) is an interior point of the ellipsoid and therefore (due to convexity of F) is a minimizer of F over Q, so that F attains its minimum over Q. ■

To proceed, let me recall to you the concept of the Legendre transformation. Given a convex function f defined on a convex subset $\mathrm{Dom}\, f$ of $\mathbf{R}^n$, one can define the *Legendre transformation* f^* of f as

$$f^*(y) = \sup_{x\in \mathrm{Dom}\, f}[y^T x - f(x)];$$

the domain of f^* is, by definition, comprised of those y for which the right hand side is finite. It is immediately seen that $\mathrm{Dom}\, f^*$ is convex and f^* is convex on its domain.

Let $\mathrm{Dom}\, f$ be open and f be $k \ge 2$ times continuously differentiable on its domain, the Hessian of f being nondegenerate. It is celarly seen that

(L.1) *if $x \in \mathrm{Dom}\, f$, then $y = f'(x) \in \mathrm{Dom}\, f^*$, and*

$$f^*(f'(x)) = (f'(x))^T x - f(x); \ x \in \partial f^*(f'(x)).$$

Since f'' is nondegenerate, by the Implicit Function Theorem the set $\mathrm{Dom}^*\, f^*$ of values of f' is open; since, in addition, f is convex, the mapping

$$x \mapsto f'(x)$$

is $(k-1)$ times continuously differentiable one-to-one mapping from $\mathrm{Dom}\, f$ onto $\mathrm{Dom}^*\, f^*$ with $(k-1)$ times continuously differentiable inverse. From (L.1) it follows that this inverse mapping also is given by gradient of some function, namely, f^*. Thus,

(L.2) *The mapping $x \mapsto f'(x)$ is a one-to-one mapping of* $\mathrm{Dom}\, f$ *onto an open set* $\mathrm{Dom}^*\, f^* \subset \mathrm{Dom}\, f^*$, *and the inverse mapping is given by $y \mapsto (f^*)'(y)$.*

As an immediate consequence of (L.2), we come to the following statement

(L.3) f^* is k times continuously differentiable on $\mathrm{Dom}^*\, f^*$, and

$$(f^*)''(f'(x)) = [f''(x)]^{-1},\ x \in \mathrm{Dom}\, f. \tag{2.15}$$

VII. Self-concordance of the Legendre transformation. *Let the Hessian of the self-concordant function F be nondegenerate at some (and then, as we know from* **II.**, *at any) point. Then* $\mathrm{Dom}\, F^* = \mathrm{Dom}^*\, F^*$ *is an open convex set, and the function F^* is self-concordant on* $\mathrm{Dom}\, F^*$.

Proof. 1^0. Let us prove first that $\mathrm{Dom}\, F^* = \mathrm{Dom}^*\, F^*$. If $y \in \mathrm{Dom}\, F^*$, then, by definition, the function $y^T x - F(x)$ is bounded from above on Q, or, which is the same, the function $F(x) - y^T x$ is below bounded on Q. This function is self-concordant (Proposition 2.1.1.(ii) and Example 2.1.1), and since it is below bounded, it attains its minimum on Q (**VI.**). At the minimizer x^* of the function we have $F'(x^*) = y$, and we see that $y \in \mathrm{Dom}^*\, F^*$. Thus, $\mathrm{Dom}\, F = \mathrm{Dom}^*\, F^*$.

2^0. The set $\mathrm{Dom}\, F^*$ is convex, and the set $\mathrm{Dom}^*\, F^*$ is open ((L.2)); from 1^0 it follows therefore that F^* is a convex function with a convex open domain $\mathrm{Dom}\, F^*$. The function is 3 times continuously differentiable on $\mathrm{Dom}\, F^* = \mathrm{Dom}^*\, F^*$ in view of (L.3). To prove self-concordance of F^*, it suffices to verify the barrier property and the differential inequality (2.1).

3^0. The barrier property is immediate: if a sequence $y_i \in \mathrm{Dom}\, F^*$ converges to a point y and the sequence $\{F^*(y_i)\}$ is bounded from above, then the functions $y_i^T x - F(x)$ are uniformly bounded from above on Q and therefore their pointwise limit $y^T x - F(x)$ also is bounded from above on Q; by definition of $\mathrm{Dom}\, F^*$ it means that $y \in \mathrm{Dom}\, F^*$, and since we already know that $\mathrm{Dom}\, F^*$ is open, we conclude that any convergent sequence of points from $\mathrm{Dom}\, F^*$ along which F^* is bounded from above converges to an interior point of $\mathrm{Dom}\, F^*$; this, of course, is an equivalent reformulation of the barrier property.

4^0. It remains to verify (2.1). From (L.3) for any fixed h we have

$$h^T (F^*)''(F'(x))h = h^T [F''(x)]^{-1} h,\ x \in Q.$$

Differentiating this identity in x in a direction g, we come to[1)]

$$D^3 F^*(F'(x))[h, h, F''(x)g] = -D^3 F(x)[[F''(x)]^{-1}h, [F''(x)]^{-1}h, g];$$

substituting $g = [F''(x)]^{-1}h$, we come to

$$|D^3 F^*(F'(x))[h,h,h]| = |D^3 F(x)[g,g,g]| \le 2\left(D^2 F(x)[g,g]\right)^{3/2} \equiv 2\left(g^T F''(x) g\right)^{3/2} =$$

[since $g = [F''(x)]^{-1}h$]

$$= 2\left(h^T [F''(x)]^{-1} h\right)^{3/2}.$$

The latter quantity, due to (L.3), is exactly $2\left(h^T (F^*)''(F'(x))h\right)^{3/2}$, and we come to

$$|D^3 F^*(y)[h,h,h]| \le 2\left(D^2 F^*(y)[h,h]\right)^{3/2}$$

for all h and all $y = F'(x)$ with $x \in Q$. When x runs over Q, y, as we alerady know, runs through the whole $\mathrm{Dom}\, F^*$, and we see that (2.1) indeed holds true. ■

[1] we use the following rule for differentiating the mapping $x \mapsto B(x) \equiv A^{-1}(x)$, $A(x)$ being a square nonsingular matrix smoothly depending on x:

$$DB(x)[g] = -B(x) DA(x)[g] B(x)$$

(to get it, differentiate the identity $B(x)A(x) \equiv I$).

VIII. Existence of minimizer, B. *F attains its minimum on Q if and only if there exists $x \in Q$ with $\lambda(F,x) < 1$, and for every x with the latter property one has*

$$F(x) - \min_Q F \leq \rho(\lambda(F,x)); \tag{2.16}$$

moreover, for an arbitrary minimizer x^* of F on Q and the above x one has

$$D^2F(x)[x^* - x, x^* - x] \leq \left(\frac{\lambda(F,x)}{1-\lambda(F,x)}\right)^2. \tag{2.17}$$

Proof. The "only if" part is evident: $\lambda(F,x) = 0$ at any minimizer x of F. To prove the "if" part, we, same as in the proof of **VI.**, can reduce the situation to the case when F is nondegenerate. Let x be such that $\lambda \equiv \lambda(F,x) < 1$, and let $y = F'(x)$. In view of (L.3) we have

$$y^T(F^*)''(y)y = (F'(x))^T[F''(x)]^{-1}F'(x) = \lambda^2 \tag{2.18}$$

(the latter relation follows from **VIa.**). Since $\lambda < 1$, we see that 0 belongs to the centered at y open Dikin ellipsoid of the self-concordant (as we know from **VII.**) function F^* and therefore (**I.**) to the domain of this function. From **VII.** we know that this domain is comprised of values of the gradient of F at the points of Q; thus, there exists $x^* \in Q$ such that $F'(x^*) = 0$, and F attains its minimum on Q. Furthermore, from (2.4) as applied to F^* and from (2.18) we have

$$F^*(0) \leq F^*(y) - y^T(F^*)'(y) + \rho(\lambda);$$

since $y = F'(x)$ and $0 = F'(x^*)$, we have (see (L.1))

$$F^*(y) = y^Tx - F(x),\ (F^*)'(y) = x,\ F^*(0) = -F^*(x^*),$$

and we come to

$$-F(x^*) \leq y^Tx - F(x) - y^Tx + \rho(\lambda),$$

which is nothing but (2.16).

Finally, setting

$$|h|_y = \sqrt{h^T(F^*)''(y)h}$$

and noticing that, by (2.18), $|y|_y = \lambda < 1$, we get for an arbitrary vector z

$$\begin{array}{rcl} |z^T(x^* - x)| & = & |z^T[(F^*)'(0) - (F^*)'(y)]| \\ & \leq & \frac{\lambda}{1-\lambda}\sqrt{z^T(F^*)''(y)z} \\ & & \text{[we have applied (2.3) to } F^* \text{ at the point } y \text{ with } h = -y] \\ & = & \frac{\lambda}{1-\lambda}\sqrt{z^T[F''(x)]^{-1}z}; \end{array}$$

substituting $z = F''(x)(x^* - x)$, we get

$$\sqrt{(x^* - x)F''(x)(x^* - x)} \leq \frac{\lambda}{1-\lambda},$$

as required in (2.17). ■

Remark 2.2.1 Note how sharp is the condition of existence of minimizer given by **VII.**: for the self-concordant on the positive ray and below unbounded function $F(x) = -\ln x$ one has $\lambda(F,x) \equiv 1$!

IX. Damped Newton method: local quadratic convergence. *Let $\lambda(F,\cdot)$ be finite, let $x \in Q$, and let x^+ be the damped Newton iterate of x (see **V.**). Then*

$$\lambda(F,x^+) \leq 2\lambda^2(F,x). \tag{2.19}$$

Besides this, if $\lambda(F,x) < 1$, then F attains its minimum on Q, and for any minimizer x^* of F one has

$$|x - x^*|_{x^*} \le \frac{\lambda(F,x)}{1-\lambda(F,x)}; \qquad (2.20)$$

$$|x - x^*|_{x} \le \frac{\lambda(F,x)}{1-\lambda(F,x)}. \qquad (2.21)$$

Proof. 1^0. To prove (2.19), denote by e the Newton direction of F at x, set

$$\lambda = \lambda(F,x), \;\; r = \frac{1}{1+\lambda},$$

and let $h \in E$. The function

$$\psi(t) = DF(x+te)[h]$$

is twice continuously differentiable on $[0,r]$; we have

$$\psi'(t) = D^2F(x+te)[h,e], \;\; \psi''(t) = D^3F(x+te)[h,e,e],$$

whence, in view of **O.**,

$$|\psi''(t)| \le 2|h|_{x+te}|e|^2_{x+te} \le$$

[in view of (2.2) and since $|e|_x = \lambda$, see (2.11)]

$$\le 2(1-t\lambda)^{-3}|h|_x|e|_x^2 = 2(1-t\lambda)^{-3}\lambda^2|h|_x.$$

It follows that

$$DF(x^+)[h] \equiv \psi(r) \le \psi(0) + r\psi'(0) + |h|_x \int_0^r \{\int_0^t 2(1-\tau\lambda)^{-3}\lambda^2 d\tau\}dt =$$

$$= \psi(0) + r\psi'(0) + \frac{\lambda^2 r^2}{1-\lambda r}|h|_x =$$

[the definition of ψ]

$$= DF(x)[h] + rD^2F(x)[h,e] + \frac{\lambda^2 r^2}{1-\lambda r}|h|_x =$$

[see (2.10)]

$$= (1-r)DF(x)[h] + \frac{\lambda^2 r^2}{1-\lambda r}|h|_x =$$

[the definition of r]

$$\frac{\lambda}{1+\lambda}DF(x)[h] + \frac{\lambda^2}{1+\lambda}|h|_x \le$$

[since $DF(x)[h] \le \lambda |h|_x$ by definition of $\lambda = \lambda(F,x)$]

$$\le 2\frac{\lambda^2}{1+\lambda}|h|_x \le$$

[see (2.2) and take into account that $|x^+ - x|_x = r|e|_x = r\lambda$]

$$\le 2\frac{\lambda^2}{1+\lambda}\frac{1}{1-r\lambda}|h|_{x^+} = 2\lambda^2|h|_{x^+}.$$

Thus, for any $h \in E$ we have $DF(x^+)[h] \le 2\lambda^2|h|_{x^+}$, as claimed in (2.19).

2^0. Let $x \in Q$ be such that $\lambda \equiv \lambda(F,x) < 1$. We already know from **VIII.** that in this case F attains its minimum on Q, and that

$$F(x) - \min_Q F \le \rho(\lambda) \equiv -\ln(1-\lambda) - \lambda. \qquad (2.22)$$

Let x^* be a minimizer of F on Q and let $r = |x - x^*|_{x^*}$. From (2.4) applied to $x = x^*$, $h = x - x^*$ it follows that

$$F(x) \geq F(x^*) + \rho(-r) \equiv F(x^*) + r - \ln(1 + r).$$

Combining this observation with (2.22), we come to

$$r - \ln(1 + r) \leq -\lambda - \ln(1 - \lambda),$$

and it immediately follows that $r \leq \frac{\lambda}{1-\lambda}$, as required in (2.20). (2.21) is identical to (2.17). ■

The main consequence of the indicated properties of self-concordant functions is the following description of the behaviour of the Damped Newton method (for the sake of simplicity, we restrict ourselves with the case of nondegenerate F):

X. Summary on the Damped Newton method. *Let F be self-concordant nondegenerate function of Q. Then*

A. [existence of minimizer] *F attains its minimum on Q if and only if it is below bounded on Q; this is for sure the case if*

$$\lambda(F, x) \equiv \sqrt{(F'(x))^T [F''(x)]^{-1} F'(x)} < 1$$

for some x.

B. *Given $x_1 \in Q$, consider the Damped Newton minimization process given by the reccurence*

$$x_{i+1} = x_i - \frac{1}{1 + \lambda(F, x_i)} [F''(x_i)]^{-1} F'(x_i). \tag{2.23}$$

The recurrency keeps the iterates in Q and possesses the following properties

B.1 [relaxation property]

$$F(x_{i+1}) \leq F(x_i) - [\lambda(F, x_i) - \ln(1 + \lambda(F, x_i))]; \tag{2.24}$$

in particular, if $\lambda(F, x_i)$ is greater than an absolute constant, then the progress in the value of F at the step i is at least another absolute constant; e.g., if $\lambda(F, x_i) \geq 1/4$, then $F(x_i) - F(x_{i+1}) \geq \frac{1}{4} - \ln \frac{5}{4} = 0.026856...$

B.2 [local quadratic convergence] *If at certain step i we have $\lambda(F, x_i) \leq \frac{1}{4}$, then we are in the region of quadratic convergence of the method, namely, for every $j \geq i$ we have*

$$\lambda(F, x_{j+1}) \leq 2\lambda^2(F, x_j) \qquad [\leq \frac{1}{2}\lambda(F, x_j)], \tag{2.25}$$

$$F(x_j) - \min_Q F \leq \rho(\lambda(F, x_j)) \qquad [\leq \frac{\lambda^2(F, x_j)}{2(1 - \lambda(F, x_j))}], \tag{2.26}$$

and for the (unique) minimizer x^* of F we have

$$|x_j - x^*|_{x^*} \leq \frac{\lambda(F, x_j)}{1 - \lambda(F, x_j)}. \tag{2.27}$$

If, in addition, $\lambda(F, x) < 1/2$, then also

$$|x_j - x^*|_{x_j} \leq \frac{\lambda(F, x_j)}{1 - 2\lambda(F, x_j)}. \tag{2.28}$$

C. *If F is below bounded, then the Newton complexity (i.e., # of steps (2.23)) of finding a point $x \in Q$ with $\lambda(F, x) \leq \kappa \leq 0.1$) does not exceed the quantity*

$$O(1) \left([F(x_1) - \min_Q F] + \ln \ln \frac{1}{\kappa} \right) \tag{2.29}$$

with an absolute constant $O(1)$.

The statements collected in **X.** in fact are already proved: **A** is given by **VIII.**; **B.1** is **V.**; **B.2** is **IX.**; **C** is an immediate consequence of **B.1** and **B.2**.

Note that the description of the convergence properties of the Newton method as applied to a self-concordant function is completely objective-independent; it does not involve any specific numeric characteristics of F.

2.3 Exercises: Around Symmetric Forms

The goal of the below exercises is to establish the statement underlying **0.**:

(P): *let $A[h_1, ..., h_k]$ be a k-linear symmetric form on $\mathbf{R}^n$ and $B[h_1, h_2]$ be a symmetric positive semidefinite 2-linear form on $\mathbf{R}^n$. Assume that for some α one has*

$$|A[h, ..., h]| \leq \alpha B^{k/2}[h, h], \;\; h \in \mathbf{R}^n. \tag{2.30}$$

Then

$$|A[h_1, ..., h_k]| \leq \alpha \prod_{i=1}^{k} B^{1/2}[h_i, h_i] \tag{2.31}$$

for all $h_1, ..., h_k$.

Let me start with recalling the terminology. A *k-linear form* $A[h_1, ..., h_k]$ on $E = \mathbf{R}^n$ is a real-valued function of k arguments $h_1, ..., h_k$, each of them varying over E, which is linear and homogeneous function with respect to every argument, the remaining arguments being set to arbitrary (fixed) values. The examples are:

- a linear form $A[h] = a^T h$ ($k = 1$);
- a bilinear form $A[h_1, h_2] = h_1^T a h_2$, a being $n \times n$ matrix ($k = 2$);
- 3-linear form of the type $A[h_1, h_2, h_3] = (a^T h_1)(h_2^T h_3)$;
- the n-linear form $A[h_1, ..., h_n] = \mathrm{Det}\,(h_1; ...; h_n)$.

A k-linear form is called *symmetric*, if it remains unchanged under every permutation of the collection of arguments.

Exercise 2.3.1 *Prove that any 2-linear form on $\mathbf{R}^n$ can be reprresented as $A[h_1, h_2] = h_1^T a h_2$ via certain $n \times n$ martix a. When the form is symmetric? Which of the forms in the above examples are symmetric?*

The restriction of a *symmetric* k-linear form $A[h_1, ..., h_k]$ onto the "diagonal" $h_1 = h_2 = ... = h_k = h$, which is a function of $h \in \mathbf{R}^n$, is called *homogeneous polynomial of full degree k* on $\mathbf{R}^n$; the definition coincides with the usual Calculus definition: "a polynomial of n variables is a finite sum of monomials, every monomial being constant times product of nonnegative integer powers of the variables. A polynomial is called homogeneous of full degree k if the sum of the powers in every monomial is equal to k".

Exercise 2.3.2 *Prove the equivalence of the aforementioned two definitions of a homogeneous polynomial. What is the 3-linear form on $\mathbf{R}^2$ which produces the polynomial xy^2 ((x, y) are coordinates on $\mathbf{R}^2$)?*

Of course, you can restrict onto diagonal an arbitrary k-linear form, not necessarily symmetric, and get certain function on E. You, anyhow, will not get something new: for any k-linear form $A[h_1, ..., h_k]$ there exists a *symmetric* k-linear form $A_S[h_1, ..., h_k]$ with the same restriction on the diagonal:

$$A[h, ..., h] \equiv A_S[h, ..., h], \;\; h \in E;$$

to get A_S, it suffices to take average, over all permutations σ of the k-element index set, of the forms $A_\sigma[h_1, ..., h_k] = A[h_{\sigma(1)}, ..., h_{\sigma(k)}]$.

From polylinearity of a k-linear form $A[h_1, ..., h_k]$ it follows that the value of the form at the collection of linear combinations

$$h_i = \sum_{j \in J} a_{i,j} u_{i,j}, \;\; i = 1, ..., k,$$

J being a finite index set, can be expressed as

$$\sum_{j_1, ..., j_k \in J} \left(\prod_{i=1}^{k} a_{i,j} \right) A[u_{1,j_1}, u_{2,j_2}, ..., u_{k,j_k}];$$

this is nothing but the usual rule for "opening the parentheses". In particular, $A[\cdot]$ is uniquely defined by its values on the collections comprised of basis vectors $e_1, ..., e_n$:

$$A[h_1, ..., h_k] = \sum_{1\le j_1,...,j_k\le n} h_{1,j_1} h_{2,j_2} ... h_{k,j_k} A[e_{j_1}, e_{j_2}, ..., e_{j_k}],$$

$h_{i,j}$ being j-th coordinate of the vector h_i with respect to the basis. It follows that a polylinear form is continuous (even C^∞) function of its arguments.

A symmetric bilinear form $A[h_1, h_2]$ is called *positive semidefinite*, if the corresponding homogeneous polynomial is nonnegative, i.e., if $A[h, h] \ge 0$ for all h. A symmetric positive semidefinite bilinear form sastisfies all requirements imposed on an inner product, except, possibly, the nondegeneracy requirements "square of nonzero vector is nonzero". If this requirement also is satisfied, i.e., if $A[h, h] > 0$ whenever $h \ne 0$, then $A[h_1, h_2]$ defines an Euclidean structure on E. As we know from Exercise 2.3.1, a bilinear form on $\mathbf{R}^n$ always can be represented by a $n \times n$ matrix a as $h_1^T a h_2$; the form is symmetric if and only if $a = a^T$, and is symmetric positive (semi)definite if and only if a is symmetric positive (semi)definite matrix.

A *symmetric* k-linear form produces, as we know, a uniquely defined homogeneous polynomial of degree k. It turns out that the polynomial "remembers everything" about the related k-linear form:

Exercise 2.3.3 $^{\#+}$ *Prove that for every k there exist:*

- *integer m,*
- *real "scale factors" $r_{1,l}, r_{2,l}, ..., r_{l,l}$, $l = 1, ..., m$,*
- *real weights w_l, $l = 1, ..., m$,*

with the following property: for any n and any k-linear symmetric form $A[h_1, ..., h_k]$ on $\mathbf{R}^n$ identically in $h_1, ..., h_k$ one has

$$A[h_1, ..., h_k] = \sum_{l=1}^{m} w_l A\left[\sum_{i=1}^{k} r_{i,l} h_i, \sum_{i=1}^{k} r_{i,l} h_i, ..., \sum_{i=1}^{k} r_{i,l} h_i\right].$$

In other words, A can be restored, in a linear fashion, via its restriction on the diagonal.

Find a set of scale factors and weights for $k = 2$ and $k = 3$.

Now let us come to the proof of (P). Of course, it suffices to consider the case when B is positive definite rather than semidefinite (replace $B[h_1, h_2]$ with $B_\epsilon[h_1, h_2] = B[h_1, h_2] + \epsilon h_1^T h_2$, $\epsilon > 0$, thus making B positive definite and preserving the assumption (2.30); given that (P) is valid for positive definite B, we would know that (2.31) is valid for B replaced with B_ϵ and would be able to pass to limit as $\epsilon \to 0$). Thus, from now on we assume that B is symmetric positive definite. In this case $B[h_1, h_2]$ can be taken as an inner product on $\mathbf{R}^n$, and in the associated "metric" terms (P) reads as follows:

(P'): *let $|\cdot|$ be a Euclidean norm on $\mathbf{R}^n$, $A[h_1, ..., h_k]$ be a k-linear symmetric form on $\mathbf{R}^n$ such that*

$$|A[h, ..., h]| \le \alpha |h|^k, \;\; h \in \mathbf{R}^n.$$

Then

$$|A[h_1, ..., h_k]| \le \alpha |h_1| ... |h_k|, \;\; h_1, ..., h_k \in \mathbf{R}^n.$$

Now, due to homogeneity of A with respect to every h_i, to prove the conclusion in (P') is the same as to prove that $|A[h_1, ..., h_k]| \le \alpha$ whenever $|h_i| \le 1$, $i = 1, ..., k$. Thus, we come to the following equivalent reformulation of (P'):

prove that for a k-linear symmetric form $A[h_1, ..., h_k]$ one has

$$\max_{|h|=1} |A[h, ..., h]| = \max_{|h_i|\le 1} |A[h_1, ..., h_k]|. \tag{2.32}$$

Note that from Exercise 2.3.3 it immediately follows that the right hand side of (2.32) is majorated by a *constant times* the left hand side, with the constant depending on k only. For this latter statement it is completely unimportant whether the norm $|\cdot|$ in question is or is not Euclidean. The point, anyhow, is that *in the case of Euclidean norm the aforementioned constant factor can be set to 1.* This is something which should be a "common knowledge"; surprisingly, I was unable to find somewhere even the statement, not speaking of the proof. I do not think that the proof presented in the remaining exercises is the simplest one, and you are welcome to find something better. We shall prove (2.32) by induction on k.

Exercise 2.3.4 *Prove the base, i.e., that (2.32) holds true for $k = 2$.*

Now assume that (2.32) is valid for $k = l - 1$ and any k-linear symmetric form A, and let is prove that it is valid also for $k = l$.

Let us fix a symmetric l-linear form A, and let us call a collection $\mathcal{T} = \{T_1, ..., T_l\}$ of one-dimensional subspaces of $\mathbf{R}^n$ *an extremal*, if for some (and then - for each) choice of *unit* vectors $e_i \in T_i$ one has

$$|A[e_1, ..., e_l]| = \omega \equiv \max_{|h_1|=...=|h_l|=1} |A[h_1, ..., h_l]|.$$

Clearly, extremals exist (we have seen that $A[\cdot]$ is continuous). Let T be the set of all extremals. To prove (2.32) is the same as to prove that T contains an extremal of the type $\{T, ..., T\}$.

Exercise 2.3.5 $^{\#+}$ *Let $\{T_1, ..., T_l\} \in \mathsf{T}$ and $T_1 \neq T_2$. Let $e_i \in T_i$ be unit vectors, $h = e_1 + e_2$, $q = e_1 - e_2$. Prove that then both $\{\mathbf{R}h, \mathbf{R}h, T_3, ..., T_l\}$ and $\{\mathbf{R}q, \mathbf{R}q, T_3, ..., T_l\}$ are extremals.*

Let T^* be the subset of T formed by the extremals of the type $\{\overbrace{T, ..., T}^{t \text{ times}}, \overbrace{S, ..., S}^{s \text{ times}}\}$ for some t and s (depending on the extremal). By virtue of the inductive assumption, T^* is nonempty (in fact, T^* contains an extremal of the type $\{T, ..., T, S\}$). For $\mathcal{T} = \{\overbrace{T, ..., T}^{t \text{ times}}, \overbrace{S, ..., S}^{s \text{ times}}\} \in \mathsf{T}^*$ let $\alpha(\mathcal{T})$ denote the angle (from $[0, \frac{\pi}{2}]$) between T and S.

Exercise 2.3.6 $^{\#+}$ *Prove that if $\mathcal{T} = \{T, ..., T, S, ..., S\}$ is an extremal of the aforementioned "2-line" type, then there exists an extremal $\mathcal{T}'$ of the same type with $\phi(\mathcal{T}') \leq \frac{1}{2}\phi(\mathcal{T})$. Derive from this observation that there exists a 2-line extremal with $\phi(\mathcal{T}) = 0$, i.e., of the type $\{T, ..., T\}$, and thus complete the inductive step.*

Exercise 2.3.7 * *Let $A[h_1, ..., h_k]$, $h_1, ..., h_k \in \mathbf{R}^n$ be a linear with respect to every argument and invariant with respect to premutations of argumens mapping taking values in certain $\mathbf{R}^l$, and let $B[h_1, h_2]$ be a symmetric positive semidefinite bilinear scalar form on $\mathbf{R}^n$ such that*

$$\| A[h, ..., h] \| \leq \alpha B^{k/2}[h, h], \quad h \in \mathbf{R}^n,$$

$\| \cdot \|$ being certain norm on $\mathbf{R}^k$. Prove that then

$$\| A[h_1, ..., h_k] \| \leq \alpha \prod_{i=1}^{k} B^{1/2}[h_i, h_i], \quad h_1, ..., h_k \in \mathbf{R}^n.$$

Chapter 3

Self-concordant barriers

We have introduced and studied the notion of a self-concordant function for an open convex domain. To complete developing of technical tools, we should investigate a specific subfamily of this family - *self-concordant barriers.*

3.1 Definition, examples and combination rules

Definition 3.1.1 *Let G be a closed convex domain in $\mathbf{R}^n$ ("domain" means "a set with a nonempty interior"), and let $\vartheta \geq 0$. A function $F : \mathrm{int}\, G \to \mathbf{R}$ is called self-concordant barrier for G with the parameter value ϑ (in short, ϑ-self-concordant barrier for G), if*

a) F is self-concordant on int G*;*

b) one has

$$|DF(x)[h]| \leq \vartheta^{1/2} \left[D^2F(x)[h,h]\right]^{1/2} \tag{3.1}$$

for all $x \in \mathrm{int}\, G$ and all $h \in \mathbf{R}^n$.

Recall that self-concordance is, basically, Lipschitz continuity of the Hessian of F with respect to the local Euclidean metric defined by the Hessian itself. Similarly, (3.1) says that F should be Lipschitz continuous, with constant $\vartheta^{1/2}$, with respect to the same local metric.

Recall also that the quantity

$$\lambda(F,x) = \max\{DF(x)[h] \mid D^2F(x)[h,h] \leq 1\}$$

was called the Newton decrement of F at x; this quantity played crucial role in our investigation of self-concordant functions. Relation (3.1) means exactly that the Newton decrement of F should be bounded from above, independently of x, by certain constant, and the square of this constant is called the parameter of the barrier.

Let us point out preliminary examples of self-concordant barriers. To this end let us look at the basic examples of self-concordant functions given in the previous lecture.

Example 3.1.1 *A constant is self-concordant barrier for $\mathbf{R}^n$ with the parameter 0.*

It can be proved that a constant is the *only* self-concordant barrier for the whole space, and the *only* self-concordant barrier with the value of the parameter less than 1. In what follows we never deal with the trivial - constant - barrier, so that you should remember that the parameters of barriers in question will always be ≥ 1.

In connection with the above trivial example, note that the known to us self-concordant on the whole space functions - linear and convex quadratic ones - are *not* self-concordant barriers, provided that they are nonconstant. This claim follows from the aforementioned general fact that the only self-concordant barrier for the whole space is a constant and also can be easily verified directly.

Another basic example of a self-concordant function known to us is more productive:

Example 3.1.2 *The function $F(x) = -\ln x$ is a self-concordant barrier with parameter 1 for the non-negative ray.*

This is seen from an immediate computation.

The number of examples can be immediately increased, due to the following simple combination rules (completely similar to those for self-concordant functions):

Proposition 3.1.1 (i) [stability with respect to affine substitutions of argument] *Let F be a ϑ-self-concordant barrier for $G \subset \mathbf{R}^n$ and let $x = Ay + b$ be affine mapping from $\mathbf{R}^k$ to $\mathbf{R}^n$ with the image intersecting* int G. *Then the inverse image of G under the mapping, i.e., the set*

$$G^+ = \{y \in \mathbf{R}^k \mid Ay + b \in G\}$$

is a closed convex domain in $\mathbf{R}^k$, and the composite function

$$F^+(y) = F(Ay + b) : \mathrm{int}\ G^+ \to \mathbf{R}$$

is a ϑ-self-concordant barrier for G^+.

(ii) [stability with respect to summation and multiplication by reals ≥ 1] *Let F_i be ϑ_i-self-concordant barriers for the closed convex domains $G_i \subset \mathbf{R}^n$ and $\alpha_i \geq 1$ be reals, $i = 1, ..., m$. Assume that the set $G = \cap_{i=1}^m G_i$ has a nonempty interior. Then the function*

$$F(x) = \alpha_1 F_1(x) + ... + \alpha_m F_m(x) : \mathrm{int}\ G \to \mathbf{R}$$

is $(\sum_i \alpha_i \vartheta_i)$-self-concordant barrier for G.

(iii) [stability with respect to direct summation] *Let F_i be ϑ_i-self-concordant barriers for closed convex domains $G_i \subset \mathbf{R}^{n_i}$, $i = 1, ..., m$. Then the function*

$$F(x_1, ..., x_m) = F_1(x_1) + ... + F_m(x_m) : \mathrm{int}\ G \to \mathbf{R},\ G \equiv G_1 \times ... \times G_m,$$

is $(\sum_i \vartheta_i)$-self-concordant barrier for G.

Proof is given by immediate and absolutely trivial verification of the definition. E.g., let us prove (ii). From Proposition 2.1.1.(ii) we know that F is self-concordant on int $G \equiv \cap_{i=1}^m$ int G_i. The verification of (3.1) is as follows:

$$|DF(x)[h]| = |\sum_{i=1}^m \alpha_i DF_i(x)[h]| \leq \sum_{i=1}^m \alpha_i |DF_i(x)[h]| \leq$$

[since F_i are ϑ_i-self-concordant barriers]

$$\leq \sum_{i=1}^m \alpha_i \vartheta_i^{1/2} \left[D^2 F_i(x)[h,h]\right]^{1/2} = \sum_{i=1}^m [\alpha_i \vartheta_i]^{1/2} \left[\alpha_i D^2 F_i(x)[h,h]\right]^{1/2} \leq$$

[Cauchy's inequality]

$$\leq \left[\sum_{i=1}^m \alpha_i \vartheta_i\right]^{1/2} \left[\sum_{i=1}^m \alpha_i D^2 F_i(x)[h,h]\right]^{1/2} = \left[\sum_{i=1}^m \alpha_i \vartheta_i\right]^{1/2} \left[D^2 F(x)[h,h]\right]^{1/2},$$

as required. ■

An immediate consequence of our combination rules is as follows (cf. Corollary 2.1.1):

Corollary 3.1.1 *Let*

$$G = \{x \in \mathbf{R}^n \mid a_i^T x - b_i \leq 0,\ i = 1, ..., m\}$$

be a convex polyhedron defined by a set of linear inequalities satisfying the Slater condition:

$$\exists x \in \mathbf{R}^n :\ a_i^T x - b_i < 0,\ i = 1, ..., m.$$

Then the standard logarithmic barrier for G given by

$$F(x) = -\sum_{i=1}^m \ln(b_i - a_i^T x)$$

is m-self-concordant barrier for G.

Proof. The function $-\ln t$ is 1-self-concordant barrier for the positive half-axis (Example 3.1.2); therefore every of the functions $F_i(x) = -\ln(b_i - a_i^T x)$ is 1-self-concordant barrier for the closed half-space $\{x \in \mathbf{R}^n \mid b_i - a_i^T x \geq 0\}$ (item (i) of Proposition; note that G_i is the inverse image of the nonnegative half-axis under the affine mapping $x \mapsto b_i - a_i^T x$), whence $F(x) = \sum_i F_i(x)$ is m-self-concordant barrier for the intersection G of these half-spaces (item (ii) of Proposition). ■

The fact stated in Corollary is responsible for 100% of polynomial time results in Linear Programming.

Now let us come to systematic investigation of properties of self-concordant barriers. Please do not be surprised by the forthcoming miscellania; everything will be heavily exploited in the mean time.

3.2 Properties of self-concordant barriers

Let G be a closed convex domain in $E = \mathbf{R}^n$, and let F be ϑ-self-concordant barrier for G.

Preliminaries: the Minkowsky function of a convex domain. Recall that, given an interior point x of G, one can define the *Minkowsky function of G with the pole at x* as

$$\pi_x(y) = \inf\{t > 0 \mid x + t^{-1}(y - x) \in G\}.$$

In other words, to find $\pi_x(y)$, consider the ray $[x, y)$ and look where this ray intersects the boundary of G. If the intersection point y' exists, then $\pi_x(y)$ is the length of the segment $[x, y']$ divided by the length of the segment $[x, y]$; if the ray $[x, y)$ is contained in G, then $\pi_x(y) = 0$. Note that the Minkowsky function is convex, continuous and positive homogeneous:

$$\pi_x(\lambda y) = \lambda \pi_x(y), \ \lambda \geq 0;$$

besides this, it is zero at x and is ≤ 1 in G, 1 on the boundary of G and > 1 outside G. Note that this function is in fact defined in purely affine terms (the lengths of segments are, of course, metric notions, but the ratio of lengths of parallel segments is metric-independent).

Now let us switch to properties of self-concordant barriers.

0. Explosure property: *Let $x \in \operatorname{int} G$ and let y be such that $DF(x)[y - x] > 0$. Then*

$$\pi_x(y) \geq \gamma \equiv \frac{DF(x)[y - x]}{\vartheta}, \tag{3.2}$$

so that the point $x + \gamma^{-1}(y - x)$ is not an interior point of G.

Proof. Let

$$\phi(t) = F(x + t(y - x)) : \Delta \to \mathbf{R},$$

where $\Delta = [0, T)$ is the largest half-interval of the ray $t \geq 0$ such that $x + t(y - x) \in \operatorname{int} G$ whenever $t \in \Delta$. Note that the function ϕ is three times continuously differentiable on Δ and that

$$T = \pi_x^{-1}(y) \tag{3.3}$$

(the definition of the Minkowsky function; here $0^{-1} = +\infty$).

From the fact that F is ϑ-self-concordant barrier for G it immediately follows (see Proposition 3.1.1.(i)) that

$$|\phi'(t)| \leq \vartheta^{1/2}\sqrt{\phi''(t)},$$

or, which is the same,

$$\vartheta \psi'(t) \geq \psi^2(t), t \in \Delta, \tag{3.4}$$

where $\psi(t) = \phi'(t)$. Note that $\psi(0) = DF(x)[y - x]$ is positive by assumption and ψ is nondecreasing (as the derivative of a convex function), so that ψ is positive on Δ. From (3.4) and the relation $\psi(0) > 0$ it follows that $\vartheta > 0$. In view of the latter relation and since $\psi(\cdot) > 0$, we can rewrite (3.4) as

$$(-\psi^{-1}(t))' \equiv \psi'(t)\psi^{-2}(t) \geq \vartheta^{-1},$$

whence

$$\psi(t) \geq \frac{\vartheta \psi(0)}{\vartheta - t\psi(0)}, \ t \in \Delta. \tag{3.5}$$

The left hand side of the latter relation is bounded on any segment $[0, T']$, $0 < T' < T$, and we conclude that

$$T \leq \frac{\vartheta}{\psi(0)}.$$

Recalling that $T = \pi_x^{-1}(y)$ and that $\psi(0) = DF(x)[y - x]$, we come to (3.2). ■

I. Semiboundedness. *For any $x \in \operatorname{int} G$ and $y \in G$ one has*

$$DF(x)[y - x] \leq \vartheta. \tag{3.6}$$

Proof. The relation is evident in the case of $DF(x)[y-x] \leq 0$; for the case $DF(x)[y-x] > 0$ the relation is an immediate consequence of (3.2), since $\pi_x(y) \leq 1$ whenever $y \in G$. ■

II. Upper bound. *Let $x, y \in \operatorname{int} G$. Then*

$$F(y) \leq F(x) + \vartheta \ln \frac{1}{1 - \pi_x(y)}. \tag{3.7}$$

Proof. For $0 \leq t \leq 1$ we clearly have

$$\pi_{x+t(y-x)}(y) = \frac{(1-t)\pi_x(y)}{1 - t\pi_x(y)};$$

from (3.6) applied to the pair $(x + t(y - x); y)$ it follows that

$$DF(x + t(y - x))[y - [x + t(y - x)]] \leq \vartheta \pi_{x+t(y-x)}(y),$$

whence

$$(1 - t)DF(x + t(y - x))[y - x] \leq \vartheta \frac{(1-t)\pi_x(y)}{1 - t\pi_x(y)},$$

or

$$DF(x + t(y - x))[y - x] \leq \vartheta \frac{\pi_x(y)}{1 - t\pi_x(y)}.$$

Integrating over $t \in [0, 1]$, we come to

$$F(y) - F(x) \leq \vartheta \ln \frac{1}{1 - \pi_x(y)},$$

as required. ■

III. Lower bound. *Let $x, y \in \operatorname{int} G$. Then*

$$F(y) \geq F(x) + DF(x)[y - x] + \ln \frac{1}{1 - \pi_x(y)} - \pi_x(y). \tag{3.8}$$

Proof. Let $\phi(t) = F(x + t(y - x))$, $-T_- < t < T \equiv \pi_x^{-1}(t)$, where T_- is the largest t such that $x - t(y - x) \in G$. By Proposition 3.1.1.(i) ϕ is a self-concordant barrier for $\Delta = [-T_-, T]$, and therefore this function is self-concordant on Δ; the closed unit Dikin ellipsoid of ϕ centered at $t \in \operatorname{int} \Delta$ should therefore belong to the closure of Δ (Lecture 2, **I.**), which means that

$$t + [\phi''(t)]^{-1/2} \leq T, \ 0 \leq t < T$$

(here $0^{-1/2} = +\infty$). We come to the inequality

$$\phi''(t) \geq (T - t)^{-2}, \ 0 \leq t < T.$$

Two sequential integrations of this inequality result in

$$F(y) - F(x) \equiv \phi(1) - \phi(0) = \int_0^1 \{ \int_0^t (T-\tau)^{-2} d\tau \} dt = \ln \frac{T}{T-1} - T^{-1};$$

substituting $T = \pi_x^{-1}(y)$, we come to (3.8). ■

IV. Upper bound on local norm of the first derivative. *Let $x, y \in \mathrm{int}\, G$. Then for any $h \in E$ one has*

$$|DF(y)[h]| \leq \frac{\vartheta}{1-\pi_x(y)} |h|_x \equiv \frac{\vartheta}{1-\pi_x(y)} \left[D^2 F(x)[h,h] \right]^{1/2}. \tag{3.9}$$

Comment: By definition, the first-order derivative of the ϑ-self-concordant barrier F at a point x in any direction h is bounded from above by $\sqrt{\vartheta}$ times the x-norm $|h|_x$ of the direction. The announced statement says that this derivative is also bounded from above by another constant times the y-norm of the direction.

Proof of IV. Since $x \in \mathrm{int}\, G$, the closed unit Dikin ellipsoid W of F centered at x is contained in G (Lecture 2, **I.**; note that G is closed). Assume, first, that $\pi_x(y) > 0$. Then there exists $w \in G$ such that

$$y = x + \pi_x(y)(w - x).$$

Consider the image V of the ellipsoid W under the dilation mapping $z \mapsto z + \pi_x(y)(w-z)$; then

$$V = \{ y + h \mid |h|_x \leq (1 - \pi_x(y)) \}$$

is an $|\cdot|_x$-ball centered at y and at the same time $V \subset G$ (since $W \subset G$ and the dilation maps G into itself). From the semiboundedness property **I.** it follows that

$$DF(y)[h] \leq \vartheta \ \ \forall h : y + h \in G,$$

and since $V \subset G$, we conclude that

$$DF(y)[h] \leq \vartheta \ \ \forall h : |h|_x \leq 1 - \pi_x(y),$$

which is nothing but (3.9).

It remains to consider the case when $\pi_x(y) = 0$, so that the ray $[x, y)$ is contained in G. From convexity of G it follows that in the case in question $y - x$ is a recessive direction of G: $u + t(y-x) \in G$ whenever $u \in G$ and $t \geq 0$. In particular, the translation $V = W + (y - x)$ of W by the vector $y - x$ belongs to G; V is nothing but the $|\cdot|_x$-unit ball centered at y, and it remains to repeat word by word the above reasoning. ■

V. Uniqueness of minimizer and Centering property. *F is nondegenerate if and only if G does not contain lines. If G does not contain lines, then F attains its minimum on* int *G if and only if G is bounded, and if it is the case, the minimizer x_F^* - the F-center of G - is unique and possesses the following Centering property:*
The closed unit Dikin ellipsoid of F centered at x_F^* is contained in G, and the $\vartheta + 2\sqrt{\vartheta}$ times larger concentric ellipsoid contains G:

$$x \in G \Rightarrow |x - x_F^*|_{x_F^*} \leq \vartheta + 2\sqrt{\vartheta}. \tag{3.10}$$

Proof. As we know from Lecture 2, **II.**, the recessive subspace E_F of any self-concordant function is also the recessive subspace of its domain: int $G + E_F =$ int G. Therefore if G does not contain lines, then $E_F = \{0\}$, so that F is nondegenerate. Vice versa, if G contains a line with direction h, then $y = x + th \in \mathrm{int}\, G$ for all $x \in \mathrm{int}\, G$ and all $t \in \mathbf{R}$, from semiboundedness (see **I.**) it immediately follows that $DF(x)[y-x] = DF(x)[th] \leq \vartheta$ for all $x \in \mathrm{int}\, G$ and all $t \in \mathbf{R}$, which implies that $DF(x)[h] = 0$. Thus, F is constant along the direction h at any point of int G, so that $D^2F(x)[h,h] = 0$ and therefore F is degenerate.

From now on assume that G does not contain lines. If G is bounded, then F, of course, attains its minumum on int G due to the standard compactness reasons. Now assume that F attains its minimum on int G; due to nondegeneracy, the minimizer x_F^* is unique. Let W be the closed unit Dikin ellipsoid of F centered at x_F^*; as we know from **I.**, Lecture 2, it is contained in G (recall that G is closed). Let us prove that the $\vartheta + 2\sqrt{\vartheta}$ times larger concentric ellipsoid W^+ contains G; this will result both in the boundedness of G and in the announced centering property and therefore will complete the proof.

Lemma 3.2.1 *Let $x \in \operatorname{int} G$ and let h be an arbitrary direction with $|h|_x = 1$ such that $DF(x)[h] \geq 0$. Then the point $x + (\vartheta + 2\sqrt{\vartheta})h$ is outside the interior of G.*

Note that Lemma 3.2.1 immediately implies the desired inclusion $G \subset W^+$, since when $x = x_F^*$ is the minimizer of F, so that $DF(x)[h] = 0$ for all h, the premise of the lemma is valid for any h with $|h|_x = 1$.
Proof of Lemma. Let $\phi(t) = D^2F(x + th)[h, h]$ and $T = \sup\{t \mid x + th \in G\}$. From self-concordance of F it follows that

$$\phi'(t) \geq -2\phi^{3/2}(t),\ 0 \leq t < T,$$

whence

$$\left(\phi^{-1/2}(t)\right)' \leq 1,$$

so that

$$\frac{1}{\sqrt{\phi(t)}} - \frac{1}{\sqrt{\phi(0)}} \leq t,\ 0 \leq t < T.$$

In view of $\phi''(0) = |h|_x^2 = 1$ we come to

$$\phi(t) \geq \frac{1}{(1+t)^2},\ 0 \leq t < T,$$

which, after integration, results in

$$DF(x + rh)[h] \equiv \int_0^r \phi(t)dt \geq \int_0^r \frac{1}{(1+t)^2}dt = \frac{r}{1+r},\ 0 \leq r < T. \tag{3.11}$$

Now, let $t \geq 1$ be such that $y = x + th \in G$. Then, as we know from the semiboundedness relation (3.2),

$$(t - r)DF(x + rh)[h] \equiv DF(x + rh)[y - (x + rh)] \leq \vartheta.$$

Combining the inequalities, we come to

$$t \leq r + \frac{(1+r)\vartheta}{r}. \tag{3.12}$$

Taking here $r = 1/2$, we get certain upper bound on t; thus, $T \equiv \sup\{t \mid x + th \in G\} < \infty$, and (3.12) is valid for $t = T$. If $T > \sqrt{\vartheta}$, then (3.12) is valid for $t = T$, $r = \sqrt{\vartheta}$, and we come to

$$T \leq \vartheta + 2\sqrt{\vartheta}; \tag{3.13}$$

this latter inequality is, of course, valid in the case of $T \leq \sqrt{\vartheta}$ as well. Thus, T always satisfies (3.13). By construction, $x + Th$ is not an interior point of G, and, consequently, $x + [\vartheta + 2\sqrt{\vartheta}]h$ also is not an interior point of G, as claimed. ■

Corollary 3.2.1 *Let h be a recessive direction of G, i.e., such that $x + th \in G$ whenever $x \in G$ and $t \geq 0$. Then F is nonincreasing in the direction h, and the following inequality holds:*

$$-DF(x)[h] \geq \sqrt{D^2F(x)[h, h]},\ \forall x \in \operatorname{int} G. \tag{3.14}$$

Proof. Let $x \in \operatorname{int} G$; since h is a recessive direction, $y = x + th \in G$ for all $t > 0$, and **I.** implies that $DF(x)[y - x] = DF(x)[th] \leq \vartheta$ for all $t \geq 0$, whence $DF(x)[h] \leq 0$; thus, F indeed is nonincreasing in the direction h at any point $x \in \operatorname{int} G$. To prove (3.14), consider the restriction $f(t)$ of F onto the intersection of the line $x + \mathbf{R}h$ with G. Since h is a recessive direction for G, the domain of f is certain ray Δ of the type $(-a, \infty)$, $a > 0$. According to Proposition 3.1.1.(i), f is self-concordant barrier for the ray Δ. It is possible that f is degenerate: $E_f \neq \{0\}$. Since f is a function of one variable, it is possible only if $\Delta = E_f = \mathbf{R}$ (see **II.**, Lecture 2), so that $f'' \equiv 0$; in this case (3.14) is an immediate consequence of already proved nonnegativity of the left hand side in the relation. Now assume that f is nondegenerate. In view of **V.** f does not attain its minimum on Δ (since f is a nondegenerate self-concordant barrier for an *unbounded* domain). From **VIII.**, Lecture 2, we conclude that $\lambda(f,t) \geq 1$ for all $t \in \Delta$. Thus,

$$1 \leq \lambda(f,0) = \frac{(f'(0))^2}{f''(0)} = \frac{(DF(x)[h])^2}{D^2F(x)[h,h]},$$

which combined with already proved nonpositivity of $DF(x)[h]$ results in (3.14). ■

VI. Geometry of Dikin's ellipsoids. *For $x \in \operatorname{int} G$ and $h \in E$ let*

$$p_x(h) = \inf\{r \geq 0 \mid x \pm r^{-1}h \in G\};$$

this is nothing but the (semi)norm of h associated with the symmetrization of G with respect to x, i.e., the norm with the unit ball

$$G_x = \{y \in E \mid x \pm y \in G\}.$$

One has

$$p_x(h) \leq |h|_x \leq (\vartheta + 2\sqrt{\vartheta})p_x(h). \tag{3.15}$$

Proof. The first inequality in (3.15) is evident: we know that the closed unit Dikin ellipsoid of F centered at x is contained in G (since F is self-concordant and G is closed, see **I**, Lecture 2). In other words, G contains the unit $|\cdot|_x$ ball $\widehat{W}_1(x)$ centered at x; by definition, the unit $p_x(\cdot)$-ball centered at x is the largest symmetric with respect to x subset of G and therefore it contains the set $\widehat{W}_1(x)$, which is equivalent to the left inequality in (3.15). To prove the right inequality, this is the same as to demonstrate that if $|h|_x = 1$, then $p_x(h) \geq (\vartheta + 2\sqrt{\vartheta})^{-1}$, or, which is the same in view of the origin of p, that at least one of the two vectors $x \pm (\vartheta + 2\sqrt{\vartheta})h$ does not belong to the interior of G. Without loss of generality, let us assume that $DF(x)[h] \geq 0$ (if it is not the case, one should replace in what follows h with $-h$). The pair x, h satisfies the premise of Lemma 3.2.1, and this lemma says to us that the vector $x + (\vartheta + 2\sqrt{\vartheta})h$ indeed does not belong to the interior of G. ■

VII. Compatibility of Hessians. *Let $x, y \in \operatorname{int} G$. Then for any $h \in E$ one has*

$$D^2F(y)[h,h] \leq \left(\frac{\vartheta + 2\sqrt{\vartheta}}{1 - \pi_x(y)}\right)^2 D^2F(x)[h,h]. \tag{3.16}$$

Proof. By definition of the Minkowski function, there exists $w \in G$ such that

$$y = x + \pi_x(y)(w - x) = [1 - \pi_x(y)]x + \pi_x(y)w.$$

Now, if $|h|_x \leq 1$, then $x + h \in G$ (since the closed unit Dikin ellipsoid of F centered at x is contained in G), so that the point

$$y + [1 - \pi_x(y)]h = [1 - \pi_x(y)](x + h) + \pi_x(y)w$$

belongs to G. We conclude that the centered at y $|\cdot|_x$-ball of the radius $1 - \pi_x(y)$ is contained in G and therefore is contained in the largest symmetric with respect to x subset of G; in other words, we have

$$|h|_x \leq 1 - \pi_x(y) \Rightarrow p_y(h) \leq 1,$$

or, which is the same,

$$p_y(h) \leq [1 - \pi_x(y)]^{-1}|h|_x, \ \ \forall h.$$

Combining this inequality with (3.15), we come to (3.16). ■

We have established the main properties of self-concordant barriers; these properties, along with the already known to us properties of general self-concordant functions, underly all our further developments. Let me conclude with the statement of another type:

VIII. Existence of a self-concordant barrier for a given domain. *Let G be a closed convex domain in $\mathbf{R}^n$. Then there exists a ϑ-self-concordant barrier for G, with*

$$\vartheta \leq O(1)n,$$

$O(1)$ being an appropriate absolute constant. If G does not contain lines, then the above barrier is given by

$$F(x) = O(1) \ln \mathrm{Vol}\{\mathcal{P}_x(G)\},$$

where $O(1)$ is an appropriate absolute constant, Vol *is the n-dimensional volume and*

$$\mathcal{P}_x(G) = \{\xi \mid \xi^T(z - x) \leq 1 \;\; \forall z \in G\}$$

is the polar of G with respect to x.

I shall not prove this theorem, since we are not going to use it. Let me stress that to apply the theory we are developing to a particular convex problem, it is necessary and more or less sufficient to point out an explicit self-concordant barrier for the corresponding feasible domain. The aforementioned theorem says that such a barrier always exists, and thus gives us certain encouragement. At the same time, the "universal" barrier given by the theorem usually is too complicated numerically, since straightforward computation of a multidimensional integral involved into the construction is, typically, an untractable task. In the mean time we shall develop certain technique for constructing "computable" self-concordant barriers; although not that universal, this technique will equip us with good barriers for feasible domains of a wide variety of interesting and important convex programs.

3.3 Exercises: Self-concordant barriers

Let us start with a pair of simple exercises which will extend our list of examples of self-concordant barriers.

Exercise 3.3.1 $^{\#+}$ *Let $f(x)$ be a convex quadratic form on $\mathbf{R}^n$, and let the set $Q = \{x \mid f(x) < 0\}$ be nonempty. Prove that*

$$F(x) = -\ln(-f(x))$$

is a 1-self-concordant barrier for $G = \operatorname{cl} Q$.

Derive from this observation that if $G \subset \mathbf{R}^n$ is defined by a system

$$f_i(x) \le 0,\ i = 1, ..., m,$$

of convex quadratic inequalities which satisfies the Slater condition

$$\exists x : \quad f_i(x) < 0,\ i = 1, ..., m,$$

then the function

$$F(x) = -\sum_{i=1}^{m} \ln(-f_i(x))$$

is an m-self-concordant barrier for G.

Note that the result in question is a natural extension of Corollary 3.1.1.

Exercise 3.3.2 *

1) Let G be a bounded convex domain in $\mathbf{R}^n$ given by m linear or convex quadratic inequalities $f_j(x) \le 0$ satisfying the Slater condition:

$$G = \{x \in \mathbf{R}^m \mid f_j(x) \le 0,\ j = 1, ..., m\}.$$

Prove that if $m > 2n$, then one can eliminate from the system at least one inequality in such a way, that the remaining system still defines a bounded domain.

2) Derive from 1) that if $\{G_\alpha\}_{\alpha \in I}$ are closed convex domains in $\mathbf{R}^n$ with bounded and nonempty intersection, then there exist an at most $2n$-element subset I' of the index set I such that the intersection of the sets G_α over $\alpha \in I'$ also is bounded.

Note that the requirement $m > 2n$ in the latter exercise is sharp, as it is immediately demonstrated by the n-dimensional cube.

Exercise 3.3.3 $^{\#+}$ *Prove that the function*

$$F(x) = -\ln \operatorname{Det} x$$

is m-self-concordant barrier for the cone $\mathbf{S}^m_+$ of symmetric positive semidefinite $m \times m$ matrices.

Those who are not afraid of computations, are kindly asked to solve the following

Exercise 3.3.4 *Let*

$$K = \{(t, x) \in \mathbf{R} \times \mathbf{R}^n \mid t \ge |x|_2\}$$

be the "ice cream" cone. Prove that the function

$$F(x) = -\ln(t^2 - |x|_2^2)$$

is a 2-self-concordant barrier for K.

My congratulations, if you have solved the latter exercise! In the mean time we shall develop technique which will allow to demonstrate self-concordance of numerous barriers (including those given by the three previous exercises) *without any computations*; those solved exercises 3.3.1 - 3.3.4, especially the latter one, will, I believe, appreciate this technique.

Now let us switch to another topic. As it was announced in Lecture 1 and as we shall see in the mean time, the value of the parameter of a self-concordant barrier is something extremely important: this quantity is responsible for the Newton complexity (i.e., # of Newton steps) of finding an ε-solution by the interior point methods associated with the barrier. This is why it is interesting to realize what the value of the parameter could be.

Let us come to the statement announced in the beginning of Lecture 3:

(P): *Let F be ϑ-self-concordant barrier for a closed convex domain $G \subset \mathbf{R}^n$. Then either $G = \mathbf{R}^n$ and $F = const$, or G is a proper subset of $\mathbf{R}^n$ and $\vartheta \geq 1$.*

Exercise 3.3.5 $^{\#*}$ *Prove that the only self-concordant barrier for $\mathbf{R}^n$ is constant.*

Exercise 3.3.6 $^{\#*}$ *Prove that if Δ is a segment with a nonempty interior on the axis which differs from the whole axis and f is a ϑ-self-concordant barrier for Δ, then $\vartheta \geq 1$. Using this observation, complete the proof of (P).*

(P) says that the parameter of any self-concordant barrier for a nontrivial (differing from the whole space) convex domain G is ≥ 1. This lower bound can be extended as follows:

(Q) *Let G be a closed convex domain in $\mathbf{R}^n$ and let u be a boundary point of G. Assume that there is a neighbourhood U of u where G is given by m independent inequalities, i.e., there exist m continuously differentiable functions $g_1, ..., g_m$ on U such that*

$$G \cap U = \{x \in U \mid g_j(x) \geq 0, \ j = 1, ..., m\}, \ \ g_j(u) = 0, \ j = 1, ..., m,$$

and the gradients of g_j at u are linearly independent. Then the parameter ϑ of any self-concordant barrier F for G is at least m.

We are about to prove (Q). This is not that difficult, but to make the underlying construction clear, let us start with the case of a simple polyhedral cone.

Exercise 3.3.7 $^{\#*}$ *Let*

$$G = \{x \in \mathbf{R}^n \mid x_i \geq 0, \, i = 1, ..., m\},$$

where x_i are the coordinates in $\mathbf{R}^n$ and m is certain positive integer $\leq n$, and let F be a ϑ-self-concordant barrier for G. Prove that for any $x \in \operatorname{int} G$ one has

$$-x_i \frac{\partial}{\partial x_i} F(x) \geq 1, \ \ i = 1, ..., m; \tag{3.17}$$

derive from this observation that the parameter ϑ of the barrier F is at least m.

Now let us look at (Q). Under the premise of this statement G locally is similar to the above polyhedral cone; to make the similarity more explicit, let us translate G to make u the origin and let us choose the coordinates in $\mathbf{R}^n$ in such a way that the gradients of g_j at the origin, taken with respect to these coordinates, will be simply the first m basic orths. Thus, we come to the situation when G contains the origin and in certain neighbourhood U of the origin is given by

$$G \cap U = \{x \in U \mid x_i \geq h_i(x), \, i = 1, ..., m\},$$

where h_i are continuously differentiable functions such that $h_i(0) = 0$, $h_i'(0) = 0$.

Those who have solved the latter exercise understand that that what we need in order to prove (Q) is certain version of (3.17), something like

$$-r \frac{\partial}{\partial x_i} F(x(r)) \geq 1 - \alpha(r), \ \ i = 1, ..., m, \tag{3.18}$$

where $x(r)$ is the vector with the first m coordinates equal to $r > 0$ and the remaining ones equal to 0 and $\alpha(r) \to 0$ as $r \to +0$.

Relation of the type (3.18) does exist, as it is seen from the following exercise:

Exercise 3.3.8 $^{\#+}$ *Let $f(t)$ be a ϑ-self-concordant barrier for an interval $\Delta = [-a, 0]$, $0 < a \leq +\infty$, of the real axis. Assume that $t < 0$ is such that the point γt belongs to Δ, where*

$$\gamma > (\sqrt{\vartheta} + 1)^2.$$

Prove that

$$-f'(t)t \geq 1 - \frac{(\sqrt{\vartheta} + 1)^2}{\gamma} \tag{3.19}$$

Derive from this fact that if F is a ϑ-self-concordant barrier for $G \subset \mathbf{R}^n$, z is a boundary point of G and x is an interior point of G such that $z + \gamma(x - z) \in G$ with $\gamma > (\sqrt{\vartheta} + 1)^2$, then

$$-DF(x)[z - x] \geq 1 - \frac{(\sqrt{\vartheta} + 1)^2}{\gamma}. \tag{3.20}$$

Now we are in a position to prove (Q).

Exercise 3.3.9 $^{\#*}$ *Prove (Q).*

Chapter 4

Basic path-following method

The results on self-concordant functions and self-concordant barriers allow us to develop the first polynomial interior point scheme - the *path-following* one; on the qualitative level, the scheme was presented in Lecture I.

4.1 Situation

Let $G \subset \mathbf{R}^n$ be a closed and *bounded* convex domain, and let $c \in \mathbf{R}^n$, $c \neq 0$. In what follows we deal with the problem of minimizing the linear objective $c^T x$ over the domain, i.e., with the problem

$$\mathcal{P}: \quad \textit{minimize } c^T x \textit{ s.t. } x \in G.$$

I shall refer to problem $\mathcal{P}$ as to a *convex programming program in the standard form*. This indeed is a universal format of a convex program, since a general-type convex problem

$$\textit{minimize } f(u) \textit{ s.t. } g_j(u) \leq 0,\ j = 1, ..., m,\ u \in H \subset \mathbf{R}^k$$

associated with convex continuous functions f, g_j on a closed convex set H always can be rewritten as a standard problem; to this end it clearly suffices to set

$$x = (t, u),\ c = (1, 0, 0, ..., 0)^T,\ G = \{(t, u) \mid\ u \in H,\ g_j(u) \leq 0,\ j = 1, ..., m,\ f(x) - t \leq 0\}.$$

The feasible domain G of the equivalent standard problem is convex and closed; passing, if necessary, to the affine hull of G, we enforce G to be a domain. In our standard formulation, G is assumed to be bounded, which is not always the case, but the boundedness assumption is not so crucial from the practical viewpoint, since we can approximate the actual problem with an unbounded G by a problem with bounded feasible domain, adding, say, the constraint $|x|_2 \leq R$ with large R.

Thus, we may focus on the case of problem in the standard form $\mathcal{P}$. What we need to solve $\mathcal{P}$ by an interior point method, is a ϑ-self-concordant barrier for the domain, and in what follows we assume that we are given such a barrier, let it be called F. The exact meaning of the words "we know F" is that, given $x \in \operatorname{int} G$, we are able to compute the value, the gradient and the Hessian of the barrier at x.

4.2 F-generated path-following method

Recall that the general path-following scheme for solving $\mathcal{P}$ is as follows: given convex smooth and nondegenerate barrier F for the feasible domain G of the problem, we associate with this barrier and the objective the *penalized family*

$$F_t(x) = tc^T x + F(x) : \operatorname{int} G \to \mathbf{R},$$

$t > 0$ being the penalty parameter, and the *path of minimizers of the family*

$$x^*(t) = \operatorname*{argmin}_{\operatorname{int} G} F_t(\cdot)$$

which is well-defined due to nondegeneracy of F and boundedness of G. The method generates a sequence $x_i \in \mathrm{int}\, G$ which approximates the sequence $x^*(t_i)$ of points of the path along certain sequence of values of the penalty parameter $t_i \to \infty$; namely, given current pair (t_i, x_i) with x_i being "close" to $x^*(t_i)$, at an iteration of the method we replace t_i by a larger value of the parameter t_{i+1} and then update x_i into an approximation x_{i+1} to our new target point $x^*(t_{i+1})$. To update x_i, we apply to the new function of our family, i.e., to $F_{t_{i+1}}$, a method for smooth unconstrained minimization, x_i being the starting point. This is the general path-following scheme. Note that a self-concordant barrier for a *bounded* convex domain does satisfy the general requirements imposed by the scheme; indeed, such a barrier is convex, C^3 smooth and nondegenerate (the latter property is given by **V.**, Lecture 3). The essence of the matter is, of course, in the specific properties of a self-concordant barrier which make the scheme polynomial.

4.3 Basic path-following scheme

Even with the barrier fixed, the path-following scheme represents a family of methods rather than a single method; to get a method, one should specify

- policy for updating the penalty parameter;
- what is the "working horse" - the optimization method used to update x's;
- what is the stopping criterion for the latter method, or, which is the same, what is the "closeness to the path $x^*(\cdot)$" which is maintained when tracing the path.

In the *basic path-following method* we are about to present the aforementioned issues are specified as follows:

- we fix certain parameter $\gamma > 0$ - *the penalty rate* - and update t's according to the rule
$$t_{i+1} = (1 + \frac{\gamma}{\sqrt{\vartheta}})t_i; \qquad (4.1)$$
- to define the notion of "closeness to the path", we fix another parameter $\kappa \in (0,1)$ - *the path tolerance* - and maintain along the sequence $\{(t_i, x_i)\}$ the *closeness relation*, namely, the predicate
$$\mathcal{C}_\kappa(t,x): \ \{t>0\}\&\{x \in \mathrm{int}\, G\}\&\{\lambda(F_t,x) \equiv \sqrt{[\nabla_x F_t(x)]^T[\nabla_x^2 F(x)]^{-1}[\nabla_x F_t(x)]} \le \kappa\} \qquad (4.2)$$
(we write $\nabla_x^2 F$ instead of $\nabla_x^2 F_t$, since F differs from F_t by a linear function);
- the updating $x_i \mapsto x_{i+1}$ is given by the damped Newton method:
$$y^{l+1} = y^l - \frac{1}{1+\lambda(F_{t_{i+1}}, y^l)}[\nabla_x^2 F(y^l)]^{-1}\nabla_x F_{t_{i+1}}(y^l); \qquad (4.3)$$
the recurrency starts at $y^0 = x_i$ and is continued until the pair (t_{i+1}, y^l) turns out to satisfy the closeness relation $\mathcal{C}_\kappa(\cdot,\cdot)$; when it happens, we set $x_{i+1} = y^l$, thus coming to the updated pair (t_{i+1}, x_{i+1}).

The indicated rules specify the method, up to the initialization rule - where to take the very first pair (t_0, x_0) satisfying the closeness relation; in the mean time we will come to this latter issue. What we are interested in now are the convergence and the complexity properties of the method.

4.4 Convergence and complexity

The convergence and the complexity properties of the basic path-following method are described by the following two propositions:

Proposition 4.4.1 [Rate of convergence] *If a pair (t, x) satisfies the closeness relation $\mathcal{P}_\kappa$ with certain $\kappa \leq 1/4$, then*

$$c^T x - c^* \leq \frac{\chi}{t}, \;\; \chi = \vartheta + \frac{\kappa}{1-\kappa}\sqrt{\vartheta}, \tag{4.4}$$

c^* being the optimal value in $\mathcal{P}$ and ϑ being the parameter of the underlying self-concordant barrier F. In particular, in the above scheme one has

$$c^T x_i - c^* \leq \frac{\chi}{t_0}\left[1 + \frac{\gamma}{\sqrt{\vartheta}}\right]^{-i} \leq \frac{\chi}{t_0}\exp\{-O(1)\frac{i}{\sqrt{\vartheta}}\}, \tag{4.5}$$

with positive constant $O(1)$ depending on γ only.

Proof. Let $x^* = x^*(t)$ be the minimizer of F_t; let us start with proving that

$$c^T x^* - c^* \leq \frac{\vartheta}{t}; \tag{4.6}$$

in other words, when we are exactly on the trajectory, the residual in terms of the objective admits an objective-independent upper bound which is inverse proportional to the penalty parameter. This is immediate; indeed, denoting by x^+ a minimizer of our objective $c^T x$ over G, we have

$$\nabla_x F_t(x^*) = 0 \Rightarrow tc = -F'(x^*) \Rightarrow t(c^T x - c^T x^+) \equiv t(c^T x - c^*) = [F'(x^*)]^T(x^+ - x^*) \leq \vartheta$$

(the concluding inequality is the Semiboundedness property **I.**, Lecture 3, and (4.6) follows.

To derive (4.5) from (4.6), let us act as follows. The function $F_t(x)$ is self-concordant on int G (as a sum of two self-concordant functions, namely, F and a linear function $tc^T x$, see Proposition 2.1.1.(ii)) and, by assumption, $\lambda \equiv \lambda(F_t, x) \leq \kappa < 1$; applying (2.20) (see Lecture 2), we come to

$$|x - x^*|_{x^*} \leq \frac{\kappa}{1-\kappa}, \tag{4.7}$$

where $|\cdot|_{x^*}$ is the Euclidean norm defined by the Hessian of F_t, or, which is the same, of F, at x^*. We now have

$$\begin{gathered} tc = -F'(x^*) \Rightarrow \\ t(c^T x - c^T x^*) = [F'(x^*)]^T(x^* - x) \leq |x^* - x|_{x^*} \sup\{DF(x^*)[h] \mid |h|_{x^*} \leq 1\} = \\ = |x^* - x|_{x^*}\lambda(F, x^*) \leq \frac{\kappa}{1-\kappa}\sqrt{\vartheta} \end{gathered}$$

(the concluding inequality follows from (4.7) and the fact that F is a ϑ-self-concordant barrier for G, so that $\lambda(F, \cdot) \leq \sqrt{\vartheta}$). Thus,

$$|c^T x - c^T x^*| \leq \frac{\kappa}{t(1-\kappa)}\sqrt{\vartheta}, \tag{4.8}$$

which combined with (4.6) results in (4.4). ∎

Now we come to the central result

Proposition 4.4.2 [Newton complexity of a step] *The updating recurrency (4.3) is well-defined, i.e., it keeps the iterates in* int G *and terminates after finitely many steps; the Newton complexity of the recurrency, i.e., the # of Newton steps (4.3) before termination, does not exceed certain constant N which depends on the path tolerance κ and the penalty rate γ only.*

Proof. As we have mentioned in the previous proof, the function $F_{t_{i+1}}$ is self-concordant on int G and is below bounded on this set (since G is bounded). Therefore the damped Newton method does keep the iterates y^l in int G and ensures the stopping criterion $\lambda(F_{t_{i+1}}, y^l) \leq \kappa$ after a finite number of steps (**IX.**, Lecture 2). What we should prove is the fact that the Newton complexity of the updating is bounded from above by something depending solely on the path tolerance and the penalty rate. To make clear why it is important here that F is a self-concordant barrier rather than an arbitrary self-concordant function, let us start with the following reasoning.

We already have associated with a point $x \in$ int G the Euclidean norm

$$|h|_x = \sqrt{h^T F''(x) h} \equiv \sqrt{h^T F_t''(x) h};$$

in our case F is nondegenerate, so that $|\cdot|_x$ is an actual norm, not a seminorm. Let $|\cdot|_x^*$ be the conjugate norm:

$$|u|_x^* = \max\{u^T h \mid |h|_x \leq 1\}.$$

By definition of the Newton decrement,

$$\lambda(F_t, x) = \max\{[\nabla_x F_t(x)]^T h \mid |h|_x \leq 1\} = |\nabla_x F_t(x)|_x^* = |tc + F'(x)|_x^*, \tag{4.9}$$

and similarly

$$\lambda(F, x) = |F'(x)|_x^*. \tag{4.10}$$

Now, (t_i, x_i) satisfy the closeness relation $\lambda(F_t, x) \leq \kappa$, i.e.

$$|t_i c + F'(x)|_{x_i}^* \leq \kappa, \tag{4.11}$$

and F is ϑ-self-concordant barrier, so that $\lambda(F, x_i) \leq \sqrt{\vartheta}$, or, which is the same in view of (4.10),

$$|F'(x_i)|_{x_i}^* \leq \sqrt{\vartheta}. \tag{4.12}$$

Combining (4.11) and (4.12), we come to

$$|t_i e|_{x_i}^* \leq \kappa + \sqrt{\vartheta},$$

whence

$$|(t_{i+1} - t_i)e|_{x_i}^* = \frac{\gamma}{\sqrt{\vartheta}} |t_i e|_{x_i}^* \leq \gamma + \frac{\gamma\kappa}{\sqrt{\vartheta}}.$$

Combining the resulting inequality with (4.11), we come to

$$\lambda(F_{t_{i+1}}, x_i) = |t_{i+1} c + F'(x_i)|_{x_i}^* \leq \gamma + [1 + \frac{\kappa}{\sqrt{\vartheta}}]\gamma \leq 3\gamma \tag{4.13}$$

(the concluding inequality follows from the fact that the parameter of any nontrivial self-concordant barrier is ≥ 1, see the beginning of Lecture 3). Thus, the Newton decrement of the new function $F_{t_{i+1}}$ at the previous iterate x_i is at most the quantity 3γ; if γ and κ are small enough, this quantity is $\leq 1/4$, so that x_i is within the region of the quadratic convergence of the damped Newton method (see **IX.**, Lecture 2), and therefore the method quickly restores the closeness relation. E.g., let the path tolerance κ and the penalty rate γ be set to the value 0.05. Then the above computation results in

$$\lambda(F_{t_{i+1}}, x_i) \leq 0.15,$$

and from the description of the local properties of the damped Newton method as applied to a self-concordant function (see (2.19), Lecture 2) it follows that the Newton iterate y^1 of the starting point $y^0 = x_i$, the Newton method being applied to $F_{t_{i+1}}$, satisfies the relation

$$\lambda(F_{t_{i+1}}, y^1) \leq 2 \times (0.15)^2 = 0.045 < 0.05 = \kappa,$$

i.e., for the indicated values of the parameters a *single* damped Newton step restores the closeness to the path after the penalty parameter is updated, so that in this particular case $N = 1$. Note that the policy for updating the penalty - which is our presentation looked as something ad hoc - in fact is a consequence of the outlined reasoning: growth of the penalty given by

$$t \mapsto (1 + \frac{O(1)}{\sqrt{\vartheta}})t$$

is the highest one which results in the relation $\lambda(F_{t_{i+1}}, x_i) \leq O(1)$.

The indicated reasoning gives an insight on what is the intrinsic nature of the method: it does not allow, anyhow, to establish the announced statement in its complete form, since it requires certain bounds on the penalty rate. Indeed, our complexity results on the behaviour of the damped Newton method bound the complexity only when the Newton decrement at the starting point is less than 1. To "globalize" the reasoning, we should look at the initial residual in terms of the objective the Newton method is applied to rather than in terms of the initial Newton decrement. To this end let us prove the following

Proposition 4.4.3 *Let t and τ be two values of the penalty parameter, and let (t,x) satisfy the closeness relation $\mathcal{C}_\kappa(\cdot,\cdot)$ with some $\kappa<1$. Then*

$$F_\tau(x)-\min_u F_\tau(u)\leq \rho(\kappa)+\frac{\kappa}{1-\kappa}|1-\frac{\tau}{t}|\sqrt{\vartheta}+\vartheta\rho(1-\frac{\tau}{t}), \tag{4.14}$$

where, as always,

$$\rho(s)=-\ln(1-s)-s.$$

Proof. The path $x^*(\tau)$ is given by the equation

$$F'(u)+\tau c=0; \tag{4.15}$$

since F'' is nondegenerate, the Implicit Function Theorem says to us that $x^*(t)$ is continuously differentiable, and the derivative of the path can be found by differentiating (4.15) in τ:

$$(x^*)'(\tau)=-[F''(x^*(\tau))]^{-1}c. \tag{4.16}$$

Now let

$$\phi(\tau)=[\tau c^Tx^*(t)+F(x^*(t))]-[\tau c^Tx^*(\tau)+F(x^*(\tau))]$$

be the residual in terms of the objective $F_\tau(\cdot)$ taken at the point $x^*(t)$. We have

$$\phi'(\tau)=c^Tx^*(t)-c^Tx^*(\tau)-[\tau c+F'(x^*(\tau))]^T(x^*)'(\tau)=c^Tx^*(t)-c^Tx^*(\tau)$$

(see (4.15)). We conclude that

$$\phi(t)=\phi'(t)=0 \tag{4.17}$$

and that $\phi'(\cdot)=c^Tx^*(t)-c^Tx^*(\tau)$ is continuously differentiable; differentiating in τ once more and taking into account (4.16), we come to

$$\phi''(\tau)=-c^T(x^*)'(\tau)=c^T[F''(x^*(\tau))]^{-1}c,$$

which combined with (4.15) results in

$$0\leq\phi''(\tau)=\frac{1}{\tau^2}[F'(x^*(\tau))]^T[F''(x^*(\tau))]^{-1}F'(x^*(\tau))=\frac{1}{\tau^2}\lambda^2(F,x^*(\tau))\leq\frac{\vartheta}{\tau^2} \tag{4.18}$$

(we have used the fact that F is ϑ-self-concordant barrier).

From (4.17), (4.18) it follows that

$$\phi(\tau)\leq\vartheta\rho(1-\frac{\tau}{t}). \tag{4.19}$$

Now let us estimate the residual invloved into our target inequality (4.14):

$$F_\tau(x)-\min_u F_\tau(u)=F_\tau(x)-F_\tau(x^*(\tau))=[F_\tau(x)-F_\tau(x^*(t))]+[F_\tau(x^*(t))-F_\tau(x^*(\tau))]=$$

$$=[F_\tau(x)-F\tau(x^*(t))]+\phi(\tau)=[F_t(x)-F_t(x^*(t))]+(t-\tau)c^T(x-x^*(t))+\phi(\tau); \tag{4.20}$$

since $F_t(\cdot)$ is self-concordant and $\lambda(F_t,x)\leq\kappa<1$, we have $F_t(x)-F_t(x^*(t))=F_t(x)-\min_u F_t(u)\leq\rho(\lambda(F_t,x))$ (see (2.16), Lecture 2), whence

$$F_t(x)-F_t(x^*(t))\leq\rho(\kappa). \tag{4.21}$$

(4.8) says to us that $|c^T(x-x^*(t))|\leq\kappa(1-\kappa)^{-1}\sqrt{\vartheta}t^{-1}$; combining this inequality, (4.20) and (4.19), we come to (4.14). ■

Now we are able to complete the proof of Proposition 4.4.2. Applying (4.14) to $x=x_i$, $t=t_i$ and $\tau=t_{i+1}=(1+\frac{\gamma}{\sqrt{\vartheta}})t_i$, we come to

$$F_{t_{i+1}}(x_i)-\min_u F_{t_{i+1}}(u)\leq\rho(\kappa)+\frac{\kappa\gamma}{1-\kappa}+\vartheta\rho(\frac{\gamma}{\sqrt{\vartheta}}),$$

and the left hand side of this inequality is bounded from above uniformly in $\vartheta\geq 1$ by certain function depending on κ and γ only (as it is immediately seen from the evident relation $\rho(s)\leq O(s^2)$, $|s|\leq\frac{1}{2}$ [1]). ■

An immediate consequence of Propositions 4.4.1 and 4.4.2 is the following

[1] here is the corresponding reasoning: if $s\equiv\gamma\vartheta^{-1/2}\leq 1/2$, then $g\equiv\vartheta\rho(\gamma\vartheta^{-1/2})\leq O(1)\gamma^2$ due to $0\leq s\leq 1/2$; if $s>1/2$, then $\vartheta\leq 4\gamma^2$, and consequently $g\leq 4\gamma^2\ln\gamma$; note that $\vartheta\geq 1$. Thus, in all cases the last term in the estimate is bounded from above by certain function of γ

Theorem 4.4.1 *Let problem $\mathcal{P}$ with a closed convex domain $G \subset \mathbf{R}^n$ be solved by the path-following method associated with a ϑ-self-concordant barrier F, let $\kappa \in (0,1)$ and $\gamma > 0$ be the path tolerance and the penalty rate used in the method, and let (t_0, x_0) be the starting pair satisfying the closeness relation $\mathcal{C}_\kappa(\cdot,\cdot)$. Then the absolute inaccuracy $c^T x_i - c^*$ of approximate solutions generated by the method admits the upper bound*

$$c^T x_i - c^* \leq \frac{2\vartheta}{t_0}(1+\frac{\gamma}{\sqrt{\vartheta}})^{-i}, \ i = 1, 2, ... \tag{4.22}$$

and the Newton complexity of each iteration $(t_i, x_i) \mapsto (t_{i+1}, x_{i+1})$ of the method does not exceed certain constant N depending on κ and γ only. In particular, the Newton complexity (total # of Newton steps) of finding an ε-solution to the problem, i.e., of finding $x \in G$ such that $c^T x - c^* \leq \varepsilon$, is bounded from above by

$$O(1)\sqrt{\vartheta}\ln\left(\frac{\vartheta}{t_0\varepsilon}+1\right),$$

with constant factor $O(1)$ depending solely on κ and γ.

4.5 Initialization and two-phase path-following method

The aforementioned description of the method is uncomplete - we know how to follow the path $x^*(\cdot)$, provided that we once came close to it, but we do not know yet how to get close to the path to start the tracing. There are several ways to resolve this *initialization difficulty*, and the simplest one is as follows. We know where the path $x^*(t)$ ends, where it tends to as $t \to \infty$ - all cluster points of the path belong to the optimal set of the problem. Let us look where the path starts, i.e., where it tends as $t \to +0$. The answer is evident - as $t \to +0$, the path

$$x^*(t) = \operatorname{argmin}[tc^T x + F(x)]$$

tends to the *analytic center of* G *with respect to* F, to the minimizer x^*_F of F over G (since G is bounded, we know from **V.**, Lecture 3, that this minimizer does exist and is unique). Thus, all F-generated paths associated with various objectives c start at the same point - the analytic center of G - and run away from this point as $t \to \infty$, each to the optimal set associated with the corresponding objective. In other words, the analytic center of G is close to all the paths generated by F, so that it is a good position to start following the path we are interested in. Now, how to come to this position? An immediate idea is as follows: the paths associated with various objectives cover the whole interior of G: if $x \neq x^*$ is an interior point of G, then a path passing through x is given by any objective of the form

$$d = -\lambda F'(x),$$

λ being positive; the path with the indicated objective passes through x when the value of the penalty parameter is exactly λ. This observation suggests the following initialization scheme: given a *starting point* $\widehat{x} \in \operatorname{int} G$, let us follow the artificial path

$$u^*(\tau) = \operatorname{argmin}[\tau d^T x + F(x)], \ \ d = -F'(\widehat{x})$$

in the "inverse time", i.e., decreasing the penalty parameter τ rather than increasing it. The artificial path clearly passes through the point $\widehat{x}$:

$$\widehat{x} = u^*(1),$$

and we can start tracing it with the pair $(\tau_0 = 1, u_0 = \widehat{x})$ which is exactly at the path. When tracing the path in the outlined manner, we in the mean time come close to the analytic center of G and, consequently, to the path $x^*(t)$ we are interested in; when it happens, we can switch to tracing this target path.

The outlined ideas underly the

Two-Phase Path-Following Method:

Input: starting point $\widehat{x} \in \operatorname{int} G$; path tolerance $\kappa \in (0,1)$; penalty rate $\gamma > 0$.

Phase 0 [approximating the analytic center] *Starting with* $(\tau_0, u_0) = (1, \widehat{x})$*, generate the sequence* $\{(\tau_i, u_i)\}$*, updating* (t_i, u_i) *into* (τ_{i+1}, u_{i+1}) *as follows:*

- $$\tau_{i+1} = \left[1 + \frac{\gamma}{\sqrt{\vartheta}}\right]^{-1} \tau_i;$$

- *to get u_{i+1}, apply to the function*

  $$\widehat{F}_{\tau_i}(x) \equiv \tau d^T x + F(x)$$

  the damped Newton method

  $$y^{l+1} = y^l - \frac{1}{1 + \lambda(\widehat{F}_{\tau_{i+1}}, y^l)} [\nabla_x^2 F(y^l)]^{-1} \nabla_x \widehat{F}_{\tau_{i+1}}(y^l)$$

  starting with $y^0 = u_i$. Terminate the method when the pair (τ_{i+1}, y^l) turns out to satisfy the predicate

  $$\widehat{\mathcal{C}}_{\kappa/2}(\tau, u) : \quad \{\tau > 0\}\&\{u \in \operatorname{int} G\}\&\{\lambda(\widehat{F}_\tau, u) \leq \kappa/2\}; \tag{4.23}$$

  when it happens, set

  $$u_{i+1} = y^l;$$

- *after (τ_{i+1}, u_{i+1}) is formed, check whether*

  $$\lambda(F, u_{i+1}) \leq \frac{3}{4}\kappa; \tag{4.24}$$

  if it happens, terminate Phase 0 and call $u^* \equiv u_{i+1}$ the result of the phase, otherwise go to the next step of Phase 0.

Initialization of Phase 1. *Given the result u^* of Phase 0, set*

$$t_0 = \max\{t \mid \lambda(F_t, u^*) \leq \kappa\}, \; x_0 = u^*, \tag{4.25}$$

thus obtaining the pair (t_0, x_0) satisfying the predicate $\mathcal{C}_\kappa(\cdot, \cdot)$.

Phase 1. [approximating optimal solution to $\mathcal{P}$] *Starting with the pair (t_0, x_0), form the sequence $\{(t_i, x_i)\}$ according to the Basic path-following scheme from Section 4.3, namely, given (t_i, x_i), update it into (t_{i+1}, x_{i+1}) as follows:*

- $$t_{i+1} = \left[1 + \frac{\gamma}{\sqrt{\vartheta}}\right] t_i;$$

- *to get x_{i+1}, apply to $F_{t_{i+1}}$ the damped Newton method*

  $$y^{l+1} = y^l - \frac{1}{1 + \lambda(F_{t_{i+1}}, x_i)} [\nabla_x^2 F(y^l)]^{-1} \nabla_x F_{t_{i+1}}(y^l), \tag{4.26}$$

  starting with $y^0 = x_i$. Terminate the method when the pair (t_{i+1}, y^l) turns out to satisfy the predicate $\mathcal{C}_\kappa(\cdot, \cdot)$; when it happens, set

  $$x_{i+1} = y^l,$$

  thus obtaining the updated pair satisfying the predicate $\mathcal{C}_\kappa$, and go to the next step of Phase 1.

The properties of the indicated method are described in the following statement:

Theorem 4.5.1 *Let problem $\mathcal{P}$ be solved by the two-phase path-following method associated with a ϑ-self-concordant barrier for the domain G (the latter is assumed to be bounded). Then*

(i) *Phase 0 is finite and is comprised of no more than*

$$N_{\text{ini}} = O(1)\sqrt{\vartheta} \ln\left(\frac{\vartheta}{1 - \pi_{x_F^*}(\widehat{x})} + 1\right) \tag{4.27}$$

iterations, with no more than $O(1)$ Newton steps (4.23) at every iteration; here and further $O(1)$ are constant factors dpending solely on the path tolerance κ and the penalty rate γ used in the method.

(ii) *For any $\varepsilon > 0$, the number of iterations of Phase 1 before an ε-solution to $\mathcal{P}$ is generated, does not exceed the quantity*

$$N_{\rm main}(\varepsilon) = O(1)\sqrt{\vartheta}\ln\left(\frac{\vartheta \mathrm{Var}_G(c)}{\varepsilon}+1\right), \tag{4.28}$$

where

$$\mathrm{Var}_G(c) = \max_{x\in G} c^T x - \min_{x\in G} c^T x,$$

with no more than $O(1)$ Newton steps (4.26) at every iteration.

In particular, the overall Newton complexity (total # of Newton steps of the both phases) of finding an ε-solution to the problem does not exceed the quantity

$$N_{\rm total}(\varepsilon) = O(1)\sqrt{\vartheta}\ln\left(\frac{\mathcal{V}}{\varepsilon}+1\right),$$

where the data-dependent constant $\mathcal{V}$ is given by

$$\mathcal{V} = \frac{\vartheta \mathrm{Var}_G(c)}{1-\pi_{x^*_F}(\widehat{x})}.$$

Proof.

1^0. Following the line of argument used in the proof of Proposition 4.4.2, one can immediately verify that the iterations of Phase 0 are well-defined and maintain along the sequence $\{(\tau_i, u_i)\}$ the predicate $\widehat{\mathcal{C}}_{\kappa/2}(\cdot,\cdot)$, while the Newton complexity of every iteration of the phase does not exceed $O(1)$. To complete the proof of (i), we should establish upper bound (4.27) on the number of iterations of Phase 0. To this end let us note that $\widehat{\mathcal{C}}_{\kappa/2}(\tau_i, u_i)$ means exactly that

$$\lambda(\widehat{F}_{\tau_i}, u_i) = |\tau_i d + F'(u_i)|^*_{u_i} \le \kappa/2, \tag{4.29}$$

(compare with (4.9)), whence

$$\lambda(F, u_i) = |F'(u_i)|^*_{u_i} \le \kappa/2 + \tau_i |d|^*_{u_i} = \kappa/2 + \tau_i |F'(\widehat{x})|^*_{u_i}. \tag{4.30}$$

We have

$$|F'(\widehat{x})|^*_{x^*_F} \equiv \max\{h^T F'(\widehat{x}) \mid |h|_{x^*_F} \le 1\} = \max\{DF(\widehat{x})[h] \mid D^2F(x^*_F)[h,h] \le 1\} \le$$

[see **IV.**, Lecture 3, namely, (3.9)]

$$\le \alpha \equiv \frac{\vartheta}{1-\pi_{x^*_F}(\widehat{x})}.$$

We see that the variation (the difference between the minumum and the maximum values) of the linear form $f(y) = y^T F'(\widehat{x})$ over the unit Dikin ellipsoid of F centered at x^*_F does not exceed 2α. Consequently, the variation of the form on the $(\vartheta + 2\sqrt{\vartheta})$-larger concentric ellipsoid W^* does not exceed $2\alpha(\vartheta + 2\sqrt{\vartheta})$. From the Centering property **V.**, Lecture 3, we know that W^* contains the whole G; in particular, W^* contains the unit Dikin ellipsoid $\widehat{W_1}(u_i)$ of F centered at u_i (**I.**, Lecture 2). Thus, the variation of the linear form $y^T F'(\widehat{x})$ over the ellipsoid $\widehat{W_1}(u_i)$, and this is nothing but twice the quantity $|F'(\widehat{x})|^*_{u_i}$, does not exceed $2\alpha(\vartheta + 2\sqrt{\vartheta})$:

$$|F'(\widehat{x})|^*_{u_i} \le \beta \equiv \frac{\vartheta(\vartheta + 2\sqrt{\vartheta})}{1-\pi_{x^*}(\widehat{x})}.$$

Substituting this estimate in (4.30), we come to

$$\lambda(F, u_i) \le \kappa/2 + \tau_i \beta.$$

Taking into account that $\tau_i = (1 + \frac{\gamma}{\sqrt{\vartheta}})^{-i}$, we conclude that the stopping criterion $\lambda(F, u_i) \le 3\kappa/4$ for sure is satisfied when i is $O(1)\ln(1 + \vartheta(1-\pi_{x^*_F}(\widehat{x}))^{-1})$, as claimed in (i).

2^0. Now let us verify that

$$t_0 \geq \frac{\kappa \mathrm{Var}_G(c)}{2}. \tag{4.31}$$

Indeed, since $c \neq 0$, it follows from the origin of t_0 (see (4.25)) that

$$\lambda(F_{t_0}, u^*) \equiv |t_0 c + F'(u^*)|^*_{u^*} = \kappa, \tag{4.32}$$

while from the termination rule for Phase 0 we know that

$$\lambda(F, u^*) \equiv |F'(u^*)|^*_{u^*} \leq \frac{3}{4}\kappa;$$

we immediately conclude that

$$t_0 |c|^*_{u^*} \geq \frac{\kappa}{2}.$$

Now, as above, $|c|^*_{u^*}$ is the variation of the linear form $y^T c$ over the closed unit Dikin ellipsoid of F centered at u^*; this ellipsoid is contained in G (**I.**, Lecture 2), whence $|c|^*_{u^*} \leq \mathrm{Var}_G(c)$. Thus,

$$t_0 \mathrm{Var}_G(c) \geq \frac{\kappa}{4},$$

and (4.31) follows.

3^0. In view of (4.32), the starting pair $(t_0, x_0 \equiv u^*)$ for Phase 1 satisfies the predicate $\mathcal{C}_\kappa$; applying Theorem 4.4.1 and taking into account (4.31), we come to (ii). ■

4.6 Concluding remarks

We have seen that the basic path-following method for solving $\mathcal{P}$ associated with a ϑ-self-concordant barrier F for feasible domain G of the problem finds an ε-solution to $\mathcal{P}$ in no more than

$$\mathcal{N}(\varepsilon) = O(1)\sqrt{\vartheta} \ln\left(\frac{\mathcal{V}}{\varepsilon}\right)$$

damped Newton steps; here $O(1)$ depends on the path tolerance κ and the penalty rate γ only, and $\mathcal{V}$ is certain data-dependent quantity (note that we include into the data the starting point $\widehat{x} \in \mathrm{int}\, G$ as well). When κ and γ are once for ever fixed absolute constants, then the above $O(1)$ also is an absolute constant; in this case we see that *if the barrier F is "computable", i.e., given x and the data vector $\mathcal{D}(p)$ identifying the problem instance, one can compute $F(x)$, $F'(x)$ and $F''(x)$ in polynomial in $l(p) \equiv \dim \mathcal{D}(p)$ number of arithmetic operations $\mathcal{M}$, then the method is polynomial (see Lecture 1), and the arithmetic cost of finding an ε-solution by the method does not exceed the quantity*

$$\mathcal{M}(\varepsilon) = O(1)[\mathcal{M} + n^3]\mathcal{N}(\varepsilon)$$

(the term n^3 is responsible for the arithmetic cost of solving the Newton system at a Newton step).

Consider, e.g., a Linear Programming problem

$$\textit{minimize } c^T x \textit{ s.t. } a_j^T x \leq b_j,\ j = 1, ..., m,\ x \in \mathbf{R}^n,$$

and assume that the system of linear inequalities $a_j^T x \leq b_j$, $j = 1, ..., m$, satisfies the Slater condition and defines a polytope (i.e., a bounded polyhedral set) G. As we know from Corollary 3.1.1, the standard logarithmic barrier

$$F(x) = -\sum_{j=1}^m \ln(b_j - a_j^T x)$$

is m-self-concordant logarithmic barrier for G. Of course, this barrier is "computable":

$$F'(x) = \sum_{j=1}^m \frac{a_j}{b_j - a_j^T x}, \quad F''(x) = \sum_{j=1}^m \frac{a_j a_j^T}{(b_j - a_j^T x)^2},$$

and we see that the arithmetic cost of computing $F(x)$, $F'(x)$ and $F''(x)$ is $O(mn^2)$, while the dimension of the data vector for a problem instance is $O(mn)$. Therefore the path-following method associated with the standard logarithmic barrier for the polytope G finds an ε-solution to the problem at the cost of

$$\mathcal{N}(\varepsilon) = O(1)\sqrt{m}\ln\left(\frac{\mathcal{V}}{\varepsilon}+1\right)$$

Newton steps, with the arithmetic cost of a step $O(1)mn^2$ (the arithmetic cost $O(n^3)$ of solving the Newton system is dominated by the cost of assembling the system, i.e., that one of computing F' and F''; indeed, since G is bounded, we have $m > n$). Thus, the overall arithmetic cost of finding an ε-solution to the problem is

$$\mathcal{M}(\varepsilon) = O(1)m^{1.5}n^2\ln\left(\frac{\mathcal{V}}{\varepsilon}+1\right),$$

so that the "arithmetic cost of an accuracy digit" is $O(m^{1.5}n^3)$. In fact the latter cost can be reduced to $O(mn^2)$ by proper implementation of the method (the Newton systems arising at the neighbouring steps of the method are "close" to each other, which allows to reduce the average over steps arithmetic cost of solving the Newton systems), but I am not going to speak about these *acceleration issues*.

What should be stressed is that the outlined method is fine from the viewpoint of its theoretical complexity; it is, anyhow, far from being appropriate in practice. The main drawback of the method is its "short-step" nature: to ensure the theoretical complexity bounds, one is enforced to increase the penalty parameter at the rate $(1 + O(1)\vartheta^{-1/2})$, so that the number of Newton steps is proportional to $\sqrt{\vartheta}$. For an LP problem of a not too large size - say, $n = 1000$, $m = 10000$, the method would require solving several hundreds, if not thousands, linear systems with 1000 variables, which will take hours - time incomparable with that one required by the simplex method; and even moderate increasing of sizes results in days and months instead of hours. You should not think that these unpleasant practical consequences are caused by the intrinsic drawbacks of the scheme; they come from our "pessimistic" approach to the implementation of the scheme. It turns out that "most of the time" you can increase the penalty at a significantly larger rate than that one given by the worst-case theoretical complexity analysis, and still will be able to restore closeness to the path by a small number - 1-2 - of Newton steps. There are very good practical implementations of the scheme which use various on-line strategies to control the penalty rate and result in a very reasonable - 20-40 - total number of Newton steps, basically independent of the size of the problem; the best examples known to me are the codes developed in the Optimization Laboratory of our faculty by Gil Roth and Michael Zibulevski. From the theoretical viewpoint, anyhow, it is important to develop computationally cheap rules for on-line adjusting the penalty rate which *ensure* the theoretical $O(\sqrt{\vartheta})$ Newton complexity of the method; in the mean time we shall speak about recent progress in this direction.

4.7 Exercises: Basic path-following method

The proof of our main rate-of-convergence statement - Proposition 4.4.1 - is based on the following fact:
(*) if x belongs to the path $x^*(t) = \mathrm{argmin}_{\mathrm{int}\, G}[tc^T x + F(x)]$: $x = x^*(t)$ for certain $t > 0$, then

$$c^T x - c^* \leq \frac{\vartheta}{t},$$

c^* being the optimal value in $\mathcal{P}$. What is responsible for this remarkable and simple inequality? The only property of a ϑ-self-concordant barrier F used in the corresponding place of the proof of Proposition 4.4.1 was the semiboundedness property:

$$DF(x)[y - x] \leq \vartheta \ \ \forall x \in \mathrm{int}\ G \ \ \forall y \in G. \tag{4.33}$$

In turn looking at the proof of this property (**0., I.**, Lecture 3), one can find out that the only properties of F and G used there were the following ones:

$S(\vartheta)$: $G \in \mathbf{R}^n$ is a closed convex domain; F is a twice continuously differentiable convex function on int G *such that*

$$DF(x)[h] \leq \vartheta^{1/2}\{D^2 F(x)[h,h]\}^{1/2} \ \ \forall x \in \mathrm{int}\ G \ \ \forall h \in \mathbf{R}^n.$$

Thus, (4.33) has nothing to do with self-concordance of F.

Exercise 4.7.1 $^{\#}$ *Verify that $S(\vartheta)$ implies (4.33).*

Exercise 4.7.2 $^{\#}$ *Prove that property $S(\cdot)$ is stable with respect to affine substitutions of argument and with respect to summation; namely, prove that*
1) if the pair $(G \subset \mathbf{R}^n, F)$ satisfies $S(\vartheta)$ and $y = \mathcal{A}(x) \equiv Ax + a$ is an affine mapping from $\mathbf{R}^k$ into $\mathbf{R}^n$ with the image intersecting int G*, then the pair $(\mathcal{A}^{-1}(G), F(\mathcal{A}(\cdot)))$ also satisfies $S(\vartheta)$;*
2) if the pairs $(G_i \subset \mathbf{R}^n, F_i)$, $i = 1, ..., m$, satisfy $S(\vartheta_i)$ and $G = \cap_i G_i$ is a domain, then the pair $(G, \sum_i \alpha_i F_i)$, $\alpha_i \geq 0$, satisfies $S(\sum_i \alpha_i \vartheta_i)$.

Now let us formulate a simple necessary and sufficient condition for a pair (G, F) to satisfy $S(\vartheta)$.

Exercise 4.7.3 $^{\#}$ *Let $\vartheta > 0$, and let $(G \subset \mathbf{R}^n, F)$ be a pair comprised of a closed convex domain and a function twice continuously differentiable on the interior of the domain. Prove that (G, F) sastisfies $S(\vartheta)$ if and only if the function* $\exp\{-\vartheta F\}$ *is concave on* int G*. Derive from this observation and the result of the previous exercise the following statement (due to Fiacco and McCormic):*
let g_i, $i = 1, ..., m$, be convex twice continuously differentiable functions on $\mathbf{R}^n$ satisfying the Slater condition. Consider the logarithmic barrier

$$F(x) = -\sum_i \ln(-g_i(x))$$

for the domain

$$G = \{x \in \mathbf{R}^n \mid g_i(x) \leq 0, \ i = 1, ..., m\}.$$

Then the pair (G, F) satisfies $S(m)$, and therefore F satisfies relation (4.33) with $\vartheta = m$. In particular, let

$$x \in \mathop{\mathrm{Argmin}}_{u \in \mathrm{int}\ G}[tc^T u + F(u)]$$

for some positive t; then $f(u) \equiv c^T u$ is below bounded on G and

$$c^T x - \inf_G f \leq \frac{m}{t}.$$

The next exercise is an ”exercise” in the direct meaning of the word.

Exercise 4.7.4 *Consider a Quadratically Constrained Quadratic Programming program*

$$\text{minimize } f_0(x) \text{ s.t. } f_j(x) \le 0,\ j = 1, ..., m,\ x \in \mathbf{R}^n,$$

where

$$f_j(x) = x^T A_j x + 2b_j^T x + c_j,\ \ j = 0, ..., m$$

are convex quadratic forms. Assume that you are given a point $\widehat{x}$ such that $f_j(\widehat{x}) < 0$, $j = 1, ..., m$, and $R > 0$ such that the feasible set of the problem is inside the ball $\{x \mid |x|_2 \le R\}$.
1) reduce the problem to the standard form with a bounded feasible domain and point out an $(m+2)$-self-concordant barrier for the domain, same as an interior point of the domain;
2) write down the algorithmic scheme of the associated path-following method. Evaluate the arithmetic cost of a Newton step of the method.

Now let us discuss the following issue. In the Basic path-following method the rate of updating the penalty parameter, i.e., the *penalty ratio*

$$\omega = t_{i+1}/t_i,$$

is set to $1 + O(1)\vartheta^{-1/2}$, ϑ being the parameter of the underlying barrier. This choice of the penalty ratio results in the best known, namely, proportional to $\sqrt{\vartheta}$, theoretical complexity bound for the method. In Lecture 4 it was explained that this fine theoretically choice of the penalty ratio in practice makes the method almost useless, since it *for sure* enforces the method to work according its theoretical worst-case complexity bound; the latter bound is in many cases too large for actual computations. In practice people normally take as the initial value of the penalty ratio certain moderate constant, say, 2 or 3, and then use various routines for on-line adjusting the ratio, slightly increasing/decreasing it depending on whether the previous updating $x_i \mapsto x_{i+1}$ took "small" or "large" (say, ≤ 2 or > 2) number of Newton steps. An immediate theoretical question here is: what can be said about the Newton complexity of a path-following method where the penalty ratio is a once for ever fixed constant $\omega > 1$ (or, more generally, varies somehow between once for ever fixed bounds $\omega_- < \omega_+$, with $1 < \omega_- \le \omega_+ < \infty$). The answer is that *in this case the Newton complexity of an iteration $(t_i, x_i) \mapsto (t_{i+1}, x_{i+1})$ is of order of ϑ rather than of order of 1.*

Exercise 4.7.5 *Consider the Basic path-following method from Section 4.3 with rule (4.1) replaced with*

$$t_{i+1} = \omega_i t_i,$$

where $\omega_- \le \omega_i \le \omega_+$ and $1 < \omega_- \le \omega_+ < \infty$. Prove that for this version of the method the statement of Theorem 4.4.1 should be modified as follows: the total # of Newton steps required to find an ε-solution to $\mathcal{P}$ can be bounded from above as

$$O(1)\vartheta \ln\left(\frac{\vartheta}{t_0\varepsilon} + 1\right),$$

with $O(1)$ depending only on $\kappa, \omega_-, \omega_+$.

Chapter 5

Conic problems and Conic Duality

In the previous lecture we dealt with the Basic path-following interior point method. It was explained that the method, being fine theoretically, is not too attractive from the practical viewpoint, since it is a routine with a prescribed (and normally close to 1) rate of updating the penalty parameter; as a result, the actual number of Newton steps in the routine is more or less the same as the number given by the theoretical worst-case analysis and for sure is proportional to $\sqrt{\vartheta}$, ϑ being the parameter of the underlying self-concordant barrier. For large-scale problems, ϑ normally is large, and the # of Newton steps turns out to be too large for practical applications. The source of difficulty is the conceptual drawback of our scheme: everything is strictly regulated, there is no place to exploit favourable circumstances which may occur. As we shall see in the mean time, this conceptual drawback can be eliminated, to certain extent, even within the path-following scheme; there is, anyhow, another family of interior point methods, the so called *potential reduction* ones, which are free of this drawback of strict regulation; some of these methods, e.g., the famous - and the very first - interior point method of Karmarkar for Linear Programming, turn out to be very efficient in practice. The methods of this potential reduction type are what we are about to investigate now; the investigation, anyhow, should be preceded by developing a new portion of tools, interesting in their own right. This development is our today goal.

5.1 Conic problems

In order to use the path-following method from the previous lecture, one should reduce the problem to the specific form of minimizing a linear objective over convex domain; we called this form *standard*. Similarly, to use a potential reduction method, one also needs to represent the problem in certain specific form, called *conic*; I am about to introduce this form.

Cones. Recall that a convex cone K in $\mathbf{R}^n$ is a nonempty convex set with the property

$$tx \in K \ \textit{whenever} \ \ x \in K \ \textit{and} \ \ t \geq 0;$$

in other words, a cone should contain with any of its points the whole ray spanned by the point. A convex cone is called *pointed*, if it does not contain lines.

Given a convex cone $K \subset \mathbf{R}^n$, one can define its *dual* as

$$K^* = \{s \in \mathbf{R}^n \mid s^T x \geq 0 \ \forall x \in K\}.$$

In what follows we use the following elementary facts about convex cones: let $K \subset \mathbf{R}^n$ be a closed convex cone and K^* be its dual. Then

- K^* is closed convex cone, and the cone $(K^*)^*$ dual to it is nothing but K.
- K is pointed if and only if K^* has a nonempty interior; K^* is pointed if and only if K has a nonempty interior. The interior of K^* is comprised of all vectors s *strictly positive* on K, i.e., such that $s^T x > 0$ for all nonzero $x \in K$.
- $s \in K^*$ is strictly positive on K if and only if the set $K(s) = \{x \in K \mid s^T x \leq 1\}$ is bounded.

An immediate corollary of the indicated facts is that a closed convex cone K is pointed and possesses a nonempty interior if and only if its dual shares these properties.

Conic problem. Let $K \subset \mathbf{R}^n$ be a closed pointed convex cone with a nonempty interior. Consider optimization problem

$$(\mathcal{P}): \quad \textit{minimize } \; c^T x \;\; \textit{s.t.} \;\; x \in \{b + L\} \cap K,$$

where

- L is a linear subspace in $\mathbf{R}^n$;
- b is a vector from $\mathbf{R}^n$.

Geometrically: we should minimize a linear objective ($c^T x$) over the intersection of an affine plane ($b+L$) with the cone K. This intersection is a convex set, so that $(\mathcal{P})$ is a convex program; let us refer to it as to *convex program in the conic form*.

Note that a program in the conic form strongly resembles a Linear Programming program in the standard form; this latter problem is nothing but $(\mathcal{P})$ with K specified as the nonnegative orthant $\mathbf{R}^n_+$. On the other hand, $(\mathcal{P})$ is a *universal* form of a convex programming problem. Indeed, it suffices to demonstrate that a standard convex problem

$$(\mathcal{S}) \quad \textit{minimize } d^T u \textit{ s.t. } u \in G \subset \mathbf{R}^k,$$

G being a closed convex domain, can be equivalently rewritten in the conic form $(\mathcal{P})$. To this end it suffices to represent G as an intersection of a closed convex cone and an affine plane, which is immediate: identifying $\mathbf{R}^k$ with the affine hyperplane

$$\Gamma = \{x = (t, u) \in \mathbf{R}^{k+1} \mid t = 1\},$$

we can rewrite $(\mathcal{S})$ equivalently as

$$(\mathcal{S}_c) \quad \textit{minimize } c^T x \textit{ s.t. } x \in \Gamma \cap K,$$

where

$$c = \begin{pmatrix} 0 \\ d \end{pmatrix}$$

and

$$K = \mathrm{cl}\{(t, x) \mid t > 0, t^{-1}x \in G\}$$

is the *conic hull* of G. It is easily seen that $(\mathcal{S})$ is equivalent to $(\mathcal{S}_c)$ and that the latter problem is conic (i.e., K is a closed convex pointed cone with a nonempty interior), provided that the closed convex domain G does not contain lines (whih actually is not a restriction at all). Thus, $(\mathcal{P})$ indeed is a universal form of a convex program.

5.2 Conic duality

The similarity between conic problem $(\mathcal{P})$ and a Linear Programming problem becomes very clear when the duality issues are concerned. This duality, which is important for developing potential reduction methods and interesting in its own right, is our now subject.

5.2.1 Fenchel dual to $(\mathcal{P})$

We are about to derive the Fenchel dual of conic problem $(\mathcal{P})$, and let me start with recalling you what is the Fenchel duality.

> Given a *convex, proper, and closed* function f on $\mathbf{R}^n$ taking values in the extended real axis $\mathbf{R} \cup \{+\infty\}$ ("proper" means that the *domain* $\mathrm{dom} f$ of the function f, i.e., the set where

f is finite, is nonempty; "closed" means that the epigraph of the function is closed[1], one can define its *congugate* (the Legendre transformation)

$$f^*(s) = \sup_{x\in\mathbf{R}^n} \{s^T x - f(x)\} = \sup_{x\in \mathrm{dom} f} \{s^T x - f(x)\},$$

which again is a convex, proper and closed function; the conjugacy is an involution: $(f^*)^* = f$.

Now, let $f_1, ..., f_k$ be convex proper and closed functions on $\mathbf{R}^n$ such that the relative interiors of the domains of the functions (i.e., the interiors taken with respect to the affine hulls of the domains) have a point in common. The *Fenchel Duality theorem* says that if the function

$$f(x) = \sum_{i=1}^k f_i(x)$$

is below bounded, then

$$-\inf f = \min_{s_1,...,s_k: s_1+...+s_k=0} \{f_1^*(s_1) + ... + f_k^*(s_k)\} \tag{5.1}$$

(note this min in the right hand side: the theorem says, in particular, that it indeed is achieved). The problem

$$\textit{minimize} \sum_{i=1}^k f_i^*(s_i) \ \ \textit{s.t.} \ \ \sum_i s_i = 0$$

is called the *Fenchel dual* to the problem

$$\textit{minimize} \sum_i f_i(x).$$

Now let us derive the Fenchel dual to the conic problem $(\mathcal{P})$. To this end let us set

$$f_1(x) = c^T x; \quad f_2(x) = \begin{cases} 0, & x \in b + L \\ +\infty, & \textit{otherwise} \end{cases}; \quad f_3(x) = \begin{cases} 0, & x \in K \\ +\infty, & \textit{otherwise} \end{cases};$$

these functions clearly are convex, proper and closed, and $(\mathcal{P})$ evidently is nothing but the problem of minimizing $f_1 + f_2 + f_3$ over $\mathbf{R}^n$. To write down the Fenchel dual to the latter problem, we should realize what are the functions f_i^*, $i = 1, 2, 3$. This is immediate:

$$f_1^*(s) = \sup\{s^T x - c^T x \mid x \in \mathbf{R}^n\} = \begin{cases} 0, & s = c \\ +\infty & \textit{otherwise} \end{cases};$$

$$f_2^*(s) = \sup\{s^T x - 0 \mid x \in \mathrm{dom} f_2 \equiv b + L\} = \begin{cases} s^T b, & s \in L^\perp \\ +\infty, & \textit{otherwise} \end{cases},$$

where $L^\perp$ is the orthogonal complement to L;

$$f_3^*(s) = \sup\{s^T x - 0 \mid x \in \mathrm{dom} f_3 \equiv K\} = \begin{cases} 0, & s \in -K^* \\ +\infty, & \textit{otherwise} \end{cases},$$

where K^* is the cone dual to K.

Now, in the Fenchel dual to $(\mathcal{P})$, i.e., in the problem of minimizing $f_1^*(s_1) + f_2^*(s_2) + f_3^*(s_3)$ over s_1, s_2, s_3 subject to $s_1 + s_2 + s_3 = 0$, we clearly can restrict s_i to be in $\mathrm{dom} f_i^*$ without violating the optimal solution; thus, we may restrict ourselves to the case when $s_1 = c$, $s_2 \in L^\perp$ and $s_3 \in -K^*$, while $s_1 + s_2 + s_3 = 0$; under these restrictions the objective in the Fenchel dual is equal to $s_2^T b$. Expressing s_1, s_2, s_3 in terms of $s = s_1 + s_2 \equiv -s_3$, we come to the following equivalent reformulation of the Fenchel dual to $(\mathcal{P})$:

$$(\mathcal{D}) \quad \textit{minimize } b^T s \ \ \textit{s.t.} \ \ s \in \{c + L^\perp\} \cap K^*.$$

[1] equivalently: f is lower semicontinuous, or: the level sets $\{x \mid f(x) \le a\}$ are closed for every $a \in \mathbf{R}$

Note that the actual objective in the Fenchel dual is $s_2^T b \equiv s^T b + c^T b$; writing down $(\mathcal{D})$, we omit the constant term $c^T b$ (this does not influence the optimal set, although varies the optimal value). Problem $(\mathcal{D})$ is called the *conic dual* to the *primal* conic problem $(\mathcal{P})$.

Note that K is assumed to be closed convex and pointed cone with a nonempty interior; therefore the dual cone K^* also is closed, pointed, convex and with a nonempty interior, so that the dual problem also is conic. Bearing in mind that $(K^*)^* = K$, one can immediately verify that the indicated duality is completely symmetric: the problem dual to dual is exactly the primal one. Note also that in the Linear Programming case the conic dual is nothing but the usual dual problem written down in terms of slack variables.

5.2.2 Duality relations

Now let us establish several useful facts about conic duality; all of them are completely similar to what we know from LP duality.

0. *Let (x, s) be a primal-dual feasible pair, i.e., a pair comprised of feasible solutions to $(\mathcal{P})$ and $(\mathcal{D})$. Then*

$$c^T x + b^T s - c^T b = x^T s \geq 0.$$

The left hand side of the latter relation is called the *duality gap*; **0.** says that the duality gap is equal to $x^T s$ and always is nonnegative. The proof is immediate: since x is primal feasible, $x - b \in L$, and since s is dual feasible, $s - c \in L^\perp$, whence

$$(x - b)^T (s - c) = 0,$$

or, which is the same,

$$c^T x + b^T s - c^T b = x^T s;$$

the right hand side here is nonnegative, since $x \in K$ and $s \in K^*$.

I. *Let $\mathcal{P}^*$ and $\mathcal{D}^*$ be the optimal values in the primal and the dual problem, respectively (optimal value is $+\infty$, if the problem is unfeasible, and $-\infty$, if it is below unbounded). Then*

$$\mathcal{P}^* + \mathcal{D}^* \geq c^T b,$$

where, for finite a, $\pm\infty + a = \pm\infty$, the sum of two infinities of the same sign is the infinity of this sign and $(+\infty) + (-\infty) = +\infty$.

This is immediate: take infimums in primal feasible x and dual feasible s in the relation $c^T x + b^T s \geq c^T b$ (see **0.**).

II. *If the dual problem is feasible, then the primal is below bounded*[2]*; if the primal problem is feasible, then the dual is below bounded.*

This is an immediate corollary of **I.**: if, say, $\mathcal{D}^*$ is $< +\infty$, then $\mathcal{P}^* > -\infty$, otherwise $\mathcal{D}^* + \mathcal{P}^*$ would be $-\infty$, which is impossible in view of **I.**

III. Conic Duality Theorem. *If one of the problems in the primal-dual pair $(\mathcal{P})$, $(\mathcal{D})$ is strictly feasible (i.e., possesses feasible solutions from the interior of the corresponding cone) and is below bounded, then the second problem is solvable, the optimal values in the problems are finite and optimal duality gap $\mathcal{P}^* + \mathcal{D}^* - c^T b$ is zero.*

If both of the problems are strictly feasible, then both of them are solvable, and a pair (x^*, s^*) comprised of feasible solutions to the problems is comprised of optimal solutions if and only if the duality gap $c^T x^* + b^T s^* - c^T b$ is zero, and if and only if the complementary slackness $(x^*)^T s^* = 0$ holds.

Proof. Let us start with the first statement of the theorem. Due to primal-dual symmetry, we can restrict ourselves with the case when the strictly feasible below bounded problem is $(\mathcal{P})$. Strict feasibility means exactly that the relative interiors of the domains of the functions f_1, f_2, f_3 (see the derivation of $(\mathcal{D})$) have a point in common, due to the description of the domains of f_1 (the whole space), f_2 (the affine plane $b+L$), f_3 (the cone K). The below boundedness of $(\mathcal{P})$ means exactly that the function $f_1+f_2+f_3$ is below bounded. Thus, the situation is covered by the premise of the Fenchel duality theorem, and

[2] i.e., $\mathcal{P}^* > -\infty$; it may happen, anyhow, that $(\mathcal{P})$ is unfeasible

according to this theorem, the Fenchel dual to $(\mathcal{P})$, which can be obtained from $(\mathcal{D})$ by substracting the constant $c^T b$ from the objective, is solvable. Thus, $(\mathcal{D})$ is solvable, and the sum of optimal values in $(\mathcal{P})$ and $(\mathcal{D})$ (which is by $c^T b$ greater than the zero sum of optimal values stated in the Fenchel theorem) is $C^T b$, as claimed.

Now let us prove the second statement of the theorem. Under the premise of this statement both problems are strictly feasible; from **II.** we conclude that both of them are also below bounded. Applying the first statement of the theorem, we see that both of the problems are solvable and the sum of their optimal values is $c^T b$. It immediately follows that a primal-dual feasible pair (x, s) is comprised of primal-dual optimal solutions if and only if $c^T x + b^T s = c^T b$, i.e., if and only if the duality gap at the pair is 0; since the duality gap equals also to $x^T s$ (see **0.**), we conclude that the pair is comprised of optimal solutions if and only if $x^T s = 0$. ■

Remark 5.2.1 The Conic duality theorem, although very similar to the Duality theorem in LP, is a little bit weaker than the latter statement. In the LP case, already (feasibility + below boundedness), not (*strict* feasibility + below boundedness), of one of the problems implies solvability of both of them and characterization of the optimality identical to that one given by the second statement of the Conic duality theorem. A "word by word" extension of the LP Duality theorem fails to be true for general cones, which is quite natural: in the non-polyhedral case we need certain qualification of constrains, and strict feasibility is the simplest (and the strongest) form of this qualification. From the exercises accompanying the lecture you can find out what are the possibilities to strengthen the Conic duality theorem, on one hand, and what are the pathologies which may occur if the assumptions are weakened too much, on the other hand.

Let me conclude this part of the lecture by saying that the conic duality is, as we shall see, useful for developing potential reduction interior point methods. It also turned out to be powerful tool for analytical - on paper - processing a problem; in several interesting cases, as we shall see in the mean time, it allows to derive (completely mechanically!) nontrivial and informative reformulations of the initial setting.

5.3 Logarithmically homogeneous barriers

To develop potential reduction methods, we need deal with conic formulations of convex programs and should equip the corresponding cones with specific self-concordant barriers - the *logarithmically homogeneous* ones. This latter issue is our current goal.

Definition 5.3.1 *Let $K \subset \mathbf{R}^n$ be a a convex, closed and pointed cone with a nonempty interior, and let $\vartheta \geq 1$ be a real. A function $F : \mathrm{int}\ K \to \mathbf{R}$ is called ϑ-logarithmically homogeneous self-concordant barrier for K, if it is self-concordant on* int K *and satisfies the identity*

$$F(tx) = F(x) - \vartheta \ln t \quad \forall x \in \mathrm{int}\ K \quad \forall t > 0. \tag{5.2}$$

Our terminology at this point looks confusing: it is not clear whether a "logarithmically homogeneous self-concordant barrier" for a cone is a "self-concordant barrier" for it. This temporary difficulty is resolved by the following statement.

Proposition 5.3.1 *A ϑ-logarithmically homogeneous self-concordant barrier F for K is a nondegenerate ϑ-self-concordant barrier for K. Besides this, F satisfies the following identities* $(x \in \mathrm{int}\ K, t > 0)$*:*

$$F'(tx) = t^{-1} F'(x); \tag{5.3}$$

$$F'(x) = -F''(x)x; \tag{5.4}$$

$$\lambda^2(F, x) \equiv -x^T F'(x) \equiv x^T F''(x)x \equiv \vartheta. \tag{5.5}$$

Proof. Since, by assumption, K does not contain lines, F is nondegenerate (**II.**, Lecture 2). Now let us prove (5.3) - (5.5). Differentiating the identity

$$F(tx) = F(x) - \vartheta \ln t \tag{5.6}$$

in x, we come to (5.3); differentiating (5.3) in t and setting $t = 1$, we obtain (5.4). Differentiating (5.6) in t and setting $t = 1$, we come to

$$-x^T F'(x) = \vartheta.$$

Due to already proved (5.4), this relation implies all equalities in (5.5), excluding the very first of them; this latter follows from the fact that x, due to (5.4), is the Newton direction $-[F''(x)]^{-1}F'(x)$ of F at x, so that $\lambda^2(F, x) = -x^T F'(x)$ (**IVa.**, Lecture 2).

Form (5.5) it follows that the Newton decrement of F is identically equal to $\sqrt{\vartheta}$; since, by definition, F is self-concordant on int K, F is ϑ-self-concordant barrier for K. ■

Let us list some examples of self-concordant barriers.

Example 5.3.1 *The standard logarithmic barrier*

$$F(x) = -\sum_{i=1}^{n} \ln x_i$$

for the nonnegative orthant $\mathbf{R}^n_+$ *is n-logarithmically homogeneous self-concordant barrier for the orthant.*

Example 5.3.2 *The function*

$$F(x) = -\ln(t^2 - |x|_2^2)$$

is 2-logarithmically homogeneous self-concordant barrier for the ice-cream cone

$$K_n^2 = \{(t, x) \in \mathbf{R}^{n+1} \mid t \geq |x|_2\}.$$

Example 5.3.3 *The function*

$$F(x) = -\ln \mathrm{Det}\, x$$

is n-logarithmically self-concordant barrier for the cone $\mathbf{S}^n_+$ *of symmetric positive semidefinite $n \times n$ matrices.*

Indeed, self-concordance of the functions listed in the above examples is given, respectively, by Corollary 2.1.1, Exercise 3.3.4 and Exercise 3.3.3; logarithmic homogeneity is evident.

The logarithmically homogeneous self-concordant barriers admit combination rules completely similar to those for self-concordant barriers:

Proposition 5.3.2 (i) [stability with respect to linear substitutions of the argument] *Let F be ϑ-logerithmically homogeneous self-concordant barrier for cone $K \subset \mathbf{R}^n$, and let $x = Ay$ be a linear homogeneous mapping from $\mathbf{R}^k$ into $\mathbf{R}^n$, with matrix A being of the rank k, such that the image of the mapping intersects* int K*. Then the inverse image $K^+ = A^{-1}(K)$ of K under the mapping is convex pointed and closed cone with a nonempty interior in $\mathbf{R}^k$, and the function $F(Ay)$ is ϑ-logarithmically homogeneous self-concordant barrier for K^+.*

(ii) [stability with respect to summation] *Let F_i, $i = 1, ..., k$, be ϑ_i-logarithmically homogeneous self-concordant barriers for cones $K_i \subset \mathbf{R}^n$, and let $\alpha_i \geq 1$. Assume that the cone $K = \cap_{i=1}^k K_i$ possesses a nonempty interior; then the function $\sum_{i=1}^k \alpha_i F_i$ is $(\sum_i \alpha_i \vartheta_i)$-logarithmically homogeneous self-concordant barrier for K.*

(iii) [stability with respect to direct summation] *Let F_i, $i = 1, ..., k$, be ϑ_i-self-concordant barriers for cones $K_i \subset \mathbf{R}^{n_i}$. Then the direct sum*

$$F_1(x_1) + ... + F_k(x_k)$$

of the barriers is $(\sum_i \vartheta_i)$-logarithmically homogeneous self-concordant barrier for the direct product $K_1 \times ... \times K_k$ of the cones.

The proposition is an immediate corollary of Proposition 3.1.1 and Definition 5.3.1.

In what follows we heavily exploit the following property of logatrithmically homogeneous self-concordant barriers:

Proposition 5.3.3 *Let $K \subset \mathbf{R}^n$ be a convex pointed closed cone with a nonempty interior, and let F be a ϑ-logarithmically homogeneous self-concordant barrier for K. Then*

(i) *The domain* $\mathrm{Dom}\, F^*$ *of the Legendre transformation of F^* of the barrier F is exactly the interior of the cone $-K^*$ anti-dual to K and F^* is ϑ-logarithmically homogeneous self-concordant barrier for this anti-dual cone. In particular, the mapping*

$$x \mapsto F'(x) \qquad (5.7)$$

is a one-to-one mapping of $\mathrm{int}\ K$ *onto* $-\,\mathrm{int}\ K^*$ *with the inverse given by* $s \mapsto (F^*)'(s)$.

(ii) *For any* $x \in \mathrm{int}\ K$ *and* $s \in \mathrm{int}\ K^*$ *the following inequality holds:*

$$F(x) + F^*(-s) + \vartheta \ln(x^T s) \geq \vartheta \ln \vartheta - \vartheta. \qquad (5.8)$$

This inequality is equality if and only if

$$s = -tF'(x) \qquad (5.9)$$

for some positive t.

Proof.

1^0. From Proposition 5.3.2 we know that F is nondegenerate; therefore F^* is self-concordant on its domain Q, and the latter is nothing but the image of int K under the one-to-one mapping (5.7), the inverse to the mapping being $s \mapsto (F^*)'(s)$ (see Lecture 2, (L.1)-(L.3) and **VII.**). Further, from (5.3) it follows that Q is an (open) cone; indeed, any point $s \in Q$, due to already proved relation $Q = F'(\mathrm{int}\ K)$, can be represented as $F'(x)$ for some $x \in \mathrm{int}\ K$, and then $ts = F'(t^{-1}x)$ also belongs to Q. It follows that $K^+ = \mathrm{cl}\, Q$ is a closed convex cone with a nonempty interior.

2^0. Let us prove that $K^+ = -K^*$. This is exactly the same as to prove that the interior of $-K^*$ (which is comprised of s strictly negative on K, i.e., with $s^T x$ being negative for any nonzero $x \in K$, see Section 5.1) coincides with $Q \equiv F'(\mathrm{int}\ K)$:

$$F'(\mathrm{int}\ K) = -\,\mathrm{int}\ K^*. \qquad (5.10)$$

2^0.1. The inclusion

$$F'(\mathrm{int}\ K) \subset -\,\mathrm{int}\ K^* \qquad (5.11)$$

is immediate: indeed, we should verify that for any $x \in \mathrm{int}\ K$ $F'(x)$ is strictly negative on K, i.e., that $y^T F'(x)$ is negative whenever $y \in K$ is nonzero. This is readily given by Corollary 3.2.1: since K is a cone, $y \in K$ is a recessive direction for K, and, due to the Corollary,

$$-y^T F'(x) \equiv -DF(x)[y] \geq \{D^2 F(x)[y, y]\}^{1/2};$$

the concluding quantity here is strictly positive, since y is nonzero and F, as we already know, is nondegenerate.

2^0.2. To complete the proof of (5.10), we need to verify the inclusion inverse to (5.11), i.e., we should prove that if s is strictly negative on K, then $s = F'(x)$ for certain $x \in \mathrm{int}\ K$. Indeed, since s is strictly negative on K, the cross-section

$$K_s = \{y \in K \mid s^T y = -1\} \qquad (5.12)$$

is bounded (Section 5.1). The restirction of F onto the relative interior of this cross-section is a self-concordant function on $\mathrm{rint} K_s$ (stability of self-concordance with respect to affine substitutions of argument, Proposition 2.1.1.(i)). Since K_s is bounded, F attains its minimum on the relative interior of K_s at certain point y, so that

$$F'(y) = \lambda s$$

for some λ, The coefficient λ is positive (since $y^T F'(y) = \lambda y^T s$ is negative in view of (5.5) and $y^T s = -1$ also is negative (recall that $y \in K_s$). Since λ is positive and $F'(y) = \lambda s$, we conclude that $F'(\lambda^{-1} y) = s$ (5.3), and s indeed is $F'(x)$ for some $x \in \mathrm{int}\ K$ (namely, $x = \lambda^{-1} y$). The inclusion (5.10) is proved.

3^0. Summarising our considerations, we see that F^* is self-concordant on the interior of the cone $-K^*$; to complete the proof of (i), it suffices to verify that

$$F^*(ts) = F(s) - \vartheta \ln t.$$

This is immediate:

$$(F^*)(ts) = \sup_{x\in \text{int } K}\{ts^T x - F(x)\} = \sup_{y\equiv tx \in \text{int } K}\{s^T y - F(y/t)\} =$$

$$= \sup_{y\in\text{int } K}\{s^T y - [F(y) - \vartheta \ln(1/t)]\} = F^*(s) - \vartheta \ln t.$$

(i) is proved.

4^0. Let us prove (ii). First of all, for $x \in \text{int } K$ and $s = -tF'(x)$ we have

$$F(x) + F^*(-s) + \vartheta \ln(x^T s) = F(x) + F^*(tF'(x)) + \vartheta \ln(-tx^T F'(x)) =$$

[since F^* is ϑ-logarithmically homogeneous due to (i) and $-x^T F'(x) = \vartheta$, see (5.5)]

$$= F(x) + F^*(F'(x)) + \vartheta \ln \vartheta =$$

[since $F^*(F'(x)) = x^T F'(x) - F(x)$ due to the definition of the Legendre transformation]

$$= x^T F'(x) + \vartheta \ln \vartheta = \vartheta \ln \vartheta - \vartheta$$

(we have used (5.5)). Thus, (5.8) indeed is equality when $s = -tF'(x)$ with certain $t > 0$.

5^0. To complete the proof of (5.8), it suffices to demonstrate that if x and s are such that

$$V(x,s) = F(x) + F^*(-s) + \vartheta \ln(s^T x) \le \vartheta \ln \vartheta - \vartheta, \tag{5.13}$$

then s is proportional, with positive coefficient, to $-F'(x)$. To this end consider the cross-section of K as follows:

$$K_s = \{y \in K \mid s^T y = s^T x\}.$$

The restriction of $V(\cdot, s)$ onto the relative interior of K_s is, up to additive constant, equal to the restriction of F, i.e., it is self-concordant (since K_s is cross-section of K by an affine hyperplane passing through an interior point of K; we have used similar reasoning in 2^0.2). Since K_s is bounded (by virtue of $s \in \text{int } K^*$), F, and, consequently, $V(\cdot, s)$ attains its minimum on the relative interior of K_s, and this minimum is unique (since F is nondegenerate). At the minimizer, let it be y, one should have

$$F'(y) = -\lambda s;$$

taking here inner product with y and using (5.5) and the inclusion $y \in K_s$, we get $\lambda > 0$. As we alerady know, the relation $F'(y) - -\lambda s$ with positive λ implies that $V(y,s) = \vartheta \ln \vartheta - \vartheta$; now from (5.13) it follows that $V(y,s) \ge V(x,s)$. Since, by construction, $x \in \text{rint}K_s$ and y is the unique minimizer of $V(\cdot, s)$ on the latter set, we conclude that $x = y$, so that $F'(x) = -\lambda s$, and we are done. ■

5.4 Exercises: Conic problems

The list of below exercises is unusually large; you are kindly asked at least to look through the formulations.

5.4.1 Basic properties of cones

Those not familiar with some of the facts on convex cones used in the lecture (see Section 5.1), are recommended to solve the exercises from this subsection; in these exercises, $K \subset \mathbf{R}^n$ is a closed convex cone and K^* is its dual.

Exercise 5.4.1 $^{\#+}$ *Prove that K^* is closed cone and $(K^*)^* = K$.*

Exercise 5.4.2 $^{\#+}$ *Prove that K possesses a nonempty interior if and only if K^* is pointed, and that K^* possesses a nonempty interior if and only if K is pointed.*

Exercise 5.4.3 $^{\#+}$ *Let $s \in \mathbf{R}^n$. Prove that the following properties of s are equivalent:*
(i) s is strictly positive on K, i.e., $s^T x > 0$ whenever $x \in K$ is nonzero;
(ii) The set $K(s) = \{x \in K \mid s^T x \leq 1\}$ is bounded;
(iii) $s \in \operatorname{int} K^*$.
Formulate "symmetric" characterization of the interior of K.

5.4.2 More on conic duality

Here we list some duality relations for the primal-dual pair $(\mathcal{P})$, $(\mathcal{D})$ of conic problems (see Lecture 5). The goal is to realize to which extent the standard properties of LP duality preserve in the general case. The forthcoming exercises are *not* accompanied by solutions, although some of then are not so simple.

Given a conic problem, let it be called $(\mathcal{T})$, with the data Q (the cone), r (the objective), $d + M$ (the feasible plane; M is the corresponding linear subspace), denote by $D(\mathcal{T})$ the feasible set of the problem and consider the following properties:

- (F): Feasibility: $D(\mathcal{T}) \neq \emptyset$;
- (B): Boundedness of the feasible set ($D(\mathcal{T})$ is bounded, e.g., empty);
- (SB): Boundedness of the solution set (the set of optimal solutions to $(\mathcal{T})$ is nonempty and bounded);
- (BO): Boundedness of the objective (the objective is below bounded on $D(\mathcal{T})$, e.g., due to $D(\mathcal{T}) = \emptyset$);
- (I): Existence of a feasible interior point ($D(\mathcal{T})$ intersects $\operatorname{int} Q$);
- (S): Solvability ($(\mathcal{T})$ is solvable);
- (WN): Weak normality (both $(\mathcal{T})$ and its conic dual are feasible, and the sum of their optimal values equals to $r^T d$).
- (N): Normality (weak normality + solvability of both $(\mathcal{T})$ and its conic dual).

Considering a primal-dual pair of conic problems $(\mathcal{P})$, $(\mathcal{D})$, we mark by superscript p, d, that the property in question is shared by the primal, respectively, the dual problem of the pair; e.g., (S_d) is abbreviation for the property "the dual problem $(\mathcal{D})$ is solvable".

Good news about conic duality:

Exercise 5.4.4 *Prove the following implications:*
1) $(\mathrm{F}_p) \Rightarrow (\mathrm{BO}_d)$
"if primal is feasible, then the dual is below bounded"; this is **II.**, Lecture 5; this is exactly as in LP*;*
2) $[(\mathrm{F}_p)\ \&\ (\mathrm{B}_p)] \Rightarrow [(\mathrm{S}_p)\ \&\ (\mathrm{WN})]$

"if primal is feasible and its feasible set is bounded, then primal is solvable, dual is feasible and below bounded, and the sum of primal and dual optimal values equals to $c^T b$"; in LP one can add to the conclusion "the dual is solvable"*;*

3) $[(\mathrm{I}_p)\ \&\ (\mathrm{BO}_p)] \Rightarrow [(\mathrm{S}_d)\ \&\ (\mathrm{WN})]$
this is exactly the Conic duality theorem*;*

4) $(\mathrm{SB}_p) \Rightarrow (\mathrm{WN})$
"if primal is solvable and its optimal set is bounded, then dual is feasible and below bounded, and the sum of primal and dual optimal values equals to $c^T b$"; in LP one can omit "optimal set is bounded" in the premise and add "dual is solvable" to the conclusion.

Formulate the "symmetric" versions of these implications, by interchanging the primal and the dual problems.

Bad news about conic duality:

Exercise 5.4.5 *Demonstrate by examples, that the following situations (which for sure do not occur in LP duality) are possible:*

1) the primal problem is strictly feasible and below bounded, and at the same time it is unsolvable (cf. Exercise 5.4.4, 2));

2) the primal problem is solvable, and the dual is unfeasible (cf. Exercise 5.4.4, 2), 3), 4));

3) the primal problem is feasible with bounded feasible set, and the dual is unsolvable (cf. Exercise 5.4.4, 2), 3));

3) both the primal and the dual problems are solvable, but there is nonzero duality gap: the sum of optimal values in the problems is strictly greater than $c^T b$ (cf. Exercise 5.4.4, 2), 3)).

The next exercise is of some interest:

Exercise 5.4.6 * *Assume that both the primal and the dual problem are feasible. Prove that the feasible set of at least one of the problems is unbounded.*

5.4.3 Complementary slackness: what it means?

The Conic duality theorem says that if both the primal problem $(\mathcal{P})$ and the dual problem $(\mathcal{D})$, see Lecture 5, are strictly feasible, then both of them are solvable, and the pair (x, s) of feasible solutions to the problems is comprised of optimal solutions if and only if $x^T s = 0$. What does the latter relation actually mean, it depends on analytic structure of the underlying cone K. Let us look what happens in several specific cases which are responsible for a wide spectrum of applications.

Recall that in Lecture 5 we have mentioned three particular (families of) cones:

- the cone $\mathbf{R}^n_+$ - the n-dimensional nonnegative orthant in $\mathbf{R}^n$; the latter space from now on is equipped with the standard Euclidean structure given by the inner product $x^T y$;
- the cone $\mathbf{S}^n_+$ of positive semidefinite symmetric $n \times n$ matrices in the space $\mathbf{S}^n$ of symmetric $n \times n$ matrices; this latter space from now on is equipped with the Frobenius Euclidean structure given by the inner product $\langle x, y\rangle = \mathrm{Tr}\{xy\}$, Tr being the trace; this is nothing but the sum, over all entries, of the products of the corresponding entries in x and in y;
- the "ice-cream" (more scientific name - second-order) cone

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

  this is a cone in $\mathbf{R}^{n+1}$, and we already have said what is the Euclidean structure the space is equipped with.

Exercise 5.4.7 $^\#$ *Prove that each of the aforementioned cones is closed, pointed, convex and with a nonempty interior and, besides this, is self-dual, i.e., coincides with its dual cone*[3]*.*

[3] self-duality, of course, makes sense only with respect to certain Euclidean structure on the embedding linear space, since this structure underlies the construction of the dual cone. We have already indicated what are these structures for the spaces where our cones live

Now let us look what means complementary slackness in the case of our standard cones.

Exercise 5.4.8 $^{\#}$ *Let K be a cone, K^* be a dual cone and let x, s satisfy the complementary slackness relation*

$$\mathcal{S}(K): \quad \{x \in K\}\&\{s \in K^*\}\&\{x^T s = 0\}.$$

Prove that

1) in the case of $K = \mathbf{R}^n_+$ the relation $\mathcal{S}$ says exactly that x and s are nonnegative n-dimensional vectors with the zero dot product $x \times s = (x_1 s_1, ..., x_n s_n)^T$;

2)$^+$ in the case of $K = \mathbf{S}^n_+$ the relation $\mathcal{S}$ says exactly that x and s are positive semidefinite symmetric matrices with zero product xs; if it is the case, then x and s commutate and possess, therefore, a common eigenbasis, and the dot product of the diagonals of x and s in this basis is zero;

3)$^+$ in the case of $K = K^2_n$ the relation $\mathcal{S}$ says exactly that $x_{n+1} = \sqrt{x_1^2 + ... + x_n^2}$, $s_{n+1} = \sqrt{s_1^2 + ... + s_n^2}$ and

$$x_1 : s_1 = x_2 : s_2 = ... = x_n : s_n = -[x_{n+1} : s_{n+1}].$$

We have presented the "explicit characterization" of complementary slackness for our particular cones which often occur in applications, sometimes as they are, and sometimes - as certain "building blocks". I mean that there are *decomposable situations* where the cone in question is a direct product:

$$K = K_1 \times ... \times K_k,$$

and the Euclidean embedding space for K is the direct product of Euclidean embedding spaces for the "component cones" K_i. In such a situation the complementary slackness is "componentwise":

Exercise 5.4.9 $^{\#}$ *Prove that in the aforementioned decomposable situation*

$$K^* = K_1^* \times ... \times K_k^*,$$

and a pair $x = (x_1, ..., x_k)$, $s = (s_1, ..., s_k)$ possesses the complementary slackness property $\mathcal{S}(K)$ if and only if each of the pairs x_i, s_i possesses the property $\mathcal{S}(K_i)$, $i = 1, ..., k$.

Thus, if we are in a decomposable situation and the cones K_i belong each to its own of our three standard families, then we are able to interpret explicitly the complementary slackness relation.

Let me complete this section with certain useful observation related to the three families of cones in question. We know form Lecture 5 that these cones admit explicit logarithmically homogeneous self-concordant barriers; on the other hand, we know that the Legendre transformation of a logarithmically homogeneous self-concordant barrier for a cone is similar barrier for the anti-dual cone. It is interesting to look what are the Legendre transformations of the particular barriers known to us. The answer is as it should be: these barriers are, basically, "self-adjoint" - their Legendre transformations coincide with the barriers, up to negating the argument and adding a constant:

Exercise 5.4.10 $^{\#}$ *Prove that*

1) the Legendre transformation of the standard logarithmic barrier

$$F(x) = -\sum_{i=1}^n \ln x_i$$

for the cone $\mathbf{R}^n_+$ is

$$F^*(s) = F(-s) - n, \ \operatorname{Dom} F^* = -\mathbf{R}^n_+;$$

2) the Legendre transformation of the standard barrier

$$F(x) = -\ln \operatorname{Det} x$$

for the cone $\mathbf{S}^n_+$ is

$$F^*(s) = F(-s) - n, \ \operatorname{Dom} F^* = -\mathbf{S}^n_+;$$

3) the Legendre transformation of the standard barrier

$$F(x) = -\ln(x_{n+1}^2 - x_1^2 - ... - x_n^2)$$

for the cone K^2_n is

$$F^*(s) = F(-s) + 2\ln 2 - 2, \ \operatorname{Dom} F^* = -K^2_n.$$

5.4.4 Conic duality: equivalent form

In many applications the "natural" form of a conic problem is

$$(\mathrm{P}): \quad \textit{minimize } \chi^T\xi \;\; s.t. \;\; \xi \in \mathbf{R}^l,\, P(\xi - p) = 0,\; \mathcal{A}(\xi) \in K,$$

where ξ is the vector of design variables, P is given $k \times l$ matrix, p is given l-dimensional vector, $\chi \in \mathbf{R}^l$ is the objective,

$$\mathcal{A}(\xi) = A\xi + b$$

is an affine *embedding* of $\mathbf{R}^l$ into $\mathbf{R}^n$ and K is a convex, closed and pointed cone with a nonempty interior in $\mathbf{R}^n$. Since $\mathcal{A}$ is an embedding (different ξ's have different images), the objective can be expressed in terms of the image $x = \mathcal{A}(\xi)$ of the vector ξ under the embedding: there exists (not necessarily unique) $c \in \mathbf{R}^n$ such that

$$c^T\mathcal{A}(\xi) = c^T\mathcal{A}(0) + \chi^T\xi$$

identically in $\xi \in \mathbf{R}^l$.

It is clear that (P) is equivalent to the problem

$$(\mathrm{P}'): \quad \textit{minimize } c^T x \textit{ s.t. } x \in \{\beta + L\} \cap K,$$

where the affine plane $\beta + L$ is nothing but the image of the affine space

$$\{\xi \in \mathbf{R}^l \mid P(\xi - p) = 0\}$$

under the affine mapping $\mathcal{A}$. Problem (P') is a conic program in our "canonical" form, and we can write down the conic dual to it, let this dual be called (D). A useful thing (which saves a lot of time in computations with conic duality) is to know how to write down this dual *directly in terms of the data involved into (P)*, thus avoiding the necessity to compute c.

Exercise 5.4.11 $^{\#}$ *Prove that (D) is as follows:*

$$\textit{minimize } \; \beta^T s \textit{ s.t. } s \in K^*,\, A^T s = \chi + P^T r \textit{ for some } r \in \mathbf{R}^k, \tag{5.14}$$

with

$$\beta = b + Ap. \tag{5.15}$$

5.5 Exercises: Truss Topology Design via Conic duality

It was said in Lecture 5 that conic duality is a powerful tool for mathematical processing a convex problem. Let us illustrate this point by considering an interesting example - Truss Topology Design problem (TTD).

"Human" formulation. We should design a *truss* - a construction, like the Eifel Tower, comprised of thin *bars* linked with each other at certain points - *nodes* of the truss. The construction is subject to loads, i.e., external forces acting at the nodes. A particular collection of these forces - each element of the collection specifying the external force acting at the corresponding node - is called *loading scenario*. A given load causes certain deformation of the truss - nodes move a little bit, the bars become shorter or longer. As a result, the truss capacitates certain energy - the *compliance*. It is reasonable to regard this compliance as the measure of rigidity of the truss under the loading in question - the larger is the compliance, the less rigid is the construction. For a given loading scenario, the compliance depends on the truss - on how thick are the bars linking the nodes. Now, the rigidity of a truss with respect to a given set of loading scenarios is usually defined as its largest, over the scenarios, compliance. And the problem is to design, given the set of scenarios and restrictions on the total mass of the construction, the most rigid truss.

More specifically, when solving the problem you are given a finite 2D or 3D set of *tentative nodes*, same as the finite set of *tentative bars*; for each of these bars it is said at which node it should start and at which node end. To specify a truss is the same as to choose the *volumes* t_i, $i = 1, ..., m$, of the tentative bars (some of these volumes may be 0, which means that the corresponding bar in fact does not present in the truss); the sum V of these volumes (proportional to the total mass of the construction) is given in advance.

Mathematical formulation. Given are

- *loading scenarios* $f_1, ..., f_k$ - vectors from $\mathbf{R}^n$; here n is the total number of degrees of freedom of the nodes (i.e., the dimension of the space of virtual nodal displacements), and the entries of f are the components of the external forces acting at the nodes.

  n is something like twice (for 2D constructions) or 3 times (for 3D ones) the number of nodes; "something like", because some of the nodes may be partially or completely fixed (say, be in the fundament of the construction), which reduces the total # of freedom degrees;

- *bar-stiffness matrices* - $n \times n$ matrices A_i, $i = 1, ..., m$, where m is the number of tentative bars. The meaning of these matrices is as follows: for a truss with bar volumes t_i virtual displacement $x \in \mathbf{R}^n$ of the nodes result in reaction forces

  $$f = A(t)x, \quad A(t) = t_1 A_1 + ... + t_m A_m.$$

  Under reasonable mechanical hypothesis, these matrices are *symmetric positive semidefinite* with positive definite sum, and in fact even dyadic:

  $$A_i = b_i b_i^T$$

  for certain vectors $b_i \in \mathbf{R}^n$ (these vectors are defined by the geometry of the nodal set). These assumptions on A_i are crucial for what follows[4].

- *total bar volume* $V > 0$ of the truss.

Now, the vector x of nodal displacements caused by loading scenario f satisfies the *equilibrium equation*

$$A(t)x = f$$

(which says that the reaction forces $A(t)x$ caused by the deformation of the truss under the load should balance the load; if the equilibrium equation has no solution, that means that the truss is unable to carry the load in question). The compliance, up to an absolute constant factor, turns out to be

$$x^T f.$$

[4] crucial are positive semidefiniteness and symmetry of A_i, not the fact that they are dyadic; this latter assumption, quite reasonable for actual trusses, is not too important, although simplifies some relations

Thus, we come to the following problem of *Multi-Loaded Truss Topology Design*:

(TTD$_{\rm ini}$): *find vector $t \in \mathbf{R}^m$ of bar volumes satisfying the constraints*

$$t \geq 0; \quad \sum_{i=1}^{m} t_i = V \tag{5.16}$$

and the displacement vectors $x_j \in \mathbf{R}^n$, $j = 1, ..., k$, satisfying the equilibrium equations

$$A(t)x_j = f_j, \ \ j = 1, ..., k, \tag{5.17}$$

which minimize the worst-case compliance

$$C(t, x_1, ..., x_k) = \max_{j=1,...,k} x_j^T f_j.$$

From our initial formulation it is not even seen that the problem is convex (since equality constraints (5.17) are bilinear in t and x_j). It is, anyhow, easy to demonstrate that in fact the problem is convex. The motivation of the reasoning is as follows: when t is strictly positive, $A(t)$ is positive definite (since A_i are positive semidefinite with positive definite sum), and the equilibrium equations can be solved explicitly:

$$x_j = A^{-1}(t)f_j,$$

so that j-th compliance, as a function of $t > 0$, is

$$c_j(t) = f_j^T A^{-1}(t) f_j.$$

This function is convex in $t > 0$, since the interior of its epigraph

$$G_j = \{(\tau, t) \mid t > 0, \, \tau > f_j^T A^{-1}(t) f_j\}$$

is convex, due to the following useful observation:

(*): *a block-diagonal symmetric matrix* $\begin{pmatrix} \tau & f^T \\ f & A \end{pmatrix}$ *(τ and A are $l \times l$ and $n \times n$ symmetric matrices, f is $n \times l$ matrix) is positive definite if and only if both the matrices A and $\tau - f^T A^{-1} f$ are positive definite.*
The convexity of G_j is an immediate consequence of this observation, since, due to it (applied with $l = 1$ and $f = f_j$) G_j is the intersection of the convex set $\{(\tau, t) \mid t > 0\}$ and the inverse image of a convex set (the cone of positive definite $(n+1) \times (n+1)$ matrices) under the affine mapping

$$(\tau, t) \mapsto \begin{pmatrix} \tau & f_j^T \\ f_j & A(t) \end{pmatrix}$$

Exercise 5.5.1 $^{\#}$ *Prove (*).*

The outlined reasoning is unsufficient for our purposes: it does not say what happens if some of t_i's are zero, which may cause degeneracy of $A(t)$. In fact, of course, nothing happens: the epigraph of the function "compliance with respect to j-th load", regarded as a function of $t \geq 0$, is simply the closure of the above G_j (and is therefore convex). Instead of proving this latter fact directly, we shall come to the same conclusion in another way.

Exercise 5.5.2 *Prove that the linear equation*

$$Ax = f$$

with symmetric positive semidefinite matrix A is solvable if and only if the concave quadratic form

$$q_f(z) = 2z^T f - z^T Az$$

is above bounded, and if this is the case, then the quantity $x^T f$, x being an arbitrary solution to the equation, coincides with $\max_z q_f(z)$.

*Derive from this observation that one can eliminate from (*TTDini*) the displacements x_j by passing to the problem*

(TTD$_1$): *find vector t of bar volumes subject to the constraint (5.16) which minimizes the objective*

$$c(t) = \max_{j=1,\dots,k} c_j(t), \;\; c_j(t) = \sup_{z\in\mathbf{R}^n} [2z^T f_j - z^T A(t) z].$$

Note that $c_j(t)$ are closed and proper convex functions (as upper bounds of linear forms; the fact that the functions are proper is an immediate consequence of the fact that $A(t)$ is positive definite for strictly positive t), so that (TTD$_1$) is a convex program.

Our next step will be to reduce (TTD$_1$) to a conic form. Let us first make the objective linear. This is immediate: by introducing an extra variable τ, we can rewrite (TTD$_1$) equivalently as

(TTD$_2$): *minimize τ by choice of $t \in \mathbf{R}^n$ and τ subject to the constraints (5.16) and*

$$\tau + z^T A(t) z - 2z^T f_j \geq 0, \;\; \forall z \in \mathbf{R}^n \;\; \forall j = 1, ..., k. \tag{5.18}$$

((5.18) clearly express the inequalities $\tau \geq c_j(t)$, $j = 1, ..., k$).

Our next step is guided by the following evident observation:

the inequality

$$\tau + z^T A z - 2z^T f,$$

τ being real, A being symmetric $n \times n$ matrix and f being a n-dimensional vector, is valid for all $z \in \mathbf{R}^n$ if and only if the symmetric $(n+1) \times (n+1)$ matrix

$$\begin{pmatrix} \tau & f^T \\ f & A \end{pmatrix} \geq 0$$

is positive semidefinite.

Exercise 5.5.3 *Prove the latter statement. Derive from this statement that* (TTD$_2$) *can be equivalently written down as*

(TTD$_p$): *minimize τ by choice of $s \in \mathbf{R}^m$ and $\tau \in \mathbf{R}$ subject to the constraint*

$$\mathcal{A}(\tau, s) \in K; \quad \sum_{i=1}^m s_i = 0,$$

where

- *K is the direct product of $\mathbf{R}^m_+$ and k copies of the cone $\mathbf{S}^{n+1}_+$;*
- *the affine mapping $\mathcal{A}(\tau, s)$ is as follows: the $\mathbf{R}^n_+$-component of $\mathcal{A}$ is*

  $$\mathcal{A}_t(\tau, s) = s + m^{-1}(V, ..., V)^T \equiv s + e;$$

  the component of $\mathcal{A}$ associated with j-th of the copies of the cone $\mathbf{S}^{n+1}_+$ is

  $$\mathcal{A}_j(\tau, s) = \begin{pmatrix} \tau & f_j^T \\ f_j & A(e) + A(s) \end{pmatrix}.$$

Note that $\mathcal{A}_t(\tau, s)$ is nothing but our previous t; the constraint $\mathcal{A}_t(\tau, s) \in \mathbf{R}^m_+$ (which is the part of the constraint $\mathcal{A}(\tau, s) \in K$) together with the constraint $\sum_i s_i = 0$ give equivalent reformulation of the constraint (5.16), while the remaining components of the constraint $\mathcal{A}(\tau, s) \in K$, i.e., the inclusions $\mathcal{A}_j(\tau, s) \in \mathbf{S}^{n+1}_+$, represent the constraints (5.18).

Note that the problem (TTD$_p$) is in fact in the conic form (cf. Section 5.4.4). Indeed, it requires to minimize a *linear* objective under the constraints that, first, the design vector (τ, s) belongs to sertain *linear* subspace E (given by $\sum_i s_i = 0$) and, second, that the image of the design vector under a given *affine* mapping belongs to certain cone (closed, pointed, convex and with a nonempty interior). Now, the objective evidently can be respresented as a linear form $c^T u$ of the image $u = \mathcal{A}(\tau, s)$ of the design vector under the mapping, so that our problem is exactly in minimizing a linear objective over the intersection of an affine plane (namely, the image of the linear subspace E under the affine mapping $\mathcal{A}$) and a given cone, which is a conic problem.

To the moment we acted in certain "clever" way; from now on we act in completely "mechanical" manner, simply writing down and straightforwardly simplifying the conic dual to (TTD$_p$).

First step: writing down conic dual to (TTD$_p$). What we should do is to apply to (TTD$_p$) the general construction from Lecture 5 and look at the result. The data in the primal problem are as follows:

- K is the direct product of $K_t = \mathbf{R}^m_+$ and k copies K_j of the cone $\mathbf{S}^{n+1}_+$; the embedding space for this cone is
$$\mathcal{E} = \mathbf{R}^n \times \mathbf{S}^{n+1} \times ... \times \mathbf{S}^{n+1};$$
we denote a point from this latter space by $u = (t, p_1, ..., p_k)$, $t \in \mathbf{R}^m$ and p_j being $(n+1) \times (n+1)$ symmetric matrices, and denote the inner product by $(\cdot, \cdot)$;

- $c \in \mathcal{E}$ is given by $c = (0, \chi, ...\chi)$, where
$$\chi = \begin{pmatrix} k^{-1} & 0 \\ 0 & 0 \end{pmatrix}$$
is $(n+1) \times (n+1)$ matrix with the only nonzero entry, which ensures the desired relation
$$(c, \mathcal{A}(\tau, s)) \equiv \tau;$$
note that there are many other ways to choose c in accordance with this relation;

- L is the image of E under the homogeneous part of the affine mapping $\mathcal{A}$;

- $b = \mathcal{A}(0, 0) = (e, \phi_1, ..., \phi_k)$, where
$$\phi_j = \begin{pmatrix} 0 & f_j^T \\ f_j & A(e) \end{pmatrix}.$$

Now let us build up the dual problem. We know that the cone K is self-dual (as a direct product of self-dual cones, see Exercises 5.4.7, 5.4.9), so that $K^* = K$. We should realize only what is $L^\perp$, in other words, what are the vectors
$$s = (r, q_1, ..., q_k) \in \mathcal{E}$$
which are orthogonal to the image of E under the homogeneous part of the affine mapping $\mathcal{A}$. This requires nothing but completely straightforward computations.

Exercise 5.5.4 $^+$ *Prove that feasible plane $c + L^\perp$ of the dual problem is comprised of exactly those $w = (r, q_1, ..., q_k)$ for which the symmetric $(n+1) \times (n+1)$ matrices q_j, $j = 1, ..., k$, are of the form*
$$q_j = \begin{pmatrix} \lambda_j & z_j^T \\ z_j & \sigma_j \end{pmatrix}, \tag{5.19}$$
with λ_j satisfying the relation
$$\sum_{j=1}^k \lambda_j = 1 \tag{5.20}$$
and the $n \times n$ symmetric matrices $\sigma_1, ..., \sigma_k$, along with the n-dimensional vector r, and a real ρ, satisfying the equations
$$r_i + \sum_{j=1}^k b_i^T \sigma_j b_i = \rho, \; i = 1, ..., m. \tag{5.21}$$

(b_i are the vectors involved into the representation $A_i = b_i b_i^T$, so that $b_i^T \sigma_j b_i = \mathrm{Tr}\{A_i \sigma_j\}$).

Derive from this observation that the conic dual to (TTD$_p$) *is the problem*

(TTD$_d$): *minimize the linear functional*

$$2\sum_{j=1}^{k} z_j^T f_j + V\rho \tag{5.22}$$

by choice of positive semidefinite matrices q_j of the form (5.19), nonnegative vector $r \in \mathbf{R}^n$ and real ρ under the constraints (5.20) and (5.21).

Second step: simplifying the dual problem. Now let us simplify the dual problem. It is immediately seen that one can eliminate the "heavy" matrix variables σ_j and the vector r by performing partial optimization in these variables:

Exercise 5.5.5 $^+$ *Prove that in the notation given by (5.19), a collection*

$$(\lambda_1, ..., \lambda_k; z_1, ..., z_k; \rho)$$

can be extended to a feasible plan $(r; q_1, ..., q_k; \rho)$ of problem (TTD$_d$) *if and only if the collection satisfies the following requirements:*

$$\lambda_j \geq 0; \ \sum_{j=1}^{k} \lambda_j = 1; \tag{5.23}$$

$$\rho \geq \sum_{j=1}^{k} \frac{(b_i^T z_j)^2}{\lambda_j} \quad \forall i \tag{5.24}$$

(a fraction with zero denominator from now on is $+\infty$), so that (TTD$_d$) *is equivalent to the problem of minimizing linear objective (5.22) of the variables $\lambda_\cdot, z_\cdot, \rho$ under constraints (5.23), (5.24).*

Eliminate ρ from this latter problem to obtain the following equivalent reformulation of (TTD$_d$)*:*

(TTD^d): *minimize the function*

$$\max_{i=1,...,m} \sum_{j=1}^{k} \left[2 z_j^T f_j + V\frac{(b_i^T z_j)^2}{\lambda_j}\right] \tag{5.25}$$

by choice of λ_j, $j = 1, ..., k$, and $z_j \in \mathbf{R}^n$ subject to the constraint

$$\lambda_j \geq 0; \ \sum_{j=1}^{k} \lambda_j = 1. \tag{5.26}$$

Note that in the important *single-load case* $k = 1$ the problem (TTD^d) is simply in minimizing, with respect to $z_1 \in \mathbf{R}^n$, the maximum over $i = 1, ..., m$ of the quadratic forms

$$\psi_i(z_1) = 2z_1^T f_1 + V(b_i^T z_1)^2.$$

Now look: the initial problem (TTD$_p$) contained m-dimensional design vector (τ, s) (the "formal" dimension of the vector is $m + 1$, but we remember that the sum of s_i should be 0). The dual problem (TTD^d) has $k(n + 1) - 1$ variables (there are k n-dimensional vectors z_j and k reals λ_j subject to a single linear equation). In the "full topology TTD" (it is allowed to link by a bar any pair of nodes), m is of order of n^2 and n is at least of order of hundreds, so that m is of order of thousands and tens of thousands. In contrast to these huge numbers, the number k of loading scenarios is, normally, a small integer (less than 10). Thus, the dimension of (TTD^d) is by order of magnitudes less than that one of (TTD$_p$). At the same time, solving the dual problem one can easily recover, via the Conic duality theorem, the optimal solution to the primal problem. As a kind of "penalty" for relatively small # of variables, (TTD^d) has a lot of inequality constraints; note, anyhow, that for many methods it is much easier to struggle with many constraints than with many variables; this is, in particular, the case with the Newton-based methods[5]. Thus, passing - *in a completely mechanical way!* - from the primal problem to the dual one, we improve the "computational tractability" of the problem.

Third step: back to primal. And now let us demonstrate how duality allows to obtain a better insight on the problem. To this end let us derive the problem *dual to* (TTD^d). This looks crazy: we know that dual to dual is primal, the problem we started with. There is, anyhow, an important point: (TTD^d) is *equivalent* to the conic dual to (TTD$_p$), not the *conic dual* itself; therefore, taking dual to (TTD^d), we should not necessarily obtain the primal problem, although we may expect that the result will be equivalent to this primal problem.

Let us implement our plan. First, we rewrite (TTD^d) in an equivalent conic form. To this end we introduce extra variables $y_{ij} \in \mathbf{R}$, $i = 1, ..., m$, $j = 1, ..., k$, in order to "localize" nonlinearities, and an extra variable f to represent the objective (5.25) (look: a minute ago we tried to eliminate as many variables as possible, and now we go in the opposite direction... This is life, isn't it?) More specifically, consider the system of constraints on the variables z_j, λ_j, y_{ij}, f (i runs from 1 to m, j runs from 1 to k):

$$y_{ij} \geq \frac{(b_i^T z_j)^2}{\lambda_j};\ \lambda_j \geq 0,\ \ i = 1, ..., m,\ j = 1, ..., k; \tag{5.27}$$

$$f \geq \sum_{j=1}^{k} \left[2z_j^T f_j + V y_{ij} \right],\ i = 1, ..., m; \tag{5.28}$$

$$\sum_{j=1}^{k} \lambda_j = 1. \tag{5.29}$$

It is immediately seen that (TTD^d) is equivalent to minimization of the variable f under the constraints (5.27) - (5.29). This latter problem is in the conic form (P) of Section 5.4.4, since (5.27) can be equivalently rewritten as

$$\begin{pmatrix} y_{ij} & b_i^T z_j \\ b_i^T z_j & \lambda_j \end{pmatrix} \geq 0,\ \ i = 1, ..., m,\ j = 1, ..., k \tag{5.30}$$

("≥ 0" for symmetric matrices stands for "positive semidefinite"); to justify this equivalence, think what is the criterion of positive semidefiniteness of a 2×2 symmetric matrix.

We see that (TTD^d) is equivalent to the problem of minimizing f under the constraints (5.28) - (5.30). This problem, let it be called (π), is of form (P), Section 5.4.4, with the following data:

- the design vector is
$$\xi = (f; \lambda.; y.; z.);$$
- K is the direct product of $\mathbf{R}^m_+$ and mk copies of the cone $\mathbf{S}^2_+$ of symmetric positive semidefinite 2×2 matrices; we denote the embedding space of the cone by $\mathcal{F}$, the vectors from $\mathcal{F}$ by $\eta = (\zeta, \{\pi_{ij}\}_{i=1,...,m, j=1,...,k})$, ζ being m-dimensional and π_{ij} being 2×2 matrices, and equip $\mathcal{F}$ with the natural inner product
$$(\eta', \eta'') = (\zeta')^T \zeta'' + \sum_{i,j} \mathrm{Tr}\{\pi'_{ij} \pi''_{ij}\};$$
- $\mathcal{A}$ is the homogeneous linear mapping with the components
$$(\mathcal{A}_\zeta)_i = f - \sum_{j=1}^{k} \left[2z_j^T f_j + V y_{ij} \right],$$
$$\mathcal{A}_{\pi_{ij}} = \begin{pmatrix} y_{ij} & b_i^T z_j \\ b_i^T z_j & \lambda_j \end{pmatrix};$$
- χ is the vector with the only nonzero component (associated with the f-component of the design vector) equal to 1.

[5] since the number of constraints influences only the complexity of assembling the Newton system, and the complexity is *linear* in this number; in contrast to this, the # of variables defines the size of the Newton system, and the complexity of solving the system is *cubic* in # of variables

- The system $P(\xi - p) = 0$ is $\sum_j \lambda_j = 1$, so that $P^T r$, $r \in \mathbf{R}$, is the vector with λ.-components equal to r and remaining components equal to 0, and p is $P^T \frac{1}{k}$.

Exercise 5.5.6 $^+$ *Prove that the conic dual, in the sense of Section 5.4.4, to problem (π) is equivalent to the following program:*

(ψ): *minimize*

$$\max_{j=1,...,k} \left[\sum_{i=1}^{m} \frac{\beta_{ij}^2}{\phi_i} \right] \tag{5.31}$$

by choice of m-dimensional vector ϕ and mk reals β_{ij} subject to the constraints

$$\phi \geq 0; \ \sum_{i=1}^{m} \phi_i = V; \tag{5.32}$$

$$\sum_{i=1}^{m} \beta_{ij} b_i = f_j, \ j = 1, ..., k. \tag{5.33}$$

Fourth step: from primal to primal. We do not know what is the actual relation between problem (ψ) and our very first problem (TTD$_{\rm ini}$) - what we can say is:

"(ψ) is equivalent to the problem which is conic dual to the problem which is equivalent to the conic dual to the problem which is equivalent to (TTD$_{\rm ini}$)";

it sounds awkful, especially taking into account that the notion of equivalency between problems has no exact meaning. At the same time, looking at (ψ), namely, at equation (5.32), we may guess that ϕ_i are nothing but our bar volumes t_i - the design variables we actually are interested in, so that (ψ) is a "direct reformulation" of (TTD$_{\rm ini}$) - the ϕ-component of optimal solution to (ψ) is nothing but the t-component of the optimal solution to (TTD$_{\rm ini}$). This actually is the case, and the proof could be given by tracing the chain wich leaded us to (ψ). There is, anyhow, a direct, simple and instructive way to establish equivalency between the initial and the final problems in our chain, which is as follows.

Given a feasible solution $(t, x_1, ..., x_k)$ to (TTD$_{\rm ini}$), consider the *bar forces*

$$\beta_{ij} = t_i x_j^T b_i;$$

these quantities are magnitudes of the reaction forces caused by elongations of the bars under the corresponding loads. The equilibrium equations

$$A(t) x_j = f_j$$

in view of $A(t) = \sum_i t_i A_i \equiv \sum_i t_i b_i b_i^T$ say exactly that

$$\sum_i \beta_{ij} b_i = f_j, \ \ j = 1, ..., k; \tag{5.34}$$

thus, we come to a feasible plan

$$(\phi, \beta.) : \ \phi = t, \ \beta_{ij} = t_i x_j^T b_i \tag{5.35}$$

to problem (ψ). What is the value of the objective of the latter problem at the indicated plan? Multiplying (5.34) by x_j^T and taking into account the origin of β_{ij}, we see that j-th compliance $c_j = x_j^T f_j$ is equal to

$$\sum_i \beta_{ij} x_j^T b_i = \sum_i t_i (x_j^T b_i)^2 = \sum_i \frac{\beta_{ij}^2}{t_i} = \sum_i \frac{\beta_{ij}^2}{\phi_i},$$

so that the value of the objective of (TTD$_{\rm ini}$) at $(t, x_1, ..., x_k)$, which is $\max_j c_j$, is exactly the value of the objective (5.31) of the problem (ψ) at the feasible plan (5.35) of the latter problem. Thus, we have establish the following proposition:

A. *Transformation (5.35) maps a feasible plan $(t, x_1, ..., x_k)$ to problem (TTS$_{\rm ini}$) into feasible plan $(\phi, \beta.)$ to problem (ψ), and the value of the objective of the first problem at the first plan is equal to the value of the objective of the second problem at the second plan.*

Are we done? Have we established the desired equivalence between the problems? No! Why do we know that images of the feasible plans to (TTD$_{\rm ini}$) under mapping (5.35) cover *the whole* set of feasible plans of (ψ)? And if it is not the case, how can we be sure that the problems are equivalent - it may happen that optimal solution to (ψ) corresponds to *no* feasible plan of the initial problem!

And the image of mapping (5.35) indeed does not cover the whole feasible set of (ψ), which is clear by dimension reasons: the dimension of the feasible domain of (TTD$_{\rm ini}$), regarded as a nonlinear manifold, is $m-1$ (this is the # of independent t_i's; x_j are functions of t given by the equilibrium equations); and the dimension of the feasible domain of (ψ), also regarded as a manifold, is $m-1$ (# of independent ϕ_i's) plus mk (# of β_{ij}) minus nk (# of scalar linear equations (5.33)), i.e., it might be *by order of magnitudes* greater than the dimension of the feasible domain of (TTD$_{\rm ini}$) (recall that normally $m >> n$). In other words, transformation (5.35) allows to obtain only those feasible plans of (ψ) where the β-part is determined, via the expressions

$$\beta_{ij} = t_i x_j^T b_i,$$

by k n-dimensional vectors x_j (which is also clear from the origin of the problem: the actual bar forces should be caused by certain displacements of the nodes), and this is in no sense a consequence of the constraints of problem (ψ): relations (5.33) say only that the sum of the reaction forces balances the external load, and says nothing on the "mehcanical validity" of the reaction forces, i.e., whether or not they are caused by certain displacements of the nodes. Our dimension analysis demonstrates that the reaction forces caused by nodal displacements - i.e., those valid mechanically - form a very small part of all reaction forces allowed by equations (5.33).

In spite of these pessimistic remarks, we *know* that the optimal value in (ψ) - which is basically dual to dual to (TTD$_{\rm ini}$) - is the same one as that one in (TTD$_{\rm ini}$), so that in fact the optimal solution to (ψ) is in the image of mapping (5.35). Can we see it directly, without referring to the chain of transformations which leaded us to (ψ)? Yes! It is very simple to verify that the following proposition holds:

B. *Let $(\phi, \beta.)$ be a feasible plan to (ψ) and ω be the corresponding value of the objective. Then ϕ can be extended to a feasible plan $(t = \phi, x_1, ..., x_k)$ to* (TTD$_{\rm ini}$)*, and the maximal, over the loads $f_1, ..., f_k$, compliance of the truss t is $\leq \omega$.*

Exercise 5.5.7 $^+$ *Prove* **B.**

From **A.** and **B.** it follows, of course, that problems (TTD$_{\rm ini}$) and (ψ) are equivalent - ε-solution to any of them can be immediately transformed into ε-solution to another.

Concluding remarks. Let me make several comments on our "truss adventure".

- Our main effort was to pass from the initial form (TTD$_{\rm ini}$) of the Truss Topology Design problem to its dual (TTD^d) and then - to the "dual to dual" *bar forces* reformulation (ψ) of the initial problem. Some steps seemed to be "clever" (convex reformulation of (TTD$_{\rm ini}$); conic reformulation of of (TTD^d) in terms of cones of positive semidefinite 2×2 matrices), but most of them were completely routine - we used in a straightforward manner the general scheme of conic duality. In fact the "clever" steps also are completely routine; small experience suffices to see immediately that the epigraph of the compliance can be represented in terms of nonnegativity of certain quadratic forms or, which is the same, in terms of positive semidefiniteness of certain matrices linearly depending on the control vectors; this is even easier to do with the constraints (5.30). I would qualify our chain of reformulations as a completely straightforward.

- Let us look, anyhow, what are the results of our effort. There are two of them:

  (a) "compressed", as far as # of variables is concerned, form (TTD^d) of the problem; as it was mentioned, reducing # of variables, we get better possibilities for numerical processing the problem;

  (b) very instructive "bar forces" reformulation (ψ) of the problem.

- After the "bar forces" formulation is guessed, one can easily establish its equivalence to the initial formulation; thus, if our only goal were to replace (TTD$_{\rm ini}$) by (ψ), we could restrict ourselves with the fourth step of our construction and skip the preceding three steps. The question, anyhow, *how to guess* that (ψ) indeed is equivalent to (TTD$_{\rm ini}$). This is not that difficult to look what are the equilibrium equations in terms of the bar forces $\beta_{ij} = t_i x_j^T b_i$; but one hardly could be courageous enough (and, to the best of our knowledge, in fact was not courageous) to conjecture that the "heart of the situation" - the restriction that the bar forces should be caused by certain displacements of the nodes - simply is redundant: in fact we can forget that the bar forces should belong to an "almost negligible", as far as dimensions are concerned, manifold (given by the equations $\beta_{ij} = t_i x_j^T b_i$), since this restriction on the bar forces is *automatically satisfied* at any optimal solution to (ψ) (this is what actually is said by **B.**).

Thus, the things are as they should be: routine transformations result in something which, in principle, could be guessed and proved directly and quickly; the bottleneck is in this "in principle": it is not difficult to justify the answer, it is difficult to guess what the answer is. In our case, this answer was "guessed" via straightforward applications of a quite routine general scheme, scheme useful in other cases as well; to demonstrate the efficiency of this scheme and some "standard" tricks in its implementation, this is exactly the goal of this text.

- To conclude, let me say several words on the "bar forces" formulation of the TTD problem. First of all, let us look what is this formulation in the single-load case $k = 1$. Here the problem becomes

  $$\textit{minimize} \quad \sum_i \frac{\beta_i^2}{\phi_i}$$

  under the constraints

  $$\phi \geq 0; \ \sum_i \phi_i = V; \sum_i \beta_i b_i = f.$$

  We can immediately perform partial optimization in ϕ_i:

  $$\phi_i = V|\beta|_i \left[\sum_i |\beta_i|\right]^{-1}.$$

  The remaining optimization in β_i, i.e., the problem

  $$\textit{minimize} \quad V^{-1}\left[\sum_i |\beta_i|\right]^2 \quad \textit{s.t.} \quad \sum_i \beta_i b_i = f,$$

  can be immediately reduced to an LP program.

  Another useful observation is as follows: above we dealt with $A(t) = \sum_{i=1}^m t_i A_i, \ A_i = b_i b_i^T$; in mechanical terms, this is the linear elastic model of the material. For other mechanical models, other types of dependencies $A(t)$ occur, e.g.,

  $$A(t) = \sum_{i=1}^m t_i^\kappa A_i, \ A_i = b_i b_i^T,$$

  where $\kappa > 0$ is given. In this case the "direct" reasoning establishing the equivalence between $(\mathrm{TTD_{ini}})$ and (ψ) remains valid and results in the following "bar forces" setting:

  $$\textit{minimize} \quad \max_{j=1,\dots,k} \sum_{i=1}^m \frac{\beta_{ij}^2}{t_i^\kappa}$$

  under the constraints

  $$t \geq 0; \ \sum_i t_i = V; \ \sum_i \beta_{ij} b_i = f_j, \ j = 1, ..., k.$$

  A bad news here is that the problem turns out to be convex in $(t, \beta.)$ if and only if $\kappa \geq 1$, and from the mechanical viewpoint, the only interesting case in this range of values of κ is that one of linear model ($\kappa = 1$).

Chapter 6

The method of Karmarkar

The goal of this lecture is to develop the method which extends onto the general convex case the very first polynomial time interior point method - the method of Karmarkar. Let me say that there is no necessity to start with the initial LP method and then pass to the extensions, since the general scheme seems to be more clear than its particular LP implementation.

6.1 Problem setting and assumptions

The method in question is for solving a convex program in the conic form:

$$(\mathcal{P}): \quad \textit{minimize} \;\; c^T x \;\; \textit{s.t.} \;\; x \in \{b + L\} \cap K, \tag{6.1}$$

where

- K is a closed convex pointed cone with a nonempty interior in $\mathbf{R}^n$;
- L is a linear subspace in $\mathbf{R}^n$;
- b and c are given n-dimensional vectors.

We assume that

A: *the feasible set*

$$K_f = \{b + L\} \cap K$$

of the problem is bounded and intersects the interior of the cone K.

B: *we are given in advance a strictly feasible solution $\widehat{x}$ to the problem, i.e., a feasible solution belonging to the interior of K;*

Assumptions **A** and **B** are more or less standard for the interior point approach. The next assumption is specific for the method of Karmarkar:

C: *the optimal value, c^*, of the problem is known.*

Assumption **C.** might look rather restrictive; in the mean time we shall see how one can eliminate it.

Our last assumption is as follows:

D: *we are given a ϑ-logarithmically homogeneous self-concordant barrier F for the cone K.*

As in the case of the path-following method, "we are given F" means that we are able, for any $x \in \mathbf{R}^n$, to decide whether $x \in \operatorname{Dom} F \equiv \operatorname{int} K$, and if it is the case, can compute the value $F(x)$, the gradient $F'(x)$ and the Hessian $F''(x)$ of the barrier at x. Note that the barrier F is the only representation of the cone used in the method.

6.2 Homogeneous form of the problem

To proceed, let us note that the feasible affine plane $b + L$ of problem $(\mathcal{P})$ can be, by many ways, represented as an intersection of a *linear space* M and an affine hyperplane $\Pi = \{x \in \mathbf{R}^n \mid e^T x = 1\}$. Indeed, our feasible affine plane always can be represented as the plane of solutions to a system

$$Px = p$$

of, say, $m + 1$ linear equations. Note that the system for sure is not homogeneous, since otherwise the feasible plane would pass through the origin; and since, in view of **A**, it intersects also the interior of the cone, the feasible set K_f would be a nontrivial cone, which is impossible, since K_f is assumed to be bounded (by the same **A**). Thus, at least one of the equations, say, the last of them, is with a nonzero right hand side; normalizing the equation, we may think that it is of the form $e^T x = 1$. Substracting this equation, with properly chosen coefficient, from the remaining m equations of the system, we may make these equations homogeneous, thus reducing the system to the form

$$Ax = 0; \;\; e^T x = 1;$$

now $b + L$ is represented in the desired form

$$b + L = \{x \in M \mid c^T x = 1\}, \;\; M = \{x \mid Ax = 0\}.$$

Thus, we can rewrite (P) as

$$\textit{minimize} \;\; c^T x \;\; s.t. \;\; x \in K \cap M, \; e^T x = 1,$$

with M being a linear subspace in $\mathbf{R}^n$.

It is convenient to convert the problem into an equivalent one where the optimal value of the objective (which, according to **C**, is known in advance) is zero; to this end it suffices to replace the initial objective c with a new one

$$\sigma = c - c^* e;$$

since on the feasible plane of the problem $e^T x$ is identically 1, this updating indeed results in equivalent problem with the optimal value equal to 0.

Thus, we have seen that $(\mathcal{P})$ can be easily rewritten in the so called *Karmarkar format*

$$(\mathcal{P}_K) \;\; \textit{minimize} \;\; \sigma^T x \;\; s.t. \;\; x \in K \cap M, \; e^T x = 1, \tag{6.2}$$

with M being a linear subspace in $\mathbf{R}^n$ and the optimal value in the problem being zero; this transformation preserves, of course, properties **A**, **B**.

Remark 6.2.1 In the original description of the method of Karmarkar, the problem from the very beginning is assumed to be in the form $(\mathcal{P}_K)$, with $K = \mathbf{R}^n_+$; moreover, Karmarkar assumes that

$$e = (1, ..., 1)^T \in \mathbf{R}^n$$

and that the given in advance strictly feasible solution $\widehat{x}$ to the problem is the barycenter $n^{-1}e$ of the standard simplex; thus, in the original version of the method it is assumed that the feasible set K_f of the problem is the intersection of the standard simplex

$$\Delta = \{x \in \mathbf{R}^n_+ \mid \sum_{i=1}^n x_i \equiv e^T x = 1\}$$

and a linear subspace of $\mathbf{R}^n$ passing through the barycenter $n^{-1}e$ of the simplex and, besides this, that the optimal value in the problem is 0.

And, of course, in the Karmarkar paper the barrier for the cone $K = \mathbf{R}^n_+$ underlying the whole construction is the standard n-logarithmically homogeneous barrier

$$F(x) = -\sum_{i=1}^n \ln x_i$$

for the nonnegative orthant.

In what follows we refer to the particular LP situation presented in the above remark as to the *Karmarkar case*.

6.3 The Karmarkar potential function

In what follows we assume that the objective $c^T x$, or, which is the same, our new objective $\sigma^T x$ is nonconstant on the feasible set of the problem (otherwise there is nothing to do: $\sigma^T \widehat{x} = 0$, i.e., the initial strictly feasible solution, same as any feasible solution, is optimal). Since the new objective is nonconstant on the feasible set and its optimal value is 0, it follows that the objective is strictly positive at any strictly feasible solution to the problem, i.e., on the relative interior $\mathrm{rint}\, K_f$ of K_f (due to $\mathbf{A}$, this relative interior is nothing but the intersection of the feasible plane and the interior of K, i.e., nothing but the set of all strictly feasible solutions to the problem). Since $\sigma^T x$ is strictly positive on the relative interior of K_f, the following *Karmarkar potential*

$$v(x) = F(x) + \vartheta \ln(\sigma^T x) : \mathrm{Dom}\, v \equiv \{x \in \mathrm{int}\ K \mid \sigma^T x > 0\} \to \mathbf{R} \tag{6.3}$$

is well-defined on $\mathrm{rint}\, K_f$; this potential is the main hero of our story.

The first observation related to the potential is that *when x is strictly feasible and the potential at x is small (negative with large absolute value), then x is a good approximate solution.*
The exact statement is as follows:

Proposition 6.3.1 *Let $x \in \mathrm{int}\ K$ be feasible for $(\mathcal{P}_K)$. Then*

$$\sigma^T x \equiv c^T x - c^* \leq \mathcal{V} \exp\{-\frac{v(\widehat{x}) - v(x)}{\vartheta}\},\ \ \mathcal{V} = (c^T \widehat{x} - c^*) \exp\{\frac{F(\widehat{x}) - \min_{\mathrm{rint}\, K_f} F}{\vartheta}\}; \tag{6.4}$$

note that $\min_{\mathrm{rint}\, K_f} F$ is well defined, since K_f is bounded (due to $\mathbf{A}$) and the restriction of F onto the relative interior of K_f is self-concordant barrier for K_f (Proposition 3.1.1.(i)).

The proof is immediate:

$$v(\widehat{x}) - v(x) = \vartheta[\ln(\sigma^T \widehat{x}) - \ln(\sigma^T x)] + F(\widehat{x}) - F(x) \leq$$

$$\leq \vartheta[\ln(\sigma^T \widehat{x}) - \ln(\sigma^T x)] + F(\widehat{x}) - \min_{\mathrm{rint}\, K_f} F,$$

and (6.4) follows. ∎

The above observation says to us that all we need is certain rule for updating strictly feasible solution x into another strictly feasible solution x^+ with a "significantly less" value of the potential; iterating this updating, we obtain a sequence of strictly feasible solutions with the potential tending to $-\infty$, so that the solutions converge in terms of the objective. This is how the method works; and the essence of the matter is, of course, the aforementioned updating which we are about to represent.

6.4 The Karmarkar updating scheme

The updating of strictly feasible solutions

$$\mathcal{K} : x \mapsto x^+$$

which underlies the method of Karmarkar is as follows:

1) Given strictly feasible solution x to problem $(\mathcal{P}_K)$, compute the gradient $F'(x)$ of the barrier F;

2) Find the Newton direction e_x of the "partially linearized" potential

$$v_x(y) = F(y) + \vartheta \frac{\sigma^T (y - x)}{\sigma^T x} + \vartheta \ln(\sigma^T x)$$

at the point x along the affine plane

$$E_x = \{y \mid y \in M,\ (y - x)^T F'(x) = 0\}$$

tangent to the corresponding level set of the barrier, i.e., set

$$e_x = \mathrm{argmin}\{h^T \nabla_y v_x(x) + \frac{1}{2} h^T \nabla_y^2 v_x(x) h \mid h \in M,\ h^T F'(x) = 0\};$$

3) Compute the reduced Newton decrement

$$\omega = \sqrt{-e_x^T \nabla_y v_x(x)}$$

and set

$$x' = x + \frac{1}{1+\omega} e_x.$$

4) The point x' belongs to the intersection of the subspace M and the interior of K. Find a point x'' from this intersection such that

$$v(x'') \leq v(x')$$

(e.g., set $x'' = x'$) and set

$$x^+ = (e^T x'')^{-1} x'',$$

thus completing the updating $x \mapsto x^+$.

The following proposition is the central one.

Proposition 6.4.1 *The above updating is well defined, maps a strictly feasible solution x to $(\mathcal{P})_K$ into another strictly feasible solution x^+ to $(\mathcal{P})$ and decreases the Karmarkar potential at least by absolute constant:*

$$v(x^+) \leq v(x) - \chi, \;\; \chi = \frac{1}{3} - \ln\frac{4}{3} > 0. \tag{6.5}$$

Proof.

0^0. Let us start with the following simple observations:

$$y \in \text{int } K \cap M \Rightarrow e^T y > 0; \tag{6.6}$$

$$y \in \text{int } K \cap M \Rightarrow \sigma^T y > 0. \tag{6.7}$$

To prove (6.6), assume, on contrary, that there exists $y \in \text{int } K \cap M$ with $e^T y \leq 0$. Consider the linear function

$$\phi(t) = e^T[\widehat{x} + t(y - \widehat{x})], \;\; 0 \leq t \leq 1.$$

This function is positive at $t = 0$ (since $\widehat{x}$ is feasible) and nonpositive at $t = 1$; therefore it has a unique root $t^* \in (0, 1]$ and is positive to the left of this root. We conclude that the points

$$x_t = \phi^{-1}(t)[\widehat{x} + t(y - \widehat{x})], \;\; 0 \leq t < t^*,$$

are well defined and, moreover, belong to K_f (indeed, since both x and y are in $K \cap M$ and $\phi(t)$ is positive for $0 \leq t < t^*$, the points x_t also are in $K \cap M$; to establish feasibility, we should verify, in addition, that $e^T x_t = 1$, which is evident).

Thus, x_t, $0 \leq t < t^*$, is certain curve in the feasible set. Let us prove that $|x_t|_2 \to \infty$ as $t \to t^* - 0$; this will be the desired contradiction, since K_f is assumed to be bounded (see **A**). Indeed, $\phi(t) \to 0$ as $t \to t^* - 0$, while $x + t(y - x)$ has a nonzero limit $x + t^*(y - x)$ (this limit is nonzero as a convex combination of two points from the interior of K and, therefore, a point from this interior; recall that K is pointed, so that the origin is not in its interior).

We have proved (6.6); (6.7) is an immediate consequence of this relation, since if there were $y \in \text{int } K \cap M$ with $\sigma^T y \leq 0$, the vector $[e^T y]^{-1} y$ would be a strictly feasible solution to the problem (since we already know that $e^T y > 0$, so that the normalization $y \mapsto [e^T y]^{-1} y$ would keep the point in the interior of the cone) with nonnegative value of the objective, which, as we know, is impossible.

1^0. Let us set

$$G = K \cap E_x \equiv K \cap \{y \mid y \in M, \; (y - x)^T F'(x) = 0\};$$

since $x \in M$ is an interior point of K, G is a closed convex domain in the affine plane E_x (this latter plane from now on is regarded as the linear space G is embedded to); the (relative) interior of G is exactly the intersection of E_x and the interior of the cone K.

2^0. Further, let $f(\cdot)$ be the restriction of the barrier F on $\operatorname{rint} G$; due to our combination rules for self-concordant barriers, namely, that one on affine substitutions of argument, f is ϑ-self-concordant barrier for G.

3^0. By construction, the "partially linearized" potential, regarded as a function on $\operatorname{rint} G$, is the sum of the barrier f and a linear form:

$$v_x(y) = f(y) + p^T(y-x) + q,$$

where the linear term $p^T(y-x)+q$ is nothing but the first order Taylor expansion of the function

$$\vartheta \ln(\sigma^T y)$$

at the point $y = x$. From (6.7) it immediately follows that this function (and therefore $v(\cdot)$) is well-defined onto $\operatorname{int} K \cap M$ and, consequently, on $\operatorname{rint} G$; besides this, the function is concave in $y \in \operatorname{rint} G$. Thus, we have

$$v(y) \leq v_x(y),\ y \in \operatorname{rint} G;\ \ v(x) = v_x(x). \tag{6.8}$$

4^0. Since v_x is sum of a self-concordant barrier and a linear form, it is self-concordant on the set $\operatorname{rint} G$. From definition of e and ω it is immediately seen that e_x is nothing but the Newton direction of $v_x(y)$ (*regarded as a function on* $\operatorname{rint} G$) at the point $y = x$, and ω is the corresponding Newton decrement; consequently (look at rule 3)) x' is the iterate of $y = x$ under the action of the damped Newton method. From Lecture 2 we know that this iterate belongs to $\operatorname{rint} G$ and that the iteration of the method decreases v_x "significantly", namely, that

$$v_x(x) - v_x(x') \geq \rho(-\omega) = \omega - \ln(1+\omega).$$

Taking into account (6.8), we conclude that

x' belongs to the intersection of the subspace M and the interior of the cone K and

$$v(x) - v(x') \geq \rho(-\omega). \tag{6.9}$$

5^0. Now comes the first crucial point of the proof: *the reduced Newton decrement ω is not too small, namely,*

$$\omega \geq \frac{1}{3}. \tag{6.10}$$

Indeed, x is the analytic center of G with respect to the barrier f (since, by construction, E_x is orthogonal to the gradient F' of the barrier F at x, and f is the restriction of F onto E_x). Since f, as we just have mentioned, is ϑ-self-concordant barrier for G, and f is nondegenerate (as a restriction of a nondegenerate self-concordant barrier F, see Proposition 5.3.1), the enlarged Dikin ellipsoid

$$W^+ = \{y \in E_x \mid |y-x|_x \leq \vartheta + 2\sqrt{\vartheta}\}$$

($|\cdot|_x$ is the Euclidean norm generated by $F''(x)$) contains the whole G (the Centering property, Lecture 3, **V.**). Now, the optimal solution x^* to $(\mathcal{P}_K)$ satisfies the relation $\sigma^T x^* = 0$ (the origin of σ) and is a nonzero vector from $K \cap M$ (since x^* is feasible for the problem). It follows that the quantity $(x^*)^T F'(x)$ is negative (since $F'(x) \in \operatorname{int}(-K^*)$, Proposition 5.3.3.(i)), and therefore the ray spanned by x^* intersects G at certain point y^* (indeed, G is the part of $K \cap M$ given by the linear equation $y^T F'(x) = x^T F'(x)$, and the right hand side in this equation is $-\vartheta$, see (5.5), Lecture 5, i.e., is of the same sign as $(x^*)^T F'(x)$). Since $\sigma^T x^* = 0$, we have $\sigma^T y^* = 0$; thus,

there exists y^* in G, and, consequently, in the ellipsoid W^+, with $\sigma^T y^* = 0$.

We conclude that the linear form

$$\psi(y) = \vartheta \frac{\sigma^T y}{\sigma^T x}$$

which is equal to ϑ at the center x of the ellipsoid W^+, attains the zero value somewhere in the ellipsoid, and therefore its variation over the ellipsoid is at least 2ϑ. Consequently, the variation of the form over the centered at x unit Dikin ellipsoid of the barrier f is at least $2\vartheta(\vartheta + 2\sqrt{\vartheta})^{-1} \geq 2/3$:

$$\max\{\vartheta \frac{\sigma^T h}{\sigma^T x} \mid h \in M, h^T F'(x) = 0, |h|_x \leq 1\} \geq \frac{1}{3}.$$

But the linear form in question is exactly $\nabla_y v_x(x)$, since $\nabla_y f(x) = 0$ (recall that x is the analytic center of G with respect to f), so that the left hand side in the latter inequality is the Newton decrement of $v_x(\cdot)$ (as always, regarded as a function on $\mathrm{rint}\, G$) at x, i.e., it is nothing but ω.

6^0. Now comes the concluding step: *the Karmarkar potential v is constant along rays: $v(tu) = v(t)$ whenever $u \in \mathrm{Dom}\, v$ and $t > 0$* [this is an immediate consequence of ϑ-logarithmic homogeneity of the barrier F][1]. As we just have seen,

$$v(x') \leq v(x) - \rho(-\frac{1}{3});$$

by construction, x'' is a point from int $K \cap M$ such that

$$v(x'') \leq v(x').$$

According to (6.6), when passing from x'' to $x^+ = [e^T x'']^{-1} x''$, we get a strictly feasible solution to the problem, and due to the fact that v remains constant along rays, $v(x^+) = v(x'')$. Thus, we come to $v(x^+) \leq v(x) - \rho(-\frac{1}{3})$, as claimed. ■

6.5 Overall complexity of the method

As it was already indicated, the method of Karmarkar as applied to problem $(\mathcal{P}_K)$ simply iterates the updating $\mathcal{K}$ presented in Section 6.4, i.e., generates the sequence

$$x_i = \mathcal{K}(x_{i-1}), \;\; x_0 = \widehat{x}, \tag{6.11}$$

$\widehat{x}$ being the initial strictly feasible solution to the problem (see **B**).

An immediate corollary of Propositions 6.3.1 and 6.4.1 is the following complexity result:

Theorem 6.5.1 *Let problem $(\mathcal{P}_K)$ be solved by the method of Karmarkar associated with ϑ-logarithmically homogeneous barrier F for the cone K, and let assumptions **A** - **C** be satisfied. Then the iterates x_i generated by the method are strictly feasible solutions to the problem and*

$$c^T x_i - c^* \leq \mathcal{V} \exp\{-\frac{v(\widehat{x}) - v(x_i)}{\vartheta}\} \leq \mathcal{V} \exp\{-\frac{i\chi}{\vartheta}\}, \;\; \chi = \frac{1}{3} - \ln\frac{4}{3}, \tag{6.12}$$

with the data-dependent scale factor $\mathcal{V}$ given by

$$\mathcal{V} = (c^T \widehat{x} - c^*) \exp\{\frac{F(\widehat{x}) - \min_{\mathrm{rint}\, K_f} F}{\vartheta}\}. \tag{6.13}$$

In particular, the Newton complexity (# of iterations of the method) of finding an ε-solution to the problem does not exceed the quantity

$$\mathcal{N}_{\mathrm{Karm}}(\varepsilon) = O(1)\vartheta \ln\left(\frac{\mathcal{V}}{\varepsilon} + 1\right) + 1, \tag{6.14}$$

$O(1)$ being an absolute constant.

Comments.

- We see that the Newton complexity of finding an ε-solution by the method of Karmarkar is proportional to ϑ; on the other hand, the restriction of F on the feasible set K_f is a ϑ-self-concordant barrier for this set (Proposition 3.1.1.(i)), and we might solve the problem by the path-following method associated with this restriction, which would result in a better Newton complexity, namely, proportional to $\sqrt{\vartheta}$. Thus, from the theoretical complexity viewpoint the method of Karmarkar is significantly worse than the path-following method; why should we be interested in the method of Karmarkar?

[1] and in fact the assumption of logarithmic homogeneity of F, same as the form of the Karmarkar potential, originate exactly from the desire to make the potential constant along rays

The answer is: due to the potential reduction nature of the method, the nature which underlies the excellent practical performance of the algorithm. Look: in the above reasoning, the only thing we are interested in is to decrease as fast as possible certain explicitly given function - the potential. The theory gives us certain "default" way of updating the current iterate in a manner which guarantees certain progress (at least by an absolute constant) in the value of the potential at each iteration, and it does not forbid as to do whatever we want to get a better progress (this possibility was explicitly indicated in our construction, see the requirements on x''). E.g., after x' is found, we can perform the line search on the intersection of the ray $[x, x')$ with the interior of G in order to choose as x'' the best, in terms of the potential, point of this intersection rather than the "default" point x'. There are strong reasons to expect that in some important cases the line search decreases the value of the potential by much larger quantity than that one given by the above theoretical analysis (see exercises accompanying this lecture); in accordance with these expectations, the method in fact behaves itself incomparably better than it is said by the theoretical complexity analysis.

- What is also important is that all "common sense" improvements of the basic Karmarkar scheme, like the aforementioned line search, do not spoil the theoretical complexity bound; and from the practical viewpoint a very attractive property of the method is that the potential gives us a clear criterion to decide what is good and what is bad. In contrast to this, in the path-following scheme we either should follow the theoretical recommendations on the rate of updating the penalty - and then for sure will be enforced to perform a lot of Newton steps - or could increase the penalty at a significantly higher rate, thus destroying the theoretical complexity bound and imposing a very difficult questions of how to choose and to tune this higher rate.

- Let me say several words about the original method of Karmarkar for LP. In fact this is exactly the particular case of the aforementioned scheme for the sutiation described in Remark 6.2.1; Karmarkar, anyhow, presents the same method in a different way. Namely, instead of processing *the same data* in varying, from iteration to iteration, plane E_x, he uses *scaling* - after a new iterate x_i is found, he performs fractional-linear substitution of the argument

$$x \mapsto \frac{X_i^{-1}x}{e^T X_i^{-1}x}, \quad X_i = \mathrm{Diag}\{x_i\}$$

(recall that in the Karmarkar situation $e = (1, ..., 1)^T$). With this substitution, the problem becomes *another* problem of the same type (with new objective σ and new linear subspace M), and the image of the actual iterate x_i becomes the barycenter $n^{-1}e$ of the simplex Δ. It is immediately seen that in the Karmarkar case to decrease by something the Karmarkar potential for the new problem at the image $n^{-1}e$ of the current iterate is the same as to decrease by the same quantity the potential of the initial problem at the actual iterate x_i; thus, scaling allows to reduce the question of how to decrease the potential to the particular case when the current iterate is the barycenter of Δ; this (specific for LP) possibility to deal with certain convenient "standard configuration" allows to carry out all required estimates (which in our approach were consequences of general properties of self-concordant barriers) via direct analysis of the behaviour of the standard logarithmic barrier $F(x) = -\sum_i \ln x_i$ in a neighbourhood of the point $n^{-1}e$, which is quite straightforward.

Let me also add that in the Karmarkar situation our general estimate becomes

$$c^T x_i - c^* \leq (c^T \widehat{x} - c^*) \exp\{-\frac{i\chi}{n}\},$$

since the parameter of the barrier in the case in question is $\vartheta = n$ and the starting point $\widehat{x} = n^{-1}e$ is the minimizer of F on Δ and, consequently, on the feasible set of the problem.

6.6 How to implement the method of Karmarkar

To the moment our abilities to solve conic problems by the method of Karmarkar are restricted by the assumptions **A** - **C**. Among these assumptions, **A** (strict feasibility of the problem and boundedness of the feasible set) is not that restrictive. Assumption **B** (a strictly feasible solution should be known in advance) is not so pleasant, but let me postpone discussing this issue - this is a common problem in interior point methods, and in the mean time we shall speak about it. And what in fact is restrictive, is assumption **C** - we should know in advance the optimal value in the problem. There are several ways to eliminate this unpleasant hypothesis; let me present to you the simplest one - the *sliding objective* approach. Assume, instead of **C**, that

C*: *we are given in advance a lower bound c_0^* for the unknown optimal value c^**

(this, of course, is by far less restrictive than the assumption that we know c^* exactly). In this case we may act as follows: at i-th iteration of the method, we use certain lower bound c_{i-1}^* for c^* (the initial lower bound c_0^* is given by $\mathbf{C}^*$). When updating x_i into x_{i+1}, we begin exactly as in the original method, but use, instead of the objective

$$\sigma = c - c^* e,$$

the "current objective"

$$\sigma_{i-1} = c - c_{i-1}^* e.$$

Now, after the current "reduced Newton decrement" $\omega = \omega_i$ is computed, we check whether it is $\geq \frac{1}{3}$. If it is the case, we proceed exactly as in the original scheme and do not vary the current lower bound for the optimal value, i.e., set

$$c_i^* = c_{i-1}^*$$

and, consequently,

$$\sigma_i = \sigma_{i-1}.$$

If it turns out that $\omega_i < 1/3$, we act as follows. The quantity ω given by rule 3) depends on the objective σ the rules 1)-3) are applied to:

$$\omega = \Omega_i(\sigma).$$

In the case in question we have

$$\Omega_i(c - te) < \frac{1}{3} \quad \textit{when} \quad t = c_{i-1}^*. \tag{6.15}$$

The left hand side of this relation is certain explicit function of t (square root of a nonnegative fractional-quadratic form of t); and as we know from the proof of Proposition 6.4.1,

$$\Omega_i(c - c^* e) \geq \frac{1}{3}. \tag{6.16}$$

It follows that the equation $\Omega_i(c - te) = \frac{1}{3}$ is solvable, and its closest to c_{i-1}^* root to the right of c_{i-1}^* separates c^* and c_{i-1}^*, i.e., this root (which can be immediately computed) is an improved lower bound for c^*. This is exactly the lower bound which we take as c_i^*; after it is found, we set

$$\sigma_i = c - c_i^* e$$

and update x_i into x_{i+1} by the basic scheme applied to this "improved" objective (for which this scheme, by construction, results in $\omega = \frac{1}{3}$).

Following the line of argument used in the proofs of Propositions 6.3.1, 6.4.1, one can verify that the modification in question produces strictly feasible solutions x_i and *nondecreasing* lower bounds $c_i^* \leq c^*$ of the unknown optimal value in such a way that the sequence of *local potentials*

$$v_i(x_i) = F(x_i) + \vartheta \ln(\sigma_i^T x_i) \equiv F(x_i) + \vartheta \ln(c^T x_i - c_i^*)$$

decreases at a reasonable rate:

$$v_i(x_i) \leq v_{i-1}(x_{i-1}) - \rho(-\frac{1}{3}),$$

which, in turn, ensures the rate of convergence

$$c^T x_i - c^* \leq \mathcal{V} \exp\{-\frac{v_0(x_0) - v_i(x_i)}{\vartheta}\} \leq \mathcal{V} \exp\{-\frac{i\chi}{\vartheta}\},$$

$$\mathcal{V} = (c^T \widehat{x} - c_0^*) \exp\{\frac{F(\widehat{x}) - \min_{\mathrm{rint}\, K_f} F}{\vartheta}\}$$

completely similar to that one for the case of known optimal value.

6.7 Exercises on the method of Karmarkar

Our first exercise is quite natural.

Exercise 6.7.1 $^{\#}$. *Justify the sliding objective approach presented in Section 6.6.*

Our next story gives a very instructive equivalent description of the method of Karmarkar (in the LP case, this description is due to Bayer and Lagarias). At a step of the method the situation is as follows: we are given a strictly feasible solution x to $(\mathcal{P}_K)$ and are struggling for updating it into a new strictly feasible solution with "significantly less" value of the potential. Now, strictly feasible solutions are in one-to-one correspondence with *strictly feasible rays* - i.e., rays $r = \{ty \mid t > 0\}$ generated by $y \in M \cap \operatorname{int} K$. Indeed, any strictly feasible solution x spans a unique ray of this type, and any strictly feasible ray intersects the relative interior of the feasible set in a unique point (since, as we know from (6.6), the quantity $e^T y$ is positive whenever $y \in M \cap \operatorname{int} K$ and therefore the normalization $[e^T y]^{-1} y$ is a strictly feasible solution to the problem). On the other hand, the Karmarkar potential v is constant along rays, and therefore it can be thought of as a function defined on the space $\mathcal{R}$ of strictly feasible rays. Thus, the goal of a step can be reformulated as follows:

given a strictly feasible ray r, find a new ray r^+ of this type with "significantly less" value of the potential.

Now let us make the following observation: there are many ways to identify strictly feasible rays with points of certain set; e.g., given a linear functional $g^T x$ which is positive on $M \cap \operatorname{int} K$, we may consider the cross-section K^g of $M \cap K$ by the hyperplane given by the equation $g^T x = 1$. It is immediately seen that any strictly feasible ray intersects the relative interior of K^g and, vice versa, any point from this relative interior spans a strictly feasible ray. What we used in the initial representation of the method, was the "parameterization" of the space $\mathcal{R}$ of strictly feasible rays by the points of the relative interior of the feasible set K_f (i.e., by the set K^e associated, in the aforementioned sense, with the constraint functional $e^T x$). Now, what happens if we use another parameterization of $\mathcal{R}$? Note that we have a natural candidate on the role of g - the objective σ (indeed, we know that $\sigma^T x$ is positive at any strictly feasible x and therefore is positive on $M \cap \operatorname{int} K$). What is the potential in terms of our new parameterization of $\mathcal{R}$, where a strictly feasible ray r is represented by its intersection $y(r)$ with the plane $\{y \mid \sigma^T y = 1\}$? The answer is immediate:

$$v(y(r)) = F(y(r)).$$

In other words, the goal of a step can be equivalently reformulated as follows:

given a point y from the relative interior of the set

$$K^\sigma = \{z \in M \cap K \mid \sigma^T z = 1\},$$

find a new point y^+ of this relative interior with $F(y^+)$ being "significantly less" than $F(y)$.

Could you guess what is the "linesearch" (with $x'' = \operatorname{argmin}_{y = x + t(x' - x)} v(y)$) version of the Karmarkar updating $\mathcal{K}$ in terms of this new parameterization of $\mathcal{R}$?

Exercise 6.7.2 $^{\#}$ *Verify that the Karmarkar updating with linesearch is nothing but the Newton iteration with linesearch as applied to the restriction of F onto the relative interior of K^σ.*

Now, can we see from our new interpretation of the method why it converges at the rate given by Theorem 6.5.1? This is immediate:

Exercise 6.7.3 $^{\#+}$ *Prove that*

- *the set K^σ is unbounded;*

- *the Newton decrement $\lambda(\phi, u)$ of the restriction ϕ of the barrier F onto the relative interior of K^σ is ≥ 1 at any point $u \in \operatorname{rint} K^\sigma$;*

- *each damped Newton iteration (and therefore - Newton iteration with linesearch) as applied to ϕ decreases ϕ at least by $1 - \ln 2 > 0$.*

Conclude from these observations that each iteration of the Karmarkar method with linesearch reduces the potential at least by $1 - \ln 2$.

Now we understand what in fact goes on in the method of Kramarkar. We start from the problem of minimizing a linear objective over a closed and bounded convex domain K_f; we know the optimal value, i.e., we know what is the hyperplane $\{c^T x = c^*\}$ which touches the feasible set; what we do not know and what should be found, is *where* the plane touches the feasible set. What we do is as follows (the below explanation is illustrated by a picture at the next page): we perform *projective transformation* of the affine hull of K_f which moves the target plane $\{c^T x = c^*\}$ to infinity (this is exactly the transformation of K_f onto K^σ given by the receipt: to find an image of $x \in \operatorname{rint} K_f$, take the intersection of the ray spanned by x with the hyperplane $\{\sigma^T y = 1\}$). The image of the feasible set K_f of the problem is an *unbounded* convex domain K^σ, and our goal is to go to infinity, staying within this image (the inverse image of the point moving in K^σ will then stay within K_f and approach the target plane $\{c^T x = c^*\}$). Now, in order to solve this latter problem, we take a self-concordant barrier ϕ for K^σ and apply to this barrier the damped Newton method (or the Newton method with linesearch). As explained in Exercise 6.7.3, the routine decreases ϕ at every step at least by absolute constant, thus enforcing ϕ to tend to $-\infty$ at certain rate. Since ϕ is convex (and therefore below bounded on any bounded subset of K^σ), this inevitably enforces the iterate to go to infinity. Rather sophisticated way to go far away, isn't it?

Our last story is related to a quite different issue - to the *anitcipated behaviour* of the method of Karmarkar. The question, unformally, is as follows: we know that a step of the method decreases the potential at least by an absolute constant; this is given by our theoretical worst-case analysis. What is the "expected" progress in the potential?

It hardly makes sense to pose this question in the general case. In what follows we restrict ourselves to the case of *semidefinite programming*, where

$$K = \mathbf{S}^n_+$$

is the cone of positive semidefinite symmetric $n \times n$ matrices and

$$F(x) = -\ln \operatorname{Det} x$$

is the standard n-logarithmically homogeneous self-concordant barrier for the cone (Lecture 5, Example 5.3.3); the below considerations can be word by word repeated for the case of LP ($K = \mathbf{R}^n_+$, $F(x) = -\sum_i \ln x_i$).

Consider a step of the method of Karmarkar with linesearch, the method being applied to a semidefinite program. Let x be the current strictly feasible solution and x^+ be its iterate given by a single step of the *linesearch* version of the method. Let us pose the following question:

(?) *what is the progress* $\alpha = v(x) - v(x^+)$ *in the potential at the step in question?*

To answer this question, it is convenient to pass to certain "standard configuration" - to perform *scaling*. Namely, consider the linear transformation

$$u \mapsto \mathcal{X}u = x^{-1/2} u x^{-1/2}$$

in the space of symetric $n \times n$ matrices.

Exercise 6.7.4 $^\#$ *Prove that the scaling* $\mathcal{X}$ *possesses the following properties:*

- *it is a one-to-one mapping of* int K *onto itself;*
- *it "almost preserves" the barrier:*

$$F(\mathcal{X}u) = F(u) + const(x);$$

in particular,

$$|\mathcal{X}h|_{\mathcal{X}u} = |h|_u, \ \ u \in \operatorname{int} K, h \in \mathbf{S}^n;$$

- *the scaling maps the feasible set K_f of problem $(\mathcal{P}_K)$ onto the feasible set of another problem $(\mathcal{P}'_K)$ of the same type; the updated problem is defined by the subspace*
$$M' = \mathcal{X}M,$$
the normalizing equation $(e', x) = 1$ with
$$e' = x^{1/2}ex^{1/2}$$
and the objective
$$\sigma' = x^{1/2}\sigma x^{1/2};$$
this problem also satisfies the assumptions **A** - **C***;*

- *let $v(\cdot)$ be the potential of the initial problem, and v' be the potential of the new one. Then the potentials at the corresponding points coincide, up to an additive constant:*
$$\mathrm{Dom}\, v' = \mathcal{X}(\mathrm{Dom}\, v); \;\; v'(\mathcal{X}u) - v(u) \equiv const, , \;\; u \in \mathrm{Dom}\, v;$$

- *$\mathcal{X}$ maps the point x onto the unit matrix I, and the iterate x^+ of x given by the linesearch version of the method as applied to the initial problem into the similar iterate I^+ of I given by the linesearch version of the method as applied to the transformed problem.*

From Exercise 6.7.4 it is clear that in order to answer the question (?), it suffices to answer the similar question (of course, not about the initial problem itself, but about a problem of the same type with updated data) *for the particular case when the current iterate is the unit matrix I*. Let us consider this special case. In what follows we use the *original* notation for the data of the *transformed* problem; this should not cause any confusion, since we shall speak about exactly one step of the method.

Now, what is the situation in our "standard configuration" case $x = I$? It is as follows:

we are given a linear subspace M passing through $x = I$ and the objective σ; what we know is that[2]

I. $(\sigma, u) \geq 0$ whenever $u \in \mathrm{int}\, K \cap M$ and there exists a nonzero matrix $x^* \in \mathrm{int}\, K \cap M$ such that $(\sigma, x^*) = 0$;

II. In order to update $x = I$ into x^+, we compute the steepest descent direction ξ of the Karmarkar potential $v(\cdot)$ at the point x along the affine plane
$$E_x = \{y \in M \mid (F'(x), y - x) = 0\},$$
the metric in the subspace being $|h|_x \equiv (F''(x)h, h)^{1/2}$, i.e., find among the unit, with respect to the indicated norm, directions parallel to E_x that one with the smallest (e.g., the "most negative") inner product onto $v'(x)$. Note that the Newton direction e_x is proportional, with positive coefficient, to the steepest descent direction ξ. Note also, that the steepest descent direction of v at x is the same as the similar direction for the function $n \ln((\sigma, u))$ at $u = x$ (recall that for the barrier in question $\vartheta = n$), since x is the minimizer of the remaining component $F(\cdot)$ of $v(\cdot)$ along E_x.

Now, in our standard configuration case $x = i$ we have $F'(x) = -I$, and $|h|_x = (h, h)^{1/2}$ is the usual Frobenius norm[3]; thus, ξ is the steepest descent direction of the linear form
$$\phi(h) = n(\sigma, h)/(\sigma, I)$$
(this is the differential of $n \ln((\sigma, u))$ at $u = I$) taken along the subspace
$$\Pi = M \cap \{h : \mathrm{Tr}\, h \equiv (F'(I), h) = 0\}$$
with respect to the standard Euclidean structure of our universe $\mathbf{S}^n$. In other words, ξ is proportional, with *negative* coefficient, to the *orthogonal projection* η *of*
$$S \equiv (\sigma, I)^{-1}\sigma$$

[2] from now on we denote the inner product on the space in question, i.e., on the space $\mathbf{S}^n$ of symmetric $n \times n$ matrices, by (x, y) (recall that this is the Frobenius inner product $\mathrm{Tr}\{xy\}$), in order to avoid confusion with the matrix products like $x^T y$

[3] due to the useful formulae for the derivatives of the barrier $F(u) = -\ln \mathrm{Det}\, u$: $F'(u) = -u^{-1}$, $F''(u)h = u^{-1}hu^{-1}$; those solved Exercise 3.3.3, for sure know these formulae, and all others are kindly asked to derive them

onto the subspace Π.

From these observations we conclude that

III. $\mathrm{Tr}\,\eta = 0$; $\mathrm{Tr}\,S = 1$ (since $\eta \in \Pi$ and Π is contained in the subspace of matrices with zero trace, and due to the origin of S, respectively);

IVa. $(S, u) > 0$ for all positive definite u of the form $I + r\eta$, $r \in \mathbf{R}$ (an immediate consequence of **I.**);

IVb. There exists positive semidefinite matrix χ^* such that $\chi^* - I \in \Pi$ and $(S, \chi^*) = 0$ (χ^* is proportional to x^* with the coefficient given by the requirement that $(F'(I), \chi^* - I) = 0$, or, which is the same, by the requirement that $\mathrm{Tr}\,\chi^* = n$; recall that $F'(I) = -I$).

Now, at the step we choose t^* as the minimizer of the potential $v(I - t\eta)$ over the set t of nonnegative T such that $I - t\eta \in \mathrm{Dom}\,v$, or, which is the same in view of **I.**, such that $I - t\eta$ is positive definite[4], and define x^+ as $(e, x'')^{-1}x''$, $x'' = I - t^*\eta$; the normalization $x'' \mapsto x^+$ does not vary the potential, so that the quantity α we are interested in is simply $v(I) - v(x'')$.

To proceed, let us look at the potential along our search ray:

$$v(I - t\eta) = -\ln \mathrm{Det}\,(I - t\eta) + n\ln((S, I - t\eta)).$$

III. says to us that $(S, I) = 1$; since η is the orthoprojection of S onto Π (see **II.**), we have also $(S, \eta) = (\eta, \eta)$. Thus,

$$\phi(t) \equiv v(I - t\eta) = -\ln \mathrm{Det}\,(I - t\eta) + n\ln(1 - t(\eta, \eta)) = -\sum_{i=1}^{n} \ln((1 - tg_i) + n\ln(1 - t|g|_2^2), \qquad (6.17)$$

where $g = (g_1, ..., g_n)^T$ is the vector comprised of the eigenvalues of the symmetric matrix η.

Exercise 6.7.5 $^{\#+}$ *Prove that*

1) $\sum_{i=1}^{n} g_i = 0$*;*

2) $|g|_\infty \geq n^{-1}$.

Now, from (6.17) it turns out that the progress in the potential is given by

$$\alpha = \phi(0) - \min_{t \in T} \phi(t) = \max_{t \in T} [\sum_{i=1}^{n} \ln(1 - tg_i) - n\ln(1 - t|g|_2^2)], \qquad (6.18)$$

where $T = \{t \geq 0 \mid 1 - tg_i > 0,\ i = 1, ..., n\}$.

Exercise 6.7.6 $^{\#+}$ *Testing the value of t equal to*

$$\tau \equiv \frac{n}{1 + n|g|_\infty},$$

demonstrate that

$$\alpha \geq (1 - \ln 2)\left(\frac{|g|_2}{|g|_\infty}\right)^2. \qquad (6.19)$$

The conclusion of our analysis is as follows:

each step of the method of Karmarkar with linesearch applied to a semidefinite program can be associated with an n-dimensional vector g (depending on the data and the iteration number) in such a way that the progress in the Karmarkar potential at a step is at least the quantity given by (6.19).

Now, the worst case complexity bound for the method comes from the worst case value of the right hand side in (6.19); this latter value (equal to $1 - \ln 2$) corresponds to the case when $|g|_2|g|_\infty^{-1} \equiv \pi(g)$ attains its minimum in g (which is equal to 1); note that $\pi(g)$ is of order of 1 only if g is an "orth-like" vector - its 2-norm comes from $O(1)$ dominating coordinates. Note, anyhow, that the "typical" n-dimensional vector is far from being an "orth-like" one, and the "typical" value of $\pi(g)$ is much larger than 1. Namely, if g is a random vector in $\mathbf{R}^n$ with the direction uniformly distributed on the unit sphere, than the "typical value" of $\pi(g)$ is of order of $\sqrt{n/\ln n}$ (the probability for π to be less than certain absolute constant

[4] recall that e_x is proportional, with positive coefficient, to ξ and, consequently, is proportional, with negative coefficient, to η

times this square root tends to 0 as $n \to \infty$; please prove this simple statement). If (if!) we could use this "typical" value of $\pi(g)$ in our lower bound for the progress in the potential, we would come to the progress per step equal to $O(n/\ln n)$ rather than to the worst-case value $O(1)$; as a result, the Newton complexity of finding ε-solution would be proportional to $\ln n$ rather than to n, which would be actually excellent! Needless to say, there is no way to *prove* something definite of this type, even after we equip the family of problems in question by a probability distribution in order to treat the vectors g arising at sequential steps as a random sequence. The diffuculty is that the future of the algorithm is strongly predetermined by its past, so that any initial symmetry seems to be destroyed as the algorithm goes on.

Note, anyhow, that impossibility to *prove* something does not necessarily imply impossibility to *understand* it. The "anticipated" complexity of the method (proportional to $\ln n$ rather than to n) seems to be quite similar to its empirical complexity; given the results of the above "analysis", one hardly could be too surprised by this phenomenon.

Chapter 7

The Primal-Dual potential reduction method

We became acquainted with the very first of the potential reduction interior point methods - with the method of Karmarkar. Theoretically, a disadvantage of the method is in not so good complexity bound - it is proportional to the parameter ϑ of the underlying barrier, not to the *square root* of this parameter, as in the case of the path-following method. There are, anyhow, potential reduction methods with the same theoretical $O(\sqrt{\vartheta})$ complexity bound as in the path-following scheme; these methods combine the best known theoretical complexity with the practical advantages of the potential reduction algorithms. Our today lecture is devoted to one of these methods, the so called *Primal-Dual algorithm*; the LP prototype of the construction is due to Todd and Ye.

7.1 The idea

The idea of the method is as follows. Consider a convex problem in the conic form

$$(\mathcal{P}): \quad \textit{minimize} \;\; c^T x \;\; \textit{s.t.} \;\; x \in \{b + L\} \cap K$$

along with its conic dual

$$(\mathcal{D}): \quad \textit{minimize} \;\; b^T s \;\; \textit{s.t.} \;\; s \in \{c + L^{\perp}\} \cap K^*,$$

where

- K is a cone (closed, pointed, convex and with a nonempty interior) in $\mathbf{R}^n$ and

  $$K^* = \{s \in \mathbf{R}^n \mid s^T x \geq 0 \;\; \forall x \in K\}$$

  is the cone dual to K;
- L is a linear subspace in $\mathbf{R}^n$, $L^{\perp}$ is its orthogonal complement and c, b are given vectors from $\mathbf{R}^n$ - the primal objective and the primal translation vector, respectively.

From now on, we assume that

A: *both primal and dual problems are strictly feasible, and we are given an initial strictly feasible primal-dual pair* $(\widehat{x}, \widehat{s})$ [i.e., a pair of strictly feasible solutions to the problems].

This assumption, by virtue of the Conic duality theorem (Lecture 5), implies that both the primal and the dual problem are solvable, and the sum of the optimal values in the problems is equal to $c^T b$:

$$\mathcal{P}^* + \mathcal{D}^* = c^T b. \tag{7.1}$$

Besides this, we know from Lecture 5 that for any pair (x, s) of feasible solutions to the problems one has

$$\delta(x, s) \equiv c^T x + b^T s - c^T b = s^T x \geq 0. \tag{7.2}$$

Substracting from this identity equality (7.1), we come to the following conclusion:

(*): *for any primal-dual feasible pair (x, s), the duality gap $\delta(x, s)$ is nothing but the sum of inaccuracies, in terms of the corresponding objectives, of x regarded as an approximate solution to the primal problem and s regarded as an approximate solution to the dual one.*

In particular, all we need is to generate somehow a sequence of primal-dual feasible pairs with the duality gap tending to zero.

Now, how to enforce the duality gap to go to zero? To this end we shall use certain potential; to construct this potential, this is our first goal.

7.2 Primal-dual potential

From now on we assume that

B: we know a ϑ-logarithmically homogeneous self-concordant barrier F for the primal cone K along with its Legendre transformation

$$F^*(s) = \sup_{x \in \mathrm{int}\ K} [s^T x - F(x)].$$

("we know", as usual, means that given x, we can check whether $x \in \mathrm{Dom}\, F$ and if it is the case, can compute $F(x)$, $F'(x)$, $F''(x)$, and similarly for F^*).

As we know from Lecture 5, F^* is ϑ-logarithmically homogeneous self-concordant barrier for the cone $-K^*$ anti-dual to K, and, consequently, the function

$$F^+(s) = F^*(-s)$$

is a ϑ-logarithmically homogeneous self-concordant barrier for the dual cone K^* involved into the dual problem. In what follows I refer to F as to the *primal*, and to F^+ - as to the *dual* barrier.

Now let us consider the following aggregate:

$$V_0(x, s) = F(x) + F^+(s) + \vartheta \ln(s^T x) \tag{7.3}$$

This function is well-defined on the direct product of the interiors of the primal and the dual cones, and, in particular, on the direct product

$$\mathrm{rint}\, K_p \times \mathrm{rint}\, K_d$$

of the relative interiors of the primal and dual feasible sets

$$K_p = \{b + L\} \cap K, \;\; K_d = \{c + L^\perp\} \cap K^*.$$

The function V_0 resembles the Karmarkar potential; indeed, when $s \in \mathrm{rint}\, K_d$ is fixed, this function, regarded as a function of primal feasible x, is, up to an additive constant, the Karmarkar potential of the primal problem, where one should replace the initial objective c by the objective s [1].

Note that we know something about the aggregate V_0: Proposition 5.3.3 says to us that

(**) *for any pair $(x, s) \in \mathrm{Dom}\, V_0 \equiv \mathrm{int}\,(K \times K^*)$, one has*

$$V_0(x, s) \geq \vartheta \ln \vartheta - \vartheta, \tag{7.4}$$

the inequality being equality if and only if $ts + F'(x) = 0$ for some positive t.

Now comes the crucial step. Let us choose a positive μ and pass from the aggregate V_0 to the *potential*

$$V_\mu(x, s) = V_0(x, s) + \mu \ln(s^T x) \equiv F(x) + F^+(s) + (\vartheta + \mu) \ln(s^T x).$$

My claim is that this potential possesses the same fundamental property as the Karmarkar potential: *when it is small (i.e., negative with large absolute value) at a strictly feasible primal-dual pair (x, s), then the pair is comprised of good primal and dual approximate solutions.*

[1] by the way, this updating of the primal objective varies it by a constant (it is an immediate consequence of the fact that s is dual feasible)

The reason for this claim is clear: before we had added to the aggregate V_0 the "penalty term" $\mu \ln(s^T x)$, the aggregate was below bounded, as it is said by (7.4); therefore *the only way for the potential to be small is to have small (negative of large modulus) value of the penalty term, which, in turn, may happen only when the duality gap (which at a primal-dual feasible pair (x, s) is exactly $s^T x$, see (7.2)) is close to zero.*

The quantitive expression of this observation is as follows:

Proposition 7.2.1 *For any strictly feasible primal-dual pair (x, s) one has*

$$\delta(x, s) \leq \Gamma \exp\{\frac{V_\mu(x, s)}{\mu}\}, \;\; \Gamma = \exp\{-\mu^{-1}\vartheta(\ln \vartheta - 1)\}. \tag{7.5}$$

The proof is immediate:

$$\ln \delta(s, x) = \ln(s^T x) = \frac{V_\mu(x, s) - V_0(x, s)}{\mu} \leq$$

[due to (7.4)]

$$\leq \frac{V_\mu(x, s)}{\mu} - \mu^{-1}\vartheta(\ln \vartheta - 1).$$

Thus, enforcing the potential to go to $-\infty$ along a sequence of strictly feasible primal-dual pairs, we enforce the sequence to converge to the primal-dual optimal set. Similarly to the method of Karmarkar, the essence of the matter is how to update a strictly feasible pair (x, s) into another strictly feasible pair (x^+, s^+) with "significantly less" value of the potential. This is the issue we come to.

7.3 The primal-dual updating

The question we address to in this section is:

given a strictly feasible pair (x, s), how to update it into a new strictly feasible pair (x^+, s^+) in a way which ensures "significant" progress in the potential V_μ?

It is natural to start with investigating possibilities to reduce the potential by changing one of our two - primal and dual - variables, not both of them simultaneously. Let us look what are our abilities to improve the potential by changing the primal variable.

The potential $V_\mu(y, v)$, regarded as a function of the primal variable, resembles the Karmarkar potential, and it is natural to improve it as it was done in the method of Karmarkar. There is, anyhow, important difference: the Karmarkar potential was constant along primal feasible rays, and in order to improve it, we first pass from the "unconvenient" fesible set K_p of the original primal problem to a more convenient set G (see Lecture 6), which is in fact the projective image of K_p. Now the potential is *not* constant along rays, and we should reproduce the Karmarkar construction in the actual primal feasible set. Well, there is nothing difficult in it. Let us write down the potential as the function of the primal variable:

$$v(y) \equiv V_\mu(y, s) = F(y) + \zeta \ln s^T y + const(s) : \mathrm{rint}\, K_p \to \mathbf{R},$$

where

$$\zeta = \vartheta + \mu, \;\; const(s) = F^+(s).$$

Now, same as in the method of Karmarkar, let us linearize the logarithmic term in $v(\cdot)$, i.e., form the function

$$v_x(y) = F(y) + \zeta \frac{s^T y}{s^T x} + const(x, s) : \mathrm{rint}\, K_p \to \mathbf{R}, \tag{7.6}$$

where, as it is immediately seen,

$$const(x, s) = const(s) + \zeta \ln s^T x - \zeta.$$

Same as in the Karmarkar situation, v_x is an upper bound for v:

$$v_x(y) \geq v(y), \; y \in \mathrm{rint}\, K_p; \;\;\; v_x(x) = v(x), \tag{7.7}$$

so that in order to update x into a new strictly feasible primal solution x^+ with improved value of the potential $v(\cdot)$, it suffices to improve the value of the upper bound $v_x(\cdot)$ of the potential. Now, v_x is the sum of a self-concordant barrier for the primal feasible set (namely, the restriction of F onto this set) and a linear form, and therefore it is self-concordant on the relative interior $\mathrm{rint}\, K_p$ of the primal feasible set; consequently, to decrease the function, we may use the damped Newton method. Thus, we come to the following

Rule 1. *In order to update a given strictly feasible pair (x, s) into a new strictly feasible pair (x', s) with the same dual component and with better value of the potential V_μ, act as follows:*
1) Form the "partially linearized" reduced potential $v_x(y)$ according to (7.6);
2) Update x into x' by damped Newton iteration applied to $v_x(\cdot)$, i.e.,
- compute the (reduced) Newton direction

$$e_x = \mathrm{argmin}\{h^T \nabla_y v_x(x) + \frac{1}{2} h^T \nabla_y^2 v_x(x) h \mid h \in L\} \tag{7.8}$$

and the (reduced) Newton decrement

$$\omega = \sqrt{-e_x^T \nabla_y v_x(x)}; \tag{7.9}$$

- set

$$x' = x + \frac{1}{1+\omega} e_x.$$

As we know from Lecture 2, the damped Newton step keeps the iterate within the domain of the function, so that $x' \in \mathrm{rint}\, K_p$, and decreases the function at least by $\rho(-\omega) \equiv \omega - \ln(1+\omega)$. This is the progress in v_x; from (7.7) it follows that the progress in the potential $v(\cdot)$, and, consequently, in V_μ, is at least the progress in v_x. Thus, we come to the following conclusion:

I. *Rule 1 transforms the initial strictly feasible primal-dual pair (x, s) into a new strictly feasible primal-dual pair (x', s), and the potential V_μ at the updated pair is such that*

$$V_\mu(x, s) - V_\mu(x', s) \geq \omega - \ln(1+\omega), \tag{7.10}$$

ω being the reduced Newton decrement given by (7.8) - (7.9).

Now, in the method of Karmarkar we proceeded by proving that the reduced Newton decrement *is not* small. This is not the case anymore; the quantity ω can be very close to zero or even equal to zero. What should we do in this unpleasant sutiation where Rule 1 fails? Here again our experience with the method of Karmarkar gives the answer. Look, the potential

$$V_\mu(y, s) = F(y) + F^+(s) + \zeta \ln s^T y$$

regarded as a function of the strictly feasible primal solution y is nothing but

$$F(y) + F^+(s) + \zeta \ln(c^T y - [c^T b - b^T s]),$$

since for primal-dual feasible (y, s) the product $s^T y$ is nothing but the duality gap $c^T y + b^T s - c^T b$ (Lecture 5). The duality gap is always nonnegative, so that the quantity

$$c^T b - b^T s$$

associated with a dual feasible s is a lower bound for the primal optimal value. Thus, the potential V_μ, regarded as a function of y, resembles the "local" potential used in the sliding objective version of the method of Karmarkar - the Karmarkar potential where the primal optimal value is replaced by its lower bound. Now, in the sliding objective version of the method of Karmarkar we also met with the situation when the reduced Newton decrement was small, and, as we remember, in this situation we were able to update the lower bound for the primal optimal value and thus got the possibility to go ahead. This is more or less what we are going to do now: we shall see in a while that *if ω turns out to be small, then* *there is a possibility to update the current dual strictly feasible solution s into a new solution s' of this type and to improve by this "significantly" the potential.*

To get the idea how to update the dual solution, consider the "worst" for Rule 1 case - the reduced Newton decrement ω is zero. What happens in this situation? The reduced Newton decrement is zero if and only if the gradient of v_x, taken at x along the primal feasible plane, is 0, or, which is the same, if the gradient taken with respect to the whole primal space is orthogonal to L, i.e., if and only if

$$F'(x) + \zeta \frac{s}{s^T x} \in L^{\perp}. \tag{7.11}$$

This is a very interesting relation. Indeed, let

$$s^* \equiv -\frac{s^T x}{\zeta} F'(x) \tag{7.12}$$

The above inclusion says that $-s^* + s \in L^{\perp}$, i.e., that $s^* \in s + L^{\perp}$; since $s \in c + L^{\perp}$, we come to the relation

$$s^* \equiv -\frac{s^T x}{\zeta} F'(x) \in c + L^{\perp}. \tag{7.13}$$

The latter relation says that the vector $-F'(x)$ can be normalized, by multiplication by a positive constant, to result in a vector s^* from the dual feasible plane. On the other hand, s^* belongs to the interior of the dual cone K^*, since $-F'(x)$ does (Proposition 5.3.3). Thus, in the case in question (when $\omega = 0$), *a proper normalization of the vector $-F'(x)$ gives us a new strictly feasible dual solution $s' \equiv s^*$.* Now, what happens with the potential when we pass from s to s^* (and do not vary the primal solution x)? The answer is immediate:

$$V_\mu(x,s) = V_0(x,s) + \mu \ln s^T x \geq \vartheta \ln \vartheta - \vartheta + \mu \ln s^T x;$$

$$V_\mu(x,s^*) = V_0(x,s^*) + \mu \ln (s^*)^T x = \vartheta \ln \vartheta - \vartheta + \mu \ln (s^*)^T x$$

(indeed, we know from (**) that $V_0(y,u) \geq \vartheta \ln \vartheta - \vartheta$, and that this inequality is an equality when $u = -tF'(y)$, which is exactly the case for the pair (x, s^*)). Thus, the progress in the potential is at least the quantity

$$\alpha = \mu[\ln s^T x - \ln (s^*)^T x] = \mu[\ln s^T x - \ln\left(\frac{s^T x}{\zeta}(-F'(x))^T x\right)] =$$

$$= \mu \ln \frac{\zeta}{(-F'(x))^T x} = \mu \ln \frac{\zeta}{\vartheta} = \mu \ln (1 + \frac{\mu}{\vartheta}) \tag{7.14}$$

(the second equality in the chain is (7.12), the fourth comes from the identity (5.5), see Lecture 5). Thus, we see that in the particular case $\omega = 0$ updating

$$(x,s) \mapsto (x, s^* = -\frac{s^T x}{\zeta} F'(x))$$

results in a strictly feasible primal-dual pair and decreases the potential at least by the quantity $\mu \ln(1 + \mu/\vartheta)$.

We have seen what to do in the case of $\omega = 0$, when Rule 1 does not work at all. This is unsifficient: we should understand also what to do when Rule 1 works, but works bad, i.e., when ω is small, although nonzero. But this is more or less clear: what is good for the limiting case $\omega = 0$, should work also when ω is small. Thus, we get an idea to use, in the case of small ω, the updating of the dual solution given by (7.12). This updating, anyhow, cannot be used directly, since in the case of positive ω it results in s^* which is *unfeasible* for the dual problem. Indeed, dual feasibility of s^* in the case of $\omega = 0$ was a consequence of two facts:

1. Inclusion $s^* \in \operatorname{int} K^*$ - since s^* is proportional, with negative coefficient, to $F'(x)$, and all vectors of this type do belong to int K^* (Proposition 5.3.3); the inclusion $s^* \in \operatorname{int} K^*$ is therefore completely independent of whether ω is large or small;

2. Inclusion $s^* \in c + L^{\perp}$. This inclusion came from (7.11), and it *does* use the hypothesis that $\omega = 0$ (and in fact is equivalent to this hypothesis).

Thus, we meet with the difficulty that 2. does not remain valid when ω is positive, although small. Ok, if the only difficulty is that s^* given by (7.12) does not belong to the dual feasible plane, we can *correct* s^* - replace it by a properly chosen projection s' of s^* onto the dual feasible plane. When $\omega = 0$, s^* is in the dual feasible plane and in the interior of the cone K^*; by continuity reasons, for small ω s^* is close to the dual feasible plane and the projection will be close to s^* and therefore, hopefully, will be still in the interior of the dual cone (so that s', which by construction is in the dual feasible plane, will be strictly dual feasible), and, besides this, the updating $(x,s) \mapsto (x,s')$ would result in "almost" the same progress in the potential as in the above case $\omega = 0$.

The outlined idea is exactly what we are going to use. The implementation of it is as follows.

Rule 2. *In order to update a strictly feasible primal-dual pair (x,s) into a new strictly feasible primal-dual pair (x,s'), act as follows. Same as in Rule 1, compute the reduced Newton direction e_x, the reduced Newton decrement ω and set*

$$s' = -\frac{s^T x}{\zeta}[F'(x) + F''(x)e_x]. \tag{7.15}$$

Note that in the case of $\omega = 0$ (which is equivalent to $e_x = 0$), updating (7.15) becomes exactly the updating (7.12). As it can be easily seen[2], *s' is the projection of s^* onto the dual feasible plane in the metric given by the Hessian $(F^+)''(s^*)$ of the dual barrier at the point s^**; in particular, *s' always* belong to the dual feasible plane, although not necesarily to the interior of the dual cone K^*; this latter inclusion, anyhow, for sure takes place if $\omega < 1$, so that in this latter case s' is strictly dual feasible. Moreover, in the case of small ω the updating given by Rule 2 decreases the potential "significantly", so that Rule 2 for sure works well when Rule 1 does not, and choosing the best of these two rules, we come to the updating which *always* works well.

The exact formulation of the above claim is as follows:

II. (i) *The point s' given by (7.15) always belongs to the dual feasible plane.*

(ii) *The point s' is in the interior of the dual cone K^* (and, consequently, is dual strictly feasible) whenever $\omega < 1$, and in this case one has*

$$V_\mu(x,s) - V_\mu(x,s') \geq \mu \ln \frac{\vartheta + \mu}{\vartheta + \omega\sqrt{\vartheta}} - \rho(\omega), \ \ \rho(r) = -\ln(1-r) - r, \tag{7.16}$$

and the progress in the potential is therefore positive for all small enough positive ω.

Proof.

1^0. By definition, e_x is the minimizer of the quadratic form

$$Q(h) = h^T[F'(x) + \gamma s] + \frac{1}{2}h^T F''(x)h,$$

$$\gamma = \frac{\zeta}{s^T x} \equiv \frac{\vartheta + \mu}{s^T x}, \tag{7.17}$$

over $h \in L$; note that

$$h^T[F'(x) + \gamma s] = h^T \nabla_y v_x(x), \ h \in L.$$

Writing down the optimality condition, we come to

$$F''(x)e_x + [F'(x) + \gamma s] \equiv \xi \in L^\perp; \tag{7.18}$$

multiplying both sides by $e_x \in L$, we come to

$$\omega^2 \equiv -e_x^T \nabla_y v_x(x) = -e_x^T[F'(x) + \gamma s] = e_x^T F''(x)e_x. \tag{7.19}$$

2^0. From (7.18) and (7.15) it follows that

$$s' \equiv -\frac{1}{\gamma}[F'(x) + F''(x)e_x] = s - \gamma^{-1}\xi \in s + L^\perp, \tag{7.20}$$

[2] we skip verification, since we do not use this fact; those interested can make the corresponding computation

and since $s \in c + L^{\perp}$ (recall that s is dual feasible), we conclude that $s' \in c + L^{\perp}$, as claimed in (i).

Besides this,

$$s^* = -\frac{1}{\gamma} F'(x) \tag{7.21}$$

(see (7.12), (7.17)), so that the equivalence in (7.20) says that

$$s' = s^* - \frac{1}{\gamma} F''(x) e_x. \tag{7.22}$$

3^0. Since $F^+(u) = F^*(-u)$ is ϑ-logarithmically homogeneous self-concordant barrier for K^* (Proposition 5.3.3), we have

$$(F^+)'(tu) = t^{-1}(F^+)'(u), \ \ u \in \operatorname{int} K, \ t > 0$$

(see (5.3), Lecture 5); differentiating in u, we come to

$$(F^+)''(tu) = t^{-2}(F^+)''(u).$$

Substituting $u = -F'(s)$ and $t = 1/\gamma$ and taking into account the relation between F^+ and the Legendre transformation F^* of the barrier F, we come to

$$(F^+)''(s^*) = \gamma^2 (F^+)''(-F'(x)) = \gamma^2 (F^*)''(F'(x)).$$

But F^* is the Legendre transformation of F, and therefore (see (L.3), Lecture 2)

$$(F^*)''(F'(x)) = [F''(x)]^{-1};$$

thus, we come to

$$(F^+)''(s^*) = \gamma^2 [F''(x)]^{-1}. \tag{7.23}$$

Combining this observation with relation (7.22), we come to

$$[s' - s^*]^T (F^+)''(s^*) [s' - s^*] = [F''(x)e_x]^T [F''(x)]^{-1} [F''(x)e_x] = e_x^T F''(x) e_x = \omega^2$$

(the concluding equality is given by (7.19)). Thus, we come to the following conclusion:

IIa. *The distance $|s' - s^*|_{F^+, s^*}$ between s^* and s' in the Euclidean metric given by the Hessian $(F^+)''(s^*)$ of the dual barrier F^+ at the point s^* is equal to the reduced Newton decrement ω. In particular, if this decrement is < 1, s' belongs to the centered at s^* open unit Dikin ellipsoid of the self-concordant barrier F^+ and, consequently, s' belongs to the domain of the barrier (**I.**, Lecture 2), i.e., to $\operatorname{int} K^*$. Since we already know that s' always belongs to the dual feasible plane (see 2^0), s' is strictly dual feasible whenever $\omega < 1$.*

We have proved all required in (i)-(ii), except inequality (7.16) related to the progress in the potential. This is the issue we come to, and from now on we assume that $\omega < 1$, as it is stated in (7.16).

4^0. Thus, let us look at the progress in the potential

$$\alpha = V_\mu(x, s) - V_\mu(x, s') = V_0(x, s) - V_0(x, s') - \mu \ln \frac{x^T s'}{x^T s}. \tag{7.24}$$

We have

$$V_0(x, s') = F(x) + F^+(s') + \vartheta \ln x^T s' = \left[F(x) + F^+(s^*) + \vartheta \ln x^T s^*\right]_1 +$$
$$+ \left[F^+(s') - F^+(s^*) + \vartheta \ln \frac{x^T s'}{x^T s^*}\right]_2; \tag{7.25}$$

since $s^* = -tF'(x)$ with some positive t, (**) says to us that

$$[\cdot]_1 = \vartheta \ln \vartheta - \vartheta. \tag{7.26}$$

Now, s', as we know from **IIa.**, is in the open unit Dikin ellipsoid of F^+ centered at s^*, and the corresponding local distance is equal to ω; therefore, applying the upper bound (2.4) from Lecture 2 (recall that F^+ is self-concordant), we come to

$$F^+(s') - F^+(s^*) \le [s' - s^*]^T (F^+)'(s^*) + \rho(\omega), \ \ \rho(r) = -\ln(1 - r) - r. \tag{7.27}$$

We have $s^* = -\gamma^{-1}F'(x)$, and since F^+ is ϑ-logarithmically homogeneous,

$$(F^+)'(s^*) = \gamma(F^+)'(-F'(x))$$

((5.3), Lecture 5); since $F^+(u) = F^*(-u)$, F^* being the Legendre transformation of F, we have

$$(F^+)'(-F'(x)) = -(F^*)'(F'(x)),$$

and the latter quantity is $-x$ ((L.2), Lecture 2). Thus,

$$(F^+)'(s^*) = -\gamma x.$$

Now, by (7.22) we have $s' - s^* = -\gamma^{-1}F''(x)e_x$, so that

$$[s' - s^*]^T(F^+)'(s^*) = x^T F''(x)e_x.$$

From this observation and (7.27) we conclude that

$$[\cdot]_2 \le x^T F'' e_x + \rho(\omega) + \vartheta \ln \frac{x^T s'}{x^T s^*},$$

which combined with (7.25) and (7.26) results in

$$V_0(x, s') \le \vartheta \ln \vartheta - \vartheta + x^T F''(x)e_x + \rho(\omega) + \vartheta \ln \frac{x^T s'}{x^T s^*}. \tag{7.28}$$

On the other hand, we know from (**) that $V_0(x, s) \ge \vartheta \ln \vartheta - \vartheta$; combining this inequality, (7.24) and (7.28), we come to

$$\alpha \ge -x^T F''(x)e_x - \rho(\omega) - \vartheta \ln \frac{x^T s'}{x^T s^*} - \mu \ln \frac{x^T s'}{x^T s}. \tag{7.29}$$

5^0. Now let us find appropriate representations for the inner products involved into (7.29). To this end let us set

$$\pi = -x^T F''(x)e_x. \tag{7.30}$$

In view of (7.22) we have

$$x^T s' = x^T s^* - \frac{1}{\gamma} x^T F''(x)e_x = x^T s^* + \frac{\pi}{\gamma}$$

and, besides this,

$$x^T s^* = -\frac{1}{\gamma} x^T F'(x) = \frac{\vartheta}{\gamma}$$

(see (7.21) and (5.5), Lecture 5). We come to

$$x^T s' = \frac{\vartheta + \pi}{\gamma},\ x^T s^* = \frac{\vartheta}{\gamma},$$

whence

$$\frac{x^T s'}{x^T s^*} = 1 + \frac{\pi}{\vartheta}, \tag{7.31}$$

and

$$\frac{x^T s'}{x^T s} = \frac{\vartheta + \pi}{\gamma x^T s} = \frac{\vartheta + \pi}{\vartheta + \mu} \tag{7.32}$$

(the concluding equality follows from the definition of γ, see (7.17)).

Substituting (7.31) and (7.32) into (7.29), we come to the following expression for the progress in potential:

$$\alpha \ge \pi - \rho(\omega) - \vartheta \ln\left(1 + \frac{\pi}{\vartheta}\right) - \mu \ln \frac{\vartheta + \pi}{\vartheta + \mu}. \tag{7.33}$$

Taking into account that $\ln(1 + z) \le z$, we derive from this inequality that

$$\alpha \ge \mu \ln \frac{\vartheta + \mu}{\vartheta + \pi} - \rho(\omega). \tag{7.34}$$

Our last task is to evaluate π, which is immediate:

$$|\pi| = |x^T F''(x)e_x| \leq \sqrt{x^T F''(x)x}\sqrt{e_x^T F''(x)e_x} \leq \omega\sqrt{\vartheta}$$

(we have used (7.19) and identity (5.5), Lecture 5). With this estimate we derive from (7.34) that

$$\alpha \geq \mu \ln \frac{\vartheta + \mu}{\vartheta + \omega\sqrt{\vartheta}} - \rho(\omega), \tag{7.35}$$

as claimed in **II.** ■

7.4 Overall complexity analysis

We have presented two rules - Rule 1 and Rule 2 - for updating a strictly feasible primal-dual pair (x, s) into a new pair of the same type. The first of the rules always is productive, although the progress in the potential for the rule is small when the reduced Newton decrement ω is small; the second of the rules, on contrary, is for sure productive when ω is small, although for large ω it may result in an unfeasible s'. And, of course, what we should do is to apply both of the rules and choose the best of the results. Thus, we come to the

Primal-Dual Potential Reduction method $PD(\mu)$**:**

form the sequence of strictly feasible primal-dual pairs (x_i, s_i)*, starting with the initial pair* $(x_0 = \widehat{x}, s_0 = \widehat{s})$ *(see* **A***), as follows:*

1) given (x_{i-1}, s_{i-1})*, apply to the pair Rules 1 and 2 to get the updated pairs* (x'_{i-1}, s_{i-1}) *and* (x_{i-1}, s'_{i-1})*, respectively.*

2) Check whether s'_{i-1} *is strictly dual feasible. If it is not the case, forget about the pair* (x_{i-1}, s'_{i-1}) *and set* $(x_i^+, s_i^+) = (x'_{i-1}, s_{i-1})$*, otherwise choose as* (x_i^+, s_i^+) *the best (with the smallest value of the potential* V_μ*) of the two pairs given by 1).*

3) The pair (x_i^+, s_i^+) *for sure is a strictly feasible primal-dual pair, and the value of the potential* V_μ *at the pair is less than at the pair* (x_{i-1}, s_{i-1})*. Choose as* (x_i, s_i) *an arbitrary strictly feasible primal-dual pair such that the potential* V_μ *at the pair is not greater than at* (x_i^+, s_i^+) *(e.g., set* $x_i = x_i^+$*,* $s_i = s_i^+$*) and loop.*

The method, as it is stated now, involves the parameter μ, which in principle can be chosen as an arbitrary positive real. Let us find out what is the reasonable choice of the parameter. To this end let us note that what we are intersted in is not the progress p in the potential V_μ per step, but the quantity $\beta = \pi/\mu$, since this is this ratio which governs the exponent in the accuracy estimate (7.5). Now, at a step it may happen that we are in the situation $\omega = O(1)$, say, $\omega = 1$, so that the only productive rule is Rule 1 and the progress in the potential, according to **I.**, is of order of 1, which results in $\beta = O(1/\mu)$. On the other hand, we may come to the situation $\omega = 0$, when the only productive rule is Rule 2, and the progress in the potential is $p = \mu \ln(1 + \mu/\vartheta)$, see (7.16), i.e., $\beta = \ln(1 + \mu/\vartheta)$. A reasonable choice of μ should balance the values of β for these two cases, which leads to

$$\mu = \kappa\sqrt{\vartheta},$$

κ being of order of 1. The complexity of the primal-dual method for this - "optimal" - choice of μ is given by the following

Theorem 7.4.1 *Assume that the primal-dual pair of conic problems* $(\mathcal{P})$*,* $(\mathcal{D})$ *(which satisfies assumption* **A***) is solved by the primal-dual potential reduction method associated with* ϑ*-logarithmically self-concordant primal and dual barriers* F *and* F^+*, and that the parameter* μ *of the method is chosen according to*

$$\mu = \kappa\sqrt{\vartheta},$$

with certain $\kappa > 0$. Then the method generates a sequence of strictly feasible primal-dual pairs (x_i, s_i), and the duality gap $\delta(x_i, x_i)$ (equal to the sum of residuals, in terms of the corresponding objectives, of the components of the pair) admits the following upper bound:

$$\delta(x_i, s_i) \le \mathcal{V} \exp\{-\frac{V_\mu(\widehat{x}, \widehat{s}) - V_\mu(x_i, s_i)}{\kappa\sqrt{\vartheta}}\} \le \mathcal{V} \exp\{-\frac{i\Omega(\kappa)}{\kappa\sqrt{\vartheta}}\}, \tag{7.36}$$

where

$$\Omega(\kappa) = \min\left\{1 - \ln 2; \inf_{0 \le \omega < 1} \max\{\omega - \ln(1+\omega); \kappa\ln(1+\kappa) - (\kappa-1)\omega + \ln(1-\omega)\}\right\} \tag{7.37}$$

is positive continuous function of $\kappa > 0$; the data-dependent scale factor $\mathcal{V}$ is given by

$$\mathcal{V} = \delta(\widehat{x}, \widehat{s}) \exp\{\frac{V_0(\widehat{x}, \widehat{s}) - [\vartheta \ln \vartheta - \vartheta]}{\kappa\sqrt{\vartheta}}\}. \tag{7.38}$$

In particular, the Newton complexity (# of iterations of the method) of finding ε-solutions to the primal and the dual problems does not exceed the quantity

$$\mathcal{N}_{\rm PrDl}(\varepsilon) \le O_\kappa(1)\sqrt{\vartheta} \ln\left(\frac{\mathcal{V}}{\varepsilon} + 1\right) + 1, \tag{7.39}$$

with the constant factor $O_\kappa(1)$ depending on κ only.

The proof is immediate. Indeed, we know from Proposition 7.2.1 that

$$\delta(x_i, s_i) \le \Gamma \exp\{\frac{V_\mu(x_i, s_i)}{\mu}\} = [\Gamma \exp\{\frac{V_\mu(\widehat{x}, \widehat{s})}{\mu}\}] \exp\{-\frac{V_\mu(\widehat{x}, \widehat{s}) - V_\mu(x_i, s_i)}{\mu}\},$$

which, after substituting the value of Γ from (7.5), results in the first inequality in (7.36), with $\mathcal{V}$ given by (7.38).

To prove the second inequality in (7.36), it suffices to demonstrate that the progress in the potential V_μ at a step of the method is at least the quantity $\Omega(\kappa)$ given by (7.37). To this end let us note that, by construction, this progress is *at least* the progress given by each of the rules 1 and 2 (when Rule 2 does not result in a strictly feasible dual solution, the corresponding progress is $-\infty$). Let ω be the reduced Newton decrement at the step in question. If $\omega \ge 1$, then the progress related to Rule 1 is at least $1 - \ln 2$, see **I.**, which clearly is $\ge \Omega(\kappa)$. Now consider the case when $\omega < 1$. Here both of the rules 1 and 2 are productive, and the corresponding reductions in the potential are, respectively,

$$p_1 = \omega - \ln(1+\omega)$$

(see **I.**) and

$$p_2 = \mu \ln \frac{\vartheta + \mu}{\vartheta + \omega\sqrt{\vartheta}} + \ln(1-\omega) + \omega = \kappa\sqrt{\vartheta} \ln \frac{1 + \kappa/\sqrt{\vartheta}}{1 + \omega/\sqrt{\vartheta}} + \ln(1-\omega) + \omega$$

(see **II.**). We clearly have

$$p_2 = \kappa\sqrt{\vartheta}\ln(1 + \kappa/\sqrt{\vartheta}) - \kappa\sqrt{\vartheta}\ln(1 + \omega/\sqrt{\vartheta}) + \ln(1-\omega) + \omega \ge$$

[since $\ln(1+z) \le z$]

$$\ge \kappa\sqrt{\vartheta}\ln(1 + \kappa/\sqrt{\vartheta}) - \kappa\omega + \ln(1-\omega) + \omega \ge$$

[since, as it is immediately seen, $z\ln(1 + a/z) \ge \ln(1+a)$ whenever $z \ge 1$ and $a > 0$]

$$\ge \kappa\ln(1+\kappa) - \kappa\omega + \ln(1-\omega) + \omega,$$

and we come to the inequality

$$\max\{p_1, p_2\} \ge \max\{\omega - \ln(1+\omega); \kappa\ln(1+\kappa) - (\kappa-1)\omega + \ln(1-\omega)\},$$

so that the progress in the potential in the case of $\omega < 1$ is at least the quantity given by (7.37).

The claim that the right hand side of (7.37) is a positive continuous function of $\kappa > 0$ is evidently true. The complexity bound (7.39) is an immediate consequence of (7.36). ■

7.5 Large step strategy

To conclude the presentation of the primal-dual method, let me briefly outline how one could exploit the advantages of the potential reduction nature of the method. Due to this nature, the only thing we are interested in is "significant" progress in the potential at a step, same as it was in the method of Karmarkar. In this latter method, the simplest way to get a better progress than that one given by the "default" theoretical step, was to perform linesearch in the direction of this default step and to find the best, in terms of the potenital, point in this direction. What is the analogy of linesearch for the primal-dual method? It is as follows. Applying Rule 1, we get certain primal feasible direction $x' - x$, which we can extend in the trivial way to a primal-dual feasible direction (i.e., a direction from $L \times L^{\perp}$) $d_1 = (x' - x, 0)$; shifting the current strictly feasible pair (x, s) in this direction, we for sure get a strictly feasible pair with better (or, in the case of $\omega = 0$, the same) value of the potential. Similraly, applying Rule 2, we get another primal-dual feasible direction $d_2 = (0, s' - s)$; shifting the current pair in this direction, we always get a pair from the primal-dual feasible plane $\mathcal{L} = \{b + L\} \times \{c + L^{\perp}\}$, although not necessarily belonging to the interior of the primal-dual cone $\mathcal{K} = K \times K^*$, What we *always* get, is certain 2-dimensional plane D (passing through (x, s) parallel to the directions d_1, d_2) which is contained in the primal-dual feasible plane $\mathcal{L}$, and one (or two, depending on whether Rule 2 was or was not productive) strictly feasible primal-dual pairs - candidates to the role of the next iterate; what we know from our theoretical analysis, is that the value of the potential at one of the candidate pairs is "significantly" - at least by the quantity $\Omega(\kappa)$ - less that the value of the potential at the previous iterate (x, s). Given this situation, a resonable policy to get additional progress in the potential at the step is 2-dimensional minimization of the potential over the intersection of the plane D with the interior of the cone $K \times K^*$. The potential is *not* convex, and it would be difficult to ensure a prescribed quality of its minimization even over the 2-dimensional plane D, but this is not the point where we must get a good minimizer; for our purposes it suffices to perform a once for ever fixed (and small) number of steps of any relaxation method for smooth minimization (the potential is smooth), running the method from the best of our candidate pairs. In the case of LP, same as in some other interesting cases, there are possibilities to implement this 2-dimensional search in a way which almost does not increase the total computational effort per step[3], and at the same time accelerates the method dramatically.

[3] this total effort normally is dominated by the cost of computing the reduced Newton direction e_x

7.6 Exercises: Primal-Dual method

The subject of the forthcoming problems is *implementation of the primal-dual method.* We shall start with some remarks related to the general situation and then consider a particular problem coming from Control.

When speaking about implementation, i.e., about algorithmical issues, we should, of course, fix somehow the way the data are represented; for a conic problem, this is, basically, the question of how the feasible subspace L is described. In most of applications known to me the situation is as follows. $b+L \subset \mathbf{R}^n$ is defined as the image of certain subspace

$$\{\xi \in \mathbf{R}^l \mid P(\xi - p) = 0\}$$

(ξ is the vector of the design variables) under a given affine mapping

$$x = \mathcal{A}(\xi) \equiv A\xi + b,$$

A being $n \times l$ and P being $k \times l$ matrices; usually one can assume that A is of full column rank, i.e., that its columns are linearly independent, and that P is of full row rank, i.e., the rows of P are linearly independent; from now on we make this regularity assumption. As far as the objective is concerned, it is a linear form $\chi^T\xi$ of the design vector.

Thus, the typical for applications form of the primal problem is

$$(\mathrm{P}): \;\; \textit{minimize} \;\; \chi^T\xi \;\; \textit{s.t.} \;\; \xi \in \mathbf{R}^l, \; P(\xi - p) = 0, \;\; x \equiv A\xi + b \in K,$$

K being a pointed closed and convex cone with a nonempty interior in $\mathbf{R}^n$. This is exactly the setting presented in Section 5.4.4.

As we know from Exercise 5.4.11, the problem dual to (P) is

$$(\mathrm{D}): \;\; \textit{minimize} \;\; \beta^T s \;\; \textit{s.t.} \;\; A^T s = \chi + P^T r, \;\; s \in K^*,$$

where the control vector is comprised of $s \in \mathbf{R}^n$ and $r \in \mathbf{R}^k$, K^* is the cone dual to K, and $\beta = \mathcal{A}(p)$.

In what follows F denotes the primal barrier - ϑ-logarithmically homogeneous self-concordant barrier for K, and F^+ denotes the dual barrier (see Lecture 7).

Let us look how the primal-dual method could be implemented in the case when the primal-dual pair of problems is in the form (P) - (D). We should answer the following basic questions

- how to represent the primal and the dual solutions;
- how to perform the updating $(x_i, s_i) \mapsto (x_{i+1}, s_{i+1})$.

As far as the first of this issues is concerned, the most natural decision is

to represent x's of the form $\mathcal{A}(\xi)$ (note that all our primal feasible x's are of this type) by storing both x (as an n-dimensional vector) and ξ (as an l-dimensional one);

to represent s's and r's "as they are" - as n- and k-dimensional vectors, respectively.

Now, what can be said about the main issue - how to implement the updating of strictly feasible primal-dual pairs? In what follows we speak about the basic version of the method only, not discussing the large step strategy from Section 7.5, since implementation of the latter strategy (and even the possibility to implement it) heavily depends on the specific analytic structure of the problem.

Looking at the description of the primal-dual method, we see that the only nontrivial issue is how to compute the Newton direction

$$e_x = \mathrm{argmin}\{h^T g + \frac{1}{2} h^T F''(x) h \mid h \in L\},$$

where (x, s) is the current iterate to be updated and $g = F'(x) + \frac{\vartheta+\mu}{s^T x} s$. Since L is the image of the linear space

$$L' = \{\zeta \in \mathbf{R}^l \mid P\zeta = 0\}$$

under the mapping $\zeta \mapsto A\zeta$, we have

$$e_x = A\eta_x$$

for certain $\eta_x \in L'$, and the problem is how to compute η_x.

Exercise 7.6.1 # *Prove that η_x is uniquely defined by the linear system of equations*

$$\begin{pmatrix} Q & P^T \\ P & 0 \end{pmatrix} \begin{pmatrix} \eta \\ u \end{pmatrix} = \begin{pmatrix} -q \\ 0 \end{pmatrix} \tag{7.40}$$

where

$$Q = A^T F''(x) A, \quad q = A^T g, \tag{7.41}$$

so that η_x is given by the relation

$$\eta_x = -Q^{-1} \left[A^T g - P^T [PQ^{-1}P^T]^{-1} PQ^{-1} A^T g \right]; \tag{7.42}$$

in the particular case when P is absent (formally, $k = 0$), η_x is given by

$$\eta_x = -Q^{-1} A^T g. \tag{7.43}$$

Note that normally k is a small integer, so that the main effort in computing η_x is to assemble and to invert the matrix Q. Usually this is the main part of the overall effort per iteration, since other actions, like computing $F(x)$, $F'(x)$, $F''(x)$, are relatively cheap.

7.6.1 Example: Lyapunov Stability Analysis

The goal of the forthcoming exercises is to develop the (principal elements of) algorithmic scheme of the primal-dual method as applied to the following interesting and important problem coming from Control theory:

(C) *given a "polytopic" linear time-varying ν-dimensional system*

$$v'(t) = V(t)v(t), \;\; V(t) \in \mathrm{Conv}\{V_1, ..., V_m\},$$

find a quadratic Lyapunov function $v^T L v$ which demonstrates stability of the system.

Let us start with explaining what we are asked to do. The system in question is a time-varying linear dynamic system with uncertainty: $v(t)$ is ν-dimensional vector-function of time t - *the trajectory*, and $V(t)$ is the time-varying matrix of the system. Note that we do not know in advance what this matrix is; all we know is that, for every t, the matrix $V(t)$ belongs to the convex hull of a given finite set of matrices V_i, $i = 1, ..., m$.

Now, the system in question is called *stable*, if $v(t) \to 0$ as $t \to \infty$ for all trajectories. A good *sufficient* condition for stability is the existence of a positive definite *quadratic Lyapunov function $v^T L v$* for the system, i.e., a positive definite symmetric $\nu \times \nu$ matrix L such that the derivative in t of the quantity $v^T(t) L v(t)$ is strictly negative for every t and every trajectory $v(t)$ with nonzero $v(t)$. This latter requirement, in view of $v'(t) = V(t)v(t)$, is equivalent to

$$[V(t)v(t)]^T L v(t) < 0 \;\; \textit{whenever} \;\; v(t) \neq 0 \;\; \textit{and} \;\; V(t) \in \mathrm{Conv}\{V_1, ..., V_m\},$$

or, which is the same (since for a given t $v(t)$ can be an arbitrary vector and $V(t)$ can be an arbitrary matrix from $\mathrm{Conv}\{V_1, ..., V_m\}$), is equivalent to the requirement

$$v^T V^T L v = \frac{1}{2} v^T [V^T L + LV] v < 0, \;\; v \neq 0, \, V \in \mathrm{Conv}\{V_1, ..., V_m\}.$$

In other words, L should be a positive definite symmetric matrix such that all the matrices of the form $V^T L + LV$ associated with $V \in \mathrm{Conv}\{V_1, ..., V_m\}$ are negative definite; matrix L with these properties will be called *appropriate*.

Our first (and extremely simple) task is to characterize the appopriate matrices.

Exercise 7.6.2 # *Prove that a symmetric $\nu \times \nu$ matrix L is appropriate if and only if it is positive definite and the matrices*

$$V_i^T L + L V_i, \;\; i = 1, ..., m$$

are negative definite.

We see that to find an appropriate matrix (and to demonstrate by this stability of (C) via a quadratic Lyapunov function) is the same as to find a solution to the following system of *strict matrix inequalities*

$$L > 0;\ V_i^T L + LV_i < 0,\ i = 1, ..., m, \tag{7.44}$$

where inequalities with symmetric matrices are understood as positive definiteness (for strict inequalities) or semidefiniteness (for non-strict ones) of the corresponding differences.

We can immediately pose our problem as a conic problem with trivial objective; to this end it suffices to treat L as the design variable (which varies over the space $\mathbf{S}^\nu$ of symmetric $\nu \times \nu$ matrices) and introduce the linear mapping

$$\mathcal{B}(L) = \mathrm{Diag}\{L; -V_1^T L - LV_1; ...; -V_m^T L - LV_m\}$$

from this space into the space $(\mathbf{S}^\nu)^{m+1}$ - the direct product of $m+1$ copies of the space $\mathbf{S}^\nu$, so that $(\mathbf{S}^\nu)^{m+1}$ is the space of symmetric block-diagonal $[(m+1)\nu] \times [(m+1)\nu]$ matrices with $m+1$ diagonal blocks of the size $\nu \times \nu$ each. Now, $(\mathbf{S}^\nu)^{m+1}$ contains the cone $\mathcal{K}$ of positive semidefinite matrices of the required block-diagonal structure; it is clearly seen that L is appropriate if and only if $\mathcal{B}(L) \in \mathrm{int}\,\mathcal{K}$, so that the set of appropriate matrices is the same as the set of strictly feasible solutions to the conic problem

$$\textit{minimize}\ \ 0\ \ \textit{s.t.}\ \ \mathcal{B}(L) \in \mathcal{K}$$

with trivial objective.

Thus, the problem in question is reduced to a conic problem involving the cone of positive semidefinite matrices of certain block-diagonal structure; the problems of this type are called *semidefinite programs* or *optimization under LMI's* (Linear Matrix Inequality constraints).

Of course, we could try to solve the problem by an interior point potential reduction method known to us, say, by the method of Karmarkar or by the primal-dual method; we immdeiately discover, anyhow, that the technique developed so far cannot be applied to our problem - indeed, in all methods known to us it was required *at least* to know in advance a strictly feasible solution to the problem, and in our particular case such a solution is exactly what should be finally found. There is, anyhow, a straightforward way to avoid the difficulty. First of all, our system (7.44) is homogeneous in L; therefore we can normalize L to be $\leq I$ (I stands for the unit matrix of the context-determined size) and pass from the initial system to the new one

$$L > 0;\ \ L \leq I;\ \ V_i^T L + LV_i < 0,\ i = 1, ..., m. \tag{7.45}$$

Now let us extend our design vector L by one variable t, so that the new design vector becomes

$$\xi = (t, L) \in E \equiv \mathbf{R} \times \mathbf{S}^n,$$

and consider the semidefinite program

$$\textit{minimize}\ \ t\ \ \textit{s.t.}\ \ L + tI \geq 0;\ \ I - L \geq 0;\ \ tI - V_i^T L - LV_i \geq 0,\ i = 1, ..., m. \tag{7.46}$$

Clearly, to solve system (7.45) is the same as to find a feasible solution to optimization problem (7.46) with *negative* value of the objective; on the other hand, in (7.46) we have no difficulties with an initial strictly feasible solution: we may set $L = \frac{1}{2}I$ and then choose t large enough to make all remaining inequalities strict.

It is clear that (7.46) is of the form (P) with the data given by the affine mapping

$$\mathcal{A}(\xi) \equiv \mathcal{A}(t, L) = \mathrm{Diag}\{L + tI; I - L; tI - V_1^T L - LV_1; ...; tI - V_m^T L - LV_m\} : E \to \mathcal{E},$$

$\mathcal{E}$ being the space $(\mathbf{S}^\nu)^{m+2}$ of block-diagonal symmetric matrices with $m+2$ diagonal blocks of the size $\nu \times \nu$ each; the cone K in our case is the cone of positive semidefinite matrices from $\mathcal{E}$, and matrix P is absent, so that our problem is

$$(\mathrm{Pr})\ \ \textit{minimize}\ \ t\ \ \textit{s.t.}\ \ \ \mathcal{A}(t, L) \in K.$$

Now let us form the method.

Exercise 7.6.3 $^{\#+}$ *Prove that*

1) the cone K is self-dual;

2) the function

$$F(x) = -\ln \mathrm{Det}\ x$$

is a $(m+2)\nu$-logarithmically homogeneous self-concordant barrier for the cone K;

3) the dual barrier F^+ associated with the barrier F is, up to an additive constant, the barrier F itself:

$$F^+(s) = -\ln \mathrm{Det}\ s - (m+2)\nu.$$

Thus, we are equipped with the primal and the dual barriers required to solve (Pr) via the primal-dual method. Now let us look what the method is. First of all, what is the dual to (Pr) problem (Dl)?

Exercise 7.6.4 $^{\#}$ *Prove that when the primal problem (P) is specified to be (Pr), the dual problem (D) becomes*

(Dl) *minimize* $\mathrm{Tr}\{s_0\}$ *under choice of* $m+2$ *symmetric* $\nu \times \nu$ *matrices* $s_{-1}, ..., s_m$ *s.t.*

$$s_{-1} - s_0 - \sum_{i=1}^{m} [V_i s_i + s_i V_i^T] = 0;$$

$$\mathrm{Tr}\{s_{-1}\} + \sum_{i=1}^{m} \mathrm{Tr}\{s_i\} = 1.$$

It is time now to think of the initialization. Could we in fact point out strictly feasible solutions $\widehat{x}$ and $\widehat{s}$ to the primal and to the dual problems? As we just have mentioned, as far as the primal problem (Pr) is concerned, there is nothing to do: we can set

$$\widehat{x} = \mathcal{A}(\widehat{t}, \widehat{L}),$$

where $\widehat{L}$ is $< I$, e.g., $\widehat{L} = \frac{1}{2}I$, and $\widehat{t}$ is large enough to ensure that $L + \widehat{t}I > 0$, $\widehat{t}I > V_i^T \widehat{L} + \widehat{L}V_i$, $i = 1, ..., m$.

Exercise 7.6.5 $^{\#}$ *Point out a strictly feasible solution $\widehat{s}$ to (Dl).*

It remains to realize what are the basic operations at a step of the method.

Exercise 7.6.6 $^{\#}$ *Verify that in the case in question the quantities involved into the description of the primal-dual method can be specified as follows:*

1) The quantities related to F are given by

$$F'(x) = -x^{-1}; \;\; F''(x)h = x^{-1}hx^{-1};$$

2) The matrix Q involved into the system for finding η_x (see Exercise 7.6.1), taken with respect to certain orthonormal basis $\{e_\alpha\}_{\alpha=1,\dots,N}$ in the space E, is given by

$$Q_{\alpha\beta} = \mathrm{Tr}\{A_\alpha x^{-1} A_\beta x^{-1}\}, \quad A_\alpha = Ae_\alpha.$$

Think about the algorithmic implementation of the primal-dual method and, in particular, about the following issues:

- *What is the dimension N of the "design space" E? What is the dimension M of the "image space" $\mathcal{E}$?*

- *How would you choose a "natural" orthonormal basis in E?*

- *Is it necessary/reasonable to store $F''(x)$ as an $M \times M$ square array? How to assemble the matrix Q? What is the arithmetic cost of the assembling?*

- *Is it actually necessary to invert Q explicitly? Which method of Linear Algebra would you choose to solve system (7.40)?*

- *What is the arithmetic cost of the step in the basic version of the primal-dual method? Where the dominating expenses come from?*
- *Are there ways to implement at a relatively low cost a large step strategy? How would you do it?*
- *When would you terminate the computations? How could you recognize that the optimal value in the problem is positive, so that you are unable to find a quadratic Lyapunov function which proves the stability? Is it possible that running the method you never will be able neither to present an appropriate L nor to come to the conclusion that it does not exist?*

Last exercise is as follows:

Exercise 7.6.7 $^{\#*}$ *Is it reasonable to replace (Pr) by "less redundant" problem*

$$(\mathrm{Pr}')\quad \textit{minimize}\ \ t\ \ \textit{s.t.}\ \ L \geq I; tI - V_i^T L - LV_i \geq 0,\ i = 1, ..., m$$

(here we normalize L in (7.44) by $L \geq I$ and, same as in (Pr), add the "slack" variable t to make the problem "evidently feasible")?

Chapter 8

Long-Step Path-Following Methods

To the moment we are acquainted with three particular interior point algorithms, namely, with the short-step path-following method and with two potential reduction algorithms. As we know, the main advantage of the potential reduction scheme is not of theoretical origin (in fact one of the potential reduction routines, the method of Karmarkar, is even worse theoretically than the path-following algorithm), but in possibility to implement ”long step” tactics. Recently it became clear that such a possibility also exists within the path-following scheme; and the goal of this lecture is to present to you the ”long step” version of the path-following method.

8.1 The predictor-corrector scheme

Recall that in the path-following scheme (Lecture 4) we were interested in the problem

$$\textit{minimize}\ \ c^T x\ \ s.t.\ \ x \in G, \tag{8.1}$$

G being a closed and bounded convex domain in $\mathbf{R}^n$. In order to solve the problem, we take a ϑ-self-concordant barrier F for the feasible domain G and trace the *path*

$$x^*(t) = \operatorname*{argmin}_{x \in \mathrm{int}\ G} F_t(x),\ \ F_t(x) = tc^T x + F(x), \tag{8.2}$$

as the penalty parameter t tends to infinity. More specifically, we generate a sequence of pairs (t^i, x^i) κ-*close* to the path, i.e., satisfying the predicate

$$\{t > 0\} \,\&\, \{x \in \mathrm{int}\ G\} \,\&\, \{\lambda(F_t, x) \equiv \sqrt{[\nabla_x F_t(x)]^T \nabla_x^2 F_t(x) \nabla_x F_t(x)} \le \kappa\}, \tag{8.3}$$

the *path tolerance* $\kappa < 1$ being the parameter of the method. The policy of tracing the path in the basic scheme of the method was very simple: in order to update $(t, x) \equiv (t^{i-1}, x^{i-1})$ into $(t^+, x^+) = (t^i, x^i)$, we first increased, in certain prescribed ratio, the value of the penalty, i.e., set

$$t^+ = t + \delta t,\ \ dt = \frac{\gamma}{\sqrt{\vartheta}} t, \tag{8.4}$$

and then applied to the new function $F_{t^+}(\cdot)$ the damped Newton method in order to update x into x^+:

$$y^{l+1} = y^l - \frac{1}{1 + \lambda(F_{t^+}, y^l)} [\nabla_x^2 F(y^l)]^{-1} \nabla_x F_{t^+}(y^l); \tag{8.5}$$

we initialized this reccurency by setting $y^0 = x$ and terminated it when the closeness to the path was restored, i.e., when $\lambda(F_{t^+}, y^l)$ turned out to be $\le \kappa$, and took the corresponding y^l as x^+.

Looking at the scheme, we immediately see at least two weak points of it: first, we use a once for ever fixed penalty rate and do not try to use larger dt's; second, when applying the damped Newton method to the function F_{t^+}, we start the reccurency at $y^0 = x$; why do not we use a better forecast for our target point $x^*(t + dt)$? Let us start with discussing this second point. The path $x^*(\cdot)$ is smooth (at least two times continuously differentiable), as it is immediately seen from the Implicit Function Theorem applied to the equation

$$tc + F'(x) = 0 \tag{8.6}$$

which defines the path. Given a tight approximation x to the point $x^*(t)$ of the path, we could try to use the first-order prediction

$$x^f(dt) = x + x'dt$$

of our target point $x^*(t+dt)$; here x' is some approximation of the derivative $\frac{d}{dt}x^*(\cdot)$ at the point t. The simplest way to get this approximation is to note that what we finally are interested in is to solve with respect to y the equation

$$(t + dt)c + F'(y) = 0;$$

a good idea is to linearize the left hand side at $y = x$ and to use, as the forecast of $x^*(t+dt)$, the solution to the linearized equation. The linearized equation is

$$(t + dt)c + F'(x) + F''(x)[y - x] = 0,$$

and we come to

$$dx(dt) \equiv y - x = -[F''(x)]^{-1}\nabla_x F_{t+dt}(x). \tag{8.7}$$

Thus, it is reasonable to start the damped Newton method with the forecast

$$x^f(dt) \equiv x + dx(dt) = x - [F''(x)]^{-1}\nabla_x F_{t+dt}(x). \tag{8.8}$$

Note that in fact we do not get anything significantly new: $x^f(dt)$ is simply the Newton (not the *damped* Newton) iterate of x with respect to the function $F_{t^+}(\cdot)$; nevertheless, this is not exactly the same as the initial implementation. The actual challenge is, of course, to get rid of the once for ever fixed penalty rate. To realize what could be done here, let us write down the generic scheme we came to:

Predictor-Corrector Updating scheme:
in order to update a given κ-close to the path $x^*(\cdot)$ pair (t, x) into a new pair (t^+, x^+) of the same type, act as follows

- Predictor step:

  1) form the *primal search line*

$$P = \{X(dt) = (t + dt, x + dx(dt) \mid dt \in \mathbf{R}\}, \tag{8.9}$$

  $dx(dt)$ being given by (8.7);

  2) choose stepsize $\delta t > 0$ and form the *forecast*

$$t^+ = t + \delta t, \; x^f = x + dx(\delta t); \tag{8.10}$$

- Corrector step:

  3) starting with $y^0 = x^f$, run the damped Newton method (8.5) until $\lambda(t^+, y^l)$ becomes $\leq \kappa$; when it happens, set $x^+ = y^l$, thus completing the updating $(t, x) \mapsto (t^+, x^+)$.

Now let us look what are the stepsizes δt acceptable for us. Of course, there is an immediate requirement that $x^f = x + dx(\delta t)$ should be strictly feasible - otherwise we simply will be unable to start the damped Newton method with x^f. There is, anyhow, a more severe restriction. Remember that the complexity estimate for the method in question heavily depended on the fact that the "default" stepsize (8.4) results in a once for ever fixed (depending on the penalty rate γ and the path tolerance κ only) Newton complexity of the corrector step. If we wish to preserve the complexity bounds - and we do wish to preserve them - we should take care of fixed Newton complexity of the corrector step. Recall that our basic results on the damped Newton method as applied to the self-concordant function $F_{t^+}(\cdot)$ (**X.**, Lecture 2) say that the number of Newton iterations of the method, started at certain point $y^0 \in \operatorname{int} G$ and runned until the relation $\lambda(F_{t^+}, y^l) \leq \kappa$ becomes true, is bounded from above by the quantity

$$O(1)\left\{[F_{t^+}(y^0) - \min_{y \in \operatorname{int} G} F_{t^+}(y)] + \ln(1 + \ln\frac{1}{\kappa})\right\},$$

$O(1)$ being an appropriate absolute constant. We see that in order to bound from above the Newton complexity of the corrector step it suffices to bound from above the residual

$$V(t^+, x^f) \equiv F_{t^+}(x^f) - \min_{y \in \mathrm{int}\, G} F_{t^+}(y),$$

i.e., to choose the stepsize δt in a way which ensures that

$$V(t + \delta t, x^f(\delta t)) \le \overline{\kappa}, \tag{8.11}$$

where $\overline{\kappa}$ is a once for ever fixed constant - the additional to the path tolerance κ parameter of the method. The problem, of course, is how to ensure (8.11). If it would be easy to compute the residual at a given pair (t^+, x^f), we could apply a linesearch in the stepsize δt in order to choose the largest stepsize compatible with a prescribed upper bound on the residual. Given a candidate stepsize δt, we normally have no problems with "cheap" computation of t^+, x^f and the quantity $F_{t^+}(x^f)$ (usually the cost of computing the value of the barrier is much less than our natural "complexity unit" - the arithmetic cost of a Newton step); the difficulty, anyhow, is that the residual invloves not only the value of F_{t^+} at the forecast, but also the unknown to us minimum value of $F_{t^+}(\cdot)$. What we are about to do is to derive certain duality-based and computationally cheap *lower bounds* for the latter minimum value, thus obtaining "computable" upper bounds for the residual.

8.2 Dual bounds and Dual search line

From now on, let us make the following *Structural assumption* on the barrier in question:

$\mathcal{Q}$: *the barrier F is of the form*

$$F(x) = \Phi(\pi x + p), \tag{8.12}$$

where Φ is a ϑ-self-concordant nondegenerate barrier for certain closed convex domain $G^+ \subset \mathbf{R}^m$ with known Legendre transformation Φ^* and $x \mapsto \pi x + p$ is an affine mapping from $\mathbf{R}^n$ into $\mathbf{R}^m$ with the image intersecting int G^+*, so that G is the inverse image of G^+ under the mapping $x \mapsto \pi x + p$.*

Note that $\mathcal{Q}$ indeed defines a ϑ-self-concordant barrier for G, see Proposition 3.1.1.(i).

Note that the essence of the Structural assumption is that we know the Legendre transformation of Φ (otherwise there would be no assumption at all - we simply could set $\Phi \equiv F$). This assumption indeed is satisfied in many important cases, e.g., in Linear Programming, where G is a polytope given by linear inequalities $a_i^T x \le b_i$, $i = 1, ..., m$, and

$$F(x) = -\sum_{i=1}^{m} \ln(b_i - a_i^T x);$$

here

$$G^+ = \mathbf{R}^m_+, \ \Phi(u) = -\sum_{i=1}^{m} \ln u_i$$

and

$$(\pi x + p)_i = b_i - a_i^T x, \ i = 1, ..., m;$$

the Legendre transformation of Φ, as it is immediately seen, is

$$\Phi^*(s) = \Phi(-s) - m, \ s \in \mathbf{R}^m_-.$$

In the mean time we shall speak about other important cases where the assumption is valid.

Now let us make the following simple and crucial observation:

Proposition 8.2.1 *Let a pair $(\tau, s) \in \mathbf{R}_+ \times \mathrm{Dom}\,\Phi^*$ satisfy the linear homogeneous equation*

$$\tau c + \pi^T s = 0. \tag{8.13}$$

Then the quantity

$$f_s(\tau) = p^T s - \Phi^*(s) \tag{8.14}$$

is a lower bound for the quantity

$$f_*(\tau) = \min_{y\in \mathrm{int}\, G} F_\tau(y)$$

and, consequently, the quantity

$$V_s(\tau, y) = F_\tau(y) - f_s(\tau) \equiv \tau c^T y + F(y) + \Phi^*(s) - p^T s \tag{8.15}$$

is an upper bound for the residual

$$V(\tau, y) = F_\tau(y) - \min F_\tau(\cdot).$$

Proof. As we know from **VII.**, Lecture 2, the Legendre transformation of Φ^* is exactly Φ. Consequently,

$$\Phi(\pi y + p) = \sup_{v\in \mathrm{Dom}\,\Phi^*} [[\pi y + p]^T v - \Phi^*(v)] \geq [\pi y + p]^T s - \Phi^*(s),$$

whence

$$F_\tau(y) \equiv \tau c^T y + F(y) \equiv \tau c^T y + \Phi(\pi y + p) \geq$$
$$\geq \tau c^T y + [\pi y + p]^T s - \Phi^*(s) = [\tau c + \pi^T s]^T y + p^T s - \Phi^*(s) = p^T s - \Phi^*(s)$$

(the concluding inequality follows from (8.13)). ■

Our next observation is that there exists a systematic way to generate *dual feasible* pairs (τ, s), i.e., the pairs satisfying the premise of the above proposition.

Proposition 8.2.2 *Let (t, x) be a primal feasible pair (i.e., with $t > 0$ and $x \in \mathrm{int}\, G$), and let*

$$u = \pi x + p,\ du(dt) = \pi dx(dt),\ s = \Phi'(u),\ ds(dt) = \Phi''(u)du(dt), \tag{8.16}$$

where $dx(dt)$ is given by (8.7). Then

(i) *Every pair $S(dt)$ on the Dual search line*

$$D = \{S(dt) = (t + dt, s^f(dt) = s + ds(dt)) \mid dt \in \mathbf{R}\}$$

satisfies equation (8.13).

(ii) *If (t, x) is κ-close to the path, then the pair $S(0)$, and, consequently, every pair $S(dt)$ with small enough $|dt|$, belongs to the domain of Φ^* and is therefore dual feasible.*

Proof.

(i): from (8.16) it follows that

$$(t + dt)c + \pi^T(s + ds(dt)) = (t + dt)c + \pi^T [\Phi'(u) + \Phi''(u)\pi dx(dt)] =$$

[since $F'(x) = \pi^T\Phi'(u)$ and $F''(x) = \pi^T\Phi''(u)\pi$ in view of (8.12) and (8.16)]

$$= (t + dt)c + F'(x) + F''(x)dx(dt) = \nabla_x F_{t+dt}(x) + F''(x)dx(dt),$$

and the concluding quantity is 0 due to the origin of $dx(dt)$, see (8.7). (i) is proved.

(ii): let us start with the following simple

Lemma 8.2.1 *One has*

$$|ds(dt)|^2_{(\Phi^*)''(s)} = |du(dt)|^2_{\Phi''(u)} = [du(dt)]^T ds(dt) \tag{8.17}$$

and

$$|ds(0)|_{(\Phi^*)''(s)} = |dx(0)|_{F''(x)} = \lambda^2(F_t, x). \tag{8.18}$$

Proof. Since $s = \Phi'(u)$ and Φ^* is the Legendre transformation of Φ, we have

$$(\Phi^*)''(s) = [\Phi''(u)]^{-1} \tag{8.19}$$

(see (L.3), Lecture 2). Besides this, $ds(dt) = \Phi''(u)du(dt)$ by (8.16), whence

$$|ds(dt)|^2_{(\Phi^*)''} \equiv [ds(dt)]^T[(\Phi^*)''][ds(dt)] = [\Phi''du(dt)]^T[\Phi'']^{-1}[\Phi''du(dt)] =$$

$$= [du(dt)]^T[\Phi''][du(dt)],$$

as claimed in the first equality in (8.17); the second inequality there is an immediate consequence of $ds(dt) = [\Phi'']du(dt)$.

To prove (8.18), note that, as we know from (8.17), $|ds(0)|^2_{(\Phi^*)''} = |du(0)|^2_{\Phi''}$; the latter quantity, in view of (8.16), is nothing but $[\pi dx(0)]^T\Phi''[\pi dx(0)]$, which, in turn, equals to $|dx(0)|^2_{F''(x)}$ in view of $F''(x) = \pi^T\Phi''(u)\pi$. We have proved the first equality in (8.18); the second is immdeiate, since $dx(0) = -[F''(x)]^{-1}\nabla_x F_t(x)$ by (8.7), and, consequently,

$$|dx(0)|^2_{F''(x)} = \left[[F''(x)]^{-1}\nabla_x F_t(x)\right]^T [F''(x)] \left[[F''(x)]^{-1}\nabla_x F_t(x)\right] =$$

$$= [\nabla_x F_t(x)]^T[F''(x)]^{-1}\nabla_x F_t(x) \equiv \lambda^2(F_t, x).$$

■

Now we can immediately complete the proof of item (ii) of the Proposition. Indeed, as we know from **VII.**, Lecture 2, the function Φ^* is self-concordant on its domain; since $s = \Phi'(u)$, we have $s \in \mathrm{Dom}\,\Phi^*$. (8.18) says that the $|\cdot|_{(\Phi^*)''(s)}$-distance between $s \in \mathrm{Dom}\,\Phi^*$ and $s^f(0)$ equals to $\lambda(F_t, x)$ and is therefore < 1 due to the premise of (ii). Consequently, $s(0)$ belongs to the centered at s open unit Dikin ellipsoid of the self-concordant function Φ^* and is therefore in the domain of the function (**I.**, Lecture 2). The latter domain is open (**VII.**, Lecture 2), so that $s^f(dt) \in \mathrm{Dom}\,\Phi^*$ for all small enough $dt \geq 0$; since $S(dt)$ always satisfies (8.13), we conclude that $S(dt)$ is dual feasible for all small enough $|dt|$. ■

Propositions 8.2.1 and 8.2.2 lead to the following

Acceptability Test:

given a κ-close to the path primal feasible pair (t, x) and a candidate stepsize dt, form the corresponding primal and dual pairs $X(dt) = (t + dt, x^f(dt) = x + dx(dt))$, $S(dt) = (t + dt, s^f(dt) = s + ds(dt))$ and check whether the associated upper bound

$$v(dt) \equiv V_{s^f(dt)}(t + dt, x^f(dt)) = (t + dt)c^T x^f(dt) + F(x^f(dt)) + \Phi^*(s^f(dt)) - p^T s^f(dt) \qquad (8.20)$$

for the residual $V(t + dt, x^f(dt))$ is $\leq \overline{\kappa}$ (by definition, $v(dt) = +\infty$ if $x^f(dt) \notin \mathrm{Dom}\,F$ or if $s^f(dt) \notin \mathrm{Dom}\,\Phi^*$).

If $v(dt) \leq \overline{\kappa}$, accept the stepsize dt, otherwise reject it.

An immediate corollary of Propositions 8.2.1, 8.2.2 is the following

Proposition 8.2.3 *If (t, x) is a κ-close to the path primal feasible pair and a stepsize dt passes the Acceptability Test, then*

$$V(t + dt, x^f(dt)) \leq \overline{\kappa}$$

and, consequently, the Newton complexity of the corrector step under the choice $\delta t = dt$ does not exceed the quantity

$$N(\kappa, \overline{\kappa}) = O(1)\left\{\overline{\kappa} + \ln\left(1 + \ln\frac{1}{\kappa}\right)\right\},$$

$O(1)$ being an absolute constant.

Now it is clear that in order to get a "long step" version of the path-following method, it suffices to equip the Predictor-Corrector Updating scheme with a linesearch-based rule for choosing the largest possible stepsize δt which passes our Acceptability Test. Such a rule for sure keeps the complexity of a corrector step at a fixed level; at the same time, the rule is computationally cheap, since to test a stepsize, we should compute the values of Φ and Φ^* only, which normally is nothing as compared to the cost of the corrector step.

The outlined approach needs, of course, theoretical justification. Indeed, to the moment we do not know what is the "power" of our Acceptability Test - does it accept, e.g., the "short" stepsizes $dt = O(t/\sqrt{\vartheta})$ used in the very first version of the method. This is the issue we come to.

8.3 Acceptable steps

Let us start with the following construction. Given a point $u \in \mathrm{Dom}\,\Phi$ and a direction $\delta u \in \mathbf{R}^m$, let us set

$$s = \Phi'(u), \quad \delta s = \Phi''(u)\delta u,$$

thus coming to the *conjugate point* $s \in \mathrm{Dom}\,\Phi^*$ and to the *conjugate direction* δs. Now, let $\rho_u^*[\delta u]$ be the remainder in the second-order Taylor expansion of the function $\Phi(v) + \Phi^*(w)$ at the point (u, s) along the direction $(\delta u, \delta s)$:

$$\rho_u^*[\delta u] = \Phi(u + \delta u) + \Phi^*(s + \delta s) -$$

$$- \left[\Phi(u) + \Phi^*(s) + [\delta u]^T \Phi'(u) + [\delta s]^T \Phi'(s) + \frac{[\delta u]^T \Phi''(u)\delta u}{2} + \frac{[\delta s]^T (\Phi^*)''(s)\delta s}{2}\right]$$

(the right hand side is $+\infty$, if $u + \delta u \notin \mathrm{Dom}\,\Phi$ or if $s + \delta s \notin \mathrm{Dom}\,\Phi^*$).

Our local goal is to establish the following

Lemma 8.3.1 *One has*

$$\zeta \equiv |\delta u|_{\Phi''(u)} = |\delta s|_{(\Phi^*)''(s)} = \sqrt{[\delta u]^T \delta s}. \tag{8.21}$$

Besides this, if $\zeta < 1$, then

$$\rho_u^*[\delta u] \le 2\rho(\zeta) - \zeta^2 = \frac{2}{3}\zeta^3 + \frac{2}{4}\zeta^4 + \frac{2}{5}\zeta^5 + ..., \ \rho(z) = -\ln(1 - z) - z. \tag{8.22}$$

Last, the third derivative of $\Phi(\cdot) + \Phi^*(\cdot)$ taken at the point (u, s) along the direction $(\delta u, \delta s)$ is zero, so that $\rho_u^*[\delta u]$ is in fact the reminder in the third-order Taylor expansion of $\Phi(\cdot) + \Phi^*(\cdot)$.

Proof. (8.21) is proved exactly as relation (8.17), see Lemma 8.2.1. From (8.21) it follows that if $\zeta < 1$, then both $u + \delta u$ and $s + \delta s$ are in the centered at u, respectively, s open unit Dikin ellipsoids of the self-concordant functions Φ, Φ^* (the latter function is self-concordant due to **VII.**, Lecture 2). Applying to Φ and Φ^* **I.**, Lecture 2, we come to

$$u + \delta u \in \mathrm{Dom}\,\Phi, \ \Phi(u + \delta u) \le \Phi(u) + [\delta u]^T \Phi'(u) + \rho(|\delta u|_{\Phi''(u)}),$$

$$s + \delta s \in \mathrm{dom}\Phi^*, \ \Phi^*(s + \delta s) \le \Phi^*(s) + [\delta s]^T (\Phi^*)'(s) + \rho(|\delta s|_{(\Phi^*)''(s)}),$$

whence

$$\rho_u^*[\delta u] \le 2\rho(\zeta) - \frac{1}{2}|\delta u|^2_{\Phi''(u)} - \frac{1}{2}|\delta s|^2_{(\Phi^*)''(s)} = 2\rho(\zeta) - \zeta^2,$$

as claimed in (8.22).

To prove that the third order derivative of $\Phi(\cdot) + \Phi^*(\cdot)$ taken at the point (u, s) in the direction $(\delta u, \delta s)$ is zero, let us differentiate the identity

$$h^T[(\Phi^*)''(\Phi'(v))]h = h^T[\Phi''(v)]^{-1}h$$

(h is fixed) with respect to v in the direction h (cf. item 4^0 in the proof of **VII.**, Lecture 2). The differentiation results in

$$D^3\Phi^*(\Phi'(v))[h, h, h] = -D^3\Phi(v)[[\Phi''(v)]^{-1}h, [\Phi''(v)]^{-1}h, [\Phi''(v)]^{-1}h];$$

substituting $v = u$, $h = \delta s$, we come to

$$D^3\Phi(u)[\delta u, \delta u, \delta u] = -D^3\Phi^*(s)[\delta s, \delta s, \delta s].$$

■

Now we are ready to prove the following central result.

Proposition 8.3.1 *Let (t, x) be κ-close to the path, and let dt, $|dt| < t$, be a stepsize. Then the quantity $v(dt)$ (see (8.20)) satisfies the inequality*

$$v(dt) \le \rho^*_u[du(dt)], \tag{8.23}$$

while

$$|du(dt)|_{\Phi''(u)} \le \omega \equiv \lambda(F_t, x) + \frac{|dt|}{t}[\lambda(F_t, x) + \lambda(F, x)] \le \kappa + \frac{|dt|}{t}[\kappa + \sqrt{\vartheta}]. \tag{8.24}$$

In particular, if $\omega < 1$, then $v(dt)$ is well-defined and is $\le 2\rho(\omega) - \omega^2$. Consequently, if

$$2\rho(\kappa) - \kappa^2 < \overline{\kappa} \tag{8.25}$$

then all stepsizes dt satisfying the inequality

$$\frac{|dt|}{t} \le \frac{\kappa^+ - \kappa}{\kappa + \lambda(F, x)}, \tag{8.26}$$

κ^+ being the root of the equation

$$2\rho(z) - z^2 = \overline{\kappa},$$

pass the Acceptability Test.

Proof. Let u, s, $du(dt)$, $ds(dt)$ be given by (8.16). In view of (8.16), s is conjugate to u and $ds(dt)$ is conjugate to $du(dt)$, so that by definition of $\rho^*_u[\cdot]$, we have, denoting $\zeta = |du(dt)|_{\Phi''(u)} = |ds(dt)|_{(\Phi^*)''(s)}$ (see (8.21))

$$\Phi(u + du(dt)) + \Phi^*(s + ds(dt)) =$$

$$= \Phi(u) + [du(dt)]^T \Phi'(u) + \Phi^*(s) + [ds(dt)]^T (\Phi^*)'(s) + \zeta^2 + \rho^*_u[du(dt)] =$$

[since $s = \Phi'(u)$ and, consequently, $\Phi(u) + \Phi^*(s) = u^T s$ and $u = (\Phi^*)'(s)$, since Φ^* is the Legendre transformation of Φ]

$$= u^T s + [du(dt)]^T s + u^T ds(dt) + \zeta^2 + \rho^*_u[du(dt)] =$$

$$= [u + du(dt)]^T [s + ds(dt)] - [du(dt)]^T ds(dt) + \zeta^2 + \rho^*_u[du(dt)] =$$

[since $[du(dt)]^T ds(dt) = \zeta^2$ by (8.21)]

$$= [u + du(dt)]^T [s + ds(dt)] + \rho^*_u[du(dt)] =$$

[since $u + du(dt) = \pi[x + dx(dt)] + p$ and, by Proposition 8.2.2, $\pi^T[s + ds(dt)] = -(t + dt)c$]

$$= p^T[s + ds(dt)] - (t + dt)c^T[x + dx(dt)] + \rho^*_u[du(dt)] =$$

[the definition of $x^f(dt)$ and $s^f(dt)$]

$$= p^T s^f(dt) - (t + dt)c^T x^f(dt) + \rho^*_u[du(dt)],$$

whence (see (8.20))

$$v(dt) \equiv (t + dt)c^T x^f(dt) + F(x^f(dt)) + \Phi^*(s^f(dt)) - p^T s^f(dt) = \rho^*_u[du(dt)],$$

as required in (8.23).

Now let us prove (8.24). In view of (8.16) and (8.12) we have

$$|du(dt)|_{\Phi''(u)} = |\pi dx(dt)|_{\Phi''(u)} = |dx(dt)|_{F''(x)} =$$

[see (8.7)]

$$= |[F''(x)]^{-1}\nabla_x F_{t+dt}(x)|_{F''(x)} \equiv \sqrt{[[F''(x)]^{-1}\nabla_x F_{t+dt}(x)]^T \, [F''(x)] \, [[F''(x)]^{-1}\nabla_x F_{t+dt}(x)]} =$$

$$= |\nabla_x F_{t+dt}(x)|_{[F''(x)]^{-1}} = |(t + dt)c + F'(x)|_{[F''(x)]^{-1}} =$$

$$= |(1 + dt/t)[tc + F'(x)] - (dt/t)F'(x)|_{[F''(x)]^{-1}} \le$$

$$\leq (1+\frac{|dt|}{t})|\nabla_x F_t(x)|_{[F''(x)]^{-1}} + \frac{|dt|}{t}|F'(x)|_{[F''(x)]^{-1}} \leq$$

[due to the definition of $\lambda(F_t, x)$ and $\lambda(F, x)$]

$$\leq (1+\frac{|dt|}{t})\lambda(F_t, x) + \frac{|dt|}{t}\lambda(F, x) = \omega \leq$$

[since (t, x) is κ-close to the path, so that $\lambda(F_t, x) \leq \kappa$, and since F is ϑ-self-concordant barrier]

$$\leq (1+\frac{|dt|}{t})\kappa + \frac{|dt|}{t}\sqrt{\vartheta}.$$

The remaining statements of Proposition are immediate consequences of (8.23), (8.24) and Lemma 8.3.1. ■

8.4 Summary

Summarizing our observations and results, we come to the following

Long-Step Predictor-Corrector Path-Following method:

- *The parameters of the method are the path tolerance $\kappa \in (0,1)$ and the treshold $\overline{\kappa} > 2\rho(\kappa) - \kappa^2$; the input to the method is a κ-close to the path primal feasible pair (t^0, x^0) .*
- *The method forms, starting with (t^0, x^0), the sequence of κ-close to the path pairs (t^i, x^i), with the updating*
$$(t^{i-1}, x^{i-1}) \mapsto (t^i, x^i)$$
being given by the Predictor-Corrector Updating scheme, where the stepsizes $\delta t^i \equiv t^i - t^{i-1}$ are nonnegative reals passing the Acceptability Test associated with the pair (t^{i-1}, x^{i-1}).

Since, as we know from Proposition 8.3.1, the stepsizes

$$\delta t^i_* = t^{i-1}\frac{\kappa^+ - \kappa}{\kappa + \lambda(F, x^{i-1})}$$

for sure pass the Acceptability Test, we may assume that the stepsizes in the above method are *at least* the default values δt^i_*:

$$\delta t^i \geq t^{i-1}\frac{\kappa^+ - \kappa}{\kappa + \lambda(F, x^{i-1})}; \tag{8.27}$$

note that to use the default stepsizes $\delta t^i \equiv \delta t^i_*$, no Acceptability Test, and, consequently, no Structural assumption on the barrier F is needed. Note also that to initialize the method (to get the initial close to the path pair (t^0, x^0)), one can trace "in the reverse time" the auxiliary path associated with a given strictly feasible initial solution $\widehat{x} \in \operatorname{int} G$ (see Lecture 4); and, of course, when tracing the auxiliary path, we also can use the long-step predictor-corrector technique.

The method in question, of course, fits the standard complexity bounds:

Theorem 8.4.1 *Let problem (8.1) be solved by the Long-Step Predict-or-Corrector Path-Following method which starts at a κ-close to the path primal feasible pair (t^0, x^0) and uses stepsizes δt^i passing the Acceptability Test and satisfying (8.27). Then the total number of Newton steps in the method before an ε-solution to the problem is found does not exceed*

$$O(1)\sqrt{\vartheta}\ln\left(\frac{\vartheta}{t_0\varepsilon} + 1\right) + 1,$$

with $O(1)$ depending on the parameters κ and $\overline{\kappa}$ of the method only.

Proof. Since (t^i, x^i) are κ-close to the path, we have $c^T x^i - \min_{x\in G} c^T x \leq O(1)\vartheta t_i^{-1}$ with certain $O(1)$ depending on κ only (see Proposition 4.4.1, Lecture 4); this inaccuracy bound combined with (8.27) (where one should take into account that $\lambda(F, x^{i-1}) \leq \sqrt{\vartheta})$ implies that $c^T x^i - \min_{x\in G} c^T x$ becomes $\leq \varepsilon$ after no more than $O(1)\sqrt{\vartheta}\ln(1+\vartheta t_0^{-1}\varepsilon^{-1})+1$ steps, with $O(1)$ depending on κ and $\overline{\kappa}$ only. It remains to note that since the stepsizes pass the Acceptability Test, the Newton complexity of a step in the method, due to Proposition 8.2.3, is $O(1)$. ■

8.5 Exercises: Long-Step Path-Following methods

Let us start with clarifying an immediate question motivated by the above construction.

Exercise 8.5.1 $^{\#*}$ *The Structural assumption requires F to be obtained from a barrier with known Legendre transformation by affine substitution of the argument. Why did not we simplify things by assuming that F itself has a known Legendre transformation?*

The remaining exercises tell us another story. We have presented certain "long step" variant of the path-following scheme; note, anyhow, that the "cost" of "long steps" is certain structural assumption on the underlying barrier. Although this assumption is automatically satisfied in many important cases, we have paid something. Can we say something definite about the advantages we have paid for? "Definite" in the previous sentence means "something which can be proved", not "something which can be supported by computational experience" (this latter aspect of the situation is more or less clear).

The answer is as follows. *As far as the worst case complexity bound is concerned, there is no progress at all*, and the current state of the theory of interior point methods do not give us any hope to get a worst-case complexity estimate better than $O(\sqrt{\vartheta}\ln(\mathcal{V}/\varepsilon))$. Thus, if we actually have got something, this is not an improvement in the worst case complexity. The goal of the forthcoming exercises is to explain what is the improvement.

Let us start with some preliminary considerations. Consider a step of a path-following predictor-corrector method; for the sake of simplicity, assume that at the beginning of the step we are *exactly* at the path rather than are close to it (what follows can be without any difficulties extended onto this latter situation). Thus, we are given $t > 0$ and $x = x^*(t)$, and our goal is to update the pair (t, x) into a new pair (t^+, x^+) close to the path with larger value of the penalty parameter. To this end we choose a stepsize $dt > 0$, set $t^+ = t + dt$ and make the predictor step

$$x \mapsto x^f = x + (x^*)'(t)dt,$$

shifting x along the tangent to the path line l. At the corrector step we apply to F_{t^+} the damped Newton method, starting with x^f, to restore closeness to the path. Assume that the method in question ensures that the residual

$$F_{t^+}(x^f) - \min_x F_{t^+}(x)$$

is $\leq O(1)$ (this is more or less the same as to ensure a fixed Newton complexity of the corrector step). Given that the method in question possesses the aforementioned properties, we may ask ourselves *what is the length of the displacement $x^f - x$ which is guaranteed by the method*. It is natural to measure the length in the local metric $|\cdot|_{F''(x)}$ given by the Hessian of the barrier. Note that in the short-step version of the method, where $dt = O(1)t(1 + \lambda(F, x))^{-1}$, we have (see (8.7))

$$dx(dt) = -dt[F''(x)]^{-1}c = t^{-1}dt[F''(x)]^{-1}F'(x)$$

(since at the path $tc + F'(x) = 0$), whence

$$|x^f(dt) - x|_{F''(x)} = |dx(dt)|_{F''(x)} = t^{-1}dt|[F''(x)]^{-1}F'(x)|_{F''(x)} =$$

$$= t^{-1}dt|F'(x)|_{[F''(x)]^{-1}} = t^{-1}dt\lambda(F, x),$$

and, substituting the expression for dt, we come to

$$\Omega \equiv |x^f(dt) - x|_{F''(x)} = O(1)\frac{\lambda(F, x)}{1 + \lambda(F, x)},$$

so that $\Omega = O(1)$, provided that $\lambda(F, x) \geq O(1)$, or, which is the same, provided that we are not too close to the analytic center of G.

Thus, the quantity Ω - let us call it the *prediction power* of the method - for the default short-step version of the method is $O(1)$. The goal of what follows is to investigate the prediction power of the long-step version of the method and to compare it with the above reference point - the $O(1)$-power of the short-step version; this is a natural formalization of the question "how long are the long steps".

First of all, let us note that there is a natural *upper* bound on the prediction power - namely, the distance (measured, of course, in $|\cdot|_{F''(x)}$) from x to the boundary of G along the tangent line l. Actually there are two distances, since there are two ways to reach ∂G along l - the "forward" and the "backward" movement. It is reasonable to speak about the shortest of these distances - about the quantity

$$\Delta \equiv \Delta(x) = \min\{|p\,(x^*)'(t)|_{F''(x)} \mid x + p\,(x^*)'(t) \notin \text{int } G\}.$$

Since G contains the centered at x unit Dikin ellipsoid of F (i.e., the centered at x $|\cdot|_{F''(x)}$-unit ball), we have

$$\Delta \geq 1.$$

Note that there is no prediction policy which *always* results in $\Omega >> 1$, since it may happen that both "forward" and "backward" distances from x to the boundary of G are of order of 1 (look at the case when G is the unit cube $\{y \in \mathbf{R}^n \mid |y|_\infty \leq 1\}$, $F(y)$ is the standard logarithmic barrier $-\sum_{i=1}^n [\ln(1-y_i) + \ln(1+y_i)]$ for the cube, $x = (0.5, 0, ..., 0)^T$ and $c = (-1, 0, ..., 0)^T$). What we can speak about is the *type of dependence* $\Omega = \Omega(\Delta)$; in other words, it is reasonable to ask ourselves "how large is Ω when Δ is large", not "how large is Ω" - the answer to this latter question cannot be better than $O(1)$.

In what follows we answer the above question for the particular case as follows:

Semidefinite Programming: the barrier Φ involved into our Structural assumption is the barrier

$$\Phi(X) = -\ln \text{Det } X$$

for the cone $\mathbf{S}^k_+$ of symmetric positive semidefinite $k \times k$ matrices

In other words, we restrict ourselves with the case when G is the inverse image of $\mathbf{S}^k_+$ under the affine mapping

$$x \mapsto \mathcal{A}(x) = \pi x + p$$

taking values in the space $\mathbf{S}^k$ of $k \times k$ symmetric matrices and

$$F(x) = -\ln \text{Det } \mathcal{A}(x).$$

Note that the Semidefinite Programming case (very important in its own right) covers, in particular, Linear Programming (look what happens when $\pi x + p$ takes values in the subspace of diagonal matrices).

Let us summarize our current knowledge on the situation in question.

- Φ is k-self-concordant barrier for the cone $\mathbf{S}^k$; the derivatives of the barrier are given by

  $$D\Phi(u)[h] = -\text{Tr}\{u^{-1}h\} = -\text{Tr}\{\widehat{h}\},\ \widehat{h} = u^{-1/2}hu^{-1/2},$$

  so that

  $$\Phi'(u) = -u^{-1}; \tag{8.28}$$

  $$D^2\Phi(u)[h,h] = \text{Tr}\{u^{-1}hu^{-1}h\} = \text{Tr}\{\widehat{h}^2\},$$

  so that

  $$\Phi''(u)h = u^{-1}hu^{-1}; \tag{8.29}$$

  $$D^3\Phi(u)[h,h,h] = -2\,\text{Tr}\{u^{-1}hu^{-1}hu^{-1}h\} = -2\,\text{Tr}\{\widehat{h}^3\}$$

  (see Example 5.3.3, Lecture 5, and references therein);

- the cone $\mathbf{S}^k_+$ is self-dual; the Legendre transformation of Φ is

  $$\Phi^*(s) = -\Phi(-s) + \mathit{const},\ \ \text{Dom}\,\Phi^* = -\text{int } \mathbf{S}^n_+$$

  (Exercises 5.4.7, 5.4.10).

Let us get more information on the barrier Φ. Let us call *an arrow* a pair (v, dv) comprised of $v \in \operatorname{int} \mathbf{S}^k_+$ and $dv \in \mathbf{S}^k$ with $|dv|_{\Phi''(v)} = 1$. Given an arrow (v, dv), let us define the *conjugate co-arrow* (v^*, dv^*) as

$$v^* = \Phi'(v) = -v^{-1}, \ dv^* = \Phi''(v)dv = v^{-1}dvv^{-1}.$$

Let also

$$\zeta(v, dv) = \sup\{p \mid \ v \pm pdv \in \mathbf{S}^k_+\}, \tag{8.30}$$

$$\zeta^*(v^*, dv^*) = \sup\{p \mid \ v^* \pm dv^* \in -\mathbf{S}^k_+\}. \tag{8.31}$$

In what follows $|w|_\infty$, $|w|_2$ are the spectral norm (maximum modulus of eigenvalues) and the Frobenius norm $\operatorname{Tr}^{1/2}\{w^2\}$ of a symmetric matrix w, respectively.

Exercise 8.5.2 *Let (v, dv) be an arrow and (v^*, dv^*) be the conjugate co-arrow. Prove that*

$$1 = |dv|_{\Phi''(v)} = |v^{-1/2}dvv^{-1/2}|_2 = |dv^*|_{(\Phi^*)''(v^*)} = \sqrt{\operatorname{Tr}\{dv\, dv^*\}} \tag{8.32}$$

and that

$$\zeta(v, dv) = \zeta^*(v^*, dv^*) = \frac{1}{|v^{-1/2}dvv^{-1/2}|_\infty}. \tag{8.33}$$

Exercise 8.5.3 * *Prove that for any positive integer j, any $v \in \operatorname{int} \mathbf{S}^k_+$ and any $h \in \mathbf{S}^k$ one has*

$$D^j\Phi(v)[h, ..., h] = (-1)^j(j-1)!\operatorname{Tr}\{\widehat{h}^j\}, \ \widehat{h} = v^{-1/2}hv^{-1/2}, \tag{8.34}$$

and, in particular,

$$|D^j\Phi(v)[h, ..., h]| \le (j-1)!|\widehat{h}|_2|\widehat{h}|_\infty^{j-2}, \ j \ge 2. \tag{8.35}$$

Let $\rho_j(z)$ be the reminder in j-th order Taylor expansion of the function $-\ln(1-z)$ at $z = 0$:

$$\rho_j(z) = \frac{1}{j+1}z^{j+1} + \frac{1}{j+2}z^{j+2} + ...$$

(so that the perfectly known to us function $\rho(z) = -\ln(1-z) - z$ is nothing but $\rho_1(z)$).

Exercise 8.5.4 + *Let (v, dv) be an arrow, and let $R^j_{(v,dv)}(r)$, $j \ge 2$, be the remainder in j-th order Taylor expansion of the function $f(r) = \Phi(v + rdv)$ at $r = 0$:*

$$R^j_{(v,dv)}(r) = f(r) - \sum_{i=0}^{j} \frac{f^{(i)}(0)}{i!}r^i$$

(the right hand side is $+\infty$, if f is undefined at r). Prove that

$$R^j_{(v,dv)}(r) \le \zeta^2(v, dv)\rho_j\left(\frac{|r|}{\zeta(v, dv)}\right), \ |r| < \zeta(v, dv) \tag{8.36}$$

(the quantity $\zeta(v, dv)$ is given by (8.30), see also (8.33)).

Exercise 8.5.5 * *Let (v, dv) be an arrow and (v^*, dv^*) be the conjugate co-arrow. Let $\mathcal{R}^j_{(v,dv)}(r)$, $j \ge 2$, be the reminder in j-th order Taylor expansion of the function $\psi(r) = \Phi(v + rdv) + \Phi^*(v^* + rdv^*)$ at $r = 0$:*

$$\mathcal{R}^j_{(v,dv)}(r) = \psi(r) - \sum_{i=0}^{j} \frac{\psi^{(i)}(0)}{i!}r^i$$

(the right hand side is $+\infty$, if ψ is undefined at r). Prove that

$$\mathcal{R}^j_{(v,dv)}(r) \le 2\zeta^2(v, dv)\rho\left(\frac{|r|}{\zeta(v, dv)}\right), \ |r| < \zeta(v, dv) \tag{8.37}$$

(the quantity $\zeta(v, dv)$ is given by (8.30), see also (8.33)).

Now let us come back to our goal - investigating the forecast power of the long step predictor-corrector scheme for the case of Semidefinite Programming. Thus, let us fix the pair (t, x) belonging to the path (so that $t > 0$ and $x = x^*(t) = \mathrm{argmin}_{y \in G}[tc^T x + F(x)]$). We use the notation as follows:

- I is the unit $k \times k$ matrix;
- $u = \pi x + p$;
- $dx \in \mathbf{R}^n$ is the $|\cdot|_{F''(x)}$-unit direction parallel to the line l, and

$$du = \pi dx$$

  is the direction of the image $\mathcal{L}$ of the line l in the space $\mathbf{S}^k$;
- $s \equiv \Phi'(u) = -u^{-1}$; $ds \equiv \Phi''(u)du = u^{-1}duu^{-1}$.

Let us first realize what the quantity $\Delta(x)$ is.

Exercise 8.5.6 $^+$ *Prove that (u, du) is an arrow, (s, ds) is the conjugate co-arrow and that*

$$\Delta = \zeta(u, du).$$

Now we are ready to answer what is the prediction power of the long step predictor-corrector scheme.

Exercise 8.5.7 $^+$ *Consider the Long-Step Predictor-Corrector Updating scheme with linesearch (which chooses, as the stepsize, the largest value of dt which passes the Acceptability Test) as applied to Semidefinite Programming. Prove that the prediction power of the scheme is at least*

$$\Omega^*(x) = O(1)\Delta^{1/2}(x),$$

with $O(1)$ depending on the treshold $\overline{\kappa}$ only[1].

Thus, the long-step scheme indeed has a "nontrivial" prediction power.

An interesting question is to bound from *above* the prediction power of an *arbitrary* predictor-corrector path-following scheme of the aforementioned type; recall that the main restrictions on the scheme were that

- in order to form the forecast x^f, we move along the tangent line l to the path [in principle we could use higher-order polynomial approximations on it; here we ignore this possibility]
- the residual $F_{t^+}(x^f) - \min_y F_{t^+}(x)$ should be $\leq O(1)$.

It can be proved that *in the case of Linear (and, consequently, Semidefinite) Programming the prediction power of any predictor-corrector scheme* subject to the above restrictions *cannot be better than* $O(1)\Delta^{2/3}(x)$ (which is slightly better than the prediction power $O(1)\Delta^{1/2}(x)$ of our method). I do not know what is the origin of the gap - drawbacks of the long-step method in question or too optimistic upper bound, and you are welcome to investigate the problem.

[1] recall that for the sake of simplicity the pair (t, x) to be updated was assumed to be *exactly* at the path; if it is κ-close to the path, then similar result holds true, with $O(1)$ depending on both κ and $\overline{\kappa}$

Chapter 9

How to construct self-concordant barriers

To the moment we are acquainted with four interior point methods; the "interior point toolbox" contains more of them, but we are enforced to stop somewhere, and I think it is a right time to stop. Let us think how could we exploit our knowledge in order to solve a convex program by one of our methods. Our actions are clear:

(a) we should reformulate our problem in the standard form

$$\textit{minimize}\ \ c^T x \ \ s.t.\ \ x \in G \tag{9.1}$$

of a problem of minimizing a linear objective over a closed convex domain (or in the conic form - as a problem of minimizing a linear objective over the intersection of a convex cone and an affine plane; for the sake of definiteness, let us speak about the standard form).

In principle (a) does not cause any difficulty - we know that both standard and conic problems are universal forms of convex programs.

(b) we should equip the domain/cone given by step (a) by a "computable" self-concordant barrier. Sometimes we need something more - e.g., to apply the potential reduction methods, we are interested in logarithmically homogeneous barriers, possibly, along with their Legendre transformations, and to use the long-step path-following scheme, we need a barrier satisfying the Structural assumption from Lecture 8.

Now, our current knowledge on the crucial issue of constructing self-concordant barriers is rather restricted. We know exactly 3 "basic" self-concordant barriers:

- (I) the 1-self-concordant barrier $-\ln x$ for the nonnegative axis (Example 3.1.2, Lecture 3);
- (II) the m-self-concordant barrier $-\ln \operatorname{Det} x$ for the cone $\mathbf{S}^m_+$ of positive semidefinite $m \times m$ matrices (Exercise 3.3.3);
- (III) the 2-self-concordant barrier $-\ln(t^2 - x^T x)$ for the second-order cone $\{(t,x) \in \mathbf{R} \times \mathbf{R}^k \mid t \geq |x|_2\}$ (Example 5.3.2, Lecture 5).

Note that the latter two examples were not justified in the lectures; and this is not that easy to prove that (III) indeed is a self-concordant barrier for the second-order cone.

Given the aforementioned basic barriers, we can produce many other self-concordant barriers by applying the combination rules, namely, by taking sums of these barriers, their direct sums and superpositions with affine mappings (Proposition 3.1.1, Lecture 3). These rules, although very simple, are surprisingly powerful; what should be mentioned first, is that *the rules allow to treat all constraints defining the feasible set G seperately.* We mean the following. Normally the feasible set G is defined by a finite number m of constraints; each of them defines its own feasible set G_i, so that the resulting feasible set G is the intersection of the G_i:

$$G = \cap_{i=1}^m G_i.$$

According to Proposition 3.1.1.(ii), in order to find a self-concordant barrier for G, it suffices to find similar barriers for all G_i and then to take the sum of these "partial" barriers. Thus, we have in our disposal the *Decomposition rule* which makes the problem of step (b) "separable with respect to constraints".

The next basic tool is the *Substitution rule* given by Proposition 3.1.1.(i):

In order to get a ϑ-self-concordant barrier F for a given convex domain G, it suffices to represent the domain as the inverse image, under certain affine mapping $\mathcal{A}$, of another domain, G^+, with known ϑ-self-concordant barrier F^+:

$$G = \mathcal{A}^{-1}(G^+) \equiv \{x \mid \mathcal{A}(x) \in G^+\}$$

(the image of $\mathcal{A}$ should intersect the interior of G^+); given such representation, you can take as F the superposition

$$F(x) = F^+(\mathcal{A}(x))$$

of F^+ and the mapping $\mathcal{A}$.

The Decomposition and the Substitution rules as applied to the particular self-concordant barriers (I) - (III) allow to obtain barriers required by several important generic Convex Programming problems, e.g., they immediately imply self-concordance of the standard logarithmic barrier

$$F(x) = -\sum_{i=1}^{m} \ln(b_i - a_i^T x)$$

for the polyhedral set

$$G = \{x \mid a_i^T x \le b_i,\ i = 1, ..., m\};$$

this latter fact covers all needs of Linear Programming. Thus, we cannot say that we are completely unequipped; at the same time, our equipment is not too rich. Consider, for example, the problem of the best $|\cdot|_p$-approximation:

(L$_p$): *given sample $u_j \in \mathbf{R}^n$, $j = 1, ..., N$, of "regressors" along with the responses $v_j \in \mathbf{R}$, find the linear model*

$$v = x^T u$$

which optimally fits the observations in the $|\cdot|_p$-norm, i.e., minimizes the quantity

$$f(x) = \sum_{j=1}^{N} |v_j - x^T u_j|^p$$

(in fact $|\cdot|_p$-criterion is $f^{1/p}(x)$, but it is, of course, the same what to minimize - f or $f^{1/p}$).

$f(\cdot)$ clearly is a convex function, so that our approximation problem is a convex program. In order to solve it by an interior point method, we can write the problem down in the standard form, which is immediate:

$$\textit{minimize}\ \ t\ \ \textit{s.t.}\ \ (t, x) \in G = \{(t, x) \mid f(x) \le t\};$$

now we need a self-concordant barrier for G, and where to take it?

At the beginning of the "interior point science" for nonlinear convex problems we were enforced to invent an "ad hoc" self-concordant barrier for each new domain we met and then were to prove that the invented barrier actually is self-concordant, which in many cases required a lot of unpleasant computations. Recently it became clear that there is a very powerful technique for constructing self-concordant barriers, which allows to obtain all previously known barriers, same as a number of new ones, without any computations "from nothing" - more exactly, from the fact that the function $-\ln x$ is 1-self-concordant barrier for the nonnegative half-axis. This technique is based on extending the Substitution rule by replacing *affine* mappings $\mathcal{A}$ by a wider family of certain *nonlinear* mappings. The essence of the matter is, of course, what are appropriate for our goals nonlinear mappings $\mathcal{A}$. It is clear in advance that these cannot be arbitrary mappings, even smooth ones - we *at least* should provide convexity of $G = \mathcal{A}^{-1}(G^+)$.

9.1 Appropriate mappings and Main Theorem

Let us fix a closed convex domain $G^+ \subset \mathbf{R}^N$. An important role in what follows is played by the *recessive cone* $\mathcal{R}(G^+)$ of the domain defined as

$$\mathcal{R}(G^+) = \{h \in \mathbf{R}^N \mid u + th \in G^+ \ \forall t \geq 0 \ \forall u \in G^+\}.$$

It is immediately seen that $\mathcal{R}(G^+)$ is a closed convex cone in $\mathbf{R}^N$.

Now we are able to define the family of mappings $\mathcal{A}$ appropriate for us.

Definition 9.1.1 *Let $G^+ \subset \mathbf{R}^N$ be closed convex domain, and let $K = \mathcal{R}(G^+)$ be the recessive cone of G^+. A mapping*

$$\mathcal{A}(x) : \mathrm{int}\, G^- \to \mathbf{R}^N$$

defined and C^3 smooth on the interior of a closed convex domain $G^- \subset \mathbf{R}^n$ is called β-appropriate for G^+ (here $\beta \geq 0$) if

(i) *$\mathcal{A}$ is concave with respect to K, i.e.,*

$$D^2\mathcal{A}(x)[h,h] \leq_K 0 \ \forall x \in \mathrm{int}\, G^- \ \forall h \in \mathbf{R}^n$$

(from now on we write $a \leq_K b$, if $b - a \in K$);

(ii) *$\mathcal{A}$ is compatible with G^- in the sense that*

$$D^3\mathcal{A}(x)[h,h,h] \leq_K -3\beta D^2\mathcal{A}(x)[h,h]$$

whenever $x \in \mathrm{int}\, G^-$ and $x \pm h \in G^-$.

For example, an affine mapping $\mathcal{A} : \mathbf{R}^n \to \mathbf{R}^N$, restricted on any closed convex domain $G^- \subset \mathbf{R}^n$, cleraly is 0-appropriate for any $G^+ \subset \mathbf{R}^N$.

The definition of compatibility looks strange; its justification is that it works. Namely, there is the following central result (it will be proved in Section 9.4):

Theorem 9.1.1 *Let*

- *$G^+ \subset \mathbf{R}^N$ be a closed convex domain;*
- *F^+ be a ϑ_+-self-concordant barrier for G^+;*
- *$\mathcal{A} : \mathrm{int}\, G^- \to \mathbf{R}^N$ be a mapping β-appropriate for G^+;*
- *F^- be a ϑ_--self-concordant barrier for G_-.*

Assume that the set

$$G^0 = \{x \in \mathrm{int}\, G^- \mid \mathcal{A}(x) \in \mathrm{int}\, G^+\}$$

is nonempty. Then G^0 is the interior of a closed convex domain

$$G \equiv \mathrm{cl}\, G^0,$$

and the function

$$F(x) = F^+(\mathcal{A}(x)) + \max[1, \beta^2] F^-(x)$$

is a ϑ-self-concordant barrier for G, with

$$\vartheta = \vartheta_+ + \max[1, \beta^2]\vartheta_-.$$

The above Theorem resembles the Substitution rule: we see that an affine mapping $\mathcal{A}$ in the latter rule can be replaced by an arbitrary nonlinear mapping (which should, anyhow, be appropriate for G^+), and the substitution $F^+(\cdot) \mapsto F^+(\mathcal{A}(\cdot))$ should be accompanied by adding to the result a self-concordant barrier for the domain of $\mathcal{A}$. Let us call this new rule "Substitution rule (N)" (nonlinear); to distinguish between this rule and the initial one, let us call the latter "Substitution rule (L)" (linear). In fact Substitution rule (L) is a very particular case of Substitution rule (N); indeed, an affine mapping $\mathcal{A}$, as we know, is appropriate for any domain G^+, and since domain of $\mathcal{A}$ is the whole $\mathbf{R}^n$, one can set $F^- \equiv 0$ (this is 0-self-concordant barrier for $\mathbf{R}^n$), thus coming to the Substitution rule (L).

9.2 Barriers for epigraphs of functions of one variable

As an immediate consequence of the Substitution rule (N), we get a number of self-concordant barriers for the epigraphs of functions on the axis. These barriers are given by the following construction:

Proposition 9.2.1 *Let $f(t)$ be a 3 times continuously differentiable real-valued concave function on the ray $\{t > 0\}$ such that*

$$|f'''(t)| \leq 3\beta t^{-1}|f''(t)|, \ t > 0.$$

Then the function

$$F(x,t) = -\ln(f(t) - x) - \max[1, \beta^2] \ln t$$

is $(1 + \max[1, \beta^2])$-self-concordant barrier for the 2-dimensional convex domain

$$G_f = \mathrm{cl}\{(x,t) \in \mathbf{R}^2 \mid t > 0, \, x \leq f(t)\}.$$

Proposition 9.2.2 *Let $f(x)$ be a 3 times continuously differentiable real-valued convex function on the ray $\{x > 0\}$ such that*

$$|f'''(x)| \leq 3\beta x^{-1} f''(x), \ x > 0.$$

Then the function

$$F(t,x) = -\ln(t - f(x)) - \max[1, \beta^2] \ln x$$

is $(1 + \max[1, \beta^2])$-self-concordant barrier for the 2-dimensional convex domain

$$G^f = \mathrm{cl}\{(t,x) \in \mathbf{R}^2 \mid x > 0, t \geq f(x)\}.$$

To prove Proposition 9.2.1, let us set

- $G^+ = \mathbf{R}_+ \ [K = \mathbf{R}_+]$,
- $F^+(u) = -\ln u \ \ [\vartheta_+ = 1]$,
- $G^- = \{(x,t) \mid t \geq 0\}$,
- $F^-(x,t) = -\ln t \ \ [\vartheta_- = 1]$,
- $\mathcal{A}(x,t) = f(t) - x$,

which results in

$$G = \mathrm{cl}\{(x,t) \mid t > 0, \, x \leq f(t)\}.$$

The assumptions on f say exactly that $\mathcal{A}$ is β-appropriate for G^+, so that the conclusion in Proposition 9.2.1 is immediately given by Theorem 9.1.1.

To get Proposition 9.2.2, it suffices to apply Proposition 9.2.1 to the image of the set G^f under the mapping $(x,t) \mapsto (t, -x)$. ■

Example 9.2.1 [epigraph of the increasing power function] *Whenever $p \geq 1$, the function*

$$-\ln t - \ln(t^{1/p} - x)$$

is 2-self-concordant barrier for the epigraph

$$\{(x,t) \in \mathbf{R}^2 \mid t \geq (x_+)^p \equiv [\max\{0, x\}]^p\}$$

of the power function $(x_+)^p$, and the function

$$-2\ln t - \ln(t^{2/p} - x^2)$$

is 4-self-concordant barrier for the epigraph

$$\{(x,t) \in \mathbf{R}^2 \mid t \geq |x|^p\}$$

of the function $|x|^p$.

The result on the epigraph of $(x_+)^p$ is given by Proposition 9.2.1 with $f(t) = t^{1/p}, \beta = \frac{2p-1}{3p}$; to get the result on the epigraph of $|x|^p$, take the sum of the already known to us barriers for the epigraphs E_+, E_- of the functions $(x_+)^p$ and $([-x]_+)^p$, thus obtaining the barrier for $E_- \cap E_+$, which is exactly the epigraph of $|x|^p$. $\blacksquare$

Example 9.2.2 [epigraph of decreasing power function] *The function*

$$\begin{cases} -\ln x - \ln(t - x^{-p}), & 0 < p \leq 1 \\ -\ln t - \ln(x - t^{-1/p}), & p > 1 \end{cases}$$

is 2-self-concordant barrier for the epigraph

$$\mathrm{cl}\{(x,t) \in \mathbf{R}^2 \mid t \geq x^{-p}, x > 0\}$$

of the function x^{-p}, $p > 0$.

The case of $0 < p \leq 1$ is given by Proposition 9.2.2 applied with $f(x) = x^{-p}, \beta = \frac{2+p}{3}$. The case of $p > 1$ can be reduced to the former one by swapping x and t. $\blacksquare$

Example 9.2.3 [epigraph of the exponent] *The function*

$$-\ln t - \ln(\ln t - x)$$

is 2*-self-concordant barrier for the epigraph*

$$\{(x,t) \in \mathbf{R}^2 \mid t \geq \exp\{x\}\}$$

of the exponent.

Proposition 9.2.1 applied with $f(t) = \ln t, \beta = \frac{2}{3}$ $\blacksquare$

Example 9.2.4 [epigraph of the entropy function] *The function*

$$-\ln x - \ln(t - x \ln x)$$

is 2-self-concordant barrier for the epigraph

$$\mathrm{cl}\{(x,t) \in \mathbf{R}^2 \mid t \geq x \ln x, x > 0\}$$

of the entropy function $x \ln x$.

Proposition 9.2.2 applied to $f(x) = x \ln x, \beta = \frac{1}{3}$ $\blacksquare$

The indicated examples allow to handle those of the constraints defining the feasible set G which are *separable*, i.e., are of the type

$$\sum_i f_i(a_i^T x + b_i),$$

f_i being a convex function on the axis. To make this point clear, let us look at the typical example - the $|\cdot|_p$-approximation problem (L_p). Introducing N additional variables t_i, we can rewrite this problem equivalently as

$$\textit{minimize} \quad \sum_{i=1}^N t_i \quad \textit{s.t.} \quad t_i \geq |v_i - u_i^T x|^p, \; i = 1, ..., N,$$

so that now there are N "simple" constraints rather than a single, but "complicated" one. Now, the feasible set of i-th of the "simple" constraints is the inverse image of the epigraph of the increasing power function under an *affine* mapping, so that the feasible domain G of the reformulated problem admits the following explicit self-concordant barrier (Example 9.2.1 plus the usual Decomposition and Substitution rules):

$$F(t,x) = -\sum_{i=1}^N [\ln(t_i^{2/p} - (v_i - u_i^T x)^2) + 2\ln t_i]$$

with the parameter $4N$.

9.3 Fractional-Quadratic Substitution

Now let me indicate an actually marvellous nonlinear substitution: the fractional-quadratic one. The simplest form of this substitution is

$$\mathcal{A}(\tau,\xi,\eta)=\tau-\frac{\xi^2}{\eta}$$

(ξ, η, τ are real variables and $\eta>0$); the general case is given by "vectorization" of the numerator and the denominator and looks as follows:

Given are

- [numerator] A symmetric bilinear mapping
$$Q[\xi',\xi'']:\mathbf{R}^n\times\mathbf{R}^n\to\mathbf{R}^N$$
so that the coordinates $Q_i[\xi',\xi'']$ of the image are of the form
$$Q_i[\xi',\xi'']=(\xi')^TQ_i\xi''$$
with symmetric $n\times n$ matrices Q_i;
- [denominator] A symmetric $n\times n$ matrix $A(\eta)$ affinely depending on certain vector $\eta\in\mathbf{R}^q$.

The indicated data define the *general fractional-quadratic mapping*

$$\mathcal{A}(\tau,\xi,\eta)=\tau-Q[A^{-1}(\eta)\xi,\xi]:\mathbf{R}^q_\eta\times\mathbf{R}^n_\xi\times\mathbf{R}^N_\tau\to\mathbf{R}^N;$$

it turns out that this mapping is, under reasonable restrictions, appropriate for domains in $\mathbf{R}^N$. To formulate the restrictions, note first that $\mathcal{A}$ is not necessarily everywhere defined, since the matrix $A(\eta)$ may, for some η, be singular. Therefore it is reasonable to restrict η to vary in certain closed convex domain $Y\in\mathbf{R}^q_\eta$; thus, from now on the mapping $\mathcal{A}$ is considered along with the domain Y where η varies. The conditions which ensure that $\mathcal{A}$ is compatible with a given closed convex domain $G^+\subset\mathbf{R}^N$ are as follows:

(A): $A(\eta)$ is positive definite for $\eta\in\operatorname{int}Y$;

(B): the bilinear form $Q[A^{-1}(\eta)\xi',\xi'']$ of ξ',ξ'' is symmetric in ξ', ξ'' for any $\eta\in\operatorname{int}Y$;

(C): the quadratic form $Q[\xi,\xi]$ takes its values in the recessive cone K of the domain G^+.

Proposition 9.3.1 *Under assumptions (A) - (C) the mappings*

$$\mathcal{A}(\tau,\xi,\eta)=\tau-Q[A^{-1}(\eta)\xi,\xi]:G^-\equiv Y\times\mathbf{R}^n_\xi\times\mathbf{R}^N_\tau\to\mathbf{R}^N$$

and

$$\mathcal{B}(\xi,\eta)=-Q[A^{-1}(\eta)\xi,\xi]:G^-\equiv Y\times\mathbf{R}^n_\xi\to\mathbf{R}^N$$

are 1-appropriate for G^+.

In particular, if F^+ is ϑ_+-self-concordant barrier for G^+ and F_Y is a ϑ_Y-self-concordant barrier for Y, then

$$F(\tau,\xi,\eta)=F^+(\tau-Q[A^{-1}(\eta)\xi,\xi])+F_Y(\eta)$$

is $(\vartheta_++\vartheta_Y)$-self-concordant barrier for the closed convex domain

$$G=\operatorname{cl}\{(\tau\xi,\eta)\mid\tau-Q[A^{-1}(\eta)\xi,\xi]\in\operatorname{int}G^+,\ \eta\in\operatorname{int}Y\}.$$

The proof of the proposition is given in Section 9.5. What we are about to do now is to present several examples.

Example 9.3.1 [epigraph of convex quadratic form] *Let $f(x)=x^TP^TPx+b^Tx+c$ be a convex quadratic form on $\mathbf{R}^n$; then the function*

$$F(t,x)=-\ln(t-f(x))$$

is 1-self-concordant barrier for the epigraph

$$\{(x,t)\in\mathbf{R}^n\times\mathbf{R}\mid t\geq f(x)\}.$$

Let the "fractional-quadratic" data be defined as follows:

- $G^+ = \mathbf{R}_+$ $[N = 1]$;
- $Q[\xi', \xi''] = (\xi')^T \xi''$, $\xi', \xi'' \in \mathbf{R}^n$;
- $\mathbf{R}^q_\eta = \mathbf{R} = Y, A(\eta) \equiv I$

  (from now on I stands for the identity operator).

Conditions (A) - (C) clearly are satisfied; Proposition 9.3.1 applied with

$$F^+(\tau) = -\ln \tau, \quad F_Y(\cdot) \equiv 0$$

says that the function

$$F(\tau, \xi, \eta) = -\ln(\tau - \xi^T \xi)$$

is 1-self-concordant barrier for the closed convex domain

$$G = \{(\tau, \xi, \eta) \mid \tau \geq \xi^T \xi\}.$$

The epigraph of the quadratic form f clearly is the inverse image of G under the affine mapping

$$(t, x) \mapsto \begin{pmatrix} \tau = t - b^T x - c \\ \xi = Px \\ \eta = 0 \end{pmatrix},$$

and it remains to apply the Substitution rule (L). ■

The result stated in the latter example is not difficult to establish directly, which hardly can be said about the following

Example 9.3.2 [barrier for the second-order cone] *The function*

$$F(t, x) = -\ln(t^2 - x^T x)$$

is 2-logarithmically homogeneous self-concordant barrier for the second order cone

$$K^2_n = \{(t, x) \in \mathbf{R} \times \mathbf{R}^n \mid t \geq \sqrt{x^T x}\}.$$

Let the "fractional-quadratic" data be defined as follows:

- $G^+ = \mathbf{R}_+$ $[N = 1]$;
- $Q[\xi', \xi''] = (\xi')^T \xi''$, $\xi', \xi'' \in \mathbf{R}^n$;
- $Y = \mathbf{R}_+ \subset \mathbf{R} \equiv \mathbf{R}^q_\eta$, $A(\eta) \equiv \eta I$.

Conditions (A) - (C) clearly are satisfied; Proposition 9.3.1 applied with

$$F^+(\tau) = -\ln \tau, \; F_Y(\eta) = -\ln \eta$$

says that the function

$$F(\tau, \xi, \eta) = -\ln(\tau - \eta^{-1} \xi^T \xi) - \ln \eta \equiv -\ln(\tau \eta - \xi^T \xi)$$

is 2-self-concordant barrier for the closed convex domain

$$G = \mathrm{cl}\{(\tau, \xi, \eta) \mid \tau > \eta^{-1} \xi^T \xi, \; \eta > 0\}.$$

The second order cone K^2_n clearly is the inverse image of G under the affine mapping

$$(t, x) \mapsto \begin{pmatrix} \tau = t \\ \xi = x \\ \eta = t \end{pmatrix},$$

and to prove that $F(t, x)$ is 2-self-concordant barrier for the second order cone, it remains to apply the Substitution rule (L). Logarithmic homogeneity of $F(t, x)$ is evident. ■

The next example originally required somewhere 15-page "brut force" justification which was by far more complicated than the justification of the general results presented in this lecture.

Example 9.3.3 [epigraph of the spectral norm of a matrix] *The function*

$$F(t,x) = -\ln \mathrm{Det}\,(tI - t^{-1}x^Tx) - \ln t$$

is $(m+1)$*-logarithmically homogeneous self-concordant barrier for the epigraph*

$$\{(t,x) \mid t \in \mathbf{R},\ x \text{ is an } m\times k \text{ matrix of the spectral norm} \le t\}.$$

of the spectral norm of $k \times m$ *matrix* x [1].

Let the "fractional-quadratic" data be defined as follows:

- $G^+ = \mathbf{S}^m_+$ is the cone of positive semidefinite $m\times m$ matrices $[N = m(m+1)/2]$;
- $Q[\xi', \xi''] = \frac{1}{2}[(\xi')^T\xi'' + (\xi'')^T\xi']$, ξ', ξ'' are $k\times m$ matrices;
- $Y = \mathbf{R}_+ \subset \mathbf{R} \equiv \mathbf{R}^q_\eta$, $A(\eta) \equiv \eta I$.

Conditions (A) - (C) clearly are satisfied; Proposition 9.3.1 applied with

$$F^+(\tau) = -\ln \mathrm{Det}\ \tau,\ F_Y(\eta) = -\ln\eta$$

says that the function

$$F(\tau,\xi,\eta) = -\ln(\tau - \eta^{-1}\xi^T\xi) - \ln\eta$$

is $(m+1)$-self-concordant barrier for the closed convex domain

$$G = \mathrm{cl}\{(\tau,\xi,\eta) \mid \tau - \eta^{-1}\xi^T\xi \in \mathrm{int}\ \mathbf{S}^m_+,\ \eta > 0\}.$$

The spectral norm of a $k\times m$ matrix x is $< t$ if and only if the maximum eigenvalue of the matrix x^Tx is $< t^2$, or, which is the same, if the $m\times m$ matrix $tI - t^{-1}x^Tx$ is positive definite; thus, the epigraph of the spectral norm of x is the inverse image of G under the affine mapping

$$(t,x) \mapsto \begin{pmatrix} \tau = tI \\ \xi = x \\ \eta = t \end{pmatrix},$$

and to prove that $F(t,x)$ is $(m+1)$-self-concordant barrier for the epigraph of the spectral norm, it suffices to apply the Substitution rule (L). The logarithmic homogeneity of $F(t,x)$ is evident. ■

The indicated examples of self-concordant barriers are sufficient for applications which will be our goal in the remaining lectures; at the same time, these examples explain how to use the general results of the lecture to obtain barriers for other convex domains.

9.4 Proof of Theorem 10.1

.

A. Let us prove that G^0 is an open convex domain in $\mathbf{R}^n$. Indeed, since $\mathcal{A}$ is continuous on int G^-, G^0 clearly is open; thus, all we need is to demonstrate that G^0 is convex. Let $x', x'' \in G^0$, so that x', x'' are in int G^- and $y' = \mathcal{A}(x')$, $y'' = \mathcal{A}(x'')$ are in int G^+, and let $\alpha \in [0,1]$. We should prove that $x \equiv \alpha x' + (1-\alpha)x'' \in G^0$, i.e., that $x \in \mathrm{int}\ G^-$ (which is evident) and that $y = \mathcal{A}(x) \in \mathrm{int}\ G$. To prove the latter inclusion, it suffices to demonstrate that

$$y \ge_K \alpha y' + (1-\alpha)y''; \tag{9.2}$$

[1] the spectral norm of a $k\times m$ matrix x is the maximum eigenvalue of the matrix $\sqrt{x^Tx}$ or, which is the same, the norm

$$\max\{|x\xi|_2 \mid \xi \in \mathbf{R}^m, |\xi|_2 \le 1\}$$

of the linear operator from $\mathbf{R}^m$ into $\mathbf{R}^k$ given by x

indeed, the right hand side in this inequality belongs to int G^+ together with y', y''; since K is the recessive cone of G^+, the translation of any vector from int G^+ by a vector form K also belongs to int G^+, so that (9.2) - which says that y is a translation of the right hand side by a direction from K would imply that $y \in \text{int } G^+$.

To prove (9.2) is the same as to demonstrate that

$$s^T y \geq s^T(\alpha y' + (1-\alpha)y'') \tag{9.3}$$

for any $s \in K^* \equiv \{s \mid s^T u \geq 0 \;\; \forall u \in K\}$ (why?) But (9.3) is immediate: the real-valued function

$$f(z) = s^T \mathcal{A}(z)$$

is concave on int G^-, since $D^2\mathcal{A}(z)[h,h] \leq_K 0$ (Definition 9.1.1.(i)) and, consequently,

$$D^2 f(z)[h,h] = s^T D^2\mathcal{A}(z)[h,h] \leq 0$$

(recall that $s \in K^*$); since $f(z)$ is concave, we have

$$s^T y = f(\alpha x' + (1-\alpha)x'') \geq \alpha f(x') + (1-\alpha)f(x'') = \alpha s^T y' + (1-\alpha)s^T y'',$$

as required in (9.3).

B. Now let us prove self-concordance of F. To this end let us fix $x \in G^0$ and $h \in \mathbf{R}^n$ and verify that

$$|D^3F(x)[h,h,h]| \leq 2\{D^2F(x)[h,h]\}^{3/2}, \tag{9.4}$$

$$|DF(x)[h]| \leq \vartheta^{1/2}\{D^2F(x)[h,h]\}^{1/2}. \tag{9.5}$$

B.1. Let us start with writing down the derivatives of F. Under notation

$$a = \mathcal{A}(x), \;\; a' = D\mathcal{A}(x)[h], \;\; a'' = D^2\mathcal{A}(x)[h,h], \;\; a''' = D^3\mathcal{A}(x)[h,h,h],$$

we have

$$DF(x)[h] = DF^+(a)[a'] + \gamma^2 DF^-(x)[h], \;\; \gamma = \max[1,\beta], \tag{9.6}$$

$$D^2F(x)[h,h] = D^2F^+(a)[a',a'] + DF^+(a)[a''] + \gamma^2 D^2F^-(x)[h,h], \tag{9.7}$$

$$D^3F(x)[h,h,h] = D^3F^+(a)[a',a',a'] + 3DF^+(a)[a',a''] + DF^+(a)[a'''] + \gamma^2 D^3F^-(x)[h,h,h]. \tag{9.8}$$

B.2. Now let us summarize our knowledge on the quantities involved into (9.6) - (9.8).

Since F^+ is ϑ_+-self-concordant barrier, we have

$$|DF^+(a)[a']| \leq p\sqrt{\vartheta_+}, \;\; p \equiv \sqrt{D^2F^+(a)[a',a']}, \tag{9.9}$$

$$|D^3F^+(a)[a',a',a']| \leq 2p^3. \tag{9.10}$$

Similarly, since F^- is ϑ_--self-concordant barrier, we have

$$|DF^-(x)[h]| \leq q\sqrt{\vartheta_-}, \;\; q \equiv \sqrt{D^2F^-(x)[h,h]}, \tag{9.11}$$

$$|D^3F^-(x)[h,h,h]| \leq 2q^3. \tag{9.12}$$

Besides this, from Corollary 3.2.1 (Lecture 3) we know that $DF^+(a)[\cdot]$ is nonpositive on the recessive directions of G^+:

$$DF^+(a)[g] \leq 0, \;\; g \in K, \tag{9.13}$$

and even that

$$\{D^2F^+(a)[g,g]\}^{1/2} \leq -DF^+(a)[g], \;\; g \in K. \tag{9.14}$$

B.3. Let us prove that

$$3\beta q a'' \leq_K a''' \leq_K -3\beta q a''. \tag{9.15}$$

Indeed, let a real t be such that $|t|q \leq 1$, and let $h_t = th$; then $D^2F^-(x)[h_t,h_t] = t^2q^2 \leq 1$ and, consequently, $x \pm h_t \in G^-$ (**I.**, Lecture 2). Therefore Definition 9.1.1.(ii) implies that

$$t^3 a''' \equiv D^3\mathcal{A}(x)[h_t,h_t,h_t] \leq_K -3\beta D^2\mathcal{A}(x)[h_t,h_t] \equiv -3\beta t^2 a'';$$

since the inequality $t^3 a''' \leq_K -3\beta t^2 a''$ is valid for all t with $|t|q \leq 1$, (9.15) follows.

Note that from (9.13) and (9.15) it follows that the quantity

$$r \equiv \sqrt{DF^+(a)[a'']} \tag{9.16}$$

is well-defined and is such that

$$|DF^+(a)[a''']| \leq 3\beta q r^2. \tag{9.17}$$

Besides this, by Cauchy's inequality

$$|D^2F^+(a)[a', a'']| \leq \sqrt{D^2F^+(a)[a', a']}\sqrt{D^2F^+(a)[a'', a'']} \leq pr^2 \tag{9.18}$$

(the concluding inequality follows from (9.14).

B.4. Subsituting (9.9), (9.11) into (9.6), we come to

$$|DF(x)[h]| \leq p\sqrt{\vartheta_+} + q\gamma^2\sqrt{\vartheta_-}; \tag{9.19}$$

substituting (9.16) into (9.7), we get

$$D^2F(x)[h, h] = p^2 + r^2 + \gamma^2 q^2, \tag{9.20}$$

while substituting (9.10), (9.12), (9.17), (9.18) into (9.8), we obtain

$$|D^2F(x)[h, h, h]| \leq 2[p^3 + \frac{3}{2}pr^2 + \frac{3}{2}\beta q r^2] + 2\gamma^2 q^3. \tag{9.21}$$

By passing from q to $s = \gamma q$, we come to inequalitites

$$|DF(x)[h]| \leq \sqrt{\vartheta_+}p + \sqrt{\vartheta_-}\gamma s, \;\; D^2F(x)[h, h] = p^2 + r^2 + s^2,$$

and

$$|D^3F(x)[h, h, h]| \leq 2[p^3 + \frac{3}{2}pr^2 + \frac{3}{2}\frac{\beta}{\gamma}sr^2] + 2\gamma^{-1}s^3 \leq$$

[since $\gamma \geq \beta$ and $\gamma \geq 1$]

$$\leq 2[p^3 + s^3 + \frac{3}{2}r^2(p + s)] \leq$$

[straightforward computation]

$$\leq 2[p^2 + r^2 + s^2]^{3/2}.$$

Thus,

$$|DF(x)[h]| \leq \sqrt{\vartheta_+ + \gamma^2\vartheta_-}\{D^2F(x)[h, h]\}^{1/2}, \;\; |D^3F(x)[h, h, h]| \leq 2\{D^2F(x)\}^{1/2}. \tag{9.22}$$

C. (9.22) says that F satisfies the differential inequalities required by the definition of a γ^2-self-concordant barrier for $G = \mathrm{cl}\, G_0$. To complete the proof, we should demonstrate that F is a barrier for G, i.e., that $F(x_i) \to \infty$ whenever $x_i \in G_0$ are such that $x \equiv \lim_i x_i \in \partial G$. To prove the latter statement, set

$$y_i = \mathcal{A}(x_i)$$

and consider two possible cases:

C.1: $x \in \mathrm{int}\ G^-$;

C.2: $x \in \partial G^-$.

In the easy case of **C.1** there exists $y = \lim_i y_i = \mathcal{A}(x)$, since $\mathcal{A}$ is continuous on the interior of G^- and, consequently, in a neighbourhood of x. Since $x \notin G^0$, $y \notin \mathrm{int}\ G^+$, so that the sequence y_i comprised of the interior points of G^+ converges to a boundary point of g^+ and therefore $F^+(y_i) \to \infty$. Since x_i converge to an interior point of G^-, the sequence $F^-(x_i)$ is bounded, and the sequence $F(x_i) = F^+(y_i) + \gamma^2 F^-(x_i)$ diverges to $+\infty$, as required.

Now consider the more difficult case when $x \in \partial G^-$. Here we know that $F^-(x_i) \to \infty$ (since x_i converge to a boundary point for the domain G^- for which F^- is a self-concordant barrier); in order to prove that $F(x_i) \equiv F^+(y_i) + \gamma^2 F^-(x_i) \to \infty$ it suffices, therefore, to prove that the sequence $F^+(y_i)$ is below bounded. From concavity of $\mathcal{A}$ we have (compare with **A**)

$$y_i = \mathcal{A}(x_i) \leq_K \mathcal{A}(x_0) + D\mathcal{A}(x_0)[x_i - x_0] \equiv z_i,$$

whence, by Corollary 3.2.1, Lecture 3,

$$F^+(y_i) \geq F^+(z_i).$$

Now below boundedness of $F^+(y_i)$ is an immediate conseqeunce of the fact that the sequence $F^+(z_i)$ is below bounded (indeed, $\{x_i\}$ is a bounded sequence, and consequently its image $\{z_i\}$ under affine mapping also is bounded; and convex function F^+ is below bounded on any bounded subset of its domain). ■

9.5 Proof of Proposition 10.1

.

A. Looking at the definition of an appropriate mapping and taking into account that $\mathcal{B}$ is the restriction of $\mathcal{A}$ onto a cross-section of the domain of $\mathcal{A}$ and an affine plane $t = 0$, we immediately conclude that it suffices to prove that $\mathcal{A}$ is 1-appropriate for G^+. Of course, $\mathcal{A}$ is 3 times continuously differentiable on the interior of G^-.

B. The coordinates of the vector $Q[A^{-1}(\eta)\xi', \xi'']$ are bilinear forms $(\xi')^T A^{-1}(\eta) Q_i \xi''$ of ξ', ξ''; by (B), they are symmetric in ξ', ξ'', so that the matrices $A^{-1}(\eta)Q_i$ are symmetric. Since both $A^{-1}(\eta)$ and Q_i are symmetric, their product can be symmetric if and only if the matrices commutate. Since $A^{-1}(\eta)$ commutate with Q_i, $\eta \in \operatorname{int} Y$, and Y is open, $A(\eta)$ commutate with Q_i for all η. Thus, we come to the following crucial conclusion:

for every $i \leq N$, the matrix $A(\eta)$ commutates with Q_i for all η.

C. Let us compute the derivatives of $\mathcal{A}$ at a point $X = (\tau, \xi, \eta) \in \operatorname{int} G^-$ in a direction $\Xi = (t, x, y)$. In what follows subscript i marks i-th coordinate of a vector from $\mathbf{R}^N$. Note that from **B.** it follows that Q_i commutates with $\alpha(\cdot) \equiv A^{-1}(\cdot)$ and therefore with all derivatives of $\alpha(\cdot)$; with this observation, we immediately obtain

$$\mathcal{A}_i(X) = \tau_i - \xi^T \alpha(\eta) Q_i \xi;$$

$$D\mathcal{A}_i(X)[\Xi] = t_i - 2x^T \alpha(\eta) Q_i \xi - \xi^T [D\alpha(\eta)[y]] Q_i \xi;$$

$$D^2\mathcal{A}_i(X)[\Xi, \Xi] = -2x^T \alpha(\eta) Q_i x - 4x^T [D\alpha(\eta)[y]] Q_i \xi - \xi^T [D^2\alpha(\eta)[y, y]] Q_i \xi;$$

$$D^3\mathcal{A}_i(X)[\Xi, \Xi, \Xi] = -6x^T [D\alpha(\eta)[y]] Q_i x - 6x^T [D^2\alpha(\eta)[y]] Q_i \xi - \xi^T [D^3\alpha(\eta)[y, y, y]] Q_i \xi.$$

Now, denoting

$$\alpha = \alpha(\eta), \;\; a' = DA(\eta)[y], \tag{9.23}$$

we immediately get

$$D\alpha(\eta)[y] = -\alpha a' \alpha, \;\; D^2\alpha(\eta)[y, y] = 2\alpha a' \alpha a' \alpha,$$

$$D^3\alpha(\eta)[y, y, y] = -6\alpha a' \alpha a' \alpha a' \alpha.$$

Substituting the expressions for the derivatives of $\alpha(\cdot)$ in the expressions for the dreivatives of $\mathcal{A}_i$, we come to

$$D^2\mathcal{A}_i(X)[\Xi, \Xi] = -2\zeta^T \alpha Q_i \zeta, \;\; \zeta = x - a' \alpha \xi, \tag{9.24}$$

and

$$D^3\mathcal{A}_i(X)[\Xi, \Xi, \Xi] = 6\zeta^T \alpha a' \alpha Q_i \zeta \tag{9.25}$$

(the simplest way to realize why "we come to" is to substitute in the latter two right hand sides the expression for ζ and to open the parentheses, taking into account that α and a' are symmetric and commutate with Q_i).

D. Now we are basically done. First, α commutates with Q_i and is positive definite in view of condition (A) (since $\alpha = A^{-1}(\eta)$ and $\eta \in \operatorname{int} Y$). It follows that $\alpha^{1/2}$ also commutates with Q_i, so that (9.24) can be rewritten as

$$D^2\mathcal{A}_i(X)[\Xi] = -2[\sqrt{\alpha}\zeta]^T Q_i [\sqrt{\alpha}\zeta],$$

which means that

$$D^2\mathcal{A}(X)[\Xi,\Xi] = -2Q[\omega,\omega]$$

for certain vector ω, so that

$$D^2\mathcal{A}(X)[\Xi,\Xi] \leq_K 0$$

according to (C). Thus, $\mathcal{A}$ is concave with respect to the recessive cone K of the domain G^+, as is required by item (i) of Definition 9.1.1.

It requires to verify item (ii) of the Definition for the case of $\beta = 1$, , i.e., to prove that

$$D^3\mathcal{A}(X)[\Xi,\Xi,\Xi] + 3D^2\mathcal{A}(X)[\Xi,\Xi] \leq_K 0$$

whenever Ξ is such that $X \pm \Xi \in G^-$. This latter inclusion means that $\eta \pm y \in Y$, so that $A(\eta \pm y)$ is positive semidefinite; since $A(\cdot)$ is affine, we conclude that

$$B = A(\eta) - DA(\eta)[y] \equiv \alpha^{-1} - a' \geq 0$$

(as always, ≥ 0 for symmetric matrices stands for "positive semidefinite"), whence also

$$\gamma = \alpha[\alpha^{-1} - a']\alpha \geq 0.$$

From (9.24), (9.25) it follows that

$$D^3\mathcal{A}_i(X)[\Xi,\Xi,\Xi] + 3D^2\mathcal{A}_i(X)[\Xi,\Xi] = -6\zeta^T\gamma Q_I\zeta,$$

and since γ is positive semidefinite and, due to its origin, commutates with Q_i (since α and a' do), we have $\zeta^T\gamma Q_i\zeta = \zeta^T\gamma^{1/2}Q_i\gamma^{1/2}\zeta$, so that

$$D^3\mathcal{A}(X)[\Xi,\Xi,\Xi] + 3D^2\mathcal{A}(X)[\Xi,\Xi] = -6Q[\gamma^{1/2}\zeta,\gamma^{1/2}\zeta] \leq_K 0$$

(the concluding inequality follows from (C)). ■

9.6 Exercises on constructing self-concordant barriers

The goal of the below exercises is to derive some new self-concordant barriers.

9.6.1 Epigraphs of functions of Euclidean norm

Exercise 9.6.1 $^{\#+}$ *Let G^+ be a closed convex domain in $\mathbf{R}^2$ which contains a point with both coordinates being positive and is "antimonotone in the x-direction", i.e., such that $(u,s) \in G^+ \Rightarrow (v,s) \in G^+$ whenever $v \le u$, and let F^+ be a ϑ_+-self-concordant barrier for G. Prove that*

1) The function

$$F^1(t,x) = F^+(x^T x, t)$$

is ϑ_+-self-concordant barrier for the closed convex domain

$$G^1 = \{(x,t) \in \mathbf{R}^n \times \mathbf{R} \mid (x^T x, t) \in G\}.$$

Derive from this observation that if $p \le 2$, then the function

$$F(t,x) = -\ln(t^{2/p} - x^T x) - \ln t$$

is 2-self-concordant barrier for the epigraph of the function $|x|_2^p$ on $\mathbf{R}^n$.

2) The function

$$F^2(t,x) = F^+(\frac{x^T x}{t}, t) - \ln t$$

is $(\vartheta_+ + 1)$-self-concordant barrier for the closed convex domain

$$G^2 = \mathrm{cl}\{(x,s) \in \mathbf{R}^n \times \mathbf{R} \mid (\frac{x^T x}{t}, t) \in G,\ t > 0\}.$$

Derive from this observation that if $1 \le p \le 2$, then the function

$$F(t,x) = -\ln(t^{2/p} - x^T x) - \ln t$$

is 3-self-concordant barrier for the epigraph of the function $|x|_2^p$ on $\mathbf{R}^n$.

9.6.2 How to guess that $-\ln \mathrm{Det}\, x$ is a self-concordant barrier

Now our knowledge on concrete self-concordant barriers is as follows. We know two "building blocks" - the barriers $-\ln t$ for the nonnegative half-axis and $-\ln \mathrm{Det}\, x$ for the cone of positive semidefinite symmetric matrices; the fact that these barriers are self-concordant was justified by straightforward computation, completely trivial for the former and not that difficult for the latter barrier. All other self-concordant barriers were given by these two via the Substitution rule (N). It turns out that the barrier $-\ln \mathrm{Det}\, x$ can be not only *guessed*, but also *derived* from the barrier $-\ln t$ via the same Substitution rule (N), so that in fact only one barrier should be guessed.

Exercise 9.6.2 $^{\#}$

1) Let

$$A = \begin{pmatrix} \tau & \xi^T \\ \xi & \eta \end{pmatrix}$$

be a symmetric matrix (τ is $p \times p$, η is $q \times q$). Prove that A is positive definite if and only if both the matrices η and $\tau - \xi^T \eta^{-1} \xi$ are positive definite; in other words, the cone $\mathbf{S}_+^{p+q}$ of positive semidefinite symmetric $(p+q) \times (p+q)$ matrices is the inverse image G, in terms of Substitution rule (N), of the cone $G^+ = \mathbf{S}^p$ under the fractional-quadratic mapping

$$\mathcal{A} : (\tau, \xi, \eta) \mapsto \tau - \xi^T \eta^{-1} \xi$$

with the domain of the mapping $\{(\tau, \xi, \eta) \mid \eta \in Y \equiv \mathbf{S}_+^q\}$.

2) Applying Proposition 9.3.1, derive from 1), that if F_p and F_q are self-concordant barriers for $\mathbf{S}^p_+$, $\mathbf{S}^q_+$ with parameters ϑ_p, ϑ_q, respectively, then the function

$$F(A) \equiv F(\tau, \xi, \eta) = F_p(\tau - \xi^T \eta^{-1} \xi) + F_q(\eta)$$

is $(\vartheta_p + \vartheta_q)$-self-concordant barrier for $\mathbf{S}^{p+q}_+$.

3) Use the observation that $-\ln \eta$ is 1-self-concordant barrier for $\mathbf{S}^1_+ \equiv \mathbf{R}_+$ to prove by induction on p that $F_p(x) = -\ln \mathrm{Det}\, x$ is p-self-concordant barrier for $\mathbf{S}^p_+$.

9.6.3 ”Fractional-quadratic” cone and Truss Topology Design

Consider the following hybride of the second-order cone and the cone $\mathbf{S}^{\cdot}_+$: let $\xi_1, ..., \xi_q$ be variable matrices of the sizes $n_1 \times m$,...,$n_q \times m$, τ be $m \times m$ variable matrix and $y_j(\eta)$, $j = 1, ..., q$, be symmetric $n_j \times n_j$ matrices which are linear homogeneous functions of $\eta \in \mathbf{R}^k$. Let Y be certain cone in $\mathbf{R}^k$ (closed, convex and with a nonempty interior) such that $y_j(\eta)$ are positive definite when $\eta \in \mathrm{int}\, Y$.

Consider the set

$$\mathcal{K} = \mathrm{cl}\{(\tau; \eta; \xi_1, ..., \xi_q) \mid \tau \geq \xi_1^T y_1^{-1}(\eta) \xi_1 + ... + \xi_q^T y_q^{-1}(\eta) \xi_q,\ \eta \in \mathrm{int}\, Y\}.$$

Let also $F_Y(\eta)$ be a ϑ_Y-self-concordant barrier for Y.

Exercise 9.6.3 $^+$ *Prove that $\mathcal{K}$ is a closed convex cone with a nonempty interior, and that the function*

$$F(\tau; \eta; \xi_1, ..., \xi_q) = -\ln \mathrm{Det}\, \left(\tau - \xi_1^T y_1^{-1}(\eta) \xi_1 - ... - \xi_q^T y_q^{-1}(\eta) \xi_q\right) + F_Y(\eta) \qquad (9.26)$$

is $(m + \vartheta_Y)$-self-concordant barrier for $\mathcal{K}$; this barrier is logarithmically homogeneous, if F_Y is.

Prove that $\mathcal{K}$ is the inverse image of the cone $\mathbf{S}^N_+$ of positive semidefinite $N \times N$ symmetric matrices, $N = m + n_1 + ... + n_q$, under the linear homogeneous mapping

$$\mathcal{L} : (\tau; \eta; \xi_1, ..., \xi_q) \mapsto \begin{pmatrix} \tau & \xi_1^T & \xi_2^T & \xi_3^T & ... & \xi_q^T \\ \xi_1 & y_1(\eta) & & & & \\ \xi_2 & & y_2(\eta) & & & \\ \xi_3 & & & y_3(\eta) & & \\ ... & ... & ... & ... & ... & ... \\ \xi_q & & & & & y_q(\eta) \end{pmatrix}$$

(blank space corresponds to zero blocks). Is the barrier (9.26) the barrier induced, via the Substitution rule (L), by the mapping $\mathcal{L}$ and the standard barrier $-\ln \mathrm{Det}\,(\cdot)$ for $\mathbf{S}^N_+$?

Now we are in a position to complete, in a sense, our considerations related to the Truss Topology Design problem (Section 5.7, Lecture 5). To the moment we know two formulations of the problem:

Dual form (TTD^d): *minimize t by choice of the vector $x = (t; \lambda_1, ..., \lambda_k; z_1, ..., z_m)$ (t and λ_j are reals, $z_i \in \mathbf{R}^n$) subject to the constraints*

$$t \geq \sum_{j=1}^{k} \left[2 z_j^T f_j + V \frac{(b_i^T z_j)^2}{\lambda_j} \right],\ i = 1, ..., m,$$

$$\lambda \geq 0;\ \sum_j \lambda_j = 1.$$

Primal form (ψ): *minimize t by choice of $x = (t; \phi; \beta_{ij})$ (t and β_{ij}, $i = 1, ..., m$, $j = 1, ..., k$ are reals, $\phi \in \mathbf{R}^m$) subject to the constraints*

$$t \geq \sum_{i=1}^{m} \frac{\beta_{ij}^2}{\phi_i},\ j = 1, ..., k;$$

$$\phi \geq 0;\ \sum_{i=1}^{m} \phi_i = V;$$

$$\sum_{i=1}^{m} \beta_{ij} b_i = f_j, \; j = 1, ..., k.$$

Both forms are respectable convex problems; the question, anyhow, is whether we are equipped enough to solve them via interior point machinery, or, in other words, are we clever enough to point out explicit self-concordant barriers for the corresponding feasible domains. The answer is positive.

Exercise 9.6.4 *Consider the problem* (TTD^d)*, and let*

$$x = (t; \lambda_1, ..., \lambda_k; z_1, ..., z_m)$$

be the design vector of the problem.

1) Prove that (TTD^d) *can be equivalently written down as the standard problem*

$$\textit{minimize} \;\; c^T x \equiv t \;\; \textit{s.t.} \;\; x \in G \subset E,$$

where

$$E = \{x \mid \sum_{j=1}^{k} \lambda_j = 1\}$$

is affine hyperplane in $\mathbf{R}^{\dim x}$ *and*

$$G = \{x \in E \mid x \textit{ is feasible for } (\mathrm{TTD}^d)\}$$

is a closed convex domain in E.

2)$^+$ *Let* $u = (s_i; t_{ij}; r_j)$ *(*i *runs over* $\{1, ..., m\}$*,* j *runs over* $\{1, ..., k\}$*,* $s.$*,* $t.$*,* $r.$ *are reals), and let*

$$\Phi(u) = -\sum_{i=1}^{m} \ln \left(s_i - \sum_{j=1}^{k} \frac{t_{ij}^2}{r_j} \right) - \sum_{j=1}^{k} \ln r_j.$$

Prove that Φ *is* $(m+k)$*-logarithmically homogeneous self-concordant barrier for the closed convex cone*

$$G^+ = \mathrm{cl}\{u \mid r_j > 0, \, j = 1, ..., k; s_i \geq \sum_{j=1}^{k} r_j^{-1} t_{ij}^2, \, i = 1, ..., m\},$$

and the Legendre transformation of the barrier is given by

$$\Phi^*(\sigma_i; \tau_{ij}; \rho_j) = -\sum_{j=1}^{k} \ln \left(-\rho_j + \sum_{i=1}^{m} \frac{\tau_{ij}^2}{4\sigma_i} \right) - \sum_{i=1}^{m} \ln(-\sigma_i) - (m+k),$$

the domain of Φ^* *being the set*

$$G^0 = \{\sigma_i < 0, \, i = 1, ..., m; -\rho_j + \sum_{i=1}^{m} \frac{\tau_{ij}^2}{4\sigma_i} > 0, \, j = 1, ..., k\}.$$

3) Prove that the domain G *of the standard reformulation of* (TTD^d) *given by 1) is the inverse image of* $G^\# = \mathrm{cl}\, G^0$ *under the affine mapping*

$$x \mapsto \pi x + p = \begin{pmatrix} s_i = t - 2\sum_{j=1}^{k} z_j^T f_j \\ t_{ij} = (b_i^T z_j)\sqrt{V} \\ r_j = \lambda_j \end{pmatrix}$$

(the mapping should be restricted onto E*).*

Conclude from this observation that one can equip G *with the* $(m+k)$*-self-concordant barrier*

$$F(x) = \Phi(\pi x + p)$$

and thus get the possibility to solve (TTD^d) *by the long-step path-following method.*

Note also that the problem

$$\textit{minimize}\ \ c^T x \equiv t\ \ \textit{s.t.}\ \ x \in E, \pi x + p \in G^+$$

is a conic reformulation of (TTD^d)*, and that* Φ *is a* $(m+k)$*-logarithmically homogeneous self-concordant barrier for the underlying cone* G^+*; since we know the Legendre transformation of* Φ*, we can solve the problem by the primal-dual potential reduction method as well.*

Note that the primal formulation (ψ) of TTD can be treated in completely similar way (since its formal structure is similar to that one of (TTD^d), up to presence of a larger number of linear equality constraints; linear *equalities* is something which does not influence our abilities to point out self-concordant barriers, due to the Substitution rule (L).

9.6.4 Geometrical mean

The below problems are motivated by by the following observation: the function ξ^2/η of two scalar variables is convex on the half-plane $\{\eta > 0\}$, and we know how to write down a self-concordant barrier for its epigraph - it is given by our marvellous fractional-quadratic substitution. How to get similar barrier for the epigraph of the function $(\xi_+)^p/\eta^{p-1}$ ($p > 1$ is integer), which, as it is easily seen, also is convex when $\eta > 0$?

The epigraph of the function $f(\xi, \eta) = (\xi_+)^p/\eta^{p-1}$ is the set

$$\mathrm{cl}\{(\tau, \xi, \eta) \mid \eta > 0,\ \tau\eta^{p-1} \geq (\xi_+)^p\}$$

This is a cone in $\mathbf{R}^3$, which clearly is the inverse image of the hypograph

$$G = \{(t, y_1, ..., y_p) \in \mathbf{R}^{p+1} \mid y_1, ..., y_p \geq 0, t \leq \phi(y) = (y_1...y_p)^{1/p}\}$$

under the affine mapping

$$\mathcal{L} : (\tau, \xi, \eta) \mapsto (\xi, \tau, \eta, \eta, ..., \eta),$$

so that the problem in question in fact is where to get a self-concordant barrier for the hypograph G of the geometrical mean. This latter question is solved by the following observation:

($\mathcal{G}$): *the mapping*

$$\mathcal{A}(t, y_1, ..., y_p) = (y_1...y_p)^{1/p} - t : G^- \to \mathbf{R},\ \ G^- = \{(t, y) \in \mathbf{R}^{p+1} \mid y \geq 0\}$$

is 1-appropriate for the domain $G^+ = \mathbf{R}_+$.

Exercise 9.6.5 $^+$ *Prove ($\mathcal{G}$).*

Exercise 9.6.6 $^+$ *Prove that the mapping*

$$\mathcal{B}(\tau, \xi, \eta) = \tau^{1/p}\eta^{(p-1)/p} - \xi : \mathrm{int}\ G^- \to \mathbf{R},\ \ G^- = \{(\tau, \xi, \eta) \mid \tau \geq 0, \eta \geq 0\}$$

is 1-appropriate for $G^+ = \mathbf{R}_+$.

Conclude from this observation that the function

$$F(\tau, \xi, \eta) = -\ln(\tau^{1/p}\eta^{(p-1)/p} - \xi) - \ln\tau - \ln\eta$$

is 3-logarithmically homogeneous self-concordant barrier for the cone

$$\mathrm{cl}\{(\tau, \xi, \eta) \mid \eta > 0,\ \tau \geq (\xi_+)^p\eta^{-(p-1)}\}$$

which is the epigraph of the function $(\xi_+)^p\eta^{-(p-1)}$.

Chapter 10

Applications in Convex Programming

To the moment we know several general schemes of polynomial time interior point methods; at the previous lecture we also have developed technique for constructing self-concordant barriers the methods are based on. It is time now to look how this machinery works. To this end let us consider several standard classes of convex programming problems. The order of exposition is as follows: for each class of problems in question, I shall present the usual description of the problem instances, the standard and conic reformulations required by the interior point approcah, the related self-concordant barriers and, finally, the complexities (Newton and arithmetic) of the resulting methods.

In what follows, if opposite is not explicitly stated, we always assume that the constraints involved into the problem satisfy the Slater condition.

10.1 Linear Programming

Consider an LP problem in the canonical form:

$$\textit{minimize} \;\; c^T x \;\; \textit{s.t.} \;\; x \in G \equiv \{x \mid Ax \leq b\}, \tag{10.1}$$

A being $m \times n$ matrix of the full column rank[1]

Path-following approach can be applied immediately:

Standard reformulation: the problem from the very beginning is in the standard form;

Barrier: as we know, the function

$$F(x) = -\sum_{j=1}^{m} \ln(b_j - a_j^T x)$$

is m-self-concordant barrier for G;

Structural assumption from Lecture 8 is satisfied: indeed,

$$F(x) = \Phi(b - Ax), \;\; \Phi(u) = -\sum_{j=1}^{m} \ln u_j : \operatorname{int} \mathbf{R}_+^m \to \mathbf{R} \tag{10.2}$$

[1] the assumption that the rank of A is n is quite natural, since otherwise the homogeneous system $Ax = 0$ has a nontrivial solution, so that the feasible domain of the problem, if nonempty, contains lines. Consequently, the problem, if feasible, is unstable: small perturbation of the objective makes it below unbounded, so that the problems of this type might be only of theoretical interest

and Φ is m-logarithmically homogeneous self-concordant barrier for the m-dimensional nonnegative orthant; the Legendre transformation of Φ, as it is immediately seen, is

$$\Phi^*(s) = -\sum_{j=1}^{m} \ln(-s_j) - m : \operatorname{int} \mathbf{R}^m_- \to \mathbf{R}. \tag{10.3}$$

Thus, to solve an LP problem, we can use both the basic and the long-step versions of the path-following method.

Complexity: as we remember, the Newton complexity of finding an ε-solution by a path-following method associated with a ϑ-self-concordant barrier is $\mathcal{M} = O(1)\sqrt{\varepsilon}\ln(\mathcal{V}\varepsilon^{-1})$, $O(1)$ being certain absolute constant[2] and $\mathcal{V}$ is a data-dependent scale factor. Consequently, the arithmetic cost of an ε-solution is $\mathcal{MN}$, where $\mathcal{N}$ is the arithmetic cost of a single Newton step. We see that the complexity of the method is completely characterized by the quantities ϑ and $\mathcal{N}$. Note that the product

$$\mathcal{C} = \sqrt{\vartheta}\mathcal{N}$$

is the factor at the term $\ln(\mathcal{V}\varepsilon^{-1})$ in the expression for the arithmetic cost of an ε-solution; thus, $\mathcal{C}$ can be thought of as the arithmetic cost of an accuracy digit in the solution (since $\ln(\mathcal{V}\varepsilon^{-1})$ can be naturally interpreted as the amount of accuracy digits in an ε-solution).

Now, in the situation in question $\vartheta = m$ is the larger size of the LP problem, and it remains to understand what is the cost $\mathcal{N}$ of a Newton step. At a step we are given an x and should form and solve with respect to y the linear system of the type

$$F''(x)y = -tc - F'(x);$$

the gradient and the Hessian of the barrier in our case, as it is immediately seen, are given by

$$F'(x) = \sum_{i=j}^{m} d_j a_j, \;\; F''(x) = A^T D^2 A,$$

where

$$d_j = [b_j - a_j^T x]^{-1}$$

are the inverse residuals in the constraints at the point x and

$$D = \mathrm{Diag}(d_1, ..., d_m).$$

It is immediately seen that the arithmetic cost of *assembling* the Newton system (i.e., the cost of computing F' and F'') is $O(mn^2)$; to solve the system after it is assembled, it takes $O(n^3)$ operations more[3]. Since $m \geq n$ (recall that $\operatorname{Rank} A = n$), the arithmetic complexity of a step is dominated by the cost $O(mn^2)$ of assembling the Newton system. Thus, we come to

$$\vartheta = m; \;\; \mathcal{N} = O(mn^2); \;\; \mathcal{C} = O(m^{3/2}n^2). \tag{10.4}$$

Potential reduction approach also is immediate:

Conic reformulation of the problem is given by

$$\textit{minimize} \;\; f^T y \;\; s.t. \;\; y \in \{L + b\} \cap K, \tag{10.5}$$

where

$$K = \mathbf{R}^m_+, \;\; L = A(\mathbf{R}^n)$$

[2] provided that the parameters of the method - i.e., the path tolerance κ and the penalty rate γ in the case of the basic method and the path tolerance κ and the treshold $\overline{\kappa}$ in the case of the long step one - are once for ever fixed

[3] if the traditional Linear Algebra is used (Gauss elimination, Choleski decomposition, etc.); there exists, at least in theory, "fast" Linear Algebra which allows to invert an $N \times N$ matrix in $O(N^\gamma)$ operations for some $\gamma < 3$ rather than in $O(N^3)$ operations

and f is m-dimensional vector which "expresses the objective $c^T x$ in terms of $y = Ax$", i.e., is such that

$$f^T Ax \equiv c^T x;$$

one can set, e.g.,

$$f = A[A^T A]^{-1} c$$

(non-singularity of $A^T A$ is ensured by the assumption that $\mathrm{Rank}\, A = n$).

The cone $K = \mathbf{R}^m_+$ clearly is self-dual, so that the conic dual to (10.5) is

$$\textit{minimize}\ \ b^T s\ \ \textit{s.t.}\ \ s \in \{L^\perp + f\} \cap \mathbf{R}^m_+; \tag{10.6}$$

as it is immediately seen, the dual feasible plane $L^\perp + f$ is given by

$$L^\perp + f = \{s \mid A^T s = c\}$$

(see Exercise 5.4.11).

Logarithmically homogeneous barrier for $K = \mathbf{R}^m_+$ is, of course, the barrier Φ given by (10.2); the parameter of the barrier is m, and its Legendre transformation Φ^* is given by (10.3). Thus, we can apply both the method of Karmarkar and the primal-dual method.

Complexity of the primal-dual method for LP is, at it is easily seen, completely similar to that one of the path-following method; it is given by

$$\vartheta = m;\ \ \mathcal{N} = O(mn^2);\ \ \mathcal{C} = O(m^{3/2} n^2).$$

The method of Karmarkar has the same arithmetic cost $\mathcal{N}$ of a step, but worse Newton complexity (proportional to $\vartheta = m$ rather than to $\sqrt{\vartheta}$), so that for this method one has

$$\mathcal{N} = O(mn^2),\ \ \mathcal{C} = O(m^2 n^2).$$

Comments. 1) *Karmarkar acceleration.* The aforementioned expressions for $\mathcal{C}$ correspond to the default assumption that we solve the sequential Newton systems "from scratch" - independently of each other. This is not the only possible policy: the matrices of the systems arising at neighbouring steps are close to each other, and therefore there is a possibility to implement the Linear Algebra in a way which results in certain progress in the average (over steps) arithmetic cost of finding Newton directions. I am not going to describe the details of the corresponding *Karmarkar acceleration*; let me say that this acceleration results in the (average over iterations) value of $\mathcal{N}$ equal to $O(m^{1/2} n^2)$ instead of the initial value $O(mn^2)$ [4]. As a result, for the accelerated path-following and primal-dual methods we have $\mathcal{C} = O(mn^2)$, and for the accelerated method of Karmarkar $\mathcal{C} = O(m^{3/2} n^2)$. Thus, the arithmetic complexity of an accuracy digit in LP turns out to be the same as when solving systems of linear *equations* by the traditional Linear Algebra technique.

2) *Practical performance.* One should be awared that the outlined complexity estimates for interior point LP solvers give very poor impression of their actual performance. There are two reasons for it:

- first, when evaluating the arithmetic cost of a Newton step, we have implicitly assumed that the matrix of the problem is dense and "unstructured"; this case *never* occurs in actual large-scale computations, so that the arithmetic cost of a Newton step normally has nothing in common with the above $O(mn^2)$ and heavily depends on the specific structure of the problem;

- second, and more important fact is that the "long-step" versions of the methods (like the potential reduction ones and the long step path following method) in practice possess much better Newton complexity than it is said by the theoretical worst-case efficiency estimate. According to the latter estimate, the Newton complexity should be proportional at least to the square root of the larger size m of the problem; in practice the dependence turns out to be much better, something like $O(\ln m)$; in the real-world range of values of sizes it means that the *Newton complexity of long step interior point methods for LP is basically independent of the size of the problem* and is something like 20-50 iterations. This is the source of "competitive potential" of the interior point methods versus the Simplex method.

[4] provided that the problem is not "too thin", namely, that $n \geq O(\sqrt{m})$

3) *Unfeasible start.* To the moment all schemes of interior point methods known to us have common practical drawback: they are indeed "interior point schemes", and to start a method, one should know in advance a strictly feasible solution to the problem. In real-world computations this might be a rather restrictive requirement. There are several ways to avoid this drawback, e.g., the following "big M" approach: to solve (10.1), let us extend x by an artificial design variable t and pass from the original problem to the new one

$$\textit{minimize}\ \ c^T x + Mt \ \ s.t.\ \ Ax + t(b-e) \le b,\ -t \le 0;$$

here $e = (1, ..., 1)^T$. The new problem admits an evident strictly feasible solution $x = 0, t = 1$; on the other hand when M is large, then the x-component of optimal solution to the problem is "almost feasible almost optimal" for the initial problem (theoretically, for large enough M the x-components of *all* optimal solutions to the modified problem are optimal solutions to the initial one). Thus, we can apply our methods to the modified problem (where we have no difficulties with initial strictly feasible solution) and thus get a good approximate solution to the problem of interest. Note that the same trick can be used in our forthcoming situations.

10.2 Quadratically Constrained Quadratic Programming

The problem here is to minimize a convex quadratic function $g(x)$ over a set given by finitely many convex quadratic constraints $g_j(x) \le 0$. By adding extra variable t and extra constraint $g(x) - t \le 0$ (note that it also is a convex quadratic constraint), we can pass from the problem to an equivalent one with a linear objective and convex quadratic constraints. It is convenient to assume that this reduction is done from the very beginning, so that the initial problem of interest is

$$\textit{minimize}\ \ c^t x \ \ s.t.\ \ x \in G = \{x \mid f_j(x) = x^T A_j x + b_j^T x + c_j \le 0,\ j = 1, ..., m\}, \tag{10.7}$$

A_j being $n \times n$ positive semidefinite symmetric matrices.

Due to positive semidefiniteness and symmetry of A_j, we always can decompose these matrices as $A_j = B_j^T B_j$, B_j being $k(B_j) \times n$ rectangular matrices, $k(B_j) \le n$; in applications, normally, we should not compute these matrices, since B_j, together with A_j, form the "matrix" part of the input data.

Path-following approach is immediate:

Standard reformulation: the problem from the very beginning is in the standard form.

Barrier: as we know from Lecture 9, the function

$$-\ln(t - f(x))$$

is 1-self-concordant barrier for the epigraph $\{t \ge f(x)\}$ of a convex quadratic form $f(x) = x^T B^T B x + b^T x + c$. Since the Lebesque set $G_f = \{x \mid f(x) \le 0\}$ of f is the inverse image of this epigraph under the linear mapping $x \mapsto (0, x)$, we conclude from the Substitution rule (L) (Lecture 9) that the function $-\ln(-f(x))$ is 1-self-concordant barrier for G_f, provided that $f(x) < 0$ at some x. Applying the Decomposition rule (Lecture 9), we see that the function

$$F(x) = -\sum_{j=1}^{m} \ln(-f_j(x)) \tag{10.8}$$

is m-self-concordant barrier for the feasible domain G of problem (10.7).

Structural assumption. Let us demonstrate that the above barrier satisfies the Structural assumption from Lecture 8. Indeed, let us set

$$r(B_j) = k(B_j) + 1$$

and consider the second order cones

$$K^2_{r(B_j)} = \{(\tau, \sigma, \xi) \in \mathbf{R} \times \mathbf{R} \times \mathbf{R}^{k(B_j)} \mid \tau \ge \sqrt{\sigma^2 + \xi^T \xi}\}.$$

Representing the quantity $b_j^T x + c_j$ as

$$b_j^T x + c_j = \left[\frac{1 + b_j^T x + c_j}{2}\right]^2 - \left[\frac{1 - b_j^T x - c_j}{2}\right]^2,$$

we come to the following representation of the set $G_j = \{x \mid f_j(x) \le 0\}$:

$$\{x \mid f_j(x) \le 0\} \equiv \{x \mid [B_j x]^T [B_j x] + b_j^T x + c_j \le 0\} =$$

$$= \left\{ x \mid \left[\frac{1 - b_j^T x - c_j}{2}\right]^2 \ge \left[\frac{1 + b_j^T x + c_j}{2}\right]^2 + [B_j x]^T [B_j x] \right\} =$$

[note that for x in the latter set $b_j^T x + c_j \le 0$]

$$= \left\{ x \mid \frac{1 - b_j^T x - c_j}{2} \ge \sqrt{\left[\frac{1 + b_j^T x + c_j}{2}\right]^2 + [B_j x]^T [B_j x]} \right\}$$

Thus, we see that G_j is exactly the inverse image of the second order cone $K^2_{r(B_j)}$ under the affine mapping

$$x \mapsto \pi_j x + p_j = \begin{pmatrix} \tau = \frac{1}{2}[1 - b_j^T x - c_j] \\ \sigma = \frac{1}{2}[1 + b_j^T x + c_j] \\ \xi = B_j x \end{pmatrix}.$$

It is immediately seen that the above barrier $-\ln(-f_j(x))$ for G_j is the superposition of the standard barrier

$$\Psi_j(\tau, \sigma, \xi) = -\ln(\tau^2 - \sigma^2 - \xi^T \xi)$$

for the cone $K^2_{r(B_j)}$ and the affine mapping $x \mapsto \pi_j x + p_j$. Consequently, the barrier $F(x)$ for the feasible domain G of our quadraticaly constrained problem can be represented as

$$F(x) = \Phi(\pi x + p), \quad \pi x + p = \begin{pmatrix} \tau_1 = \frac{1}{2}[1 - b_1^T x - c_1] \\ \sigma_1 = \frac{1}{2}[1 + b_1^T x + c_1] \\ \xi_1 = B_1 x \\ \dots \\ \tau_m = \frac{1}{2}[1 - b_m^T x - c_m] \\ \sigma_m = \frac{1}{2}[1 + b_m^T x + c_m] \\ \xi_m = B_m x \end{pmatrix}, \tag{10.9}$$

where

$$\Phi(\tau_1, \sigma_1, \xi_1, \dots, \tau_m, \sigma_m, \xi_m) = -\sum_{j=1}^m \ln(\tau_j^2 - \sigma_j^2 - \xi_j^T \xi_j) \tag{10.10}$$

is the direct sum of the standard self-concordant barriers for the second order cones $K^2_{r(B_j)}$; as we know from Proposition 5.3.2.(iii), Φ is $(2m)$-logarithmically homogeneous self-concordant barrier for the direct product K of the cones $K^2_{r(B_j)}$. The barrier Φ possesses the immediately computable Legendre transformation

$$\Phi^*(s) = \Phi(-s) + 2m \ln 2 - 2m \tag{10.11}$$

with the domain $-\operatorname{int} K$.

Complexity. The complexity characteristics of the path-following method associated with barrier (10.8), as it is easily seen, are given by

$$\vartheta = m; \;\; \mathcal{N} = O([m + n]n^2); \;\; \mathcal{C} = O(m^{1/2}[m + n]n^2) \tag{10.12}$$

(as in the LP case, expressions for $\mathcal{N}$ and $\mathcal{C}$ correspond to the case of dense "unstructured" matrices B_j; in the case of sparse matrices with reasonable nonzero patterns these characteristics become better).

Potential reduction approach also is immediate:

Conic reformulation of the problem is a byproduct of the above considerations; it is

$$\textit{minimize}\ \ f^T y\ \ \textit{s.t.}\ \ y \in \{L + p\} \cap K, \tag{10.13}$$

where $K = \prod_{j=1}^{m} K^2_{r(B_j)}$ is the above product of second order cones, $L + b$ is the image of the above affine mapping $x \mapsto \pi x + p$ and f is the vector which "expresses the objective $c^T x$ in terms of $y = \pi x$", i.e., such that

$$f^T \pi x = c^T x;$$

it is immediately seen that such a vector f does exist, provided that the problem in question is solvable.

The direct product K of the second order cones is self-dual (Exercise 5.4.7), so that the conic dual to (10.13) is the problem

$$\textit{minimize}\ \ p^T s\ \ \textit{s.t.}\ \ s \in \{L^{\perp} + f\} \cap K \tag{10.14}$$

with the dual feasible plane $L^{\perp} + f$ given by

$$L^{\perp} + f = \{s \mid \pi^T s = c\}$$

(see Exercise 5.4.11).

Logarithmically homogeneous self-concordant barrier with parameter $2m$ for the cone K is, as it was already mentioned, given by (10.10); the Legendre transformation of Φ is given by (10.11). Thus, we have in our disposal computable primal and dual barriers for (10.13) - (10.14) and can therefore solve the problems by the method of Karmarkar or by the primal-dual method associated with these barriers.

Complexity: it is immediately seen that the complexity characteristics of the primal-dual method are given by (10.12); the characteristics $\mathcal{N}$ and $\mathcal{C}$ of the method of Karmarkar are $O(\sqrt{m})$ times worse than the corresponding characteristics of the primal-dual method.

10.3 Approximation in L_p norm

The problem of interest is

$$\textit{minimize}\ \ \sum_{j=1}^{m} |v_j - u_j^T x|^p, \tag{10.15}$$

where $1 < p < \infty$, $u_j \in \mathbf{R}^n$ and $v_j \in \mathbf{R}$.

Path-following approach seems to be the only one which can be easily carried out (in the potential reduction scheme there are difficulties with explicit formulae for the Legendre transformation of the primal barrier).

Standard reformulation of the problem is obtained by adding m extra variables t_j and rewriting the problem in the equivalent form

$$\textit{minimize}\ \ \sum_{j=1}^{m} t_j\ \ \textit{s.t.}\ \ (t, x) \in G = \{(t, x) \in \mathbf{R}^{m+n} \mid |v_j - u_j^T x|^p \le t_j,\ j = 1, ..., m\}. \tag{10.16}$$

Barrier: self-concordant barrier for the feasible set G of problem (10.16) was constructed in Lecture 9 (Example 9.2.1, Substitution rule (L) and Decomposition rule):

$$F(t, x) = \sum_{j=1}^{m} F_j(t_j, x),\ \ F_j(t, x) = -\ln(t_j^{2/p} - (v_j - u_j^T x)^2) - 2\ln t_j,\ \ \vartheta = 4m.$$

Complexity of the path-following method associated with the indicated barrier is characterized by

$$\vartheta = 4m;\ \ \mathcal{N} = O([m + n]n^2);\ \ \mathcal{C} = O(m^{1/2}[m + n]n^2).$$

The above expression for the arithmetic complexity $\mathcal{N}$ needs certain clarification: our barrier depends on $m+n$ variables, and its Hessian is therefore an $(m+n)\times(m+n)$ matrix; how it could be that we can assemble and invert this matrix at the cost of $O(n^2[m+n])$ operations, not at the "normal" cost $O([m+n]^3)$?

The estimate for $\mathcal{N}$ is given by the following reasoning. Since the barrier is separable, its Hessian H is the sum of Hessians of the "partial barriers" $F_j(t,x)$; the latter Hessians, as it is easily seen, can be computed at the arithmetic cost $O(n^2)$ and are of very specific form: the $m\times m$ block corresponding to t-variables contains only one nonzero entry (coming from to $\frac{\partial^2}{\partial t_j \partial t_j}$). It follows that H can be computed at the cost $O(mn^2)$ and is $(m+n)\times(m+n)$ matrix of the form

$$H = \begin{pmatrix} T & P^T \\ P & Q \end{pmatrix},$$

where the $m\times m$ block T corresponding to t-variables is *diagonal*, P is $n\times m$ and Q is $n\times n$. It is immediately seen that the gradient of the barrier can be computed at the cost $O(mn)$. Thus, the arithmetic cost of *assembling* the Newton system is $O(mn^2)$, and the system itself is of the type

$$\begin{array}{ll} Tu + P^T v & = p \\ Pu + Qv & = q \end{array}$$

with m-dimensional vector of unknowns u, n-dimensional vector of unknowns v and diagonal T. To solve the system, we can express u via v:

$$u = T^{-1}[p - P^T v]$$

and substitute this expression in the remaining equations to get a $n\times n$ system for u:

$$[Q - PT^{-1}P^T]u = q - PT^{-1}p.$$

To assemble this latter system it clearly costs $O(mn^2)$ operations, to solve it - $O(n^3)$ operations, and the subsequent computation of u takes $O(mn)$ operations, so that the total arithmetic cost of assembling and solving the entire Newton system indeed is $O([m+n]n^2)$.

What should be noticed here is not the particular expression for $\mathcal{N}$, but the general rule which is illustrated by this expression: the Newton systems which arise in the interior point machinery normally possess nontrivial structure, and a reasonable solver should use this structure in order to reduce the arithmetic cost of Newton steps.

10.4 Geometrical Programming

The problem of interest is

$$\textit{minimize}\;\; f_0(x) = \sum_{i\in\mathcal{I}_0} c_{i0}\exp\{a_i^T x\} \;\; \textit{s.t.}\;\; f_j(x) = \sum_{i\in\mathcal{I}_j} c_{ij}\exp\{a_i^T x\} \le d_j,\; j=1,...,m. \tag{10.17}$$

Here $x\in\mathbf{R}^n$, $\mathcal{I}_j$ are subsets of the index set $\mathcal{I}=\{1,...,k\}$ and all coefficients c_{ij} are positive, $j=1,...,m$.

Note that in the standard formulation of a Geometrical Programming program the objective and the constraints are sums, with nonnegative coefficients, of "monomials" $\xi_1^{\alpha_1}...\xi_n^{\alpha_n}$, ξ_i being the design variables (which are restricted to be positive); the exponential form (10.17) is obtained from the "monomial" one by passing from ξ_i to the new variables $x_i = \ln \xi_i$.

Here it again is difficult to compute the Legendre transformation of the barrier associated with the conic reformulation of the problem, so that we restrict ourselves with the *Path-following approach* only.

Standard reformulation: to get it, we introduce k additional variables t_i, one per each of the exponents $\exp\{a_i^T x\}$ involved into the problem, and rewrite (10.17) in the following equivalent form:

$$\textit{minimize}\;\; \sum_{i\in\mathcal{I}_0} c_{i0}t_i \;\; \textit{s.t.}\;\; (t,x)\in G, \tag{10.18}$$

with

$$G = \{(t,x) \in \mathbf{R}^k \times \mathbf{R}^n \mid \sum_{i \in \mathcal{I}_j} c_{ij} t_j \le d_j,\ j = 1,...,m; \exp\{a_i^T x\} \le t_i,\ i = 1,...,k\}.$$

Barrier. The feasible domain G of the resulting standard problem is given by a number of linear constraints and a number of exponential inequalities $\exp\{a_i^T x\} \le t_i$. We know how to penalize the feasible set of a linear constraint, and there is no difficulty in penalizing the feasible set of an exponential inequality, since this set is inverse image of the epigraph

$$\{(\tau, \xi) \mid \tau \ge \exp\{\xi\}\}$$

under an affine mapping.

Now, a 2-self-concordant barrier for the epigraph of the exponent, namely, the function

$$\Psi(\tau, \xi) = -\ln(\ln \tau - \xi) - \ln \tau$$

was found in Lecture 9 (Example 9.2.3). Consequently, the barrier for the feasible set G is

$$F(t,x) = \sum_{i=1}^{k} \Psi(t_i, a_i^T x) - \sum_{j=1}^{m} \ln\left(d_j - \sum_{i \in \mathcal{I}_j} c_{ij} t_j\right) = \Phi\left(\pi \begin{pmatrix} t \\ x \end{pmatrix} + p\right),$$

where

$$\Phi(\tau_1, \xi_1, ..., \tau_k, \xi_k; \tau_{k+1}, \tau_{k+2}, ..., \tau_{k+m}) = \sum_{i=1}^{k} \Psi(\tau_i, \xi_i) - \sum_{j=1}^{m} \ln \tau_{k+j}$$

is self-concordant barrier with parameter $2k+m$ and the affine substitution $\pi \begin{pmatrix} t \\ x \end{pmatrix} + p$ is given by

$$\pi \begin{pmatrix} t \\ x \end{pmatrix} + p = \begin{pmatrix} \tau_1 = t_1 \\ \xi_1 = a_1^T x \\ ... \\ \tau_k = t_k \\ \xi_k = a_k^T x \\ \tau_{k+1} = d_1 - \sum_{i \in \mathcal{I}_1} c_{i1} t_i \\ \tau_{k+2} = d_2 - \sum_{i \in \mathcal{I}_2} c_{i2} t_i \\ ... \\ \tau_{k+m} = d_m - \sum_{i \in \mathcal{I}_m} c_{im} t_i \end{pmatrix}.$$

Structural assumption. To demonstrate that the indicated barrier satisfies the Structural assumption, it suffices to point out the Legendre transformation of Φ; since this latter barrier is the direct sum of k copies of the barrier

$$\Psi(\tau, \xi) = -\ln(\ln \tau - \xi) - \ln \tau$$

and m copies of the barrier

$$\psi(\tau) = -\ln \tau,$$

the Legendre transformation of Φ is the direct sum of the indicated number of copies of the Legendre transformations of Ψ and ψ. The latter transformations can be computed explicitly:

$$\Psi^*(\sigma, \eta) = (\eta + 1) \ln\left(\frac{\eta + 1}{-\sigma}\right) - \eta - \ln \eta - 2,\ \operatorname{Dom} \Psi^* = \{\sigma < 0, \eta > 0\},$$

$$\psi^*(\sigma) = -\ln(-\sigma) - 1,\ \operatorname{Dom} \psi^* = \{\sigma < 0\}.$$

Thus, we can solve Geometrical programming problems by both the basic and the long-step path-following methods.

Complexity of the path-following method associated with the aforementioned barrier is given by

$$\vartheta = 2k + m;\ \mathcal{N} = O(mk^2 + k^3 + n^3);\ \mathcal{C} = O((k+m)^{1/2}[mk^2 + k^3 + n^3]).$$

10.5 Exercises on applications of interior point methods

The below problems deal with a topic from Computational Geometry - with computing extremal ellipsoids related to convex sets.

There are two basic problems on extremal ellipsoids:

(Inner): *given a solid $Q \subset \mathbf{R}^n$ (a closed and bounded convex domain with a nonempty interior), find the ellipsoid of the maximum volume contained in Q.*

(Outer): *given a solid $Q \subset \mathbf{R}^n$, find the ellipsoid of the minimum volume containing Q.*

Let us first explain where the problems come from.

I know exactly one source of problem (Inner) - the *Inscribed Ellipsoid method* InsEll for general convex optimization. This is an algorithm for solving problems of the type

$$\textit{minimize} \;\; f(x) \;\; \textit{s.t.} \;\; x \in Q,$$

where Q is a polytope in $\mathbf{R}^n$ and f is convex function. The InsEll, which can be regarded as a multidimensional extension of the usual bisection, generates a decreasing sequence of polytopes Q_i which cover the optimal set of the problem; these localizers are defined as

$$Q_0 = Q; \;\; Q_{i+1} = \{x \in Q_i \mid (x - x_i)^T f'(x_i) \le 0\},$$

where x_i is the center of the maximum volume ellipsoid inscribed into Q_i.

It can be proved that in this method the inaccuracy $f(x^i) - \min_Q f$ of the best (with the smallest value of f) among the search points $x_1, ..., x_i$ admits the upper bound

$$f(x^i) - \min_Q f \le \exp\{-\kappa \frac{i}{n}\}[\max_Q f - \min_Q f],$$

$\kappa > 0$ being an absolute constant; it is known also that the indicated rate of convergence is the best, in certain rigorous sense, rate a convex minimization method can achieve, so that InsEll is optimal. And to run the method, you should solve at each step an auxiliary problem of the type (Inner) related to a *polytope* Q given by list of linear inequalities defining the polytope.

As for problem (Outer), the applications known to me come from Control. Consider a discrete time linear controlled plant given by

$$x(t + 1) = Ax(t) + Bu(t), \;\; t = 0, 1, ...,$$

where $x(t) \in \mathbf{R}^n$ and $u(t) \in \mathbf{R}^k$ are the state of the plant and the control at moment t and A, B are given $n \times n$ and $n \times k$ matrices, A being nonsingular. Assume that $u(\cdot)$ can take values in a polytope $U \subset \mathbf{R}^k$ given as a convex hull of finitely many points $u_1, ..., u_m$:

$$U = \mathrm{Conv}\{u_1, ..., u_m\}.$$

Let the initial state of the plant be known, say, be zero. The question is: what is the set X_T of possible states of the plant at a given moment T?

This is a difficult question which, in the multi-dimensional case, normally cannot be answered in a "closed analytic form". One of the ways to get certain numerical information here is to compute *outer ellipsoidal approximations* of the sets X_t, $t = 0, ..., T$ - ellipsoids E_t which cover the sets X_t. The advantage of this approach is that these approximations are of once for ever fixed "tractable" geometry, in contrast to the sets X_t which may become more and more complicated as t grows. There is an evident possibility to form E_t's in a recurrent way: indeed, if we already know that X_t belongs to a known ellipsoid E_t, then the set X_{t+1} for sure belongs to the set

$$\widehat{E}_t = AE_t + BU.$$

Since U is the convex hull of $u_1, ..., u_m$, the set $\widehat{E}_t$ is nothing but the convex hull Q_{t+1} of the union of E_t^i, $i = 1, ..., m$. Thus, a convex set contains $\widehat{E}_t$ if and only if it contains Q_{t+1}.

Now, it is, of course, reasonable to look for "tight" approximations, i.e., to choose E_{t+1} as close as possible to the set Q_{t+1} (unfortunately, Q_{t+1} usually is not an ellipsoid, so that in any case E_{t+1} will be redundant). A convenient integral measure of the quality of outer approximation is the volume of the approximating set - the less it is, the better is the approximation. Thus, to approximate the sets X_t, we should solve a sequence of problems (Outer) with Q given as the convex hull of a union of ellipsoids.

10.5.1 (Inner) and (Outer) as convex programs

Problems (Inner) and (Outer) can be reformulated as convex programs. To this end recall that there are two basic ways to describe an ellipsoid

- an ellipsoid $W \subset \mathbf{R}^n$ is the image of the unit Euclidean ball under a one-to-one affine mapping of $\mathbf{R}^n$ onto itself:

(I) $W = I(x, X) \equiv \{y = x + Xu \mid u^T u \leq 1\}$;

here $x \in \mathbf{R}^n$ is the center of the ellipsoid and X is a nonsingular $n \times n$ matrix. This matrix is defined uniquely *up to multiplication from the right by an orthogonal matrix*; under appropriate choice of this orthogonal "scale factor" we may make X to be symmetric positive definite, and from now on our convention is that *the matrix X involved into (I) is symmetric positive definite*. Thus, (I) allows to parameterize n-dimensional ellipsoids by the pairs (x, X), with $x \in \mathbf{R}^n$ and X being $n \times n$ positive definite symmetric matrix.

It is worthy to recall that the volume of ellipsoid (I) is $\kappa_n \,\mathrm{Det}\, X$, κ_n being the volume of the n-dimensional Euclidean ball.

- an ellipsoid W is the set given by strictly convex quadratic inequality:

(II) $W = E(r, x, X) \equiv \{u \mid u^T X u + 2x^T u + r \leq 0\}$;

here X is a positive definite symmetric $n \times n$ matrix, $x \in \mathbf{R}^n$ and $r \in \mathbf{R}$. The above relation can be equivalently rewritten as

$$W = \{u \mid (u + X^{-1}x)^T X (u + X^{-1}x) + r - x^T X^{-1} x \leq 0;$$

thus, it indeed defines an ellipsoid if and only if

$$\delta(r, x, X) \equiv x^T X^{-1} x - r > 0.$$

The representation of W via r, x, X is not unique (proportional triples define the same ellipsoid). Therefore we always can enforce the quantity δ to be ≤ 1, and in what follows this is our default convention on the parameterization in question.

It is clearly seen that the volume of the ellipsoid $E(r, x, X)$ is nothing but

$$\kappa_n \delta^{n/2}(r, x, X) \mathrm{Det}^{-1/2} X.$$

Now let us look at problem (Inner). From the above discussion we see that it can be written down as

(Inner') *minimize* $F(X) = -\ln \mathrm{Det}\, X$ *s.t.* $(x, X) \in G_{\mathrm{I}}$,

with

$$G_{\mathrm{I}} = \{(x, X) \mid X \in \mathbf{S}^n_+, I(x, X) \subset Q\};$$

here $\mathbf{S}^n_+$ is the cone of positive semidefinite matrices in the space $\mathbf{S}^n$ of symmetric $n \times n$ matrices.

To get (Inner'), we have passed from the problem of maximizing

$$\mathrm{Vol}_n(I(x, X)) = \kappa_n \,\mathrm{Det}\, X$$

to the equivalent problem of minimizing $-\ln \mathrm{Det}\, X$.

Exercise 10.5.1 *Prove that (Inner') is a convex program: its feasible domain G_{I} is closed and bounded convex set with a nonempty interior in the space $\mathbf{R}^n \times \mathbf{S}^n$, and the objective is a continuous convex function (taking values in $\mathbf{R} \cup \{+\infty\}$) on G_{I} and finite on the interior of the domain G_{I}.*

Similarly, (Outer) also can be posed as a convex program

(Outer') *minimize* $-\ln \mathrm{Det}\, X$ *s.t.* $(r, x, X) \in G_{\mathrm{O}} = \mathrm{cl}\, G'$,

$$G' = \{(r, x, X) \in \mathbf{R} \times \mathbf{R}^n \times \mathrm{int}\, \mathbf{S}^n_+ \mid \delta(r, x, X) \leq 1,\ E(r, x, X) \supset Q\}.$$

Exercise 10.5.2 $^+$ *Prove that (Outer') is a convex programming program: $G_{\rm O}$ is closed convex domain, and F is continuous convex function on $G_{\rm O}$ taking values in $\mathbf{R}\cup\{+\infty\}$ and finite on* int $G_{\rm O}$. *Prove that the problem is equivalent to (Outer).*

Thus, both (Inner) and (Outer) can be reformulated as convex programs. *This does not, anyhow, mean that the problems are computationally tractable.* Indeed, the minimal "well posedness" requirement on a convex problem which allows to speak about it numerical solution is as follows:

(!) *given a candidate solution to the problem, you should be able to check whether the solution is feasible, and if it is the case, you should be able to compute the value of the objective at this solution*[5].

Whether (!) is satisfied or not for problems (Inner) and (Outer), it depends on what is the set Q and how it is represented; and, as we shall see in a while, "well posed" cases for one of our problems could be "ill posed" for another. Note that "well posedness" for (Inner) means a possibility, given an ellipsoid W to check whether W is contained in Q; for (Outer) you should be able to check whether W contains Q.

Consider a couple of examples.

- Q is a polytope given "by facets", more exactly, by a list of linear inequalities (not all of them should represent facets, some may be redundant).

  This leads to well-posed (Inner) (indeed, to check whether W is contained in Q, i.e., in the intersection of a given finite family of half-spaces, is the same as to check whether W is contained in each of the half-spaces, and this is immediate). In contrast to this, in the case in question (Outer) is ill-posed: to check whether, say, a Euclidean ball W contains a polytope given by a list of linear inequalities is, basically, the same as to maximize a convex quadratic form (namely, $|x|_2^2$) under linear inequality constraints, and this is an NP-hard problem.

- Q is a polytope given "by vertices", i.e., represented as a convex hull of a given finite set S.

  Here (Outer) is well-posed (indeed, W contains Q if and only if it contains S, which can be immediately verified), and (Inner) is ill-posed (it is NP-hard).

As we shall see in a while, in the case of a polytope Q our problems can be efficiently solved by interior point machinery, provided that they are well-posed.

10.5.2 Problem (Inner), polyhedral case

In this section we assume that

$$Q = \{x \mid a_j^T x \le b_j,\ j = 1, ..., m\}$$

is a polytope in $\mathbf{R}^n$ given by m linear inequalities.

Exercise 10.5.3 *Prove that in the case in question problem (Inner) can be equivalently formulated as follows:*

(Inner_Lin) *minimize* t *s.t.* $(t, x, X) \in G$,
with

$$G = \{(t, x, X) \mid |Xa_j|_2 \le b_j - a_j^T x,\ j = 1, ..., m;\ \ X \in \mathbf{S}_+^n; -\ln \mathrm{Det}\ X \le t\}.$$

To solve (Inner_Lin) by interior point machinery, we need self-concordant barrier for the feasible set of the problem. This set is given by a number of constraints, and in our "barrier toolbox" we have self-concordant barriers for the feasible sets of all of these constraints, except the latter of them. This shortcoming, anyhow, can be immediately overcome.

Exercise 10.5.4 * *Prove that the function*

$$\Phi(t, X) = -\ln(t + \ln\det X) - \ln \mathrm{Det}\ X$$

is $(n+1)$*-self-concordant barrier for the epigraph*

$$\mathrm{cl}\{(t, X) \in \mathbf{R} \times \mathrm{int}\ \mathbf{S}_+^n \mid t + \ln \mathrm{Det}\ X \ge 0\}$$

[5] to apply interior point methods, you need, of course, much stronger assumptions: you should be able to point out a "computable" self-concordant barrier for the feasible set

of the function $-\ln \mathrm{Det}\; X$*. Derive from this observation that the function*

$$F(t,x,X) = -\sum_{j=1}^{m} \ln([b_j - a_j^T x]^2 - a_j^T X^T X a_j) - \ln(t + \ln \mathrm{Det}\; X) - \ln \mathrm{Det}\; X$$

is $(2m+n+1)$*-self-concordant barrier for the feasible domain* G *of problem* (Inner_Lin)*. What are the complexity characteristic of the path-following method associated with this barrier?*

10.5.3 Problem (Outer), polyhedral case

Now consider problem (Outer) with the set Q given by

$$Q = \{\sum_{j=1}^{m} \lambda_j a_j \mid \lambda \geq 0 \sum_j \lambda_j = 1\}.$$

Exercise 10.5.5 *Prove that in the case in question problem (Outer') becomes the problem*

(Outer_Lin) *minimize* t *s.t.* $(t,r,x,X) \in G$*,*

with

$$G = \{(t,r,x,X) \mid$$
$$a_j^T X a_j + 2x^T a_j + r \leq 0,\; j = 1,...,m;\; X \in \mathbf{S}^n_+;\; -\ln \mathrm{Det}\; X \leq t; \delta(r,x,X) \leq 1\}.$$

Prove$^+$ *that the function*

$$F(t,r,x,X) = -\sum_{j=1}^{m} \ln(-a_j^T X a_j - 2x^T a_j - r) -$$

$$-\ln(1 + r - x^T X^{-1} x) - \ln(t + \ln \mathrm{Det}\; X) - 2\ln \mathrm{Det}\; X$$

is $(m+2n+2)$*-self-concordant barrier for* G*. What are the complexity characteristics of the path-following method associated with this barrier?*

10.5.4 Problem (Outer), ellipsoidal case

The polyhedral versions of problems (Inner) and (Outer) considered so far are, in a sense, particular cases of "ellipsoidal" versions, where Q is an intersection of a finite family of ellipsoids (problem (Inner)) or convex hull of a finite number of ellipsoids (problem (Outer); recall that our motivation of this latter problem leads to the "ellipsoidal" version of it). Indeed, the polyhedral (Inner) relates to the case when Q is an intersection of a finite family of half-spaces, and a half-space is nothing but a "very large" ellipsoid. Similarly, polyhedral (Outer) relates to the case when Q is a convex hull of finitely many points, and a point is nothing but a "very small" ellipsoid. What we are about to do is to develop polynomial time methods for the ellipsoidal version of (Outer). The basic question of well-posedness here reads as follows:

(?) *Given two ellipsoids, define whether the second of them contains the first one*

This question can be efficiently answered, and the nontrivial observation underlying this answer is, I think, more important than the question itself.

We shall consider (?) in the situation where the first ellipsoid is given as $E(r,x,X)$, and the second one - as $E(s,y,Y)$. Let us start with equivalent reformulation of the question.

The ellipsoid $E(r,x,X)$ is contained in $E(s,y,Y)$ if and only if every solution u to the inequality

$$u^T X u + 2x^T u + r \leq 0$$

satisfies the inequality

$$u^T Y u + 2y^T u + s \leq 0.$$

Substituting $u = v/t$, we can reformulate this as follows:

$E(r,x,X) \subset E(s,y,Y)$ if and only if from the inequality

$$v^T X v + 2tx^T v + rt^2 \leq 0$$

and from $t \neq 0$ it always follows that

$$v^T Y v + 2ty^T v + st^2 \leq 0.$$

In fact we can omit here "$t \neq 0$", since for $t = 0$ the first inequality can be valid only when $v = 0$ (recall that X is positive definite), and the second inequality then also is valid. Thus, we come to the conclusion as follows:

$E(r,x,X) \subset E(s,y,Y)$ if and only if the following implication is valid:

$$w^T S w \leq 0 \Rightarrow w^T R w \leq 0,$$

where

$$S = \begin{pmatrix} X & x \\ x^T & r \end{pmatrix}, \; R = \begin{pmatrix} Y & y \\ y^T & s \end{pmatrix}.$$

We have reduced (?) to the following question

(??) *given two symmetric matrices R and S of the same size, detect whether all directions w where the quadratic form $w^T S w$ is nonpositive are also the directions where the quadratic form $w^T R w$ is nonpositive:*

$$\text{(Impl)} \qquad w^T S w \leq 0 \Rightarrow w^T R w \leq 0.$$

In fact we can say something additional about the quadratic forms S and R we actually are interested in:

(*) *in the case of matrices coming from ellipsoids there is a direction w with negative $w^T S w$, and there is a direction w' with positive $(w')^T R w'$.*

Exercise 10.5.6 $^+$ *Prove (*).*

Now, there is an evident sufficient condition which allos to give a positive answer to (??): if $R \leq \lambda S$ with some nonnegative λ, then, of course, (Impl) is valid. It is a kind of miracle that this sufficient condition is also necessary, provided that $w^T S w < 0$ for some w:

Exercise 10.5.7 * *Prove that if S and R are symmetric matrices of the same size such that the implication* (Impl) *is valid and S is such that $w^T S w < 0$ for some w, then there exists nonnegative λ such that*

$$R \leq \lambda S;$$

if, in addition, $(w')^T R w' > 0$ for some w', then the above λ is positive.

Conclude from the above, that if S and R are symmetric matrices of the same size such that $w_S^T S w_S < 0$ for some w_S and $w_R^T R w_R > 0$ for some w_R, then implication (Impl) is valid if and only if

$$R \leq \lambda S$$

for some positive λ.

It is worthy to explain why the statement given in the latter exercise is so amazing. (Impl) says exactly that the quadratic form $f_1(w) = -w^T R w$ is nonnegative whenever the quadratic form $f_2(w) = w^T S w$ is nonpositive, or, in other words, that the function

$$f(w) = \max\{f_1(w), f_2(w)\}$$

is nonegative everywhere and attains therefore its minimum at $w = 0$. If the functions f_1 and f_2 were convex, we could conclude from this that certain convex combination $\mu f_1(w) + (1-\mu) f_2(w)$ of these functions also attains its minimum at $w = 0$, so that $-\mu R + (1-\mu) S$ is positive semidefinite; the conclusion is exactly what is said by our statement (it says also that $\mu > 0$, so that the matrix inequality can be rewritten as $R \leq \lambda S$ with $\lambda = (1-\mu)\mu^{-1}$; this additional information is readily given by the assumption that $w^T S w < 0$ and causes no surprise). Thus, the conclusion is the same as in the situation of convex f_1 and f_2; but we did not assume the functions to be convex! Needless to say, the "statement" of the type

$$\max\{f_1, f_2\} \geq 0 \text{ everywhere} \Rightarrow \exists \mu \in [0,1]: \ \mu f_1 + (1-\mu) f_2 \geq 0 \text{ everywhere}$$

fails to be true for arbitrary f_1 and f_2, but, as we have seen, it is true for homogeneous quadratic forms. Let me add that the implication

$$\max\{w^T S_1 w, ..., w^T S_k w\} \geq 0 \ \forall w \quad \Rightarrow \quad \textit{certain convex combination of } S_i \textit{ is } \geq 0$$

is valid only for $k = 2$.

Now we are ready to apply interior point machinery to the ellipsoidal version of (Outer).

Consider problem (Outer) with Q given as the convex hull of ellipsoids $E(p_i, a_i, A_i)$, $i = 1, ..., m$. An ellipsoid $E(r, x, X)$ is a convex set; therefore it contains the convex hull Q of our ellipsoids if and only if it contains each of the ellipsoids. As we know from Exercise 10.5.7 and (*), the latter is equivalent to existence of m positive reals $\lambda_1, ..., \lambda_m$ such that

$$R(r, x, X) \equiv \begin{pmatrix} X & x \\ x^T & r \end{pmatrix} \leq \lambda_i S_i,$$

where $S_i = R(p_i, a_i, A_i)$.

Exercise 10.5.8 *Prove that in the case in question problem (Outer) can be equivalently formulated as the following convex program:*

(Outer_Ell) *minimize* t *s.t.* $(t, r, x, X, \lambda) \in G$,

where

$$G = \mathrm{cl}\{(t, r, x, X, \lambda) \mid$$
$$X \in \mathrm{int}\, \mathbf{S}^n_+, t + \ln \mathrm{Det}\ X \geq 0, \delta(r, x, X) \leq 1, R(r, x, X) \leq \lambda_i S_i,\ i = 1, ..., m\}.$$

Prove that the function

$$F(t, r, x, X, \lambda) = -\ln(t + \ln \mathrm{Det}\ X) - 2 \ln \mathrm{Det}\ X - \ln(1 + r - x^T X^{-1} x) -$$
$$- \sum_{i=1}^{m} \ln \mathrm{Det}\,(\lambda_i S_i - R(r, x, X))$$

is $([m+2]n+2)$*-self-concordant barrier for the feasible domain* G *of the problem. What are the complexity characteristics of the path-following method associated with this barrier?*

Chapter 11

Semidefinite Programming

This concluding lecture is devoted to an extremely interesting and important class of convex programs - the so called *Semidefinite Programming*.

11.1 A Semidefinite program

The canonical form of a semidefinite program is as follows:

(SD) *minimize linear objective $c^T x$ of $x \in \mathbf{R}^n$ under Linear Matrix Inequality constraints*

$$A_j(x) \geq 0,\ j = 1, ..., M,$$

where $A_j(x)$ are symmetric matrices affinely depending on x (i.e., each entry of $A_j(\cdot)$ is an affine function of x), and $A \geq 0$ for a symmetric matrix A stands for "A is positive semidefinite".

Note that a system of m Linear Matrix Inequality constraints (LMI's) $A_j(x) \geq 0,\ j = 1, ..., M,$ is equivalent to a *single* LMI

$$A(x) \geq 0,\ A(x) = \mathrm{Diag}\{A_1(x), ..., A_M(x)\} = \begin{pmatrix} A_1(x) & & & \\ & A_2(x) & & \\ ... & ... & ... & ... \\ & & & A_M(x) \end{pmatrix}$$

(blank space corresponds to zero blocks). Further, an affine in x matrix-valued function $A(x)$ can be represented as

$$A(x) = A_0 + \sum_{i=1}^{n} x_i A_i,$$

A_0,...,A_n being fixed matrices of the same size; thus, in a semidefinite program we should minimize a linear form of $x_1, ..., x_n$ provided that a linear combination of given matrices A_i with the coefficients x_i plus the constant term A_0 is positive semidefinite.

The indicated problem seems to be rather artificial. Let me start with indicating several examples of important problems covered by Semidefinite Programming.

11.2 Semidefinite Programming: examples

11.2.1 Linear Programming

Linear Programming problem

$$\textit{minimize}\ \ c^T x\ \ \textit{s.t.}\ \ a_j^T x \leq b_j,\ j = 1, ..., M$$

is a very particular semidefinite program: the corresponding matrix $A(x)$ is $M \times M$ diagonal matrix with the diagonal entries $b_j - a_j^T x$ (indeed, a diagonal matrix is positive semidefinite if and only if its diagonal entries are nonnegative, so that $A(x) \geq 0$ if and only if x is feasible in the initial LP problem).

11.2.2 Quadratically Constrained Quadratic Programming

A convex quadratic constraint

$$f(x) \equiv x^T B^T Bx + b^T x + c \leq 0,$$

B being $k \times n$ matrix, can be expressed in terms of positive semidefiniteness of certain *affine* in x $(k+1) \times (k+1)$ symmetric matrix $A_f(x)$, namely, the matrix

$$A_f(x) = \begin{pmatrix} -c - b^T x & [Bx]^T \\ Bx & I \end{pmatrix}.$$

Indeed, it is immediately seen that a symmetric matrix

$$A = \begin{pmatrix} P & R^T \\ R & Q \end{pmatrix}$$

with positive definite block Q is positive semidefinite if and only if the matrix $P - R^T Q^{-1} R$ is positive semidefinite[1]; thus, $A_f(x)$ is positive semidefinite if and only if $-c - b^T x \geq x^T B^T Bx$, i.e., if and only if $f(x) \leq 0$.

Thus, a convex quadratic constraint can be equivalently represented by an LMI; it follows that a convex quadratic quadratically constrained problem can be resresented as a problem of optimization under LMI constraints, i.e., as a semidefinite program.

The outlined examples are not that convincing: there are direct ways to deal with LP and QCQP, and it hardly makes sense to reduce these problems to evidently more complicated semidefinite programs. In the forthcoming examples LMI constraints come from the nature of the problem in question.

11.2.3 Minimization of Largest Eigenvalue and Lovasz Capacity of a graph

The *Linear Eigenvalue problem* is to find x which minimizes the maximum eigenvalue of symmetric matrix $B(x)$ affinely depending on the design vector x (there are also nonlinear versions of the problem, but I am not speaking about them). This is a traditional area of Convex Optimization; the problem can be immediately reformulated as the semidefinite program

$$\textit{minimize} \ \ \lambda \ \ \textit{s.t.} \ \ A(\lambda, x) = \lambda I - B(x) \geq 0.$$

As an application of the Eigenvalue problem, let us look at computation of the *Lovasz capacity number* of a graph. Consider a graph Γ with the set of vertices V and set of arcs E. One of the fundamental characteristics of the graph is its *inner stability number* $\alpha(\Gamma)$ - the maximum cardinality of an independent subset of vertices (a subset is called independent, if no two vertices in it are linked by an arc). To compute $\alpha(\Gamma)$, this is an NP-hard problem.

There is another interesting characteristic of a graph - the *Shannon capacity number* $\sigma(\Gamma)$ defined as follows. Let us interpret the vertices of Γ as letters of certain alphabet. Assume that we are transmitting words comprised of these letters via an unreliable communication channel; unreliability of the channel is described by the arcs of the graph, namely, letter i on input can become letter j on output if and only if i and j are linked by an arc in the graph. Now, what is the maximum number s_k of k-letter words which you can send through the channel without risk that one of the words will be converted to another? When $k = 1$, the answer is clear - exactly $\alpha(\Gamma)$; you can use, as these words, letters from (any) maximal independent set V^* of vertices. Now, $s_k \geq s_1^k$ - the words comprised of letters which cannot be "mixed" also cannot be mixed. In fact s_k can be greater than s_1^k, as it is seen from simple examples. E.g., if Γ is the 5-letter graph-pentagon, then $s_1 = \alpha(\Gamma) = 2$, but $s_2 = 5 > 4$ (you can draw the 25 2-letter words in our alphabet and find 5 of them which cannot be mixed). Similarly to the inequality $s_k \geq s_1^k$, you can prove that $s_{p \times q} \geq s_p^q$ (consider s_p p-letter words which cannot be mixed as your new alphabet and note that the words comprised of these q "macro-letters" also cannot be mixed). From the relation $s_{p \times q} \geq s_p^q$ (combined with the evident relation $s_p \leq |V|^p$) it follows that there exists

$$\sigma(\Gamma) = \lim_{p \to \infty} s_p^{1/p} = \sup_p s_p^{1/p};$$

[1] to verify this statement, note that the minimum of the quadratic form $v^T Pv + 2v^T R^T u + u^T Qu$ with respect to u is given by $u = -Q^{-1} Rv$, and the corresponding minimum value is $v^T Pv - v^T R^T Q^{-1} Rv$; A is positive semidefinite if and only if this latter quantity is ≥ 0 for all v

this limit is exactly the Shannon capacity number. Since $\sigma(\Gamma) \geq s_p^{1/p}$ for every p, and, in particular, for $p = 1$, we have

$$\sigma(\Gamma) \geq \alpha(\Gamma);$$

for the above 5-letter graph we also have $\sigma(\Gamma) \geq \sqrt{s_2} = \sqrt{5}$.

The Shannon capacity number is an upper bound for the inner stability number, which is a good news; a bad news is that $\sigma(\Gamma)$ is even less computationally tractable than $\alpha(\Gamma)$. E.g., for more than 20 years nobody knew whether the Shannon capacity of the above 5-letter graph is equal to $\sqrt{5}$ or is greater than this quantity.

In 1979, Lovasz introduced a "computable" upper bound for $\sigma(\Gamma)$ (and, consequently, for $\alpha(\Gamma)$) - the *Lovasz capacity number* $\theta(\Gamma)$ which is defined as follows: let N be the number of vertices in the graph, and let the vertices be numbered by 1,...,N. Let us associate with each arc γ in the graph its own variable x_γ, and let $B(x)$ be the following symmetric matrix depending on the collection x of these variables: $B_{ij}(x)$ is 1, if either $i = j$, or the vertices i and j are not adjacent; if the vertices are linked by arc γ, then $B_{ij}(x) = x_\gamma$. For the above 5-letter graph, e.g.,

$$B(x) = \begin{pmatrix} 1 & x_{12} & 1 & 1 & x_{51} \\ x_{12} & 1 & x_{23} & 1 & 1 \\ 1 & x_{23} & 1 & x_{34} & 1 \\ 1 & 1 & x_{34} & 1 & x_{45} \\ x_{51} & 1 & 1 & x_{45} & 1 \end{pmatrix}.$$

Now, by definition the Lovasz capacity number is the minimum, over all x's, of the maximum eigenvalue of the matrix $B(x)$. Lovasz has proved that his capacity number is an upper bound for the Shannon capacity number and the inner stability number:

$$\theta(\Gamma) \geq \sigma(\Gamma) \geq \alpha(\Gamma).$$

Thus, Lovasz capacity number (which can be computed via solving a semidefinite program) gives important information on the fundamental combinatorial characteristic of a graph. In many cases the information is complete, as it happens in our example, where $\theta(\Gamma) = \sqrt{5}$; consequently, $\sigma(\Gamma) = \sqrt{5}$, since we know that for the graph in question $\sigma(\Gamma) \geq \sqrt{5}$; and since $\alpha(\Gamma)$ is integer, we can rewrite the Lovasz inequality as $\alpha(\Gamma) \leq \lfloor \theta(\Gamma) \rfloor$ and get for our example the correct answer $\alpha(\Gamma) = 2$.

11.2.4 Dual bounds in Boolean Programming

Consider another application of semidefinite programming in combinatorics. Assume that you should solve a Boolean Programming problem

$$\textit{minimize} \ \sum_{j=1}^{k} d_j u_j \ \ s.t. \ \sum_{j=1}^{k} p_{ij} u_j = q_i, \ i = 1, ..., n, u_j \in \{0; 1\}.$$

One of the standard ways to solve the problem is to use the branch-and-bound scheme, and for this scheme it is crucial to generate lower bounds for the optimal value in the subproblems arising in course of running the method. These subproblems are of the same structure as the initial problem, so that we may think of how to bound from below the optimal value in the problem. The traditional way here is to pass from the Boolean problem to its Linear Programming relaxation by replacing the Boolean restrictions $u_j \in \{0; 1\}$ with linear inequalities $0 \leq u_j \leq 1$. Some years ago Shor suggested to use nonlinear relaxation which is as follows. We can rewrite the Boolean constraints equivalently as quadratic equalities

$$u_j(1 - u_j) = 0, \ j = 1, ..., k;$$

further, we can add to our initial linear equations their quadratic implications like

$$[q_i - \sum_{j=1}^{k} p_{ij} u_j][q_{i'} - \sum_{j=1}^{k} p_{i'j} u_j] = 0, \ \ i, i' = 1, ..., n.$$

thus, we can equivalently rewrite our problem as a problem of continuous optimization with linear objective and quadratic *equality* constraints

$$\textit{minimize} \;\; d^T u \;\; \textit{s.t.} \;\; K_i(u) = 0, \, i = 1, ..., N, \tag{11.1}$$

where all K_i are quadratic forms. Let us form the Lagrange function

$$L(u, x) = d^T u + \sum_{i=1}^{N} x_i K_i(u) = u^T A(x) u + 2b^T(x) u + c(x),$$

where $A(x)$, $b(x)$, $c(x)$ clearly are affine functions of the vector x of Lagrange multipliers. Now let us pass to the "dual" problem

$$\textit{maximize} \;\; f(x) \equiv \inf_u L(u, x). \tag{11.2}$$

If our primal problem (11.1) were convex, the optimal value c_* in the dual, under mild regularity assumptions, would be the same as the optimal value in the primal problem; our situation has nothing in common with convexity, so that we should not hope that c_* is the optimal value in (11.1); anyhow, independently of any convexity assumptions c_* *is a lower bound for the primal optimal value*[2]; this is the bound suggested by Shor.

Let us look how to compute Shor's bound. We have

$$f(x) = \inf_u \{u^T A(x) u + 2b^T(x) u + c(x)\},$$

so that $f(x)$ is the largest real f for which the quadratic form of u

$$u^T A(x) u + 2b^T(x) u + [c(x) - f]$$

is nonnegative for all u; substituting $u = t^{-1} v$, we see that the latter quadratic form of u is nonnegative for all u if and only if the homogeneous quadratic form of v, t

$$v^T A(x) v + 2b^T(x) v t + [c(x) - f] t^2$$

is nonnegative whenever $t \neq 0$. By continuity reasons the resulting form is nonnegative for all v, t with $t \neq 0$ if and only if it is nonnegative for *all* v, t, i.e., if and only if the matrix

$$\mathcal{A}(f, x) = \begin{pmatrix} c(x) - f & b^T(x) \\ b(x) & A(x) \end{pmatrix}$$

is positive semidefinite. Thus, $f(x)$ is the largest f for which the matrix $\mathcal{A}(f, x)$ is positive semidefinite; consequently, the quantity $\sup_x f(x)$ we are interested in is nothing but the optimal value in the following semidefinite program:

$$\textit{maximize} \;\; f \;\; \textit{s.t.} \;\; \mathcal{A}(f, x) \geq 0.$$

It can be easily seen that the lower bound c_* given by Shor's relaxation is not worse than that one given by the usual LP relaxation. Normally the "semidefinite" bound is better, as it is the case, e.g., in the following toy problem

$$\begin{array}{rcrcrcrl} 40x_1 & + & 90x_2 & + & 28x_3 & + & 22x_4 & \to \min \\ 30x_1 & + & 27x_2 & + & 11x_3 & + & 33x_4 & = 41 \\ 28x_1 & + & 2x_2 & + & 46x_3 & + & 46x_4 & = 74 \end{array}$$

$$x_1, x_2, x_3, x_4 = 0, 1$$

with optimal value 68 ($x_1^* = x_3^* = 1, x_2^* = x_4^* = 0$); here Shor's bound is 43, and the LP-based bound is 40.

[2] the proof is immediate: if u is primal feasible, then, for any x, $L(x, u) = d^T u$ (since $K_i(u) = 0$) and therefore $f(x) \leq d^T u$; consequently, $c_* = \sup_x f(x) \leq d^T u$. Since the latter inequality is valid for all primal feasible u, c_* is $\leq$ the primal optimal value, as claimed

11.2.5 Problems arising in Control

An extremely powerful source of semidefinite problems is modern Control; there are tens of problems which are naturally formulated as semidefinite programs. Let me present two generic examples.

Proving Stability via Quadratic Lyapunov function[3]. Consider a *polytopic differential inclusion*

$$x'(t) \in Q(x(t)), \tag{11.3}$$

where

$$Q(x) = \mathrm{Conv}\{Q_1 x, ..., Q_M x\},$$

Q_i being $k \times k$ matrices. Thus, every vector $x \in \mathbf{R}^k$ is associated with the polytope $Q(x)$, and the trajectories of the inclusion are differentiable functions $x(t)$ such that their derivatives $x'(t)$ belong, for any t, to the polytope $Q(x(t))$. When $M = 1$, we come to the usual linear time-invariant system

$$x'(t) = Q_1 x(t).$$

The general case $M > 1$ allows to model *time-varying systems with uncertainty*; indeed, a trajectory of the inclusion is the solution to the time-varying equation

$$x'(t) = A(t)x(t), \;\; A(t) \in \mathrm{Conv}\{Q_1, ..., Q_M\},$$

and the trajectory of any time-varying equation of this type clearly is a trajectory of the inclusion.

One of the most fundamental questions about a dynamic system is its stability: what happens with the trajectories as $t \to \infty$ - do they tend to 0 (this is the stability), or remain bounded, or some of them go to infinity. A natural way to prove stability is to point out a quadratic Lyapunov function $f(x) = x^T L x$, L being positive definite symmetric matrix, which ”proves the decay rate α of the system”, i.e., satisfies, for some α, the inequality

$$\frac{d}{dt} f(x(t)) \le -\alpha f(x(t))$$

along *all* trajectories $x(t)$ of the inclusion. From this differential inequality it immediately follows that

$$f(x(t)) \le f(x(0)) \exp\{-\alpha t\};$$

if $\alpha > 0$, this proves stability (the trajectories approach the origin at a known exponential rate); if $\alpha = 0$, the trajectories remain bounded; if $\alpha < 0$, we do not know whether the system is stable, but we have certain upper bound for the rate at which the trajectories may go to infinity. It is worthy to note that in the case of linear time-invariant system the existence of quadratic Lyapunov function which ”proves a negative decay rate” is a *necessary and sufficient* stability condition (this is stated by the famous Lyapunov Theorem); in the general case $M > 1$ this condition is only sufficient, and is not anymore necessary.

Now, where could we take a quadratic Lyapunov function which proves stability? The derivative of the function $x^T(t)Lx(t)$ in t is $2x^T(x)Lx'(t)$; if L proves the decay rate α, this quantity should be $\le -\alpha x^T(t)Lx(t)$ for all trajectories $x(\cdot)$. Now, $x(t)$ can be an arbitrary point of $\mathbf{R}^k$, and for given $x = x(t)$ the vector $x'(t)$ can be an arbitrary vector from $Q(x)$. Thus, L ”proves decay rate α” if and only if it is symmetric positive definite (this is our a priori restriction on the Lyapunov function) and is such that

$$2x^T L y \le -\alpha x^T L x$$

for all x and for all $y \in Q(x)$; since the required inequality is linear in y, it is valid for all $y \in Q(x)$ if and only if it is valid for $y = Q_i x$, $i = 1, ..., M$ (recall that $Q(x)$ is the convex hull of the points $Q_i x$). Thus, positive definite symmetric L proves the decay rate α if and only if

$$x^T[LQ_i + Q_i^T L]x \equiv 2x^T L Q_i x \le -\alpha x^T L x$$

for all x, i.e., if and only if L satisfies the system of Linear Matrix Inequalities

$$\alpha L + LQ_i + O_i^T L \le 0, \; i = 1, ..., M; \;\; L > 0.$$

[3] this example was the subject of exercises to Lecture 7, see Section 7.6.1

Due to homogeneity with respect to L, we can impose on L nonstrict inequality $L \geq I$ instead of strict (and therefore inconvenient) inequality $L > 0$, and thus come to the necessity to solve the system

$$L \geq I;\ \alpha L + LQ_i + Q_i^T L \leq 0,\ i = 1, ..., M, \tag{11.4}$$

of Linear Matrix Inequalities, which is a positive semidefinite program with trivial objective.

Feedback synthesis via quadratic Lyapunov function. Now let us pass from differential inclusion (11.3) to a controlled plant

$$x'(t) \in Q(x(t), u(t)), \tag{11.5}$$

where

$$Q(x, u) = \mathrm{Conv}\{Q_1 x + B_1 u, ..., Q_M x + B_M u\}$$

with $k \times k$ matrices Q_i and $k \times l$ matrices B_i. Here $x \in \mathbf{R}^k$ denotes state of the system and $u \in \mathbf{R}^l$ denotes the control. Our goal is to "close" the system by a *linear time-invariant feedback*

$$u(t) = Kx(t),$$

K being $k \times l$ *feedback matrix*, in a way which ensures stability of the closed-loop system

$$x'(t) \in Q(x(t), Kx(t)). \tag{11.6}$$

Here again we can try to achieve our goal via quadratic Lyapunov function $x^T Lx$. Namely, if, for some given $\alpha > 0$, we are able to find simultaneously a $k \times l$ matrix K and a positive definite symmetric $k \times k$ matrix L in such a way that

$$\frac{d}{dt}(x^T(t)Lx(t)) \leq -\alpha x^T(t)Lx(t) \tag{11.7}$$

for all trajectories of (11.6), then we will get both the stabilizing feedback and a sertificate that it indeed stabilizes the system.

Same as above, (11.7) and the initial requirement that L should be positive definite result in the system of matrix inequalities

$$[Q_i + B_i K]^T L + L[Q_i + B_i K] \leq -\alpha L,\ i = 1, ..., M;\ L > 0; \tag{11.8}$$

the unknowns in the system are both L and K. The system is *not* linear in (L, K); nevertheless, the LMI-based approach still works. Namely, let us perform *nonlinear* substitution:

$$(L, K) \mapsto (R = L^{-1}, P = KL^{-1}) \quad [L = R^{-1}, K = PR^{-1}].$$

In the new variables the system becomes

$$Q_i^T R^{-1} + R^{-1} Q_i + R^{-1} P^T B_i^T R^{-1} + R^{-1} B_i P R^{-1} \leq -\alpha R^{-1},\ i = 1, ..., M;\ R > 0,$$

or, which is the same (multiply by R from the left and from the right)

$$RQ_i^T + Q_i R + P^T B_i^T + B_i P \leq -\alpha R,\ i = 1, ..., M; R > 0.$$

Due to homogeneity with respect to R, P, we can reduce the latter system to

$$RQ_i^T + Q_i R + P^T B_i + B_i P \leq -\alpha R,\ i = 1, ..., M; R \geq I,$$

which is a system of LMI's in variables R, P, or, which is the same, a semidefinite program with trivial objective.

There are many other examples of semidefinite problems arising in Control (and in other areas like Structural Design), but I believe that the already indicated examples demonstrate that Semidefinite Programming possesses a wide variety of important applications.

11.3 Interior point methods for Semidefinite Programming

Semidefinite Programming is a nice field for interior point methods; in fact this family of problems, due to some intrinsic mathematical properties, is very similar to Linear Programming. Let us look how the interior point methods can be applied to a semidefinite program

$$\text{minimize } c^T x \text{ s.t. } x \in G = \{x \in \mathbf{R}^n \mid A(x) \geq 0\}, \tag{11.9}$$

$A(x)$ being $m \times m$ symmetric matrix affinely depending on $x \in \mathbf{R}^n$:

$$A(x) = A_0 + \sum_{i=1}^n x_i A_i.$$

It is reasonable to assume that $A(\cdot)$ possesses certain structure, namely, that it is is block-diagonal matrix with certain number M of diagonal blocks, and the blocks are of the row sizes $m_1, ..., m_M$. Indeed, normally $A(\cdot)$ represents a *system* of LMI's rather than a single LMI; and when assembling system of LMI's

$$A_i(x) \geq 0,\ i = 1, ..., M$$

into a single LMI

$$A(x) = \mathrm{Diag}\{A_1(x), ..., A_M(x)\} \geq 0,$$

we get block-diagonal A. Note also that the "unstructured" case ($A(\cdot)$ has no nontrivial block-diagonal structure, as, e.g., in the problem associated with the Lovasz capacity number) is also covered by our assumption (it corresponds to $M = 1$, $m_1 = m$).

Path-following approach is immediate:

Standard reformulation of the problem: problem from the very beginning is in the standard form.

Barrier: by definition, the feasible set of the problem is the inverse image of the cone $\mathbf{S}^\mu_+$ of all positive semidefinite symmetric $m \times m$ matrices belonging to the space $\mathbf{S}^\mu$ of symmetric matrices of the block-diagonal structure

$$\mu = (m_1, ..., m_M)$$

(M diagonal blocks of the sizes $m_1, ..., m_M$) under the mapping

$$x \mapsto A(x) : \mathbf{R}^n \to \mathbf{S}^\mu.$$

Due to our standard combination rules, the function

$$\Phi(X) = -\ln \mathrm{Det}\, X : \mathrm{int}\ \mathbf{S}^\mu_+ \to \mathbf{R}$$

is m-logarithmically homogeneous self-concordant barrier for the cone $\mathbf{S}^\mu_+$; by construction, G is the inverse image of the cone under the affine mapping

$$x \mapsto A(x),$$

so that the function

$$F(x) = \Phi(A(x))$$

is a m-self-concordant barrier for G.

Structural assumption is satisfied simply by the origin of the barrier F: it comes from the m-logarithmically homogeneous self-concordant barrier Φ for $\mathbf{S}^\mu_+$, and the latter barrier possesses the explicit Legendre transformation

$$\Phi^*(S) = \Phi(-S) - m.$$

Complexity. The only complexity characteristic which needs special investigation is the arithmetic cost $\mathcal{N}$ of a Newton step. Let us look what is, computationally, this step. First of all, a straigtforward computation results in the following expressions for the derivatives of the barrier Φ:

$$D\Phi(X)[H] = -\mathrm{Tr}\{X^{-1}H\};\ \ D^2\Phi(X)[H,H] = \mathrm{Tr}\{X^{-1}HX^{-1}H\}.$$

Therefore the derivatives of the barrier $F(x) = \Phi(A(x))$ are given by the relations

$$\frac{\partial}{\partial x_i} F(x) = -\operatorname{Tr}\{A^{-1}(x)A_i\}$$

(recall that $A(x) = A_0 + \sum_{i=1}^n x_i A_i$),

$$\frac{\partial^2}{\partial x_i \partial x_j} F(x) = \operatorname{Tr}\{A^{-1}(x)A_i A^{-1}(X)A_j\}.$$

We see that in order to *assemble* the Newton system

$$F''(x)y = -tc - F'(x)$$

we should perform computations as follows (the expressions in brackets $\{\cdot\}$ represent the arithmetic cost of the computation; for the sake of clarity, I omit absolute constant factors):

- given x, compute $X = A(x)$ $\{n\sum_{i=1}^M m_i^2$ - you should multiply n block-diagonal matrices A_i by x_i's and take the sum of these matrices and the matrix $A_0\}$;
- given X, compute X^{-1} $\{\sum_{i=1}^M m_i^3$; recall that X is block-diagonal$\}$;
- given X^{-1}, compute n components $-\operatorname{Tr}\{X^{-1}A_i\}$ of the vector $F'(x)$ $\{n\sum_{i=1}^M m_i^2\}$;
- given X^{-1}, compute n matrices $\widehat{A}_i = X^{-1}A_iX^{-1}$ $\{n\sum_{i=1}^M m_i^3\}$ and then compute $n(n+1)/2$ quantities $F''(x)_{ij} = \operatorname{Tr}\{\widehat{A}_iA_j\}$, $1 \le i \le j \le n$ $\{n^2\sum_{i=1}^M m_i^2\}$.

The total arithmetic cost of assembling the Newton system is therefore

$$\mathcal{N}_{\rm ass} = O(n^2\sum_{i=1}^M m_i^2 + n\sum_{i=1}^M m_i^3).$$

It takes $O(n^3)$ operations more to solve the Newton system after it is assembled. Note that we may assume that $A(\cdot)$ is an embedding - otherwise the feasible set G of the problem contains lines, and the problem is unstable - small perturbation of the objective makes the problem below unbounded. Assuming from now on that $A(\cdot)$ is an embedding (as a byproduct, this assumption ensures nonsingularity of $F''(\cdot)$), we see that $n \le \sum_{i=1}^M m_i(m_i+1)/2$ - simply because the latter quantity is the dimension of the space where the mapping $A(\cdot)$ takes its values. Thus, here, as in the (dense) Linear Programming case, the cost of assembling the Newton system (which is at least $O(n^2\sum_{i=1}^M m_i^2)$) dominates the cost $O(n^3)$ of solving the system, and we come to $\mathcal{N} = O(\mathcal{N}_{\rm ass})$. Thus, the complexity characteristics of the path-following method for solving semidefinite programs are

$$\vartheta = m = \sum_{i=1}^M m_i;\ \mathcal{N} =)(n^2\sum_{i=1}^M m_i^2 + n\sum_{i=1}^M m_i^3);\ \mathcal{C} = \mathcal{N}\sqrt{m}. \tag{11.10}$$

Potential reduction approach also is immediate: *Conic reformulation of the problem* is given by

$$\textit{minimize}\ \ \operatorname{Tr}\{fy\}\ \ \textit{s.t.}\ \ y = A(x) \in \mathbf{S}^\mu_+, \tag{11.11}$$

where $f \in \mathbf{S}^\mu$ "represents the objective x^Tc in terms of $y = \sum_{i=1}^n x_iA_i$", i.e., is such that

$$\operatorname{Tr}\{fAi\} = c_i,\ i = 1, ..., n.$$

The conic dual to (11.11) is, as it is easily seen, the problem

$$\textit{minimize}\ \ \operatorname{Tr}\{A_0s\}\ \ \textit{s.t.}\ \ s \in \mathbf{S}^\mu_+,\ \operatorname{Tr}\{A_is\} = c_i,\ i = 1, ..., n. \tag{11.12}$$

Logarithmically homogeneous self-concordant barrier: we already know that $\mathbf{S}^\mu_+$ admits explicit m-

logarithmically homogeneous self-concordant barrier $\Phi(X) = -\ln \mathrm{Det}\, X$ with explicit Legendre transformation $\Phi^*(S) = \Phi(-S) - m$; thus, we have no conceptual difficulties with applying the methods of Karmarkar or the primal-dual method.

Complexity: it is easily seen that the complexity characteristics of the primal-dual method associated with the indicated barrier are given by (11.10); the characteristic $\mathcal{C}$ for the method of Karmarkar is $O(\sqrt{m})$ times worse than that one given by (11.10). **Comments.** One should take into account that in the case of

Semidefinite Programming, same as in the Linear Programming case, complexity characteristics (11.10) give very poor impression of actual performance of the algorithms. The first source of this phenomenon is that "real-world" semidefinite programs normally possess additional structure which was ignored in our evaluation of the arithmetic cost of a Newton step; e.g., for the Lyapunov Stability problem (11.4) we have $m_i = k$, $i = 1, ..., M$, k being the dimension of the state space of the system, $n = O(k^2)$ (# of design variables equals to # of free entries in a $k \times k$ symmetric matrix L). Our general considerations result in

$$\mathcal{N} = O(k^6 M)$$

(see (11.10) and in the qualitative conclusion that the cost of a step is dominated by the cost of assembling the Newton system. It turns out, anyhow, that the structure of our LMI's allows to reduce $\mathcal{N}_{\text{ass}}$ to $O(k^4 M)$, which results in $\mathcal{N} = O(k^6 + k^4 M)$; in particular, if $M << k^2$, then the cost of assembling the Newton system is negligible as compared to the cost of solving the system.

Further, numerical experiments demonstrate that the Newton complexity of finding an ε-solution of a semidefinite program by a long-step path-following or a potential reduction interior point method normally is significantly less than its theoretical $O(\sqrt{m})$ upper bound; in practice # of Newton steps looks like a moderate constant (something 30-60). Thus, Semidefinite Programming is, basically, as computationally tractable as Linear Programming.

11.4 Exercises on Semidefinite Programming

The goal of the below exercises is to demonstrate additional abilities to represent convex problems via semidefinite restrictions. Let us start with a useful definition:

let G be a closed convex domain in $\mathbf{R}^n$. We call G semidefinite representable (SDR), if there exists an affine mapping

$$A_G(x,u) : \mathbf{R}^n_x \times \mathbf{R}^l_u \to \mathbf{S}^k$$

taking values in the space $\mathbf{S}^k$ of symmetric matrices of certain row size k such that the image of A_G intersects the interior of the cone $\mathbf{S}^k_+$ of positive semidefinite symmetric $k \times k$ matrices and

$$G = \{x \mid \exists u : A_G(x,u) \in \mathbf{S}^k_+\}.$$

The above A_G is called semidefinite representation of G.

Example: the mapping

$$A(x,u) = \begin{pmatrix} u_3 - x_5 & & & & & & \\ & u_2 & u_3 & & & & \\ & u_3 & u_1 & & & & \\ & & & x_4 & u_2 & & \\ & & & u_2 & x_3 & & \\ & & & & & x_2 & u_1 \\ & & & & & u_1 & x_1 \end{pmatrix} : \mathbf{R}^5_x \times \mathbf{R}^3_u \to \mathbf{S}^7$$

(blank space corresponds to zero entries) represents the hypograph

$$G = \{x \in \mathbf{R}^5 \mid x_1, x_2, x_3, x_4 \geq 0, x_5 \leq [x_1 x_2 x_3 x_4]^{1/4}\}$$

of the geometric mean of four variables $x_1, ..., x_4$.

Indeed, positive semidefiniteness of $A(x,u)$ says that the north-western entry $u_3 - x_5$ is nonnegative, i.e.,

$$x_5 \leq u_3,$$

and that the remaining 2×2 diagonal blocks of A are positive semidefinite symmetric matrices, i.e., say that $x_1, ..., x_4, u_1, u_2$ are nonnegative and

$$u_1 \leq \sqrt{x_1 x_2},\ u_2 \leq \sqrt{x_3 x_4},\ u_3 \leq \sqrt{u_1 u_2}.$$

It is clear that a given x can be extended, by certain u, to a collection satisfying the indicated inequalities if and only if $x_1, ..., x_4$ are nonnegative and $x_5 \leq [x_1 ... x_4]^{1/4}$, i.e., if and only if $x \in G$.

The relation of the introduced notion to Semidefinite Programming is clear from the following

Exercise 11.4.1 *$i^{\#}$ Let G be an SDR domain with semidefinite representation A_G. Prove that the convex program*

$$\textit{minimize } c^T x \text{ s.t. } x \in G$$

is equivalent to the semidefinite program

$$\textit{minimize } c^T x \text{ s.t. } A_G(x,u) \geq 0.$$

SDR domains admit a kind of calculus:

Exercise 11.4.2 $^{\#}$. *1) Let $G^+ \subset \mathbf{R}^n$ be SDR, and let $x = B(y)$ be an affine mapping from $\mathbf{R}^l$ into $\mathbf{R}^n$ with the image intersecting* int G^+. *Prove that $G = B^{-1}(G^+)$ is SDR, and that a semidefinite representation of G^+ induces, in an explicit manner, a semidefinite representation of G.*

2) Let $G = \cap_{i=1}^m G_i$ be a closed convex domain in $\mathbf{R}^n$, and let all G_i be SDR. Prove that G also is SDR, and that semidefinite representations of G_i induce, in an explicit manner, a semidefinite representation of G.

3) Let $G_i \subset \mathbf{R}^{n_i}$ be SDR, $i = 1, ..., m$. Prove that the direct product $G = G_1 \times G_2 \times ... \times G_m$ is SDR, and that semidefinite representations of G_i induce, in an explicit manner, a semidefinite representation of G.

The above exercises demonstrate that the possibilities to pose convex problems as semidefinite programs are limited only by our abilities to find semidefinite representations for the constraints involved into the problem. The family of conves sets which admit explicit semidefinite representations is surprisingly wide. Lecture 11 already gives us a number of examples which are summarized in the following

Exercise 11.4.3 $^{\#}$ *Verify that the below sets are SDR and point out their explicit semidefinite representations:*

- *half-space*
- *Lebesque set* $\{x \mid f(x) \leq 0\}$ *of a convex quadratic form, such that* $f(x) < 0$ *for some* x
- *the second order cone* $K^2 = \{(t,x) \in \mathbf{R} \times \mathbf{R}^n \mid t \geq |x|_2\}$
- *the epigraph* $\{(t,X) \in \mathbf{R} \times \mathbf{S}^k \mid t \geq \lambda_{\max}(X)\}$ *of the masimal eigenvalue of a symmetric* $k \times k$ *matrix* X

Now some more examples.

Exercise 11.4.4 *Prove that*

$$A(t,x) = \mathrm{Diag}\{t - x_1, t - x_2, ..., t - x_n\}$$

is SDR for the epigraph

$$\{(t,x) \in \mathbf{R} \times \mathbf{R}^n \mid t \geq x_i,\, i = 1, ..., n\}$$

of the function $\max\{x_1, ..., x_n\}$.

Exercise 11.4.5 *Prove that*

$$A(t,x) = \begin{pmatrix} t & x^T \\ x & X \end{pmatrix}$$

is SDR for the epigraph

$$\mathrm{cl}\{(t, xi, X) \in \mathbf{R} \times \mathbf{R}^n \times (\mathrm{int}\, \mathbf{S}^n_+) \mid t \geq x^T X^{-1} x\}$$

of fractional-quadratic funtion $x^T X^{-1} x$ *of vector* x *and symmetric positive semidefinite matrix* X.

Exercise 11.4.6 *The above Example gives a SDR of the hypograph of the geometrical mean* $[x_1...x_4]^{1/4}$ *of four nonnegative variables. Find SDR for the hypograph of the geometrical mean of* 2^l *nonnegative variables.*

Exercise 11.4.7 *Find semidefinite representation of the epigraph*

$$\{(t,x) \in \mathbf{R}^2 \mid p \geq (x_+)^p\},\ x_+ = \max[0,x],$$

of the power function for
1) $p = 1$; *2)* $p = 2$; *3)* $p = 3$; *4) arbitrary integer* $p > 0$.

11.4.1 Sums of eigenvalues and singular values

For a symmetric $k \times k$ matrix X let $\lambda_1(X) \geq \lambda_2(X) \geq ... \geq \lambda_k(X)$ be the eigenvalues of X written down with their multiplicities in the descent order. To the moment all we know about convexity of eigenvalues is that the maximum eigenvalue $\lambda_1(X)$ is convex; we know even a SDR for this function (Exercise 11.4.3). the remaining eigenvalues $\lambda_i(X)$, $i \geq 2$, simply are non convex in X. Nevertheless, they possess nice property of monotonicity:

$$X, X' \in \mathbf{S}^k,\ X \leq X' \to \lambda_i(X) \leq \lambda_i(X'),\ i = 1, ..., k.$$

This is an immediate corollary of the *Courant-Fisher characterization of eigenvalues*[4]:

$$\lambda_i(X) = \max_{E \in \mathcal{E}_i} \min_{u \in E, |u|=1} u^T X u,$$

[4] I strongly recommend to those who do not know this characterization pay attention to it; a good (and not difficult) exercise if to prove the characterization

$\mathcal{E}_i$ being the family of all linear subspaces in $\mathbf{R}^k$ of the dimension i.

An important fact is that the functions

$$S_m(x) = \sum_{i=1}^{m} \lambda_i(X),\ 1 \le m \le k,$$

are convex.

Exercise 11.4.8 $^+$ *Prove that*

$$A_m(t, X; \tau, U) = \begin{pmatrix} t - m\tau - \operatorname{Tr} U & 0 & 0 \\ 0 & \tau I + U - X & 0 \\ 0 & 0 & U \end{pmatrix}$$

(τ is scalar, U is symmetric $k \times k$ matrix) is a SDR for the epigraph

$$\{(t, X) \in \mathbf{R} \times \mathbf{S}^k \mid t \ge S_m(X)\};$$

in particular, $S_m(x)$ is convex (since its epigraph is SDR and is therefore convex) monotone function.

For an *arbitrary* $k \times k$ matrix X let $\sigma_i(X)$ be the *singular values* of X, i.e., square roots of the eigenvalues of the matrix $X^T X$. In what follows we always use the descent order of singular values:

$$\sigma_1(X) \ge \sigma_2(X) \ge ... \ge \sigma_k(X).$$

Let also

$$\Sigma_m(X) = \sum_{i=1}^{m} \sigma_i(X).$$

The importance of singular values is seen from the following fundamental Singular Value Decomposition Theorem (which for non-symmetric matrices plays basically the same role as the theorem that a symmetric matrix is orthogonally equivalent to a diagonal matrix):

If X is a $k \times k$ matrix with singular values σ_1,..., σ_k, then there exist pair of orthonormal basises $\{e_i\}$ and $\{f_i\}$ such that

$$X = \sum_{i=1}^{k} \sigma_i e_i f_i^T$$

(geometrically: the mapping $x \to Xx$ takes the coordinates of x in the basis $\{f_i\}$, multiplies them by the singular values and makes the result the coordinates of Xx in the basis $\{e_i\}$).

In particular, the spectral norm of X (the quantity $\max_{|x|_2 \le 1} |Xx|_2$) is nothing but the largest singular value σ_1 of X.

In the symmetric case we, of course, have $e_i = \pm f_i$ (plus corresponds to eigenvectors f_i of X with positive, minus - to those with negative eigenvalues).

What we are about to do is to prove that the functions $\Sigma_m(X)$ are convex, and to find their SDR's. To this end we make the following important observation:

let A and B be two $k \times k$ matrices. Then the sequences of eigenvalues (counted with their multiplicities) of the matrices AB and BA are equal (more exactly, become equal under appropriate reordering). The proof is immediate: we should prove that the characteristic polynomials $\operatorname{Det}(\lambda I - AB)$ and $\operatorname{Det}(\lambda I - BA)$ are equal to each other. By continuouty reasons, it suffices to establish this identity when A is nondegenerate. But then it is evident:

$$\operatorname{Det}(\lambda I - AB) = \operatorname{Det}(A(\lambda I - BA)A^{-1}) = (\operatorname{Det}\ A)\operatorname{Det}(\lambda I - BA)(\operatorname{Det}(A^{-1})) = \operatorname{Det}(\lambda I - BA).$$

Now we are enough equipped to construct SDR's for sums of singular values.

Exercise 11.4.9 $^+$ *Given a $k \times k$ matrix X, form the symmetric $2k \times 2k$ matrix*

$$Y(X) = \begin{pmatrix} 0 & X \\ X^T & 0 \end{pmatrix}.$$

Prove that the eigenvalues of this matrix are as follows: the first k of them are $\sigma_1(X)$,$\sigma_2(X)$,..., $\sigma_k(X)$, and the remaining k are $-\sigma_k(X)$, $-\sigma_{k-1}(X)$,...,$-\sigma_1(X)$. Derive from this observation that

$$\Sigma_m(X) = S_m(Y(X))$$

and use SDR's for $S_m(\cdot)$ given by Exercise 11.4.8 to get SDR's for $\Sigma_m(X)$.

The results stated in the exercises from this subsection play the central role in constructing semidefinite representations for the epigraphs of functions of eigenvalues/singular values of symmetric/arbitrary matrices.

Concluding remark. We see that the family of SDR sets is very rich, and, consequently, the family of convex problems which can be reformulated as semidefinite programs also is very rich. Nevertheless, the family of SDR convex domains is smaller than the family of *all* convex domains. Indeed, a SDR domain $G \subset \mathbf{R}^n_x$, by definition, is a projection of the inverse image $G^+ \subset \mathbf{R}^n_x \times \mathbf{R}^l_u$ of the cone of positive semidefinite matrices of certain dimension under an affine mapping $(u, x) \mapsto A_G(u, x)$. The cone of positive semidefinite matrices is a *semianalytic* set, i.e., a set given by finitely many nonstrict polynomial inequalities (to say that a symmetric matrix is positive semidefinite is the same as to say that all its principal minors are nonnegative). Consequently, the inverse image of the cone under affine mapping - our G^+ - also is semianalytic. Now, there exists fundamental Tarski's Theorem which states that a projection of a semianalytic set again is semianalytic. Thus, a *necessary* condition for a convex domain to be SDR is semianalyticity of the domain. I do not know whether this condition is sufficient. I do not know the answer even to the following

Open question: Let $p(x)$ be convex polynomial on the axis. Is it true that the epigraph of p is SDR?

Hints to Exercises

Hints to Section 2.3

Exercise 2.3.7: apply (P) to scalar symmetric forms $u^T A[h_1, ..., h_k]$, u being a vector with

$$\| u \|_* \equiv \sup_{v \in \mathbf{R}^k \min \|v\| \le 1} u^T v \le 1.$$

Hints to Section 3.3

Exercise 3.3.2$^+$:
1): the function

$$F(x) = -\sum_{i=j}^{m} \ln(-f_j(x)) \equiv \sum_{j=1}^{n} F_j(x)$$

is self-concordant barrier for G (Exercise 3.3.1). Since G is bounded, F attains its minimum on int G at certain point x^* (**V.**, Lecture 3). Choosing appropriate coordinates in $\mathbf{R}^n$, we may assume that $F''(x^*)$ is the unit matrix. Now let j^* be the index of that one of the matrices $F''_j(x^*)$ which has the minimal trace; eliminate j^*th of the inequalities and look at the Newton decrement of the self-concordant function $\sum_{j \neq j^*} F_j(x)$ at x^*.

2): we clearly can eliminate from the list of the sets G_α all elements which coincide with the whole space, without violating boundedness of the intersection. Now, every closed convex set which differs from the whole space is intersection of closed half-spaces, and these half-spaces can be chosen in such a way that their interiors have the same intersection as the half-spaces themselves. Representing all G_α as intersections of the above type, we see that the statement in question clearly can be reduced to a similar statement with all G_α being closed half-spaces such that the intersection of the interiors of these half-spaces is the same as the intersection of the half-spaces themselves. Prove that if $\cap_{\alpha \in I} G_\alpha$ is bounded and nonempty, then there exists a finite $I' \subset I$ such that $\cap_{\alpha \in I'} G_\alpha$ also is bounded (and, of course, nonempty); after this is proved, apply 1).

Exercise 3.3.5: this is an immediate consequence of **II.**, Lecture 3.

Exercise 3.3.6: without loss of generality we may assume that $\Delta = (a, 0)$ with some $a < 0$. Choose an arbitrary $x \in \Delta$ and look what are the conclusions of **II.**, **III.**, Lecture 3, when $y \to -0$.

To complete the proof of (P), note that if G differs from $\mathbf{R}^n$, then the intersection of G with certain line is a sgement Δ with a nonempty interior which is a proper part of the line, and choose as f the restriction of F onto Δ (this restriction is a ϑ-self-concordant barrier for Δ in view of Proposition 3.1.1.(i)).

Exercise 3.3.7: note that the standard basis orths e_i are recessive directions of G (see Corollary 3.2.1) and therefore, according to the Corollary,

$$-DF(x)[e_i] \geq \{D^2 F(x)[e_i, e_i]\}^{1/2}. \tag{12.13}$$

To prove (3.17), combine (12.13) and the fact that $D^2 F(x)[e_i, e_i] \geq x_i^{-2}$, $1 \leq i \leq m$ (since $x - x_i e_i \notin$ int G, while the open unit Dikin ellipsoid of F centered at x is contained in int G (**I.**, Lecture 2)).

To derive from (3.17) the lower bound $\vartheta \geq m$, note that, in view of **II.**, Lecture 3, it should be

$$\vartheta \geq DF(x)[0 - x],$$

while (3.17) says that the latter quantity is at least m. ■

Exercise 3.3.9: as it was already explained, we can reduce the situation to the case of

$$G \cap U = \{x \in U \mid x_i \geq h_i(x),\ i = 1, ..., m\},$$

where $h_i(0) = 0$, $h'_i(0) = 0$. It follows that the interval

$$x(r) = r \sum_{i=1}^{m} e_i,\ 0 < r < r_0,$$

associated with certain $r_0 > 0$, belongs to G; here e_i are the standard basis orths. Now, let $\Delta_i(r)$ be the set of those t for which the vector $x(r) - (t + r)e_i$ belongs to G. Prove that $\Delta_i(r)$ is of the type $[-a_i(r), b_i(r)]$ which contains in its interior r, and that $b_i(r)/r \to 0$, $a_i(r)/r \to \infty$ as $r \to +0$. Derive from these observations and the statement of Exercise 3.3.8 that $-DF(x(r))[e_i]r \geq 1 - \alpha(r)$, $i = 1, ..., m$, with certain $\alpha(r) \to 0$, $r \to +0$. To complete the proof of (Q), apply the Semiboundedness inequality **I.**, Lecture 3, to $x = x(r)$ and $y = 0$.

Hints to Section 7.6

Exercise 7.6.7: (Pr') could be used, but not when we intend to solve it by the primal-dual method. Indeed, it is immediately seen that if (7.44) is solvable, i.e., in the case we actually are interested in, the objective in (Pr') is below unbounded, so that the problem dual to (Pr') is unfeasible (why?) Thus, we simply would be unable to start the method!

Hints to Section 8.5

Exercise 8.5.1: we could, of course, assume that the Legendre transformation F^* of F is known; but it would be less restrictive to assume instead that the solution to the problem is given in advance. Indeed, knowledge of F^* means, in particular, ability to solve "in one step" any equation of the type $F'(x) = d$ (the solution is given by $x = (F^*)'(d)$); thus, setting $x = (F^*)'(-10^{20}c)$, we could get - in one step - the point of the path $x^*(\cdot)$ associated with $t = 10^{20}$.
Exercise 8.5.3: to get (8.34), prove by induction that

$$D^j\Phi(v)[h,...,h] = (-1)^j(j-1)!\,\mathrm{Tr}\{[v^{-1}h]^j\}$$

(use the rule $\frac{d}{dt}|_{t=0}(v+th)^{-1} = -v^{-1}hv^{-1}$). To derive (8.35) from (8.34), pass to the eigenbasis of $\widehat{h}$.
Exercise 8.5.5: combine the result of Exercise 5.4.4, the "symmetric" to this result statement and the result of Exercise 8.5.2.

Hints to Section 10.5

Exercise 10.4: prove$^+$ that the mapping

$$\mathcal{A}(t,X) = t + \ln \operatorname{Det} \, X : \mathbf{R} \times \operatorname{int} \mathbf{S}^n_+ \to \mathbf{R}$$

is $\frac{2}{3}$-appropriate for the domain $G^+ = \mathbf{R}_+$ and apply Superposition rule (N) from Lecture 9.
Exercise 10.5.7$^+$: let for a vector v the set L_v on the axis be defined as

$$L_v = \{\lambda \geq 0 \mid \; v^T R v \leq \lambda v^T S v\}.$$

This is a closed convex set, and the premise of the statement we are proving says that the set is nonempty for every v; and the statement we should prove is that all these sets have a point in common. Of course, the proof should use the Helley Theorem; according to this theorem, all we should prove is that

(a) $L_v \cap L_{v'} \neq \emptyset$ for any pair v, v';

(b) L_v is bounded for some v.

Solutions to Exercises

Solutions to Section 2.3

Exercise 2.3.3: let $\mathcal{A}$ be the set of all multiindices $\alpha = (\alpha_1, ..., \alpha_k)$ with nonnegative integer entries α_i and the sum of entries equal to k, let S_k be the # of elements in $\mathcal{A}$, and let for $\alpha \in \mathcal{A}$

$$A_\alpha[h_1, ..., h_k] = A[\overbrace{h_1, ..., h_1}^{\alpha_1 \text{ times}}, \overbrace{h_2, ..., h_2}^{\alpha_2 \text{ times}}, ..., \overbrace{h_k, ..., h_k}^{\alpha_k \text{ times}}]$$

For k-dimensional vector $r = (r_1, ..., r_k)$ we have, identically in $h_1, ..., h_k \in \mathbf{R}^n$:

$$A[\sum_{i=1}^{k} r_i h_i, \sum_{i=1}^{k} r_i h_i, ..., \sum_{i=1}^{k} r_i h_i] = \sum_{\alpha \in \mathcal{A}} \omega_\alpha(r) A_\alpha[h_1, ..., h_k] \tag{13.14}$$

(open parentheses and take into account symmetry of A), with $\omega_\alpha(r)$ being certain polynomials of r.

What we are asked to do is to find certain number m of vectors r^1, r^2,...,r^m and certain weights $w_1, ..., w_m$ in such a way that when substituting $r = r^l$ into (13.14) and taking sum of the resulting identities with the weights $w_1, ..., w_m$, we get in the right hand side the only term $A[h_1, ..., h_k] \equiv A_{(1,...,1)}[h_1, ..., h_k]$, with the unit coefficient; then the resulting identity will be the required representation of $A[h_1, ..., h_k]$ as a linear combination of the restriction of $A[\cdot]$ onto the diagonal.

Our reformulated problem is to choose m vectors from the family

$$\mathcal{F} = \{\widehat{\omega}(r) = (\omega_\alpha(r) \mid \alpha \in \mathcal{A})\}_{r \in \mathbf{R}^k}$$

of S_k-dimensional vectors in such a way that certain given S_k-dimensional vectors (unit at certain specified place, zeros at the remaining places) will be a linear combination of the selected vectors. This for sure is possible, with $m = S_k$, if the linear span of vectors from $\mathcal{F}$ is the entire space $\mathbf{R}^{S_k}$ of S_k-dimensional vectors; and we are about to prove that this is actually the case (this will complete the proof). Assume, on contrary, that the linear span of $\mathcal{F}$ is a proper subspace in $\mathbf{R}^{S_k}$. Then there exists a nonzero linear functional on the space which vanishes on $\mathcal{F}$, i.e., there exists a set of coefficients λ_α, not all zeros, such that

$$p(r) \equiv \sum_{\alpha \in \mathcal{A}} \lambda_\alpha \omega_\alpha(r) = 0$$

identically in $r \in \mathbf{R}^k$. Now, it is immediately seen what is ω_α:

$$\omega_\alpha(r) = \frac{k!}{\alpha_1! \alpha_2! ... \alpha_k!} r_1^{\alpha_1} r_2^{\alpha_2} ... r_k^{\alpha_k}.$$

It follows that the partial derivative $\frac{\partial^k}{\partial^{\alpha_1} r_1 \partial^{\alpha_2} r_2 ... \partial^{\alpha_k} r_k}$ of $p(\cdot)$ is identically equal to λ_α; if $p \equiv 0$, then all these derivatives, and, consequently, all λ_α's, are zero, which is the desired contradiction. ■

Exercise 2.3.5: first of all, e_1 and e_2 are linearly independent since $T_1 \neq T_2$, therefore $h \neq 0$, $q \neq 0$. Let $(Qx, y) = A[x, y, e_3, ..., e_l]$; then Q is a symmetric matrix.

Since $\{T_1, ..., T_l\}$ is an extremal, we have

$$\omega = |(Qe_1, e_2)| = \max\{|(Qu, v)| \mid \| u \|, \| v \| \leq 1\}.$$

Therefore if $E^+ = \{x \in \mathbf{R}^n \mid Qx = \omega x\}$, $E^- = \{x \in \mathbf{R}^n \mid Qx = -\omega x\}$ and $E = (E^+ + E^-)^\perp$, then at least one of the subspaces E^+, E^- is nonzero, $\| Qx \| \leq \omega' \| x \|$, $x \in E$, where $\omega' < \omega$. $\mathbf{R}^n$ is the direct sum of E^+, E^- and E. Let $x = x^+ + x^- + x'$ be the decomposition of $x \in \mathbf{R}^n$ corresponding to the decomposition $\mathbf{R}^n = E^+ + E^- + E$. Since each of the subspaces E^+, E^- and E is invariant for Q,

$$\begin{aligned}
\omega &= |(Qe_1, e_2)| \leq |\omega(e_1^+, e_2^+) - \omega(e_1^-, e_2^-)| + \omega' \| e_1' \| \| e_2' \| \\
&\leq \omega(\| e_1^+ \| \| e_2^+ \| + \| e_1^- \| \| e_2^- \|) + \omega' \| e_1' \| \| e_2' \| \\
&\leq \omega\{\| e_1^+ \|^2 + \| e_1^- \|^2\}^{1/2}\{\| e_2^+ \|^2 + \| e_2^- \|^2\}^{1/2} + \omega' \| e_1' \| \| e_2' \| \\
&\leq \omega
\end{aligned}$$

(we have taken into account that $\| e_i^+ \|^2 + \| e_i^- \|^2 + \| e_i' \|^2 = 1, \quad i = 1, 2$). We see that all the inequalities in the above chain are equalities. Therefore we have

$$\| e_1' \| = \| e_2' \| = 0; \quad \| e_1^+ \| = \| e_2^+ \|; \quad \| e_1^- \| = \| e_2^- \|;$$

moreover, $|(e_1^+, e_2^+)| = \| e_1^+ \| \| e_2^+ \|$ and $|(e_1^-, e_2^-)| = \| e_1^- \| \| e_2^- \|$, which means that $e_1^+ = \pm e_2^+$ and $e_1^- = \pm e_2^-$. Since e_1 and e_2 are linearly independent, only two cases are possible:

(a) $e_1^+ = e_2^+ \neq 0$, $e_1^- = -e_2^- \neq 0$, $e_1' = e_2' = 0$;

(b) $e_1^+ = -e_2^+ \neq 0$, $e_1^- = e_2^- \neq 0$, $e_1' = e_2' = 0$.

In case (a) h is proportional to e_1^+, q is proportional to e_1^-, therefore

$$\{\mathbf{R}h, \mathbf{R}h, T_3, ..., T_l\} \in \mathsf{T}$$

and

$$\{\mathbf{R}q, \mathbf{R}q, T_3, ...T_l\} \in \mathsf{T}.$$

The same arguments can be used in case (b). ■

Exercise 2.3.6: let $e \in T$ and $f \in S$ be unit vectors with the angle between them being equal to $\alpha(\mathcal{T})$. Without loss of generality we can assume that $t \leq s$ (note that reordering of an extremal leads to an extremal, since A is symmetric). By virtue of Exercise 2.3.5 in the case of $\alpha(\mathcal{T}) \neq 0$ the collection

$$\mathcal{T}' = \{\overbrace{\mathbf{R}(e+f), ..., \mathbf{R}(e+f)}^{2t \text{ times}}, \overbrace{S, ..., S}^{s-t \text{ times}}\}$$

belongs to T^* and clearly $\alpha(\mathcal{T}') = \alpha(\mathcal{T})/2$. Thus, either T^* contains an extremal $\mathcal{T}$ with $\alpha(\mathcal{T}) = 0$, or we can find a sequence $\{\mathcal{T}_i \in \mathsf{T}^*\}$ with $\alpha(\mathcal{T}_i) \to 0$. In the latter case the sequence $\{\mathcal{T}_i\}$ contains a subsequence converging (in the natural sense) to certain collection $\mathcal{T}$, which clearly belongs to T^*, and $\alpha(\mathcal{T}) = 0$. Thus, T contains an extremal $\mathcal{T}$ with $\alpha(\mathcal{T}) = 0$, or, which is the same, an extremal of the type $\{T, ..., T\}$. ■

Solutions to Section 3.3

Exercise 3.3.1: F clearly is C^3 smooth on $Q = \mathrm{int}\, G$ and possesses the barrier property, i.e., tends to ∞ along every sequence of interior points of G converging to a boundary point. Let $x \in Q$ and $h \in \mathbf{R}^n$. We have

$$F(x) = -\ln(-f(x)); \; DF(x)[h] = -\frac{Df(x)[h]}{f(x)};$$

$$D^2F(x)[h,h] = \frac{[Df(x)[h]]^2}{f^2(x)} - \frac{D^2f(x)[h,h]}{f(x)} = [DF(x)[h]]^2 + \frac{D^2f(x)[h,h]}{|f(x)|};$$

$$D^3F(x)[h,h,h] = -2\frac{[Df(x)[h]]^3}{|f|^3(x)} + 3\frac{Df(x)[h]D^2f(x)[h,h]}{f^2(x)}.$$

Since f is convex, we immediately conclude that

$$D^2F(x)[h,h] = r^2 + s^2, \; r = \sqrt{\frac{D^2f(x)[h,h]}{|f(x)|}}, \; s = \frac{|Df(x)[h]|}{|f(x)|},$$

$$|DF(x)[h]| = s \le \sqrt{D^2F(x)[h,h]}$$

and

$$|D^3F(x)[h,h,h]| \le 2s^3 + 3sr^2 \le 2(s^2+r^2)^{3/2}$$

(verify the concluding inequality yourself!). The resulting bounds on DF and D^2F demonstrate that F is self-concordant and that $\lambda(F,\cdot) \le 1$, so that F is a 1-self-concordant barrier for G.

The concluding statement of the exercise in question follows from the already proved one and Proposition 3.1.1. ∎

Exercise 3.3.2:

1): according to Exercise 3.3.1, F is self-concordant barrier for G; since G is bounded, F is nondegenerate (**II.**, Lecture 2) and attains its minimum at certain point x^* (**V.**, Lecture 3). Choosing appropriate coordinates in $\mathbf{R}^n$, we may assume that $F''(x^*) = I$, I being the unit matrix. Now let $F_j(x) = -\ln(-f_j(x))$, $Q_j = F_j''(x^*)$, so that $F = \sum_j F_j$ and $I = \sum_j Q_j$. We have $\sum_{j=1}^m \mathrm{Tr}\, Q_j = \mathrm{Tr}\, I = n$, so that $\mathrm{Tr}\, Q_{j^*} \le n/m$, j^* being the index of Q_j with the smallest trace. To simplify notation, in what follows we assume that $j^* = 1$. An immediate computation implies that

$$Q_1 = gg^T + H, \quad g = F_1'(x^*), \quad H = \frac{f_1''}{|f_1(x^*)|};$$

it is seen that $H \ge 0$, so that $\frac{n}{m} \ge \mathrm{Tr}\, Q_1 \ge \mathrm{Tr}\{gg^T\} = |g|_2^2$.

Now let us compute the Newton decrement of the function

$$\Phi(x) = \sum_{j=2}^m F_j(x)$$

at the point x^*. Since the gradient of F at the point is 0, the gradient of Φ is $-g$; since the Hessian of F at x^* is I, the Hessian of Φ is $I - Q_1 \ge (1 - \frac{n}{m})I$ (the latter inequality immediately follows from the fact that $Q_1 \ge 0$ and $\mathrm{Tr}\, Q_1 \le \frac{n}{m}$. We see that

$$\lambda^2(\Phi, x^*) = [\Phi'(x^*)]^T[\Phi''(x^*)]^{-1}\Phi'(x^*) = g^T[\Phi''(x^*)]^{-1}g \le |g|_2^2(1-\frac{n}{m})^{-1} \le \frac{n}{m-n} < 1$$

(we have used the already proved estimate $|g|_2^2 \le \frac{n}{m}$ and the fact that $m > 2n$). Thus, the Newton decrement of a nondegenerate (in view of $\Phi''(x^*) > 0$) self-concordant barrier (in view of Exercise 3.3.1) Φ for the convex domain $G^+ = \{x \in \mathbf{R}^n \mid f_j(x) \le 0, j = 2, ..., m\}$ is < 1; therefore Φ attains its minimum on $\mathrm{int}\, G^+$ (**VII.**, Lecture 2). Since Φ is a nondegenerate self-concordant barrier for G^+, the latter is possible only when G^+ is bounded (**V.**, Lecture 3). ∎

2): as explained in Hints, we can reduce the situation to that one with G_α being closed half-spaces such that the intersection of the interiors of these half-spaces coincides with the intersection of the half-spaces themselves; in particular, the intersection of any finite subfamily of the half-spaces G_α possesses a nonempty interior. Let us first prove that there exists a finite $I' \subset I$ such that $\cap_{\alpha\in I'} G_\alpha$ is bounded. Without loss of generality we may assume that $0 \in G_\alpha$, $\alpha \in I$ (since the intersection of all G_α is nonempty). Assume that for every finite subset I' of I the intersection $G^{I'} = \cap_{\alpha\in I'} G_\alpha$ is unbounded. Then for every $R > 0$ and every I' the set $G^{I'}_R = \{x \in G^{I'} \mid |x|_2 = R\}$ is a nonempty compact set; these compact sets form a nested family and therefore their intersection is nonempty, which means that $\cap_{\alpha\in I} G_\alpha$ contains, for every $R > 0$, a vector of the norm R and is therefore an unbounded set, which in fact is not the case.

Thus, we can reduce the situation to a similar one for a *finite* family of closed half-spaces G_α with the intersection of the interiors being bounded and nonempty; for this case the required statement is given by 1). ■

Remark 13.0.1 I do not think that the above proof of item 1) of Exercise 3.3.2 is the simplest one; please try to find a better proof.

Exercise 3.3.3: it is clear that F is C^3 smooth on the interior of $\mathbf{S}^m_+$ and possesses the barrier property, i.e., tends to ∞ along every sequence of interior point of the cone converging to a boundary point of it. Now, let x be an interior point of $\mathbf{S}^m_+$ and h be an arbitrary direction in the space $\mathbf{S}^m$ of symmetric $m \times m$ matrices, which is the embedding space of the cone. We have

$$F(x) = -\ln \mathrm{Det}\, x;$$

$$DF(x)[h] = \frac{\partial}{\partial t}|_{t=0}[-\ln \mathrm{Det}\,(x+th)] = \frac{\partial}{\partial t}|_{t=0}[-\ln \mathrm{Det}\; x - \ln \mathrm{Det}\,(I + tx^{-1}h)] =$$

$$= -\frac{\frac{\partial}{\partial t}|_{t=0}\,\mathrm{Det}\,(I+tx^{-1}h)}{\mathrm{Det}\,(I)} = -\,\mathrm{Tr}(x^{-1}h)$$

(to understand the concluding step, look at the matrix $I + tx^{-1}h$; its diagonal entries are $1 + t[x^{-1}h]_{ii}$, and the entries outside the diagonal are of order of t. Representing the determinant as the sum of products, we obtain $m!$ terms, one of them being $\prod_i (1 + t[x^{-1}h]_{ii})$ and the remaining being of the type $t^k p$ with $k \geq 2$ and p independent of t. These latter terms no not contribute to the derivative with respect to t at $t = 0$, and the contribution of the "diagonal" term is exactly $\sum_i [x^{-1}h]_{ii} = \mathrm{Tr}(x^{-1}h)$).

Thus,

$$DF(x)[h] = -\,\mathrm{Tr}(x^{-1}h),$$

whence

$$D^2F(x)[h,h] = \mathrm{Tr}(x^{-1}hx^{-1}h)$$

(we have already met with the relation $DB(x)[h] = -B(x)hB(x)$, $B(x) \equiv x^{-1}$; to prove it, differentiate the identity $B(x)x \equiv I$).

Differentiating the expression for D^2F, we come to

$$D^3F(x)[h,h,h] = -2\,\mathrm{Tr}(x^{-1}hx^{-1}hx^{-1}h)$$

(we again have used the rule for differentiating the mapping $x \mapsto x^{-1}$). Now, x is positive definite symmetric matrix; therefore there exists a positive semidefinite symmetric y such that $x^{-1} = y^2$. Replacing x^{-1} by y and taking into account that $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$, we come to the expressions

$$DF(x)[h] = -\,\mathrm{Tr}\,\xi,\ D^2F(x)[h,h] = \mathrm{Tr}\,\xi^2,\ D^3F(x)[h,h,h] = -2\,\mathrm{Tr}\,\xi^3,\ \ \xi = yhy$$

(compare these relations with the expressions for the derivatives of the function $-\ln t$). The matrix ξ clearly is symmetric; expressing the traces via the eigenvalues $\lambda_1, ..., \lambda_m$ of the matrix ξ, we come to

$$DF(x)[h] = -\sum_{i=1}^m \lambda_i;\ D^2F(x)[h,h] = \sum_{i=1}^m \lambda_i^2;\ D^3F(x)[h,h,h] = -2\sum_{i=1}^m \lambda_i^3,$$

which immediately implies the desired inequalities

$$|DF(x)[h]| \leq \sqrt{m}\sqrt{D^2F(x)[h,h]}$$

and

$$|D^3F(x)[h,h,h]| \leq 2\left[D^2F(x)[h,h]\right]^{3/2}.$$

■

Exercise 3.3.8: If $\Delta = (-\infty, 0]$, then the statement in question is given by Corollary 3.2.1. From now on we assume that Δ is finite (i.e., that $a < +\infty$). Then f attains its minimum on int Δ at a unique point t^* (**V.**, Lecture 3), and t^* partitiones Δ in the ratio not exceeding $(\vartheta + 2\sqrt{\vartheta}) : 1$ (this is the centering property stated by the same **V.**). Thus, $t^* \leq -a/(\vartheta + 2\sqrt{\vartheta} + 1)$; the latter quantity is $< t$, since $\gamma t \in \Delta$ and therefore $t \geq -a/\gamma$. Since $t^* < t$, we have $f'(t) > 0$. Note that we have also establish that

$$t/t^* \leq \frac{(1+\sqrt{\vartheta})^2}{\gamma}.$$

Let λ be the Newton decrement of a self-concordant function f at t; since $f'(t) > 0$, we have

$$\lambda = f'(t)/\sqrt{f''(t)}.$$

Note that $f''(t) \geq t^{-2}$ (because the open Dikin ellipsoid of f centered at t should be contained in int Δ and 0 is a boundary point of Δ), and therefore

$$\lambda \leq -tf'(t). \tag{13.15}$$

It is possible, first, that $\lambda \geq 1$. If it is the case, then (3.19) is an immediate consequence of (13.15).

It remains to consider the case when $\lambda < 1$. In this case, in view of **VIII.**, Lecture 2, we have

$$f(t) \leq f(t^*) + \rho(\lambda), \quad \rho(s) = -\ln(1-s) - s.$$

On the other hand, from the Lower bound on f (**III.**, Lecture 3) it follows that

$$f(t) \geq f(t^*) + f'(t^*)(t-t^*) - \ln(1-\pi_{t^*}(t)) - \pi_{t^*}(t) \equiv f(t^*) + \rho(\pi_{t^*}(t)).$$

Thus, we come to

$$\rho(\lambda) \geq \rho(\pi_{t^*}(t)),$$

whence

$$\lambda \geq \pi_{t^*}(t) \equiv |(t-t^*)/t^*| \geq 1 - \frac{(1+\sqrt{\vartheta})^2}{\gamma}.$$

Combining this inequality with (13.15), we come to

$$-tf'(t) \geq 1 - \frac{(1+\sqrt{\vartheta})^2}{\gamma},$$

as required in (3.19). (3.20) is nothing but (3.19) applied to the restriction of F onto the contained in G part of the line passing through x and z. ■

Solutions to Section 5.4

Exercise 5.5.4: let $\alpha(\tau, s)$ be the homogeneous part of the affine mapping $\mathcal{A}$. A vector $w = (r, q_1, ..., q_k)$ is in $c + L^\perp$ if and only if

$$(w, \alpha(\tau, s)) = (c, \alpha(\tau, s))$$

identically in $(\tau, s) \in E$. This relation, in view of the construction of $\mathcal{A}$, can be rewritten as

$$r^T s + \sum_{j=1}^{k} [\lambda_j \tau + \mathrm{Tr}\{\sigma_j A(s)\}] = \tau$$

identically in (τ, s) with the zero sum of s_i, which immediately results in (5.20), (5.21).

To complete the derivation of the dual problem, we should realize what is (b, w) for $w \in c + L^\perp$. This is immediate:

$$(b, w) = [e^T r + \sum_{j=1}^{k} \mathrm{Tr}\{A(e)\sigma_j\}] + 2\sum_{j=1}^{k} z_j^T f_j,$$

and the quantity in the parentheses [ ] is nothing but $V\rho$ in view of (5.21). ■

Exercise 5.5.5: let us perform in (TTD$_d$) partial optimization over σ_j and r. Given a feasible plan of (TTD$_d$), we have in our standard notation:

$$\lambda_j \geq 0; \ \lambda_j = 0 \Rightarrow z_j = 0; \ \sigma_j \geq \lambda_j^{-1} z_j z_j^T$$

(these relations say exactly that the symmetric matrix $q_j = \begin{pmatrix} \lambda_j & z_j^T \\ z_j & \sigma_j \end{pmatrix}$ is positive semidefinite, cf. Exercise 5.5.3).

From these observations we immediately conclude that replacing in the feasible plan in question the matrices σ_j by the matrices $\sigma_j' = \lambda_j^{-1} z_j z_j^T$ for $\lambda_j > 0$ and zero matrices for $\lambda_j = 0$, we preserve positive semidefiniteness of the updated matrices q_j and ensure that $\sum_j b_i^T \sigma_j' b_i \leq \sum_j b_i^T \sigma_j b_i$; these latter quantities were equal to $\rho - r_i$ with nonnegative r_i, so that the former ones also can be represented as $\rho - r_i'$ with nonnegative r_i'. Thus, we may pass from a feasible plan of (TTD$_d$) to another feasible plan with the same value of the objective, and with σ_j being of the dyadic form $\lambda_j^{-1} z_j z_j^T$; the remaining simplifications are straightforward.

Exercise 5.5.6: as we know, K is self-dual, so that the formalism presented in Exercise 5.4.11 results in the following description of the problem dual to (π):

minimize $\beta^T \eta$ by choice of

$$\eta = (\zeta, \pi.) \in K$$

and real r subject to the constraint that the equality

$$(\eta, \mathcal{A}(\xi)) = \chi^T \xi + krp^T \xi \tag{13.16}$$

holds true identically in ξ; here

$$\beta = \mathcal{A}(p).$$

Indeed, the requirement that (13.16) is identity in ξ is exactly the same as the relation

$$A^T \eta = \chi + P^T r,$$

A being the matrix of the mapping $\mathcal{A}$ (in our case this mapping is linear homogeneous); we have taken into account that $P^T r = krp$, see the description of the data of (π).

Now, using in the straightforward manner the description of the data in (π) and denoting

$$\pi_{ij} = \begin{pmatrix} \alpha_{ij} & \beta_{ij} \\ \beta_{ij} & \gamma_{ij} \end{pmatrix},$$

we can rewrite identity (13.16) as the following identity with respect to f, λ, y_{ij} and z_j (in what follows i varies from 1 to m, j varies from 1 to k):

$$\sum_i \left\{ \zeta_i \left[f - \sum_j [2z_j^T f_j + V y_{ij}] \right] \right\} + \sum_{i,j} \{ y_{ij}\alpha_{ij} + 2\beta_{ij} b_i^T z_j + \lambda_j \gamma_{ij}] = f + r \sum_j \lambda_j,$$

which results in the following equations on η:

$$\sum_i \zeta_i = 1; \tag{13.17}$$

$$V\zeta_i = \alpha_{ij}; \tag{13.18}$$

$$(\sum_i \zeta_i) f_j = \sum_i \beta_{ij} b_i; \tag{13.19}$$

$$\sum_i \gamma_{ij} = r. \tag{13.20}$$

Now, the constraint $\eta \in K$ is equivalent to

$$\zeta_i \geq 0; \;\; \pi_{ij} \equiv \begin{pmatrix} \alpha_{ij} & \beta_{ij} \\ \beta_{ij} & \gamma_{ij} \end{pmatrix} \geq 0, \tag{13.21}$$

and the objective $\beta^T \eta \equiv (\mathcal{A}(p))^T \eta$ is nothing but

$$k^{-1} \sum_{ij} \gamma_{ij}.$$

Expressing via equations (13.17) - (13.20) all components of η via in terms of variables $\phi_i \equiv V\zeta_i$, β_{ij} and r and taking into account that the condition $\pi_{ij} \geq 0$ is equivalent to $\alpha_{ij} \geq 0$, $\gamma_{ij} \geq 0$, $\alpha_{ij}\gamma_{ij} \geq \beta_{ij}^2$, and eliminating in the resulting problem the variables γ_{ij} by partial optimization with respect to these variables, we immediately come to the desired formulation of the problem dual to (π). ■

Exercise 5.5.7: let $(\phi, \beta_\cdot)$ be a feasible solution to (ψ), and let I be the set of indices of nonzero ϕ_i. Then $\beta_{ij} = 0$ whenever $i \notin I$ - otherwise the objective of (ψ) at the solution would be infinite (this is our rule for interpreting fractions with zero denominators), and the solution is assumed to be feasible. Let us fix j and consider the following optimization problem:

$$(P_j): \quad \textit{minimize} \; \sum_{i \in I} v_i^2 \phi_i^{-1} \;\; \textit{s.t.} \;\; \sum_{i \in I} v_i b_i = f_j,$$

v_i being the control variables. The problem clearly is feasible: a feasible plan is given by $v_i = \beta_{ij}$, $i \in I$. Now, (P$_j$) is a quadratic problem with nonnegative objective and linear equality constraints; therefore it is solvable. Let β_{ij}^*, $i \in I$, be an optimal solution to the problem, and let $\beta_{ij}^* = 0$ for $i \notin I$. From the optimality conditions for (P$_j$) it follows that there is an n-dimensional vector $2x_j$ - the vector of Lagrange multipliers for the equality constraints - such that β_{ij}^*, $i \in I$, is an optimal solution to the unconstrained problem

$$\textit{minimize} \; \sum_{i \in I} v_i^2 \phi_i^{-1} + 2x_j^T (f_j - \sum_i v_i b_i),$$

so that for $i \in I$ one has

$$\beta_{ij}^* = \phi_i x_j^T \beta_i; \tag{13.22}$$

this relation, of course, is valid also for $i \notin I$ (where both sides are zero). Since $\beta_{i\cdot}^*$ is feasible for (P$_j$), we have $\sum_i \beta_{ij}^* b_i = f_j$, which in view of (13.22) implies that

$$f_j = (\sum_i \phi_i (b_i b_i^T)) x_j \equiv A(\phi) x_j. \tag{13.23}$$

This latter relation combined with (13.22) says that *the plan* $(\phi, \beta_\cdot^*)$ *is the image of the feasible plan* $(\phi, x_1, ..., x_k)$ *under the mapping (5.35).*

What are the compliances c_j associated with the plan $(\phi, x_1, ..., x_k)$? In view of (13.22) - (13.23) we have

$$c_j = x_j^T f_j = x_j^T \sum_i \beta^*_{ij} b_j = \sum_{i \in I} \beta^*_{ij} (x_j^T b_j) = \sum_{i \in I} [\beta^*_{ij}]^2 \phi_j^{-1};$$

and since β_{ij} form a feasible, and β^*_{ij} - an optimal plan to (P$_j$), we come to

$$c_j \leq \sum_i \beta^2_{ij} \phi_i^{-1}.$$

Thus, the value of the objective (i.e., $\max_j c_j$) of (TTD$_{\text{ini}}$) at the plan $(\phi, x_1, ..., x_k)$ does not exceed the value of the objective of (ψ) at the plan $(\phi, \beta.)$, and we are done. ■

Solutions to Section 6.7

Exercise 6.7.3: if the set $K^\sigma = \{y \in K \cap M \mid \sigma^T y = 1\}$ were bounded, the set $K(\sigma) = \{y \in K \cap M \mid \sigma^T y \le 1\}$ also would be bounded (since, as we know from (6.7), $\sigma^T y$ is positive on $M \cap \mathrm{int}\, K$). From this latter fact it would follow that σ is strictly positive on the cone $K' = K \cap M$ (see basic statements on convex cones in Lecture 5). The optimal solution x^* is a nonzero vector from the cone K' and we know that $\sigma^T x^* = 0$; this is the desired contradiction.

All remaining statements are immediate: ϕ is nondegenerate self-concordant barrier for K^σ (regarded as a domain in its affine hull) due to Proposition 5.3.1; $\mathrm{Dom}\,\phi$ is unbounded and therefore ϕ is below unbounded on its domain (**V.**, Lecture 3); since ϕ is below unbounded, its Newton decrement is ≥ 1 at any point (**VIII.**, Lecture 2) and therefore the damped Newton step decreases ϕ at least by $\rho(-1) = 1 - \ln 2$ (**V.**, Lecture 2). ■

Exercise 6.7.5: 1) is an immediate consequence of **III.**. To prove 2), note that $(S, \chi^*) = 0$ for certain positive semidefinite $\chi^* = I - \delta$ with $\delta \in \Pi$ (**IVb.**). Since $(S, I) = 1$ (**III.**), we have $(\delta, S) = 1$; since η is the orthoprojection of S onto Π and $\delta \in \Pi$, we have $(\delta, \eta) = (\delta, S)$, whence $(\delta, \eta) = 1$. Now, $(\eta, I) = 0$ (recall that $\eta \in \Pi$ and Π is contained in the subspace of matrices with zero trace, see **II.**). Thus, we come to $(I - \delta, \eta) \equiv (\chi^*, \eta) = -1$. Writing down the latter relation in the eigenbasis of η, we come to

$$\sum_{i=1}^n \chi_i g_i = -1,$$

χ_i being the diagonal entries of χ^* with respect to the basis; since $\chi_i \ge 0$ (recall that χ^* is positive semidefinite) and $\sum_i \chi_i^* = n$ (see **IVb.**), we conclude that $\max_i |g_i| \ge n^{-1}$. ■

Exercise 6.7.6: one clearly has $\tau \in T$, and, consequently, $\tau \in \mathrm{Dom}\,\phi$. We have

$$\phi(0) - \phi(\tau) = \sum_i \ln(1 - \tau g_i) - n \ln(1 - \tau |g|_2^2) \ge$$

[due to concavity of ln]

$$\ge \sum_i \ln(1 - \tau g_i) + n\tau |g|_2^2 = \sum_{j=1}^\infty \sum_i j^{-1}(-\tau g_i)^j + n\tau |g|_2^2 =$$

[since $\sum_i g_i = 0$, see Exercise 6.7.5, 1)]

$$= \sum_{j=2}^\infty \sum_i j^{-1}(-\tau g_i)^j + n\tau|g|_2^2 \ge$$

$$\ge -\sum_{j=2}^\infty j^{-1}[\tau|g|_2]^2[\tau|g|_\infty]^{j-2} + n\tau|g|_2^2 =$$

$$= -\frac{|g|_2^2}{|g|_\infty^2} \sum_{j=2}^\infty j^{-1}(\tau|g|_\infty)^j + n\tau|g|_2^2 =$$

$$= \frac{|g|_2^2}{|g|_\infty^2}\left[\ln(1 - \tau|g|_\infty) + \tau|g|_\infty\right] + n\tau|g|_2^2.$$

Substituting into the resulting lower bound for $\phi(0) - \phi(\tau)$ the value of τ indicated in the exercise, we come to the lower bound

$$\alpha \ge \frac{|g|_2^2}{|g|_\infty^2}\left[n|g|_\infty - \ln(1 + n|g|_\infty)\right];$$

it remains to use Exercise 6.7.5, 2). ■

Solutions to Section 7.6

Exercise 7.6.3: by construction, K is the direct product of $M + r$ copies of the cone $\mathbf{S}^{\nu}_{+}$ of positive semidefinite symmetric $\nu \times \nu$ matrices. The latter cone is self-dual (Exercise 5.4.7), and therefore K also is self-dual (Exercise 5.4.9). Now, $-\ln \mathrm{Det}\, y$ is a ν-vartheta logarithmically homogeneous self-concordant barrier for the cone $\mathbf{S}^{\nu}_{+}$ (Example 5.3.3, Lecture 5), and the Legendre transformation of this barrier is $-\ln \mathrm{Det}\,(-r) - \nu$ (Exercise 5.4.10). From Proposition 5.3.2.(iii) it follows that the direct sum of the above barriers for the direct factors of K, which is nothing but the barrier $F(x) = -\ln \mathrm{Det}\, x$, is $(M + 2)\nu$-logarithmically homogeneous self-concordant barrier for K. The Legendre transformation of direct sum clearly is direct sum of the Legendre transformations. ■

Solutions to Section 8.5

Exercise 8.5.4: by definition of $\zeta \equiv \zeta(v, dv)$ we have $v + rdv \in \operatorname{int} \mathbf{S}^k_+$ whenever $|r| < \zeta$, so that $f(r)$ is well-defined. Now, the function $f(r)$ clearly is analytic on its domain, and its Taylor expansion at $r = 0$ is

$$\sum_{i=0}^{\infty} \frac{f^{(i)}(0)}{i!} r^i = \sum_{i=0}^{\infty} \frac{D^i\Phi(v)[dv, ..., dv]}{i!} r^i =$$

[Exercise 8.5.3]

$$= f(0) + f'(0)r + \sum_{i=2}^{\infty} (-1)^i \frac{\mathrm{Tr}\{\widehat{h}^i\}}{i} r^i, \ \widehat{h} = v^{-1/2} dv v^{-1/2}.$$

In view of (8.35) the absolute values of the coefficients in the latter series are bounded from above by $i^{-1}|\widehat{h}|_2^2 |\widehat{h}|_\infty^{i-2}$, so that the series converges (and, consequently, represents f - recall that f is analytic on its domain) when $r < \zeta(v, dv) \equiv |h|_\infty^{-1}$ (see (8.33). It follows that the reminder for the aforementioned r is bounded from above by the series

$$\frac{|\widehat{h}|_2^2}{|\widehat{h}|_\infty^2} \sum_{i=j+1}^{\infty} i^{-1}(r|\widehat{h}|_\infty)^i,$$

and, taking into account that $|\widehat{h}|_2 = 1$ in view of (8.32), we come to (8.36). ■

Exercise 8.5.6: Since $x \in \operatorname{int} G$, we have $u \in \operatorname{int} \mathbf{S}^k_+$; further,

$$|du|_{\Phi''(u)} = |\pi dx|_{\Phi''(u)} = |dx|_{\pi^T \Phi''(u) \pi} = |dx|_{F''(x)} = 1,$$

so that (u, du) indeed is an arrow; by construction, (s, ds) is the conjugate to (u, du) co-arrow.

It remain to note that by definition of Δ and due to the normalization $|dx|_{F''(x)} = 1$ we have

$$\Delta = \max\{p \mid x \pm pdx \in G\} = \max\{p \mid u \pm pdu \in \mathbf{S}^k_+\} \equiv \zeta(u, du).$$

■

Exercise 8.5.7: by Lemma 8.3.1 and Proposition 8.3.1, the upper bound $v(r)$ for the residual $F_{t+dt}(x + dx(dt)) - \min_y F_{t+dt}(y)$ is bounded from above by the reminder $\rho^*(r)$ in the *third order* Taylor expansion of the function $\Phi(u + rdu(dt)) + \Phi^*(s + rds(dt))$; here dt is an arbitrary positive scale factor, and we are in our right to choose dt in a way which ensures that $|dx(dt)|_{F''(x)} = 1$; with this normalization, $\Omega = \Omega(x)$ will be exactly the quantity $\delta t/dt$, where δt is the stepsize given by the linesearch. The quantity Ω is therefore such that $v(\Omega) = O(1)$ (since we use linesearch to get the largest r which results in $v(r) \le \overline{\kappa}$); consequently, $\rho^*(\Omega) \ge O(1)$. On the other hand, in view of Exercises 8.5.5 and 8.5.6, $\rho^*(r)$ is exactly $\mathcal{R}^3_{(u,du)}(r)$; combining (8.37) and the inequality $\rho^*(\Omega) \ge O(1)$, we come to

$$\zeta^2(u, du)\rho_3(\Omega/\zeta(u, du)) \ge O(1),$$

and since $\zeta(u, du) = \Delta \equiv \Delta(x)$ by Exercise 8.5.6, we obtain

$$\rho_3(\Omega/\Delta) \ge O(1)\Delta^{-2}.$$

Since $\rho_3(z) \le O(1)z^4$, $|z| \le 1/2$, we conclude that

$$\Omega/\Delta \le 1/2 \Rightarrow \Omega \ge O(1)\sqrt{\Delta};$$

the resulting inequality for sure is true if $\Omega/\Delta > 1/2$, since, as we know, $\Delta \ge 1$. ■

Solutions to Section 9.6

Exercise 9.6.1:
1): the "general" part is an immediate consequence of the Substitution rule (N) as applied to the mapping

$$\mathcal{B} : (t, x) \mapsto \begin{pmatrix} x^T x \\ t \end{pmatrix} \qquad [G^- = \mathbf{R} \times \mathbf{R}^n]$$

which is 1-appropriate for G^+ in view of Proposition 9.3.1.

The "particular" part is given by the general one as applied to

$$G^+ = \{(u, s) \mid u \le s^{2/p},\ s \ge 0\}$$

and the 2-self-concordant barrier $F^+(u, s) = -\ln(s^{2/p} - u) - \ln s$ for G^+, see Example 9.2.1.

2): the "general" part is an immediate consequence of the Substitution rule (N) applied to the mapping

$$\mathcal{B} : (t, x) \mapsto \begin{pmatrix} \frac{x^T x}{t} \\ t \end{pmatrix} \qquad [G^- = \mathbf{R}_+ \times \mathbf{R}^n]$$

which is appropriate for G^+ in view of Proposition 9.3.1.

The "particular" part is given by the general one as applied to

$$G^+ = \{(u, s) \mid u \le s^{2/p-1},\ s \ge 0\},$$

the 2-self-concordant barrier

$$F^+(u, s) = -\ln(s^{2/p-1} - u) - \ln s$$

for G^+ (see Example 9.2.1) and the 1-self-concordant barrier

$$F^-(t, x) = -\ln t$$

for the domain G^- of the mapping $\mathcal{B}$.

Exercise 9.6.3: apply Proposition 9.3.1 to the data

- $G^+ = \mathbf{S}^m_+$, $F^+(\tau) = -\ln \mathrm{Det}\ \tau$;
- $Q[\xi', \xi''] = \frac{1}{2} \sum_{j=1}^q [(\xi_j')^T \xi_j'' + (\xi_j'')^T \xi_j']$,
  $\xi = (\xi_1, ..., \xi_q)$;
- $A(\eta)\xi = (y_1(\eta)\xi_1, ..., y_q(\eta)\xi_q)$,
  $F^-(\eta) = F_Y(\eta)$.

Exercise 9.6.4.2): specify the data in Exercise 9.6.3 as

- $q = k$, $n_1 = ... = n_k = m$;
- $Y = \mathbf{R}^k_+$, $y_j(\eta) = \eta_j I$, $j = 1, ..., k$;
- $F_Y(\eta) = -\sum_{j=1}^k \ln \eta_j$.

The resulting cone $\mathcal{K}$ clearly is comprised of collections $(\tau; \eta; \xi_j)$ (τ is $m \times m$ symmetric matrix, $\eta \in \mathbf{R}^k$, ξ_j are $m \times m$ matrices), for which

$$\eta \ge 0; \quad \tau - \sum_{j=1}^k \eta_j^{-1} \xi_j^T \xi_j \ge 0.$$

The cone G^+ is the inverse image of the "huge" cone $\mathcal{K}$ under the linear mapping

$$(s_i; t_{ij}; r_j) \mapsto \begin{pmatrix} \tau = \mathrm{Diag}\{s_1, ..., s_m\} \\ \xi_j = \mathrm{Diag}\{t_{1j}, ..., t_{mj}\} I \\ \eta_j = r_j \end{pmatrix},$$

and Φ is nothing but the superposition of the barrier F for $\mathcal{K}$ given by the result of Exercise 9.6.3 and this mapping. ∎

Exercise 9.6.5: let us compute derivatives of $\mathcal{A}$ at a point $u = (t, y) \in \text{int } G^-$ in a direction $du = (dt, dy)$ such that $u \pm du \in G^-$; what we should prove is that

$$D^2\mathcal{A}(u)[du, du] \leq 0 \tag{13.24}$$

and that

$$D^3\mathcal{A}(u)[du, du, du] \leq -3D^2\mathcal{A}(u)[du, du]. \tag{13.25}$$

Let us set $\eta_i = dy_i/y_i$, $\sigma_k = \sum_{i=1}^p \eta_i^k$, so that, in the clear notation, $d\sigma_k = -k\sigma_{k+1}$, and let $\phi(t, y) = (y_1...y_p)^{1/p}$. We have

$$DA(t,y)[du] = p^{-1}\sigma_1\phi(t,y) - dt;$$

$$D^2\mathcal{A}(t,y)[du,du] = p^{-2}\sigma_1^2\phi(t,y) - p^{-1}\sigma_2\mathcal{A}(t,y) = p^{-2}[\sigma_1^2 - p\sigma_2]\phi(t,y),$$

$$D^3\mathcal{A}(t,y)[du,du,du] = -p^{-2}[2\sigma_1\sigma_2 - 2p\sigma_3]\phi(t,y) + p^{-3}\sigma_1[\sigma_2^2 - p\sigma_2].$$

Now, let

$$\lambda = p^{-1}\sigma_1, \;\; \alpha_i = \eta_i - \lambda.$$

We clearly have

$$\sigma_1 = p\lambda;\; \sigma_2 = \sum_{i=1}^p \eta_i^2 = p\lambda^2 + \sum_{i=1}^p \alpha_i^2;\; \sigma_3 = \sum_{i=1}^p \eta_i^3 = p\lambda^3 + 3\lambda\sum_{i=1}^p \alpha_i^2 + \sum_{i=1}^p \alpha_i^3. \tag{13.26}$$

Substituting these expressions for σ_k in the expressions for the second and the third derivative of $\mathcal{A}$, we come to

$$d_2 \equiv -D^2\mathcal{A}(t,y)[du,du] = p^{-1}\phi(t,y)\sum_{i=1}^p \alpha_i^2 \geq 0, \tag{13.27}$$

as required in (13.24), and

$$d_3 \equiv D^3\mathcal{A}(t,u)[du,du,du] = -2p^{-2}\phi(u)[p^2\lambda^3 + p\lambda\sum_{i=1}^p \alpha_i^2 - p^2\lambda^3 - 3p\lambda\sum_{i=1}^p \alpha_i^2 - p\sum_{i=1}^p \alpha_i^3] -$$

$$-p^{-1}\phi(u)\lambda\sum_{i=1}^p \alpha_i^2 =$$

$$= \frac{3}{p}\phi(u)\lambda\sum_{i=1}^p \alpha_i^2 + \frac{2}{p}\phi(u)\sum_{i=1}^p \alpha_i^3 = \frac{3}{p}\phi(u)\sum_{i=1}^p [\lambda + \frac{2}{3}\alpha_i]\alpha_i^2 =$$

$$= \frac{3}{p}\phi(u)\sum_{i=1}^p [\frac{1}{3}\lambda + \frac{2}{3}\eta_i]\alpha_i^2. \tag{13.28}$$

Now, the inclusion $u \pm du \in G^-$ means exactly that $-1 \leq \eta_i \leq 1$, $i = 1, ..., p$, whence also $|\lambda| \leq 1$; therefore $|\frac{1}{3}\lambda + \frac{2}{3}\eta_i| \leq 1$, and comparing (13.28) and (13.27), we come to (13.25). ∎

Exercise 9.6.6: The mapping $\mathcal{B}(\cdot)$ is the superposition $\mathcal{A}(\mathcal{L}(\cdot))$ of the mapping

$$\mathcal{A}(t, y_1, ..., y_p) = (y_1...y_p)^{1/p} - t : H \to \mathbf{R}$$

with the domain

$$H = \{(t, y_1, ..., y_p) \mid y \geq 0\}$$

and the *linear mapping*

$$\mathcal{L}(\tau, \xi, \eta) = (\xi, \tau, \eta, ..., \eta) : \mathbf{R}^3 \to \mathbf{R}^{p+1};$$

namely, the set G^- is exactly $\mathcal{L}^{-1}(H)$, and on the interior of G^- we have $\mathcal{B}(\cdot) \equiv \mathcal{A}(\mathcal{L}(\cdot))$.

From Exercise 9.6.5 we know that $\mathcal{A}$ is 1-appropriate for $\mathbf{R}_+$; the fact that $\mathcal{B}$ laso is 1-appropriate for $\mathbf{R}^+$ is given by the following immediate observation:

Let $\mathcal{A} : \mathrm{int}\, H \to \mathbf{R}^N$ (H is a closed convex domain in $\mathbf{R}^K$) be β-appropriate for a closed convex domain $G^+ \subset \mathbf{R}^N$, let $\mathcal{L}$ be an affine mapping in certain $\mathbf{R}^M$, and let G^- be a closed convex domain in the latter space such that $\mathcal{L}(\mathrm{int}\, G^-) \subset \mathrm{int}\, H$. Then the composite mapping

$$\mathcal{B}(x) = \mathcal{A}(\mathcal{L}(x)) : \mathrm{int}\, G^- \to \mathbf{R}^N$$

is β-appropriate for G^+.

Thus, our particular $\mathcal{B}$ indeed is 1-appropriate with $\mathbf{R}_+$; the remaining claims of the Exercise are given by Theorem 9.1.1 applied with $F^+(z) = -\ln z$ and $F^-(\tau, \xi, \eta) = -\ln \tau - \ln \eta$. ■

Solutions to Section 10.5

Exercise 10.5.2: what we should prove is that $G_{\rm O}$ is convex and that the solutions to (Outer') are exactly the minimum volume ellipsoids which contain Q.

To prove convexity, assume that (r',x',X') and (r'',x'',X'') are two points of G', $\lambda \in [0,1]$ and $(r,x,X) = \lambda(r',x',X') + (1-\lambda)(r'',x'',X'')$; we should prove that $(r,x,X) \in G'$. Indeed, by the definition of G' we have for all $u \in Q$

$$u^T(X')u + 2(x')^T u + r' \le 0, \; u^T(X'')u + 2(x'')^T u + r'' \le 0,$$

whence, after taking weighted sum,

$$u^T X u + 2x^T u + r \le 0.$$

Thus, the points of Q indeed satisfy the quadratic inequality associated with (r,x,X); since X clearly is symmetric positive definite and Q possesses a nonempty interior, this quadratic inequality does define an ellipsoid, and, as we have seen, this ellipsoid $E(r,x,X)$ contains Q. It remains to prove that the triple (r,x,X) satisfies the normalizing condition $\delta(r,x,X) \le 1$; but this is an immediate consequence of convexity of the function $x^T X^{-1} x - r$ on the set (r,x,X) with $X \in \operatorname{int} \mathbf{S}^n_+$ (see the section on the fractional-quadratic mapping in Lecture 9).

It remains to prove that optimal solutions to (Outer') represent exactly minimum volume ellipsoids which cover Q. Indeed, let (r,x,X) be a feasible solution to (Outer') with finite value of the objective. I claim that $\delta(r,x,X) > 0$. Indeed, X is positive definite (since it is in $\mathbf{S}^n_+$ and F is finite at X), therefore the set $E(r,x,X)$ is empty, a point or an ellipsoid, depending on whether $\delta(r,x,X)$ is negative, zero or positive; since $(r,x,X) \in G_{\rm O}$, the set $E(r,x,X)$ contains Q, and is therefore neither empty nor a point (since $\operatorname{int} Q \ne \emptyset$), so that $\delta(r,x,X)$ must be positive. Thus, feasible solutions (r,x,X) to (Outer') with finite value of the objective are such that the sets $E(r,x,X)$ are ellipsoids containing Q; it is immediately seen that every ellipsoid with the latter property comes from certain feasible solution to (Outer'). Note that the objective in (Outer') is "almost" (monotone transformation of) the objective in (Outer):

$$\ln \operatorname{Vol}(E(r,x,X)) = \ln \kappa_n + \frac{n}{2} \ln \delta(r,x,X) - \frac{1}{2} \ln \operatorname{Det} X,$$

and the objective in (Outer') is $F(X) = -\ln \operatorname{Det} X$. We conclude that (Outer) is equivalent to the problem (Outer") which is obtained from (Outer') by replacing the *inequality* $\delta(r,x,X) \le 1$ with the *equation* $\delta(r,x,X) = 1$. But this is immediate: if (r,x,X) is a feasible solution to (Outer') with finite value of the objective, then, as we know, $\delta(r,x,X) > 0$; setting $\gamma = \delta^{-1}(r,x,X)$ and $(r',x',X') = \gamma(r,x,X)$, we come to $E(r,x,X) = E(r',x',X')$, $\delta(r',x',X') = 1$, so that $(r',x',X') \in G_{\rm O}$, and $F(X') = F(X) + n \ln \gamma \le F(X)$. From this latter observation it immediately follows that (Outer') is equivalent to (Outer"), and this latter problem, as we just have seen, is nothing but (Outer). ■

Exercise 10.5.4: to prove that $\mathcal{A}$ is $\frac{2}{3}$-appropriate for G^+, note that a direct computation says that for positive definite symmetric X and any (dt, dX) one has

$$d_2 \equiv D^2\mathcal{A}(t,X)[(dt,dX),(dt,dX)] = -\operatorname{Tr}\{X^{-1}dX X^{-1}dX\} = -\operatorname{Tr}\{[\delta X]^2\},$$
$$\delta X = X^{-1/2} dX X^{-1/2}$$

and

$$d_3 \equiv D^3\mathcal{A}(t,x)[(dt,dX),(dt,dX),(dt,dX)] =$$
$$= 2\operatorname{Tr}\{X^{-1}dX X^{-1}dX X^{-1}dX\} = 2\operatorname{Tr}\{[\delta X]^3\}.$$

Since the recessive cone K of G^+ is the nonnegative ray, the evident relation $d_2 \le 0$ says that $\mathcal{A}$ is concave with respect to K. Besides this, if $X \pm dX$ is positive semidefinite, then $-I \le \delta X \le I$, whence $\operatorname{Tr}\{[\delta X]^3\} \le \operatorname{Tr}\{[\delta X]^2\}$ (look what happens in the eigenbasis of δX), so that $d_3 \le -2d_2$. Thus, $\mathcal{A}$ indeed is $\frac{2}{3}$-appropriate for G^+. ■

Exercise 10.5.5: the feasible set in question is given by the following list of constraints:

$$a_j^T X a_j + 2x^T a_j + r \le 0, \; j = 1, ..., m$$

(corresponding 1-self-concordant barriers are $-\ln(-a_j^T X a_j - 2x^T a_j - r)$);

$$-\ln \operatorname{Det} X \le t$$

(corresponding $(n+1)$-self-concordant barrier is $-\ln(t+\ln\mathrm{Det}\ X)-\ln\mathrm{Det}\ X$, Exercise 10.5.4); and, finally,

$$\mathrm{cl}\{X\in\mathrm{int}\ \mathbf{S}^n_+,\ 1-r+x^TX^{-1}x\geq 0\}.$$

The set H defined by the latter constraint is the inverse image of $G^+=\mathbf{R}_+$ under the nonlinear mapping

$$(r,x,X)\mapsto 1+r-x^TX^{-1}x:\mathrm{int}\ G^-\to\mathbf{R},$$

$G^-=\{(r,x,X)\mid X\in\mathbf{S}^n_+\}$. Proposition 9.3.1 says that the function

$$-\ln(1+r-x^TX^{-1}x)-\ln\mathrm{Det}\ X$$

is $(n+1)$-self-concordant barrier for H.

To get a self-concordant barrier for G, it remains to take the sum of the indicated barriers.

Exercise 10.5.6: since Y is positive definite, any direction w' of the type $(u,0)$ is such that $(w')^TRw'>0$. Now, $E(r,x,X)$ is an ellipsoid, not a point or the empty set, and therefore there is a vector v such that

$$v^TXv+2x^Tv+r<0;$$

setting $w=(v,1)$, we get $w^TSw<0$. ■

Exercise 10.5.7: (b) is immediate: we know that $w^TSw<0$ for some w, so that L_w clearly is bounded. A nontrivial task is to prove (a). Thus, let us fix v and v' and prove that L_v and $L_{v'}$ have a point in common.

1^0. Consider the quadratic forms

$$S[p,q]=(pv+qv')^TS(pv+qv'),\ \ R[p,q]=(pv+qv')^TR(pv+qv')$$

on $\mathbf{R}^2$, and let

$$\mathcal{S}=\begin{pmatrix}a & d\\ d & b\end{pmatrix},\ \mathcal{R}=\begin{pmatrix}\alpha & \delta\\ \delta & \beta\end{pmatrix}$$

be the matrices of these forms. What we should prove is that there exists nonnegative λ such that

$$\alpha\leq\lambda a,\ \ \beta\leq\lambda b. \tag{13.29}$$

The following four cases are possible:

Case A: $a>0,b>0$. In this case (13.29) is valid for all large enough positive λ.

Case B: $a\leq 0,b\leq 0$. Since $a=v^TSv$ and $\alpha=v^TRv$, in the case of $a\leq 0$ we have also $\alpha\leq 0$ (this is given by (Impl)). Similarly, $b\leq 0\Rightarrow\beta\leq 0$. Thus, in the case in question $\alpha\leq 0$, $\beta\leq 0$, and (13.29) is satisfied by $\lambda=0$.

Case C: $a\leq 0$, $b>0$; *Case D:* $a>0$, $b\leq 0$. These are the only nontrivial cases which we should consider; due to the symmetry, we may restrict ourselves with the case C only. Thus, from now on $a\leq 0$, $b>0$.

2^0. Assume (case C.1) that $a<0$. Then the determinant $ab-d^2$ of the matrix $\mathcal{S}$ is negative, so that in appropriate coordinates p', q' on the plane the matrix $\mathcal{S}'$ of the quadratic form $S[\cdot]$ becomes $\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}$. Let $\begin{pmatrix}\xi & \zeta\\ \zeta & -\eta\end{pmatrix}$ be the matrix of the form $R[\cdot]$ in the coordinates p',q'. (Impl) says to us that for any 2-dimensional vector $z=(p',q')^T$ we have

$$z^T\mathcal{S}'z\equiv(p')^2-(q')^2\leq 0\Rightarrow z^T\mathcal{R}'z=\xi(p')^2+2\zeta p'q'-\eta(q')^2\leq 0. \tag{13.30}$$

The premise in this implication is satisfied by $z=(0,1)^T$, $z=(1,1)^T$ and $z=(1,-1)^T$, and the conclusion of it says to us that $\eta\geq 0$, $\xi-\eta\pm 2\zeta\leq 0$, whence

$$\eta\geq 0;\ \eta-\xi\geq 2|\zeta|. \tag{13.31}$$

2^0.1. Assume, first, that the quantity

$$\lambda=\frac{\eta+\xi}{2}$$

is nonnegative. Then the matrix

$$\lambda\mathcal{S}' - \mathcal{R}' = \begin{pmatrix} \frac{\eta-\xi}{2} & -\zeta \\ -\zeta & \frac{\eta-\xi}{2} \end{pmatrix}$$

is positive semidefinite (see (13.31)), and, consequently, the matrix

$$\lambda\mathcal{S} - \mathcal{R}$$

is positive semidefinite, so that λ satisfies (13.29).

2^0.2. Now assume that $\eta + \xi < 0$. In the case in question $-\xi = |\xi| > \eta \geq 0$ (the latter inequality is given by (13.31)). Let $\rho_\varepsilon = \sqrt{(\eta+\varepsilon)|\xi|^{-1}}$, where $\varepsilon > 0$ is so small that $0 \leq \rho_\varepsilon \leq 1$. The premise in (13.30) is satisfied by $z = (\rho_\varepsilon, \pm 1)^T$, so that from the conclusion of the implication it follows that

$$-|\xi|\rho_\varepsilon^2 - \eta \pm 2\zeta\rho_\varepsilon \leq 0,$$

or

$$|\zeta| \leq \frac{(2\eta+\varepsilon)\sqrt{|\xi|}}{2\sqrt{\eta+\varepsilon}}$$

for all small enough positive ε. Passing to limit as $\varepsilon \to 0$, we come to $|\zeta| \leq \sqrt{\eta|\xi|}$. Thus, in the case in question $\mathcal{R}' = \begin{pmatrix} -|\xi| & \zeta \\ \zeta & -|\eta| \end{pmatrix}$ is 2×2 matrix with nonpositive diagonal entries and nonnegative determinant; consequently, this matrix is negative semidefinite, so that $\mathcal{R}$ also is negative semidefinite, and (13.29) is satisfied by $\lambda = 0$.

3^0. It remains to consider the case C.2 when $b > 0$, $a = 0$. Here we have $a = v^T S v = 0$, so that $\alpha = v^T R v \leq 0$ by (Impl). Since $b > 0$, (13.29) is satisfied for all large enough positive λ. ■

Solutions to Section 11.4

Exercise 11.4.8: we should prove that

(i) if $A_m(t, X; \tau, U)$ is positive semidefinite, then $S_m(X) \le t$;

(ii) if $S_m(X) \le t$, then there exist τ and U such that $A_m(t, X; \tau, U)$ is positive semidefinite.

Let us start with (i). Due to construction of $A_m(\cdot)$, both matrices $\tau I + U - X$ and U are positive semidefinite; in particular, $X \le \tau I + U$, whence, due to monotonicity of $S_m(\cdot)$, $S_m(X) \le S_m(\tau I + U)$, The latter quantity clearly is $m\tau + S_m(U) \le m\tau + \operatorname{Tr} U$. Thus, $S_m(X) \le m\tau + \operatorname{Tr} U$, while $t \ge m\tau + \operatorname{Tr} U$, again due to the construction of $A_m(\cdot)$, Thus, $S_m(X) \le t$, as required.

To prove (ii), let us denote by $\lambda_1 \ge ... \ge \lambda_k$ the eigenvalues of X, and let U have the same eigenvectors as X and the eigenvalues

$$\lambda_1 - \lambda_m, \lambda_2 - \lambda_m ..., \lambda_{m-1} - \lambda_m, 0, ..., 0.$$

Set also $\tau = \lambda_m$. The U is positive senidefinite, while $\tau I + U - X$ is the matrix with the eigenvalues

$$0, 0, ..., 0, \lambda_m - \lambda_{m+1}, ..., \lambda_m - \lambda_k,$$

so that it also is positive semidefinite. At the same time $m\tau + \operatorname{Tr} U = S_m(X) \le t$, so that $A_m(t, X; \tau, U)$ is positive semidefinite. ∎

Exercise 11.4.9: let λ_i, $i = 1, ..., 2k$, be the eigenvalues of $Y(X)$, and $\sigma_1, ..., \sigma_k$ be the singular values of X. It is immediately seen that $Y^2(X) = \begin{pmatrix} XX^T & 0 \\ 0 & X^T X \end{pmatrix}$. We know that the sequence of eigenvalues of XX^T is the sasme of sequence of eigenvalues of $X^T X$, and the latter sequence is the same as the sequence of squared eigenvalues of X, by definition of the singular values. Since $Y^2(X)$ is block diagonal with diagonal blocks XX^T and $X^T X$, and both blocks have the same seqeunces of eigenvalues, to get the sequence of eigenvalues of $Y^2(X)$, you should twicen the multiplicity of each eigenvalue of $X^T X$. Thus, the sequence of eigenvalues of $Y^2(X)$ is

$$(I) \quad \sigma_1^2, \sigma_1^2, \sigma_2^2, \sigma_2^2, ..., \sigma_k^2, \sigma_k^2.$$

On the other hand, the sequence of eigenvalues of $Y^2(x)$ is comprised of (possibly, reordered) squared eigenvalues of $Y(x)$. Thus, the sequence

$$\lambda_1^2, \lambda_2^2, ..., \lambda_{2k}^2$$

differs from (I) only by order. To derive from this intermediate conclusion the statement in question, it suffices to prove that if certain $\lambda \neq 0$ is an eigenvalue of $Y(X)$ of certain multiplicity s, then $-\lambda$ also is an eigenvalue of the same multiplicity s. But this is simple. Let L be the eigenspace of $Y(X)$ associated with the eigenvalue λ. In other words, L is comprised of all vectors $\begin{pmatrix} u \\ v \end{pmatrix}$, $u, v \in \mathbf{R}^k$, for which

$$\begin{pmatrix} Xv \\ X^T u \end{pmatrix} = \lambda \begin{pmatrix} u \\ v \end{pmatrix}. \tag{13.32}$$

Now consider the space

$$L_- = \{\begin{pmatrix} Xv \\ -X^T u \end{pmatrix} \mid \begin{pmatrix} u \\ v \end{pmatrix} \in L\}.$$

It is immediately seen that L_- is comprised of eigenvectors of $Y(X)$ with the eigenvalue $-\lambda$:

$$\begin{pmatrix} 0 & X \\ X^T & 0 \end{pmatrix} \begin{pmatrix} Xv \\ -X^T u \end{pmatrix} = \begin{pmatrix} -XX^T u \\ X^T X v \end{pmatrix} = -\lambda \begin{pmatrix} Xv \\ -X^T u \end{pmatrix}.$$

If we could prove that the mapping $\begin{pmatrix} u \\ v \end{pmatrix} \mapsto \begin{pmatrix} Xv \\ -X^T u \end{pmatrix}$, restricted to L, has no kernel, we could conclude that dim $L_- \ge$ dim L, so that the multiplicity of the eigenvalue $-\lambda$ is at least that one of the eigenvalue λ; by swapping λ and $-\lambda$, we would conclude that the multiplicities of both the eigenvalues are equal, as required. Thus, it remains to verify that if $\begin{pmatrix} u \\ v \end{pmatrix} \in L$ and $Xv = 0, X^T u = 0$, then u and v are both zeros. But this is an immediate consequence of (13.32) and the assumption that $\lambda \neq 0$. ∎

Appendix I: Surface-Following Methods

Multi-Parameter Surfaces of Analytic Centers and Long-step Surface-Following Interior Point Methods

Yu. Nesterov[1] and A. Nemirovski[2]

1

Abstract

We develop a long-step polynomial time version of the Method of Analytic Centers for nonlinear convex problems. The method traces a multi-parameter surface of analytic centers rather than the usual path, which allows to handle cases with non-centered and possibly infeasible starting point.

The paper is an extended version of [5].

14.1 Introduction

Consider a convex program in the following standard form:

$$\textit{minimize}\ \ c^T x\ \ \textit{s.t.}\quad x \in G; \tag{14.33}$$

here G is a closed and bounded convex subset of $\mathbf{R}^n$ with a nonempty interior. One of the most attractive theoretically ways to solve the problem is to trace the path of *analytic centers*, i.e., the minimizers over $x \in \operatorname{int} G$ of the penalized family of functions

$$F^t(x) = F(x) + tc^T x; \tag{14.34}$$

here F is a barrier (interior penalty function) for G. Under the above parameterization of the path, in order to converge to the optimal set one should trace the path as $t \to \infty$. The path of analytic centers, however, can be parameterized in another way, say, as the path of minimizers of the family

$$F_t(x) = F(x) - \vartheta \ln(t - c^T x); \tag{14.35}$$

($\vartheta > 0$ is fixed); here in order to get close to the optimal set one should approach the parameter t to the optimal value of the problem. As it is well-known, both the parameterizations, under appropriate choice of F, imply polynomial-time interior-point methods for (14.33). If G is a polytope, then it is reasonable to choose as F the standard logarithmic barrier for G; polynomiality of the associated path-following methods for Linear Programming was first established in the seminal papers of Renegar ([8], parameterization (14.35)) and Gonzaga ([1], parameterization (14.34)). For the nonpolyhedral case polynomial

[1] CORE, Louvain-la-Neuve, Belgium; e-mail nesterov@core.ucl.ac.be

. [2] Faculty of Industrial Engineering & Management at Technion, Haifa, Israel; e-mail nemirovs@ie.technion.ac.il

. Research was partly supported by the Vice-President for Research Fund for the promotion of research at Technion, grant # 191-142

time results for both the parameterizations can be obtained if F is a *self-concordant barrier* for G (see below), as it is the case with the standard logarithmic barrier for a polytope.

Now, in order to trace the path of analytic centers F one should once get close to the path; this is the aim of a special *preliminary phase* of a path-following method, which, theoretically, is of the same complexity as following the path itself. Moreover, to initialize the preliminary phase one should know in advance a strictly feasible solution $\widehat{x} \in \operatorname{int} G$. To get such a point, it, generally speaking, again requires an additional phase of the method; at this phase we, basically, solve an auxiliary problem of the same type as (14.33), but with a known in advance strictly feasible solution. There are numerous strategies of combining all these phases; one of the main goals of this paper is to present a kind of a general framework, based on the notion of a "multi-parameter surface of analytic centers", for these combined strategies. The notion is introduced in Section 14.2, along with motivating the advantages of tracing surface as compared to tracing the usual single-parameter path. Section 14.2 contains also a generic predictor-corrector scheme of tracing a surface of analytic centers. When tracing a surface, one should decide, first, *where* to move - what should be the strategy of choosing the subsequent search directions - and, second, *how* to move - what should be the tactics of choosing the stepsize in the chosen direction in order to move as fast as possible. The "tactics" issues are discussed in Sections 14.3 and 14.4. In Section 14.3 we develop, under some reasonable assumptions on the structure of the underlying barriers, a duality-based technique which, roughly speaking, allows to adjust the stepsizes to the "local curvature" of the surface and thus results, under favourable circumstances, in "long steps". Main theoretical properties of the resulting scheme are presented in Section 14.4. In particular, we demonstrate that under reasonable assumptions on the underlying barriers "long steps" indeed are long – they form a significant fraction of the way to the boundary of the feasible set. Note that our "long steps" technique is closely related to the one recently developed in [4] for path-following methods as applied to primal-dual conic formulations of convex problems, and the results on the length of the steps are similar to those of [6, 7]. The advantage of our approach as compared to [4, 6, 7] is not only in the fact that now we are able to trace surfaces rather than paths, but also that now we need neither explicit conic reformulation of the initial problem, nor the self-scaled property of the associated cone; this allows to avoid necessity to increase the number of variables and enables to work with problems (e.g., the Geometric Programming ones) which cannot be covered by methods of [6, 7].

In Section 14.5 we present a strategy of tracing the surface for both the cases of feasible and infeasible start; the strategy in question fits the standard polynomial time complexity bounds and seems to be computationally reasonable.

As it was already mentioned, the "long step" technique presented in the paper requires certain assumption on the structure of the barriers in question; Section 14.6 presents a number of barriers satisfying this assumption and thus allows to understand what might be the applications of the developed technique.

14.2 Surfaces of analytic centers: preliminaries

We start with specifying the basic for what follows notions of a self-concordant function/barrier ([3], Chapter 2; what is called below a self-concordant function, is a "strongly self-concordant function" in the terminology of [3]).

14.2.1 Self-concordant functions and barriers

Definition 14.2.1 Let Q be an open nonempty convex domain in certain $\mathbf{R}^N$. A function $\Psi : Q \to \mathbf{R}$ is called *self-concordant* (s.-c. for short), if Ψ is a C^3 smooth convex function on Q which tends to ∞ along every sequence of points from Q converging to a boundary point of Q and satisfies the differential inequality

$$\left|D^3\Psi(u)[h,h,h]\right| \leq 2\left(D^2\Psi(u)[h,h]\right)^{3/2}, \ u \in Q, h \in \mathbf{R}^N; \tag{14.36}$$

from now on $D^sF(u)[h_1, ..., h_s]$ denotes s-th differential of a function F taken at a point u along the set of directions $h_1, ..., h_s$.

If, in addition,

$$|D\Psi(u)[h]|^2 \leq \vartheta D^2\Psi(u)[h,h], \ u \in Q, h \in \mathbf{R}^N, \tag{14.37}$$

for some $\vartheta \geq 1$, we say that Ψ is a *ϑ-self-concordant barrier* (ϑ-s.-c.b. for short) for the closed convex domain $G = \mathrm{cl}\, Q$.

Let $\alpha \geq 1$. We say that a s.-c. function Ψ is *α-regular* on its domain Q, if Ψ is C^4 function such that

$$\left|D^4\Psi(u)[h,h,h,h]\right| \leq \alpha(\alpha+1)D^2\Psi(u)[h,h] \parallel h \parallel_{Q,u}^2, \;\; u \in Q, h \in \mathbf{R}^N, \tag{14.38}$$

where

$$\parallel h \parallel_{Q,u} = \inf\{t^{-1} \mid t > 0, u \pm th \in Q\}$$

is the (semi)norm on $\mathbf{R}^N$ with the unit ball being the closure of the symmeterization $Q \cap (2u - Q)$ of Q with respect to u.

E.g., the standard logarithmic barrier $-\sum_{i=1}^m \ln(b_i - a_i^T u)$ for a nonempty polytope $G = \mathrm{cl}\{u \mid a_i^T u < b_i,\, i = 1, ..., m\}$ is both m-s.-c.b. for G and 2-regular s.-c. function on int G.

The important for us properties of self-concordant functions/barriers are as follows (for proofs, see [3] , Chapter 2):

Proposition 14.2.1 [Combination rules]

(i) [summation] *Let Q_i, $i = 1, ..., k$, be open convex domains in $\mathbf{R}^N$ with a nonempty intersection Q and let Ψ_i be s.-c. functions on Q_i. Then the function $\Psi(x) = \sum_i \Psi_i(x)$ is s.-c. on Q. If all Ψ_i are ϑ_i-s.-c.b.'s for $\mathrm{cl}\, Q_i$, then Ψ is a $(\sum_i \vartheta_i)$-s.-c.b. for $\mathrm{cl}\, Q$, and if all Ψ_i are α-regular, then so is Ψ.*

(ii) [direct summation] *Let $Q_i \subset \mathbf{R}^{N_i}$ be open convex domains, $i = 1, ..., k$, and let Ψ_i be s.-c. on Q_i. Then the function*

$$\Psi(u_1 ..., u_k) = \sum_i \Psi_i(u_i) : Q \equiv \prod_i Q_i \to \mathbf{R}$$

is s.-c. on Q. If all Ψ_i are ϑ_i-s.-c.b.'s for $\mathrm{cl}\, Q_i$, then Ψ is a $(\sum_i \vartheta_i)$-s.-c.b. for $\mathrm{cl}\, Q$, and if all Ψ_i are α-regular, then so is Ψ.

(iii) [affine substitutions of argument] *Let Q^+ be an open convex set in $\mathbf{R}^M$ and $\mathcal{A}$ be an affine mapping from $\mathbf{R}^N$ into $\mathbf{R}^M$ with the image intersecting Q^+. Let Ψ^+ be s.-c. on Q^+; then $\Psi(\cdot) \equiv \Psi^+(\mathcal{A}(\cdot))$ is s.-c. on $Q = \mathcal{A}^{-1}(Q^+)$. If Ψ^+ is a ϑ-s.-c.b. for $\mathrm{cl}\, Q^+$, then Ψ is a ϑ-s.-c.b. for $\mathrm{cl}\, Q$, and if Ψ^+ is α-regular, then so is Ψ.*

From now on, for a positive semidefinite symmetric matrix A and a vector h of the corresponding dimension,

$$|h|_A = (h^T A h)^{1/2}.$$

Proposition 14.2.2 *Let Q be a nonempty open convex domain in $\mathbf{R}^N$, $G = \mathrm{cl}\, Q$ and Ψ be a s.-c. function on Q. Then*

(i) [behaviour in Dikin's ellipsoid] *For any $u \in Q$ the Dikin ellipsoid*

$$W_\Psi(u) = \{v \mid |v - u|_{\Psi''(u)} < 1\}$$

is contained in Q, and and

$$\Psi(u + h) \leq \Psi(u) + h^T \Psi'(u) + \rho(|h|_{\Psi''(u)}); \tag{14.39}$$

from now on,

$$\rho(r) = -\ln(1 - r) - r. \tag{14.40}$$

(ii) [nondegeneracy] *If $\Psi''(u)$ is nondegenerate at certain $u \in Q$, then Ψ'' is nondegenerate everywhere on Q; this for sure is the case when Q does not contain lines. If Q is bounded, then Ψ attains its minimum over Q, and the minimizer is unique.*

(iii) [stability with respect to Legendre transformation] *Let $\Psi''(u)$ be nondegenerate for some $u \in Q$ (and then, by (ii), for any $u \in Q$). Consider the Legendre transformation*

$$\Psi_*(v) = \sup\{u^T v - \Psi(u) \mid u \in Q\}$$

regarded as a function on the domain Q^* comprised of those v for which the right hand side is finite. Then Q^* is an open nonempty convex domain, the mapping $x \mapsto \Psi'(x)$ is a one-to-one correspondence between Q and Q^* and Ψ_* is s.-c. and with nondegenerate Hessian on Q^*; the Legendre transformation of Ψ_* is exactly Ψ.

14.2.2 Surface of analytic centers

Let F be a ϑ-s.-c.b. for a closed and *bounded* convex domain $G \subset \mathbf{R}^n$ with a nonempty interior. Aside from the parameterization issues, the path of analytic centers associated with the barrier can be defined as the set of points $x \in \text{int } G$ where $-\nabla F(x) = \lambda c$ for some positive λ. A natural "multi-parameter" extension of this description is as follows: let us fix k nonzero vectors $c_1, ..., c_k$ and associate with this collection the "surface" $\mathcal{S}_k = \mathcal{S}_k(c_1, ..., c_k)$ comprised of all points $x \in \text{int } G$ where

$$-\nabla F(x) = \sum_{i=1}^{k} \lambda_i c_i$$

with certain positive λ_i. A convenient way to parameterize the surface is to introduce the k-dimensional "parameter"

$$t = (t_1, ..., t_k)^T$$

and associate with this parameter the barrier (cf. (14.35))

$$F_t(x) = \vartheta \sum_{i=1}^{k} \psi_i(t_i - c_i^T x) + F(x), \quad \psi_i(r) = -\ln r,$$

for the convex set

$$G_t = \{x \in G \mid c_i^T x \leq t_i, \, i = 1, ..., k\}.$$

In what follows we are interested only in those values of t for which int $G_t \neq \emptyset$; the corresponding set $T = T_k(e_1, ..., e_k)$ of values of t clearly is a nonempty open convex and monotone ($t' \geq t \in T \Rightarrow t' \in T$) subset in $\mathbf{R}^k$. Now, since F is a ϑ-s.-c.b. for G, the functions F_t, $t \in T$, are ϑ_*-s.-c.b.'s for the domains G_t with

$$\vartheta_* = (k+1)\vartheta \tag{14.41}$$

(Proposition 14.2.1; note that the function $-\vartheta \ln(s)$ for $\vartheta \geq 1$ clearly is a ϑ-s.-c.b. for $\mathbf{R}_+$). Since G_t is bounded, F_t attains its minimum over int G_t, and the corresponding minimizer $x_k^*(t)$ - the *analytic center of* G_t - is unique (Proposition 14.2.2.(ii)). At this minimizer, of course, $-\nabla F(x) = \vartheta \sum_{i=1}^{k} (t_i - c_i^T x)^{-1} c_i$, so that $x_k^*(t) \in \mathcal{S}_k$. Vice versa, every point of $\mathcal{S}_k$ is $x_k^*(t)$ for certain $t \in T$: immediate computation demonstrates that

$$\{\lambda_i > 0, \, i = 1, ..., k\} \& \{-\nabla F(x) = \sum_{i=1}^{k} \lambda_i c_i\} \Rightarrow x = x_k^*(\vartheta\lambda_1^{-1} + c_1^T x, ..., \vartheta\lambda_k^{-1} + c_k^T x).$$

Thus, we do have defined certain parameterization of $\mathcal{S}_k$.

The following property of the surfaces of analytic centers is immediate:

Lemma 14.2.1 *Let $\widehat{t} \in T_k(e_1, ..., e_k)$ and let $t^j \geq \widehat{t}$, $j = 1, 2, ...$ be such that $t_i^j = \widehat{t}_i$ for $i \in I \subset \{1, ..., k\}$ and $t_i^j \to \infty, j \to \infty$, for $i \notin I$. Then the points $x_k^*(t^j)$ converge as $j \to \infty$ to the point $x_l^*(\{\widehat{t}_i\}_{i \in I})$ of the surface $\mathcal{S}_l(\{c_i\}_{i \in I})$, $l = \text{card } I$ (from now on $\mathcal{S}_0$ is the 0-dimensional surface comprised of the (unique) minimizer x_F^* of F over* int G*). Thus, the surfaces of dimensions $< k$ obtained from $\mathcal{S}_k(c_1, ..., c_k)$ by eliminating some of the vectors c_i are contained in the closure of $\mathcal{S}_k$. In particular, the closures of all surfaces of analytic centers have a point in common, namely, x_F^*.*

14.2.3 Tracing surfaces of analytic centers: motivation

To solve problem (14.33), we may take an arbitrary set $c_2, ..., c_k$ of vectors, set $c_1 = c$ and trace the surface $\mathcal{S}_k(c_1, ..., c_k)$ along certain sequence $\{t^i\}$ of values of the parameter, i.e., to produce approximations $x^i \in \text{int } G_{t^i}$ of the analytic centers $x^*(t^i)$. If the sequence $\{t^i\}$ is such that $t_1^i \to c^*$ as $i \to \infty$, c^* being the optimal value in (14.33), then x^i clearly form a sequence of feasible solutions to (14.33) converging to the optimal set. In what follows we demonstrate that there are basically the same possibilities to trace the surface $\mathcal{S}_k$ as in the standard case when $\mathcal{S}_k$ is a single-parameter path; with this in mind, let us explain what are the advantages of tracing a multi-parameter surface $\mathcal{S}_k$ rather than the usual path $\mathcal{S}_1(c)$.

1. Difficulty of initialization. As it was already mentioned, in the usual path-following method we trace the path $\mathcal{S}_1(c)$; to start the process, we should, anyhow, come close to the path. Now assume that we are given an initial strictly feasible solution $\widehat{x}$. In the standard path-following scheme, to get close to $\mathcal{S}_1(c)$ we trace the auxiliary path $\mathcal{S}_1(d)$, $d = -\nabla F(\widehat{x})$ which clearly passes through $\widehat{x}$. According to Lemma 14.2.1, both the paths $\mathcal{S}_1(c)$ and $\mathcal{S}_1(d)$ approach, as the parameter tends to ∞, the minimizer x^*_F of F over G; therefore, tracing the auxiliary path as $t \to \infty$, we in the mean time come close to the path $\mathcal{S}_1(c)$ and then may switch to tracing this latter path. On the other hand, given $\widehat{x}$ and arbitrary $\widehat{t}_1 > c^T\widehat{x}$, we can easily find vector c_2 and real $\widehat{t}_2$ such that the 2-dimensional surface of analytic centers $\mathcal{S}_2(c_1 \equiv c, c_2)$ would pass, as $t = (\widehat{t}_1, \widehat{t}_2)$, through $\widehat{x}$; it suffices to set

$$c_2 = \vartheta^{-1}d - (\widehat{t}_1 - c^T\widehat{x})^{-1}c, \quad \widehat{t}_2 = c_2^T\widehat{x} + 1.$$

Now we have a two-dimensional surface of analytic centers which "links" $\widehat{x}$ with the optimal set, and we may use various policies of tracing the surface, starting at $\widehat{x}$, in order to approach the optimal set. Note that our "main path" $\mathcal{S}_1(c)$, due to Lemma 14.2.1, lies in the closure of $\mathcal{S}_2$, while the "auxiliary path" $\mathcal{S}_1(d)$, as it is immediately seen, simply belongs to the surface. Thus, the standard path-following scheme - first trace $\mathcal{S}_1(d)$ and then $\mathcal{S}_1(c)$ - is nothing but a specific way to trace the two-dimensional surface of analytic centers $\mathcal{S}_2(c, c_2)$. After this is realized, it becomes clear that there is no necessity to restrict ourselves with the above specific route; why not to move in a more "direct" manner, thus avoiding the preliminary phase where we do not take care of the objective at all?

2. Infeasible start. Now assume that we do not know in advance an initial strictly feasible solution to the problem. What should we do? Note that normally the situation, under appropriate renaming of the data, is as follows. We need to solve the problem

$$(P'): \quad \textit{minimize } \; c^T x \;\; s.t. \;\; x \in G, \, x_n = 0,$$

where G is a solid in $\mathbf{R}^n \cap \{x_n \geq 0\}$ with a known in advance interior point $\widehat{x}$. In other words, normally we can represent the actual feasible set as a kind of a "facet" in a higher-dimensional convex solid with known in advance interior point. To support this claim, consider a standard form convex program

$$\text{(CP)} \quad f_0(u) \to \min \mid \; f_i(u) \leq 0, \, i = 1, ..., m \quad [u \in \mathbf{R}^q],$$

(f_i, $0 \leq i \leq m$, are convex lower semicontinuous functions) and assume that we know in advance

- a point u_0 such that all f_i are finite in a neighbourhood of u_0,
- an upper bound $R > |u_0|$ on the Euclidean norm of the optimal solution to (CP),
- an upper bound $V > f_0(u_0)$ on the optimal value of the problem.

Then we can equivalently rewrite (CP) in the form of (P') with the design vector $x = (u, v, w) \in \mathbf{R}^{q+2}$, the objective $c^T x \equiv v$ and

$$G = \{(u, v, w) \mid \; f_0(u) \leq v \leq V; \, u^T u \leq R^2; f_i(u) \leq w, \, i = 1, ..., m; \, 0 \leq w \leq W\},$$

where W is an arbitrary constant which is greater than $\widehat{f}(u_0) \equiv \max\{0, f_1(u_0), ..., f_m(u_0)\}$. Note that G indeed is a solid in $\mathbf{R}^{q+2}$ and that there is no difficulty to point out an interior point x_0 in G: one can set $x_0 = (u_0, v_0, w_0)$ with arbitrarily chosen $v_0 \in (f_0(u_0), V)$ and $w_0 \in (\widehat{f}(u_0), W)$.

Now assume that we are given a ϑ-s.-c.b. F for G [2]. In this situation the standard "big M" approach to (P') is to apply an interior point method to the problem

$$(P): \quad \textit{minimize } \; (c + Mf)^T x \;\; s.t. \;\; x \in G \quad \left[f^T x \equiv x_n\right],$$

[2] Note that in the case of problem (P') coming from (CP) such a barrier is readily given by ϑ_i-s.-c.b.'s $F_i(v, u)$ for the epigraphs $\{(u, v) \mid \; v \geq f_i(u)\}$ of the functions f_i, $i = 0, ..., m$:

$$F(u, v, w) = F_0(u, v) + \sum_{i=1}^m F_i(u, w) - \ln(R^2 - u^T u) - \ln(V - v) - \ln(W - w) - \ln w \qquad \left[\vartheta = 4 + \sum_{i=0}^m \vartheta_i\right].$$

where M is a "large enough" constant. Here we meet with unpleasant question how big should be the "big M". Now note that the path $\mathcal{S}_1(c+Mf)$ which is traced in the "big M" scheme clearly belongs to the two-dimensional surface $\mathcal{S}_2(c,f)$, which is independent of the particular value of M we choose. Thus, the "big M" approach is nothing but a specific way of tracing certain 2-dimensional surface of analytic centers. After this is realized, we may ask ourselves why should we trace the surface in this particular manner rather than to use more flexible strategies.

Note that for the Linear Programming case (G is a polytope, F is the standard logarithmic barrier for G) the surface $\mathcal{S}_2(c,f)$ was introduced and studied in details, although from a slightly different viewpoint, in [2].

Now, to trace $\mathcal{S}_2(c,f)$, we should first get close to the surface. Here again the traditional way would be to note that all surfaces $\mathcal{S}_k(c_1,...,c_k)$ in G come close to each other, so that tracing the path $\mathcal{S}_1(-\nabla F(\widehat{x}))$ (which passes through the given point $\widehat{x}$) and pushing the parameter to ∞, we come close to x_F^* and, consequently, to $\mathcal{S}_2(c,f)$ and then can switch to tracing the surface $\mathcal{S}_2(c,f)$. But this is nothing but a particular way to trace the 3-parameter surface $\mathcal{S}_3(c,f,d)$ in G given by

$$d = -\vartheta^{-1}\nabla F(\widehat{x}) - (\widehat{t}_1 - c^T\widehat{x})^{-1}c - (\widehat{t}_2 - \widehat{x}_n)^{-1}f$$

($\widehat{t}_1 > c^T\widehat{x}, \widehat{t}_2 > \widehat{x}_n$ are arbitrary). The surface $\mathcal{S}_3$ clearly passes through $\widehat{x}$ and links $\widehat{x}$ with the optimal set of the initial problem. After this is realized, why should we restrict ourselves with certain particular route?

14.2.4 The "surface-following" scheme

We believe that the aforementioned discussion demonstrates that it makes sense to trace not only *paths* of analytic centers, but also *multi-parameter surfaces* of these centers, at least 2- and 3-parameter ones. The point is, of course, how to trace such a surface; this is the issue we address in this section.

14.2.5 Surface of analytic centers: general definition

To make the presentation more compact it is convenient to get rid of particular structure of the surfaces introduced so far and speak about general situation as follows. Assume that G^+ is a closed convex domain with a nonempty interior in certain $\mathbf{R}^N$ and Ψ is a ϑ_*-s.-c.b. for G^+ with nondegenerate Ψ''. Let, further, π and σ be $N\times n$ and $N\times k$ matrices, respectively, and let $\epsilon \in \mathbf{R}^N$. Consider the affine mapping

$$(t,x) \mapsto \mathcal{A}(t,x) = \sigma t + \pi x + \epsilon : \mathbf{R}^k \times \mathbf{R}^n \to \mathbf{R}^N,$$

and assume that

(A) the image of the mapping $\mathcal{A}$ intersects the interior of G^+.

(A) implies that the set

$$Q = \{(t,x) \mid \mathcal{A}(t,x) \in \operatorname{int} G^+\}$$

is a nonempty open convex subset of $\mathbf{R}^k \times \mathbf{R}^n$. Let T be the projection of Q onto the "parameter space" $\mathbf{R}^k$; for $t \in T$ let

$$Q_t = \{x \in \mathbf{R}^n \mid (t,x) \in Q\},\ t \in T; \quad G_t = \operatorname{cl} Q_t,$$

so that Q_t is a nonempty open convex set in $\mathbf{R}^n$.

Our second, and for the time being the last, assumption is

(B) for some (and, consequently, for all) $t \in T$ the set Q_t is bounded.

Since Ψ is a ϑ_*-s.-c.b. for G^+, (A) implies that the function

$$F(t,x) = \Psi(\mathcal{A}(t,x))$$

is a ϑ_*-s.-c.b. for $\operatorname{cl} Q$, and for $t \in T$ the function

$$F_t(x) \equiv F(t,x)$$

is a ϑ_*-s.-c.b. for G_t (Proposition 14.2.1). Since G_t is bounded, the Hessian of $F_t(\cdot)$ is nondegenerate at any point from $Q_t = \operatorname{int} G_t$, and $F_t(\cdot)$ attains its minimum over Q_t at exactly one point $x^*(t)$ (Proposition 14.2.2.(ii)). From now on we call the set

$$\mathcal{S} = \{(t,x) \in Q \mid x = x^*(t)\}$$

the *surface of analytic centers* associated with the data G^+, Ψ, $\mathcal{A}(\cdot)$. Note that the surface of analytic centers $\mathcal{S}_k(c_1, ..., c_k)$ is obtained from the general definition by setting

$$G^+ = \mathbf{R}^k_+ \times G,\ \Psi(u_1, ..., u_k, x) = -\vartheta \sum_{i=1}^k \ln u_i + F(x),\ \mathcal{A}(t,x) = \begin{pmatrix} t_1 - c_1^T x \\ \cdots \\ t_k - c_k^T x \\ x \end{pmatrix},\ \vartheta_* = (k+1)\vartheta, \quad (14.42)$$

ϑ being the parameter of the s.-c.b. F.

14.2.6 Tracing a surface of analytic centers: basic scheme

Our general scheme of tracing the surface $\mathcal{S}$ associated with the data $G^+, \Psi, \mathcal{A}$ is as follows. First, we fix the "tolerances"

$$\kappa \in (0, 0.125], \quad \overline{\kappa} > 2\rho(\kappa) - \kappa^2.$$

We say that a pair (t,x) is *κ-close* to $\mathcal{S}$ if it satisfies the predicate

$\mathcal{P}_\kappa(t,x)$: $\qquad (t,x) \in Q$ *and* $\lambda(t,x) \leq \kappa$,

where

$$\lambda(t,x) = \left((\nabla_x F(t,x))^T [\nabla_x^2 F(t,x)]^{-1} \nabla_x F(t,x)\right)^{1/2}$$

is the *Newton decrement* of the s.-c.b. $F_t(\cdot)$ at x; the quantity is well-defined, since, as it was already mentioned, boundedness of G_t implies nondegeneracy of $\nabla^2 F_t(x)$ at any $x \in Q_t$.

We say that a pair (t,x) is *$\overline{\kappa}$-good* with respect to $\mathcal{S}$, if it satisfies the predicate

$\mathcal{R}_{\overline{\kappa}}$: $\qquad \{(t,x) \in Q\} \,\&\, \{V(t,x) \equiv F(t,x) - \min_{u \in Q_t} F(t,u) \leq \overline{\kappa}\}$.

When tracing $\mathcal{S}$, at each step we are given a κ-close to $\mathcal{S}$ pair (t,x) and transform it into a new pair (t^+, x^+) with the same property according to the following

Basic Updating Scheme:

1. Choose a *search direction* $\delta t \in \mathbf{R}^k$, "lift" it to the direction

$$(\delta t, \delta x) \in \Pi(t,x) \equiv \{(dt, dx) \in \mathbf{R}^k \times \mathbf{R}^n \mid \frac{\partial}{\partial t} \nabla_x F(t,x) dt + \nabla_x^2 F(t,x) dx = 0\} =$$

$$= \{(dx, dt) \mid dx = -[\nabla_x^2 F(t,x)]^{-1} \frac{\partial^2}{\partial t \partial x} F(t,x) dt\} =$$

$$= \{(dt, dx) \mid \pi^T \Psi''(\sigma t + \pi x + \epsilon)[\sigma dt + \pi dx] = 0\} \quad (14.43)$$

and define the *primal search line*

$$R = \{X(p) = (t + p\delta t, x - d_x(t,x) + p\delta x) \mid p \in \mathbf{R}\},\ d_x(t,x) = [\nabla_x^2 F(t,x)]^{-1} \nabla_x F(t,x). \quad (14.44)$$

2. [predictor step] Choose a stepsize $r > 0$ along the primal search line and form the forecast $(t^+, \widetilde{x}) \equiv X(r)$ which should be $\overline{\kappa}$-good with respect to $\mathcal{S}$ (this is a restriction on the stepsize; a stepsize satisfying this restriction will be called *proper*, and it will be proved that proper stepsizes do exist).

3. [corrector step] Apply to the function $F(t^+, \cdot)$ the damped Newton minimization

$$y^{s+1} = y^s - \frac{1}{1 + \lambda(t^+, y^s)} [\nabla_x^2 F(t^+, y^s)]^{-1} \nabla_x F(t^+, y^s), \quad y^0 = \widetilde{x}. \quad (14.45)$$

(14.45) is terminated when it turns out that $\lambda(t^+, y^s) \leq \kappa$; the corresponding y^s is taken as x^+, which ensures $\mathcal{P}_\kappa$. The updating $(t,x) \mapsto (t^+, x^+)$ is completed.

Comment. The origin of equations (14.43) and (14.44) is clear: (14.43) is the equation in variations corresponding to the equation $\nabla_x F(s,y)=0$ of the surface $\mathcal{S}$, so that $\Pi(t,x)$ is the "approximate tangent plane" to the surface at the point (t,x). The primal search line R is given by the linearization

$$\nabla_x F(t,x)+\frac{\partial}{\partial t}\nabla_x F(t,x)dt+\nabla_x^2 F(t,x)dx=0$$

of the equation of the surface at the point (t,x): it is comprised of the points $(t+dt,x+dx)$ with dx and dt linked by the above equation and dt is proportional to δt.

We do not discuss here how to choose the direction δt; it depends on an "upper-level" strategy of tracing the surface, the issue to be discussed in Section 14.5.

Since $(t^+,\widetilde{x})$ is a $\overline{\kappa}$-good pair, the number of iterations (14.45) at a corrector step - the *Newton complexity* of the step - can be bounded as follows:

Proposition 14.2.3 [[3] , Theorem 2.2.3 and Proposition 2.2.2] *Process (14.45) is well-defined (i.e., it keeps the iterates y^s in* int G_{t^+}*) and results in y^s such that $\mathcal{P}_\kappa(t^+,y^s)$ is satisfied in no more than*

$$N=O(1)\left\{\overline{\kappa}+\ln\ln(1/\kappa)\right\} \tag{14.46}$$

Newton iterations (14.45); here and further $O(1)$ denote appropriately chosen absolute constants.

The point is, of course, how to choose the "large" stepsize r for which the forecast $X(r)=(t,x)+r(\delta t,\delta x)$ satisfies the predicate $\mathcal{R}_{\overline{\kappa}}$. To this end it is natural to use line search in r. A straightforward line search is impossible, since $V(\tau,y)$ involves the implicitly defined quantity

$$f^*(\tau)=\min_{u\in Q_\tau}F(\tau,u).$$

What we intend to do is to derive "computationally cheap" *lower bound* for the latter quantity. This is the issue we are coming to.

14.3 Dual bounds

14.3.1 Basic assumption

From now on we make the following assumption on the barrier Ψ under consideration:

(C) we know the Legendre transformation

$$\Psi_*(s)=\sup_u\{s^Tu-\Psi(u)\} \tag{14.47}$$

of the barrier Ψ.

"We know Ψ_*" means that, given s, we are able to check whether s belongs to the domain $\mathrm{Dom}\,\Psi_*$ of the Legendre transformation, and if it is the case, are able to compute $\Psi_*(s)$.

Note that by assumption Ψ'' is nondegenerate, so that the domain of Ψ_* is an open convex set and Ψ_* is s.-c. on its domain (Proposition 14.2.2, (ii) and (iii)).

14.3.2 Dual bounds

Let us start with the following simple observation

Lemma 14.3.1 *Let $s\in\mathrm{Dom}\,\Psi_*$ satisfy the linear homogeneous equation*

$$\pi^T s=0. \tag{14.48}$$

Then for any $\tau\in T$ we have

$$f^*(\tau)\geq[\sigma\tau+\epsilon]^T s-\Psi_*(s). \tag{14.49}$$

Proof. Since Ψ is the Legendre transformation of Ψ_* (Proposition 14.2.2.(iii)), we have for any $y \in Q_\tau$

$$F(\tau, y) = \Psi(\sigma\tau + \pi y + \epsilon) \geq [\sigma\tau + \pi y + \epsilon]^T s - \Psi_*(s) = [\sigma\tau + \epsilon]^T s - \Psi_*(s)$$

(we have used that $\pi^T s = 0$). ∎

According to Lemma, each *dual feasible vector* s (a vector s from $\mathrm{Dom}\,\Psi_*$ satisfying (14.48)) results in an affine lower bound for the function $f^*(\cdot)$, and these are the bounds we intend to use in order to ensure $\mathcal{R}$. Note that dual feasible vectors belong to the subspace $\mathcal{D} \subset \mathbf{R}^N$ of all solutions to linear equation (14.48); the vectors from this subspace will be called *dual feasible directions*. The set $\mathcal{D}^*$ of dual feasible vectors clearly is an open convex subset of $\mathcal{D}$.

We are about to present a systematic way to generate dual feasible directions and dual feasible vectors.

14.3.3 Dual search line

Lemma 14.3.2 *Given a primal search line*

$$R = \{X(p) = (t + p\delta t, x - d_x(t,x) + p\delta x) \mid p \in \mathbf{R}\} \tag{14.50}$$

($(t,x) \in Q$, $\delta t \in \mathbf{R}^k$, δx and $d_x(t,x)$ are given by (14.43) and (14.44)), set

$$u = \sigma t + \pi x + \epsilon, \quad s(t,x) = \Psi'(u), \quad d_s(t,x) = \Psi''(u)\pi d_x(t,x), \quad \delta s = \Psi''(u)[\sigma\delta t + \pi\delta x] \tag{14.51}$$

and define the dual search line as

$$R^* = \{S(p) = s(t,x) - d_s(t,x) + p\delta s \mid p \in \mathbf{R}\}. \tag{14.52}$$

Then all vectors from R^* are dual feasible directions:

$$\pi^T S(p) = 0, \ p \in \mathbf{R}. \tag{14.53}$$

Moreover,

$$\lambda(t,x) < 1 \Rightarrow S(0) \in \mathrm{Dom}\,\Psi_*, \tag{14.54}$$

so that under the premise of (14.54) all points $S(p)$ corresponding to small enough $|p|$ are dual feasible vectors.

Proof. To simplify notation, let us omit explicit indicating the argument values in the below computations; the values of all quantities related to Ψ are taken at the point u, and the values of the quantities related to Ψ_* are taken at the point $s = s(t,x) = \Psi'(u)$.

By virtue of (14.44), (14.51) and (14.52) we have

$$\pi^T S(p) = \pi^T \Psi' - \pi^T \Psi'' \pi [\nabla_x^2 F(t,x)]^{-1} \nabla_x F(\tau, x) + p\pi^T \Psi''[\sigma\delta t + \pi\delta x] =$$

[since $\nabla_x^2 F(t,x) = \pi^T \Psi'' \pi$, $\nabla_x F(t,x) = \pi^T \Psi'$ and due to (14.43)]

$$= p\pi^T \Psi''[\sigma\delta t + \pi\delta x] = 0,$$

as required in (14.53).

To prove (14.54), note that

$$\lambda^2(t,x) = |\nabla_x F(t,x)|^2_{[\nabla_x^2 F(t,x)]^{-1}} = \left|[\nabla_x^2 F(t,x)]^{-1} \nabla_x F(t,x)\right|^2_{\nabla_x^2 F(t,x)} =$$

$$= |d_x(t,x)|^2_{\pi^T \Psi'' \pi} = |\pi d_x(t,x)|^2_{\Psi''} =$$

[see (14.51) and take into account that $[\Psi'']^{-1} = \Psi''_*$]

$$= \left|[\Psi'']^{-1} d_s(t,x)\right|^2_{\Psi''} = |d_s(t,x)|^2_{[\Psi'']^{-1}} = |d_s(t,x)|^2_{\Psi''_*}.$$

Thus, we come to

$$|d_s(t,x)|_{\Psi''_*} = |\pi d_x(t,x)|_{\Psi''} = \lambda(t,x). \tag{14.55}$$

It remains to note that, as we know, Ψ_* is s.-c. on its domain and that $s = \Psi' \in \mathrm{Dom}\,\Psi_*$, so that (14.55) combined with Proposition 14.2.2.(i) (applied to Ψ_*) implies that if $\lambda(t,x) < 1$, then $S(0) = s - d_s(t,x) \in \mathrm{Dom}\,\Psi_*$. ∎

Now we are ready to present the "computationally cheap" sufficient condition for a forecast $X(r)$ to satisfy the predicate $\mathcal{R}$:

Basic Test: *given r, compute $X(r)$ and $S(r)$ (see (14.50) - (14.52)) and verify whether $X(r) \in Q$ and $S(r) \in \mathrm{Dom}\,\Psi_*$ (then $S(r)$ is a dual feasible vector, see Lemma 14.3.2). If one of these inclusions is not valid, reject r, otherwise check the inequality*

$$v(r) \equiv F(t + r\delta t, x - d_x(t,x) + r\delta x) + \Psi_*(S(r)) - [\sigma[t + r\delta t] + \epsilon]^T S(r) \le \overline{\kappa}. \tag{14.56}$$

If it is satisfied, accept r, otherwise reject it.

An immediate consequence of lemmas 14.3.1 and 14.3.2 is as follows.

Proposition 14.3.1 *Let $(t,x) \in Q$, $(\delta t, \delta x) \in \Pi(t,x)$, and let $r \ge 0$ be such that r passes the Basic Test. Then the forecast $X(r) = (t + r\delta t, x - d_x(t,x) + r\delta x)$ satisfies the predicate $\mathcal{R}_{\overline{\kappa}}$.*

We should, of course, prove that the test is "reasonable", namely, that it accepts at least "small" steps in the parameters leading to the standard overall complexity of the algorithm. This is the issue we are coming to.

14.4 Main results on tracing a surface

14.4.1 Acceptable stepsizes

Our main observation is that a stepsize r such that the displacement $r\delta t$ is not too large in the Euclidean metric defined by the matrix $\nabla_t^2 F(t,x)$ for sure passes the Basic Test. We start from the following simple

Proposition 14.4.1 *Given $u \in \mathrm{Dom}\,\Psi$ and $du \in \mathbf{R}^N$, let us set $s = \Psi'(u)$, so that $s \in \mathrm{Dom}\,\Psi_*$, and $ds = \Psi''(u)du$. Let also*

$$\begin{array}{rcl}\rho_u^*[du] & = & \Psi(u+du) + \Psi_*(s+ds) \\ & & -\left[\Psi(u) + \Psi_*(s) + (du)^T\Psi'(u) + (ds)^T\Psi_*'(s) + \frac{1}{2}(du)^T\Psi''(u)du + \frac{1}{2}(ds)^T\Psi_*''(s)ds\right]\end{array}$$

be the remainder in the second-order Taylor expansion of the function $\Psi + \Psi_*$ at the point (u,s) along the displacement (du, ds) (if this displacement moves the point outside the domain of the function, then $\rho^* = \infty$). Then

(i) $|du|^2_{\Psi''(u)} = |ds|^2_{\Psi_*''(s)} = (du)^T ds$,

(ii) *if $\chi \equiv |du|_{\Psi''(u)} < 1$, then $\rho_u^*[du]$ is well-defined and, moreover,*

$$\rho_u^*[du] \le 2\rho(\chi) - \chi^2, \tag{14.57}$$

($\rho(\cdot)$ is given by (14.40)) and

(iii) *The third derivative of the function $\Psi + \Psi_*$ taken at the point (u,s) along the direction (du, ds) is zero, so that $\rho_u^*(du)$ is in fact the remainder in the third* order *Taylor expansion of $\Psi + \Psi_*$ at the point (u,s) along the direction (du, ds).*

Proof. (i) is an immediate consequence of the relations $ds = \Psi''(u)du$ and $\Psi_*''(s) = [\Psi''(u)]^{-1}$. (i) combined with the upper bound on the remainder in the first-order Taylor expansion of a s.-c. function (Proposition 14.2.2.(i)) results in (ii). To verify (iii), let us differentiate the identity (y varies, h is fixed)

$$h^T\Psi_*''(\Psi'(y))h = h^T[\Psi''(y)]^{-1}h$$

at the point $y = u$ in the direction du, which results in

$$D^3\Psi_*(s)[h, h, \Psi''(u)du] = -D^3\Psi(u)[[\Psi''(u)]^{-1}h, [\Psi''(u)]^{-1}h, du];$$

substituting $h = \Psi''(u)du \equiv ds$, we come to $D^3\Psi_*(s)[ds, ds, ds] = -D^3\Psi(u)[du, du, du]$, as required in (iii). ∎

Theorem 14.4.1 *Let (t,x) satisfy $\mathcal{P}_\kappa$, $(\delta t,\delta x)\in\Pi(t,x)$, let*

$$R^*=\{S(p)=s(t,x)-d_s(t,x)+p\delta s\mid p\in\mathbf{R}\}$$

be the dual search line associated with the primal search line

$$R=\{X(p)=(t,x-d_x(t,x))+p(\delta t,\delta x)\mid p\in\mathbf{R}\}$$

(see Lemma 14.3.2) and let

$$u=\sigma t+\pi x+\epsilon,\ du(r)=r\sigma\delta t+\pi[-d_x(t,x)+r\delta x].$$

Then

(i) *The vector δx is the minimizer of the quadratic form $|\sigma\delta t+\pi h|^2_{\Psi''(u)}$ over $h\in\mathbf{R}^n$, and, in particular,*

$$\zeta\equiv|\sigma\delta t+\pi\delta x|_{\Psi''(u)}\le|\sigma\delta t|_{\Psi''(u)}. \tag{14.58}$$

(ii) *One has*

$$\chi(r)\equiv|du(r)|_{\Psi''(u)}=\sqrt{r^2\zeta^2+\lambda^2(t,x)} \tag{14.59}$$

and

$$v(r)\equiv F(X(r))+\Psi_*(S(r))-[\sigma(t+r\delta t)+\epsilon]^TS(r)=\rho_u^*[du(r)]; \tag{14.60}$$

in particular,

$$\chi(r)<1\Rightarrow v(r)\le 2\rho(\chi(r))-\chi^2(r), \tag{14.61}$$

so that if

$$\omega_*=\max\left[\omega\mid 2\rho(\sqrt{\kappa^2+\omega^2})-\kappa^2-\omega^2\le\overline{\kappa}\right] \tag{14.62}$$

then all stepsizes r with

$$|r|\le\frac{\omega_*}{|\sigma\delta t|_{\Psi''(u)}} \tag{14.63}$$

for sure pass the Basic Test.

Proof. From now on the quantities related to F, Ψ and Ψ_*, if no argument values are explicitly indicated, are taken at the points (t,x), $u=\sigma t+\pi x+\epsilon$, $s=s(t,x)=\Psi'(u)$, respectively.

1^0. The minimizer of the quadratic form $|\sigma\delta t+\pi h|^2_{\Psi''}=[\sigma\delta t+\pi h]^T\Psi''[\sigma\delta t+\pi h]$ of $h\in\mathbf{R}^n$ is given by

$$\pi^T\Psi''[\sigma\delta t+\pi h]=0;$$

this is exactly the equation defining δx, see (14.43), which proves the first statement of the theorem.

2^0. By definition of $d_x(t,x)$ (see (14.44)), $d_s(t,x)$ (see (14.51)) and the correspondence between δx and δs given by (14.51) we have

$$d_s(t,x)=\Psi''\pi d_x(t,x),\ \delta s=\Psi''[\sigma\delta t+\pi\delta x] \tag{14.64}$$

whence

$$ds(r)\equiv-d_s(t,x)+r\delta s=\Psi''du(r). \tag{14.65}$$

By construction $s=\Psi'(u)$. From (14.55) we know that

$$|d_s(t,x)|_{\Psi''_*}=|\pi d_x(t,x)|_{\Psi''}=\lambda(t,x);$$

besides this, $[\pi d_x(t,x)]^T\Psi''[\sigma\delta t+\pi\delta x]=0$ in view of (14.43), so that

$$\chi^2(r)\equiv|du(r)|^2_{\Psi''}=|-\pi d_x(t,x)+r[\sigma\delta t+\pi\delta x]\,|^2_{\Psi''}=\lambda^2(t,x)+r^2\zeta^2, \tag{14.66}$$

as claimed in (14.59).

3^0. We have (see the definitions of $X(r)$, $S(r)$, $du(r)$, $ds(r)$)

$$F(X(r))+\Psi_*(S(r))=\Psi(u+du(r))+\Psi_*(s+ds(r))=$$

[definition of $\rho^*_u[\cdot]$ combined with (14.65)]

$$= \Psi(u) + \Psi_*(s) + (du(r))^T\Psi' + (ds(r))^T\Psi'_* + \frac{1}{2}|du(r)|^2_{\Psi''} + \frac{1}{2}|ds(r)|^2_{\Psi''_*} + \rho^*_u[du(r)] =$$

[Proposition 14.4.1.(i) as applied to $du = du(r)$ combined with (14.65); relations $\Psi' = s, \Psi'_* = u$]

$$= u^T s + (du(r))^T s + (ds(r))^T u + |du(r)|^2_{\Psi''} + \rho^*_u[du(r)] =$$

$$= (u + du(r))^T(s + ds(r)) - (du(r))^T ds(r) + |du(r)|^2_{\Psi''} + \rho^*_u[du(r)] =$$

[Proposition 14.4.1.(i)]

$$= (u + du(r))^T(s + ds(r)) + \rho^*_u[du(r)]. \tag{14.67}$$

By construction,

$$(u + du(r))^T(s + ds(r)) = [\sigma(t + r\delta t) + \epsilon]^T S(r) + [-d_x(t,x) + r\delta x]^T \pi^T S(r) =$$

[Lemma 14.3.2]

$$= [\sigma(t + r\delta t) + \epsilon]^T S(r),$$

so that (14.67) implies (14.60). Relation (14.61) follows from (14.57) and (14.60), and the concluding statement of the Theorem is a corollary of (14.58) and (14.61). ■

How long are "long steps"

Theorem 14.4.1 says that if the tolerances κ and $\overline{\kappa}$ are chosen reasonably (say, $\kappa = 0.125$ and $\overline{\kappa} = 2$) and (t,x) is κ-close to $\mathcal{S}$, then any step

$$(t,x) \mapsto (t^+, \widetilde{x}) = (t + \delta t, x - d_x(t,x) + \delta x)$$

"along the surface" (i.e., with $(\delta t, \delta x) \in \Pi(t,x)$) of "$O(1)$-local length", namely, with

$$\zeta \equiv |\sigma\delta t + \pi\delta x|_{\Psi''(\sigma t + \pi x + \epsilon)} \leq 0.89$$

results in a $\overline{\kappa}$-good forecast $(t^+, \widetilde{x})$ (indeed, for $\kappa = 0.125$ and $\zeta = 0.89$ the right hand side in (14.61) is < 2). The natural question is whether we could ensure a larger step, still resulting in a $\overline{\kappa}$-good forecast. For the sake of simplicity, let us answer this question for the case when the point (t,x) belongs to $\mathcal{S}$ (i.e., $\lambda(t,x) = 0$) rather than is κ-close to the surface (modifications required in the case of small positive $\lambda(t,x)$ are quite straightforward).

When answering the question, we can normalize the direction $(\delta t, \delta x) \in \Pi(t,x)$ to have unit local length:

$$|du|_{\Psi''} = 1, \;\; du = \sigma\delta t + \pi\delta x \tag{14.68}$$

(here and in what follows all quantities related to Ψ, Ψ_* are evaluated at the points $u = \sigma t + \pi x + \epsilon$ and $s = \Psi'(u)$, respectively). Recall that we have associated with the data $(t, x, \delta t)$ the primal and the dual search lines R and R^*; it is convenient to aggregate these lines in a single "primal-dual" line

$$R_{pd} = (u,s) + \mathbf{R}(du, ds) \subset \mathbf{R}^{2N}_{pd} = \mathbf{R}^N_p \times \mathbf{R}^N_d,$$

where $ds = \Psi'' du$; the projection of R_{pd} onto the space $\mathbf{R}^n_d$ of dual variables is R^*, while the projection of R_{pd} onto the space $\mathbf{R}^N_p$ of the primal variables is the image of the primal search line R under the embedding $(\tau, y) \mapsto \sigma\tau + \pi y + \epsilon$. It is convenient to equip the primal-dual space $\mathbf{R}^{2N}_{pd}$ with the Euclidean norm

$$|(v_p, v_d)|_{pd} = \sqrt{|v_p|^2_{\Psi''} + |v_d|^2_{\Psi''_*}};$$

this is nothing but the local norm $|\cdot|_{\Xi''(z_0)}$ given by the Hessian of the s.-c. function

$$\Xi(v_p, v_d) = \Psi(v_p) + \Psi_*(v_d)$$

at the point $z_0 = (u, s)$.

Let us define T as the distance from z_0 to the boundary of the domain $\mathrm{Dom}\,\Xi = (\mathrm{Dom}\,\Psi)\times(\mathrm{Dom}\,\Psi_*)$ along the line R_{pd}:

$$T = \min\{|z - z_0|_{pd} \mid z \in R_{pd}\backslash \mathrm{Dom}\,\Xi\};$$

note that $T \geq 1$ due to Proposition 14.2.2.(i). Now, when choosing a stepsize r, forming the corresponding forecast and subjecting it to the Basic Test, we in fact generate and process the point $z_r = (u+rdu, s+rds)$ on the primal-dual search line R_{pd}, i.e., perform certain step along R_{pd}. The $|\cdot|_{pd}$-length of this step $|z_r - z_0|_{pd}$ is simply $|r|\sqrt{2}$ (indeed, $ds = \Psi'' du$, so that $|ds|_{\Psi''_*} = |du|_{\Psi''}$ by Proposition 14.4.1, while du is normalized by (14.68)). It follows that T is a natural upper bound on the acceptable step $|z_r - z_0|_{pd}$ – when $r = 2^{-1/2}T$ and du is "badly oriented", the stepsize r results in $z_r \notin \mathrm{Dom}\,\Xi$ and is therefore rejected by the Basic Test. With these remarks, the above question "how long are long steps" can be posed as follows:

Which fraction of T indeed is accepted by the Basic Test, i.e., which fraction of the way to the boundary of the primal-dual feasible set $\mathrm{Dom}\,\Xi$ along the direction (du, ds) can we cover in one step of the Basic Updating scheme ?

According to the above discussion, we for sure can move towards the boundary by the fixed distance $0.89\sqrt{2}$; this is a "short step" allowed in any path-following interior point method, and to get a result of this type, no structural assumption **(C)** on the s.-c.b. in question and no dual bounding are needed. If T is large, then a short step covers small part of the way to the boundary, and a short-step method becomes slow.

In fact our approach in many cases enables much larger steps:

Proposition 14.4.2 *Let both the barrier Ψ and its Legendre transformation Ψ_* be α-regular (Definition 14.2.1), let $(t, x) \in \mathcal{S}$ and let δt be such that (14.68) takes place. Then all stepsizes satisfying*

$$|r| \leq r^*(\alpha)\sqrt{T}, \tag{14.69}$$

same as all stepsizes satisfying

$$|r| \leq r^*(\alpha)T\vartheta_*^{-1/4} \tag{14.70}$$

are accepted by the Basic Test. Here $r^*(\alpha) > 0$ depends on α only and ϑ_* is the parameter of the s.-c.b. Ψ.

Proof. Let $R_{pd} = z_0 + \mathbf{R}dz$, $dz = (du, ds)$, be the primal-dual search line associated with $t, x, \delta t$, and let $\Delta = \{r \mid z_0 + 2^{-1/2}rdz \in \mathrm{cl}\,\mathrm{Dom}\,\Xi\}$, so that Δ is a closed convex set on the axis containing the segment $[-T, T]$. By Proposition 14.2.1, the function

$$\phi(r) = \Xi(z_0 + 2^{-1/2}rdz)$$

is self-concordant and α-regular on int Δ. Since we clearly have $\parallel 1 \parallel_{\mathrm{int}\ \Delta, r} \leq (T - |r|)^{-1}$ for $|r| < T$ (for notation, see Definition 14.2.1), inequality (14.38) implies that the function $\psi(r) \equiv \phi''(r)$ satisfies

$$|\psi''(r)| \leq \frac{\alpha(\alpha+1)}{(T - |r|)^2}\psi(r), \ \ |r| < T. \tag{14.71}$$

Besides this, we have

$$\psi(0) = 1, \quad \psi'(0) = 0, \tag{14.72}$$

the first relation coming from $\psi(0) = |2^{-1/2}dz|^2_{pd} = 1$ (note that $|dz|_{pd} = \sqrt{2}$ due to the discussion preceding the Proposition), and the second relation being given by Proposition 14.4.1.

We claim that

$$0 \leq \phi''(r) \equiv \psi(r) \leq \frac{T^\alpha}{(T - |r|)^\alpha}, \ \ |r| < T. \tag{14.73}$$

The left inequality follows from convexity of ϕ. By symmetry reasons, it suffices to establish the right inequality for $0 \leq r < T$. To this end note that the function

$$\omega(r) = \frac{T^\alpha}{(T - r)^\alpha}$$

clearly satisfies the relations

$$\omega''(r) = \frac{\alpha(\alpha+1)}{(T-r)^2}\omega(r),\ 0 \le r < T;\ \omega(0) = 1;\ \omega'(0) > 0. \tag{14.74}$$

From (14.74), (14.71), (14.72) we see that the function $\xi(r) = \omega(r) - \psi(r)$ satisfies the relations

$$\xi''(r) \ge \frac{\alpha(\alpha+1)}{(T-r)^2}\xi(r),\ 0 \le r < T;\quad \xi(0) = 0;\quad \xi'(0) > 0. \tag{14.75}$$

To establish (14.73) for $0 \le r < T$ is the same as to prove that $\xi \ge 0$ on $[0,T)$, which is immediate: since $0 = \xi(0) < \xi'(0)$, ξ is positive on $(0,\widehat{r})$ with some positive $\widehat{r}$. Consequently, if $\xi < 0$ somewhere on $[0,T)$, then ξ possesses a zero on $[\widehat{r},T)$. Let r^* be the smallest of zeros of ξ on $[\widehat{r},T)$; then ξ is nonnegative on $[0,r^*]$, whence, due to (14.75), ξ is convex on $[0,r^*]$. This observation combined with $\xi(0) = 0$, $\xi'(0) > 0$ contradicts the relation $\xi(r^*) = 0$.

Combining (14.73) and (14.71), we come to

$$|\phi^{(4)}(r)| \le \frac{\alpha(\alpha+1)T^\alpha}{(T-|r|)^{\alpha+2}},\quad |r| < T. \tag{14.76}$$

According to Theorem 14.4.1.(ii) and Proposition 14.4.1.(ii), the quantity $v(2^{-1/2}r)$, $v(\cdot)$ being defined by (14.56), is the remainder in the third-order Taylor expansion of $\phi(\cdot)$ at the point $r = 0$. From (14.76) we therefore conclude that

$$v(r) \le \frac{\alpha(\alpha+1)T^\alpha r^4}{6(T-|r|\sqrt{2})^{\alpha+2}},\quad \sqrt{2}|r| < T. \tag{14.77}$$

Since the Basic Test accepts all stepsizes with $v(r) \le \overline{\kappa}$, we see from (14.77) that it accepts all stepsizes satisfying (14.69), with properly chosen $r^*(\alpha)$ (recall that $T \ge 1$).

In view of already proved part of the statement, in order to demonstrate acceptability of the stepsizes (14.70) with properly chosen $r^*(\alpha)$, it suffices to verify that

$$T \le \gamma(\alpha)\sqrt{\vartheta_*}. \tag{14.78}$$

This inequality is an immediate consequence of the following

Lemma 14.4.1 *Let Ψ be α-regular ϑ_*-s.c.b. for convex domain G^+, let $u \in \operatorname{int} G^+$, let du be such that $|du|_{\Psi''(u)} = 1$, and let T be such that $u \pm T du \in G^+$. Then*

$$T \le 2^{2+\alpha/2}\sqrt{\vartheta_*}. \tag{14.79}$$

Proof. Without loss of generality we may assume that $(du)^T\Psi'(u) \ge 0$ (otherwise we could replace du with $-du$). Let

$$\phi(r) = \Psi(u + r du),\ \ r \in \Delta = \{r \mid u + r du \in G^+\} \supset [-T,T];$$

by Proposition 14.2.1, ϕ is α-regular ϑ_*-s.-c.b. for Δ. We claim that if $r \in \operatorname{int}\Delta$ and $d > 0$ is such that $r \pm 2d \in \operatorname{int}\Delta$, then

$$0 \le \phi''(r-d) \le 2^{\alpha+1}\phi''(r). \tag{14.80}$$

Indeed, by Proposition 14.2.1 the function $\chi(s) = \phi(r-s) + \phi(r+s)$ is α-regular on the set $\Delta_r = \{s \mid r \pm s \in \operatorname{int}\Delta\} \supset [-2d,2d]$ and is such that $\chi'(0) = \chi'''(0) = 0$. From these properties, same as above (cf. (14.73)), one can derive the inequality

$$\chi''(s) \le \frac{(2d)^\alpha}{(2d-|s|)^\alpha}\chi''(0),\ \ |s| < 2d.$$

Substituting $s = d$, we get

$$\phi''(r-d) + \phi''(r+d) = \chi''(d) \le 2^\alpha\chi''(0) = 2^{\alpha+1}\phi''(r),$$

which, in view of the convexity of ϕ, implies (14.80).

Now let $0 \le r < T/3$. Applying (14.80) to $d = r$, we get

$$\phi''(r) \ge 2^{-\alpha-1}\phi''(0) = 2^{-\alpha-1},\ 0 < r < T/3 \qquad (14.81)$$

(the equality is given by $|du|_{\Psi''(u)} = 1$), whence, in view of $\phi'(0) = (du)^T\Psi'(u) \ge 0$,

$$\phi'(r) \ge 2^{-\alpha-1}r,\ 0 < r < T/3.$$

Now note that ϕ is ϑ_*-self-concordant barrier for Δ, whence, by [3] , Proposition 2.3.2, $\phi'(r)(s-r) \le \vartheta_*$ for all $s \in \Delta$. Applying this inequality to $s = T \in \Delta$ and $r \in (0, T/3)$, we get $2^{-\alpha-1}r(T-r) \le \vartheta_*$, $0 < r < T/3$, and (14.79) follows. ■

Results similar to those stated by Proposition 14.4.2 were recently obtained in [6, 7] for the predictor-corrector interior point methods associated with the *self-scaled* cones. Note that the property of Ψ and Ψ_* to be α-regular seems to be less restrictive that the one imposed in [6, 7]; we shall see in Section 14.6 that the property of 6-regularity is shared by the functions Ψ and Ψ_* responsible for Linear, Quadratically Constrained Quadratic and Semidefinite Programming (these applications are covered by the results of [6, 7] as well), same as by those responsible for Geometric Programming (where [6, 7] is inapplicable).

14.4.2 Centering property

To proceed, we need certain centering property of the surface of analytic centers given by the following

Proposition 14.4.3 *Let (t, x) satisfy $\mathcal{P}_\kappa$ with $\kappa \le 0.125$, and let $t' \in \operatorname{cl} T$. Then*

$$[\Psi'(\sigma t + \pi x + \epsilon)]^T\sigma(t'-t) \le (1+6\kappa)\vartheta_* + 4\kappa. \qquad (14.82)$$

Proof. By evident reasons, it suffices to consider the case $t' \in T$. Let $x' \in Q_{t'}$. From general properties of s.-c.b.'s ([3] , Proposition 2.3.2) it follows that

$$(\nabla F(t,x))^T(t'-t, x'-x) \le \vartheta_*,$$

so that (in what follows the derivatives of Ψ are taken at $\sigma t + \pi x + \epsilon$)

$$(\Psi')^T[\sigma(t'-t) + \pi(x'-x)] \le \vartheta_*.$$

Therefore

$$(\Psi')^T\sigma(t'-t) \le \vartheta_* + (\Psi')^T\pi(x-x') = \vartheta_* + (\nabla_x F(t,x))^T(x-x'). \qquad (14.83)$$

On the other hand, let x^* be the minimizer of $F(t,\cdot)$, and let

$$|h| = (h^T\nabla_x^2 F(t,x^*)h)^{1/2},\ \ |h|_* = (h^T[\nabla_x^2 F(t,x^*)]^{-1}h)^{1/2}.$$

From the relation $\lambda(t,x) \le \kappa \le 0.125$ it follows (see [3] , Theorem 2.2.2 and Proposition 2.3.2) that

$$|x-x^*| \le 1,\ \ |x'-x^*| \le 1 + 3\vartheta_*,\ \ |\nabla_x F(t,x)|_* \le 2\kappa.$$

Therefore the concluding expression in (14.83) does not exceed $|x-x'||\nabla_x F(t,x)|_* \le 2\kappa(3\vartheta_*+2) + \vartheta_*$, and we come to (14.82). ■

What we are interested in is the following consequence of the latter theorem:

Corollary 14.4.1 *Let $F(t,x)$ be the barrier associated with a surface of the type $\mathcal{S}_k(c_1,...,c_k)$:*

$$F(t,x) = -\vartheta\sum_{i=1}^k \ln(t_i - c_i^T x) + F(x),$$

where F is a ϑ-s.-c.b. for a closed and bounded convex domain $G \subset \mathbf{R}^n$, and let (t,x) satisfy $\mathcal{P}_\kappa$ with some $\kappa \le 0.125$. If $t' \le t$ belongs to $\operatorname{cl} T$, then

$$\sum_{i=1}^k \frac{t_i - t_i'}{\Delta_i} \le 2(k+1),\ \ \Delta_i = \Delta_i(t,x) = t_i - c_i^T x. \qquad (14.84)$$

Geometrically: The part of $\mathrm{cl}\, T$ *"to the left of* t*", i.e., comprised of* $t' \le t$*, belongs to the simplex*

$$\Delta(t,x) = \{t' \le t \mid \sum_{i=1}^{k} \frac{t_i - t'_i}{\Delta_i} \le 2(k+1)\}$$

and contains the box

$$C(t,x) = \{t' \le t \mid \frac{t_i - t'_i}{\Delta_i} \le 1,\ i = 1,...,k\}.$$

In particular, if

$$t^*_i(t) = \inf\{c_i^T y \mid y \in G,\ c_j^T y \le t_j,\ j = 1,...,k\},$$

then, for every $i \le k$*,*

$$t_i - t^*_i(t) \le 2(k+1)\Delta_i(t,x). \tag{14.85}$$

Proof. In the case in question the left hand side in (14.82) is equal to $\vartheta \sum_{i=1}^{k}(t_i - t'_i)\Delta_i^{-1}$, while the right hand side is $(1+6\kappa)\vartheta_* + 4\kappa \le 1.75(k+1)\vartheta + 0.5 \le 2\vartheta(k+1)$ (recall that $\vartheta \ge 1$, see Definition 14.2.1). Thus, (14.82) implies (14.84). The inclusion $C(t,x) \subset \mathrm{cl}\, T$ is evident, since for $t' \in \mathrm{int}\ C(t,x)$ we simply have $c_i^T x < t'_i$, $i = 1,...,k$, so that $(t',x) \in Q = \{(\tau,y) \mid y \in \mathrm{int}\ G,\ \tau_i > c_i^T y,\ i = 1,...,k\}$. Relation (14.85) is an immediate consequence of the preceding statements of the Corollary. ■

14.5 Solving convex programs via tracing surfaces

To the moment we know what are our local abilities to trace a surface of analytic centers, but did not discuss the "strategy" - where to move in order to solve the problem the surface is associated with. This question does not occur in the usual path-following approach, since there is a unique reasonable strategy: to vary the parameter in the only direction of interest at the highest possible rate compatible with the restriction on the Newton complexity of the corrector steps. In the multi-parameter case the strategy of tracing the surface requires special investigation; this is the issue we are coming to.

We intend to apply our surface-following scheme to convex programs

$$\begin{array}{ll} (P) & \textit{minimize } c^T x \textit{ s.t. } x \in G, \\ (P') & \textit{minimize } c^T x \textit{ s.t. } x \in G,\ f^T x \le 0; \end{array}$$

in both of the problems, G is a closed and bounded convex domain in $\mathbf{R}^n$ represented by a ϑ-s.-c.b. F, and we are given a starting point $\widehat{x} \in \mathrm{int}\ G$. In the second problem it is assumed that the quantity

$$f^* = \min_{x \in G} f^T x \tag{14.86}$$

is nonnegative (the case of feasible (P') clearly corresponds to the case of $f^* = 0$).

To make presentation more compact, we shall focus on (evidently more general) problem (P'); to get constructions and results for (P), it suffices to set in what follows $f = 0$.

In Section 14.2.3 problem (P') was associated with the barrier

$$F(t,x) = -\vartheta \ln(t_1 - c^T x) - \vartheta \ln(t_2 - f^T x) - \vartheta \ln(t_3 - d^T x) + F(x) \tag{14.87}$$

and the 3-parameter surface $\mathcal{S}_3(c,f,d)$; here d is readily given by the requirement that the pair $(\widehat{t},\widehat{x})$, with certain explicit $\widehat{t}$, belongs to the surface. In what follows we deal with the setup

$$\begin{array}{rcl} \widehat{t}_1 &=& c^T\widehat{x} + \left[c^T[\nabla^2 F(\widehat{x})]^{-1}c\right]^{1/2}, \\ \widehat{t}_2 &=& f^T\widehat{x} + \left[f^T[\nabla^2 F(\widehat{x})]^{-1}f\right]^{1/2}, \\ d &=& -\vartheta^{-1}\nabla F(x) - [\widehat{t}_1 - c^T\widehat{x}]^{-1}c - [\widehat{t}_2 - f^T\widehat{x}]^{-1}f, \\ \widehat{t}_3 &=& d^T\widehat{x} + 1; \end{array} \tag{14.88}$$

Note that $\widehat{t}_j$, $j = 1,2$, are nothing but the maxima of the linear forms $c^T x$, respectively, $f^T x$, over the closed Dikin ellipsoid $W_F(\widehat{x})$, see Proposition 14.2.2; according to this Proposition, the ellipsoid is contained in G, so that

$$\widehat{t}_1 \le \max_{x \in G} c^T x; \quad \widehat{t}_2 \le \max_{x \in G} f^T x. \tag{14.89}$$

Below c^* denotes the optimal value in the problem in question. When measuring accuracy of an approximate solution $x \in G$, we normalize the residuals $c^T x - c^*$ and $f^T x$ by the variations of the corresponding linear forms on the domain of the problem, the variation of a linear form $e^T x$ on a bounded set U being defined as

$$\mathrm{V}_U(e) = \max_{x\in U} e^T x - \min_{x\in U} e^T x.$$

When solving (P') via tracing the surface $\mathcal{S}_3(c,f,d)$, our goal is to enforce the "objective parameter" t_1 and the "constraint parameter" t_2 to converge to the optimal value c^* and to 0, respectively. As for the "centering parameter" t_3, all we need is to control it in a way which allows us to achieve the indicated goals, and a reasonable policy is to push the parameter to ∞, since with a "small" value τ of the centering parameter the artificial constraint $d^T x \le \tau$ may vary the optimal value in the problem.

14.5.1 Assumption on the structure of F

In order to use the long-step tactics presented in Section 14.3, from now on we make the following assumption on the structure of the barrier F for the domain G:

$\mathcal{Q}$: *we are given a closed convex domain $H \in \mathbf{R}^M$, a ϑ-s.-c.b. Φ for H such that Φ'' is nondegenerate and the Legendre transformation Φ_* is known, and an affine mapping $B(x) = \pi x + \epsilon : \mathbf{R}^n \to \mathbf{R}^M$ with the image of the mapping intersecting* int H, *such that*

$$G = B^{-1}(H),\ F(x) = \Phi(B(x)).$$

Note that under this assumption any barrier of the type

$$F(t,x) = \vartheta \sum_{i=1}^{k} \psi_i(t_i - c_i^T x) + F(x),\ \psi_i(s) = -\ln s, \tag{14.90}$$

and, in particular, the barrier underlying the surface $\mathcal{S}_3(c,f,d)$, satisfies assumptions (A) - (C) from sections 14.2.5 and 14.3.1. The corresponding data are

$$G^+ = \mathbf{R}^k_+ \times H,\ \Psi(u_1,...,u_k,u) = \vartheta \sum_{i=1}^{k} \psi_i(u_i) + \Phi(u),\ \vartheta_* = (k+1)\vartheta,$$

$$\mathcal{A}(t,x) = (t_1 - c_1^T x, ..., t_k - c_k^T x, B(x));$$

the Legendre transformation of Ψ is

$$\Psi_*(s_1,...,s_k,s) = \vartheta \sum_{i=1}^{k} \psi_i(-s_i) + \Phi_*(s) + k\vartheta(\ln\vartheta - 1),$$

$$\mathrm{Dom}\,\Psi_* = \{(s_1,...,s_k) < 0\} \times \mathrm{Dom}\,\Phi_*.$$

14.5.2 Preliminary remarks

In what follows we speak about tracing the surface $\mathcal{S}_3(c,f,d)$. Sometimes we write c_i instead of the i-th vector identifying the surface (so that $c_1 = c$, $c_2 = f$, $c_3 = d$).

Let us fix the tolerances κ, $\overline{\kappa}$ such that

$$\kappa \in (0, 0.125];\ \overline{\kappa} > 2\rho(\kappa) - \kappa^2.$$

When tracing the surface $\mathcal{S}_3$, we form a sequence of pairs (t^i, x^i) satisfying the predicate $\mathcal{P}_\kappa$ associated with the surface. To update the pair (t^{i-1}, x^{i-1}) into the new pair (t^i, x^i), we use the Basic Updating Scheme equipped with the Basic Test for choosing a proper stepsize r_i in the current direction $(\delta t^i, \delta x^i)$, so that the forecast we use is

$$(t^i, \widetilde{x}^i) = (t^{i-1}, x^{i-1} - d_x(t^{i-1}, x^{i-1})) + r_i(\delta t^i, \delta x^i).$$

The above remarks specify the method up to the following two "degrees of freedom":

(I) *strategy of choosing the direction* δt^i*;*
(II) *tactics of choosing a proper stepsize* r_i.

(I) is the main subject of this section. As about (II), we are not going to be too specific. The only assumption on r_i is that it is *at least the "short step"* r_i^* *given by Theorem 14.4.1*, i.e. (see (14.63)),

$$r_i \geq r_i^* \equiv \frac{\omega_*}{\sqrt{\vartheta \sum_{j=1}^{3} (\delta t_j^i)^2 / \Delta_j^2(t^{i-1}, x^{i-1})}}, \;\; \Delta_j(t,x) = t_j - c_j^T x \tag{14.91}$$

(from now on, the superscript in notation like δt_j^i denotes the number of the step, and the subscript is the coordinate index). In view of the origin of ω_* (see (14.62)) and Theorem 14.4.1 (applied to barrier (14.87)), the default value r_i^* of the stepsize for sure passes the Basic Test (so that to use the default stepsize no Basic Test, and, consequently, no assumptions on the structure of F are needed). One is welcome to combine the Basic Test with any kind of line search to get a larger (proper) stepsize. Note that in what follows we sometimes impose certain "safety" upper bounds on r_i; each time it can be immediately verified that these bounds are valid for $r_i = r_i^*$, so that the safety bounds are consistent with the aforementioned lower bound on r_i.

With the above remarks, we may completely focus on the "strategic" issue (I).

As it was already explained, when tracing surface $\mathcal{S}_3(c,f,d)$, we should get rid of the centering parameter; the simplest way is to push it to ∞. How to ensure this, this is the issue we start with.

Consider a surface $\mathcal{S}_k(c_1,...,c_k)$ associated with barrier (14.90), and let (t,x) satisfy the predicate $\mathcal{P}_\kappa$ associated with the surface. Let

$$q = \frac{2k}{4k+3}. \tag{14.92}$$

We say that a direction $\delta t = (\delta t_1, ..., \delta t_k)$ in the space of parameters is *k-safe*, if

$$\delta t_k \geq 0, \frac{\delta t_k}{\Delta_k(t,x)} \geq 2(2k+1)\frac{-\delta t_i}{\Delta_i(t,x)}, \;\; i = 1,...,k-1, \tag{14.93}$$

where, same as in Corollary 14.4.1,

$$\Delta_i(t,x) = t_i - c_i^T x, \;\; i = 1,...,k,$$

and we say that a stepsize $r \geq 0$ in the direction δt is *k-safe*, if

$$t + q^{-1} r \delta t \in T \equiv \{t \in \mathbf{R}^k \mid \exists y \in \operatorname{int} G : t_i > c_i^T y, \, i = 1,...,k\}. \tag{14.94}$$

Lemma 14.5.1 *Let* (t,x) *satisfy* $\mathcal{P}_\kappa$ *with respect to* $\mathcal{S}_k(c_1,...,c_k)$*,* δt *be a safe direction,* r *be a safe stepsize, and let* $t^+ = t + r\delta t$*. Then*

$$t_k^+ - t_k^*(t^+) \geq \min\{1 + \frac{1}{2k+2}; 1 + \frac{r\delta t_k}{4(k+1)\Delta_k(t,x)}\}(t_k - t_k^*(t)), \tag{14.95}$$

where, same as in Corollary 14.4.1, $t_i^*(t) = \min\{c_i^T y \mid y \in G, \, c_j^T y \leq t_j, \, j = 1,...,k\}$.

Proof is given in Appendix A.

14.5.3 How to trace $\mathcal{S}_3(c,f,d)$

The algorithm

Our strategy for solving (P') is as follows. Given (t,x) κ-close to $\mathcal{S} \equiv \mathcal{S}_3(c,f,d)$, we say that (t,x) is *good*, if

$$t_2 \leq 16\Delta_2(t,x) \tag{14.96}$$

and is *bad* otherwise; here and in what follows, as always,

$$\Delta_1(t,x) = t_1 - c^T x, \, \Delta_2(t,x) = t_2 - f^T x, \, \Delta_3(t,x) = t_3 - d^T x.$$

At step i (where $z^{i-1} \equiv (t^{i-1}, x^{i-1})$ is updated into $z^i \equiv (t^i, x^i)$), we act as follows.

1) If the current iterate (t^{i-1}, x^{i-1}) is bad, we apply

Infeasibility Test: *If both the inequalities*

$$\Delta_1(t^{i-1}, x^{i-1}) > 9\vartheta_* \left(c^T [\nabla_x^2 F(t^{i-1}, x^{i-1})]^{-1} c\right)^{1/2} \tag{14.97}$$

and

$$\Delta_3(t^{i-1}, x^{i-1}) > 9\vartheta_* \left(d^T [\nabla_x^2 F(t^{i-1}, x^{i-1})]^{-1} d\right)^{1/2} \tag{14.98}$$

are valid, claim that (P') is infeasible and terminate.

2) If the method is not terminated by 1), we set

$$\delta t^i = \left\{ \begin{array}{llll} (-\Delta_1(z^{i-1}), & -\Delta_2(z^{i-1}), & 16\Delta_3(z^{i-1})), & z^{i-1} \text{ is good} \\ (\ \ \Delta_1(z^{i-1}), & \quad 0 \quad , & 16\Delta_3(z^{i-1})), & \text{otherwise} \end{array} \right. . \tag{14.99}$$

After the direction δt^i is determined, we use the Basic Updating Scheme to update (t^{i-1}, x^{i-1}) into (t^i, x^i); in this updating, we subject the stepsize r_i to the "safety restriction"

$$r_i |\delta t^i_j| \le \frac{1}{8} \Delta_j(t^{i-1}, x^{i-1}),\ j = 1, 2 \tag{14.100}$$

(this does not forbid "long steps": the short step r_i^* in our case is by factor $O(\sqrt{\vartheta})$ less than the upper bound in (14.100)).

The recurrence is started at $(t^0, x^0) = (\widehat{t}, \widehat{x}) \in \mathcal{S}_3$, see (14.88).

Remark 14.5.1 In the case of problem (P) – according to our convention, it means that $f = 0$ – all (t^{i-1}, x^{i-1}) are good, since $t_2^{i-1} = \Delta_2(t^{i-1}, x^{i-1})$, so that (14.99) always results in $\delta t_1^i < 0$: we decrease the parameter of interest, as it should be in the case of feasible start.

Complexity

To describe the rate of convergence of the resulting method, let us denote by N^* the index of the iteration where the method terminates (if it happens; otherwise $N^* = \infty$), and let $N(\varepsilon)$, $0 < \varepsilon \le 1$, be the index i of the first iteration starting with which all iterates are ε-solutions to (P'):

$$c^T x^j - c^* \le \varepsilon \mathrm{V}_G(c), \quad f^T x^j \le \varepsilon \mathrm{V}_G(f),\ \forall j,\ N^* \ge j > i$$

(note that in the case of $N^* < \infty$ the latter relations are for sure satisfied when $i = N^*$, so that $N(\varepsilon) \le N^*$). The efficiency estimate for the presented method is given by the following.

Theorem 14.5.1 *The method never claims a feasible problem (P') to be infeasible, and if (P') is infeasible, this is detected in no more than*

$$N^* = O(1)\omega_*^{-1}\sqrt{\vartheta} \ln\left(\frac{2\vartheta[G\!:\!\widehat{x}](\mathrm{V}_G(f) + f^*)}{f^*}\right) \tag{14.101}$$

iterations; here ω_* is given by (14.62), f^* is given by (14.86) and

$$[G\!:\!\widehat{x}] = \max\{s \mid \exists y \notin G : \widehat{x} + s(\widehat{x} - y) \in G\}$$

is the asymmetry coefficient of G with respect to $\widehat{x}$.

If (P) is feasible, then

$$N(\varepsilon) \le O(1){\omega_*}^{-1}\sqrt{\vartheta} \ln\left(\frac{2\vartheta[G\!:\!\widehat{x}]}{\varepsilon}\right) \quad \forall \varepsilon \in (0, 1). \tag{14.102}$$

The Newton complexity of any corrector step of the method does not exceed

$$O(1)\{\overline{\kappa} + \ln\ln(1/\kappa)\}.$$

Proof. 1^0. The upper bound on the Newton complexity of corrector steps is given by Propositions 14.2.3 and 14.3.1.
2^0. Note that (14.91), (14.99) and (14.100) result in

$$\frac{1}{8} \geq r_i \geq O(1)\omega_* \vartheta^{-1/2}. \tag{14.103}$$

Let, same as before,

$$G_t = \{x \in G \mid c^T x \leq t_1, f^T x \leq t_2, d^T f \leq t_3\}, \quad T = \{t \in \mathbf{R}^3 \mid \text{int } G_t \neq \emptyset\}$$

and

$$\begin{array}{rcl} t_1^*(t) & = & \min\{c^T x \mid x \in G, f^T x \leq t_2, d^T x \leq t_3\} \\ t_2^*(t) & = & \min\{f^T x \mid x \in G, c^T x \leq t_1, d^T x \leq t_3\} \\ t_3^*(t) & = & \min\{d^T x \mid x \in G, c^T x \leq t_1, f^T x \leq t_2\} \end{array}, \quad t \in T.$$

3^0. Let us prove that if the method terminates at certain step i, then (P') is infeasible. Let x be an arbitrary point of the domain $G_{t^{i-1}}$, let x^* be the minimizer of $F(t^{i-1}, \cdot)$, and let $H_* = \nabla_x^2 F(t^{i-1}, x^*)$, $H = \nabla_x^2 F(t^{i-1}, x^{i-1})$. As we know, $F(t^{i-1}, \cdot)$ is ϑ_*-s.-c.b. for $G_{t^{i-1}}$ and $\lambda(t^{i-1}, x^{i-1}) \leq \kappa \leq 0.125$; from these observations by [3] , Theorem 2.2.2.(iii) and Proposition 2.3.2, it follows that

$$|x^* - x|_{H_*} \leq 3\vartheta_* + 1, \quad |x^{i-1} - x^*|_H \leq 2\kappa, \quad H_* \leq (1 - 2\kappa)^{-2} H,$$

whence $|x - x^{i-1}|_H \leq 9\vartheta_*,\ x \in G_{t^{i-1}}$. Therefore for an arbitrary vector e we have

$$\max_{x \in G_{t^{i-1}}} e^T x \leq c^T x^{i-1} + 9\vartheta_* |e|_{H^{-1}}.$$

Applying this relation to $e = c$ and $e = d$ and taking into account (14.97) and (14.98), we see that

$$t_1^{i-1} > \max_{x \in G_{t^{i-1}}} c^T x, \quad t_3^{i-1} > \max_{x \in G_{t^{i-1}}} d^T x.$$

Thus, the constraints $c^T x \leq t_1^{i-1}$ and $d^T x \leq t_3^{i-1}$ are redundant in the description of $G_{t^{i-1}}$, whence

$$t_2^*(t^{i-1}) = \min\{f^T x \mid x \in G\}.$$

By Corollary 14.4.1 applied with $k = 3$, we have (see (14.85))

$$t_2^{i-1} - t_2^*(t^{i-1}) \leq 8\Delta_2(t^{i-1}, x^{i-1});$$

since the Infeasibility Test was applied at the step i, the pair (t^{i-1}, x^{i-1}) is bad, so that $t_2^{i-1} > 16\Delta_2(t^{i-1}, x^{i-1})$, and we come to

$$\min_{x \in G} f^T x = t_2^*(t^{i-1}) \geq t_2^{i-1} - 8\Delta_2(t^{i-1}, x^{i-1}) > 8\Delta_2(t^{i-1}, x^{i-1}) > 0,$$

and (P') is infeasible, as claimed.

The algorithm in question terminates when the Infeasibility Test detects that (P') is infeasible. In what follows it is however more convenient to think that we ignore the "reports on infeasibility", if any, and continue the process as if there were no Infeasibility Test at all.
4^0. Note that all directions δt^i satisfy (14.93) with $k = 3$. Besides this, (14.100) ensures that the stepsizes r_i are 3-safe. Indeed, in our case $q^{-1} = 5/2$, so that by (14.100) we have for $j = 1, 2$:

$$t_j^{i-1} + q^{-1} r_i \delta t_j^i \geq t_j^{i-1} - \Delta_j(t^{i-1}, x^{i-1}) = \begin{cases} c^T x^{i-1}, & j = 1 \\ f^T x^{i-1}, & j = 2 \end{cases};$$

since $\delta t_3^i > 0$, we also have $t^{i-1} + q^{-1} r_i \delta t_3^i > d^T x^{i-1}$, whence $t^{i-1} + q^{-1} \delta t^i \in T$.
5^0. Applying Lemma 14.5.1 and taking into account (14.103), we observe that

$$t_3^i - t_3^*(t^i) \geq (1 + r_i)(t_3^{i-1} - t_3^*(t^{i-1})) \geq (1 + O(1)\omega_* \vartheta^{-1/2})(t_3^{i-1} - t_3^*(t^{i-1})). \tag{14.104}$$

Let us derive from this observation that there exists the first moment i, let it be called i^*, when $t_3^i \geq \max_{x\in G} d^T x$, and that

$$i^* \leq O(1)\omega_*^{-1}\vartheta^{1/2}\ln(2\vartheta[G\!:\!\widehat{x}]). \tag{14.105}$$

To derive (14.105) from (14.104), it clearly suffices to verify that

$$\mathrm{V}_G(d) \leq 14\vartheta[G\!:\!\widehat{x}](\widehat{t}_3 - t_3^*(\widehat{t})). \tag{14.106}$$

To get (14.106), note that by (14.89) the domain

$$G_{\rm ini} = \{x \in G \mid c^T x \leq \widehat{t}_1,\ f^T x \leq \widehat{t}_2\}$$

contains the closed Dikin ellipsoid

$$W = \{y \mid |y - \widehat{x}|_{F''(\widehat{x})} \leq 1\}.$$

Let $\widehat{G} = G \cap (2\widehat{x} - G)$ be the symmeterization of G with respect to $\widehat{x}$; by definition of the quantity $[G\!:\!\widehat{x}]$, we have

$$G \subset \widehat{x} + [G\!:\!\widehat{x}](\widehat{G} - \widehat{x}). \tag{14.107}$$

On the other hand, the function

$$\widehat{F}(x) = F(x) + F(2\widehat{x} - x)$$

is $\widehat{\vartheta}$-s.-c.b. for $\widehat{G}$, $\widehat{\vartheta} = 2\vartheta$ (Proposition 14.2.1.(i)), and clearly $\nabla\widehat{F}(\widehat{x}) = 0$. From the latter inequality it follows ([3] , Theorem 2.2.2.(iii)) that $\widehat{G}$ is contained in $|\cdot|_{\widehat{F}''(\widehat{x})}$-ball of the radius $1 + 3\widehat{\vartheta} \leq 7\vartheta$ centered at $\widehat{x}$, whence

$$\widehat{G} - \widehat{x} \subset 7\vartheta(W - \widehat{x});$$

combining this relation with (14.107), we get

$$G \subset \widehat{x} + 7\vartheta[G\!:\!\widehat{x}](W - \widehat{x}).$$

It follows that for an arbitrary vector e one has

$$\mathrm{V}_G(e) \leq 7\vartheta[G\!:\!\widehat{x}]\mathrm{V}_W(e) = 14\vartheta[G\!:\!\widehat{x}](e^T\widehat{x} - \min_{x\in W} e^T x). \tag{14.108}$$

Taking into account that

$$\widehat{t}_3 - t_3^*(\widehat{t}) = \widehat{t}_3 - \min\{d^T x \mid x \in G,\, c^T x \leq \widehat{t}_1,\, f^T x \leq \widehat{t}_2,\, d^T x \leq \widehat{t}_3\} =$$

$$= \widehat{t}_3 - \min\{d^T x \mid x \in G, c^T x \leq \widehat{t}_1,\, f^T x \leq \widehat{t}_2\} = \widehat{t}_3 - \min_{x\in G_{\rm ini}} d^T x \geq$$

[since $\widehat{t}_3 \geq d^T\widehat{x}$ and $W \subset G_{\rm ini}$]

$$\geq d^T\widehat{x} - \min_{x\in W} d^T x$$

and applying (14.108) to $e = d$, we come to (14.106).

6^0. Let

$$\Omega_i = \frac{t_1^i - t_1^*(t^i)}{(t_2^i)^{32}}.$$

Our key argument is as follows: for properly chosen $O(1)$ and all i one has

$$\Omega_i/\Omega_{i-1} \geq 1 + O(1)\omega_*\vartheta^{-1/2}. \tag{14.109}$$

To establish the inequality, let us fix i and consider separately the cases of good and bad (t^{i-1}, x^{i-1}).

6^0.1. Assume that (t^{i-1}, x^{i-1}) is bad. According to (14.99), in the case in question

$$t_1^i = t_1^{i-1} + r_i\Delta_1(t^{i-1}, x^{i-1}), \quad t_2^i = t_2^{i-1}, \quad t_3^i \geq t_3^{i-1}.$$

Since $t_1^*(t)$ clearly depends only on t_2, t_3 and is nonincreasing in (t_2, t_3), we have $t_1^*(t^i) \leq t_1^*(t^{i-1})$, whence, in view of $t_2^i = t_2^{i-1}$,

$$\frac{\Omega_i}{\Omega_{i-1}} = \frac{t_1^i - t_1^*(t^i)}{t_1^{i-1} - t_1^*(t^{i-1})} \geq \frac{t_1^{i-1} + r_i \Delta_1(t^{i-1}, x^{i-1}) - t_1^*(t^{i-1})}{t_1^{i-1} - t_1^*(t^{i-1})} = 1 + r_i \frac{\Delta_1(t^{i-1}, x^{i-1})}{t_1^{i-1} - t_1^*(t^{i-1})}.$$

According to Corollary 14.4.1 applied with $k = 3$ (see (14.85)), the concluding quantity is $\geq 1 + \frac{1}{8} r_i$. Taking into account (14.103), we come to (14.109).

6^0.2. Now assume that (t^{i-1}, x^{i-1}) is good. For the sake of brevity, let us write (t, x) instead of (t^{i-1}, x^{i-1}), r instead of r_i, and let $t^+ = t^i$. By definition of $t_1^*(\cdot)$, there exists $u \in G$ such that

$$c^T u = t_1^*(t); \quad f^T u \leq t_2; \quad d^T u \leq t_3; \tag{14.110}$$

by definition of $\Delta_j(\cdot, \cdot)$ and since $x \in G_t$, we have

$$c^T x = t_1 - \Delta_1(t, x); \quad f^T x = t_2 - \Delta_2(t, x); \quad d^T x \leq t_3, \tag{14.111}$$

and since (t, x) is good, we have

$$t_2 \leq 16 \Delta_2(t, x). \tag{14.112}$$

Let

$$v = (1 - r) u + r x,$$

and let us verify that $v \in G_{t^+}$. Indeed, by (14.110) - (14.111) and due to $t_3^+ \geq t_3$ we have $d^T v \leq t_3^+$. Further, by (14.110), (14.111) and (14.99)

$$f^T v = (1 - r) f^T u + r f^T x \leq (1 - r) t_2 + r(t_2 - \Delta_2(t, x)) = t_2 - r \Delta_2(t, x) = t_2^+.$$

Last, by (14.110) and (14.111)

$$c^T v = (1 - r) c^T u + r c^T x \leq c^T x = t_1 - \Delta_1(t, x) \leq t_1^+$$

(concluding inequality is given by (14.100)).

Thus, $v \in G_{t^+}$; consequently,

$$t_1^*(t^+) \leq c^T v = (1 - r) c^T u + r c^T x = (1 - r) t_1^*(t) + r(t_1 - \Delta_1(t, x)). \tag{14.113}$$

With this inequality, we have

$$t_1^+ - t_1^*(t^+) = t_1 - r \Delta_1(t, x) - t_1^*(t^+) \geq t_1 - r \Delta_1(t, x) - (1 - r) t_1^*(t) - r(t_1 - \Delta_1(t, x)) =$$

$$= (1 - r)(t_1 - t_1^*(t)),$$

whence

$$\frac{\Omega_i}{\Omega_{i-1}} \geq (1 - r) \left(\frac{t_2}{t_2^+} \right)^{32} = (1 - r) \left(\frac{t_2}{t_2 - r \Delta_2(t, x)} \right)^{32} \geq$$

[see (14.112)]

$$\geq (1 - r) \left(\frac{1}{1 - \frac{r}{16}} \right)^{32} \geq (1 - r) \left(1 + \frac{r}{16} \right)^{32} \geq (1 - r)(1 + 2r) \geq 1 + \frac{3}{4} r$$

(we have taken into account that $r \leq 1/8$, see (14.103)); this inequality combined with (14.103) results in (14.109).

7^0. Let (P') be infeasible, and let us prove (14.101). Indeed, we have $t_2^i \geq f^* > 0$, whence $\Omega_i \leq \Delta_1(t^i, x^i)(f^*)^{-32}$. We conclude from (14.109) that

$$\Delta_1(t^i, x^i) \geq (1 + O(1) \omega_* \vartheta^{-1/2})^i \Delta_1(\widehat{t}, \widehat{x}) \left(\frac{f^*}{\Delta_2(\widehat{t}, \widehat{x})} \right)^{32}.$$

Taking into account that $\Delta_2(\widehat{t},\widehat{x}) \le \mathrm{V}_G(f)$ due to $\widehat{t}_2 \le \max_{x\in G} f^T x$ (see (14.89)) and that

$$\Delta_1(\widehat{t},\widehat{x}) \ge (14\vartheta[G\!:\!\widehat{x}])^{-1}\mathrm{V}_G(c)$$

(apply (14.108) to $e = c$ and note that $c^T\widehat{x} - \min_{x\in W} c^T x = \widehat{t}_1 - c^T\widehat{x}$ by (14.88)), we come to

$$\Delta_1(t^i,x^i) \ge (1+O(1)\omega_*\vartheta^{-1/2})^i \frac{\mathrm{V}_G(c)}{14\vartheta[G\!:\!\widehat{x}]}\left(\frac{f^*}{\mathrm{V}_G(f)}\right)^{32}.$$

It follows that

$$i \ge i_c = O(1)\omega_*^{-1}\sqrt{\vartheta}\ln\left(\frac{2\vartheta[G\!:\!\widehat{x}]\mathrm{V}_G(f)}{f^*}\right) \Rightarrow \Delta_1(t^i,x^i) \ge 5\vartheta_*\mathrm{V}_G(c). \qquad (14.114)$$

Further, from (14.85), (14.104) and (14.106) we have

$$\Delta_3(t^i,x^i) \ge \frac{1}{8}(t_3^i - t_3^*(t_i)) \ge \frac{1}{8}(1+O(1)\omega_*\vartheta^{-1/2})^i(\widehat{t}_3 - t_3^*(\widehat{t})) \ge$$
$$\ge (1+O(1)\omega_*\vartheta^{-1/2})^i(112\vartheta[G\!:\!\widehat{x}])^{-1}\mathrm{V}_G(d),$$

whence

$$i \ge i_d = O(1)\omega_*^{-1}\sqrt{\vartheta}\ln(2\vartheta[G\!:\!\widehat{x}]) \Rightarrow \Delta_3(t^i,x^i) \ge 5\vartheta_*\mathrm{V}_G(d). \qquad (14.115)$$

Last, $t_2^i \le t_2^{i-1}$ for all i, and if (t^{i-1},x^{i-1}) is good, then

$$t_2^i = (1 - r_i\Delta_2(t^{i-1},x^{i-1})/t_2^{i-1})t_2^{i-1} \ge (1-r_i/16)t_2^{i-1} \ge (1-O(1)\omega_*\vartheta^{-1/2})t_2^{i-1}$$

(we have used (14.103)). We clearly have $t_2^i \ge f^*$, while $\widehat{t}_2 \le f^* + \mathrm{V}_G(f)$ by (14.89). Combining these observations, we see that the total number i_f of those i with good (t^{i-1},x^{i-1}) can be bounded as follows:

$$i_f \le O(1)\omega_*^{-1}\sqrt{\vartheta}\ln\left(\frac{2\vartheta(\mathrm{V}_G(f)+f^*)}{f^*}\right). \qquad (14.116)$$

From (14.114), (14.115) and (14.116) we conclude that there exists

$$i^+ \le O(1)\omega_*^{-1}\sqrt{\vartheta}\ln\left(\frac{2\vartheta[G\!:\!\widehat{x}](\mathrm{V}_G(f)+f^*)}{f^*}\right)$$

such that (t^{i^+-1},x^{i^+-1}) is bad and

$$\Delta_1(t^{i^+-1},x^{i^+-1}) \ge 5\vartheta_*\mathrm{V}_G(c), \quad \Delta_3(t^{i^+-1},x^{i^+-1}) \ge 5\vartheta_*\mathrm{V}_G(d). \qquad (14.117)$$

Now let $H = \nabla_x^2 F(t^{i^+-1},x^{i^+-1})$. The Dikin ellipsoid $\{x \mid |x - x^{i^+-1}|_H \le 1\}$ is contained in G, so that the variation $2\sqrt{e^T H^{-1} e}$ of a linear form $e^T x$ on the ellipsoid is $\le \mathrm{V}_G(e)$. Thus, (14.117) implies (14.97) and (14.98); since (t^{i^+-1},x^{i^+-1}) is bad, we see that the Infeasibility Test detects infeasibility at the iteration i^+, and (14.101) follows.

8^0. It remains to consider the case when (P') is feasible, as it is assumed from now on. We need the following observation:

Lemma 14.5.2 *Let i be such that $t_3^{i-1} \ge \max_{x\in G} d^T x$. If (t^{i-1},x^{i-1}) is bad, then both t_1^{i-1} and t_1^i are $\le c^*$.*

Proof. For the sake of brevity, let us write (t,x) instead of (t^{i-1},x^{i-1}) and t^+ instead of t^i. Since $t_3 \ge \max_{x\in G} d^T x$, we have

$$t_2^*(\tau,\tau',t_3) = \phi(\tau) \equiv \min\{f^T x \mid x \in G, c^T x \le \tau\}, \ \ (\tau,\tau',t_3) \in T.$$

Since G is compact convex set, $\phi(\tau)$ is continuous nonincreasing convex function on the ray $[\min_{x\in G} c^T x, \infty)$; this function is positive to the left of c^* and is zero to the right of c^* (recall that since (P') is feasible, we have $\min_{x\in G} f^T x = 0$ and $c^* = \min_{x\in G: f^T x\le 0} c^T x\})$.

By Corollary 14.4.1 and since (t,x) is bad, we have

$$16\Delta_2(t,x) < t_2 = t_2^*(t) + [t_2 - t_2^*(t)] \le t_2^*(t) + 8\Delta_2(t,x) = \phi(t_1) + 8\Delta_2(t,x),$$

whence

$$\phi(t_1) \ge \frac{1}{2}t_2 > 8\Delta_2(t,x) > 0. \tag{14.118}$$

Consequently, $t_1 < c^*$. Besides this,

$$c^Tx = t_1 - \Delta_1(t,x), \quad f^Tx = t_2 - \Delta_2(t,x),$$

whence

$$\phi(t_1 - \Delta_1(t,x)) \le t_2 - \Delta_2(t,x).$$

Since ϕ is convex, we have

$$\phi(t_1 + \Delta_1(t,x)) \ge 2\phi(t_1) - \phi(t_1 - \Delta_1(t,x)) \ge 2\phi(t_1) + \Delta_2(t,x) - t_2 \ge$$

[see (14.118)]

$$\ge \Delta_2(t,x) > 0,$$

whence $t_1 + \Delta_1(t,x) < c^*$. It remains to note that in view of (14.99), (14.103)

$$t_1^+ = t_1 + r_i\Delta_1(t,x) < t_1 + \Delta_1(t,x). \quad \blacksquare$$

9^0. According to Lemma 14.5.2, the behaviour of the sequence $\{t_1^i\}$ starting with the moment i^* is as follows: if (t^i,x^i) is bad, then both t_1^i and t_1^{i+1} are less than c^*, and if (t^i,x^i) is good, then, by (14.99), $t_1^{i+1} < t_1^i$. Consequently, the sequence

$$\{\varepsilon_i = \max[t_1^i - c^*, 0]\}_{i \ge i^*}$$

is nonincreasing and

$$\phi_i \equiv t_1^i - \min_{x\in G} c^Tx \le \phi^* \equiv \max[c^*, t_1^{i^*}] - \min_{x\in G} c^Tx, \ \ i \ge i^*. \tag{14.119}$$

Since one clearly has $t_1^i - t_1^*(t^i) \le \phi_i$, we come to

$$t_1^i - t_1^*(t^i) \le \phi^*, \ \ i \ge i^*. \tag{14.120}$$

By definition of Ω_i, for $i \ge i^*$ we have

$$f^Tx^i \le t_2^i = \left[\frac{\Omega_0}{\Omega_i}\,\frac{t_1^i - t_1^*(t^i)}{\widehat{t}_1 - t_1^*(\widehat{t})}\right]^{1/32} \widehat{t}_2 \le$$

[see (14.109), (14.89) and (14.120)]

$$\le (1 - O(1)\omega_*\vartheta^{-1/2})^i \left[\frac{\phi^*}{\widehat{t}_1 - t_1^*(\widehat{t})}\right]^{1/32} \max_{x\in G} f^Tx.$$

Applying (14.108) to $e = c$ and taking into account that by definition of $\widehat{t}_1$ one has

$$\widehat{t}_1 - t_1^*(\widehat{t}) \ge \widehat{t}_1 - c^T\widehat{x} = c^T\widehat{x} - \min_{x\in W} c^Tx,$$

we get

$$i \ge i^* \ \Rightarrow\ f^Tx^i \le (1 - O(1)\omega_*\vartheta^{-1/2})^i(\phi^{**})^{1/32}\max_{x\in G} f^Tx, \quad \phi^{**} = 14\vartheta[G\,{:}\,\widehat{x}]\frac{\phi^*}{\mathrm{V}_G(c)}. \tag{14.121}$$

10^0. Let us prove that

$$\phi^* \le 20\vartheta[G\,{:}\,\widehat{x}]\mathrm{V}_G(c). \tag{14.122}$$

Let $d_i = t_3^i - t_3^*(t^i)$. By (14.104) we have $d_i \geq (1+r_i)d_{i-1}$, while (see (14.119), (14.99))

$$\frac{\phi_i}{\phi_{i-1}} \leq 1 + r_i \frac{\Delta_1(t^{i-1}, x^{i-1})}{\phi_{i-1}} \leq 1 + r_i$$

(we have taken into account that $\phi_{i-1} > \Delta_1(t^{i-1}, x^{i-1}) = t_1^{i-1} - c^T x^{i-1}$). Consequently,

$$\phi_i \leq \phi_0 \frac{d_i}{d_0} \leq \mathrm{V}_G(c)\frac{d_i}{d_0}$$

(the second inequality is given by (14.89)). In the case of $i^* = 0$ the resulting relation clearly implies $\phi_{i^*} \leq \mathrm{V}_G(c)$. If $i^* > 0$, then the above inequalities lead to

$$\phi_{i^*} \leq (1 + r_{i^*})\phi_{i^*-1} \leq (1 + r_{i^*})\mathrm{V}_G(c)\frac{d_{i^*-1}}{d_0} \leq$$

[see (14.103) and note that $d_{i^*-1} \leq \mathrm{V}_G(d)$ by definition of i^*, while $d_0 \geq (14\vartheta[G\!:\!\widehat{x}])^{-1}\mathrm{V}_G(d)$ by (14.106)]

$$\leq 20\vartheta[G\!:\!\widehat{x}]\mathrm{V}_G(c).$$

To get (14.122), it remains to note that, by definition,

$$\phi^* = \max[c^*, t_1^{i^*}] - \min_{x\in G} c^T x = \max[c^* - \min_{x \in G} c^T x, \phi_{i^*}] \leq \max[\mathrm{V}_G(c), \phi_{i^*}].$$

Combining (14.122) and (14.121). we come to

$$i \geq i^* \Rightarrow f^T x^i \leq O(1)\vartheta[G\!:\!x](1 - O(1)\zeta^*\vartheta^{-1/2})^i \max_{x\in G} f^T x. \tag{14.123}$$

11^0. Now let i^{**} be the first $i \geq i^*$ such that (t^i, x^i) is bad (if no such i exists, we set $i^{**} = +\infty$). As it was explained in the beginning of 9^0, we have

$$c^T x^i - c^* \leq \varepsilon_i \equiv \begin{cases} 0, & i \geq i^{**} \\ \max\left[t_1^i - c^*, 0\right], & i^* \leq i < i^{**} \end{cases} \tag{14.124}$$

According to Corollary 14.4.1,

$$t_1^{i-1} - t_1^*(t^{i-1}) \leq 8\Delta_1(t^{i-1}, x^{i-1});$$

when $i > i^*$, we have

$$t_1^*(t^{i-1}) = \min\{c^T x \mid x \in G, f^T x \leq t_2^{i-1}\} \leq c^*,$$

so that for $i^* < i < i^{**}$

$$t_1^i - c^* \leq t_1^{i-1} - r_i\Delta_1(t^{i-1}, x^{i-1}) - c^* \leq t_1^{i-1} - c^* - \frac{r_i}{8}(t_1^{i-1} - t_1^*(t^{i-1})) \leq \left(1 - \frac{r_i}{8}\right)(t_1^{i-1} - c^*),$$

whence, due to (14.103),

$$\varepsilon_i \leq (1 - O(1)\omega_*\vartheta^{-1/2})\varepsilon_{i-1}, \ \ i > i^*.$$

Combining this observation with (14.124), we come to

$$c^T x^i - c^* \leq (1 - O(1)\omega_*\vartheta^{-1/2})^{i-i^*}\varepsilon_{i^*}, \ \ i > i^*,$$

whence, in view of $\varepsilon_{i^*} \leq \phi^*$ and (14.122)

$$c^T x^i - c^* \leq O(1)\vartheta[G\!:\!\widehat{x}](1 - O(1)\omega_*\vartheta^{-1/2})^{i-i^*}\mathrm{V}_G(c), \ \ i \geq i^*. \tag{14.125}$$

Combining (14.125), (14.123) and taking into account (14.105), we come to (14.102). ■

14.6 Application examples

Our "long step" technique for tracing a surface of analytic centers heavily exploits assumption $\mathcal{Q}$ (Section 14.5.1) on structure of the s.-c.b. F for the domain G of problems (P), (P'); let us call barriers satisfying this assumption *good*. The goal of this section is to demonstrate that the s.-c.b.'s responsible for many important applications indeed are good.

14.6.1 Combination rules

Let us start from the following general remark. The desired structure is "stable with respect to intersections". Namely, assume that G is represented as an intersection $\cap_{i=1}^m G_i$ of closed convex domains; we shall say that G_i represents (or simply is) *i-th constraint of the problem.* The aforementioned stability means that if every G_i admits a good ϑ_i-s.-c.b. $F_i(x) = \Phi_i(\pi_i x + \epsilon_i)$ (so that Φ_i is a ϑ_i-s.-c.b. for certain closed convex domain $H_i \subset \mathbf{R}^{q_i}$ and we know the Legendre transformation $\Phi_{i,*}$ of Φ_i), then a good $(\sum_i \vartheta_i)$-s.c.b. for G is

$$F(x) = \Phi(\pi x + \epsilon),$$

where

$$\Phi(u_1, ..., u_m) = \sum_{i=1}^m \Phi_i(u_i) : \mathrm{int}\,(H_1 \times ... \times H_m) \to \mathbf{R}, \ \ \pi x + \epsilon = (\pi_1 x + \epsilon_1, ..., \pi_m x + \epsilon_m);$$

note that

$$\Phi_*(s_1, ..., s_m) = \sum_{i=1}^m \Phi_{i,*}(s_i)$$

(see Proposition 14.2.1).

Thus, our assumption is "separable with respect to the constraints" involved into the description of G.

The structure in question is also stable with respect to affine substitutions of argument: if G is the inverse image of certain closed convex domain G^+ under an affine mapping $\mathcal{A}$ (the image of the mapping intersects int G^+) and we know a good ϑ-s.-c.b. F^+ for G^+, then we can equip G with the ϑ-s.-c. barrier $F(x) = F^+(\mathcal{A}(x))$, and this barrier clearly is good.

14.6.2 "Building blocks"

The indicated combination rules can be applied to a number of "building blocks", i.e., good barriers for certain standard convex domains. These blocks are as follows:

1. *Nonnegative half-axis* $\mathbf{R}_+$: The standard 1-s.-c.b. $\Phi(x) = -\ln x$ for $\mathbf{R}^+$ is good: its Legendre transformation is $\Phi_*(s) = -\ln(-s) - 1$, $s < 0$; both Φ and Φ_* are 2-regular (regularity of all functions mentioned in this section is proved in Appendix B).

This elementary observation, in view of the combination rules, allows to handle arbitrary linear inequality constraints and, in particular, covers all needs of Linear Programming.

2. *Convex domain $G \subset \mathbf{R}^n$ which is a connectedness component of the Lebesgue set* $\mathrm{cl}\{x \mid f(x) < 0\}$ *of a quadratic function f*: Such a domain can be represented as the inverse image of the second-order cone

$$H = \{u \in \mathbf{R}^q \mid u_q \geq (\sum_{i=1}^{q-1} u_i^2)^{1/2}\}$$

under an easily computable affine mapping $x \mapsto u = \pi x + \epsilon$ with the image of the mapping intersecting the interior of H; setting

$$\Phi(u) = -\ln(u_q^2 - \sum_{i=1}^{q-1} u_i^2),$$

we obtain a 2-s.-c.b. for H ([3] , Chapter 5) with the explicit Legendre transformation

$$\Phi_*(s) = -\ln(s_q^2 - \sum_{i=1}^{q-1} s_i^2) - 2 + 2\ln 2, \ s \in -H,$$

and consequently can equip G with the good 2-s.-c.b. $F(x) = \Phi(\pi x + \epsilon)$; both Φ and Φ_* are 6-regular.

This observation covers convex quadratically constrained problems and even more general family of convex programs (note that f should not necessarily be convex, e.g., we may handle the hyperbolic domain of the type $\sum_{i=1}^{n-1} x_i^2 + 1 \leq x_n^2$, $x_n > 0$).

3. *Geometrical Programming in the exponential form*: Assume that among the constraints defining the feasible domain of a convex program there is a constraint of the form

$$\sum_{i=1}^{q} \exp\{a_i^T x + b_i\} \leq p^T x + r.$$

Adding q extra variables $y_1, ..., y_q$, one may pass from the initial problem to an equivalent one where the indicated constraint is represented by the system of convex inequalities

$$\exp\{a_i^T x\} \leq y_i,\ i = 1, ..., q;\ \sum_{i=1}^{q} \exp\{b_i\} y_i \leq p^T x + r.$$

We already know how to handle the concluding linear constraint, and all we need is to understand how to deal with the exponential inequality

$$\exp\{a_i^T x\} \leq y_i.$$

In order to penalize the latter constraint by a good barrier, it suffices to point out a good barrier for the epigraph

$$G = \{(\tau, x) \in \mathbf{R}^2 \mid \tau \geq \exp\{x\}\}$$

of the exponent and to use the combination rule related to affine substitutions of argument. A good 2-s.-c.b. for G can be written down explicitly ([3] , Chapter 5):

$$\Phi(\tau, x) = -\ln(\ln \tau - x) - \ln t,\ \Phi_*(s, \xi) = (\xi + 1) \ln \left(\frac{\xi + 1}{-s} \right) - \xi - \ln \xi - 2,$$

$$\mathrm{Dom}\, \Phi_* = \{(s, \xi) \in \mathbf{R}^2 \mid s < 0, \xi > 0\}.$$

Both Φ and Φ_* turn out to be 6-regular.

4. *Linear Matrix Inequality constraint:* A constraint of this type arises in numerous applications and defines the domain of the form

$$G = \{x \mid \mathcal{A}(x) \text{ is positive semidefinite}\},$$

where $\mathcal{A}(x)$ is an affine in x matrix-valued function taking values in the space of symmetric matrices of a given row size m. A good m-s.-c.b. for G is given by

$$F(x) = \Phi(\mathcal{A}(x)),\ \Phi(y) = -\ln \mathrm{Det}\ y;$$

Φ is the standard m-s.-c.b. for the cone $\mathbf{S}_+^m$ of symmetric $m \times m$ positive semidefinite matrices (see [3] , Chapter 5) with the Legendre transformation

$$\Phi_*(s) = -\ln \mathrm{Det}\,(-s) - m,\ \mathrm{Dom}\, \Phi_* = -\,\mathrm{int}\ \mathbf{S}_+^m;$$

both Φ and Φ_* are 2-regular. This example covers all needs of Semidefinite Programming.

We see that our assumption on the structure of F is compatible with a wide spectrum of important Convex Programming problems.

Appendix A: Proof of Lemma 14.5.1

Proof of Lemma 14.5.1. The function $t_k^*(\cdot)$ clearly depends on the first $k-1$ components of argument only; let us denote the vector comprised of these first $k-1$ components by τ, so that $t_k^* = t^*(\tau)$. Since G is bounded, $t^*(\tau)$ is a convex continuous function on the closure T' of the projection of T on the plane of the first $k-1$ parameters; this projection is monotone ($\tau' \geq \tau \in T' \Rightarrow \tau' \in T'$), and $t^*(\tau)$ is monotonically nonincreasing on T'.

Let

$$\tau = (t_1, ..., t_{k-1}), \quad \delta\tau = ((\delta t_1)_-, (\delta t_2)_-, ..., (\delta t_{k-1})_-),$$

where $a_- = \min\{a, 0\}$, and let $\Delta\tau = (\delta t_1 - \delta\tau_1, ..., \delta t_{k-1} - \delta\tau_{k-1})$; then $\Delta\tau \geq 0$. It is possible that $\delta\tau = 0$; then the statement in question is evident, since here

$$t_k^*(t + r\delta t) = t^*(\tau + r\Delta\tau) \leq t^*(\tau) = t_k^*(t)$$

(we have used the monotonicity of $t^*(\cdot)$), so that

$$t_k + r\delta t_k - t_k^*(t + r\delta t) \geq t_k + r\delta t_k - t_k^*(t) = (t_k - t^*(t))\left(1 + \frac{r\delta t_k}{t_k - t_k^*(t)}\right) \geq$$

$$\geq \left(1 + r\frac{\delta t_k}{2(k+1)\Delta_k(t,x)}\right)(t_k - t_k^*(t))$$

(the concluding inequality follows from (14.85)), and we come to (14.95).

Now consider the case when $\delta\tau \neq 0$. Let z be the largest real $p \geq 0$ such that

$$|p\delta\tau_i| \leq \Delta_i(t,x), \ i = 1, ..., k-1.$$

As we know from Corollary 14.4.1, one has $(\tau + z\delta\tau, t_k - \Delta_k(t,x)) \in \operatorname{cl} T$, and consequently

$$t^*(\tau + z\delta\tau) \leq t_k - \Delta_k(t,x). \tag{14.126}$$

Let, further,

$$\widehat{r} = \min\{r, z\}, \ \theta = \widehat{r}/z;$$

since t^* is convex, we have

$$t^*(\tau + \widehat{r}\delta\tau) \leq t^*(\tau) + \theta(t^*(\tau + z\delta\tau) - t^*(\tau)) \leq t^*(\tau) + \theta[t_k - \Delta_k(t,x) - (t_k - 2(k+1)\Delta_k(t,x))]$$

(the latter inequality follows from (14.126) and (14.85)). Thus, we come to

$$t^*(\tau + \widehat{r}\delta\tau) \leq t^*(\tau) + (2k+1)\theta\Delta_k(t,x),$$

and since $t^*(\cdot)$ is monotonically nonincreasing, we have also

$$t_k^*(t + \widehat{r}\delta t) = t^*(\tau + \widehat{r}\delta\tau + \widehat{r}\Delta\tau) \leq t^*(\tau) + (2k+1)\theta\Delta_k(t,x),$$

or

$$t_k^*(t + \widehat{r}\delta t) \leq t_k^*(t) + (2k+1)\theta\Delta_k(t,x). \tag{14.127}$$

On the other hand, by definition of z one has $z|\delta t_i| = \Delta_i(t,x)$ for certain $i < k$ with $\delta t_i < 0$; since, by assumption, δta k-safe,

$$\frac{\delta t_k}{\Delta_k(t,x)} \geq (4k+2)\frac{|\delta t_i|}{\Delta_i(t,x)},$$

and we come to

$$z\delta t_k \geq (4k+2)z|\delta t_i|\Delta_k(t,x)\Delta_i^{-1}(t,x) = (4k+2)\Delta_k(t,x), \tag{14.128}$$

whence

$$\Delta_k(t,x) \leq (4k+2)^{-1}z\delta t_k.$$

Combining this inequality with (14.127), we come to

$$(t_k + \widehat{r}\delta t_k) - t_k^*(t + \widehat{r}\delta t) \geq \widehat{r}\delta t_k + (t_k - t_k^*(t)) - \theta z(4k+2)^{-1}(2k+1)\delta t_k =$$

[since $\theta z = \widehat{r}$]

$$= \frac{\widehat{r}}{2}\delta t_k + (t_k - t_k^*(t)),$$

whence, again in view of $t_k - t_k^*(t) \leq 2(k+1)\Delta_k(t,x)$,

$$(t_k + \widehat{r}\delta t_k) - t_k^*(t + \widehat{r}\delta t) \geq (1 + \frac{\widehat{r}\delta t_k}{4(k+1)\Delta_k(t,x)})(t_k - t_k^*(t)). \tag{14.129}$$

It is possible that $\widehat{r} = r$; in this case (14.95) immediately follows from (14.129). It remains to consider the case when $\widehat{r} < r$. By definition of $\widehat{r}$, it means that $\widehat{r} = z < r$, and in view of (14.128) relation (14.129) implies that

$$\Delta \equiv (t_k + \widehat{r}\delta t_k) - t_k^*(t + \widehat{r}\delta t) \geq (1 + \frac{4k+2}{4k+4})(t_k - t_k^*(t)) = \frac{4k+3}{2k+2}(t_k - t_k^*(t)). \tag{14.130}$$

Now consider the function

$$g(l) = (t_k + lr\delta t_k) - t_k^*(t + lr\delta t);$$

since r is a safe stepsize, it is a well-defined concave nonnegative function on the segment $[0, 1/q]$; the value of this function at the point $\widehat{l} = \widehat{r}/r \leq 1$ is, as we know from (14.130), $\geq (4k+3)(2k+2)^{-1}g(0)$, and from concavity it follows that

$$g(1) \geq \frac{q^{-1} - 1}{q^{-1} - \widehat{l}} g(\widehat{l}) \geq (1-q)g(\widehat{l}),$$

whence

$$t_k + r\delta t_k - t_k^*(t + r\delta t) = g(1) \geq (1-q)\frac{4k+3}{2k+2}g(0) = (1-q)\frac{4k+3}{2k+2}(t_k - t_k^*(t)),$$

as required in (14.95). ∎

Appendix B: Regularity of some self-concordant functions

Example 14.6.1 *The function*

$$f(X) = -\ln \mathrm{Det}\ X$$

is 2-regular on the interior $\mathbf{S}^n_{++}$ *of the cone* $\mathbf{S}^n_+$ *of* $n \times n$ *positive semidefinite symmetric matrices.*

Proof. Let $X \in \mathbf{S}^n_{++}$, and let dX be a symmetric $n \times n$ matrix. Direct computation results in

$$\begin{array}{rcl} D^k f(X)[dX, ..., dX] & = & (-1)^k (k-1)! \,\mathrm{Tr}\left[(X^{-1}dX)^k\right] \\ & = & (-1)^k (k-1)! \,\mathrm{Tr}\left[X^{-1}dXX^{-1}dX...X^{-1}dX\right] \\ & = & (-1)^k (k-1)! \,\mathrm{Tr}\left[X^{-1/2}dXX^{-1}...X^{-1}dXX^{-1/2}\right] \\ & = & (-1)^k (k-1)! \,\mathrm{Tr}\left[(dZ)^k\right], \quad dZ = X^{-1/2}dXX^{-1/2}. \end{array}$$

Since the mapping $Y \mapsto X^{-1/2}YX^{-1/2}$ is a linear one-to-one mapping of the space $\mathbf{S}^n$ of symmetric $n \times n$ matrices onto itself which maps X onto the unit matrix I, maps dX onto dZ and maps $Q \equiv \mathbf{S}^n_{++}$ onto itself, we have

$$|dX|_{Q,X} = |dZ|_{Q,I}.$$

Now let $\lambda = (\lambda_1, ..., \lambda_k)$ be the vector comprised of the eigenvalues of dZ. We have for $k \geq 2$:

$$\begin{array}{rcl} |D^k f(X)[dX, dX, ..., dX]| & = & (k-1)!|\,\mathrm{Tr}\left[(dZ)^k\right]| \\ & = & (k-1)!|\sum_{i=1}^n \lambda_i^k| \\ & \leq & (k-1)!|\lambda|_\infty^{k-2}|\lambda|_2^2 = (k-1)!|\lambda|_\infty^{k-2}\,\mathrm{Tr}\left[(dZ)^2\right] \\ & = & (k-1)!|\lambda|_\infty^{k-2} D^2 f(X)[dX, dX]. \end{array}$$

It remains to note that by evident reasons,

$$[|dX|_{Q,X} =] \quad |dZ|_{Q,I} = [\max\{t \mid 1 \pm t\lambda_i \geq 0\}]^{-1} = |\lambda|_\infty. \qquad ∎$$

Example 14.6.2 *The standard logarithmic barrier*

$$f(t, x) = -\ln(t^2 - x^T x)$$

for the second-order cone

$$\mathbf{K}^n = \{(t, x) \in \mathbf{R} \times \mathbf{R}^n \mid t \geq \sqrt{x^T x}\}$$

is 6-regular on int $\mathbf{K}^n$.

Proof. Let $g(x,t) = -t^{-1}x^T x$, so that

$$f(t,x) = -\ln(t + g(t,x)) - \ln t.$$

Let $z = (t,x) \in \operatorname{int} \mathbf{K}^n$, and let $dz = (dt, dx)$ be such that $z \pm dz \in \mathbf{K}^n$; what we should prove is the inequality

$$|D^4 f(z)[dz,dz,dz,dz]| \le 42 D^2 f(z)[dz,dz]. \tag{14.131}$$

Denoting for the sake of brevity

$$d^k f = D^k f(z)[dz,...,dz], \quad d^k g = D^k g(z)[dz,...,dz], \quad \tau = \frac{dt:}{t},$$

we have

$$\begin{array}{rcl} df & = & -\frac{dt+dg}{t+g} - \tau, \\ d^2 f & = & \frac{(dt+dg)^2}{(t+g)^2} - \frac{d^2 g}{t+g} + \tau^2, \\ d^3 f & = & -2\frac{(dt+dg)^3}{(t+g)^3} + 3\frac{(dt+dg)d^2 g}{(t+g)^2} - \frac{d^3 g}{t+g} - 2\tau^3, \\ d^4 f & = & 6\frac{(dt+dg)^4}{(t+g)^4} - 12\frac{(dt+dg)^2 d^2 g}{(t+g)^3} + 3\frac{(d^2 g)^2}{(t+g)^2} \\ & & +4\frac{(dt+dg)d^3 g}{(t+g)^2} - \frac{d^4 g}{t+g} + 6\tau^4, \end{array}$$

and

$$\begin{array}{rcl} dg & = & -2\frac{(dx)^T x}{t} + \frac{x^T x}{t}\tau, \\ d^2 g & = & -2\frac{(dx)^T dx}{t} + 4\frac{(dx)^T x}{t}\tau - 2\frac{x^T x}{t}\tau^2 \\ & = & -2\frac{w^T w}{t}, \\ w & \equiv & \tau x - dx. \end{array}$$

We have

$$dw = -\tau^2 x + \tau dx = -\tau w,$$

whence

$$\begin{array}{rcl} d^3 g & = & +4\frac{w^T w}{t}\tau + 2\frac{w^T w}{t}\tau \\ & = & 6\frac{w^T w}{t}\tau, \\ d^4 g & = & -12\frac{w^T w}{t}\tau^2 - 6\frac{w^T w}{t}\tau^2 - 6\frac{w^T w}{t}\tau^2 \\ & = & -24\frac{w^T w}{t}\tau^2. \end{array}$$

Denoting

$$\phi = \frac{dt + dg}{t + g},$$

we come to

$$d^2 f = \phi^2 + \tau^2 + 2\frac{w^T w}{t(t+g)}, \tag{14.132}$$

$$d^4 f = 6\phi^4 + 6\tau^4 + 2\frac{w^T w}{t(t+g)}\left[12\phi^2 + 12\phi\tau + 12\tau^2 + 6\frac{w^T w}{t(t+g)}\right]. \tag{14.133}$$

Since $z \pm dz \in \mathbf{K}^n$, we have $t \pm dt \ge 0$, whence $|\tau| \le 1$. Now, the function $h(t,x) = t + g(t,x)$ clearly is concave in (t,x) on the set $\{(t,x) \mid t > 0\}$. Since $z \pm dz \in \mathbf{K}^n$, the function is nonnegative at the points $z \pm dz$, whence

$$t + g \pm (dt + dg) \ge h(z \pm dz) \ge 0,$$

so that

$$|\phi| \le 1; \quad |\tau| \le 1. \tag{14.134}$$

Now let us verify that

$$\frac{w^T w}{t(t+g)} \le 1 - \tau^2. \tag{14.135}$$

Indeed, since $z \pm dz \in \mathbf{K}^n$, we have

$$\begin{array}{rcl} x^T x + 2x^T dx + (dx)^T dx & \le & (t+dt)^2 \equiv t^2(1+\tau)^2 \\ x^T x - 2x^T dx + (dx)^T dx & \le & (t-dt)^2 \equiv t^2(1-\tau)^2 \end{array};$$

multiplying the first inequality by $(1-\tau)/2$, the second – by $(1+\tau)/2$ and taking sum of the resulting inequalities, we come to

$$x^Tx - 2\tau x^T dx + (dx)^T dx \le t^2(1-\tau^2),$$

or

$$w^Tw \equiv \tau^2 x^Tx - 2\tau x^T dx + (dx)^T dx \le (t^2 - x^Tx)(1-\tau^2) = t(t+g)(1-\tau^2),$$

as required in (14.135).

From (14.133), (14.134) and (14.135) it follows that quantity

$$[0 \le] \quad 12\phi^2 + 12\phi\tau + 12\tau^2 \le 12\phi^2 + 12\phi\tau + 12\tau^2 + 6\frac{w^Tw}{t(t+g)} \le$$

$$\le 12\phi^2 + 12\phi\tau + 6\tau^2 + 6 \le 36,$$

so that (14.132) and (14.133) result in (14.131). ■

Example 14.6.3 *The standard barrier*

$$F(t,s) = -\ln(\ln t - s) - \ln s$$

for the epigraph

$$G = \{x = (t,s) \in \mathbf{R}^2 \mid t \ge \exp\{s\}\}$$

of the exponent is 6-regular on $Q = \operatorname{int} G$.

Proof. Let $x = (t,s) \in Q$, and let $dx = (dt, ds) \in \mathbf{R}^2$. We should prove that

$$|D^4F(x)[dx,dx,dx,dx]| \le 42 D^2F(x)[dx,dx]|dx|^2_{Q,x},$$

or, which is the same, that

$$x \pm dx \in Q \Rightarrow |D^4F(x)[dx,dx,dx,dx]| \le 42 D^2[dx,dx]. \tag{14.136}$$

Let (x, dx) satisfy the premise in (14.136). Direct computation results in

$$\begin{array}{rcl} D^2F(x)[dx,dx] & = & \frac{(ds - dt/t)^2}{(\ln t - s)^2} + \frac{(dt/t)^2}{\ln t - s} + (dt/t)^2, \\ D^4F(x)[dx,dx,dx,dx] & = & 6\frac{(ds-dt/t)^4}{(\ln t - s)^4} + 12\frac{(ds-dt/t)^2(dt/t)^2}{(\ln t - s)^3} \\ & & -8\frac{(ds-dt/t)(dt/t)^3}{(\ln t - s)^2} + 3\frac{(dt/t)^4}{(\ln t - s)^2} + 6\frac{(dt/t)^4}{\ln t - s} + 6(dt/t)^4. \end{array}$$

Setting

$$\sigma = \ln t - s > 0,\ d\sigma = ds,\ d\tau = dt/t,$$

we get from the premise in (14.136) that

$$\exp\{-\sigma + d\sigma\} \le 1 + d\tau, \exp\{-\sigma - d\sigma\} \le 1 - d\tau, \tag{14.137}$$

$$d^2F \equiv D^2F(x)[dx,dx] = \frac{(d\sigma - d\tau)^2}{\sigma^2} + \frac{(d\tau)^2}{\sigma} + (d\tau)^2, \tag{14.138}$$

$$\begin{array}{rcl} |d^4F| \equiv |D^4F(x)[dx,dx,dx,dx]| & \le & 6\frac{(d\sigma - d\tau)^4}{\sigma^4} + 12\frac{(d\sigma-d\tau)^2(d\tau)^2}{\sigma^3} + 8\frac{|d\sigma - d\tau|(d\tau)^3}{\sigma^2} \\ & & +3\frac{(d\tau)^4}{\sigma^2} + 6\frac{(d\tau)^4}{\sigma} + 6(d\tau)^4 \\ & \equiv & 6a_1 + 12a_2 + 8a_3 + 3a_4 + 6a_5 + 6a_6. \end{array} \tag{14.139}$$

What we should prove is that

$$|d^4F| \le 42 d^2F. \tag{14.140}$$

When proving this statement, one can assume that $d\tau \geq 0$ (since both the premise (14.137) and the quantities d^2F and $|d^4F$ remain invariant under the substitution $(d\sigma, d\tau) \mapsto (-d\sigma, -d\tau)$).
1^0. Since $\exp\{u\} \geq 1+u$, we get from (14.137) $-\sigma + d\sigma \leq d\tau$, $\ -\sigma - d\sigma \leq -d\tau$, whence

$$|d\sigma - d\tau| \leq \sigma. \tag{14.141}$$

Besides this, from (14.137) and our assumption that $d\tau \geq 0$ it of course follows

$$0 \leq d\tau \leq 1. \tag{14.142}$$

2^0. From (14.141)-(14.142) it immediately follows that

$$\begin{array}{lcccc}
a_1 & = & \frac{(d\sigma - d\tau)^4}{\sigma^4} & \leq & \frac{(d\sigma - d\tau)^2}{\sigma^2} \\
a_2 & = & \frac{(d\sigma - d\tau)^2 (d\tau)^2}{\sigma^3} & \leq & \frac{(d\tau)^2}{\sigma} \\
a_3 & = & \frac{|d\sigma - d\tau| (d\tau)^2}{\sigma^2} & \leq & \frac{(d\tau)^2}{\sigma} \\
a_5 & = & \frac{(d\tau)^4}{\sigma} & \leq & \frac{(d\tau)^2}{\sigma} \\
a_6 & = & (d\tau)^4 & \leq & (d\tau)^2
\end{array}$$

whence

$$|d^4F| \leq [6a_1 + 12a_2 + 8a_3 + 6a_5 + 6a_6] + 3a_4 \leq 6\frac{(d\sigma - d\tau)^2}{\sigma^2} + 24\frac{(d\tau)^2}{\sigma} + 6(d\tau)^2 + 3a_4. \tag{14.143}$$

3^0. We have

$$a_4 = \frac{(d\tau)^4}{\sigma^2} \leq \frac{(d\tau)^2}{\sigma}\omega, \quad \omega = \frac{(d\tau)^2}{\sigma}. \tag{14.144}$$

Let us prove that

$$\omega \leq 18. \tag{14.145}$$

To this end let us first assume that

$$2|d\sigma - d\tau| \geq d\tau.$$

In this case, in view of (14.141),

$$\omega \leq \frac{2|d\sigma - d\tau| d\tau}{\sigma} \leq 2d\tau,$$

so that, by (14.142),

$$\omega \leq 2. \tag{14.146}$$

Note also that (14.146) is valid in the case of $\sigma > \frac{1}{2}$ in view of (14.142). Now consider the case when

$$0 < \sigma \leq \frac{1}{2}, \quad 2|d\sigma - d\tau| < d\tau,$$

whence also $2d\tau - 2d\sigma < d\tau$, or

$$d\tau \leq 2d\sigma. \tag{14.147}$$

From (14.137) (take the sum of the inequalities) it follows that

$$\cosh\{d\sigma\} \leq \exp\{\sigma\}, \tag{14.148}$$

whence, due to $0 < \sigma \leq 1/2$, also

$$|d\sigma| \leq \ln(2\exp\{\sigma\}) \leq \frac{1}{2} + \ln 2.$$

On the segment $|r| < 1/2 + \ln 2$ one clearly has

$$\cosh\{r\} \geq \exp\{\gamma r^2\}, \ \ \gamma = \frac{1}{2(1/2 + \ln 2)^2} \geq \frac{2}{9}.$$

Combining the resulting inequality with (14.148), we get $\gamma|d\sigma|^2 \leq \sigma$, whence $|d\sigma|^2 \leq \frac{9}{2}\sigma$ and consequently (see (14.147)) $(d\tau)^2 \leq 18\sigma$; The resulting inequality implies that in the case in question $\omega = (d\tau)^2/\sigma \leq 18$; taking into account (14.146), we come to (14.145).
4^0. Combining (14.145), (14.144) and (14.143), we get

$$|d^4F| \leq 6\frac{(d\sigma - d\tau)^2}{\sigma^2} + 42\frac{(d\tau)^2}{\sigma} + 6(d\tau)^2 \leq 42d^2F,$$

and (14.140) follows. ■

Example 14.6.4 *The Legendre transformation*

$$F_*(\tau,\sigma) \equiv \sup_{t,s:t\geq\exp\{s\}} [s\sigma - t\tau + \ln(\ln t - s) + \ln t] \equiv$$

$$\equiv (\sigma+1)\ln(\sigma+1) - (\sigma+1)\ln\tau - \ln\sigma - 2 : \mathbf{R}^2_{++} \to \mathbf{R}$$

of the barrier from Example 14.6.3 is 5*-regular on* $Q \equiv \mathbf{R}^2_{++}$.

Proof. Let $\xi = (\tau,\sigma) \in Q$, $d\xi = (d\tau, d\sigma) \in \mathbf{R}^2$ be such that $\xi \pm d\xi \in Q$, i.e., such that

$$\tau \pm d\tau \geq 0, \quad \sigma \pm d\sigma \geq 0; \tag{14.149}$$

we should prove that

$$|d^4F_*| \equiv |D^4F_*(\xi)[d\xi,d\xi,d\xi,d\xi]| \leq 30d^2F_*, \quad d^2F_* \equiv D^2F_*(\xi)[d\xi,d\xi]. \tag{14.150}$$

Direct computation results in

$$d^2F_* = (1+\sigma)(d\omega)^2 + \frac{(d\sigma)^2}{\sigma^2}, \quad d\omega = \frac{d\sigma}{1+\sigma} - \frac{d\tau}{\tau}, \tag{14.151}$$

$$d^4F_* = 4(1+\sigma)(d\omega)^2(d\zeta)^2 + 2(1+\sigma)(d\omega)^3d\zeta - 3d\sigma(d\omega)^2d\zeta + 6\frac{(d\sigma)^4}{\sigma^4}, \quad d\zeta = \frac{d\sigma}{1+\sigma} + \frac{d\tau}{\tau}. \tag{14.152}$$

From (14.149) it follows that $|d\omega| \leq 2$, $|d\zeta| \leq 2$, $|d\sigma/\sigma| \leq 1$, and therefore (14.152) implies that

$$|d^4F_*| \leq 16(1+\sigma)(d\omega)^2 + 8(1+\sigma)(d\omega)^2 + 6\sigma(d\omega)^2 + 6\frac{(d\sigma)^2}{\sigma^2} \leq 30\left[(1+\sigma)(d\omega)^2 + \frac{(d\sigma)^2}{\sigma^2}\right] \leq$$

[see (14.151)]

$$\leq 30d_2F_*,$$

as required in (14.150). ■

Bibliography

[1] Gonzaga, C. (1988) "Polynomial time algorithm for linear programming" - Report ES-139/88, COPPE, Universidade Federal do Rio de Janeiro; appeared in: N.Megiddo, Ed. *Progress in Mathematical Programming: Interior Point and related methods*, Springer-Verlag, 1989

[2] Mizuno, S., Todd, M.J., and Ye, Y. (1993) "A surface of analytic centers and infeasible-interior-point algorithms in linear programming" - to appear in *MOR*

[3] Nesterov, Yu., and Nemirovski. A. *Interior point polynomial methods in Convex Programming: theory and applications - SIAM Publications*, Dec. 1993

[4] Nesterov, Yu. "Long-step strategies in interior point potential reduction methods" - Preprint, Departement d'Economie Commericale et Industrielle, Universite de Geneve, 102 Bd. Carl Vogt, CH-1211 Geneve 4, Switzerland, September 1993

[5] Nesterov, Yu., and Nemirovski, A. "Multi-parameter surfaces of analytic centers and long-step path-following interior point methods" - Research Report # 2/94, Optimization Laboratory, Faculty of Industrial Engineering & Management, Technion - Isreal Institute of Technology, April 1994.

[6] Nesterov, Yu., and Todd, M. "Self-scaled cones and interior-point methods in nonlinear programming" - CORE Discussion Paper # 9462 (1994), Louvain-la-Neuve, Belgium.

[7] Nesterov, Yu., and Todd, M. "Primal-dual interior point methods for self-scaled cones" - CORE Discussion Paper # 9544 (1995), Louvain-la-Neuve, Belgium.

[8] Renegar, J. (1986) "A polynomial time algorithm, based on Newton's method, for linear programming" - Research Report, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY; appeared in *Mathematical Programming* v. 40 (1988), 55-93

Appendix II: Robust Truss Topology Design

Robust Truss Topology Design via Semidefinite Programming

A. Ben-Tal[1)] and A. Nemirovski[1)]
Faculty of Industrial Engineering and Management
Technion - the Israel Institute of Technology, Technion City, Haifa 32000, Israel
e-mail: morbt@techunix.technion.ac.il   nemirovs@ie.technion.ac.il

Abstract

We present and motivate a new model of the Truss Topology Design problem, where the rigidity of the resulting truss with respect both to given loading scenarios and small "occasional" loads is optimized. It is shown that the resulting optimization problem is a Semidefinite Program. We derive and analyze several equivalent reformulations of the problem and present illustrative numerical examples.

Key words: structural optimization, truss topology design, robustness, semidefinite programming, interior point methods.

15.1 Introduction

Truss Topology Design (TTD) deals with the selection of optimal configuration for structural systems (mechanical, civil engineering, aerospace) and constitutes one of the newest and most rapidly growing fields of Structural Design (see the excellent survey paper by Rozvany, Bendsøe and Kirsch [12]). The TTD problem was studied extensively, both mathematically and algorithmically, in [1, 2, 3, 4, 5].

In this paper we bring forth the issue of the *robustness* of the truss; here we say that a truss is *robust*, if it is reasonable rigid with respect both to the given set of loading scenarios and to all small uncertain (in size and direction) load which may act at any of the *active* nodes of the truss, i.e., those which are linked at least by one bar. In the engineering literature rigidity is modeled by considering different *loading scenarios* on the structure (the multi-load TTD problem) or by imposing upper and lower bounds on nodal displacements. The first approach depends on the engineer's ability to "guess right" the relevant scenarios, while the second approach leads to a mathematical problem which is not tractable computationally. Here we suggest a new modeling approach which circumvents both of the above mentioned difficulties.

The paper is organized as follows. Section 15.2 describes the modeling approach in question. The preliminary Section 15.2.1 presents the basic notions related to the TTD problem and the traditional

[1] The research was partly supported by the Israeli Academy of Sciences grant # and grant # 191-142 of The Fund for Promotion of Research at Technion

formulations of the problem. We demonstrate by simple example (Section 15.2.2) that robustness restrictions (which are basically ignored in the traditional formulations) are critical to obtain reasonable designs; this observation motivates our modeling approach presented in Section 15.2.3. Its computational tractability is demonstrated in Section 15.2.4, where we show that the TTD problem in our new formulation can be equivalently cast as a *semidefinite program*. This brings the problem into the realm of Convex Programming for which efficient (polynomial time) interior point algorithms can be employed. Sections 15.3 – 15.5 are devoted to mathematical processing of the semidefinite program of Section 15.2.4; the goal is to get a program better suited for interior point algorithms. Possibilities for Robust Truss Topology Design by these algorithms are discussed in Section 15.6. We end up (Section 15.7) with illustrating usefulness of our approach by considering several examples of optimal trusses with and without robustness considerations. We show that at least for these examples *robustness* can be gained without sacrificing much in the optimality of the resulting trusses. Concluding Section 15.8 contains remarks on the possibility to extend the idea of "robust reformulation" of an optimization program from the particular case of the TTD problem to other problems of Mathematical Programming.

15.2 Truss Topology Design with Robustness Constraints

15.2.1 Trusses, loads, compliances

Informally, a *truss* is a 2D or 3D construction comprised of thin elastic *bars* linked with each other at *nodes* – points from finite *nodal set* $\mathcal{V}$ given in advance in 2D plane, respectively, 3D space. When subjected to a given *load* – distribution of external forces applied at the nodes – the construction deformates, until the reaction forces caused by deformations of the bars compensate the external load. The deformated truss capacitates certain potential energy, and this energy – the *compliance* – measures stiffness of the truss, its ability to withstand the load; the less is compliance, the more rigid is the truss with respect to the load.

In the usual Truss Topology Design (TTD) problem we are given the nodal set and one (*single-load* TTD) or several (*multi-load* TTD) loads, along with total volume of the bars. The displacements of some of the nodes are completely or partially fixed, so that the space R_v of virtual displacements of node v is certain linear subspace and the problem is to distribute the given volume of the truss between the bars in order to get the most rigid construction, i.e., the one which minimizes the maximal compliance over the given set of loads. Some of the bars can get zero volume, i.e., be eliminated from the resulting construction, so that in fact the topology of the construction is optimized as well (this is the origin of the term "Topology Design").

The mathematical formulation of the problem, in its simplest form, is as follows.

Given are:

- graph $(\mathcal{V}, \mathcal{B})$ (ground structure) with the nodal set $\mathcal{V} \subset \mathbf{R}^D$ $(D = 2, 3)$ comprised of $\widehat{n}$ nodes and with arc set $\mathcal{B}$ of m tentative bars;

- collection of linear subspaces $R_v \subset \mathbf{R}^D$, $v \in \mathcal{V}$ – the spaces of virtual displacements of the nodes.

  We refer to the quantity $n = \sum_{v \in \mathcal{V}} \dim R_v$ as the number of degrees of freedom of the nodal set and call the space $\mathbf{R}^n = \prod_{v \in \mathcal{V}} \mathbf{R}_v$ the *space of nodal displacements*. A vector $x \in \mathbf{R}^n$ can be naturally interpreted as collection of virtual displacements of the nodes. Similarly, a load – collection of external forces applied at the nodes – can be interpreted as a vector from $\mathbf{R}^n$ (one can ignore the components of the forces orthogonal to the subspaces of virtual nodal displacements, since these components are compensated by supports restricting virtual displacements of nodes; the remaining components of the forces can be naturally assembled in a vector from $\mathbf{R}^n$).

- When designing the truss, we are given a finite set $F \subset \mathbf{R}^n$ of *loading scenarios*; the truss should be able to carry the load for each of the scenarios.

- The design variables in the problem are *bar volumes* t_i, $i = 1, ..., m$; along with the nodal set $\mathcal{V}$, they completely determine the truss. We are given the total volume $V > 0$ of the bars, so that the set of all admissible vectors of bar volumes is the simplex

$$T = \{t \in \mathbf{R}^m \mid t \geq 0, \sum_{i=1}^{m} t_i = V\}.$$

With the elastic model of the bars, deformation of truss accompanied by displacement $x \in \mathbf{R}^n$ of the nodes results in the vector of reaction forces $A(t)x$, where t is the vector of bar volumes and

$$A(t) = \sum_{i=1}^{m} t_i A_i$$

is the $n \times n$ *bar-stiffness matrix* of the truss. The *bar-stiffness matrix* A_i of the i-th bar is readily given by the geometry of the nodal set, and involves the Young modulus of the material. What is crucial for us, is that, for all i,

$$A_i = b_i b_i^T \tag{15.153}$$

is a rank 1 positive semidefinite symmetric matrix (for explanations and details, see, e.g., [1, 2, 3]).

Given $t \in T$ and a load $f \in F$, one can associate with this pair the equilibrium equation

$$A(t)x = f \tag{15.154}$$

(as was explained, x is the vector of nodal displacements caused by the load f, provided that the vector of bar volumes is t). Solvability of this equation means that the truss is capable of carrying the load f, and if this is the case, then the *compliance*[2)]

$$c_f(t) \equiv f^T x = \sup_{u \in \mathbf{R}^n} \left[2f^T u - u^T A(t) u \right] \tag{15.155}$$

is regarded as a measure of internal work done by the truss with respect to the load f; the smaller is the compliance, the larger is the stiffness of the truss. If the equilibrium equation (15.154) for a given t is unsolvable, then it is convenient to define the compliance $c_f(t)$ as $+\infty$, which is compatible with the second equality in (15.155).

The problem of optimal minmax Truss Topology Design is to find the vector of bar volumes which results in the smallest possible worst-case compliance:

$(\mathrm{TD}_{\mathrm{minmax}})$: *find* $t \in T$ *which minimizes the worst-case compliance* $c^F(t) = \sup_{f \in F} c_f(t)$.

From now on we assume that the problem is *well-posed*, i.e., that

A. The matrix $\sum_{i=1}^{m} A_i$ is positive definite

(this actually means that the supports prevent rigid body motion of the truss).

15.2.2 Robustness constraint: Motivation

The "standard" case of problem $(\mathrm{TD}_{\mathrm{minmax}})$ is the one when F is a singleton (*single-load TTD problem*) or a finite set comprised of small number (3-5) of loads (*multi-load TTD problem*). An evident shortcoming of both these settings is that they do not take "full" care of the robustness of the resulting truss. The associated optimal design ensures reasonable (in fact the best possible) behaviour of the truss under the loads from the list of scenarios F; it may happen however that a load not from this set, even a "small" one, will cause an inappropriately large deformation of the truss. Consider, e.g., the following toy example. Fig. 1 represents 6-element nodal set with 2 fixed nodes ($R_v = \{0\}$) and 4 free nodes ($R_v = \mathbf{R}^2$), the "ground structure" – the set of all tentative bars, and the load f which is the unique element of F.

[2)] The "true" compliance, as defined in Mechanics, is one half of the quantity given by (15.155); we rescale the compliance in order to avoid multiple fractions $\frac{1}{2}$

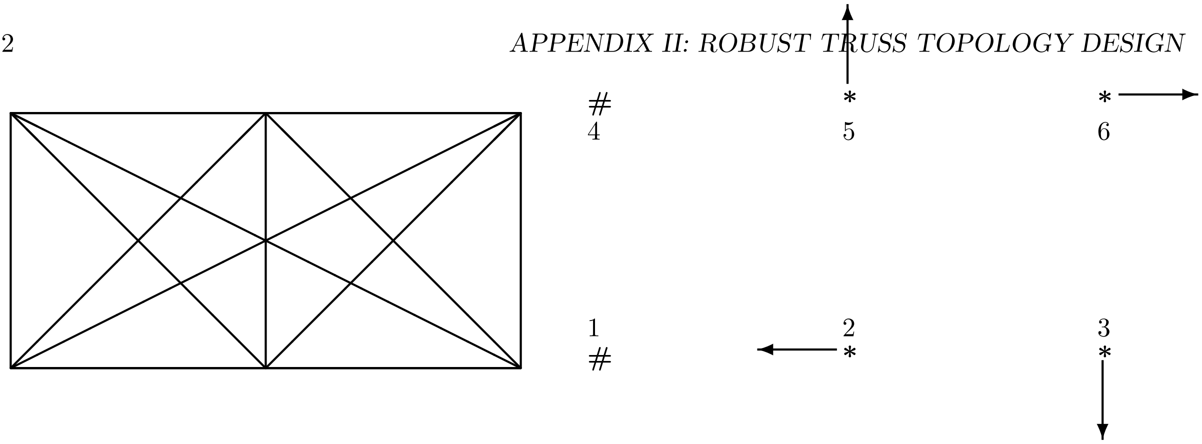

Fig. 1: Ground structure and loading scenario
* – free nodes; # – fixed nodes; arrows - forces

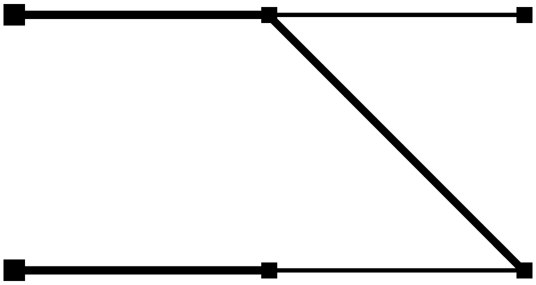

Fig. 2: Optimal single-load design

Fig. 2 shows the results of the usual single-load design which results in the optimal compliance 16.000. Note that the resulting truss is completely unstable: e.g., the bar linking nodes 5 and 6 can rotate around node 5, so that arbitrarily small non-horizontal force applied at node 6 will cause infinite compliance.

It seems that a "good" design should ensure reasonable compliances under *all* tentative loads of reasonable magnitude acting at the nodes of the resulting truss, not only "the best possible" compliance under the small list of loads in F of primary interest.

The indicated requirement can be modeled as follows. When formulating the problem, the engineer embeds a small finite set of loads $F = \{f_1, ..., f_q\}$ he is especially interested in ("primary" loads) into a "more massive" set M containing F, but also "occasional loads" of perhaps much smaller magnitude ("secondary" loads), and looks for the truss $t \in T$ which minimizes the worst-case compliance $c^M(t)$ taken with respect to this extended set M of loading scenarios.

In order to get a computationally tractable problem, in what follows we restrict ourselves to the case where M is an ellipsoid centered at the origin[3)].

$$M = QW_q \equiv \{Qe \mid e \in \mathbf{R}^q, \; e^T e \leq 1\}.$$

Here Q is a given $n \times q$ "scale" matrix, and W_q is the unit Euclidean ball in $\mathbf{R}^q$. Note that we allow the case $q < n$ as well, where M is "flat" q-dimensional ellipsoid.

The corresponding modification of $(\mathrm{TD}_{\mathrm{minmax}})$ is as follows:

$(\mathrm{TD}_{\mathrm{robust}})$: *find $t \in T$ which minimizes the compliance*

$$c^M(t) = \max_{e^T e \leq 1} \max_{x \in \mathbf{R}^n} \left[2(Qe)^T x - x^T A(t) x \right].$$

[3] the only other case when the indicated problem is computationally tractable seems to be that one of a polytope M given by the list of its vertices. This case hardly deserves a special consideration, since it leads to the standard multi-load TTD problem

15.2.3 Selection of scale matrix Q

Problem (TD$_{\rm robust}$) takes care of all loads $f \in M$, M being the image of the unit q-dimensional Euclidean ball under the mapping $e \mapsto Qe$. It follows that if a load $f \in M$ has a nonzero force acting at certain node l, then this node will for sure be present in the resulting construction. This observation means that we should be very careful when forming Q – otherwise we enforce incorporating into the final construction the nodes which in fact are redundant. There are two ways to meet the latter requirement:

- **A.** We could use the indicated approach as a postoptimality analysis; after we have found the solution to the usual multi-load TTD problem, given the resulting nodal structure, we can improve the robustness of the solution by solving (TD$_{\rm robust}$) associated with this nodal structure.
- **B.** We know in advance some nodes which for sure will present in the solution (certainly the nodes where the forces from the given loading scenarios are applied) and it seems to be natural to require rigidity with respect to all properly scaled forces acting at these "active" nodes.

Let us discuss in more details the latter possibility. Let $F = \{f_1, ..., f_k\}$ be the given set of loading scenarios. We say that a node $v \in \mathcal{V}$ is *active* with respect to F if the projection of certain load f_j on the space $\mathbf{R}_v$ of virtual displacements of the node is nonzero. Let $\mathcal{V}^*$ be the set of all active nodes. Our goal is to embed F into a "reasonably chosen" ellipsoid M in the space $\mathbf{R}^q = \prod_{v\in\mathcal{V}^*} \mathbf{R}_v$ (which for sure will be the part of the displacement space in the final construction). According to our motivation, M should contain

- the set F of given loads;
- the ball $B = \{f \in \mathbf{R}^q \mid f^T f \le r^2\}$ of all "occasional" loads of prescribed magnitude r.

The most adequate to our motivation setup $M = F \cup B$ is inappropriate – as it was explained, we need M to be an ellipsoid in order to get a computationally tractable problem, so that we should look for "the smallest possible" ellipsoid M containing $F \cup B$. The simplest interpretation of "the smallest possible" here is in terms of q-dimensional volume. Thus, it is natural to choose as M the ellipsoid in $\mathbf{R}^q$ centered at the origin and containing $F \cup B$ of the minimum q-dimensional volume. To form the indicated *ellipsoidal envelope* M of F and B is a convex problem; since normally q is not large, there is no difficulty to solve the problem numerically. Note, however, that there exists an "easy case" where M can be pointed out explicitly. Namely, let $L(F) \subset \mathbf{R}^k$ be the linear span of F. Assume that

- the loads $f_1, ..., f_k$ are linearly independent;
- the convex hull $\widehat{F}$ of the set $F \cup (-F)$ contains the k-dimensional ball $B' = B \cap L(F)$.

Note that in actual design both these assumptions normally are satisfied.

Lemma 15.2.1 *Under the indicated assumptions the ellipsoidal envelope of F and B is*

$$M = QW_q, \quad Q = [f_1; ...; f_k; re_1; ...; re_{q-k}], \tag{15.156}$$

where $e_1, ..., e_{q-k}$ is an orthonormal basis in the orthogonal complement to $L(F)$ in $\mathbf{R}^q$.

Proof. We can choose an orthonormal basis in $\mathbf{R}^q$ in such a way that the first k vectors of the basis span $L(F)$ and the rest $q-k$ vectors span the orthogonal complement $L^\perp(F)$ to $L(F)$ in $\mathbf{R}^q$. Let $x = (u, v)$ be the coordinates of a vector in this basis (u are the first k and v are the rest $q-k$ coordinates). A centered at the origin ellipsoid E in $\mathbf{R}^q$ can be parameterized by a positive definite symmetric $q \times q$ matrix A:

$$E = \{x \mid x^T Ax \le 1\};$$

the squared volume of E is inversely proportional to $\det A$. The matrix A_* corresponding to the minimum volume centered at the origin ellipsoid containing F and B is therefore an optimal solution to the following convex program:

$$\ln\det A \to \max \mid \; A = A^T > 0, \; x^T Ax \le 1 \;\forall x \in B \cup \widehat{F}. \tag{15.157}$$

The problem clearly is solvable, and since its objective is strictly concave on the cone of positive definite symmetric $q \times q$ matrices, the solution is unique. On the other hand, let J be the matrix of the mapping

$(u, v) \mapsto (u, -v)$; then the mapping $A \mapsto J^T AJ$ clearly is a symmetry of (15.157): this mapping preserves feasibility and does not vary the value of the objective. We conclude that the optimal solution is invariant with respect to the indicated mapping: $A_* = JA_*J$, whence A_* is block diagonal with $k \times k$ diagonal block U_* and $(q-k) \times (q-k)$ diagonal block V_*. Since the ellipsoid $\{x \mid x^T A_* x \leq 1\}$ contains $B \cup \widehat{F}$, the k-dimensional ellipsoid $M' = \{u \mid u^T U_* u \leq 1\}$ in $L(F)$ contains $\widehat{F}$, while the $(q-k)$-dimensional ellipsoid $M'' = \{v \mid v^T V_* v \leq 1\}$ in $L^\perp(F)$ contains the ball B'' centered at the origin of the radius r in $L^\perp(F)$.

Now let $U = U^T > 0$ and $V = V^T > 0$ be $k \times k$ and $(q-k) \times (q-k)$ matrices such that the ellipsoids $E' = \{u \mid u^T U u \leq 1\}$ and $E'' = \{v \mid v^T V v \leq 1\}$ contain $\widehat{F}$ and B'', respectively. We claim that then the ellipsoid $\{x \mid x^T Ax \leq 1\}$, $A = \mathrm{Diag}(U, V)$, contains $B \cup \widehat{F}$. Indeed, the ellipsoid clearly contains $\widehat{F}$, and all we need is to verify that if $x = (u, v) \in B$, i.e., $u^T u + v^T v \leq r^2$, then $u^T U u + v^T V v \leq 1$. This is immediate: since $E' \supset \widehat{F} \supset B'$, we have $u^T U u \leq 1$ whenever $u^T u \leq r^2$, or, which is the same, $u^T U u \leq r^{-2} u^T u$ for all u. Similarly, $E'' \supset B''$ implies that $v^T V v \leq r^{-2} v^T v$, so that $u^T u + v^T v \leq r^2$ indeed implies $u^T U u + v^T V v \leq 1$.

The above observations combined with the identity $\ln \det A = \ln \det U + \ln \det V$ for positive definite symmetric $A = \mathrm{Diag}(U, V)$ demonstrate that the block U_* of the optimal solution to (15.157) corresponds to the minimum volume ellipsoid in $L(F)$ containing $\widehat{F}$, and similarly for V_*, $L^\perp(F)$ and B''. In other words, M is the "ellipsoidal product" of the ellipsoid M' of the minimum volume in $L(F)$ containing $F \cup (-F)$ and the ball B'' in $L^\perp(F)$: if $M' = Q' W_k$, then

$$M = [Q'; re_1; ...; re_{q-k}]\, W_q.$$

To conclude the proof, it suffices to verify that one can choose, as Q', the matrix $[f_1; ...; f_k]$, which is immediate. Indeed, let $s_1, ..., s_k$ be an orthonormal basis in $L(F)$, and let D be the linear transformation of $L(F)$ which maps s_i onto f_i, $i = 1, ..., k$. Since the ratio of k-dimensional volumes of solids in $L(F)$ remains invariant under the transformation D, $M' = DN'$, where N' is the minimum volume ellipsoid centered at the origin in $L(F)$ containing $s_1, ..., s_k$. The latter ellipsoid is clearly $[s_1; ...; s_k]\, W_k$, whence

$$M' = DN' = \{D(\sum_{i=1}^{k} \lambda_i s_i) \mid \lambda \in W_k\} = \{\sum_{i=1}^{k} \lambda_i f_i \mid \lambda \in W_k\} = [f_1; ...; f_k]\, W_k. \quad \blacksquare$$

Remark 15.2.1 Evident modification of the proof of Lemma 15.2.1 demonstrates that the minimum volume ellipsoid in $\mathbf{R}^q$ centered at the origin and containing $F \cup B$ always is the "ellipsoidal product" of the minimum volume ellipsoid M' in $L(F)$ containing $F \cup (-F) \cup B'$ and the ball B'' in $L^\perp(F)$: if $M' = Q' W_{\widehat{k}}$, $\widehat{k} = \dim\, L(F)$, then $M = \left[Q'; re_1, ..., re_{q-\widehat{k}}\right] W_q$, $e_1, ..., e_{q-\widehat{k}}$ being an orthonormal basis in $L^\perp(F)$. Thus, to find M is, basically, the same as to find M', and this latter convex problem normally is of quite a small dimension, since $\widehat{k} \leq k$ and typically $k \leq 5$.

The outlined way of modeling the robustness constraint is, perhaps, more reasonable than the usual multi-load setting of the TTD problem. Indeed, the new model enforces certain level of rigidity of the resulting construction with respect not only to the primary loads, but also to loads associated with "active" nodes. At the same time, it turns out, as we are about to demonstrate, that the resulting problem $(\mathrm{TD_{robust}})$ is basically not more computationally demanding than the usual multi-load TTD problem of the same size (i.e., with the same ground structure and the number of scenario loads equal to the dimension of the loading ellipsoid used in $(\mathrm{TD_{robust}})$).

15.2.4 Semidefinite reformulation of $(\mathrm{TD_{robust}})$

Our goal now is to rewrite $(\mathrm{TD_{robust}})$ equivalently as a so called *semidefinite program*. To this end we start with the following simple result.

Lemma 15.2.2 *Let A be a positive semidefinite $n \times n$ matrix, and let*

$$c = \max_{x \in \mathbf{R}^n; e \in \mathbf{R}^q : e^T e \leq 1} \left[2(Qe)^T x - x^T Ax\right]. \tag{15.158}$$

Then the inequality $c \le \tau$ is equivalent to positive semidefiniteness of the matrix

$$\mathcal{A} = \begin{pmatrix} \tau I_q & Q^T \\ Q & A \end{pmatrix},$$

I_q being the unit $q \times q$ matrix.

Proof. We have

$$c \le \tau \Leftrightarrow \forall (x \in \mathbf{R}^n, e \in \mathbf{R}^q, e^T e \le 1): \quad \tau - 2(Qe)^T x + x^T A x \ge 0 \Leftrightarrow$$

[by homogeneity reasons]

$$\forall (\lambda > 0, x \in \mathbf{R}^n, e \in \mathbf{R}^q, e^T e \le 1): \quad \tau\lambda^2 - 2(Q\lambda e)^T(\lambda x) + (\lambda x)^T A(\lambda x) \ge 0 \Leftrightarrow$$

[set $\lambda e = f, \lambda x = y$]

$$\forall (\lambda > 0, y \in \mathbf{R}^n, f \in \mathbf{R}^q, f^T f \le \lambda^2): \quad \tau\lambda^2 - 2(Qf)^T y + y^T A y \ge 0 \Rightarrow$$

$$\forall (\begin{pmatrix} f \\ y \end{pmatrix} \in \mathbf{R}^{q+n}): \quad \begin{pmatrix} f \\ y \end{pmatrix}^T \begin{pmatrix} \tau I_q & Q^T \\ Q & A \end{pmatrix} \begin{pmatrix} f \\ y \end{pmatrix} \equiv \tau f^T f - 2(Qf)^T y + y^T A y \ge 0.$$

Thus, $\tau \ge c \Rightarrow \mathcal{A} \ge 0$. Vice versa, if $\mathcal{A} \ge 0$, then clearly $\tau \ge 0$, and therefore the implication $\Rightarrow$ in the above chain can be inverted. ■

Remark 15.2.2 It is well-known that a symmetric matrix $\begin{pmatrix} U & Q^T \\ Q & A \end{pmatrix}$ with positive definite U is positive semidefinite if and only if $A \ge QU^{-1}Q^T$. Applying this observation to the case of $U = \tau I_q$, we can reformulate the result of Lemma 15.2.2 as follows:

The compliance c of a truss t with respect to the ellipsoid of loads $M = QW_q$ is $\le \tau$ if and only if $A(t) \ge \tau^{-1}QQ^T$.

In the particular case when QQ^T is the orthoprojector P onto the linear span L of the columns of Q, the above observation can be reformulated as follows:

$c \le \tau$ if and only if the minimum eigenvalue of the restriction of $A(t)$ onto L is $\ge \tau^{-1}$

(in the general case, the interpretation is similar, but instead of the usual minimum eigenvalue of the restriction we should speak about minimum eigenvalue of the matrix pencil $(A\,|_L, QQ^T\,|_L)$ on L).

In view of Lemma 15.2.2, problem $(\mathrm{TD_{robust}})$ can be rewritten equivalently as the following *Semidefinite Program*:

$$\begin{array}{lrcl}
(\mathrm{TD_{sd}}) & & & \\
& \min\limits_{t \in \mathbf{R}^m, \tau \in \mathbf{R}} & & \tau \\
s.t. & & & \\
& \begin{pmatrix} \tau I_q & Q^T \\ Q & A(t) \end{pmatrix} & \ge & 0, \\
& t & \ge & 0 \\
& \sum_{i=1}^m t_i & = & V
\end{array}$$

(here and in what follows the inequality $A \ge B$ between symmetric matrices means that the matrix $A - B$ is positive semidefinite).

15.3 Deriving a dual problem to $(\mathrm{TD_{sd}})$

Here we derive the Fenchel-Rockafellar [11] dual to the problem $(\mathrm{TD_{sd}})$. The latter problem is of the form

$$\min\{\tau : \ \mathcal{A}(\tau, t) + B \in \mathbf{S}_+, \ t \in T\},$$

where

$$\mathcal{A}(\tau, t) = \begin{pmatrix} \tau I_q & 0 \\ 0 & A(t) \end{pmatrix}$$

is a linear mapping from $\mathbf{R} \times \mathbf{R}^n$ to the space $\mathbf{S}$ of symmetric $(n+q) \times (n+q)$ matrices equipped with the standard Frobenius Euclidean structure $\langle X, Y \rangle = \mathrm{Tr}(XY)$, $\mathbf{S}_+$ is the cone of positive semidefinite matrices from $\mathbf{S}$ and

$$B = \begin{pmatrix} 0 & Q^T \\ Q & 0 \end{pmatrix} \in \mathbf{S}.$$

We write the problem in the Fenchel-Rockafellar primal scheme:

$$\text{(P)} \qquad \min \left\{ f(\tau, t) - g(\mathcal{A}(\tau, t)) \right\},$$

where

$$f(\tau, t) = \tau + \delta(t \mid T), \quad g(X) = -\delta(X + B \mid \mathbf{S}_+)$$

and $\delta(x \mid W)$ is the indicator function of a set W. To derive the dual to (P), we need to compute the conjugates f^* and g_* of the convex function f and the concave function g, which is quite straightforward:

$$f^*(\sigma, s) = \sup_{\tau, t} \{\sigma\tau + s^T t - \tau \mid \ t \in T\} = \begin{cases} V \max_{1 \le i \le n} s_i & ,\sigma = 1 \\ +\infty & ,\text{otherwise} \end{cases};$$

$$g_*(R) = \inf_S \{\mathrm{Tr}(SR) \mid S + B \in \mathbf{S}_+\} = \inf\{\mathrm{Tr}((Z - B)R) \mid Z \in \mathbf{S}_+\} =$$

$$= \begin{cases} -\mathrm{Tr}(BR), & R \in \mathbf{S}_+ \\ -\infty, & \text{otherwise} \end{cases}$$

(we have used the well-known fact that the cone of positive semidefinite matrices is self-conjugate with respect to the Frobenius Euclidean structure).

The Fenchel-Rockafellar dual to (P) is

$$\text{(D)} \qquad \sup\nolimits_{R \in \mathbf{S}} \left\{ g_*(R) - f^*(\mathcal{A}^* R) \right\},$$

where $\mathcal{A}^* : \mathbf{S} \to \mathbf{R} \times \mathbf{R}^n$ is the adjoint to $\mathcal{A}$.

Representing $R \in \mathbf{S}$ in the block form

$$R = \begin{pmatrix} \Lambda & X^T \\ X & Y \end{pmatrix}$$

(Λ is $q \times q$, Y is $n \times n$), we get

$$\mathcal{A}^* R = \begin{pmatrix} \tau = \mathrm{Tr}\,\Lambda \\ t_1 = \mathrm{Tr}(A_1 Y) \\ \ldots \\ t_n = \mathrm{Tr}(A_n Y) \end{pmatrix}.$$

Substituting the resulting expressions for f^*, g_* and A^*, we come to the following explicit formulation of the dual problem (D):

$$\text{(D)} \qquad \max \left[-2\,\mathrm{Tr}(QX^T) - V \max\nolimits_{i=1,\ldots,m} [\mathrm{Tr}(A_i Y)] \right],$$

s.t.

$$\begin{pmatrix} \Lambda & X^T \\ X & Y \end{pmatrix} \ge 0, \ \mathrm{Tr}\,\Lambda = 1,$$

the design variables being: symmetric $q \times q$ and $n \times n$ matrices Λ, Y, respectively, and $n \times q$ matrix X.

Note that the functions f and g in (P) are clearly closed convex and concave, respectively. Moreover, from the well-posedness assumption **A**, it immediately follows that (P) is strictly feasible (i.e., the relative interiors of the domains of $f(\tau, t)$ and $\phi(\tau, t) = g(\mathcal{A}(\tau, t))$ have nonempty intersection, and the image of the mapping $\mathcal{A}$ intersects the interior of the domain of g); to see this, choose arbitrary positive $t \in T$ and enforce τ to be large enough). Of course (P) is bounded below (the compliance always is nonnegative), thus, all requirements of the Fenchel-Rockafellar Duality Theorem are satisfied, and we come to

Proposition 15.3.1 *(D) is solvable, and the optimal values in (P) and (D) are equal to each other.*

Remark 15.3.1 To the moment, we dealt with the TTD problem with *simple constraints* on the bar volumes:

$$t \in T = \{t \in \mathbf{R}^n \mid t \geq 0, \sum_{i=1}^{n} t_i = V\}.$$

In the case when there are also lower and upper bounds on the bar volumes, so that the constraints on t are

$$t \in T^+ = \{t \in T \mid L \leq t \leq U\},$$

($U > L \geq 0$ are given n-dimensional vectors), the above derivation results in a dual problem as follows:

$(\mathrm{D_b})$ $\max \left[-2\,\mathrm{Tr}(QX^T) - \lambda V - \sum_{i=1}^{n} \max\left[(\mathrm{Tr}(YA_i) - \lambda)L_i; (\mathrm{Tr}(YA_i) - \lambda)U_i\right]\right]$

s.t.

$$\begin{pmatrix} \Lambda & X^T \\ X & Y \end{pmatrix} \geq 0,\ \mathrm{Tr}\,\Lambda = 1,$$

the design variables being real λ, symmetric $q \times q$ matrix Λ, symmetric $n \times n$ matrix Y and $n \times q$ matrix X.

15.4 A simplification of the dual problem (D)

Our next goal is to simplify problem (D), derived in the previous section, by eliminating the matrix variable Y. To this end it suffices to note that (D) can be rewritten as

$$\begin{array}{lrcl}
(\mathrm{TD_{dl}}) & & & \\
\min\limits_{X \in \mathbf{R}^{n\times q}, \Lambda = \Lambda^T \in \mathbf{R}^{q\times q}, Y = Y^T \in \mathbf{R}^{n\times n}, \rho \in \mathbf{R}} & 2\,\mathrm{Tr}(QX^T) + V\rho & & \\
s.t. & & & \\
(\alpha) & \mathrm{Tr}(YA_i) & \leq & \rho,\ i = 1, ..., m \\
(\beta) & \begin{pmatrix} \Lambda & X^T \\ X & Y \end{pmatrix} & \geq & 0 \\
(\gamma) & \mathrm{Tr}(\Lambda) & = & 1
\end{array}$$

(we have replaced the maximization problem (D) by an equivalent minimization one). Note that $(\mathrm{TD_{dl}})$ is strictly feasible – there exists a feasible solution where all scalar inequality constraints and the matrix inequality one are strict (take $\Lambda = q^{-1}I_q$, $Y = I_n$ and enforce ρ to be large enough).

The matrix inequality (β) clearly implies that Λ is positive semidefinite. Thus, we do not vary $(\mathrm{TD_{dl}})$ when adding (in fact, redundant) inequality $\Lambda \geq 0$. Now let us strengthen, for a moment, the latter inequality to one

$$\Lambda > 0 \tag{15.159}$$

i.e. positive definiteness of Λ; it is immediately seen from strict feasibility of $(\mathrm{TD_{dl}})$ that this transformation does not violate the optimal value of the problem, although it may cut off the optimal solution (anyhow, from the computational viewpoint the exact solution is nothing but a fiction). Thus, we may focus on the problem $(\mathrm{TD'_{dl}})$ obtained from $(\mathrm{TD_{dl}})$ by adding to the list of constraints inequality (15.159).

The pair of matrix inequalities (β), (15.159) which are present among the constraints of $(\mathrm{TD'_{dl}})$ is equivalent to the pair of matrix inequalities

$$\Lambda > 0; \quad Y \geq Y^*(\Lambda, X) = X\Lambda^{-1}X^T.$$

Now let (Λ, X, Y, ρ) be a feasible solution to $(\mathrm{TD'_{dl}})$; then, as we just have mentioned, $Y \geq Y^*(\Lambda, X)$ and the collection $(\Lambda, X, Y^* = Y^*(\Lambda, X), \rho)$ satisfies (β), (γ) and (15.159). Moreover, since A_i are symmetric positive semidefinite and $Y \geq Y^*$, we have $\mathrm{Tr}(YA_i) \geq \mathrm{Tr}(Y^*A_i)$, so that the updated collection satisfies (α) as well, and (Λ, X, Y^*, ρ) is feasible for $(\mathrm{TD'_{dl}})$. Note that the transformation $(\Lambda, X, Y, \rho) \mapsto (\Lambda, X, Y^*(\Lambda, X), \rho)$ does not affect the objective function of the problem. We conclude that $(\mathrm{TD'_{dl}})$ can be equivalently rewritten as

$$\min_{X \in \mathbf{R}^{n\times q}, \Lambda = \Lambda^T \in \mathbf{R}^{q\times q}, \rho \in \mathbf{R}} 2\,\mathrm{Tr}(QX^T) + V\rho \ s.t.\ \Lambda > 0,\ \mathrm{Tr}(\Lambda) = 1,\ \rho \geq \mathrm{Tr}(X\Lambda^{-1}X^TA_i),\ i = 1, ..., m.$$

Substituting $A_i = b_i b_i^T$ (see (15.153)), we can rewrite the constraints

$$\rho \geq \mathrm{Tr}(X\Lambda^{-1}X^T A_i)$$

as

$$\rho \geq (X^T b_i)^T \Lambda^{-1}(X^T b_i),$$

which is the same (since $\Lambda = \Lambda^T > 0$), as

$$\begin{pmatrix} \Lambda & X^T b_i \\ b_i^T X & \rho \end{pmatrix} \geq 0.$$

With this substitution, the problem $(\mathrm{TD}'_{\mathrm{dl}})$ becomes

$$\min_{X\in\mathbf{R}^{n\times q},\Lambda=\Lambda^T\in\mathbf{R}^{q\times q},\rho\in\mathbf{R}} \quad 2\,\mathrm{Tr}(QX^T) + V\rho \; s.t. \; \Lambda > 0,\ \mathrm{Tr}(\Lambda) = 1,\ \begin{pmatrix} \Lambda & X^T b_i \\ b_i^T X & \rho \end{pmatrix} \geq 0,\ i = 1, ..., m.$$

When replacing the strict inequality $\Lambda > 0$ in the latter problem with the nonstrict one $\Lambda \geq 0$, we clearly do not vary the optimal value of the problem; in the modified problem, the inequality $\Lambda \geq 0$ is in fact redundant (it follows from positive semidefiniteness of any of the matrices $\begin{pmatrix} \Lambda & X^T b_i \\ b_i^T X & \rho \end{pmatrix}$). With these modifications, we come to the final formulation of the problem dual to $(\mathrm{TD}_{\mathrm{robust}})$:

$$(\mathrm{TD}_{\mathrm{fn}}) \quad \begin{array}{lll} \min\limits_{\Lambda=\Lambda^T\in\mathbf{R}^{q\times q},X\in\mathbf{R}^{n\times q},\rho\in\mathbf{R}} & 2\,\mathrm{Tr}(QX^T) + V\rho & \\ s.t. & & \\ & \begin{pmatrix} \Lambda & X^T b_i \\ b_i^T X & \rho \end{pmatrix} \geq 0, & i = 1, ..., m, \\ & \mathrm{Tr}(\Lambda) = 1 & \end{array}.$$

Note that $(\mathrm{TD}_{\mathrm{fn}})$ is very similar to the standard multi-load TTD problem in dual setting [5]; the only difference is that in the latter problem Λ is further restricted to be diagonal.

15.5 Recovering the bar volumes

To the moment, the only relation between the initial primal problem $(\mathrm{TD}_{\mathrm{robust}})$ and the dual one $(\mathrm{TD}_{\mathrm{fn}})$ is that their optimal values are negations of each other (note that when coming to $(\mathrm{TD}_{\mathrm{fn}})$ from the maximization problem $(\mathrm{TD}_{\mathrm{dl}})$ which has the same optimal value as $(\mathrm{TD}_{\mathrm{sd}})$, we have changed the sign of the objective and have replaced maximization with minimization). Thus, the problem arises: how to restore good approximate solutions to $(\mathrm{TD}_{\mathrm{robust}})$ via good approximate solutions to $(\mathrm{TD}_{\mathrm{fn}})$. To resolve this problem, we first derive the Fenchel-Rockafellar dual $(\mathrm{TD}^*_{\mathrm{fn}})$ to $(\mathrm{TD}_{\mathrm{fn}})$ and recognize in it the initial problem $(\mathrm{TD}_{\mathrm{robust}})$, and then use the well-known relation in Interior-Point Theory between "central path" approximate solutions to $(\mathrm{TD}_{\mathrm{fn}})$ and approximate solutions to $(\mathrm{TD}^*_{\mathrm{fn}})$.

15.5.1 A dual problem to $(\mathrm{TD}_{\mathrm{fn}})$

Similar to the above, we represent problem $(\mathrm{TD}_{\mathrm{fn}})$ in the Fenchel-Rockafellar scheme:

$$(\mathrm{PI}) \qquad \min\left\{f(\Lambda, X, \rho) - g(\mathcal{A}(\Lambda, X, \rho))\right\},$$

where

$$f(\Lambda, X, \rho) = 2\,\mathrm{Tr}(QX^T) + V\rho + \delta(\mathrm{Tr}(\Lambda) \mid \{1\}),$$

$$\mathcal{A}(\Lambda, X, \rho) = \mathrm{Diag}\left\{\begin{pmatrix} \Lambda & X^T b_i \\ b_i^T X & \rho \end{pmatrix}, i = 1, ..., m\right\}$$

is the linear mapping from the space of design variables of $(\mathrm{TD}_{\mathrm{fn}})$ to the space $\mathbf{S}$ of block-diagonal symmetric matrices with m diagonal blocks of the sizes $(q + 1) \times (q + 1)$ each, and

$$g(W) = -\delta(W \mid \mathbf{S}_+),$$

$\mathbf{S}_+$ being the cone of positive semidefinite matrices from $\mathbf{S}$.

The dual to (P) is

$$\text{(DI)} \qquad \max_{R\in\mathbf{S}} \left\{g_*(R) - f^*(\mathcal{A}^*R)\right\},$$

where $\mathcal{A}^*$ is the operator adjoint to $\mathcal{A}$. Here

$$f^*(L,\Xi,r) = \sup_{\Lambda,X,\rho}\left[\operatorname{Tr}(\Lambda L) + \operatorname{Tr}(\Xi X^T) + r\rho - f(\Lambda,X,\rho)\right] =$$

$$= \sup_{\Lambda}\left[\operatorname{Tr}(\Lambda L) - \delta(\operatorname{Tr}(\Lambda) \mid \{1\})\right] + \sup_{X}\left[\operatorname{Tr}(\Xi X^T) - 2\operatorname{Tr}(QX^T)\right] + \sup_{\rho}\left[r\rho - V\rho\right] =$$

$$= \frac{1}{q}\operatorname{Tr}(L) + \delta((L,\Xi,r) \mid \{(L = \lambda I_q, 2Q, V) \mid \lambda\in\mathbf{R}\}) =$$

$$= \begin{cases} \lambda, & \text{if } L = \lambda I_q \text{ for some } \lambda\in\mathbf{R} \text{ and } \Xi = 2Q,\ r = V \\ \infty, & \text{otherwise} \end{cases}$$

and

$$g_*(R) = \inf_{S}\left[\operatorname{Tr}(SR) + \delta(S \mid \mathbf{S}_+)\right] = -\delta(R \mid \mathbf{S}_+)$$

(here we again used the fact that the cone $\mathbf{S}_+$ is self-dual with respect to the Frobenius Euclidean structure of $\mathbf{S}$).

Denoting a generic element of $\mathbf{S}$ as

$$R = \operatorname{Diag}\left\{\begin{pmatrix} L_i & d_i \\ d_i^T & t_i \end{pmatrix},\ i = 1,...,m\right\}$$

(L_i are symmetric $q\times q$ matrices, d_i are q-dimensional vectors, t_i are reals) it can be seen that:

$$\mathcal{A}^*R = (L = \sum_{i=1}^m L_i, \Xi = 2\sum_{i=1}^m b_i d_i^T, r = \sum_{i=1}^m t_i).$$

With these relations, the dual (DI) to (PI) becomes

$$\begin{array}{lrcl}
(\mathrm{TD}^*_{\mathrm{fn}}) & & & \\
 & \min\limits_{\lambda\in\mathbf{R}, L_i = L_i^T\in\mathbf{R}^{q\times q}, d_i\in\mathbf{R}^q, t_i\in\mathbf{R}} & & \lambda \\
\text{s.t.} & & & \\
(\alpha) & \sum_{i=1}^m L_i & = & \lambda I_q, \\
(\beta) & \sum_{i=1}^m b_i d_i^T & = & Q, \\
(\gamma) & \sum_{i=1}^m t_i & = & V, \\
(\delta) & \begin{pmatrix} L_i & d_i \\ d_i^T & t_i \end{pmatrix} & \geq & 0,\ i = 1,...,m
\end{array}$$

(we again have replaced a maximization problem with the equivalent minimization one).

Problem $(\mathrm{TD}_{\mathrm{fn}})$ clearly satisfies the assumption of the Fenchel-Rockafellar Duality Theorem, and this together with Proposition 15.3.1 proves

Proposition 15.5.1 *Problem* $(\mathrm{TD}^*_{\mathrm{fn}})$ *is solvable, and its optimal value* λ^* *is equal to the optimal value* c^* *of the initial problem* $(\mathrm{TD}_{\mathrm{robust}})$.

It is not difficult to guess that the variables t_i involved into $(\mathrm{TD}^*_{\mathrm{fn}})$ can be interpreted as our initial bar volumes t_i. The exact statement is given by the following

Theorem 15.5.1 *Let* $R = \{\lambda; L_i, d_i, t_i,\ i = 1,...,m\}$ *be a feasible solution to* $(\mathrm{TD}^*_{\mathrm{fn}})$. *Then the vector* $t = (t_1,...,t_m)$ *is a feasible solution to* $(\mathrm{TD}_{\mathrm{robust}})$, *and the value of the objective of the latter problem at* t *is less than or equal to* λ. *In particular, if* R *is an* ϵ*-solution to* $(\mathrm{TD}^*_{\mathrm{fn}})$ *(i.e.,* $\lambda - \lambda^* \leq \epsilon$*), then* t *is an* ϵ*-solution to* $(\mathrm{TD}_{\mathrm{robust}})$ *(i.e.,* $c^M(t) - c^* \leq \epsilon$*).*

Proof. The "in particular" part of the statement follows from its first part due to Proposition 15.5.1, and all we need is to prove the first part. From the positive semidefiniteness constraints (δ) in $(\mathrm{TD}^*_{\mathrm{fn}})$ it follows that $t \geq 0$, which combined with (γ) implies the inclusion $t \in T$. To complete the proof, we should verify that $c^M(t) \leq \lambda$.

Let $e \in \mathbf{R}^q$, $e^T e \leq 1$. From (β) we have

$$Qe = \sum_{i=1}^m (d_i^T e) b_i.$$

Let $x \in \mathbf{R}^n$. Due to $A_i = b_i b_i^T$, we have

$$\phi_e(x) \equiv 2(Qe)^T x - x^T A(t) x = \sum_{i=1}^m 2(d_i^T e)(b_i^T x) - t_i \sum_{i=1}^m (b_i^T x)^2 =$$

$$= \sum_{i=1}^m \left[2(d_i^T e)(b_i^T x) - t_i (b_i^T x)^2\right] =$$

[denoting $s_i = -b_i^T x$]

$$-\sum_{i=1}^m \left[e^T L_i e + 2(d_i^T e) s_i + t_i s_i^2\right] + \sum_{i=1}^m e^T L_i e = -\sum_{i=1}^m \begin{pmatrix} e \\ s_i \end{pmatrix}^T \begin{pmatrix} L_i & d_i \\ d_i^T & t_i \end{pmatrix} \begin{pmatrix} e \\ s_i \end{pmatrix} + \sum_{i=1}^m e^T L_i e \leq$$

[by (δ)]

$$\leq \sum_{i=1}^m e^T L_i e =$$

[by (α)]

$$= \lambda.$$

Thus, $\phi_e(x) \leq \lambda$ for all x. By definition, $c^M(t)$ is the upper bound of $\phi_e(x)$ over x, and the inequality $c^M(t) \leq \lambda$ then follows. ∎

Remark 15.5.1 Note that $(\mathrm{TD}^*_{\mathrm{fn}})$ is a natural modification of the "bar-forces" formulation of the usual multi-load Truss Topology Design problem, see [5].

15.6 Solving $(\mathrm{TD}_{\mathrm{fn}})$ and $(\mathrm{TD}^*_{\mathrm{fn}})$ via interior point methods

Among numerical methods available for solving semidefinite programs like $(\mathrm{TD}_{\mathrm{fn}})$ and $(\mathrm{TD}^*_{\mathrm{fn}})$, the most attractive (in fact the only meaningful in the large scale case) are the recent interior point algorithms (for relevant general theory, see [10]). Here we discuss the corresponding possibilities. In what follows we restrict ourselves with outlining the main elements of the construction, since our goal now is not to present detailed description of the algorithms, but to demonstrate that

I. From the above semidefinite programs related to Truss Topology Design with robustness constraints, the most convenient for numerical processing by interior point methods is the problem $(\mathrm{TD}_{\mathrm{fn}})$

II. Solving $(\mathrm{TD}_{\mathrm{fn}})$ by *interior point path-following methods*, one has the possibility of generating, as a byproduct, good approximate solutions to the problem of interest $(\mathrm{TD}^*_{\mathrm{fn}})$, i.e., of recovering the primal design variables (bar volumes).

When solving a generic semidefinite program

$$\text{(SP)} \qquad \sigma^T \xi \to \min \mid \mathcal{A}(\xi) \in \mathbf{S}_+,$$

$\xi \in \mathbf{R}^N$ being the design vector, $\mathcal{A}(\xi)$ being an affine mapping from $\mathbf{R}^N$ to the space $\mathbf{S}$ of symmetric matrices of certain fixed block-diagonal structure, and $\mathbf{S}_+$ being the cone of positive semidefinite matrices from $\mathbf{S}$, by a path-following interior point method, one defines the family of barrier-type functions

$$F_s(\xi) = s\sigma^T \xi + \Phi(\mathcal{A}(\xi)), \;\; \Phi(\Xi) = -\ln \mathrm{Det}\, \Xi,$$

and traces the *central path* – the path of minimizers

$$\xi^*(s) = \operatorname*{argmin}_{\xi\in \mathrm{Dom}\, F_s} F_s(\xi).$$

If (SP) is strictly feasible (i.e., $\mathcal{A}(\xi)$ is positive definite for certain ξ) and the level sets

$$\{\xi \in \mathbf{R}^N \mid \mathcal{A}(\xi) \in \mathbf{S}_+, \sigma^T\xi \le a\},$$

$a \in \mathbf{R}$, are bounded, then the path ξ^* is well-defined and converges, as $s \to \infty$, to the optimal set of the problem. In the path-following scheme, one generates close (in certain exact sense) approximations ξ_i to the points $\xi^*(s_i)$ along certain sequence $\{s_i\}$ of penalty parameters "diverging to ∞ fast enough", thus generating a sequence of strictly feasible approximate solutions converging to the optimal set. Updating $(s_i, \xi_i) \mapsto (s_{i+1}, \xi_{i+1})$ is as follows: first, we increase, according to certain rule, the current value s_i to a larger value s_{i+1}. Second, we restore closeness to the path of the new point $\xi^*(s_{i+1})$ by running the *damped Newton method* – the recurrence

$$y \mapsto y^+ = y - (1+\lambda(F_s, y))^{-1}[\nabla^2_y F_s(y)]^{-1}\nabla_y F_s(y), \quad \lambda(F_s, y) = \sqrt{\nabla^T_y F_s(y)[\nabla^2_y F_s(y)]^{-1}\nabla_y F_s(y)}, \tag{15.160}$$

with s set to s_{i+1}. The recurrence is started at $y = \xi_i$ and is terminated when, for the first time, it turns out that $\lambda(F_{s_{i+1}}, y) \le \kappa$, $\kappa \in (0,1)$ being a once for ever fixed threshold. (Thus, the exact meaning of "closeness of a point ξ to the point $\xi^*(s)$ is given by the inequality $\lambda(F_s, \xi) \le \kappa$. In what follows, for the sake of definiteness, it is assumed that $\kappa = 0.1$). The resulting y is chosen as ξ_{i+1}, and the process is iterated.

It is known that

- it is possible to trace the path "quickly": with reasonable policy of updating the values of the penalty parameter, it takes, for any $T > 2$, no more than

  $$M = M(T) = O(1)\sqrt{\mu}\ln T$$

  Newton steps (15.160) to come from a point ξ_0 close to $\xi^*(s_0)$ to a point ξ_M close to $\xi^*(s_M)$, with $s_M \ge Ts_0$; here μ is the total row size of the matrices from $\mathbf{S}$ and $O(1)$ is an absolute constant;

- if ξ is close to $\xi(s)$, then the quality of ξ as an approximate solution to (SP) can be expressed via the value of s alone:

  $$\sigma^T\xi - \sigma^* \le \frac{2\mu}{s}, \tag{15.161}$$

  σ^* being the optimal value in (SP);

- being close to the path, it is easy to come "very close" to it: if $\lambda \equiv \lambda(F_s, y) \le 0.1$, then (15.160) results in

  $$\lambda^+ \equiv \lambda(F_s, y^+) \le 2.5\lambda^2. \tag{15.162}$$

Although the indicated remarks deal with the path-following scheme only, the conclusions related to the number of "elementary steps" required to solve a semidefinite program to a given accuracy and to the complexity of a step (dominated by the computational cost of the Newton direction, see (15.160)) are valid for other interior point methods for Semidefinite Programming. The "integrated" complexity characteristic of an interior point method for (SP) is the quantity

$$\mathcal{C} = \sqrt{\mu}\mathcal{C}_{\rm Nwt},$$

where $\mathcal{C}_{\rm Nwt}$ is the arithmetic cost of computing the Newton direction. Indeed, according to the above remarks, it takes $O(1)\sqrt{\mu}$ Newton steps to increase the value of the penalty by an absolute constant factor, or, which is the same, to reduce by the same factor the (natural upper bound for) inaccuracy of the current approximate solution.

Now let us look at the complexity characteristic $\mathcal{C}$ for the semidefinite programs related to (TD$_{\rm robust}$) . In the table below we write down the principal terms of the corresponding quantities (omitting absolute constant factors); it is assumed (as it is normally the case for Truss Topology Design) that

$$m = O(n^2); \quad q << n.$$

The expression for $\mathcal{C}_{\rm Nwt}$ corresponds to the "explicit" policy when we first assemble, in the natural manner, the Hessian matrix $\nabla^2_\xi F_s(\cdot)$ and then solve the resulting Newton system by traditional direct Linear Algebra routines like Choleski decomposition. It turns out that the specific structure of matrix inequalities in our problems[4)] allows to assemble the Hessians at relatively low cost, so that the cost of a single Newton step is dominated by the complexity of Choleski factorization of the Hessian, i.e., by cube of the design dimension of the corresponding problem. With this remark, we come to the results as follows:

Model	μ	$\mathcal{C}_{\rm Nwt}$	$\mathcal{C}$
$({\rm TD}_{\rm sd})$	m	m^3	$m^{3.5} \approx n^7$
$({\rm TD}_{\rm dl})$	m	m^3	$m^{3.5} \approx n^7$
$({\rm TD}_{\rm fn})$	qm	q^3n^3	$q^{3.5}n^4$
$({\rm TD}^*_{\rm fn})$	qm	q^6m^3	$q^{6.5}m^{3.5} \approx q^{6.5}n^7$

The reader should be aware that there are "implicit" schemes of computing the Newton direction in in $({\rm TD}^*_{\rm fn})$ with arithmetic cost $O(q^3n^3)$ (the same as in $({\rm TD}_{\rm fn})$). Thus, in fact the primal and dual problems in primal-dual pairs $(({\rm TD}_{\rm sd}),({\rm TD}_{\rm dl}))$, $(({\rm TD}_{\rm fn}),({\rm TD}^*_{\rm fn}))$ are theoretically equivalent in complexity; moreover, there are "symmetric" primal-dual methods which solve simultaneously the primal-dual pair of the problems at the complexity, respectively, $O(n^7)$ and $O(q^{3.5}n^4)$. Nevertheless, we believe that at the moment practical considerations still are in favour of "purely primal" methods as applied to $({\rm TD}_{\rm sd})$ in the first primal-dual pair and to $({\rm TD}_{\rm fn})$ in the second pair. The reason is that the feasible planes $\mathcal{L}$ in the "unfavourable" problems of the above pairs are given by linear equalities, while in the "favourable" components of the pairs they are parameterized (from the very beginning they are represented as images of affine mappings). Now, the theoretically efficient way to compute the Newton direction for an "unfavourable" problem represents the direction as the difference of certain "exactly known" vector and its projection on the orthogonal complement to $\mathcal{L}$. Such a computation is relatively unstable: rounding errors make the actually computed Newton directions non-parallel to $\mathcal{L}$, and the iterates eventually become far from the feasible plane. In order to overcome this instability, in the existing software for Semidefinite problems "expensive" Linear Algebra routines, like QR factorization, are used, at least at the final phase of computations. In contrast to this, in the "favourable" problems the Newton direction is computed in the space of parameters identifying a point on the feasible plane, so that there is no danger of being kicked off this plane.

With the above remarks, it is clear that among the semidefinite programs we introduced, the most convenient for numerical processing by interior point methods is $({\rm TD}_{\rm fn})$, as it was claimed in **I**. There is, however, an a priori drawback of this approach: what we need, are the bar volumes, and they "are not seen" at all in $({\rm TD}_{\rm fn})$. We are about to demonstrate that in order to overcome this difficulty it suffices to solve $({\rm TD}_{\rm fn})$ not by an arbitrary interior point method, but with a path-following one.

Assume that we are applying a path-following method to $({\rm TD}_{\rm fn})$ and have computed a point $\xi = (\Lambda, X, \rho)$ close (in the aforementioned sense) to the point $\xi^*(s)$. From (15.162) it follows that a small number of steps of the recurrence (15.160) started at ξ allows to come "very close" to $\xi^*(s)$ (6 steps of the recurrence restore $\xi^*(s)$ within machine accuracy). We may, therefore, assume for the sake of simplicity that we can "stand at the path", i.e., operate with $\xi^*(s)$ itself rather than with a tight approximation of the point[5]. It turns out that given $\xi^*(s)$, one can explicitly generate a feasible solution to $({\rm TD}^*_{\rm fn})$ of inaccuracy $\leq O(1/s)$. The exact statement is as follows:

Proposition 15.6.1 *Let $s > 0$, and let $\xi^*(s) = (\Lambda_s, X_s, \rho_s)$ be the minimizer of the function*

$$F_s(\Lambda, X, \rho) = s\left[2\,{\rm Tr}(QX^T) + V\rho\right] + \Phi(\mathcal{A}(\Lambda, X, \rho)) \tag{15.163}$$

over the set of strictly feasible solutions to $({\rm TD}_{\rm fn})$*. Here*

$$\Phi(S) = -\ln {\rm Det}\; S : {\rm int}\; \mathbf{S}_+ \to \mathbf{R}, \tag{15.164}$$

[4)] in particular, the fact that in TTD design each of the vectors b_i has $O(1)$ nonzero entries – at most 4 in the case of 2D and at most 6 in the case of 3D trusses

[5] This is an idealization, of course, but it is as well-motivated as the standard model of precise real arithmetic. We could replace in the forthcoming considerations $\xi^*(s)$ by its tight approximation, with minor modification of the construction, but we do not think it makes sense

$\mathbf{S}$ *is the space of block-diagonal symmetric matrices with m diagonal blocks of the size $(q+1)\times(q+1)$ each, and*

$$\mathcal{A}(\Lambda,X,\rho)=\mathrm{Diag}\left\{\begin{pmatrix}\Lambda & X^Tb_i\\ b_i^TX & \rho\end{pmatrix},\ i=1,...,m\right\}, \tag{15.165}$$

Then the matrix

$$R(s)\equiv\mathrm{Diag}\left\{\begin{pmatrix}L_i & d_i\\ d_i^T & t_i\end{pmatrix}\ i=1,...,m\right\}=s^{-1}\mathcal{A}^{-1}(\Lambda_s,X_s,\rho_s)\quad\left[=-s^{-1}\nabla_S|_{S=\mathcal{A}(\Lambda_s,X_s,\rho_s)}\Phi(S)\right] \tag{15.166}$$

is such that $\sum_{i=1}^m L_i=\lambda_sI_q$ for some real λ_s, and $(R(s),\lambda_s)$ is a feasible solution to $(\mathrm{TD}^*_{\mathrm{fn}})$ with the value of the objective

$$\lambda_s\le c^*+\frac{\mu}{s}, \tag{15.167}$$

where c^* is the optimal value in $(\mathrm{TD}^*_{\mathrm{fn}})$ and $\mu=m(q+1)$ is the total row size of the matrices from $\mathbf{S}$.

The proposition is an immediate consequence of general results of [10]; to make the paper self-contained, below we present a direct proof.

Let us set $Y=\mathcal{A}(\Lambda_s,X_s,\rho_s)$, $Z=Y^{-1}$, so that

$$R(s)=s^{-1}Z;\quad \nabla\Phi(Y)=-Z.$$

The set G of strictly feasible solutions to $(\mathrm{TD}_{\mathrm{fn}})$ is comprised of all triples $\xi=(\Lambda,X,\rho)$ which correspond to positive definite $\mathcal{A}(\xi)$ and are such that $\mathrm{Tr}\,\Lambda=1$; this is an open convex subset in the hyperplane given by the equation $\mathrm{Tr}\,\Lambda=1$. Since $\xi^*(s)=(\Lambda_s,X_s,\rho_s)$ is the minimizer of F_s over G, we have, for certain real p,

$$\nabla_\Lambda F_s(\xi^*(s))=pI_q;\quad \nabla_X F_s(\xi^*(s))=0;\quad \nabla_\rho F_s(\xi^*(s))=0.$$

Substituting the expression for F_s and $\mathcal{A}$, we obtain

$$\begin{array}{rcccccc}\sum_{i=1}^m L_i & \equiv & [\mathcal{A}^*R(s)]_\Lambda & \equiv & -s^{-1}\left[\mathcal{A}^*\nabla\Phi(Y)\right]_\Lambda & = & -s^{-1}pI_q,\\ 2\sum_{i=1}^m b_id_i^T & \equiv & [\mathcal{A}^*R(s)]_X & \equiv & -s^{-1}\left[\mathcal{A}^*\nabla\Phi(Y)\right]_X & = & 2Q,\\ \sum_{i=1}^m t_i & \equiv & [\mathcal{A}^*R(s)]_\rho & \equiv & -s^{-1}\left[\mathcal{A}^*\nabla\Phi(Y)\right]_\rho & = & V\end{array}$$

(here $[\cdot]_\Lambda$, $[\cdot]_X$ and $[\cdot]_\rho$ denote, respectively, the Λ-, the X- and the ρ-component of the design vector of $(\mathrm{TD}_{\mathrm{fn}})$). Note also that Y (and therefore Z) is positive definite. We see that $(R(s),\lambda\equiv -s^{-1}p)$ indeed is a feasible solution of $(\mathrm{TD}^*_{\mathrm{fn}})$.

Now, if (Λ,X,ρ) is a feasible solution to $(\mathrm{TD}_{\mathrm{fn}})$, and

$$\left(R\equiv\mathrm{Diag}\left\{\begin{pmatrix}M_i & c_i\\ c_i^T & r_i\end{pmatrix},\ i=1,...,m\right\},\lambda\right)$$

is a feasible solution to $(\mathrm{TD}^*_{\mathrm{fn}})$, then

$$2\,\mathrm{Tr}(QX^T)+V\rho=$$

[since $[\mathcal{A}^*R]_\Lambda=\lambda I_q$, $[\mathcal{A}^*R]_X=2Q$, $[\mathcal{A}^*R]_\rho=V$ by the constraints of $(\mathrm{TD}^*_{\mathrm{fn}})$ and $\mathrm{Tr}\,\Lambda=1$ by the constraints of $(\mathrm{TD}_{\mathrm{fn}})$]

$$=\left[\mathrm{Tr}([\mathcal{A}^*R]_XX^T)+[\mathcal{A}^*R]_\rho\rho+\mathrm{Tr}([\mathcal{A}^*R]_\Lambda\Lambda)\right]-\lambda=$$
$$=\mathrm{Tr}(R\mathcal{A}(\Lambda,X,\rho))-\lambda,$$

whence

$$[2\,\mathrm{Tr}(QX^T)+V\rho]+\lambda=\mathrm{Tr}(R\mathcal{A}(\Lambda,X,\rho)).$$

Since the optimal values in $(\mathrm{TD}_{\mathrm{fn}})$ and $(\mathrm{TD}^*_{\mathrm{fn}})$ are, as we know from Fenchel-Rockafellar Duality Theorem, negations of each other, we come to

$$\epsilon[\Lambda,X,\rho]+\epsilon^*[R,\lambda]=\mathrm{Tr}(R\mathcal{A}(\Lambda,X,\rho)); \tag{15.168}$$

here $\epsilon[\Lambda,X,\rho]$ is the accuracy of the feasible solution (Λ,X,ρ) of $(\mathrm{TD}_{\mathrm{fn}})$ (i.e. the value of the objective of $(\mathrm{TD}_{\mathrm{fn}})$ at (Λ,X,ρ) minus the optimal value of the problem), and $\epsilon^*[\cdot]$ is similar accuracy in $(\mathrm{TD}^*_{\mathrm{fn}})$.

Specifying (Λ,X,ρ) as (Λ_s,X_s,ρ_s) and (R,λ) as $(R(s),\lambda_s)$, we make the right hand side of (15.168) equal to

$$\mathrm{Tr}(R(s)Y)=s^{-1}\,\mathrm{Tr}(ZY)=s^{-1}\,\mathrm{Tr}(Y^{-1}Y)=s^{-1}\mu,$$

and with this equality (15.168) implies (15.167). ■

15.7 Numerical Examples

Let us illustrate the developed approach by few examples.

Example 1. Our first example deals with the toy problem presented on Fig. 1; as it was explained in Section 15.2.2, here the single-load optimal design results in unstable truss capable to carry only very specific loads; the compliance of the truss with respect to the given load is 16.000. Now let us apply approach **B** from Section 15.2.3, where the robustness constraint is imposed before solving the problem and corresponds to "active" nodes – those where the given load is applied. When imposing robustness requirement, we choose Q as explained in Section 15.2.3. Namely, in our case we have 2 fixed and 4 free nodes, so that the dimension n of the space of virtual nodal displacements is $2 \times 4 = 8$. Since all free nodes are active, the ellipsoid of loads in robust setting is full-dimensional ($q = n = 8$); this ellipsoid is chosen as explained in Section 15.2.3 – one of the half-axis is the given load, and the remaining 7 half-axes are 10 times smaller. The corresponding matrix (rounded to 3 decimal places after the dot) is

$$Q = \begin{pmatrix} 2.000 & 0.014 & -0.026 & 0.117 & -0.063 & 0.170 & -0.264 & -0.054 \\ 0 & 0.235 & 0.216 & 0.125 & -0.032 & -0.161 & -0.070 & 0.104 \\ 0 & -0.040 & -0.107 & 0.099 & 0.311 & -0.158 & -0.117 & -0.035 \\ 2.000 & 0.045 & 0.137 & -0.263 & 0.162 & 0.039 & 0.002 & 0.043 \\ 0 & -0.202 & 0.148 & -0.081 & -0.111 & -0.190 & -0.124 & -0.164 \\ -2.000 & 0.149 & -0.108 & -0.203 & -0.030 & 0.006 & -0.210 & -0.009 \\ -2.000 & -0.089 & 0.219 & 0.057 & 0.129 & 0.203 & -0.052 & -0.003 \\ 0 & 0.173 & 0.028 & 0.020 & 0.042 & 0.035 & 0.098 & -0.341 \end{pmatrix}$$

(to relate Q to the nodal structure presented on Fig. 1, note that the coordinates of virtual displacements are ordered as 2X,2Y,3X,3Y,5X,5Y,6X,6Y, where, say, 3X corresponds to the displacement of node #3 along the X-axis).

The result of "robust" design is presented on Fig. 3.

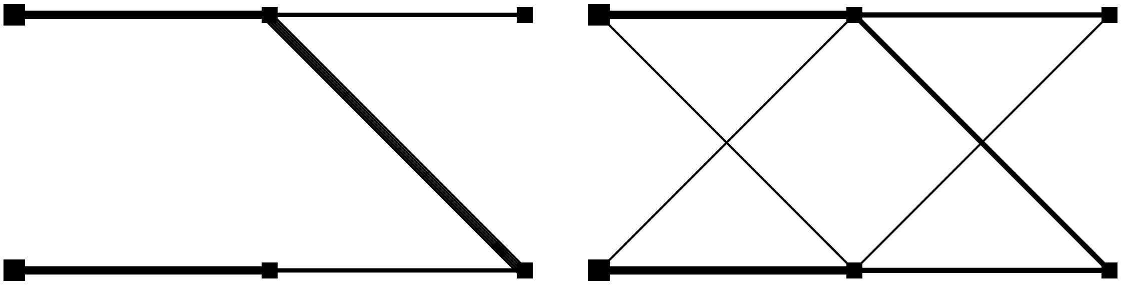

Fig. 3: Optimal design without (left) and with (right) robustness constraints

Problem setting	Compliance	Bars, node : node	Bar volumes, %
without robustness constraints	16.000	1 : 2	25.00
		4 : 5	25.00
		3 : 5	25.00
		5 : 6	12.50
		2 : 3	12.50
with robustness constraints	17.400	4 : 5	24.48
		1 : 2	24.48
		3 : 5	23.68
		2 : 3	11.95
		5 : 6	11.95
		2 : 4	1.27
		1 : 5	1.27
		2 : 6	0.92

Now the maximum over the 8-dimensional loading ellipsoid compliance becomes 17.400 (8.75% growth). But the compliance of the truss with respect to the load f is 16.148, i.e., it is only by 0.9% larger than for the truss given by single-load setting.

Example 2: "Console". The second example deals with approach **A** from Section 15.2.3, where the robustness constraint is used for postoptimality analysis. The left part of Fig. 4 represents optimal single-load design for 9×9 nodal grid on 2D plane; nodes from the very left column are fixed, the remaining nodes are free, and the load is the unit force acting down and applied at the mid-node of the very right column (long arrow). The compliance of the resulting truss w.r.t. f^*, in appropriate scale, is 1.00. Now note that the compliance of t with respect to very small (of magnitude $0.005 \parallel f^* \parallel$) "occasional" load (short arrow) applied at properly chosen node is > 8.4 ! Thus, in fact t is highly unstable.

The right part of Fig. 4 represents the truss obtained via postoptimality design with robustness constraint. We marked the nodes incident to the bars of t (there were only 12 of them) and formed a new design problem with the nodal set comprised of these marked nodes, and the tentative bars given by all 66 possible pair connections in this nodal set (in the original problem, there were 2040 tentative bars). The truss represented in the right part corresponds to optimal design with robustness constraint imposed at all 10 free nodes of this ground structure in the same way as in the previous example (i.e., the first column in the 20×20 matrix Q is the given load f^*, and the remaining 19 columns formed orthogonal basis in the orthogonal complement to f^* in of 20-dimensional space of virtual displacements of the construction; the Euclidean lengths of these additional columns were set to 0.1 (10% of the magnitude of f^*).

The maximal compliance, over the resulting ellipsoid of loads, of the "robust" truss is now 1.03, and its compliance with respect to f is 1.0024 – i.e., it is only by 0.24% larger than the optimal compliance c^* given by the single-load design; at the same time, the compliance of the new truss with respect to all "occasional" loads of magnitude 0.1 is at most by 3% greater than c^*.

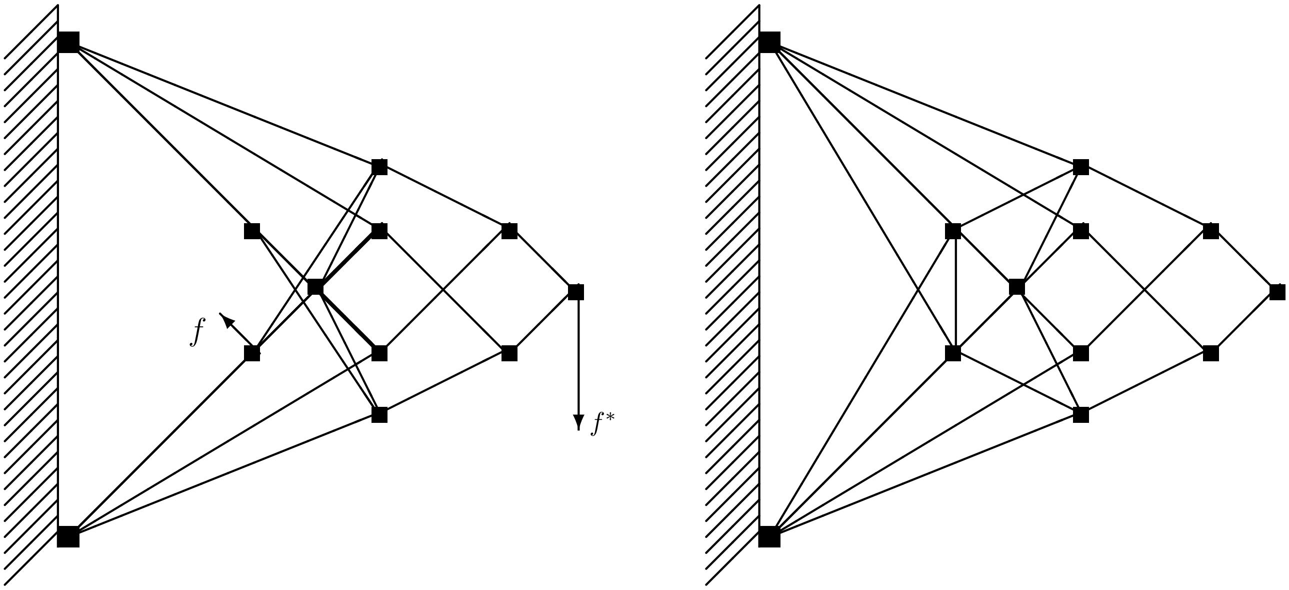

Fig. 4: Single-load optimal design (left) and its postoptimal "robust correction" (right)

Example 3: "$N \times 2$- truncated pyramids". The examples below deal with simple 3D trusses. The nodal set is comprised of $2N$ points. N "ground" nodes are the vertices of equilateral N-polygon in the plane $z = 0$:

$$x_i = \cos(2\pi i/N),\ y_i = \sin(2\pi i/N),\ z_i = 0,\ i = 1, ..., N,$$

and N "top" nodes are the vertices of twice smaller concentric polygon in the plane $z = 2$:

$$x_i = \cos(2\pi i/N),\ y_i = \sin(2\pi i/N),\ z_i = 2,\ i = N+1, ..., 2N.$$

The ground nodes are fixed, the top ones are free. The ground structure is comprised of all pair connections of the nodes, except connections between the ground – fixed – ones.

We dealt with two kinds of loading scenarios, referred to, respectively, as "$N \times 2$s-" and "$N \times 2$m" – design data. $N \times 2$s-data corresponds to a singleton scenario set, where the load is comprised of N nearly horizontal forces acting at the top nodes and "rotating" the construction: the force acting at i-th node, $i = N+1, ..., 2N$, is

$$f_i = \alpha(\sin(2\pi i/N), -\cos(2\pi i/N), -\rho),\ i = N+1, ..., 2N \qquad (15.169)$$

where ρ is small parameter and α is normalizing coefficient which makes the Euclidean length of the load equal to 1 (i.e., $\alpha = 1/\sqrt{N(1+\rho^2)}$). $N \times 2$m-data correspond to N-scenario design where the forces (15.169) act nonsimultaneously (and are renormalized to be of unit length, i.e., $\alpha = 1/\sqrt{1+\rho^2}$).

Along with the traditional "scenario design" (single-load in the case of s-data and multi-load in the case of m-data), we carried out "robust design" where we minimized the maximum compliance with respect to a full-dimensional ellipsoid of loads M_θ – the "ellipsoidal envelope" of the unit ball in the linear span $L(F)$ of the scenario loads and the ball of radius θ in the orthogonal complement of $L(F)$ in the $3N$-dimensional space of virtual displacements of the nodal set. In other words, dim $L(F)$ of the principal half-axes of M_θ are of unit length and span $L(F)$, and the remaining principal half-axes are of length θ. In our experiments, we used $\theta = 0.3$.

The resulting structures Fig. 5 (data $N \times 2$s) and Fig. 6 (data $N \times 2$m), and the corresponding compliances – in Table 1.

Table 1. Compliances

Topo-file	Scenario design			Robust design	
	Compl(Scen)	Compl(0.1)	Compl(0.3)	Compl(Scen)	Compl(0.3)
3x2s	1.0000	7.5355	67.820	1.0029	1.0029
4x2s	1.0000	12.209	109.88	1.00280	1.0028
5x2s	1.0000	2.7311	24.580	1.0022	1.0022
3x2m	1.0000	1.2679	1.2679	1.0942	1.0943
4x2m	1.0000	4.1914	37.722	1.2903	1.2903
5x2m	1.0000	1.5603	1.6882	1.5604	1.5604

In Table 1, Compl(Scen) means the maximum compliance of the designed structure w.r.t. the set of loading scenarios given by the corresponding data, while Compl(θ), $\theta = 0.1, 0.3$ is the maximum compliance with respect to the ellipsoid M_θ. In order to make the comparison more clear, we normalize the data in each row to make the compliance of the truss given by Scenario design with respect to the underlying set of scenarios equal to 1.

The summary of the numerical results in question is as follows.

- $N \times 2$s design data. Although the trusses given by the scenario and the robust designs have the same topology and differ only in bar sizes, this difference is quite significant. The scenario design results in highly unstable constructions: appropriately chosen "occasional" load 10 times less than the scenario may result in 2.6 - 13.0 times larger compliance than the one with respect to the scenario load; when the occasional load is allowed to be 30% of the scenario one, the ratio in question may become 15 - 100. Note that bad robustness of the trusses given by the scenario design has very simple origin: in the limiting case of $\rho = 0$ (purely horizontal rotating load) the top horizontal bars (see Fig. 5) disappear at all, and the optimal truss given by the usual single-load design becomes completely unstable.

  The robust design associated with the ellipsoid $M_{0.3}$ ("occasional" loads may be as large as 30% of the scenario one) results in trusses nearly optimal with respect to the scenario load ("nonoptimality" is at most 0.3%). Surprisingly enough, for the trusses given by the robust design the maximum compliance with respect to the ellipsoid of loads is the same as their compliance with respect to the scenario load. Thus, in the case in question the robustness is "almost costless".

- $N \times 2$m design data. Here the trusses given by the scenario design are of course much more stable than in the case of $N \times 2$s data, and both kinds of design possess their own advantages and drawbacks. On one hand, the maximum, over the ellipsoid $M_{0.3}$ of loads, compliance of the truss given by the scenario design is considerably larger than the optimal value of this quantity (by 27% for $N = 3$, by factor 37.7 for $N = 4$, and by 56% for $N = 5$). On the other hand, the maximum, over the scenario set, compliance of the truss given by robust design also is considerably larger than the optimal value of this quantity (by 9% for $N = 3$, by 29% for $N = 4$, and by 56% for $N = 5$). Thus, it is difficult to say which design – the scenario or the robust one – results in better construction.

  The example in question suggests a seemingly better approach to ensuring robustness than those mentioned in Section 15.2.3, namely, as follows. Given a finite scenario set F, we, as in Section

15.2.3, embed it into an ellipsoid $M = \{Qu \mid u \in \mathbf{R}^q, u^T u \leq 1\}$ and solve the resulting problem $(\mathrm{TD}_{\mathrm{robust}})$; let c^*_{robust} be the corresponding optimal value. After this value is found, we increase it in certain fixed proportion $1+\chi$, say, by 10%, and solve the problem

$$\begin{array}{c}\text{find } t \in T \text{ which minimizes the compliance } c^F(t) = \max_{f\in F} c_f(t) \\ \text{subject to } c^M(t) \equiv \max_{f\in M} c_f(t) \leq (1+\chi)c^*_{\mathrm{robust}}.\end{array}$$

Note that the latter problem can be posed as a semidefinite program which only slightly differs from $(\mathrm{TD}_{\mathrm{sd}})$:

The corresponding primal semidefinite program is

$$\begin{array}{lrcl} & \min\limits_{t\in\mathbf{R}^m,\tau\in\mathbf{R}} & \tau & \\ \text{s.t.} & & & \\ & \begin{pmatrix} \tau & f^T \\ f & \sum_{i=1}^m t_i A_i \end{pmatrix} & \geq & 0,\ \forall f \in F \\ & \begin{pmatrix} a & Q^T \\ Q & \sum_{i=1}^m t_i A_i \end{pmatrix} & \geq & 0 \end{array}$$

where

$$a = (1+\chi)c^*_{\mathrm{robust}}.$$

The dual to the latter problem is the computationally more convenient program

$$\begin{array}{l} \min \left\{ a\,\mathrm{Tr}(\Lambda) + 2\,\mathrm{Tr}(QX^T) + 2\sum_{f\in F} f^T x_f + V\rho \right\} \\ \text{s.t.} \\ \begin{array}{rcl} \begin{pmatrix} \Lambda & X^T b_i \\ b_i^T X & \sigma_i \end{pmatrix} & \geq & 0,\ i=1,...,m \\ \sigma_i + \sum_{f\in F} \frac{(b_i^T x_f)^2}{\lambda_f} & \leq & \rho,\ i=1,...,m \\ \lambda_f & \geq & 0,\ f \in F \\ \sum_{f\in F}^k \lambda_f & = & 1 \end{array} \end{array}$$

the design variables being $\Lambda \in \mathbf{S}^k$, $X \in \mathbf{R}^{n\times q}$, $\sigma \in \mathbf{R}^n$, $\{(\lambda_f, x_f) \in \mathbf{R}\times\mathbf{R}^n\}_{f\in F}$, and $\rho \in \mathbf{R}$.

3x2s, Scenario design

3x2s, Robust design

4x2s, Scenario design

4x2s, Robust design

5x2s, Scenario design

5x2s, Robust design

Fig. 5. Scenario design vs. Robust design, single "rotating" load
($\rho = 0.001$ for 3×2s and 4×2s, $\rho = 0.01$ for 5×2s)

`3x2m`, Scenario design `3x2m`, Robust design

`4x2m`, Scenario design `4x2m`, Robust design

`5x2m`, Scenario design `5x2m`, Robust design

Fig. 6. Scenario design vs. Robust design, multiple "extending" loads
($\rho = 0.001$ for 3×2m and 4×2m, $\rho = 0.01$ for 5×2m)

The reported numerical experiments were carried out with the LMI Control toolbox [7], the only software for Semidefinite Programming available for us for the moment. The Projective interior point method ([10], Chapter 5) implemented in the Toolbox is of the potential reduction rather than of the path-following type, and we were enforced to add to the Toolbox solver a "centering" interior point routine which transforms a good approximate solution to $(\mathrm{TD}_{\mathrm{fn}})$ into another solution of the same quality belonging to the central path, which enabled us to recover the optimal truss, as it is explained in Section 15.6. The complexity of solving $(\mathrm{TD}_{\mathrm{fn}})$ by the toolbox solver was moderate, as it is seen from the following table:

Table 2. Computational complexity

Problem	Scenario design			Robust design		
	$(N_{\mathrm{dsg}}, N_{\mathrm{LMI}}, N_{\mathrm{img}})$	Nwt	CPU	$(N_{\mathrm{dsg}}, N_{\mathrm{LMI}}, N_{\mathrm{img}})$	Nwt	CPU
Example 2	(146,2041,6121)	75	$3'58''$	(611,67,15247)	95	$24'42''$
`3x2s`	(11,13,37)	14	$0.2''$	(127,13,661)	62	$14.5''$
`4x2s`	(14,23,67)	16	$0.4''$	(223,23,2003)	77	$1'18''$
`5x2s`	(17,36,106)	17	$0.6''$	(346,36,4761)	59	$3'13''$
`3x2m`	(31,13,121)	16	$0.4''$	(127,13,661)	101	$24''$
`4x2m`	(53,23,331)	23	$1.5''$	(223,23,2003)	65	$1'6''$
`5x2m`	(81,36,736)	23	$3''$	(346,36,4761)	65	$3'32''$

In the table:

N_{dsg}	– number of design variables in $(\mathrm{TD}_{\mathrm{fn}})$
N_{LMI}	– number of Linear Matrix Inequalities in $(\mathrm{TD}_{\mathrm{fn}})$
N_{ing}	– total image dimension of $(\mathrm{TD}_{\mathrm{fn}})$, i.e., the dimension of the corresponding semidefinite cone
Nwt	– number of Newton steps performed by the interior point solver when solving $(\mathrm{TD}_{\mathrm{fn}})$
CPU	– solution time (workstation RS 6000)

15.8 Concluding remarks

Uncertainty of the data is a generic property of optimization models of the real world origin; consequently, "robust reformulation" of an optimization model as a way to improve applicability of the resulting solution is a very traditional idea in Mathematical Programming, and different approaches to implement this idea were proposed. One of the best known approaches is *Stochastic Programming*, where uncertainty is assumed to be of stochastic nature. Another approach is *robust optimization* (see [9] and references therein); here, roughly speaking, the "robust solution" should not necessarily be feasible for all "allowed" data, and the "optimal robust solution" minimizes the sum of the original objective and a penalty for infeasibilities, the infeasibilities being taken over a finite set of scenarios. The approach used in our paper is somewhat different: a solution to the "stabilized" problem should be feasible for all allowed data. This approach is exactly the one used in Robust Control. The goal of this concluding section is to demonstrate that the approach developed in the paper can be naturally extended to other Mathematical Programming problems. To this end let us look what in fact was done in Section 15.2.

I. We start with an optimization program in the "conic" form

$$(P) \qquad cTu \to \min \mid Au \in K,\ u \in E,$$

where u is the design vector, A is $M \times N$ matrix, K is closed convex cone in $\mathbf{R}^M$ and E is an affine plane in $\mathbf{R}^N$.

This is exactly the form of a single-load TTD problem $\min\{\sigma \mid \sigma \geq c_f(t), t \in T\}$ (see Section 15.2.1): to cast TTD as (P) it suffices to specify (P) as follows:

- $u = (t, \tau, \sigma) \in \mathbf{R}^m \times \mathbf{R} \times \mathbf{R}$;
- $E = \{(t, \tau, \sigma) \mid \tau = 1, \sum_{i=1}^m t_i = V\}$;
- K is the direct product of the cone of positive semidefinite symmetric $(n+1) \times (n+1)$ matrices ("matrix part") and $\mathbf{R}^m_+$ ("vector part")
- the "vector" part of the linear mapping $(t, \tau, \sigma) \mapsto A(t, \tau, \sigma)$ is t, and the "matrix" part is $\begin{pmatrix} \sigma & \tau f^T \\ \tau f & A(t) \end{pmatrix}$, f being the load in question.

II. We say that the data in (P) (entries in the data matrix A) are inexact (in TTD, these are entries associated with the load vector f). We model the corresponding uncertainty by the assumption hat $A \in \mathcal{U}$, where $\mathcal{U}$ is certain ellipsoid in the space of $M \times N$ matrices[6]. Accordingly, we impose on the decision u the requirement to be *robust feasible*, i.e., to satisfy the inclusions $u \in A$ and $Au \in K$ for *all* possible data matrices $A \in \mathcal{U}$. This leads to our *robust reformulation*

$$(P_{\rm st}) \qquad c^T u \to \min \mid \ u \in E, \ Au \in K \ \forall A \in \mathcal{U}$$

of (P).

Note that this is a general form of the approach we have used in Section 15.2;and the goal of the remaining sections was to realize, for the case when (P) is the single load TTD problem, what is $(P_{\rm st})$ as a Mathematical Programming problem and how to solve it efficiently.

Now note that (P) is a quite general form of a Convex Programming problem; the advantage of this conic form is that it allows to separate the "structure" of the problem (c, K, E) and the "data" (A)[7]. The data now become a quite tractable entity – simply a matrix. Whenever program in question can be naturally posed in the conic form, we can apply the above approach to get a "robust reformulation" of (P). Let us look at a couple of examples.

Robust Linear Programming. Let K in (P) be the nonnegative orthant; this is exactly the case when (P) is a Linear Programming problem in the canonical form[8]. It is shown in [6] that $(P_{\rm st})$ is a conic quadratic program (i.e., a conic program with K being a direct product of the second order cones).

Robust Quadratic Programming. Let K be a direct product of the second order cones, so that (P) is a conic quadratic program (a natural extension of the usual quadratically constrained convex quadratic program). It can be verified (see [6]) that in this case, under mild restrictions on the structure of the uncertainty ellipsoid $\mathcal{U}$, the problem $(P_{\rm st})$ can be equivalently rewritten as a semidefinite program (a conic program with K being the cone of positive semidefinite symmetric matrices).

Note that in these examples $(P_{\rm st})$ is quite tractable computationally, in particular, it can be efficiently solved by interior point methods.

A somewhat "arbitrary" element in the outlined general approach is that we model uncertainty as an *ellipsoid*. Note, anyhow, that *in principle* the above scheme can be applied any other uncertainty set $\mathcal{U}$, and the actual "bottleneck" is our ability to solve efficiently the resulting problem $(P_{\rm st})$. Note that the robust problem $(P_{\rm st})$ always is convex, so that there is a sufficient condition for its "efficient solvability". The condition, roughly speaking (for the details, see [8]), is that we should be able to equip the feasible domain

$$G = \{u \mid \ u \in E, Au \in K \ \forall A \in \mathcal{U}\}$$

of $(P_{\rm st})$ with a *Separation oracle* – a "computationally efficient" routine which, given on input u, reports on output whether $u \in G$, and if it is not the case, returns a linear form which separates G and u. Whether this sufficient condition is satisfied or not depends on the geometry of $\mathcal{U}$ and K, and the "more complicated" is $\mathcal{U}$, the "simpler" should K be. When $\mathcal{U}$ is very simple (a polytope given as a convex hull of a finite set), K could be an arbitrary "tractable" cone (one which can be equipped with a Separation oracle); when $\mathcal{U}$ is an ellipsoid, K for sure could be the nonnegative orthant or a direct product of the second order cones. On the other hand, if K is simple (the nonnegative orthant, as in the Linear Programming case), $\mathcal{U}$ could be more complicated than an ellipsoid – e.g., it could be an intersection of finitely many ellipsoids. Under mild regularity assumptions, in the latter case $(P_{\rm st})$ turns out to be a conic quadratic program [6]. In other words, there is a "tradeoff" between the *flexibility* and the *tractability*, i.e., between the ability to express uncertainties, on one hand, and the ability to produce computationally tractable problems $(P_{\rm st})$, on the other hand.

[6] here, as in the main body of the paper, a k-dimensional ellipsoid in $\mathbf{R}^M$ is, by definition, the image of the unit Euclidean ball in $\mathbf{R}^k$ under an affine embedding of $\mathbf{R}^k$ into $\mathbf{R}^M$.

[7] in some applications the objective c should be treated as a part of the data rather than the structure. One can easily reduce this case to the one in question by evident equivalent reformulation of (P).

[8] up to the fact that the mapping $u \mapsto Au$ is assumed to be linear rather than affine. This assumption doe not restrict generality, since we incorporate into the model the affine constraint $u \in E$; at the same time, the homogeneous form $Au \in K$ of the nonnegativity constraints allows to handle both uncertainties in the matrix of the linear inequality constraints and those in the right hand side vector.

We strongly believe that the approach advocated here is promising and worths investigation, and we intend to devote to it a separate paper.

Bibliography

[1] Achtziger, W., Bendsøe, M.P., Ben-Tal, A., and Zowe, J. "Equivalent displacement-based formulations for maximum strength truss topology design", *Impact of Computing in Science and Engineering* **4** (1992), 315-345.

[2] Bendsøe, M.P., Ben-Tal, A., and Zowe, J. "Optimization methods for Truss Geometry and Topology Design", *Structural Optimization*, **7** (1994), 141-159.

[3] Ben-Tal, A. and Bendsøe, M.P. "A new method for optimal Truss Topology Design", *SIAM Journal of Optimization* **3** (1993), 322-358.

[4] Ben-Tal, A., Kočvara, M., and Zowe, J. "Two nonsmooth approaches to simultaneous geometry and topology design of trusses" - in: *Topology Design of Structures*, Bendsøe, M.P. (Ed.), Proceedings of NATO-ARW, Sesimbra, Portugal, 1992.

[5] Ben-Tal, A. and Nemirovski, A. "Potential reduction polynomial time method for Truss Topology Design", *SIAM Journal of Optimization* **4** (1994), 596-612.

[6] Ben-Tal, A., and Nemirovski, A. "Robust Convex Programming" – manuscript, Optimization Laboratory, Faculty of Industrial Engineering and Management at Technion, October 1995

[7] Gahinet, P., Nemirovski, A., Laub, A.J., and Chilali, M. "LMI Control Toolbox", The MathWorks Inc., 1995.

[8] Grötschel, M., Lovasz, L., and Schrijver, A. *The Ellipsoid Method and Combinatorial Optimization*, Springer, Heidelberg, 1988.

[9] Mulvey, J.M., Vanderbei, R.J., and Zenios, S.A., "Robust optimization of large-scale systems", *Operations Research* **43** (1995), 264-281.

[10] Nesterov, Yu., and Nemirovski, A. *Interior point polynomial methods in Convex Programming*, SIAM Series in Applied Mathematics, Philadelphia, 1994.

[11] Rockafellar, R.T. *Convex Analysis*, Princeton University Press, 1970.

[12] Rozvany, G., Bendsøe, M.P., and Kirsch, U., "Layout optimization of structures", *Applied Mechanics Reviews* **48** (1955) No.2, 41-119.

Appendix III: Minicourse on Polynomial Time Optimization Algorithms

Polynomial time methods in Convex Programming

What follows is a mini-course aimed to give some impression on *polynomial time methods* in Convex Programming. We start with discussing the notion of a polynomial time method, then present basic results on the methods of this type for general oracle-represented convex programs and mainly focus on *interior point* polynomial time methods for well-structured convex programs.

16.1 Polynomial methods in Convex Programming: what is it?

16.1.1 The concept

The notion of a *polynomial time method* – a method solving a computational problem in polynomial in the size of the problem number of elementary steps – was introduced explicitly in 1965 by Cobham and Edmonds [2, 3], in the context of Discrete Mathematics. The essence of the definition is as follows. A generic problem in question is formalized as a set $\mathcal{P}$ of pairs (x, y) of finite words in certain finite alphabet, say, the 0-1 one. Problem instances are identified by the inputs – finite words x, and an algorithm $\mathcal{A}$ solving the problem must, given such an input x, convert it into a solution to the input – a word y such that $(x, y) \in \mathcal{P}$ – or to detect that no such y exists. Here the word "algorithm" can be specified in any way known from Mathematical Logic. The algorithm $\mathcal{A}$ solving the problem $\mathcal{P}$ is called *polynomial*, if its *running time* $\mathcal{T}_{\mathcal{A}}(x)$ – number of elementary steps on input x – is bounded by a polynomial of the length of the input:

$$(\forall x): \quad \mathcal{T}_{\mathcal{A}}(x) \leq \pi(l(x)),$$

where $\pi(\cdot)$ is a polynomial and $l(x)$ is the *size* of the input – number of letters in the word x. Problem $\mathcal{P}$ is called *polynomially solvable*, if it admits a polynomial time solution algorithm; this property of the problem is independent of what is the particular model of an algorithm we use.

E.g., The Linear Algebra problem

(EqSys) *given a $m \times n$ matrix A and m-dimensional vector b with rational entries, find a rational solution to the system of linear equations $Ax = b$ or detect that no such solution exists*

can be easily included in the outlined framework and turns out to be polynomially solvable. Indeed, there is no difficulty to encode the rational data of the problem by a single finite binary word, same as encode in this manner candidate solutions. And it turns out that the Gauss elimination algorithm admits polynomial time implementation, so that the problem in question is polynomially solvable.

The idea behind the partitioning of problems and algorithms into polynomial and non-polynomial is to distinguish between "efficiently solvable" problems and "efficient" algorithms, on one side, and "intractable" problems and "inefficient" algorithms, on another, and the notion of polynomiality proved itself to be quite satisfactory for this purpose, at least from the theoretical viewpoint. At the same time, this notion can be used directly only in the context of Discrete Mathematics, where we have no difficulties with finite encoding of problem instances and their solutions. In "continuous" computational problems, however, this notion as it is hardly can be used: how could we speak about finite encoding of continuous data and continuous candidate solutions; besides this, in most of the cases we have no hope to get an exact solution in finitely many elementary steps, at least with a reasonable notion of a step.

There are, of course, continuous problems for which the above machinery works. For example, the already indicated Linear Algebra problem is, in fact, continuous, but its rational version, as it was explained, can be captured by the approach in question; after Khachiyan's paper of 1979, we know that the same is true for systems of *linear inequalities* with rational data – for Linear Programming; this problem also is covered by the indicated approach and also is polynomially solvable. Note, however, that already in these examples we were enforced to restrict somehow the problem – to pass from its natural setting as a problem with real data to the one with rational data – and heavily exploited the very specific fact that a solvable instance here admits a rational solution; this latter property is lost when we pass, e.g., to "evidently efficiently solvable" quadratic equations with one unknown.

In order to speak about efficient methods for "continuous" computational problems, we need therefore to extend the notion of "polynomial time" method in order to capture at least the following two features of continuous problems

- real, and thus not admitting natural finite encoding, data and candidate solutions;
- impossibility, in general, to find an *exact* solution in finitely many steps.

The corresponding formalization of the notions of a polynomial time algorithm and efficiently solvable problem within the framework of *real arithmetic model of computations* was proposed and developed during last years, starting with the seminal paper of L. Blum, M. Shub and S. Smale [1]. It makes no sense to present here general definitions of this type; what we are about to do is to give a version of these definitions for the situation we are interested in – nonlinear convex optimization. In order to distinguish between the "real arithmetic based" polynomial time algorithms we are going to define and the usual polynomial time algorithms of Discrete Mathematics, we shall call the latter algorithms "fully polynomial".

16.1.1.1. **Convex optimization problems**

Problem instances we are interested in are convex optimization programs of the following canonical form

$$P: \quad f_0(x) \to \min \mid \; f_i(x) \leq 0, \; i = 1, ..., m; \quad x \in B \equiv \{x \in \mathbf{R}^n \mid \| \; x \; \|_2 \leq 1\}.$$

From now on we assume that

- the functions f_1,...,f_m are convex and continuous on B;
- P is feasible (and consequently, due to compactness reasons, is solvable).

Note that the feasibility assumption is made mainly for simplicity reasons; we could avoid it without conceptual difficulties. Note also that we could replace the unit Euclidean ball B in our canonical form of a convex program by any other "simple" solid, like a box, or a simplex; we are speaking about the ball just for the sake of definiteness. What indeed is crucial is that the variables are from the very beginning restricted to belong to certain known solid for which we can easily solve the *separation problem* – to check, given a point, whether it belongs to the solid, and if not, to point out a hyperplane separating the point and the solid.

Let $\epsilon > 0$. We say that a point $x \in \mathbf{R}^n$ is an ϵ-solution to P, if $x \in B$, and the difference between the value of the objective at x and the optimal value of the objective f_0^*, same as the violations of the constraints at the point, do not exceed ϵ:

$$x \in G; \quad f_0(x) - f_0^* \leq \epsilon; \quad f_i(x) \leq \epsilon, \; i = 1, ..., m.$$

Perhaps it makes not that many sense to measure all function values in the same units; note, anyhow, that we can come to this case by rescaling properly the objective and the constraints.

In what follows we consider two models of representation of a problem instance on input to a solution method and, consequently, two notions of polynomial time methods for solving P.

16.1.1.2. **Oracle-represented programs and polynomiality modulo oracle**

We start with the *black box* represented instances. In this model, a particular problem P is represented by an *oracle* $\mathcal{O}$ – a subroutine which is capable, given on input a vector $x \in \text{int}\, B$, to return on output the values

$$f(x) = (f_0(x), ..., f_m(x))$$

of the objective and the constraints at x along with certain subgradients

$$f'(x) = (f_0'(x), ..., f_m'(x))$$

of the functions at the point x.

A *solution algorithm* is a code for idealized computer capable to perform operations of exact real arithmetic (for the sake of simplicity, we treat computation of elementary functions like square roots, exponents, etc., as arithmetic ones). The code gets on input the sizes n and m of the problem instance to be solved, the required accuracy $\epsilon > 0$ and is given access to the oracle. When running, the code calls the oracle at finitely many sequentially generated inputs $x_1, x_2, ..., x_N$, the inputs being formed in finitely many arithmetic operations on the basis of the previous answers of the oracle. At certain moment N the method must terminate, and the N-th point x_N generated by the method must be an ϵ-solution to the problem instance P.

An algorithm $\mathcal{A}$ of the indicated type (an *oracle-based* one) is called *polynomial modulo oracle*, if, for each $\epsilon > 0$, the total number of arithmetic operations $\mathcal{T}_{\mathcal{A}}(P, \epsilon)$ in course of solving any problem instance P to an accuracy $\epsilon > 0$ is bounded from above by a polynomial of the dimensions n, m of the instance and the minus logarithm of the *relative accuracy*

$$\nu(P, \epsilon) = \max \left[\frac{\epsilon}{\mathrm{V}(f_0)}, \frac{\epsilon}{\mathrm{V}(f_1)}, ..., \frac{\epsilon}{\mathrm{V}(f_m)} \right],$$

where

$$\mathrm{V}(f) = \max_B f - \min_B f$$

is the variation of a function f over the domain B of our program P. Thus, for a polynomial modulo oracle algorithm we should have

$$\forall (P, \epsilon > 0): \quad \mathcal{T}_{\mathcal{A}}(P, \epsilon) \leq \pi\left(m, n, \ln\left(2 + \frac{1}{\nu(P, \epsilon)}\right)\right),$$

π being a polynomial.

Note that we count in T only the computational effort of the algorithm itself, not the one spent by the oracle to process requests of the algorithm. A nontrivial (and unclear in advance) fact is that polynomial modulo oracle algorithms *do exist*.

16.1.1.3. **Data-represented programs and polynomiality over reals**

The notions of oracle-based algorithm and polynomiality modulo oracle are well-suited when we are interested in the most general convex optimization problems which do not possess any tractable and known in advance structure. Typically this is not the case: the problems arising in applications normally are well-structured, so that a problem instance can be identified with a finite set of coefficients of certain generic and known in advance analytical expressions, as it is the case in Linear Programming, Quadratic Programming, etc. The coefficients specifying the instance normally form the input to the algorithm, so that the algorithm from the very beginning possesses complete information on the instance. Normally we can use this information to build the oracle representing the problem and thus reduce the case in question to the previous one, but it is hardly the most efficient way to use the complete information on the problem instance we have in our disposal. The most natural way to formalize the situation in question seems to be as follows.

We consider a generic optimization problem $\mathcal{P} = \{P\}$ of "a given analytical structure"; by definition, this means that a problem instance $P \in \mathcal{P}$ can be specified by a finite-dimensional *data vector* $\mathrm{Data}(P)$ with real entries. A solution algorithm $\mathcal{A}$ is a code for an idealized computer capable to perform operations of exact real arithmetic. As applied to a problem instance $P \in \mathcal{P}$, the algorithm takes on input the data vector of the instance along with the desired accuracy ϵ and after finitely many operations $\mathcal{T}_{\mathcal{A}}(P, \epsilon)$ reports on output n-dimensional vector which must be an ϵ-solution to the instance. The algorithm is called *polynomial*, if

$$\forall (P \in \mathcal{P}, \epsilon > 0): \quad \mathcal{T}_{\mathcal{A}}(P, \epsilon) \leq \pi\left(\dim \mathrm{Data}(P), \ln\left(2 + \frac{V(P)}{\epsilon}\right)\right),$$

where

- $\pi(\cdot, \cdot)$ is certain polynomial;
- $V(P)$ is certain *scale factor* which may depend on the data.

This is not an actual definition, since we have not specified the scale factor. In applications a generic problem of a particular analytical structure is equipped with certain specific scale factors, most convenient in the context; after this is done, the notion of a "polynomial time" algorithm becomes indeed well-defined. Of course, uncertainty with the scale factor is, theoretically, a severe drawback of the indicated approach; at the same time, this uncertainty is not that crucial from the practical viewpoint, since the scale factor is under log and therefore its exact value is not that important, provided, of course, that it is in a reasonable range. In what follows we refer to just defined algorithms as to *algorithms polynomial over reals*.

To get a natural interpretation of the introduced notion, note that for the majority of known polynomial time algorithms the polynomial π in the complexity bound is linear with respect to the second argument, so that the bound becomes

$$\mathcal{T}_{\mathcal{A}}(P, \epsilon) \leq p(\dim \mathrm{Data}(P)) \ln\left(2 + \frac{V(P)}{\epsilon}\right),$$

p being a polynomial. The quantity $\ln(V(P)/\epsilon)$ is nothing but the number of accuracy digits in an ϵ-solution, the accuracy being measured in an appropriate (given by the scale factor) relative scale. With this interpretation of the log-factor, our complexity bound simply means that the "arithmetic cost per accuracy digit" for the algorithm in question is bounded from above by polynomial of the *dimension of the data instance*, i.e., the dimension of the corresponding data vector; in particular, twice larger computational effort results in twice larger number of accuracy digits.

We believe that the introduced notion of a polynomial time algorithm for continuous optimization problems is a good analogy to the basic notion of a fully polynomial algorithm – the usual polynomial time algorithm of Discrete Mathematics. The main differences are that now we are speaking about more powerful elementary operations – those of exact real arithmetic rather than the "bitwise operations" of the basic definition – and our goals are more restricted – we are looking for approximate solutions rather than for the precise ones, and the complexity bound is allowed to depend, in certain prescribed fashion, on the accuracy. As a compensation, the number of elementary steps in our new definition should be polynomial in the *dimension* of the data vector, while in the original definition it should be bounded by a polynomial of a much larger quantity – the bit size of the (rational) data vector. Because of these differences, the two notions of polynomial time algorithms are *not* equivalent even in simple cases when both are applicable, as it is, e.g., for systems of linear equations with rational data. The Gauss elimination algorithm is polynomial both over reals (as it is called, *strongly polynomial*) and in the bitwise sense, but these statements are in "general position" – they say different things and require different proofs. Polynomiality of Gauss elimination over reals means that the algorithm is capable to find exact solution in number of exact real arithmetic operations bounded by a polynomial of the number of equations and unknowns, while "full polynomiality" of the algorithm means that the number of *bitwise* operations sufficient to solve a system with rational data is bounded by a polynomial of the *total bit length* of the data; no one of these statements automatically implies the other one.

What we intend to overview in this mini-course are polynomial modulo oracle algorithms for general-type convex programs and polynomial over reals algorithms for convex programs of "good" analytical structure.

16.1.2 Algorithms polynomial modulo oracle

As it was already mentioned, an important, although simple, fact is that algorithms polynomial modulo oracle do exist. All known for the moment algorithms of this type belong to the so called *cutting plane scheme* which is multi-dimensional extension of the usual bisection.

16.1.2.1. **Generic cutting plane scheme**

A generic cutting plane algorithm is as follows:

Algorithm 16.1.1 [Generic Cutting Plane algorithm]
Initialization: *Choose as* G_0 *an arbitrary compact containing* B *and set* $\nu_0 = +\infty$
Step t: *Given* G_{t-1}*, choose somehow* $x_t \in \mathbf{R}^n$.

- **A** [choosing cutting plane]:

  A.1. *If* $x_t \notin \operatorname{int} B$*, call step t non-productive, set*

$$e_t = x_t,$$

so that e_t *separates* x_t *and* int B*:*

$$(\forall x \in \operatorname{int} B): \quad (x - x_t)^T e_t < 0,$$

set

$$\nu_t = \nu_{t-1}$$

and go to **B***;*

A.2. *If* $x_t \in \operatorname{int} B$*, call the oracle,* x_t *being the input, and get from the oracle the values* $f_i(x_t)$ *and certain subgradients* $f_i'(x_t)$ *of the objective and the constraints at the point* x_t*,* $i = 0, ..., m$*.*

If there exists $i, 1 \leq i \leq m$*, such that*

$$f_i(x) > \epsilon, \tag{16.170}$$

call step t non-productive, set

$$e_t = f_i'(x_t),$$

where $i \in \{1, ..., m\}$ satisfies (16.170), set

$$V_t = \max_{y \in B} (y - x_t)^T e_t = \| e_t \|_2 - x_t^T e_t; \quad \nu_t = \min \left[\nu_{t-1}, \frac{\epsilon}{V_t} \right]$$

and go to **B***;*

A.3. *If $x_t \in \operatorname{int} B$ and $f_i(x_t) \le \epsilon$, $1 \le i \le m$, call step t productive, set*

$$e_t = f_0'(x_t),$$

set

$$V_t = \max_{y \in B} (y - x_t)^T e_t = \| e_t \|_2 - x_t^T e_t; \quad \nu_t = \min \left[\nu_{t-1}, \frac{\epsilon}{V_t} \right]$$

and go to **B**.

- **B** [performing a cut] *Choose, as G_t, an arbitrary compact set containing the set*

$$G_t^+ = \{x \in G_t \mid (x - x_t)^T e_t < 0\},$$

and compute an upper bound κ_t for the quantity

$$\left(\frac{\operatorname{mes}_n G_t}{\operatorname{mes}_n B} \right)^{1/n}.$$

If

$$\kappa_t \ge \frac{\nu_t}{1 + \nu_t}, \quad \left[\frac{+\infty}{1 + \infty} = 1 \right]$$

loop, otherwise terminate, reporting as the result the best (with the smallest value of the objective) of the points x_τ, $\tau \le t$, associated with productive steps τ.

The main property of the indicated generic method is given by the following simple statement.

Proposition 16.1.1 *If a cutting plane algorithm terminates at a convex program P, then the result formed by the algorithm is an ϵ-solution to P.*

Besides this, in course of running the method one always have

$$\nu_t \ge \nu(P, \epsilon),$$

so that the method for sure terminates at a step t with

$$\kappa_t < \frac{\nu(P, \epsilon)}{1 + \nu(P, \epsilon)}.$$

Proof. The second statement of the proposition is evident, since for every t with well-defined V_t one clearly has

$$V_t \le \max_{i=0,\dots,m} \mathrm{V}(f_i).$$

Let us prove the first statement. Thus, assume that the method terminates at certain step N; we should prove that the result, let it be called z, is well-defined and that it is an ϵ-solution to P.

1^0. According to our termination rule, we have

$$\left(\frac{\operatorname{mes}_n G_N}{\operatorname{mes}_n B} \right)^{1/n} < \frac{\nu_N}{1 + \nu_N};$$

the right hand side in this inequality is ≤ 1, and therefore we can find $\lambda \in (0, 1)$ such that

$$\theta \equiv \left(\frac{\operatorname{mes}_n G_n}{\operatorname{mes}_n B} \right)^{1/n} < \lambda < \frac{\nu_N}{1 + \nu_N}.$$

Let x^* be an optimal solution to P, and let

$$B' = (1-\lambda)x^* + \lambda \,\mathrm{int}\; B.$$

Note that

$$\operatorname*{mes}_n B' = \lambda^n \operatorname*{mes}_n B > \theta^n \operatorname*{mes}_n B = \operatorname*{mes}_n G_N,$$

so that the set $B' \backslash G_N$ is nonempty. Let $u \in B' \backslash G_N$, so that

$$u = (1-\lambda)x^* + \lambda z \quad [z \in \mathrm{int}\; B]. \tag{16.171}$$

2^0. Since $u \in B \subset G_0$ and $u \notin G_N$, there exists $t \leq N$ such that $u \in G_{t-1}$ and $u \notin G_t$. According to **B**, this is possible only when

$$(u - x_t)^T e_t \geq 0. \tag{16.172}$$

3^0. Let us verify that the step t is productive. Indeed, at this step for sure $x_t \in \mathrm{int}\; B$, since otherwise, according to **A.1**, $(x - x_t)^T e_t < 0$ for all $x \in \mathrm{int}\; B$, and we know that at $x = u \in \mathrm{int}\; B$ the opposite is true. Now let us lead to a contradiction the assumption that $x_t \in \mathrm{int}\; B$ and $f_i(x_t) > \epsilon$ for certain i, $1 \leq i \leq m$. Under this assumption $e_t = f_i'(x_t)$ for (one of) the indicated i, so that (16.172) and (16.171) imply that

$$\begin{array}{rcl} f_i(x_t) & \leq & f_i(x_t) + (u - x_t)^T e_t \\ & = & (1-\lambda)(f_i(x_t) + (x^* - x_t)^T e_t) + \lambda(f_i(x_t) + (z - x_t)^T e_i) \\ & & \quad [\text{since } f_i(x_t) + (x^* - x_t)^T e_t \leq f_i(x^*) \leq 0] \\ & \leq & \lambda(f_i(x_t) + (z - x_t)^T e_t) \leq \lambda(f_i(x_t) + V_t), \end{array}$$

whence

$$f_i(x_t) \leq \frac{\lambda}{1-\lambda} V_t.$$

On the other hand, $f_i(x_t) > \epsilon$, and we come to

$$\frac{\epsilon}{V_t} \leq \frac{\lambda}{1-\lambda}.$$

this is a contradiction: according to our rules for ν_t, the left hand side in this inequality is $\geq \nu_N$, while the right hand side one, due to $\lambda < \nu_N/(1+\nu_N)$, is $< \nu_N$.

4^0. Thus, t is a productive step and, consequently, the result w returned by the algorithm is well-defined. By construction of the result, w is the best of the points x_τ associated with productive steps $\tau \leq N$; since we already know that the step t is productive, we conclude that

$$f_0(w) \leq f_0(x_t);$$

besides this, by the definition of a productive step,

$$w \in \mathrm{int}\; B; \quad f_i(z) \leq \epsilon,\ i = 1, ..., m.$$

Thus, in order to prove that w is an ϵ-solution to P it suffices to verify that $f_0(x_t) \leq f_0(x^*) + \epsilon$. This is immediate: the same computation as in 2^0 applied to f_0 results in

$$f_0(x_t) \leq (1-\lambda)(f_0(x_t) + (x^* - x_t)^T e_t) + \lambda(f_0(x_t) + V_t),$$

and since $e_t = f_0'(x_t)$ (t is productive!) and f_0 is convex, this inequality implies that

$$f_0(x_t) \leq (1-\lambda) f(x^*) + \lambda(f_0(x_t) + V_t),$$

whence

$$f_0(x_t) - f_0(x^*) \leq \frac{\lambda}{1-\lambda} V_t \leq \nu_N V_t \leq \nu_t V_t \leq \epsilon. \quad \blacksquare$$

16.1.2.2. **Polynomial implementations**

According to Proposition 16.1.1, all we need in order to get a polynomial modulo oracle cutting plane algorithm is to find policies for choosing

- (1) sequential points x_t
- (2) transformations $(G_{t-1}, x_t, e_t) \mapsto G_t \supset \{x \in G_t \mid (x - x_t)^T e_t < 0\}$

ensuring that

(A) $\kappa_t \leq \mathcal{Q}(m,n) \exp\{-\frac{t}{\mathcal{R}(n,m)}\}$, $t = 0, 1, ...$

for certain polynomials $\mathcal{Q}$ and $\mathcal{R}$

(B) *the number of arithmetic operations required to perform (1) and (2) at a single step is bounded by polynomial $\mathcal{C}(m,n)$ of m and n.*

Proposition 16.1.1 says that if a cutting plane method fits (A) and (B), then both the number of oracle calls $\mathcal{N}(P,\epsilon)$ in course of solving a convex program P and the arithmetic effort of the method $\mathcal{T}(P,\epsilon)$ admit polynomial upper bounds:

$$\begin{array}{rcl} \mathcal{N}(P,\epsilon) & \leq & 1 + \mathcal{R}(m,n) \ln \left(\mathcal{Q}(m,n) \frac{1+\nu(P,\epsilon)}{\nu(P,\epsilon)} \right), \\ \mathcal{T}(P,\epsilon) & \leq & \mathcal{C}(m,n)\mathcal{N}(P,\epsilon). \end{array}$$

To the moment we know several cutting plane methods fitting (A) and (B); let us look what are these methods.

The Center of Gravity method. The first cutting plane algorithm satisfying (A) – the *Center of Gravity method* – was proposed independently by A.Yu. Levin and J. Newman as early as in 1965. In this algorithm, $G_0 = B$, x_t is the center of gravity of G_{t-1}:

$$x_t = \frac{1}{\mathrm{mes}_n\, G_{t-1}} \int_{G_{t-1}} x dx$$

and

$$G_t = \mathrm{cl}\, G_t^+ = \mathrm{cl}\{x \in G_{t-1} \mid (x - x_t)^T e_t < 0\}.$$

For this method,

$$\mathcal{Q}(m,n) = 1; \quad \mathcal{R}(m,n) = 2.2n;$$

these are the best, up to absolute constant factors, polynomials $\mathcal{Q}$ and $\mathcal{R}$ one could hope for. A bad news about the Center of Gravity method is that this scheme does not fit (B), since we do not know how to compute in polynomial time the center of gravity of a given solid, say, of a general-type polytope; this latter problem seems to be NP-hard. Note, anyhow, that several years ago Cannam, Dyer and Freeze developed a *randomized* algorithm which, in particular, is capable to approximate in polynomial time the center of gravity of a polytope with accuracy sufficient for the needs of the Center of Gravity method, so that this algorithm admits randomized polynomial modulo oracle implementation. The order of polynomial $\mathcal{C}(\cdot,\cdot)$ arising in this implementation, however, is too high to be of any practical interest.

The Ellipsoid method. The first implementation of the cutting plane algorithm which fits both the requirements (A) and (B) was the Ellipsoid method (Nemirovski and Yudin, 1976; Shor, 1977). In this implementation, $G_0 = B$ and all G_t are ellipsoids represented as the images of the unit ball B under affine mapping:

$$G_t = c_t + A_t B,$$

A_t being a nonsingular $n \times n$ matrix, $x_t = c_{t-1}$ is the center of previous ellipsoid and G_t is the ellipsoid of the smallest volume containing the "half-ellipsoid" G_t^+. For this method,

$$\mathcal{Q}(m,n) = 1; \quad \mathcal{R}(m,n) = 2n^2,$$

so that the method fits the requirement (A), although with $\mathcal{R}$ $O(n)$ times worse than the one for the Center of Gravity method. At the same time, there are explicit formulae for the transformation

$$(A_{t-1}, x_{t-1}, e_t) \mapsto (A_t, x_t)$$

of arithmetic cost $\mathcal{C}(m,n) = O(mn + n^2)$, so that the method fits (B) as well.

The Outer Simplex method. This method is very similar to the Ellipsoid one; it was proposed in 1982 independently by Levin & Yamnistki and Bulatov & Shepot'ko. In this method, all G_t are simplices, G_0 is the smallest volume simplex containing B and x_t is the barycenter of the simplex G_{t-1}. With certain explicit policy for (2), the method has the same arithmetic cost per oracle call as the Ellipsoid method and fits (A) with

$$\mathcal{Q}(m,n) = n; \quad \mathcal{R}(m,n) = O(n^3)$$

($\mathcal{R}$ is $O(n)$ times worse than for the Ellipsoid method).

The Inscribed Ellipsoid method. This method was developed by Khachiyan, Tarasov and Erlikh in 1988. Here G_0 is the unit cube, $G_t = \mathrm{cl}\, G_t^+$ and x_t is the center of the maximum volume ellipsoid contained in the polytope G_{t-1}, or, more exactly, a tight enough approximation of the latter point. For this method, similarly to the Center of Gravity one,

$$\mathcal{Q}(m,n) = n; \quad \mathcal{R}(m,n) = O(1)n.$$

Now, to find x_t, given G_{t-1}, this is a specific convex program which can be solved to the required accuracy in polynomial time. The best known so far upper bound for arithmetic cost of solving this auxiliary problem is, up to logarithmic in n factors,

$$\mathcal{C}(m,n) = O(n^{3.5}).$$

The Method of Volumetric Centers. This is the cutting plane algorithm due to Vaidya (1990); it possesses the best known so far complexity characteristics:

$$\mathcal{Q}(m,n) = n; \quad \mathcal{R}(m,n) = O(n)$$

(basically the same as for the Center of Gravity and the Inscribed Ellipsoid methods) and

$$\mathcal{C}(m,n) = O(n^3)$$

(compare with $O(n^{3.5})$ in the Inscribed Ellipsoid method).

16.1.2.3. **Polynomial Solvability of Convex Programming**

The fact that Convex Programming admits polynomial modulo oracle solution algorithms has important theoretical consequences. The most important and an immediate one is the existence of polynomial over reals solution algorithms for all "normal" families of well-structured convex programs:

Theorem 16.1.1 *Let $\mathcal{P}$ be a family of well-structured convex programs, so that every instance $P \in \mathcal{P}$ is specified by a finite-dimensional data vector* Data(P). *Assume that problem instances from $\mathcal{P}$ are "polynomially computable", i.e., there exists an algorithm $\mathcal{O}$ which, given on input the data vector* Data(P) *and an $n(P)$-dimensional vector x, $\| x \|_2 < 1$ ($n(P)$ is the design dimension of program P), computes in polynomial in* dim Data(P) *number of arithmetic operations the values $f_i(x)$ and some subgradients $f_i'(x)$, $i = 0, ..., m$, of the objective and the constraints of the instance at the point x.*

Then $\mathcal{P}$ equipped with the scale factors

$$V(P) = \max_{0 \le i \le m(P)} [\max_{B_P} f_i - \min_{B_P} f_i]$$

($f_0, ..., f_{m(P)}$ are the objective and the constraints of the instance P, B_P is the domain of the instance) admits a polynomial over reals solution algorithm.

Proof is immediate: to get the desired algorithm, solve the input instance by, say, the Ellipsoid method, using the algorithm $\mathcal{O}$ as an oracle. Due to the already known properties of the Ellipsoid method and the assumptions on polynomial computability of the instances from $\mathcal{P}$, the overall arithmetic effort required to find an ϵ-solution to (any) problem instance $P \in \mathcal{P}$ will be bounded from above by a polynomial of $n(P)$, $m(P)$ and $\zeta \equiv \ln(2 + \nu^{-1}(P, \epsilon))$. Since, on one hand, the arithmetic effort of $\mathcal{O}$ is bounded by a polynomial in dim Data(P), and, on the other hand, this effort cannot be less than $m(P) + n(P)$ ($\mathcal{O}$ should at least read $n(P)$ input entries of x and write down $m(P) + 1$ output values of $f_i(x)$), we conclude that $m(P)$ and $n(P)$ are bounded by polynomials of dim Data(P), so that the overall arithmetic complexity of our method is in fact bounded by a polynomial of dim Data(P) and ζ. ■

In some cases polynomial over reals algorithms can further be converted into "fully polynomial" algorithms of Discrete Mathematics. The most widely known example of this type is the result on polynomial solvability (in the "completely finite" setting) of a general Linear Programming problem with rational data (Khachiyan, 1979). The main points of the construction are as follows:

- it is not difficult to demonstrate that the problem of *solving* an LP program with n variables and m inequality constraints of total bit length L can be reduced to a series of m *feasiblility problems* of the form:

  > *given a system of m linear inequalities with n unknowns of total bit length $O(1)L$, check whether the system is solvable*,

  so that it suffices to establish polynomial solvability of the feasibility problem.

- Due to particular analytical structure of the feasibility problem, it is not difficult to demonstrate that to solve this problem is the same as to decide whether the optimal value in certain associated convex program P without functional constraints is nonpositive or is greater than $2^{-O(1)L}$; to make the correct decision, it, of course, suffices to find an ϵ-solution to P with $\epsilon = 2^{-O(1)L}$. Now, it is immediately seen that

  - program P belongs to certain "well-structured" family $\mathcal{P}$ and is specified in this family by the data vector of polynomial in m and n dimension (namely, $m(n + 1)$);
  - the family $\mathcal{P}$ is comprised of polynomially computable convex programs;
  - the scale factor $V(P)$ from Theorem 16.1.1 is at most $2^{O(1)L}$.

- According to the above remarks and Theorem 16.1.1, we can solve P to the desired accuracy in polynomial in n, m, L (and, consequently, in $L > \max[m, n]$) number of operations of *exact real arithmetic*, and all we need to get a "fully polynomial" solution algorithm for feasibility problem is to pass from exact real arithmetic operations to "bitwise" ones. A straightforward, although involving, analysis demonstrates that replacing in the algorithm exact operations by inexact ones keeping in the result $O(1)mL$ digits before and after the decimal dot, we still get an approximate solution of the desired quality, and with this substitution of operations we do get a fully polynomial solution algorithm for the feasibility problem.

To the moment there are several fully polynomial algorithms for LP, and the structure of all these algorithms is similar to the just outlined: we start with a polynomial over reals algorithm for Convex or Linear Programming programs and subject it to (now quite standard) modifications to make the method "completely finite" on rational instances.

The existence of polynomial modulo oracle cutting plane algorithms has further theoretical consequences. The underlying feature of the scheme is that to run a cutting plane algorithm, we in fact do not need explicit access to all functional constraints. What is sufficient for us, is a subroutine $\mathcal{S}$ which checks, given on input a point x, whether all the constraints are satisfied at x within accuracy ϵ; if it is the case, the routine returns the value and a subgradient of the objective at x, otherwise – the value and a subgradient of (any) "essentially violated" constraint. Whenever we have a "well-structured" generic convex problem $\mathcal{P}$ which can be equipped by subroutine $\mathcal{S}$ of the indicated type with polynomial in dim Data(P) running time, we still are able to point out a polynomial over reals solution algorithm for $\mathcal{P}$. As it was shown by the authors of this approach Grötshel, Lovasz and Shrijver (1984; see [5]), this scheme allows to get in a unified manner fully polynomial algorithms for all combinatorial problems known to be polynomially solvable (in several cases, the algorithms developed on the basis of the scheme were the first polynomial algorithms for the corresponding problems).

In spite of their theoretical universality, or, better to say, because of this universality, cutting plane polynomial time algorithms are not that attractive for practical computations: their arithmetic complexity grows quickly with the design dimension of the problem. In a sense, this phenomenon is unavoidable for general-purpose convex optimization algorithms using only local information on problem instances: one can point out families $\mathcal{P}_n$, $n = 1, 2, ...$ of simple convex programs

$$P: \quad f(x) \to \min \mid \parallel x \parallel_2 \leq 1 \quad [x \in \mathbf{R}^n]$$

with $V(P) \leq 1$ for all instances $P \in \mathcal{P}_n$ such that the worst-case, over $P \in \mathcal{P}_n$, number of calls to a local oracle for *any* method solving the instances within accuracy n^{-1} is at least $O(1)n$. It follows that the efficiency estimate (in terms of number of oracle calls) of any polynomial modulo oracle algorithm is at least proportional to the design dimension of the problem; the number of arithmetic operations (modulo oracle) "per accuracy digit" for the most efficient known algorithms of this type grows at least as the fourth degree of the design dimension. Therefore the outlined methods are well-suited only for problems with moderate (not more then several tens) number of variables and are basically useless in the large scale case. As some people say, the role of these algorithms is to provide us with the proof of polynomial solvability of Convex Programming; to get an actually efficient method for a large scale problem, we need specific algorithms which are adjusted to the particular analytical structure of the problem in question, and large scale convex problems arising in applications normally do possess such a structure. The most promising approach to utilizing structure of a convex program seems to be the one based on recent progress in the theory of polynomial time interior point methods; this approach is the subject of the forthcoming sections.

16.2 Interior point polynomial methods: introduction

The most powerful general approach to the design of polynomial time convex methods capable to utilize information on the structure of the convex programs to be solved is offered by the recent advances in the theory of *polynomial time interior point methods*. The first method of this type was developed in the seminal paper of N. Karmarkar (1984) for Linear Programming; this paper initiated outstanding activity, during the last decade mainly focused on LP and now being extended on well-structured nonlinear convex problems as well. It is impossible to mention even the most important contributions of many researchers to the area (the incomplete bibliography of the subject due to Dr. Kranich contains over 1,500 entries), but there is one contribution which cannot be avoided – the papers of J. Renegar (1986) and C. Gonzaga (1988), where the first *path-following* interior point methods for LP were proposed. These papers not only improved the complexity results for LP obtained by Karmarkar, but also linked the new-born area with a quite traditional optimization scheme; this contributed a lot to better understanding of the nature of new methods and brought the area into a good position for further extensions.

As it was already mentioned, the activity in the area of polynomial time interior point methods was initially focused on Linear Programming. The extensions onto more general nonlinear convex optimization problems at this stage were restricted mainly to linearly constrained convex quadratic programming, where one could use without significant difficulties basically the same constructions and proofs as in Linear Programming. The "most nonlinear" convex problems which were successfully studied at this stage were convex quadratically constrained quadratic programs (Jarre, 1987; Mizuno & Sun, 1988); here one also can use the tools already known from LP. Further extensions required better understanding of the intrinsic nature of interior point methods in LP. In 1988, two close, although not completely identical, ideas on this intrinsic nature and, consequently, on how to extend the interior point methods from LP onto more general convex problems, were proposed. One of them is the "relative Lipschitz condition" of F. Jarre, another – the self-concordance-based approach originating from Yu. Nesterov. In the mean time it turned out that the second approach is more convenient and general, and now it became a kind of standard; it underlies a general theory of interior point polynomial time methods in Convex Programming we possess now, and this theory allows to explain all polynomial time constructions and results known in Linear Programming and to extend these constructions and results on the nonlinear case.

16.2.1 Self-concordance-based approach

To get a very brief overview of the general theory as it is developed in [3] , let us start with a problem

$$P: \qquad c^T x \to \min \mid x \in G \subset \mathbf{R}^n, \tag{16.173}$$

of minimizing a linear objective over a convex solid (closed and bounded convex set with a nonempty interior) in the particular case when the solid is a polytope:

$$G = \{x \in \mathbf{R}^n \mid a_i^T x \leq b_i,\ i = 1, ..., m\} \tag{16.174}$$

(we assume that all a_i are nonzero). In the LP interior point methods for the Linear Programming program P, (16.174) the key role is played by the *logarithmic barrier*

$$F(x) = -\sum_{i=1}^{m} \ln(b_i - a_i^T x)$$

for the feasible polytope G of the problem. The crucial observation of Nesterov today can be expressed as follows: among all numerous features of this logarithmic barrier, in fact only the following three are responsible for the polynomiality of the associated interior point methods for LP:

- **A** [barrier property]: *F is a barrier for G: $F(x_i) \to \infty$ along every sequence $\{x_i\}$ of interior points of G converging to a boundary point of G;*

- **B** [self-concordance]: *F is self-concordant, i.e., is a* C^3 *smooth convex function on* int G *satisfying the differential inequality*

$$\forall (x \in \text{int } G, h \in \mathbf{R}^n): \quad |D^3 F(x)[h,h,h]| \leq 2 \left(D^2 F(x)[h,h]\right)^{3/2}$$

[from now on $D^k F(x)[h_1, ..., h_k]$ denotes k-th differential of a function F taken at a point x along the directions $h_1, ..., h_k$]; *in other words, the second-order derivative of F is Lipschitz continuous with respect to the local Euclidean metric defined by this derivative itself;*

- **C** [boundedness of the Newton decrement]: *F satisfies the differential inequality*

$$\forall (x \in \operatorname{int} G, h \in \mathbf{R}^n) : \quad |DF(x)[h]| \leq \vartheta^{1/2}(F) \left(D^2 F(x)[h,h]\right)^{1/2};$$

in other words, F itself is Lipschitz continuous with respect to the local Euclidean metric given by the second order derivative of F.

Now, properties **A** – **C** do not address explicitly polyhedral structure of G; given *an arbitrary* closed convex domain with a nonempty interior, we may try to equip it with a barrier F satisfying these properties; such a barrier will be called *self-concordant*, the quantity $\vartheta(F)$ being called the *parameter of the barrier*. And it turns out that

Theorem 16.2.1 *Given a self-concordant barrier F for an arbitrary solid $G \subset \mathbf{R}^n$ and a starting point $\widehat{x} \in \operatorname{int} G$, one can associate with these entities an interior point method $\mathcal{IP}_F$ for solving problem P of minimizing a linear objective on G. This method solves P to a prescribed accuracy $\epsilon > 0$ – i.e., generates a point $x_\epsilon \in \operatorname{int} G$ such that*

$$c^T x_\epsilon - \min_{x \in G} c^T x \leq \epsilon$$

at the cost of assembling and solving no more than

$$\mathcal{N} \equiv \mathcal{N}_{F,\widehat{x}}(\epsilon) = O(1)\sqrt{\vartheta(F)} \ln \left(2\vartheta(F)[G:\widehat{x}] \frac{1+\nu(P,\epsilon)}{\nu(P,\epsilon)} \right)$$

Newton systems of the form

$$S_i : \qquad \nabla^2 F(x^i) u = \nabla F(x^i) + t^i c.$$

Here

- *$O(1)$ is absolute constant;*
- *$[G:\widehat{x}]$ is the asymmetry coefficient of G with respect to the starting point:*

$$[G:\widehat{x}] = \sup\{t \mid \exists y \notin G : x - t(y-x) \in G\};$$

- *$\nu(P,\epsilon) = \epsilon / (\max_{x \in G} c^T x - \min_{x \in G} c^T x)$ is the relative accuracy of ϵ-solution (cf. Section 1);*
- *x^i, t^i, $i = 1, ..., \mathcal{N}$, are generated by the method interior points of G and scalar parameters, respectively.*

To update (x^i, t^i) into (x^{i+1}, t^{i+1}) is, computationally, exactly the same as

- *to assemble the system S_i and to solve it with respect to the unknown u (the system for sure is solvable);*
- *to perform $O(n)$ arithmetic operations more to update x^i, t^i, u into x^{i+1}, t^{i+1}.*

Note that in the Linear Programming case, when G is given by (16.174) and F is the logarithmic barrier for the polytope, F turns out to be self-concordant with parameter $\vartheta(F)$ equal to m – the number of linear inequalities defining G. In this case the above statement yields the standard Renegar-type efficiency estimate for the LP interior point methods.

How to associate with a self-concordant barrier for the feasible domain of a problem (16.173) interior point methods fitting efficiency estimate of the above (or similar) type – this is one piece of the theory in question; here it turns out that basically all constructions and results related to interior point methods in Linear Programming can be naturally extended onto the general nonpolyhedral case.

Now, problem (16.173) of linear minimization over a solid is a universal, in the natural sense, Convex Programming program. Indeed, a convex program in the canonical form we dealt with in the previous Section:

$$P : \quad f_0(u) \to \min \mid \; f_i(u) \leq 0, \; i = 1, ..., m; \quad u \in B = \{u \in \mathbf{R}^k \mid \parallel u \parallel_2 \leq 1\}$$

(f_i, $i = 0, ..., m$, are convex and continuous on B) can be in many ways converted to (16.173). E.g., if we know in advance that P satisfies the Slater condition and are able to form an upper bound $\widehat{f_0}$ on absolute values of f_0 in B, we can pass from the initial design vector u to the extended vector $x = (t, u) \in \mathbf{R}^{k+1}$ and to rewrite P equivalently in the form of (16.173) as

$$\begin{array}{rl} c^T x \equiv t \to \min \mid x \in G_P = \{(t,u) \mid & u \in B,\ f_0(u) \le t \le 2\widehat{f_0}, \\ & f_i(u) \le 0,\ i = 1, ..., m\}. \end{array}$$

After this reformulation, we could try to equip G_P with a self-concordant barrier F and to solve the problem with the associated interior point method. The complexity characteristics of this method mainly depend on the magnitude of $\vartheta(F)$ and on the "computational complexity" of the barrier – the arithmetic cost at which one can compute the gradient and the Hessian of F at a given point; there are also "secondary" effects caused by the necessity to start the method in an interior point of G_P not too close to the boundary of G_P. When P comes from a well-structured family $\mathcal{P}$ of convex programs and both these problems – the one of finding an "efficiently computable" self-concordant barrier with reasonable value of the parameter for G_P and the one of finding a "reasonably centered" starting point $\widehat{x} \in \operatorname{int} G_P$ – can be properly solved, we end up with a polynomial over reals algorithm for the family. Now, the second of the indicated issues – initialization policy – can be easily resolved in a universal and a quite satisfactory manner, so that in fact our possibilities to exploit the outlined approach as a source of polynomial over reals algorithms for generic well-structured convex programs are limited only by our abilities to point out "computable" self-concordant barriers for the corresponding feasible domains. In principle, *every* n-dimensional convex domain can be equipped with a self-concordant barrier with parameter $O(1)n$ (from now on, all $O(1)$'s are absolute constants); this *universal* barrier is given by the formula

$$F_G(x) = O(1) \ln \operatorname*{mes}_n \operatorname{Polar}(G, x),$$

where

$$\operatorname{Polar}(G, x) = \{y \in \mathbf{R}^n \mid y^T(x' - x) \le 1 \ \forall x' \in G\}$$

is the polar of G with respect to x. Thus, *in principle* the outlined approach can be applied to *any* convex optimization problem. Unfortunately, the universal barrier usually is "computationally intractable" – it cannot be computed at a reasonable arithmetic cost. There exists, however, a kind of calculus of "efficiently computable" barriers, which yields barriers of this type for a wide family of convex domains arising in applications; this "barrier calculus" is the second large piece of the theory. And the third piece is formed by applications, where we put things together, namely, take a particular well-structured convex problem $\mathcal{P}$ and work out a policy for converting problem instances into the standard form required by the general interior point schemes and for equipping the corresponding feasible domains with computable self-concordant barriers, thus coming to the associated with our generic problem interior point methods.

This was a very brief overview of the theory; its detailed representation is far beyond the scope of this mini-course. What we are about to do is to present just two general interior point schemes, skipping all technical considerations, and to outline several generic applications of the second, more efficient of the schemes.

16.2.2 Preliminaries: Newton method and self-concordance

16.2.2.1. Self-concordant functions

As it was already explained, the main hero of the story to be told is self-concordant barrier for a closed convex domain $G \subset \mathbf{R}^n$. As we remember, this is a function satisfying simple barrier property and two differential inequalities linking the first, the second and the third directional derivatives of the function. The role of the second of these inequalities is relatively restricted, although finally crucial, and it makes sense not to impose it as long as possible. Thus, we start with *self-concordant functions* – those possessing only properties **A** and **B**:

Definition 16.2.1 [Self-concordant function] *Let $G \subset \mathbf{R}^n$ be a closed convex domain with a nonempty interior. We say that a function F is self-concordant on G, if*

- *F is three times continuously differentiable convex function on the interior of G with the barrier property:*

$$F(x_i) \to \infty \text{ whenever } x_i \in \operatorname{int} G \text{ converge to } x \in \partial G;$$

- *F satisfies the differential inequality*

$$\forall (x \in \operatorname{int} G, h \in \mathbf{R}^n): \quad |D^3F(x)[h,h,h]| \le 2\left(D^2F(x)[h,h]\right)^{3/2} \tag{16.175}$$

There is nothing too specific about constant 2 in (16.175): both sides of inequality (16.175) are of the same homogeneity degree with respect to h, as it should be for an affine invariant relation, but are of different homogeneity degrees with respect to F. It follows that if F satisfies (16.175) with constant factor 2 replaced by any other constant factor, we could rescale F – multiply it by a positive constant – to enforce the rescaled function to satisfy similar inequality with a desired constant factor. The particular normalization we choose is the most convenient: with this normalization of inequality (16.175) the function $-\ln t$ turns out to be self-concordant on the nonnegative axis, the logarithmic barrier for a polytope turns out to be self-concordant on the polytope, etc.

For the sake of simplicity, in what follows we deal only with *nondegenerate* self-concordant functions F – those with nonsingular at any point of the domain Hessians F''.

Proposition 16.2.1 *The necessary and sufficient condition of nondegeneracy of a self-concordant function is to have a nonsingular Hessian at at least one point.*

A sufficient condition for nondegeneracy is that the domain of the function does not contain straight lines (which for sure is the case when the domain of the function is bounded).

Note that if F is a nondegenerate self-concordant function on $G \subset \mathbf{R}^n$ and $x \in \operatorname{int} G$, then the second order differential of F at x, being a positive definite quadratic form, defines a Euclidean norm on $\mathbf{R}^n$:

$$\| h \|_{F''(x)} = \left(D^2F(x)[h,h]\right)^{1/2} \quad [= (h^TF''(x)h)^{1/2}],$$

same as the *conjugate norm*:

$$\| \eta \|_{[F''(x)]^{-1}} = \max\{\eta^T h \mid \| h \|_{F''(x)} \le 1\} \quad [= (\eta^T[F''(x)]^{-1}\eta)^{1/2}].$$

Basic properties of self-concordant functions. The central role in all interior point schemes is played by the Newton minimization method, and the role of self-concordance in our coming considerations is based upon the very nice behaviour of this method as applied to a self-concordant function. This nice behaviour, in turn, comes from the fact that a self-concordant function is fairly well approximated locally by its second order Taylor expansion. The corresponding result is as follows:

Proposition 16.2.2 *Let F be self-concordant on G, and let $x \in \operatorname{int} G$. Then the unit Dikin ellipsoid of F at x – the set*

$$D_{F,x} = \{y \mid \| y - x \|_{F''(x)} \le 1\}$$

is contained in G. For all $y \in \mathbf{R}^n$ we have

$$\begin{array}{rcl} \Omega(-\| y - x \|_{F''(x)}) & \le & F(y) \\ & & -\left[F(x) + (y-x)^TF'(x) + \frac{1}{2}(y-x)^TF''(x)(y-x)\right] \le \\ & \le & \Omega(\| y - x \|_{F''(x)}), \end{array} \tag{16.176}$$

where

$$\Omega(s) = -\ln(1-s) - s - \frac{s^2}{2} \quad \left[= \frac{s^3}{3} + \frac{s^4}{4} + ..., |s| < 1\right]$$

(F is extended outside int *G by the value $+\infty$, and by definition $\Omega(s) = +\infty$ for $s \ge 1$, so that (16.176) makes sense for all y).*

To express the convergence properties of the Newton method as applied to a self-concordant function, we need to measure somehow the "distance" from the current iterate to the minimizer of the function; since we are interested in affine invariant description (note that both the notion of self-concordance and the Newton method are affine invariant entities), this distance should be defined in terms of F itself. There are two natural "distances" of this type:

- Residual in terms of F:
$$F(x) - \inf_{\mathrm{int}\ G} F;$$
- Newton decrement
$$\lambda(F,x) = \| F'(x) \|_{[F''(x)]^{-1}} \qquad \left[= \sqrt{(F'(x))^T [F''(x)]^{-1} F'(x)} \right].$$

Both these quantities, as it should be for distances, are zero if x is the minimizer of F and are positive otherwise. The advantage of the Newton decrement is that this quantity is "observable" – given x, we can explicitly compute $\lambda(F,x)$. These two distances are closely related to each other:

Proposition 16.2.3 *Let F be self-concordant and nondegenerate on $G \subset \mathbf{R}^n$. Then the following relations are equivalent to each other:*

- *F attains its minimum on* int G*, the minimizer being unique;*
- *F is below bounded on* int G*;*
- *there exists $x \in \mathrm{int}\ G$ such that $\lambda(F,x) < 1$.*

Moreover, for any $x \in \mathrm{int}\ G$ one has

$$\lambda(F,x) - \ln(1+\lambda(F,x)) \le F(x) - \min_{\mathrm{int}\ G} F \le -\lambda(F,x) - \ln(1-\lambda(F,x)),$$

the right hand side in the second inequality being $+\infty$ when $\lambda(F,x) \ge 1$.

The proposition, in particular, says that the "distances" $d_1(x) = F(x) - \min_{\mathrm{int}\ G} F$ and $d_2(x) = \frac{1}{2}\lambda^2(F,x)$ from a point $x \in \mathrm{int}\ G$ to the minimizer of F are equivalent to each other when they are small; e.g.,

$$\begin{array}{rcl} d_1(x) < 1/3 & \Rightarrow & d_2(x) \le \frac{d_1(x)}{1-1.2\sqrt{d_1(x)}}, \\ d_2(x) < 1/2 & \Rightarrow & d_1(x) \le \frac{d_2(x)}{1-0.95\sqrt{d_2(x)}}. \end{array} \qquad (16.177)$$

It is time now to look at the behaviour of the Newton minimization method as applied to a self-concordant function. We shall deal with the following version of the method:

$$x_{t+1} = x_t - \frac{1}{1+\lambda(F,x_t)} [F''(x_t)]^{-1} F'(x_t); \qquad (16.178)$$

this is the usual Newton method with certain specific rule for stepsizes. The main result on this recurrence is as follows:

Proposition 16.2.4 *Let F be a nondegenerate self-concordant function on $G \subset \mathbf{R}^n$, and let the recurrence (16.178) be started at a point $x_0 \in$* int G*. Then the iterates x_t are well-defined, belong to the interior of G and for all t one has*

$$\lambda(F,x_{t+1}) \le 2\lambda^2(F,x_t); \quad F(x_t) - F(x_{t+1}) \ge \lambda(F,x_t) - \ln(1+\lambda(F,x_t)).$$

Proposition says that if we are minimizing a (nondegenerate) self-concordant function F by the damped Newton method (16.178), then the method converges, and the "convergence pattern" admits problem-independent description as follows:

- I. Generally speaking, there is an initial stage – until the first moment t when it happens that $\lambda(F, x_t) \le 1/4$. At this stage, each Newton step decreases the function at least by the absolute constant $\frac{1}{4} - \ln\frac{5}{4} = 0.0269....$

- II. Starting with the moment when $\lambda(F, x_t) \le 1/4$, we are in the region of quadratic convergence of the method: each step basically squares the Newton decrement:

  $$\lambda(F, x_t) \le 1/4 \Rightarrow \lambda(F, x_{t+1}) \le 2\lambda^2(F, x_t) \quad [\le \lambda(F, x_t)/2].$$

  According to (16.177), at this phase also

  $$F(x_{t+1}) - \min_{\mathrm{int}\, G} F \le 8[F(x_t) - \min_{\mathrm{int}\, G} F]^2 \quad \left[\le 0.4[F(x_t) - \min_{\mathrm{int}\, G} F]\right].$$

An immediate corollary of these results is as follows:

Corollary 16.2.1 *Let a nondegenerate below bounded self-concordant on G function F be minimized by the damped Newton method (16.178) started from a point $x_0 \in \mathrm{int}\, G$, and let $\kappa \in (0, \frac{1}{8})$, Then the number t of steps of the method until we for the first time get the inequality*

$$\lambda(F, x_t) \le \kappa$$

does not exceed the quantity

$$40(F(x_0) - \min_{\mathrm{int}\, G} F) + 2\ln\ln\frac{1}{\kappa}.$$

Indeed, the number of steps at the initial stage, due to I, does not exceed

$$(F(x_0) - \min F)/0.0269 \le 40(F(x_0) - \min F),$$

and from II it takes no more than $2\ln\ln(1/\kappa)$ steps of the final stage to reach the target value of the Newton decrement.

16.2.2.2. **Self-concordant barriers**

Now let us look at *self-concordant barriers*. As we know from Section 16.2.1, a self-concordant barrier F for a closed convex domain $G \subset \mathbf{R}^n$ is a self-concordant on G function which, in addition, satisfies the inequality

$$\forall (x \in \mathrm{int}\, G, h \in \mathbf{R}^n): \quad |DF(x)[h]| \le \vartheta^{1/2}(F) \left(D^2F(x)[h,h]\right)^{1/2}; \tag{16.179}$$

here $\vartheta(F) < \infty$ is the *parameter* of the barrier. It can be proved that the only barrier with the parameter less than 1 is the constant barrier for the entire $\mathbf{R}^n$; this "barrier" will be of no interest for us, so that from now on we assume that the parameter of any barrier in question is ≥ 1.

Restricting ourselves for the sake of simplicity to *nondegenerate* self-concordant barriers, we can rewrite (16.179) equivalently as

$$(\forall x \in \mathrm{int}\, G): \quad \lambda(F, x) \le \sqrt{\vartheta(F)}.$$

Thus, a self-concordant barrier for G is a self-concordant on G function with bounded Newton decrement.

Combination of self-concordance and inequality (16.179) implies a lot of nontrivial relations which are satisfied by a self-concordant barrier. Among these relations let us stress the following two:

Proposition 16.2.5 *Let F be a nondegenerate self-concordant barrier for a closed convex domain $G \subset \mathbf{R}^n$. Then*

- (i) [semiboundedness] *One has*

  $$\forall (y \in G, x \in \mathrm{int}\, G): \quad (y - x)^T F'(x) \le \vartheta(F);$$

- (ii) [centering property] *F is below bounded on G if and only if G is bounded; this is the case if and only if $\lambda(F,x) < 1$ for certain $x \in \mathrm{int}\, G$. If G is bounded, then the Dikin ellipsoid of F with center at the minimizer $x(F)$ of F approximates G within factor $\vartheta(F)$:*

$$\{y \mid \| y - x(F) \|_{F''(x(F))} \leq 1\} \subset G \subset \{y \mid \| y - x(F) \|_{F''(x(F))} \leq 3\vartheta(F) + 1\}$$

(it was shown by F. Jarre that $3\vartheta(F) + 1$ in (ii) can be replaced with $\vartheta(F) + 2\sqrt{\vartheta(F)}$).

It is time now to present several basic examples of self-concordant barriers:

Example 16.2.1 *The function $-\ln t$ is 1-self-concordant barrier for the nonnegative axis.*

This is a particular case of the following

Example 16.2.2 *The function*

$$F(x) = -\ln \mathrm{Det}\, x$$

is m-self-concordant barrier for the cone $\mathbf{S}^m_+$ of positive definite symmetric $m \times m$ matrices.

Example 16.2.3 *Let f be a convex quadratic form on $\mathbf{R}^n$. Then the function*

$$F(t,x) = -\ln(t - f(x))$$

is 1-self-concordant barrier for the epigraph $\{(t,x) \mid t \geq f(x)\}$ of the form.

The function

$$F(t,x) = -\ln(t - \frac{x^T x}{t}) - \ln t = -\ln(t^2 - x^T x) \quad [x \in \mathbf{R}^n]$$

is 2-self-concordant barrier for the epigraph of the Euclidean norm – for the second order cone

$$\mathbf{K}^n = \{(t,x) \in \mathbf{R} \times \mathbf{R}^n \mid t \geq \| x \|_2\}.$$

The second part of the latter example is a particular case of the following

Example 16.2.4 *The function*

$$F(x) = -\ln \mathrm{Det}\,(tI_m - \frac{x^T x}{t}) - \ln t \quad [x \text{ is } n \times m \text{ matrix, } I_m \text{ is the unit } m \times m \text{ matrix}]$$

is $(m+1)$-self-concordant barrier for the epigraph of the spectral norm of an $n \times m$ matrix – for the cone

$$\{(t,x) \in \mathbf{R} \times \mathbf{R}^{n \times m} \mid \| xs \|_2 \leq t \| s \|_2 \ \ \forall s \in \mathbf{R}^m\}.$$

In the mean time we shall extend the list of basic examples of self-concordant barriers. What should be stressed is that the "barrier calculus" we possess now allows to justify self-concordance of these and forthcoming barriers without any computations (in fact allows even to *derive* them rather than to guess) on the basis of the only evident fact – that the function $-\ln t$ indeed is a 1-self-concordant barrier for the nonnegative axis.

Given a set of basic examples of self-concordant barriers, we can construct more examples applying the following *combination rules*

Proposition 16.2.6 [Elementary Barrier Calculus]

(i) [summation] *Let F_i be self-concordant on $G_i \subset \mathbf{R}^n$, $i = 1, ..., m$, and let $G = \cap_{i=1}^m G_i$ be a set with a nonempty interior. Then, for all $\alpha_i \geq 1$, the function*

$$F(x) = \sum_{i=1}^m \alpha_i F_i(x)$$

is self-concordant on G; if all F_i are ϑ_i-self-concordant barriers for G_i, then F is $(\sum_{i=1}^m \alpha_i \vartheta_i)$-self-concordant barrier for G.

(ii) [direct summation] *Let F_i be self-concordant on $G_i \subset \mathbf{R}^{n_i}$, $i = 1, ..., m$, and let $\alpha_i \geq 1$. Then the function*

$$F(x_1, ..., x_m) = \sum_{i=1}^{m} \alpha_i F_i(x_i)$$

is self-concordant on $G = G_1 \times ... \times G_m \subset \mathbf{R}^{n_1+\cdots+n_m}$. If all F_i are ϑ_i-self-concordant barriers for G_i, then F is $(\sum_{i=1}^{m} \alpha_i \vartheta_i)$-self-concordant barrier for G.

(iii) [superposition with affine mappings] *Let F^+ be self-concordant on $G^+ \subset \mathbf{R}^n$, and let $x \mapsto A(x)$ be an affine mapping from $\mathbf{R}^k$ to $\mathbf{R}^n$ with the image intersecting* int G^+. *Then the function $F(\cdot) = F^+(A(\cdot))$ is self-concordant on the inverse image $G = A^{-1}(G^+)$ of G^+. If F^+ is ϑ-self-concordant barrier for G^+, then so is F for G.*

As an immediate application example of these combination rules (which are evident consequences of the definitions of a self-concordant function/barrier), let us prove that the standard logarithmic barrier

$$F(x) = -\sum_{i=1}^{m} \ln(b_i - a_i^T x)$$

for polytope $G = \{x \in \mathbf{R}^n \mid a_i^T x \leq b_i, \, i = 1, ..., m\}$ (the system of linear equalities defining G satisfies the Slater condition) is an m-self-concordant barrier for the polytope. Indeed,

$-\ln t$ is 1-s.-c.b. for $\mathbf{R}_+ \Rightarrow$ [Prop. 16.2.6.(iii)]
$-\ln(b_i - a_i^T x)$ is 1-s.c.b. for $G_i = \{x \mid a_i^T x \leq b_i\} \Rightarrow$ [Prop. 16.2.6.(i)]
$F(x) = -\sum_{i=1}^{m} \ln(b_i - a_i^T x)$ is m-s.-c.b. for $G = \cap_{i=1}^{m} G_i$. ■

16.2.3 First fruits: the method of Karmarkar

Now we have developed enough machinery to present interior point methods. We start with nonpolyhedral extension of the very first LP interior point method – the method of Karmarkar [7].

Consider the problem

$$c^T x \to \min \mid x \in G \subset \mathbf{R}^n,$$

where G is a solid (closed and bounded convex set with a nonempty interior). Same as in the original method of Karmarkar, let us make the following assumptions:

K.1. We know in advance an interior point $\widehat{x}$ of G
[without loss of generality, we may assume that $\widehat{x} = 0$];

K.2. We know in advance the optimal value c^* of the problem
[note that $c^* \leq c^T \widehat{x} = 0$, the case of equality here being of no interest, so that we always can assume that $c^* < 0$; by further normalization, we can assume that $c^* = -1$].

The geometry of the method is fairly simple. We should approximate from inside G the set where G touches the hyperplane $\Pi = \{x \mid c^T x + 1 = 0\}$. Let us perform the projective transformation

$$x \mapsto \frac{x}{c^T x + 1};$$

this transformation pushes Π to infinity and makes G an *unbounded* closed convex domain G^+. To approach Π from inside G is the same as to approach infinity from inside G^+; thus, all we need is to point out a procedure which allows to move quickly towards ∞, staying inside G^+. To this end let us equip G^+ with a ϑ-self-concordant barrier F, and let us apply to this barrier the damped Newton minimization (16.178), starting it at the point $y_0 = 0$. It is immediately seen that G^+ does not contain lines (since G is bounded), and therefore F is nondegenerate (Proposition 16.2.1). Since G^+ is unbounded, Proposition 16.2.5.(ii) says that $\lambda(F, y) \geq 1$ for all $y \in$ int G, so that by Proposition 16.2.4 each step of the method decreases F at least by $\kappa = 1 - \ln 2$. It follows that F diverges to $-\infty$ along the trajectory $\{y_t\}$ of the method. But F is convex and therefore below bounded on every bounded subset of G^+, so that $F(y_t)$ can diverge to $-\infty$ only when $y_t \in G^+$ diverge to infinity, and this is exactly what we need. This qualitative reasoning can easily be quantified to get the rate of convergence.

Indeed, consider the ray $R_t = [y_t, 0)$. We claim that this ray intersects the boundary of G^+ at certain point z_t. Indeed, from Proposition 16.2.5.(i) it follows that a self-concordant barrier never increases along any ray contained in the domain of the barrier, so that assumption $R_t \subset G$ would imply $F(y_t) \geq F(0)$; this is not the case, since, as it was already explained,

$$F(y_t) \leq F(0) - t\kappa \quad [\kappa = 1 - \ln 2]. \tag{16.180}$$

Introducing coordinate r along the ray R_t in such a way that $r = 0$ corresponds to the point 0 of the ray, and $r = 1$ corresponds to the point z_t, denoting by $-T$ the coordinate of y_t (clearly $T > 0$) and setting $f(r) = F(rz_t)$, we get

$$-T \leq r \leq 0 \Rightarrow f'(r) = z_t^T F'(rz_t) = \frac{1}{1-r}(z_t - rz_t)^T F'(rz_t) \leq \frac{1}{1-r}\vartheta,$$

the concluding inequality being given by Proposition 16.2.5.(i). Integrating the resulting inequality, we get

$$F(0) - F(y_t) = f(0) - f(-T) = \int_{-T}^{0} f'(r)dr \leq \vartheta \int_{-T}^{0} \frac{1}{1-r}dr = \vartheta \ln(1+T),$$

which combined with (16.180) implies the lower bound

$$T + 1 \geq \left(\frac{e}{2}\right)^{t/\vartheta}. \tag{16.181}$$

Now let

$$x_t = \frac{y_t}{1 - c^T y_t}; \quad w_t = \frac{z_t}{1 - c^T z_t}$$

be the inverse images of y_t and z_t in G. Since 0 is an interior point of the segment $[y_t, z_t]$, it is also an interior point of the segment $[x_t, w_t]$, and since z_t is a boundary point of G^+, w_t is a boundary point of G. Consequently,

$$\parallel x_t \parallel_2 : \parallel w_t \parallel_2 \leq [G:0]$$

(the right hand side is the asymmetry coefficient of G with respect to $\widehat{x} = 0$, see Theorem 16.2.1), whence, since x_t and w_t are collinear,

$$|c^T x_t| \leq [G:0]|c^T w_t|. \tag{16.182}$$

We have $c^T y_t = -Tc^T z_t$, or, substituting $y_t = \dfrac{x_t}{1 + c^T x_t}$, $z_t = \dfrac{w_t}{1 + c^T w_t}$,

$$c^T w_t = -\frac{c^T x_t}{T + (T+1)c^T x_t};$$

substituting this relation into (16.182), we come to

$$|c^T x_t| \leq [G:0]\frac{|c^T x_t|}{|T + (T+1)c^T x_t|};$$

assuming for the time being that $c^T x_t \neq 0$, we conclude from the resulting inequality that $T + (T+1)c^T x_t \leq [G:0]$, whence

$$c^T x_t + 1 \leq \frac{1 + [G:0]}{T+1};$$

the same inequality in fact is valid also under the assumption that $c^T x_t = 0$ (to see it, it suffices to replace in the above reasoning y_t by a close to y_t point y_t' with $c^T y_t' \neq 0$ and then to pass to limit as $y_t' \to y_t$). The left hand side in the resulting inequality is nothing but the residual in terms of the objective at x_t; taking into account (16.181), we finally come to the following polynomial time convergence estimate:

$$\frac{c^T x_t - \min_{x \in G} c^T x}{c^T \widehat{x} - \min_{x \in G} c^T x} \leq (1 + [G:\widehat{x}])\left(\frac{e}{2}\right)^{-t/\vartheta}, \ t = 1, 2, ...$$

(the estimate in this form is valid independently of our normalization assumptions $\widehat{x} = 0$, $c^* = -1$).

In the Linear Programming case, when G is a polytope

$$G = \{x \in \mathbf{R}^n \mid a_i^T x \leq b_i,\ i = 1, ..., m\}$$

and the situation is normalized by the relation $\widehat{x} = 0$, we have

$$G^+ = \{y \in \mathbf{R}^n \mid a_i^T y \leq b_i(1 + (c^*)^{-1} c^T y),\ i = 1, ..., m;\ 1 + (c^*)^{-1} c^T y \geq 0\},$$

and we can take, as F, the $(m+1)$-self-concordant barrier

$$F(y) = -\sum_{i=1}^{m} \ln \left((b_i(1 + (c^*)^{-1} c^T y) - a_i^T y \right) - \ln \left(1 + (c^*)^{-1} c^T y \right);$$

this is, essentially, the potential function of the original algorithm of Karmarkar, and with this choice of F the outlined method becomes this very algorithm (in its equivalent form found by Bayer and Lagarias).

Now we understand that the only "problem-dependent" element of the construction is the choice of barrier F for the transformed domain G^+. It can be shown that such a barrier can be obtained straightforwardly from a self-concordant barrier for the original domain G, so that already known to us self-concordant barriers allow to build the method of Karmarkar for quadratically constrained programs, semidefinite programs, etc.

16.3 Path-following interior point methods

The method of Karmarkar is of the *potential reduction* nature: as we remember from the analysis of the method, the only thing we are interested in is to decrease the barrier F for the (transformed) domain of the problem as fast as possible; the damped Newton method provides us with certain basic policy of this type, but we are welcome to use all kinds of tricks to get better progress in the barrier, say, to minimize the barrier along the Newton direction instead of performing a step of certain prescribed size. From the computational viewpoint, this is an important advantage of the method – due to linesearch, the practical behaviour of the algorithm is, typically, much better than the worst-case behaviour predicted by the theory. At the same time, from theoretical viewpoint, the method of Karmarkar is not the most efficient one; since the seminal papers of Renegar [9] and Gonzaga [1] we know that there are theoretically more powerful *path-following* interior point schemes in Linear Programming. In early times of the interior point science, the path-following methods, in spite of their nice theoretical properties, were thought to be impractical as compared to the method of Karmarkar or to the more advanced potential reduction methods. The reason was that the early theory of the path-following methods enforced them to work exactly as the (typically very conservative) worst-case efficiency estimate says, not allowing the methods to utilize favourable circumstances they can met. Now the situation is being changed: there are not only extremely efficient practical implementations of the path-following scheme, but also theoretically valid ways for on-line adjustment of the methods to the local geometry of the problem. What we are about to do is to explain briefly the essence of the path-following scheme and to present in more details one of the methods from this family; the choice of the method is mainly determined by the personal preferences of the author.

16.3.1 The standard path-following scheme

The essence of the path-following approach is clearly seen from the following basic description.

In order to solve convex program in the standard form

$$c^T x \to \min \mid x \in G \subset \mathbf{R}^n, \quad [c \neq 0] \tag{16.183}$$

G being a solid, we equip the feasible domain of the problem with a ϑ-self-concordant barrier F and introduce the following single-parameter family of barriers:

$$F_t(x) = -\vartheta \ln(t - c^T x) + F(x);$$

here the parameter t should be greater than the optimal value c^* of the problem.

Since $-\ln(t - c^T x)$ is 1-self-concordant barrier for the half-space $\{x \mid c^T x \leq t\}$ (Example 16.2.1 + Proposition 16.2.6.(iii)), the function F_t is (2ϑ)-self-concordant barrier for the domain

$$G_t = \{x \in G \mid c^T x \leq t\}$$

(Proposition 16.2.6.(i)). Since the domain is bounded along with G, the barrier F_t possesses a unique minimizer $x^*(t)$ on int G_t (Proposition 16.2.3). Thus, we come to the *path of analytic centers*

$$x^*(t) = \operatorname*{argmin}_{\text{int } G_t} F_t(\cdot) : (c^*, +\infty) \to \text{int } G.$$

As t approaches from above the optimal value of the problem, the path of analytic centers clearly converges to the optimal set. And in the method of analytic centers we *trace* the path – given an iterate (t^i, x^i) with $t^i > c^*$ and $x^i \in \text{int } G_{t^i}$ being "close enough" to $x^*(t^i)$, we set the parameter to a smaller value t^{i+1} and then update x^i into a tight enough approximation x^{i+1} to $x^*(t^{i+1})$.

More exactly, let us fix a tolerance $\kappa \in (0, 0.125]$, and let us say that a pair (t, x) is *close* to the path of analytic centers, if

$$(t > c^*) \& (x \in \text{int } G_t) \ \& \ (\lambda(F_t, x) \leq \kappa).$$

At i-th step of the method we update pair (t^{i-1}, x^{i-1}) close to the path into a new pair (t^i, x^i) with the same property; namely, we set

$$t^i = t^{i-1} - dt^i, \quad dt^i = \frac{1}{2}\left(\frac{\partial^2}{\partial t^2} F_t(x)\right)^{-1/2}$$

and then apply to $F_{t^i}(\cdot)$ the damped Newton method (16.178), the method being started at x^{i-1}. The method is terminated when the current iterate, let it be called y, turns out to satisfy the "closeness relation" $\lambda(F_{t^i}, y) \le \kappa$. When it happens, we choose y as x^i and loop.

It can be proved that if we initialize the outlined procedure at (any) close to the path of analytic centers pair (t^0, x^0), then

- the number of Newton steps (16.178) at any iteration of the method is bounded from above by certain quantity depending on κ only; setting κ to a once for ever fixed value, say, 0.125, we enforce the *Newton complexity* of an iteration – the number of Newton steps in updating $x^{i-1} \mapsto x^i$ – to be above bounded by an absolute constant;

- the residual $t^i - c^*$ – the upper bound for inaccuracy of x^i in terms of the objective value – goes to 0 linearly with the rate depending on the parameter of the barrier only:

$$(t^i - c^*) \le \left(1 - \frac{O(1)}{\sqrt{\vartheta}}\right)^i (t^0 - c^*),\ i = 1, 2, ...$$

where, as always, $O(1)$ is a positive absolute constant.

As a result, the Newton complexity of ϵ-solution to (16.183) (total # of Newton steps in course of forming the solution) can be bounded from above as

$$\mathcal{N}(\epsilon) \le O(1)\sqrt{\vartheta} \ln\left(2 + \frac{t_0 - c^*}{\epsilon}\right),$$

basically as it is announced in Theorem 16.2.1.

The outlined constructions and results extend onto the nonpolyhedral case the results of the seminal paper of J. Renegar [9] where the first path-following polynomial time method for LP was developed.

16.3.1.1. **Difficulties and extensions**

The standard path-following scheme as presented now possesses several drawbacks, same as asks for several immediate improvements. As about improvements, the most obvious one is as follows: it is clearly seen that the path of analytic centers is smooth. Why should we start approaching the new "target point" $x^*(t^i)$ from the approximation x^{i-1} to the previous target $x^*(t^{i-1})$ rather than to use a natural forecast of $x^*(t^i)$? Implementing this forecast, we come to what is called the *predictor-corrector scheme*; we present a detailed description of the scheme a little bit later.

Now, the most severe drawbacks of the initial path-following scheme are as follows:

- The scheme is a *short step* one: the rate of updating the parameter t comes from the worst-case analysis aimed to ensure fixed Newton complexity of an iteration. Since we hardly could meet with "the worst case" at every iteration, our policy seems to be too conservative. We may hope that typically a significantly larger step in the parameter also will not elaborate the iteration too much; and what we need is a *long-step* tactics of tracing the path – one which gives us certain on-line computationally cheap tools for approximating "the largest possible" stepsize in the parameter still compatible with the desired Newton complexity of iteration.

- In order to start tracing the path of analytic centers, we should once come close to it, and in the aforementioned description there was not a single word on how to start our process.

In fact there is a very simple way to get close to the path. Namely, it is immediately seen that the path converges, as $t \to \infty$, to the *analytic center* of the domain – to the (unique) minimizer $x(F)$ of F over int G. This behaviour is shared by the paths of analytic centers associated with all possible objectives; and it is clearly seen that the paths associated with these objectives cover the entire int G. Indeed, if $\widehat{x} \in \text{int } G$ and

$$d = -\vartheta^{-1} F'(\widehat{x}),$$

then the path

$$x_d^*(t) = \operatorname*{argmin}_{x \in \text{int } G} [-\vartheta \ln(t - d^T x) + F(x)]$$

passes through $\widehat{x}$ as the parameter is $\widehat{t} = 1 + d^T\widehat{x}$. Now assume that we know in advance a starting point $\widehat{x} \in \text{int } G$. Then we can form the corresponding auxiliary path $x_d^*(t)$ and trace it in the aforementioned manner, but now pushing the parameter to $+\infty$ rather than decreasing it. With this policy, we shall eventually come close to $x^*(F)$ and thus – to the "path of interest", which in our now notation is $x_c^*(t)$; coming close to $x_c^*(\cdot)$, we can switch to tracing this latter path.

The presented "two-phase path-following scheme" is fine but, in turn, possesses unpleasant drawbacks. First of all, it requires a priori knowledge an interior point of int G. How to find such a point? This new difficulty can be overcome with the same path-following technique, now applied to certain auxiliary problem of the same structure as (16.183), but with known in advance interior initial solution. The overall process, however, becomes not that attractive from the practical viewpoint: we end up with something like four-phase method, with its own path to be traced at each of the phases. There were several proposals to combine all these numerous phases. What we are about to do is to present certain unified framework for these combined strategies – a *long-step surface-following* scheme capable to meet all our needs. The below constructions and results originate from [8].

16.3.2 Surface-following scheme

We start with introducing our main subject – the one of the *surface of analytic centers.*

16.3.2.1. **Surface of analytic centers**

The path of analytic centers

$$x_c^*(t) = \operatorname*{argmin}_{\text{int } G_t} F_t(x) \quad [F_t(x) = -\vartheta \ln(t - c^T x) + F(x), G_t = \{x \in G \mid c^T x \le t\}]$$

associated with a ϑ-self-concordant barrier for a convex solid G geometrically can be defined as a set of points $x \in \text{int } G$ where $-F'(x)$ is proportional, with positive coefficient, to the objective c. A natural "multi-parameter" extension of this definition is as follows: let us fix k vectors $c_1, ..., c_k$ and consider the set of all points $x \in \text{int } G$ where $-F'(x)$ can be represented as a combination of the vectors c_i with positive coefficients. We can easily parameterize this set in the same way as the path $x_c^*(t)$ is parameterized by t. Namely, let us introduce k-dimensional parameter vector t, let

$$T = \{t \mid \exists x \in G : c_i^T x < t_i,\, i = 1, ..., k\}, \quad G_t = \{x \in G \mid c_i^T x \le t_i,\, i = 1, ..., k\},$$

and let, finally,

$$F_t(x) = -\vartheta \sum_{i=1}^{k} \ln(t_i - c_i^T x) + F(x) : \text{int } G_t \to \mathbf{R} \quad [t \in T].$$

For every $t \in T$ the function $F(t, \cdot)$ is ϑ_*-self-concordant barrier,

$$\vartheta_* = (k+1)\vartheta,$$

for the convex solid G_t (Proposition 16.2.6 (i) and (iii) + Example 16.2.1); consequently, $F(t, x)$ possesses a unique minimizer $x^*(t)$ in int G_t (Proposition 16.2.3). At this minimizer, of course, $-F'$ is a combination of c_i with positive coefficients. It is easily seen that the inverse also is true: each point $x \in \text{int } G$ where $-F'(x)$ is a combination of c_i with positive coefficients is $x^*(t)$ for some $t \in T$. Thus, we have defined parameterization of the set in question and may speak about the *k-parameter surface of analytic centers*

$$\begin{array}{rcl}
\S(c_1, ..., c_k) & = & \{(t, x^*(t)) \in \mathbf{R}^k \times \mathbf{R}^n \mid t \in T, x^*(t) = \operatorname{argmin}_{x \in \text{int } G_t} F_t(x)\}, \\
\text{where} & & \\
T & = & \{t \in \mathbf{R}^k \mid \exists x \in G : c_i^T x < t_i,\, i = 1, ..., k\}, \\
G_t & = & \{x \in G \mid c_i^T x \le t_i,\, i = 1, ..., k\}, \\
F_t(x) & = & -\vartheta \sum_{i=1}^{k} \ln(t_i - c_i^T x) + F(x).
\end{array}$$

Surface of analytic centers associated with convex problem. Consider a solvable convex optimization program in the the same canonical form as in Section 1:

$$P: \quad f_0(u) \to \min \mid f_i(u) \le 0,\, i = 1, ..., m; \ \| u \|_2 \le 1 \quad [u \in \mathbf{R}^p],$$

f_i, $i = 0, ..., m$, being continuous convex functions on the Euclidean ball $B = \{u \in \mathbf{R}^p \mid \| u \|_2 \leq 1\}$. Assume that we know in advance an upper bound $W(P) \in (0, \infty)$ on the absolute values of all f_i, $i = 0, ..., m$, in B. Then we can rewrite P equivalently in the following form:

$$\widehat{P}: \quad c^T x \to \min \mid f^T x \leq 0, \; x \in G \subset \mathbf{R}^n,$$

where

- $n = p+2$ and $x = (u, v, w)$ is obtained form the design vector u of P by adding two scalar variables v and w;
- G is defined as

$$\begin{array}{rcl} G & = & \{x = (u, v, w) \in \mathbf{R}^n \mid f_0(u) \leq v \leq 5W(P); \; f_i(x) \leq w, \; i = 1, ..., m; \\ & & 0 \leq w \leq 3W(P); \; \| u \|_2 \leq 1\}; \end{array}$$

- c and f are given by

$$c^T x = v; \quad f^T x = w \quad [x = (u, v, w)].$$

Note that G clearly is a solid contained in the "cylinder"

$$Q^+ = \{(u, v, w) \mid \| u \|_2 \leq 1; \; -W(P) \leq w \leq 5W(P); \; 0 \leq w \leq 3W(P)\}$$

centrally symmetric with respect to the point

$$\widehat{x} = (u = 0; v = 2W(P); w = \frac{3}{2}W(P))$$

and containing three times smaller concentric cylinder

$$Q^- = \{(u, v, w) \mid \| u \|_2 \leq \frac{1}{3}; \; W(P) \leq v \leq 3W(P); \; W(P) \leq w \leq \frac{3}{2}W(P)\}.$$

In particular, G is "almost symmetric" with respect to $\widehat{x}$:

$$[G : \widehat{x}] \leq 3.$$

Assumptions. In what follows we treat $\widehat{P}$ as our original problem, not necessarily coming from certain P; we assume only that

- The domain G of the problem is a solid with known in advance interior point $\widehat{x}$;
- $\min_{x \in G} f^T x = 0$

(these assumptions are automatically satisfied when $\widehat{P}$ is obtained in the aforementioned manner from a canonical convex program P).

From now on we also assume – and this assumption is crucial – that

(A) we know a ϑ-self-concordant barrier for the domain G.

Under this assumption we can explicitly point out a 3-parameter surface of analytic centers

$$\S(c_1, c_2, c_3) = \{(t, x^*(t))\}$$

associated with the barrier F in such a way that

- I. The surface passes through $\widehat{x}$: $\widehat{x} = x^*(\widehat{t})$ for some known in advance $\widehat{t}$;
- II. The surface "links" the starting point $\widehat{x}$ with the optimal set of problem $\widehat{P}$: $x^*(t)$ converges to the optimal set of $\widehat{P}$ along a properly chosen sequence of values of t.

Indeed, it suffices to specify the surface and $\widehat{t}$ as follows:

$$\begin{array}{rcl}
\widehat{t}_1 &=& c^T\widehat{x} + \sqrt{c^T[F''(\widehat{x})]^{-1}c};\\
\widehat{t}_2 &=& f^T\widehat{x} + \sqrt{f^T[F''(\widehat{x})]^{-1}f};\\
c_1 &=& c;\\
c_2 &=& f;\\
c_3 &=& -\vartheta^{-1}F'(\widehat{x}) - (\widehat{t}_1 - c^T\widehat{x})^{-1}c - (\widehat{t}_2 - f^Tx)^{-1}f;\\
\widehat{t}_3 &=& c_3^T\widehat{x} + 1.
\end{array}$$

It is immediately seen that if the parameter $t \in T$ varies in such a way that $t_3 \to \infty$, $t_2 \to 0$ and $t_1 \to c^*$, c^* being the optimal value in $\widehat{P}$, then $x^*(t)$ converges to the optimal set of $\widehat{P}$. Thus, we can approximate the optimal set in question by tracing the surface (cf. the path-following scheme), enforcing the parameters to vary in the just indicated way. The advantage of this surface-following approach is that we have no problems with the starting point – the process is started at a "nearly central" explicitly given interior point of G, and this point belongs to the surface we should trace.

In order to implement the presented idea of tracing multi-parameter surface, thus avoiding all difficulties with initialization, we should develop

- tactics: rules for updating current "close to the surface" pair (t^{i-1}, x^{i-1}) into a new pair (t^i, x^i) with the same property, provided that we already know *where* to move – what is the direction of the vector $t^i - t^{i-1}$;
- strategy: rules governing the choice of sequential directions of movement in the 3D-space of parameters.

16.3.2.2. **Tactics of tracing surface: the predictor-corrector scheme**

Same as in the path-following situation, we are going to travel along the surface of analytic centers staying close to it. Let us fix one for ever a tolerance $\kappa \in (0, 0.125]$, and let us say that a pair (t, x) is *close* to the surface $\mathcal{S}$, if the following predicate is true:

$$\mathcal{U}_\kappa : \quad (t \in T) \ \& \ (x \in \operatorname{int} G_t) \ \& \ (\lambda(F_t, x) \le \kappa).$$

At a step i of our method we are given

- a close to the surface current iterate $(t, x) \equiv (t^{i-1}, x^{i-1})$;
  $[(t^0, x^0) = (\widehat{t}, \widehat{x})$; this pair for sure is close to the surface$]$
- a direction $dt = dt^i$ in the parameter space $\mathbf{R}^3$,

and our goal is to update this pair into a new, also close to the surface, pair $(t^+, x^+) \equiv (t^i, x^i)$ with $t^i - t^{i-1}$ being proportional, with positive coefficient r_i, to dt^i. The generic scheme of the updating

$$(t, x, dt) \mapsto (t^+, x^+)$$

is as follows:

Predictor-corrector scheme:

Predictor step:

1. "Lift" the direction dt in the space of parameters to a direction (dx, dt) in the (t, x)-space by setting

$$dx = -\left[\frac{\partial^2}{\partial x^2}F_t(x)\right]^{-1}\left[\frac{\partial^2}{\partial t \partial x}F_t(x)\right]dt$$

and define the *primal search line*

$$\begin{array}{rcl}
R &=& \{(t(r), x(r)) = (t + rdt, x - d_x(t, x) + rdx) \mid r \in \mathbf{R}\},\\
d_x(t, x) &=& \left[\frac{\partial^2}{\partial x^2}F_t(x)\right]^{-1}\frac{\partial}{\partial x}F_t(x).
\end{array}$$

Comment: R is comprised of "forecasts" of the points $(t', x^*(t'))$ of the surface with $t' - t = rdt$. Indeed, the surface is defined by the equation

$$\frac{\partial}{\partial y} F_\tau(y) = 0.$$

Linearizing this equation at $(\tau, y) = (t, x)$, we get equation in variations

$$\frac{\partial}{\partial x} F_t(x) + \left[\frac{\partial^2}{\partial x^2} F_t(x)\right] \Delta x + \left[\frac{\partial^2}{\partial t \partial x} F_t(x)\right] \Delta t = 0;$$

solving the equation with respect to Δx, Δt being set to rdt, we get exactly $x(r) - x$.

2. Choose a stepsize $r > 0$ ensuring the predicate

$$\mathcal{R}: \quad \big(x(r) \in \operatorname{int} G_{t(r)}\big) \, \& \, \big(F_{t(r)}(x(r)) - \min F_{t(r)} \leq 2\big)$$

and set

$$t^+ = t(r) \quad [= t + rdt]; \qquad \widetilde{x} = x(r).$$

Corrector step:

3. Starting with $y^0 = \widetilde{x}$, minimize $F_{t^+}(\cdot)$ with the damped Newton method

$$y^{l+1} = y^l + \frac{1}{1 + \lambda(F_{t^+}, y^l)} \left[\frac{\partial^2}{\partial x^2} F_{t^+}(y^l)\right]^{-1} \frac{\partial}{\partial x} F_{t^+}(y^l); \tag{16.184}$$

terminate the recurrence when it for the first time happens that $\lambda(F_{t^+}, y^l) \leq \kappa$ and set $x^+ = y^l$, thus ensuring that (t^+, x^+) satisfies $\mathcal{U}_\kappa$. Updating $(t, x, dt) \mapsto (t^+, x^+)$ is complete.

Since the stepsize r in 2. is subject to the restriction $\mathcal{R}$ which states that the residual $F_{t^+}(y^0) - \min F_{t^+}$ of the starting point for (16.184) is ≤ 2 and F_{t^+} is self-concordant, we obtain from Corollary 16.2.1 the following important fact:

Proposition 16.3.1 *The Newton complexity (# of Newton steps (16.184)) of the Predictor-Corrector Scheme is bounded from above by $O(1) \ln\ln(1/\kappa)$.*

The just indicated statement explains the origin of the restriction on the stepsize expressed by the predicate $\mathcal{R}$: this restriction is certain indirect way to control the Newton complexity of the corrector step (and thus the overall complexity of the method). The point, of course, is how to ensure $\mathcal{R}$ with "large enough" stepsizes. This is the issue we are coming to.

Acceptability test. First of all, it turns out that $\mathcal{R}$ can be satisfied by "short steps" similar to those used in the worst-case oriented path-following methods:

Proposition 16.3.2 *Let (t, x) be close to the surface. Then $\mathcal{R}$ for sure is satisfied by the stepsize*

$$r = r_*(t, x) = 0.89 \left((dt)^T \left[\frac{\partial^2}{\partial t^2} F_t(x)\right] dt\right)^{-1/2}. \tag{16.185}$$

Already short steps are sufficient to get the standard overall complexity bound for the method; but of course we would prefer to have tools for "long steps" – computationally cheap on-line routines for choosing larger stepsizes still compatible with $\mathcal{R}$. To this end we intend to use certain duality-based *upper bounds* for the residual

$$V(\tau, y) = F_\tau(y) - \min F_\tau(\cdot).$$

To use these bounds, we need the following structural assumption on the self-concordant barrier F in question:

(B) F is represented as

$$F(x) = \Phi(\pi x + p),$$

where

- $x \mapsto \Pi(x) = \pi x + p$ is a given affine mapping from $\mathbf{R}^n$ into certain $\mathbf{R}^N$;
- Φ is a nondegenerate ϑ-self-concordant barrier for a closed convex domain $H \subset \mathbf{R}^n$ such that the image of Π intersects int H and $G = \Pi^{-1}(H)$;
- we know the Legendre transformation

$$\Phi_*(\eta) = \sup_y \left[\eta^T y - \Phi(y)\right]$$

of the barrier Φ.

"We know" Φ_* means that, given $\eta \in \mathbf{R}^N$, we are able to detect whether $\eta \in \operatorname{Dom}\Phi_*$ and if it is the case are able to compute $\Phi_*(\eta)$.

In the mean time we shall see that (B) is satisfied in many important applications.

Note that under assumption (B) our aggregate $F_t(x)$ can be represented as

$$F_t(x) = \Psi(\sigma t + \pi x + p),$$

where

- $\Psi(z) = \Phi(y) - \vartheta \sum_{i=1}^k \ln(w_i)$, $z = (y, w_1, ..., w_k) \in \mathbf{R}^{N+k}$ is nondegenerate self-concordant barrier for the domain
$$Q = H \times \mathbf{R}^k_+,$$
the parameter of the barrier being $\vartheta_* = (k+1)\vartheta$;
- $(t,x) \mapsto Z(t,x) = \sigma t + \pi x + p \equiv \begin{pmatrix} y = \pi x + p \\ w_1 = t_1 - c_1^T x \\ \dots \\ w_k = t_k - c_k^T x \end{pmatrix}$ is affine mapping.

Note that we know the Legendre transformation Ψ_* of Ψ:

$$\Psi_*(\zeta) = \Phi_*(\eta) - \vartheta \sum_{i=1}^k \ln(-\omega_i) + k\vartheta(\ln\vartheta - 1) \quad [\zeta = (\eta, \omega_1, ..., \omega_k)].$$

Under assumption (B) we can equip our Predictor-corrector scheme with the following

Acceptability test:

Given a pair (t,x) close to the surface and a direction dt in the parameter space along with the associated primal search line

$$R = \{(t(r), x(r)) = (t + rdt, x - d_x(t,x) + rdx) \mid r \in \mathbf{R}\},$$

lift R to $\mathbf{R}^{N+k}$ with the help of the affine mapping Z, thus getting the line

$$\begin{array}{rcl} L & = & \{z(r) \equiv z + rdz - d_z \mid r \in \mathbf{R}\}, \\ \text{where} & & \\ z & = & Z(t,x) = \sigma t + \pi x + p, \\ dz & = & \sigma dt + \pi dx, \\ d_z & = & \pi d_x(t,x), \end{array}$$

and form the dual search line

$$L_* = \{\zeta(r) \equiv \zeta + rd\zeta - d_\zeta \mid r \in \mathbf{R}\}, \quad \zeta = \Psi'(z), \quad d\zeta = \Psi''(z)dz, \quad d_\zeta = \Psi''(z)d_z.$$

In order to check whether a candidate stepsize r fits $\mathcal{R}$, verify the inequality

$$\Phi(z(r)) + \Psi_*(\zeta(r)) - [\sigma(t + rdt) + p]^T \zeta(r) \le 2 \qquad (16.186)$$

(the right hand side is $+\infty$ if $z(r) \notin \operatorname{Dom}\Psi$ or if $\zeta(r) \notin \operatorname{Dom}\Psi_*$). If this inequality is satisfied, accept the stepsize, otherwise reject it.

Explanation. Basic facts underlying the Acceptability test are as follows:

1^0. The Legendre transformation of a nondegenerate self-concordant function also is nondegenerate and self-concordant; twice taken Legendre transformation results in the original function

2^0. Let $\xi \in \mathrm{Dom}\,\Psi_*$ satisfy the linear equation

$$\pi^*\xi = 0.$$

Then for all (τ, y) the quantity

$$\Psi(\sigma\tau + \pi y + p) + \Psi_*(\xi) - [\sigma\tau + p]^T s \tag{16.187}$$

is an upper bound for the residual $V(\tau, y) = F_\tau(y) - \min F_\tau(\cdot)$ (all functions are extended outside their domains by the value $+\infty$).
[2^0 is given by the following computation:

$$\begin{array}{rcl} \min F_\tau(\cdot) & = & \min_u \Psi(\sigma\tau + \pi u + p) \\ & & \quad \text{[since, by } 1^0\text{, } \Psi \text{ is the Legendre transform of } \Psi_*] \\ & \geq & \min_u \left[[\sigma\tau + \pi u + p]^T\xi - \Psi_*(\xi)\right] \\ & & \quad [\text{since } \pi^T\xi = 0] \\ & = & [\sigma\tau + p]^T\xi - \Psi_*(\xi), \end{array}$$

and (16.187) follows.]

3^0. Let us fix a candidate stepsize r, and let

$$\Delta z = -d_z + r[\sigma dt + \pi dx]; \quad \Delta\zeta = \Psi''\Delta z$$

(here and below derivatives of Ψ, Ψ_* are taken at z, ζ, respectively), so that $z(r) = z + \Delta z$ and $\zeta(r) = \zeta + \Delta\zeta$. Direct computation taking into account the origin of the involved quantities implies that

a) $\pi^*\zeta(r) = 0$, so that the left hand side of (16.186), by 2^0, is an upper bound for the residual $V(t+rdt, x(r))$. It follows that *the Acceptability test is valid: if it accepts a candidate stepsize, the stepsize for sure satisfies $\mathcal{R}$;*

b) $\zeta \in \mathrm{Dom}\,\Psi_*$ and

$$\| \Delta z \|^2_{\Psi''(z)} = \| \Delta\zeta \|^2_{\Psi''_*(\zeta)} = (\Delta\zeta)^T\Delta z;$$

the common value of all these quantities is

$$\rho^2(r) \equiv \lambda^2(F_t, x) + r^2 \min_h \| \sigma dt + \pi h \|^2_{\Psi''(z)};$$

in particular, we have

$$\rho^2(r) \leq \kappa^2 + r^2 \| \sigma dt \|^2_{\Psi''(z)} \equiv \kappa^2 + r^2(dt)^T\left[\frac{\partial^2}{\partial t^2}F_t(x)\right]dt.$$

c) The left hand side $v(r)$ of (16.186) is the remainder in the second-order Taylor expansion of the function $\Xi(w,\xi) = \Psi(w) + \Psi_*(\xi)$, the expansion being taken at the point (z,ζ) along the displacement $(\Delta z, \Delta\zeta)$; according to (16.176) and b), we have

$$v(r) \leq 2\Omega(\rho(r)) \leq 2\Omega\left(\sqrt{\kappa^2 + r^2(dt)^T\left[\frac{\partial^2}{\partial t^2}F_t(x)\right]dt}\right). \tag{16.188}$$

Moreover, the third order derivative of Ξ taken at the point (z,ζ) along the direction $(\Delta z, \Delta\zeta)$ is zero, so that $v(r)$ is in fact the remainder in the third order Taylor expansion of Ξ.

We see that a candidate stepsize r which passes the Acceptability test indeed implies $\mathcal{R}$, and that r for sure passes the test (and therefore satisfies $\mathcal{R}$) if the right hand side in (16.188) is ≤ 2. By the way, this conclusion justifies acceptability of "short steps" given by Proposition 16.3.2

The Acceptability test, when available (i.e. when (B) is satisfied), allows to equip the Predictor-corrector scheme with on-line tools for choosing "large" stepsizes; to this end we can use line search to identify the largest stepsize still accepted by the test.

How long are "long steps". A natural question is: whether the outlined "long step" tactics indeed results in something "significantly better" than the default short steps (16.185). For the sake of simplicity, let us answer this question for the ideal case when the pair (t,x) under consideration belongs to the surface of analytic centers ($\lambda(F_t,x)=0$) rather than is close to it. In this situation the question we are interested in can be reformulated as follows.

Let $\mathcal{L}=L\times L_*$ be the "primal-dual search line" in the primal-dual space $\mathcal{E}\equiv\mathbf{R}^{N+k}_z\times\mathbf{R}^{N+k}_\zeta$, let $\mathcal{G}$ be the domain of the "primal-dual potential"

$$\Xi(\mathcal{X})=\Psi(u)+\Psi_*(\xi)\quad[\mathcal{X}=(u,\xi)]$$

and let $\widehat{\mathcal{X}}=(z,\zeta)$ (we use the notation from the description of the Acceptability test). The point $\widehat{\mathcal{X}}$ belongs to $\mathcal{G}$ (see b)), and since (t,x) is on the surface, the line $\mathcal{L}$ passes through $\widehat{\mathcal{X}}$:

$$\mathcal{L}=\{\mathcal{X}=\widehat{\mathcal{X}}+rd\mathcal{X}\mid r\in\mathbf{R}\}.$$

It is convenient to equip the primal-dual space $\mathcal{E}$ with the Euclidean norm

$$\|\mathcal{X}\|\equiv\|\mathcal{X}\|_{\Xi''(\widehat{\mathcal{X}})}\,.$$

Let T be the distance from $\widehat{\mathcal{X}}$ to the boundary of $\mathcal{G}$ along the line $\mathcal{L}$, so that the line interval $\{\mathcal{X}\in\mathcal{L}\mid\|\mathcal{X}-\widehat{\mathcal{X}}\|<T\}$ is contained in $\mathcal{G}$ and at least one of its endpoints is outside $\mathcal{G}$. Note that Ξ is self-concordant, so that the open unit Dikin ellipsoid of the function centered at $\widehat{\mathcal{X}}$ – i.e., the centered at $\widehat{\mathcal{X}}$ open unit $\|\cdot\|$-ball – is contained in $\mathcal{G}$ (Proposition 16.2.2), whence $T\geq 1$ (in fact $T\geq\sqrt{2}$ due to the "direct sum" nature of Ξ and the special orientation of $\mathcal{L}$).

Now, when using our Predictor-corrector scheme equipped with the Acceptability test, we in fact perform a step along the primal-dual search line $\mathcal{L}$ staying within $\mathcal{G}$ (this gives us the forecast along with justification that it fits $\mathcal{R}$); and what we are interested in is the size of the step. It follows that T is a "natural upper bound" for the step we can take – if the direction of movement is badly oriented, larger step simply would push us out of $\mathcal{G}$. With these preliminaries, the question "how long are long steps" can be posed as follows:

> *Which fraction of the distance T to the boundary of the primal-dual feasible domain $\mathcal{G}$ can we cover in one step?*

E.g., it can be easily seen from b) and c) that we for sure can perform the "short step" $|r|\,\|d\mathcal{X}\|=0.89\sqrt{2}=1.285...$; however, when T is large, this default step covers only a small fraction of T. Can we do something better? The answer depends on the "degree of regularity" of the barrier Φ underlying the entire construction and is as follows.

Regular self-concordant functions. Let Q be a closed convex domain in $\mathbf{R}^n$, let $x\in\operatorname{int}Q$, and let

$$Q_x=(Q-x)\cap(x-Q)$$

be the symmeterization of Q with respect to x (translated further to make x the origin). Q_x is a convex symmetric neighbourhood of the origin, and therefore it defines a seminorm:

$$|h|_{Q,x}=[\max\{r:x\pm rh\in Q\}]^{-1}\,.$$

Now let F be a self-concordant function on Q, and let $\alpha\geq 1$. We say that F is $\underline{\alpha\text{-regular}}$, if it is four times continuously differentiable on int Q and satisfies the differential inequality

$$\forall(x\in\operatorname{int}Q,h\in\mathbf{R}^n):\quad|D^4F(x)[h,h,h,h]|\leq\alpha(\alpha+1)D^2F(x)[h,h]|h|^2_{Q,x}.$$

Example 16.3.1 *The standard self-concordant barrier (cf. Example 16.2.2)*

$$F(x) = -\ln \operatorname{Det} x \tag{16.189}$$

for the cone $\mathbf{S}^m_+$ of $m \times m$ positive semidefinite symmetric matrices is 2-regular.

> In the situation of Example 16.3.1 $|h|_{\mathbf{S}^m_+,x}$ is just the spectral norm $|\cdot|$ of the symmetric matrix $x^{-1/2}hx^{-1/2}$, and the statement of the Example is equivalent to the standard inequality
>
> $$\operatorname{Tr}(x^4) \le |x|^2 \operatorname{Tr}(x^2),$$
>
> x being a symmetric matrix.

The property of α-regularity with reasonable value of α is shared by many standard self-concordant functions; e.g., all self-concordant barriers indicated in Section 2, same as their Legendre transformations are (at most) 6-regular. Note also that α-regularity is stable with respect to the elementary composition rules mentioned in Proposition 16.2.6.

The role of regularity in our current considerations is seen from the following

Proposition 16.3.3 *Let both ϑ-self-concordant barrier Φ involved into (B) and its Legendre transformation Φ_* be α-regular. Then all stepsizes r satisfying the inequality*

$$|r| \parallel d\mathcal{X} \parallel \le \gamma_\alpha \sqrt{T},$$

same as all stepsizes r satisfying the inequality

$$|r| \parallel d\mathcal{X} \parallel \le \gamma_\alpha T \vartheta^{-1/4}$$

for sure pass the Acceptability test; here $\gamma_\alpha > 0$ depends on α only.

In many important "regular" cases (e.g., in Linear and Semidefinite Programming; both these examples belong to the situation when Φ is given by (16.189)), the "typical" value of T is essentially larger than 1 (e.g., for the barrier (16.189) this typical value is $O(\sqrt{m})$), and whenever it is the case, we may hope that our long-step policy results in stepsizes much larger than the default "short" ones.

16.3.2.3. **Strategy of tracing surface: where to move**

Now it is time to resolve the "strategy issue" – how to choose sequential directions dt^i in the parameter space. If we were tracing the *path* of analytic centers rather than the 3-parameter surface, this issue would not bother us at all – there were the only parameter of interest, and we should decrease at the highest rate compatible with our upper bound on the Newton complexity of the corrector steps (recall that this bound is indirectly given by the predicate $\mathcal{R}$ which should be satisfied by the stepsizes). When tracing *surface* of analytic centers, we do have to decide where to move; this is the price we pay for avoiding all kinds of initialization difficulties. As we shall see, this price seems to be quite appropriate.

Recall that we have associated with the problem

$$\widehat{P}: \quad c^T x \to \min \mid f^T x \le 0,\ x \in G \subset \mathbf{R}^n$$

the 3-parameter surface of analytic centers

$$\begin{array}{rcl} \mathcal{S} & = & \{(t, x^*(t)) \mid t \in T\}, \\ x^*(t) & = & \operatorname*{argmin}_{x \in \operatorname{int} G_t} [-\vartheta \sum_{j=1}^{3} \ln(t_j - c_j^T x) + F(x)] \\ G_t & = & \{x \in G \mid c_1^T x \equiv c^T x \le t_1;\ c_2^T x \equiv f^T x \le t_2; \\ & & c_3^T x \le t_3\}, \\ T & = & \{t \mid \exists x \in G : c_j^T x < t_j,\ j = 1,2,3\}. \end{array}$$

In order to approximate optimal solution to $\widehat{P}$, we should enforce the parameter t_2 responsible for violation of the constraint $f^T x \le 0$ to approach zero, while the parameter t_1 should approach the optimal value

c^* of the problem. As about the "centering" parameter t_3 (introduced, along with c_3, in order to enforce the surface to pass through a given starting point $\widehat{x} \in \operatorname{int} G$), the only thing we are interested in is to vary it in a way which allows the "parameters of interest" t_1 and t_2 to approach their target values, and a safe decision here is to push t_3 to ∞, since with too small value of t_3 the artificial constraint $c_3^T x \leq t_3$ can change the optimal value in the problem.

We already know what are our abilities to move: if we currently are at a point $t^{i-1} \in T$, have computed x^{i-1} such that (t^{i-1}, x^{i-1}) is close to the surface and have chosen direction dt^i of movement in the space of parameters, then our new position in the parameter space will be

$$t^i = t^{i-1} + r_i dt^i,$$

$r_i > 0$ being the stepsize. We for sure can use as r_i the "short step" $r_*(t^{i-1}, x^{i-1})$ given by Proposition 16.3.2; direct computation results in

$$r_{*,i} \equiv r_*(t^{i-1}, x^{i-1}) = \frac{0.89}{\sqrt{\vartheta \sum_{j=1}^{3} (dt_j^i / \Delta_j(t^{i-1}, x^{i-1}))^2}}, \quad \Delta_j(t,x) = t_j - c_j^T x \tag{16.190}$$

(superscripts and subscripts denote the iteration and the coordinate indices, respectively). We see from (16.190) that the natural measure of variation of t_j is the *relative variation* $\delta t_j^i = dt_j^i / \Delta_j(t^{i-1}, x^{i-1})$, not dt_j^i itself.

An important fact (completely similar to the one established in the LP case by Renegar [9]) is as follows:

Proposition 16.3.4 *Let (t,x) be close to $\mathcal{S}$. Then*

$$\Delta_j(t,x) \leq t_j - \min_{x \in G_t} c_j^T x \leq 8\Delta_j(t,x). \tag{16.191}$$

Note that in fact the "weights" ϑ at the log-terms in our aggregate $F_t(\cdot)$ are aimed exactly to make true a relation like (16.191).

Now we are able to present a "good", both from theoretical and seemingly practical viewpoint, policy of where to move.

> *given current close to the surface $\mathcal{S}$ iterate $z^{i-1} = (t^{i-1}, x^{i-1})$, choose the direction dt^i according to the rule*
>
> $$dt^i = \left\{ \begin{array}{llll} (-\Delta_1(z^{i-1}), & -\Delta_2(z^{i-1}), & 16\Delta_3(z^{i-1})), & t_2^{i-1} \leq 16\Delta_2(z^{i-1}) \\ (\ \ \Delta_1(z^{i-1}), & \quad 0 \quad , & 16\Delta_3(z^{i-1})), & \text{otherwise} \end{array} \right. . \tag{16.192}$$
>
> *Motivation* behind the indicated policy is as follows. First of all, we need to get rid of the "centering parameter" t_3 pushing it to ∞ as fast as possible; it turns out that to this end it suffices to enforce δt_3^j to be by absolute constant times greater than $|\delta t_1^i|$ and $|\delta t_2^i|$, and this is exactly what (16.192) does. With such a policy, t_3^i grows exponentially with j, so that the constraint $c_3^T x \leq t_3^i$, starting with certain not too large moment i, becomes redundant in the description of G_{t^i}.
>
> To explain the policy for updating the "parameters of interest" t_1 and t_2, it is worthy to forget about the "centering inequality" $c_3^T x \leq t_3$, as if we were traveling along 2-parameter surface associated with c_1 and c_2; it was already explained that this is basically the situation we eventually come to. We should enforce $t_2 \to 0$, so that it makes sense to decrease it; as about t_1, it should approach the optimal value c^* which we do not know. A natural idea is to decrease both t_2 and t_3 at the highest possible rate until it becomes evident that t_1 is too small – is less than c^*. When it happens, we should increase t_1 in order "to release t_2" – to allow t_2 to approach its target value 0. This is exactly what is done by the branching in (16.192). Indeed, if we see that $t_2^i > 16\Delta_2(t^{i-1}, x^{i-1})$, we can conclude that $t_1^{i-1} < c^*$, since then Proposition 16.3.4 implies that
>
> $$\min_{G_{t^{i-1}}} f^T x \equiv \min[f^T x \mid x \in G,\ c^T x \leq t_1^{i-1}] > 0$$
>
> (recall that we have agreed to ignore the constraint $c_3^T x \leq t_3^{i-1}$ in the description of $G_{t^{i-1}}$), and the resulting inequality is possible only if $t_2^{i-1} < c^*$.

After the direction dt^i is chosen, we use the Predictor-corrector scheme to choose a stepsize r_i in this direction. Besides our crucial restriction to be compatible with $\mathcal{R}$, we subject the stepsize to the restrictions

$$r_{*,i} \le r_i \le \frac{1}{8};$$

the lower bound on r_i simply says that we do not allow to move slower than with short steps (16.190); the upper bound is a kind of safeguard; it would be too time consuming to explain the origin of this bound. Note that the upper bound on r_i is by order of magnitudes larger than the lower one, since (16.192) and (16.190) result in

$$r_{*,i} \le 0.0557\vartheta^{-1/2};$$

thus, there is enough room for "long steps".

Now the description of the method is complete, and it is time to evaluate its complexity.

16.3.2.4. **Complexity**

To describe the rate of convergence of the resulting method, let us denote by $\mathcal{N}(\epsilon)$ the first i such that all x^j, $j \ge i$, are ϵ-solutions to $\widehat{P}$, so that

$$j \ge i \Rightarrow (c^T x^j - c^* \le \epsilon) \ \& \ (f^T x^j \le \epsilon).$$

Theorem 16.3.1 *For all $\epsilon > 0$ one has*

$$\begin{array}{rcl} \mathcal{N}(\epsilon) & \le & O(1)\sqrt{\vartheta}\ln\left(2\vartheta[G\!:\!\widehat{x}]\dfrac{1+\nu(\widehat{P},\epsilon)}{\nu(\widehat{P},\epsilon)}\right), \\ \nu(\widehat{P},\epsilon) & = & \min\left[\dfrac{\epsilon}{\max_{x\in G} c^T x - \min_{x\in G} c^T x}, \dfrac{\epsilon}{\max_{x\in G} f^T x}\right]. \end{array} \tag{16.193}$$

The Newton complexity of any corrector step of the method does not exceed $O(1)\ln\ln(1/\kappa)$.

Remark 16.3.1 In the above form, our surface-following method of analytic centers is a converging infinite process. We can easily equip the method with on-line computationally cheap termination rules capable to terminate the process and to output an ϵ-solution to the problem, $\epsilon > 0$ being a given on input value of the accuracy; the number of iterations in this "finite" routine still fits the complexity bound (16.193).

To the moment we assume $\widehat{P}$ to be feasible; this assumption also can be eliminated: the method can be equipped with an on-line computationally cheap test which correctly detects infeasibility; with this test, infeasibility of (infeasible) $\widehat{P}$ is detected in no more than

$$N^* = O(1)\sqrt{\vartheta}\ln\left(2\vartheta[G\!:\!\widehat{x}]\frac{\min_{x\in G} f^T x + \max_{x\in G} f^T x}{\min_{x\in G} f^T x}\right)$$

iterations.

16.4 Applications

In this concluding Section we present interior point based complexity bounds for several "well-structured" generic convex problems. We process the programs in question as explained in Section 16.3.2, in particular, use the same notation. Speaking about a particular problem, we sequentially

- present the initial formulation of the problem,
- explain how a problem instance is converted to the standard form

$$\widehat{P}: \quad c^T x \to \min \mid f^T x \leq 0; \, x \in G,$$

- equip the arising domain G with self-concordant barrier (trying to fit the structural assumption (B) underlying the long-step tactics from Section 16.3.2) and point out a "nearly centered" starting point $\widehat{x} \in \operatorname{int} G$,
- indicate the Newton complexity of finding an ϵ-solution to $\widehat{P}$ by the surface-following method from Section 16.3.2, same as the arithmetic cost of the solution; for the sake of definiteness, the tolerance parameter κ of the method is set to 0.125. The estimates of the arithmetic cost correspond to the case when matrices are multiplied and inverted by the traditional Linear Algebra routines.

In what follows all $O(1)$'s are absolute constants which do not depend on any parameter of situation. Note that the arithmetic cost of an ϵ-solution corresponds to the case when problem instances have no additional data structure; in large-scale applications, however, this typically is not the case, so that normally the arithmetic cost of ϵ-solution is by order of magnitudes less than the upper bound we present.

16.4.1 Linear Programming

16.4.1.1. **Family of problems:**

Problem instance: solvable LP program

$$P: \quad e^T u \to \min \mid a_i^T u \leq b_i, \; i = 1, ..., m; \; \| u \|_\infty \leq 1 \quad [u \in \mathbf{R}^n];$$

Data:

$$\begin{array}{l} \mathrm{Data}(P) = [m; n; e; a_1, b_1; ...; a_m, b_m], \\ \dim \mathrm{Data}(P) = (m+1)(n+1) + 1. \end{array}$$

ϵ-solution: any u such that

$$\begin{array}{rcl} \| u \|_\infty & \leq & 1, \\ a_i^T u & \leq & b_i + \epsilon, \; i = 1, ..., m, \\ e^T u & \leq & e^* + \epsilon \quad [e^* \text{ is the optimal value in } P]. \end{array}$$

Scale factor: $\mathrm{V}(P) = \max[\| e \|_1, |b_1| + \| a_1 \|_1, ..., |b_m| + \| a_m \|_1]$;

16.4.1.2. **Standard reformulation:**

$[x = (u, w) \in \mathbf{R}^{n+1}]$:

$$\begin{array}{l} \widehat{P}: \\ \qquad c^T x \equiv e^T u \to \min \\ \text{s.t.} \\ \qquad f^T x \equiv w \leq 0; \\ \qquad x \in G = \{(u, w) \mid \quad \| u \|_\infty \leq 1; \\ \qquad\qquad\qquad\qquad\qquad a_i^T u - w \leq b_i, \; i = 1, ..., m; \\ \qquad\qquad\qquad\qquad\qquad 0 \leq w \leq 3\mathrm{V}(P)\}. \end{array}$$

16.4.1.3. **Self-concordant barrier for G:**

$$\begin{array}{rcl}
F(x) & = & \Phi(\pi x + p), \\
\Phi(y) & = & -\sum_{i=1}^{N} \ln y_i, \\
y & = & (y_1, ..., y_N) \in \mathbf{R}^N, \\
\dim y & = & N \equiv m + 2n + 2, \\
\vartheta(\Phi) & = & m + 2n + 2, \\
\pi \begin{pmatrix} u \\ w \end{pmatrix} + p & = & \begin{pmatrix} b_1 + w - a_1^T u \\ \cdots \\ b_m + w - a_m^T u \\ w \\ 3\mathrm{V}(P) - w \\ u_1 + 1 \\ 1 - u_1 \\ \cdots \\ u_n + 1 \\ 1 - u_n \end{pmatrix} \\
\Phi_*(\eta) & = & -\sum_{i=1}^{N} \ln(-\eta_i) - N, \\
\eta & = & (\eta_1, ..., \eta_N) \in \mathbf{R}^N.
\end{array}$$

F is the standard log-barrier for the polytope G. Note that both Φ and Φ_* are 2-regular.

16.4.1.4. **Starting point $\widehat{x}$:**

$$\widehat{x} = (\widehat{u} = 0, \widehat{w} = \frac{3}{2}\mathrm{V}(P)) \quad [[G : \widehat{x}] \le 3] \tag{16.194}$$

16.4.1.5. **Newton complexity of ϵ-solution:**

$$\mathcal{N}(\epsilon) = O(1)\sqrt{m+n} \ln\left(2(m+n)\frac{1 + \nu(P,\epsilon)}{\nu(P,\epsilon)}\right), \quad \nu(P,\epsilon) = \frac{\epsilon}{\mathrm{V}(P)}.$$

16.4.1.6. **Arithmetic complexity of ϵ-solution:**

$$\mathcal{C}(\epsilon) = O(1)(m+n)^{3/2} n^2 \ln\left(2(m+n)\frac{1 + \nu(P,\epsilon)}{\nu(P,\epsilon)}\right). \tag{16.195}$$

Remark 16.4.1 The arithmetic cost of ϵ-solution given by (16.195) corresponds to the case when the Newton systems arising in course of running the method are assembled and solved “from scratch”. It is known, anyhow, that one can ensure (16.194) replacing the exact solutions to the Newton systems by their tight approximations, and that it can be done in a manner utilizing the results of Linear Algebra computations at preceding steps. With this *Karmarkar acceleration* of Linear Algebra (originating from [7] and incorporated for the first time in the Renegar path-following method by Vaidya) the arithmetic cost of ϵ-solution becomes *cubic* in the dimension of the problem:

$$\mathcal{C}(\epsilon) = O(1)(m+n)(n^2 + n\sqrt{m+n}) \ln\left(2(m+n)\frac{1 + \nu(P,\epsilon)}{\nu(P,\epsilon)}\right).$$

16.4.2 Quadratically Constrained Convex Quadratic Programming

16.4.2.1. **Family of problems:**

Problem instance: solvable program

$$P: \quad f_0(u) \to \min \mid f_i(u) \le 0,\ i = 1, ..., m;\ \| u \|_2 \le 1 \quad [u \in \mathbf{R}^n],$$

where

$$f_i(u) = \frac{1}{2} u^T A_i^T A_i u + b_i^T u + c_i, \ \ i = 0, ..., m,$$

are convex quadratic forms (A_i is $k_i \times n$ matrix of rank k_i, $i = 0, ..., m$).

Data:

$$\begin{array}{l} \mathrm{Data}(P) = [m; n; A_0, b_0, c_0; ...; A_m, b_m, c_m], \\ \dim \mathrm{Data}(P) = (\sum_{i=0}^{m} k_i)n + (m+1)(n+1) + 2. \end{array}$$

ϵ-solution: any u such that

$$\begin{array}{rcl} \| u \|_2 & \le & 1, \\ f_i(u) & \le & \epsilon, \ i = 1, ..., m, \\ f_0(u) & \le & f_0^* + \epsilon \quad [f_0^* \text{ is the optimal value in } P]. \end{array}$$

Scale factor: $\mathrm{V}(P) = \max_{i=0,...,m} \left[\frac{1}{2} |A_i|^2 + \| b_i \|_2 + |c_i| \right]$, $|A|$ being the operator norm (the largest singular value) of a matrix A.

16.4.2.2. **Standard reformulation:**

$[x = (u, v, w) \in \mathbf{R}^{n+2}]$:

$$\begin{array}{ll} \widehat{P}: & \\ & c^T x \equiv v \to \min \\ \text{s.t.} & \\ & f^T x \equiv w \le 0; \\ & x \in G = \{(u, v, w) \mid \quad \| u \|_2 \le 1; \\ & \qquad\qquad\qquad\qquad\quad f_0(u) \le v; \\ & \qquad\qquad\qquad\qquad\quad f_i(u) \le w,\ i = 1, ..., m; \\ & \qquad\qquad\qquad\qquad\quad v \le 5\mathrm{V}(P); \\ & \qquad\qquad\qquad\qquad\quad 0 \le w \le 3\mathrm{V}(P)\}. \end{array}$$

16.4.2.3. **Self-concordant barrier for G:**

$$\begin{array}{rcl} F(x) & = & \Phi(\pi x + p), \\ \Phi(y) & = & -\sum\limits_{i=0}^{m+1} \ln\left(s_i - \frac{1}{2} r_i^T r_i\right) - \sum\limits_{\kappa=1}^{3} \ln q_\kappa \\ y & = & (s_0, r_0; ...; s_{m+1}, r_{m+1}; q_1, q_2, q_3) : \\ & & s_i \in \mathbf{R},\ r_i \in \mathbf{R}^{k_i}\ [k_{m+1} = n],\ q_\kappa \in \mathbf{R}, \\ \dim y & = & 5 + n + m + \sum_{i=0}^m k_i, \\ \vartheta(\Phi) & = & m + 5, \\ \pi \begin{pmatrix} u \\ v \\ w \end{pmatrix} + p & = & \begin{pmatrix} s_0 & = & v - b_0^T u - c_0 \\ r_0 & = & A_0 u \\ s_1 & = & w - b_1^T u - c_1 \\ r_1 & = & A_1 u \\ \cdots & & \cdots \\ s_m & = & w - b_m^T x - c_m \\ r_m & = & A_m u \\ s_{m+1} & = & \frac{1}{2} \\ r_{m+1} & = & u \\ q_1 & = & 5\mathrm{V}(P) - v \\ q_2 & = & 3\mathrm{V}(P) - w \\ q_3 & = & w \end{pmatrix} \end{array}$$

$$\begin{array}{rcl} \Phi_*(\eta) & = & -\sum\limits_{i=0}^{m+1} \left[\ln(-\sigma_i) + \frac{\rho_i^T \rho_i}{2\sigma_i}\right] - \sum\limits_{\kappa=1}^{3} \ln(-\theta_\kappa) - m - 5, \\ \eta & = & (\sigma_0, \rho_0; ...; \sigma_{m+1}, \rho_{m+1}; \theta_1, \theta_2, \theta_3) :\ \sigma_i \in \mathbf{R},\ \rho_i \in \mathbf{R}^{k_i},\ \theta_\kappa \in \mathbf{R}. \end{array}$$

F comes from log-barriers for linear and convex quadratic constraints (Examples 16.2.1, 16.2.3). Note that both Φ and Φ_* are 6-regular.

16.4.2.4. **Starting point $\widehat{x}$:**

$$\widehat{x} = (\widehat{u} = 0, \widehat{v} = 2\mathrm{V}(P), \widehat{w} = \frac{3}{2}\mathrm{V}(P)) \quad [[G : \widehat{x}] \le 3]$$

16.4.2.5. **Newton complexity of ϵ-solution:**

$$\mathcal{N}(\epsilon) = O(1)\sqrt{m+1} \ln\left(2(m+1)\frac{1 + \nu(P,\epsilon)}{\nu(P,\epsilon)}\right), \quad \nu(P,\epsilon) = \frac{\epsilon}{\mathrm{V}(P)}.$$

16.4.2.6. **Arithmetic complexity of ϵ-solution:**

$$\mathcal{C}(\epsilon) = O(1)(m+1)^{1/2} n(n^2 + m + \sum_{i=0}^m k_i^2) \ln\left(2(m+1)\frac{1 + \nu(P,\epsilon)}{\nu(P,\epsilon)}\right).$$

16.4.3 Semidefinite Programming

16.4.3.1. **Family of problems:**

Problem instance: solvable program

$$P: \quad e^T u \to \min \mid \ A_i(u) \geq 0, \ i=1,...,m; \ \| u \|_2 \leq 1 \quad [u \in \mathbf{R}^n],$$

where

$$A_i(u) = A_{i0} + \sum_{j=1}^n u_j A_{ij}$$

with symmetric $k_i \times k_i$ matrices A_{ij}; $A \geq 0$ for symmetric matrix A means that A is positive semidefinite.

Data:

$$\begin{array}{l} \mathrm{Data}(P) = [m; n; e; A_{ij}, i=1,...,m, j=0,...,n], \\ \dim \mathrm{Data}(P) = 2 + n + (n+1) \sum_{i=1}^k \frac{k_i(k_i+1)}{2}. \end{array}$$

ϵ-solution: any u such that

$$\begin{array}{rcl} \| u \|_2 & \leq & 1, \\ A_i(u) + \epsilon I_{k_i} & \geq & 0, \ i=1,...,m \quad [I_k \text{ is the } k \times k \text{ unit matrix}], \\ e^T u & \leq & e^* + \epsilon \quad [e^* \text{ is the optimal value in } P]. \end{array}$$

Scale factor: $\mathrm{V}(P) = \max[\| e \|_2; |A_{ij}|, i=1,...,m, j=0,...,n]$, $|A|$ being the operator norm of matrix A.

16.4.3.2. **Standard reformulation:**

$[x = (u, w) \in \mathbf{R}^{n+1}]$:

$$\begin{array}{ll} \widehat{P}: & \\ & c^T x \equiv e^T u \to \min \\ \text{s.t.} & \\ & f^T x \equiv w \leq 0; \\ & x \in G = \{(u,w) \mid \quad \begin{array}{l} \| u \|_2 \leq 1; \\ A_i(u) + w I_{k_i} \geq 0, \ i=1,...,m; \\ 0 \leq w \leq 3\mathrm{V}(P)\}. \end{array} \end{array}$$

16.4.3.3. **Self-concordant barrier for G:**

$$\begin{array}{rcl} F(x) & = & \Phi(\pi x + p), \\ \\ \Phi(y) & = & -\sum_{i=1}^m \ln \mathrm{Det}\ y_i - \ln\left(s - \frac{1}{2} r^T r\right) - \sum_{\kappa=1}^2 \ln q_\kappa, \\ \\ y & = & (y_1,...,y_m; s, r; q_1, q_2): \ y_i \in \mathbf{S}^{k_i}, \ s \in \mathbf{R}, \ r \in \mathbf{R}^n, \ q_\kappa \in \mathbf{R}, \\ & & \quad [\mathbf{S}^k \text{ is the space of symmetric } k \times k \text{ matrices}], \\ \dim y & = & 3 + n + \sum_{i=1}^m \frac{k_i(k_i+1)}{2}, \\ \\ \vartheta(\Phi) & = & 3 + \sum_{i=1}^m k_i, \\ \\ \pi \begin{pmatrix} u \\ w \end{pmatrix} + p & = & \begin{pmatrix} y_1 & = & A_1(u) + w I_{k_1} \\ \cdots & & \cdots \\ y_m & = & A_m(u) + w I_{k_m} \\ s & = & \frac{1}{2} \\ r & = & u \\ q_1 & = & w \\ q_2 & = & 3\mathrm{V}(P) - w \end{pmatrix} \end{array}$$

$$\begin{array}{rcl} \Phi_*(\eta) & = & -\sum_{i=1}^m \ln \mathrm{Det}\,(-\eta_i) - \left[\ln(-\sigma) + \frac{\rho^T \rho}{2\sigma}\right] - \sum_{\kappa=1}^2 \ln(-\theta_\kappa) \\ & & -3 - \sum_{i=1}^m k_i, \\ \eta & = & (\eta_1,...,\eta_m; \sigma, \rho; \theta_1, \theta_2): \ \eta_i \in \mathbf{S}^{k_i}, \ \sigma \in \mathbf{R}, \ \rho \in \mathbf{R}^n, \ \theta_\kappa \in \mathbf{R}. \end{array}$$

F comes from the logDet-barrier for the cone of positive semidefinite matrices (Example 16.2.2) and log-barriers for linear constraints (Example 16.2.1) and the quadratic bound (Example 16.2.3). Note that both Φ and Φ_* are 6-regular.

16.4.3.4. **Starting point $\widehat{x}$:**

$$\widehat{x} = (\widehat{u} = 0, \widehat{w} = \frac{3}{2}\mathrm{V}(P)) \quad [[G : \widehat{x}] \leq 3]$$

16.4.3.5. **Newton complexity of ϵ-solution:**

$$\mathcal{N}(\epsilon) = O(1)\sqrt{1 + \sum_{i=1}^{m} k_i} \ln\left(2\left[1 + \sum_{i=1}^{m} k_i\right]\frac{1 + \nu(P,\epsilon)}{\nu(P,\epsilon)}\right), \quad \nu(P,\epsilon) = \frac{\epsilon}{\mathrm{V}(P)}.$$

16.4.3.6. **Arithmetic complexity of ϵ-solution:**

$$\mathcal{C}(\epsilon) = O(1)(1 + \sum_{i=1}^{m} k_i)^{1/2} n(n^2 + n\sum_{i=1}^{m} k_i^2 + \sum_{i=1}^{m} k_i^3) \ln\left(2(m+n)\frac{1 + \nu(P,\epsilon)}{\nu(P,\epsilon)}\right).$$

16.4.4 Geometric Programming

16.4.4.1. **Family of problems:**

Problem instance: solvable program

$$P: \quad f_0(u) \to \min \mid f_i(u) \leq 0, \; i = 1, ..., m, \; \| u \|_2 \leq 1 \quad [u \in \mathbf{R}^n],$$

where $f_i(x)$ are logarithms of exponential posynomials:

$$f_i(x) = \ln\left(\sum_{j=1}^{k_i} \exp\{b_{ij} + a_{ij}^T u\}\right).$$

Data:

$$\begin{array}{l} \mathrm{Data}(P) = [m; n; k_0, \{b_{0j}, a_{0j}\}_{j=1,\ldots,k_1}; \ldots; k_m, \{b_{mj}, a_{mj}\}_{j=1,\ldots,k_m}], \\ \dim \mathrm{Data}(P) = 3 + m + (n+1)K, \\ K = \sum_{i=0}^{m} k_i. \end{array}$$

ϵ-solution: any u such that

$$\begin{array}{rcl} \| u \|_2 & \leq & 1, \\ f_i(u) & \leq & \epsilon, \; i = 1, ..., m, \\ f_0(u) & \leq & f_0^* + \epsilon, \quad [f_0^* \text{ is the optimal value in } P]. \end{array}$$

Scale factor: $\mathrm{V}(P) = \max_{i=0,\ldots,m}\left[\ln(3k_i) + \max_{j=1,\ldots,k_i}[|b_{ij}| + \| a_{ij} \|_2]\right].$

16.4.4.2. **Standard reformulation:**

$$x = (u, \{u_{ij} | i = 0, ..., m, j = 1, ..., k_i\}, v, w) \in \mathbf{R}^{n+K+2};$$

$$
\begin{array}{lll}
\widehat{P}: & & \\
& c^T x \equiv v \to \min & \\
\text{s.t.} & & \\
& f^T x \equiv w \le 0; & \\
& x \in G = \{(u, \{u_{ij}\}, v, w) \mid & \| u \|_2 \le 1; \\
(\alpha) & & \exp\{b_{0j} + a_{0j}^T u - v\} \le u_{0j},\ j = 1, ..., k_0; \\
(\beta) & & \exp\{b_{ij} + a_{ij}^T u - w\} \le u_{ij},\ \begin{array}{c} i = 1, ..., m \\ j = 1, ..., k_i \end{array}; \\
(\gamma) & & \sum_{j=1}^{k_i} u_{ij} \le 1,\ i = 0, ..., m; \\
& & 0 \le w \le 3\mathrm{V}(P); \\
& & v \le 5\mathrm{V}(P)\}.
\end{array}
$$

Note that an ϵ-solution to $\widehat{P}$ immediately yields a similar solution for P. Indeed, if u is an ϵ-solution to P, then setting

$$u_{ij} = \exp\{b_{ij} + a_{ij}^T u - f_i(u)\},\ v = f_0(u),\ w = \max_{i=1,\dots,m} f_i(u),$$

we clearly extend u to a solution $x \in G$ to $\widehat{P}$ with $f^T x \le \epsilon$ and $c^T x = v$. Vice versa, if $x = (u, \{u_{ij}\}, v, w) \in G$, then from the inequalities (α), (β) and (γ) it follows that $f_0(u) \le v$ and $f_i(u) \le w$, $i = 1, ..., m$. These observations say that the optimal values in both the problems are equal to each other, and that the u-part of an ϵ-solution $x \in G$ to $widehatP$ (solution with $v \le f_0^* + \epsilon$ and $f^T x \le \epsilon$) is an ϵ-solution to P, as claimed.

16.4.4.3. **Self-concordant barrier for G:**

$$
\begin{array}{rcl}
F(x) & = & \Phi(\pi x + p) \\
\Phi(y) & = & -\displaystyle\sum_{i=0}^{m} \sum_{j=1}^{k_i} [\ln(s_{ij}) + \ln(\ln(s_{ij}) - r_{ij})] \\
& & -\displaystyle\sum_{i=0}^{m} \ln t_i - \sum_{\kappa=1}^{3} \ln q_\kappa - \ln(s - \frac{1}{2} r^T r), \\
y & = & (\{s_{ij}, r_{ij}\}_{i=0,\dots,m, j=1,\dots,k_i}; t_0, ..., t_m; s, r; q_1, q_2, q_3): \\
& & s_{ij}, r_{ij}, t_i, s, q_\kappa \in \mathbf{R},\ r \in \mathbf{R}^n, \\
\dim y & = & m + n + 2K + 4 \\
\vartheta(\Phi) & = & m + 2K + 5
\end{array}
$$

$$
\pi \begin{pmatrix} u \\ \{u_{ij}\} \\ v \\ w \end{pmatrix} + p = \begin{pmatrix}
s_{ij} & = & u_{ij},\ i = 0, ..., m, j = 1, ..., k_i \\
r_{0j} & = & b_{0j} + a_{0j}^T u - v,\ j = 1, ..., k_0 \\
r_{ij} & = & b_{ij} + a_{ij}^T u - w,\ i = 1, ..., m,\ j = 1, ..., k_i \\
t_i & = & 1 - \sum_{j=1}^{k_i} v_{ij},\ i = 0, ..., m \\
q_1 & = & 5\mathrm{V}(P) - v \\
q_2 & = & 3\mathrm{V}(P) - w \\
q_3 & = & w \\
s & = & \frac{1}{2} \\
r & = & u
\end{pmatrix}
$$

$$
\begin{array}{rcl}
\Phi_*(\eta) & = & -\sum_{i=0}^{m} \sum_{j=1}^{k_i} \left[\rho_{ij} + \ln \rho_{ij} - (\rho_{ij} + 1) \ln\left(\frac{\rho_{ij}+1}{-\sigma_{ij}}\right)\right] \\
& & -\sum_{i=0}^{m} \ln(-\tau_i) - \sum_{\kappa=1}^{3} \ln(-\theta_\kappa) \\
& & -\left[\ln(-\sigma) + \frac{\rho^T \rho}{2\sigma}\right] - m - 5 - 2K \\
\eta & = & (\{\sigma_{ij}, \rho_{ij}\}_{i=0,\dots,m, j=1,\dots,k_i}, \tau_0, ..., \tau_m, \sigma, \theta_1, \theta_2, \theta_3, \rho): \\
& & \sigma_{ij}, \rho_{ij}, \tau_i, \sigma, \theta_\kappa \in \mathbf{R}, \rho \in \mathbf{R}^n
\end{array}
$$

The "main ingredient" in the barrier Φ is the 2-self-concordant barrier

$$-\ln(\ln t - s) - \ln t$$

for the epigraph of the exponent. Both Φ and Φ_* are 6-regular.

16.4.4.4. **Starting point $\widehat{x}$:**

$$\widehat{x} = (\widehat{u} = 0, \widehat{u}_{ij} = \frac{1}{2k_i}\ \forall(i,j), \widehat{v} = 2\mathrm{V}(P), \widehat{w} = \frac{3}{2}\mathrm{V}(P)) \quad [[G:\widehat{x}] \le 6K]\,.$$

16.4.4.5. **Newton complexity of ϵ-solution:**

$$\mathcal{N}(\epsilon) = O(1)\sqrt{K}\ln\left(2K\frac{1+\nu(P,\epsilon)}{\nu(P,\epsilon)}\right), \quad \nu(P,\epsilon) \equiv \frac{\epsilon}{\mathrm{V}(P)}.$$

16.4.4.6. **Arithmetic complexity of ϵ-solution:**

$$\mathcal{C}(\epsilon) = O(1)K^{1/2}(K+n)(m+n)^2\ln\left(2K\frac{1+\nu(P,\epsilon)}{\nu(P,\epsilon)}\right).$$

16.4.5 Approximation in $\|\cdot\|_\gamma$

16.4.5.1. **Family of problems:**

Problem instance: solvable program

$$P: \quad f(u) \equiv\parallel Ax - b \parallel_\gamma \to \min \mid \parallel u \parallel_2 \le 1 \quad [u \in \mathbf{R}^n],$$

where $\gamma \in (1,\infty)$, $A = \begin{pmatrix} a_1^T \\ \dots \\ a_m^T \end{pmatrix}$ is an $m \times n$ matrix and $b = \begin{pmatrix} b_1 \\ \dots \\ b_m \end{pmatrix} \in \mathbf{R}^m$.

Data:

$$\begin{array}{l}\mathrm{Data}(P) = [m;n;A,b],\\ \dim \mathrm{Data}(P) = m(n+1)+2.\end{array}$$

ϵ-solution: any u such that

$$\begin{array}{rcl}\parallel u \parallel_2 & \le & 1,\\ f(u) & \le & f^* + \epsilon, \quad [f^* \text{ is the optimal value in } P].\end{array}$$

Scale factor: $\mathrm{V}(P) = \max_{i=1,\dots,m} \parallel a_i \parallel_2 + \parallel b \parallel_\infty$.

16.4.5.2. **Standard reformulation:**

$$x = (u, v, w_1, \dots, w_m) \in \mathbf{R}^{n+m+1};$$

$$\begin{array}{lll}\widehat{P}: & & \\ & c^T x \equiv v \to \min & \\ \text{subject to} & [\alpha = \frac{1}{\gamma}, \beta = 1-\alpha] & \\ & x \in G = \{(u,v,w_1,\dots,w_m) \mid & \parallel u \parallel_2 \le 1;\\ & & v \ge 0;\\ & & w_i \ge 0,\ i = 1,\dots,m;\\ & & |b_i - a_i^T u| \le w_i^\alpha v^\beta;\\ & & \sum_{i=1}^m w_i \le v;\\ & & v \le 5m\mathrm{V}(P)\}.\end{array}$$

Problem $\widehat{P}$ indeed is equivalent to P. To see it, note that P clearly can be rewritten as

$$v \to \min \mid \left(\sum_{i=1}^{m} |b_i - a_i^T u|^\gamma\right)^{1/\gamma} \leq v \leq 5m\mathrm{V}(P),\ \| u \|_2 \leq 1 \tag{16.196}$$

(the upper bound imposed on v is simply redundant). Now, if (u, v) is a feasible solution to (16.196), then the vector

$$(u, v, w_i = |b_i - a_i^T u|^\gamma v^{1-\gamma},\ i = 1, ..., m) \quad [\text{here } 0/0 = 0]$$

clearly is a feasible solution to $\widehat{P}$. Vice versa, if $(u, v, w_1, ..., w_m)$ is a feasible solution to $\widehat{P}$, then

$$\left(\sum_{i=0}^{m} |b_i - a_i^T x|^\gamma\right)^{1/\gamma} \leq \left(\sum_{i=1}^{m} w_i v^{\gamma-1}\right)^{1/\gamma} \leq v \quad [\leq 5m\mathrm{V}(P)],$$

so that (u, v) is a feasible solution to (16.196). It remains to note that the indicated correspondence between feasible solutions to $\widehat{P}$ and to (16.196) preserves the value of the objective.

16.4.5.3. **Self-concordant barrier for G:**

$$\begin{array}{rcl}
F(x) & = & \Phi(\pi x + p), \\
\Phi(y) & = & \sum_{i=1}^{m} \left[\phi(w_{1,i}, v_{1,i}, z_{1,i}) + \phi(w_{2,i}, v_{2,i}, z_{2,i})\right] \\
 & & - \sum_{i=1}^{2} \ln(q_i) - \ln\left(s - \frac{1}{2} r^T r\right) \\
 & & \quad \left[\phi(w, v, z) = -\ln(w^\alpha v^\beta - z) - \ln w - \ln v\right], \\
y & = & (\{w_{\kappa,i}, v_{\kappa,i}, z_{\kappa,i}\}_{\kappa=1,2,i=1,...,m}; s, r; q_1, q_2) : \\
 & & w_{\kappa,i}, v_{\kappa,i}, z_{\kappa,i}, s, q_j \in \mathbf{R},\ r \in \mathbf{R}^n, \\
\dim y & = & 3 + n + 6m, \\
\vartheta(\Phi) & = & 3 + 6m, \\
\pi \begin{pmatrix} u \\ v \\ w_1 \\ \cdots \\ w_m \end{pmatrix} + p & = & \begin{pmatrix} w_{1,i} = w_{2,i} = w_i,\ i = 1, ..., m \\ v_{1,i} = v_{2,i} = v,\ i = 1, ..., m \\ z_{1,i} = b_i - a_i^T u,\ i = 1, ..., m \\ z_{2,i} = a_i^T u - b_i,\ i = 1, ..., m \\ q_1 = v - \sum_{i=1}^{m} w_i \\ q_2 = 5m\mathrm{V}(P) - v \\ s = \frac{1}{2} \\ r = u \end{pmatrix}
\end{array}$$

$$\begin{array}{rcl}
\Phi_*(\eta) & = & \sum_{i=1}^m \left[\phi_*(\omega_{1,i},\nu_{1,i},\zeta_{1,i}) + \phi_*((\omega_{2,i},\nu_{2,i},\zeta_{2,i})\right] \\
& & -\sum_{i=1}^2 \ln(-\theta_i) - \left[\ln(-\sigma) + \frac{\rho^T\rho}{2\sigma}\right] - 3, \\
\eta & = & (\{\omega_{\kappa,i},\nu_{\kappa,i},\zeta_{\kappa,i}\}_{\kappa=1,2,i=1,...,m};\sigma,\rho;\theta_1,\theta_2) : \\
& & \omega_{\kappa,i},\nu_{\kappa,i},\zeta_{\kappa,i},\sigma,\theta_j \in \mathbf{R},\ \rho \in \mathbf{R}^n, \\
\phi_*(\omega,\nu,\zeta) & = & -\ln(\zeta) - \ln\left(\frac{-\omega}{1+\alpha\Lambda_\gamma(\zeta^{-1}|\omega|^\alpha|\nu|^\beta)}\right) - \ln\left(\frac{-\nu}{1+\beta\Lambda_\gamma(\zeta^{-1}|\omega|^\alpha|\nu|^\beta)}\right), \\
& & \mathrm{Dom}(\phi_*) = \{\omega < 0, \nu < 0, \zeta > 0, \zeta^{-1}|\omega|^\alpha|\nu|^\beta > \alpha^\alpha\beta^\beta\}, \\
& & \Lambda_\gamma(\xi),\ 0 < \xi < \alpha^\alpha\beta^\beta,\ \text{is the positive root of the equation}
\end{array}$$

$$(1+\alpha x)^\alpha (1+\beta x)^\beta = \xi x.$$

The structure of the barrier Φ is as follows:

- the function
$$-\ln(w^\alpha v^\beta - z) - \ln w - \ln v$$
can be proved to be a 3-self-concordant barrier for the set
$$\{(w,v,z) \in \mathbf{R}^3 \mid w \geq 0, v \geq 0, w^\alpha v^\beta \geq z\}$$
so that the terms in F coming from $\phi(w_{1,i},v_{1,i},z_{1,i}) + \phi(w_{2,i},v_{2,i},z_{2,i})$ penalize the constraints
$$w_i^\alpha v^\beta \geq |b_i - a_i^T u|, \quad w_i \geq 0,\ v \geq 0$$
in $\widehat{P}$;
- the remaining part of Φ penalizes in already known to us way the constraints
$$\sum_{i=1}^m w_i \leq v,\ v \leq 5m\mathrm{V}(P),\ \| u \|_2 \leq 1.$$

16.4.5.4. **Starting point $\widehat{x}$:**

$$\widehat{x} = (\widehat{u} = 0; \widehat{v} = \frac{5}{2}m\mathrm{V}(P); \widehat{w}_i = 2\mathrm{V}(P),\ i = 1,...,m) \quad [[G:\widehat{x}] \leq 10m]$$

16.4.5.5. **Newton complexity of ϵ-solution:**

$$\mathcal{N}(\epsilon) = O(1)\sqrt{m}\ln\left(2m\frac{1+\nu(P,\epsilon)}{\nu(P,\epsilon)}\right), \quad \nu(P,\epsilon) = \frac{\epsilon}{\mathrm{V}(P)}.$$

16.4.5.6. **Arithmetic complexity of ϵ-solution:**

$$\mathcal{C}(\epsilon) = O(1)m^{1/2}(m+n)n^2\ln\left(2m\frac{1+\nu(P,\epsilon)}{\nu(P,\epsilon)}\right).$$

Bibliography

[1] L. Blum, M. Shub and S. Smale, *On the theory of computation and complexity over the real numbers: NP-completeness, recursive functions and universal machines*, Bulletin (New Series) of the American Mathematical Society **21** (1989), 1–46.

[2] A. Cobham *The intrinsic computational difficulty of functions.* in: Y. Bar-Hillel, Ed. *Logic, Methodology and Philosophy of Science*, Proc. Intern. Congress, North-Holland, Amsterdam, 1965, pp. 24–30.

[3] J. Edmonds *Paths, trees and flowers.* Canadian Journal of Mathematics **17** (1965), 449–467.

[4] C. Gonzaga, C *Polynomial time algorithm for linear programming* in: N. Megiddo, Ed. *Progress in Mathematical Programming: Interior Point and related methods*, Springer-Verlag, 1989.

[5] M. Grötshel, L. Lovasz and A. Shrijver *The Ellipsoid Method and Combinatorial Optimization*, Springer, 1986.

[6] Yu. Nesterov and A. Nemirovski *Interior point polynomial methods in Convex Programming: theory and applications*, SIAM Series in Applied Mathematics, SIAM, 1994.

[7] N. Karmarkar, *A new polynomial-time algorithm for linear programming*, Combinatorica **4** (1984), 373–395.

[8] Yu. Nesterov and A. Nemirovski *Multi-parameter surfaces of analytic centers and long-step surface-following interior point methods*, Research Report # 3/95, Optimization Laboratory, Faculty of Industrial Engineering & Management, Technion - Israel Institute of Technology, April 1995.

[9] J. Renegar *A polynomial time algorithm, based on Newton's method, for linear programming*, Mathematical Programming **40** (1988), 55–93.