TECHNION - THE ISRAEL INSTITUTE OF TECHNOLOGY
FACULTY OF INDUSTRIAL ENGINEERING & MANAGEMENT

INFORMATION-BASED COMPLEXITY

OF

CONVEX PROGRAMMING

A. Nemirovski

Fall Semester 1994/95

Information-Based Complexity of Convex Programming

Goals: given a class of Convex Optimization problems, one may look for an *efficient* algorithm for the class, i.e., an algorithm with a "good" (best possible, polynomial time,...) theoretical worst-case efficiency estimate on the class. The goal of the course is to present a number of efficient algorithms for several standard classes of Convex Optimization problems.

The course deals with the black-box setting of an optimization problem (all known in advance is that the problem belongs to a given "wide" class, say, is convex, convex of a given degree of smoothness, etc.; besides this a priory qualitative information, we have the possibility to ask an "oracle" for quantitative local information on the objective and the constraints, like their values and derivatives at a point). We present results on the associated with this setting *complexity* of standard problem classes (i.e. the best possible worst-case # of oracle calls which allows to solve any problem from the class to a given accuracy) and focus on the corresponding *optimal* algorithms.

Duration: one semester

Prerequisites: knowledge of elementary Calculus, Linear Algebra and of the basic concepts of Convex Analysis (like convexity of functions/sets and the notion of subgradient of a convex function) is welcomed, although is not absolutely necessary.

Contents:
Introduction: problem complexity and method efficiency in optimization
Methods with linear dimension-dependent convergence

from bisection to the cutting plane scheme
how to divide a n-dimensional pie: the Center-of-Gravity method
the Outer Ellipsoid method
polynomial solvability of Linear Programming
the Inner Ellipsoid method
convex-concave games and variational inequalities with monotone operators

Large-scale problems and methods with dimension-independent convergence

subradient and mirror descent methods for nonsmooth convex optimization
optimal methods for smooth convex minimization
strongly convex unconstrained problems

How to solve a linear system: optimal iterative methods for unconstrained convex quadratic minimization

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

Lecture 1

Introduction: what the course is about

What we are interested in the course are theoretically efficient methods for convex optimization problems. Almost each word in the previous sentence should be explained, and this explanation, that is, formulation of our goals, is the main thing I am going to speak about today. I believe that the best way to explain what we are about to do is to start with a simple example - one-dimensional convex minimization - where everything is seen.

1.1 Example: one-dimensional convex problems

Consider one-dimensional convex problems

$$\textit{minimize} \quad f(x) \quad \textit{s.t.} \quad x \in G = [a, b],$$

where $[a, b]$ is a given finite segment on the axis. It is also known that our objective f is a continuous convex function on G; for the sake of simplicity, assume that we know bounds, let them be 0 and V, for the values of the objective on G. Thus, all we know about the objective is that it belongs to the family

$$\mathcal{P} = \{f : [a, b] \to \mathbf{R} \mid f \textit{ is convex and continuous}; \ 0 \le f(x) \le V, x \in [a, b]\}.$$

And what we are asked to do is to find, for a given positive ε, an ε-solution to the problem, i.e., a point $\bar{x} \in G$ such that

$$f(\bar{x}) - f^* \equiv f(\bar{x}) - \min_G f \le \varepsilon.$$

Of course, our a priori knowledge on the objective given by the inclusion $f \in \mathcal{P}$, is, for small ε, far from being sufficient for finding an ε-solution, and we need some source of quantitative information on the objective. The standard assumption here which comes from the optimization practice is that we can compute the value and a subgradient of the objective at a point, i.e., we have access to a subroutine, *an oracle* $\mathcal{O}$, which gets, as an input, a point x from our segment and returns the value $f(x)$ and a subgradient $f'(x)$ of the objective at the point.
We have subject the input to the subroutine to the restriction $a < x < b$, since the objective, generally speaking, is not defined outside the segment $[a, b]$, and its subgradient might be undefined at the endpoints of the segment as well. I should also add that the oracle is not uniquely

defined by the above description; indeed, at some points f may have a "massive" set of subgradients, not a single one, and we did not specify how the oracle at such a point chooses the subgradient to be reported. We need exactly one hypothesis of this type, namely, we assume the oracle to be *local*: the information on f reported at a point x must be uniquely defined by the behaviour of f in a neighbourhood of x:

$$\{f, \bar{f} \in \mathcal{P}, x \in \text{int } G,\ f \equiv \bar{f} \ \textit{ in a neighbourhood of } x\ \} \Rightarrow \mathcal{O}(f,x) = \mathcal{O}(\bar{f},x).$$

What we should do is to find a method which, given on input the desired value of accuracy ε, after a number of oracle calls produces an ε-solution to the problem. And what we are interested in is the most efficient method of this type. Namely, given a method which solves every problem from our family to the desired accuracy in finite number of oracle calls, let us define the worst-case complexity N of the method as the maximum, over all problems from the family, of the number of calls; what we are looking for is exactly the method of the minimal worst-case complexity. Thus, the question we are interested in is

> *Given*
> *- the family $\mathcal{P}$ of objectives f,*
> *- a possibility to compute values and subgradients of f at a point of (a,b),*
> *- desired accuracy ε,*
> *what is the minimal #, $\text{Compl}(\varepsilon)$, of computations of f and f' which is sufficient, for all $f \in \mathcal{P}$, to form an ε-minimizer of f? What is the corresponding - i.e., the optimal - minimization method?*

Of course, to answer the question we should first specify the notion of a method. This is very simple task. Indeed, let us think what a method, let it be called $\mathcal{M}$, could be. It should perform sequential calls for the oracle, at i-th step forwarding to it certain input $x_i \in (a,b)$, let us call this input *i-th search point*. The very first input x_1 is generated by the method when the method has no specific information on the particular objective f the method is applied to; thus, the first search point should be objective-independent:

$$x_1 = S_1^{\mathcal{M}}. \tag{1.1.1}$$

Now, the second search point is generated after the method knows the value and a subgradient of the objective at the first search point, and x_2 should be certain function of this information:

$$x_2 = S_2^{\mathcal{M}}(f(x_1), f'(x_1)). \tag{1.1.2}$$

Similarly, i-th search point is generated by the method when it already knows the values and the subgradients of f at the previous search points, and this is all the method knows about f so far, so that i-th search point should be certain function of the values and the subgradients of the objective at the previous search points:

$$x_i = S_i^{\mathcal{M}}(f(x_1), f'(x_1); ...; f(x_{i-1}), f'(x_{i-1})). \tag{1.1.3}$$

We conclude that the calls to the oracle are defined by certain recurrence of the type (1.1.3); the rules governing this recurrence, i.e., the functions $S_i^{\mathcal{M}}(\cdot)$, are specific for the method and form a part of its description.

I have said that the *search rules* $S_i^{\mathcal{M}}$ form only a part of the description of the method; indeed, the method should sometime stop and form the result. The moment of termination, same as the result found by the method, also should depend only on information accumulated to the corresponding moment; we may assume that there is a sequence of termination tests - functions

$$T_i^{\mathcal{M}}(\cdot) \in \{\text{STOP}, \text{CONTINUE}\} \tag{1.1.4}$$

taking values in the indicated two-element set, and the method terminates at the very first moment i when it turns out that

$$T_i^{\mathcal{M}}(f(x_1), f'(x_1); ...; f(x_i), f'(x_i)) = \text{STOP}.$$

At this moment the method forms the result of its work:

$$\bar{x}(\mathcal{M}, f) = R_i^{\mathcal{M}}(f(x_1), f'(x_1); ...; f(x_i), f'(x_i)); \tag{1.1.5}$$

the termination tests $T_i^{\mathcal{M}}$, same as the rules $R_i^{\mathcal{M}}$ for forming the result, also are a part of the description of the method.

Given the search rules, the termination tests and the rules for forming the result, we can completely define the behaviour of the method on every problem from our family, and we may identify the method with the collection of these rules. Thus, by definition

> *a method is a collection of the search rules S_i, the termination tests T_i and the rules R_i for forming the result*

In order to get the widest possible definition, we do not subject the rules comprising a method to any further restrictions like "computability in finitely many arithmetic operations"; the rules might be arbitrary functions of the information on the problem accumulated to the step when the rule should be used.

Now, the number of steps performed by a method $\mathcal{M}$ as applied to a problem f from our family is called the *complexity* $\text{Compl}(\mathcal{M}, f)$ of the method at f (this is a positive integer or $+\infty$ depending on whether the method stops at f or works at this problem forever). The complexity of the method on the whole family of problems is, by definition, the maximum of the complexity at a problem f over all f:

$$\text{Compl}(\mathcal{M}) = \max_{f \in \mathcal{P}} \text{Compl}(\mathcal{M}, f).$$

This is simply the worst-case # of oracle calls made by the method.

Similarly, we define the *accuracy* (it would be better to say inaccuracy) of the method $\mathcal{M}$ at a problem f as

$$\text{Accur}(\mathcal{M}, f) = f(\bar{x}(\mathcal{M}, f)) - \min_G f,$$

i.e., as the residual, in the values of the objective, of the result obtained by the method as applied to f, and define the accuracy of the method at the whole family looking at the worst case:

$$\text{Accur}(\mathcal{M}) = \max_{f \in \mathcal{P}} \text{Accur}(\mathcal{M}, f).$$

With our now terminology, the problem of finding the optimal method is

given the family

$$\mathcal{P} = \{f : G = [a,b] \to \mathbf{R} \mid \ f \text{ is convex and continuous on } G \ \ , 0 \le f \le V\}$$

of problems and an $\varepsilon > 0$, find among the methods $\mathcal{M}$ with the accuracy on the family not worse than ε that one with the smallest possible complexity on the family.

The complexity of the associated - optimal - method, i.e., the function

$$\mathrm{Compl}(\varepsilon) = \min\{\mathrm{Compl}(\mathcal{M}) \mid \ \mathrm{Accur}(\mathcal{M}) \le \varepsilon\}$$

is called the *complexity* of the family. Thus, the question we are interested in is to find the complexity of the family $\mathcal{P}$ of one-dimensional convex minimization problems along with the associated optimal method.

The answer to the question is given by the following statement.

Theorem 1.1.1 *The complexity of the family in question satisfies the inequalities*

$$\frac{1}{5}\log_2(\frac{V}{\varepsilon}) - 1 < \mathrm{Compl}(\varepsilon) \le \rfloor \log_2(\frac{V}{\varepsilon})\lfloor, \ 0 < \varepsilon < V. \tag{1.1.6}$$

The method associated with the upper bound (and thus optimal in complexity, up to an absolute constant factor), is the usual bisection terminated after $N = \rfloor \log_2(V/\varepsilon)\lfloor$ steps.

Note that the range of values of ε in our statement is $(0, V)$, and this is quite natural: since all functions from the family take their values between 0 and V, any point of the segment solves every problem from the family to the accuracy V, so that a nontrivial optimization problem occurs only when $\varepsilon < V$.

Now let us prove the theorem.

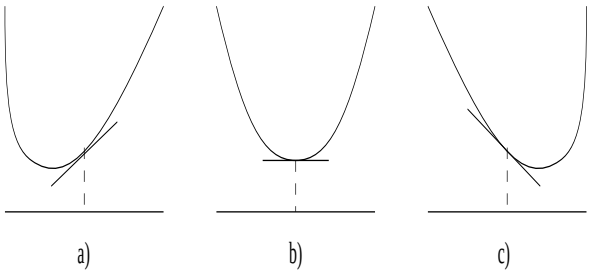
10. *Upper bound.* I hope all of you remember what is the bisection method. In order to minimize f, we start with the midpoint of our segment, i.e., with

$$x_1 = \frac{a+b}{2},$$

and ask the oracle about the value and a subgradient of the objective at the point. If the subgradient is zero, we are done - we have found an optimal solution. If the subgradient is positive, then the function, due to convexity, to the right of x_1 is greater than at the point, and we may cut off the right half of our initial segment - the minimum for sure is localized in the remaining part. If the subgradient is negative, then we may cut off the left half of the initial segment: Thus, we either terminate with an optimal solution, or find a new segment, twice smaller than the initial domain, which for sure localizes the set of optimal solutions. In this latter case we repeat the procedure, with the initial domain replaced by the new localizer, and so on. After we have performed the number of steps indicated in the formulation of the theorem we terminate and form the result as the best - with the minimal value of f - of the search points we have looked through:

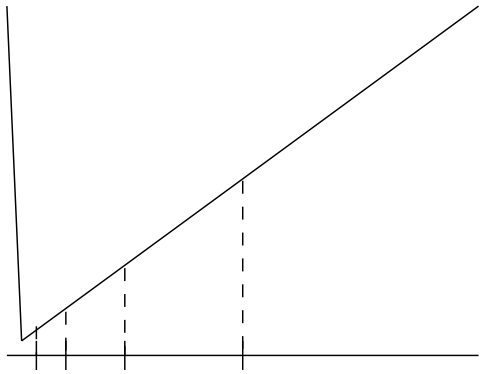
$$\bar{x} \in \{x_1, ..., x_N\}; \ f(\bar{x}) = \min_{1 \le i \le N} f(x_i).$$

Let me note that traditionally the approximate solution given by the bisection method is identified with the last search point (which is clearly at at the distance at most $(b-a)2^{-N}$ from



The bisection step. a) $f'(x_1) > 0$; b) $f'(x_1) = 0$; c) $f'(x_1) < 0$
[the dashed segment is the new localizer of the optimal set]

the optimal solution), rather than with the best point found so far. This traditional choice has small in common with our accuracy measure (we are interested in small values of the objective rather than in closeness to optimal solution) and is simply dangerous, as you can see from the following example:



Here during the first $N-1$ steps everything looks as if we were minimizing $f(x) = x$, so that the N-th search point is $x_N = 2^{-N}$; our experience is misleading, as you see from the picture, and the relative accuracy of x_N as an approximate solution to the problem is very bad, something like $V/2$.

By the way, we see from this example that the evident convergence of the search points to the optimal set at the rate at least 2^{-i} does not imply automatically certain fixed rate of convergence in terms of the objective; it turns out, anyhow, that such a rate exists, but for the best points found so far rather than for the search points themselves.

Let me present you an extremely simple and rather instructive proof of the rate of convergence in terms of the objective.

Let us start with the observation that if G_N is the final localizer of optimum found during the bisection, then outside the localizer the value of the objective is at least that one at the best of the search points, i.e., at least the value at the approximate solution $\bar{x}$ found by bisection:

$$f(x) \geq f(\bar{x}) \equiv \min_{1 \leq i \leq N} f(x_i), \ x \in G \backslash G_N.$$

Indeed, at a step of the method we cut off only those points of the domain G where f is at least as large as at the current search point and is therefore $\geq$ its value at the best of the search points, that is, to the value at $\bar{x}$; this is exactly what was claimed.

Now, let x^* be an optimal solution to the problem, i.e., a minimizer of f; as we know, such a minimizer does exist, since our continuous objective for sure attains its minimum over the finite segment G. Let α be a real greater than 2^{-N} and less than 1, and consider α-contraction of the segment G to x^*, i.e., the set

$$G^\alpha = (1-\alpha)x^* + \alpha G \equiv \{(1-\alpha)x^* + \alpha z \mid z \in G\}.$$

This is a segment of the length $\alpha(b-a)$, and due to our choice of α the length is greater than that one of our final localizer G_N. It follows that G^α cannot be inside the localizer, so that there is a point, let it be y, which does not belong to the final localizer and belongs to G^α:

$$\exists y \in G^\alpha : y \notin \text{int } G_N.$$

Since y belongs to G^α, we have

$$y = (1-\alpha)x^* + \alpha z$$

for some $z \in G$, and from convexity of f it follows that

$$f(y) \le (1-\alpha)f(x^*) + \alpha f(z),$$

which can be rewritten as

$$f(y) - f^* \le \alpha\,(f(z) - f^*) \le \alpha V,$$

so that y is an αV-solution to the problem. On the other hand, y is outside G_N, where, as we already know, f is at least $f(\bar{x})$:

$$f(\bar{x}) \le f(y).$$

We conclude that

$$f(\bar{x}) - f^* \le f(y) - f^* \le \alpha V.$$

Since α can be arbitrary close to 2^{-N}, we come to

$$f(\bar{x}) - f^* \le 2^{-N}V = 2^{-\rfloor \log_2(V/\varepsilon)\lfloor}V \le \varepsilon.$$

Thus,

$$\text{Accur}(\textit{Bisection}_N) \le \varepsilon.$$

The upper bound is justified.

2^0. *Lower bound.* At the first glance, it is not clear where from could one get a lower bound for the complexity of the family. Indeed, to find an upper bound means to invent a method and to evaluate its behaviour on the family - this is what people in optimization are doing all their life. In contrast to this, to find a lower bound means to say something definite about all methods from an extremely wide class, not to investigate a particular method. Nevertheless, as we shall see in a while, the task is quite tractable.

To simplify explanation, consider the case when the domain G of our problems is the segment $[-1,1]$ and the objectives vary from 0 to 1, i.e., let V be equal to 1; due to similarity reasons, with these assumptions we do not loose generality. Thus, given an $\varepsilon \in (0,1)$ we should prove that the complexity of our family for this value of ε is greater than

$$\frac{1}{5}\log_2(\frac{1}{\varepsilon}) - 1.$$

In other words, we should prove that if $\mathcal{M}$ is a method which solves all problems from the family to the accuracy ε, then the complexity K' of the method on the family, i.e., the worst case number of steps is at least the aforementioned quantity. Let $\mathcal{M}$ solve every problem to the accuracy ε in no more than K' steps. By adding redundant steps, we may assume that $\mathcal{M}$ performs exactly

$$K = K' + 1$$

steps at every problem and that the result of $\mathcal{M}$ as applied to a problem is nothing but the last, K-th search point.

Now let us prove the following basic lemma:

Lemma 1.1.1 *For any i, $0 \le i \le K$, our family contains an objective f_i with the following two properties:*

(1_i) : there is a segment $\Delta_i \subset G = [-1, 1]$ - the active segment of f_i - of the length

$$\delta_i = 2^{1-2i}$$

where f_i is a modulus-like function, namely,

$$f_i(x) = a_i + 2^{-3i}|x - c_i|, \ x \in \Delta_i,$$

c_i being the midpoint of Δ_i;

(2_i) : The first i points $x_1, ..., x_i$ generated by $\mathcal{M}$ as applied to f_i, are outside Δ_i.

Proof will be given by induction on i.

Base $i = 0$ is immediate: we set

$$f_0(x) = |x|, \quad \Delta_0 = G = [-1, 1],$$

thus ensuring (1_0). Property (2_0) holds true by trivial reasons - when $i = 0$, then there are no search points to be looked at.

Step $i \Rightarrow i+1$: Let f_i be the objective given by our inductive hypothesis, let Δ_i be the active segment of this objective and let c_i be the midpoint of the segment.

Let also $x_1, ..., x_i, x_{i+1}$ be the first $i + 1$ search points generated by $\mathcal{M}$ as applied to f_i. According to our inductive hypothesis, the first i of these points are outside the active segment.

In order to obtain f_{i+1}, we modify the function f_i in its active segment and do not vary the function outside the segment. The way we modify f_i in the active segment depends on whether x_{i+1} is to the right of the midpoint c_i of the segment ("right" modification), or this is not the case and x_{i+1} either coincides with c_i or is to the left of the point ("left" modification).

The "right" modification is as follows: we replace the modulus-like in its active segment function f_i by a piecewise linear function with three linear pieces, as is shown on Figure 1.1. Namely, we do not change the slope of the function in the initial 1/14 part of the segment, then change the slope from -2^{-3i} to $-2^{-3(i+1)}$ and make a new breakpoint at the end c_{i+1} of the first quarter of the segment Δ_i. Starting with this breakpoint and till the right endpoint of the active segment, the slope of the modified function is $2^{-3(i+1)}$. It is easily seen that the modified function at the right endpoint of Δ_i comes to the same value as that one of f_i and that the modified function is convex on the whole axis.

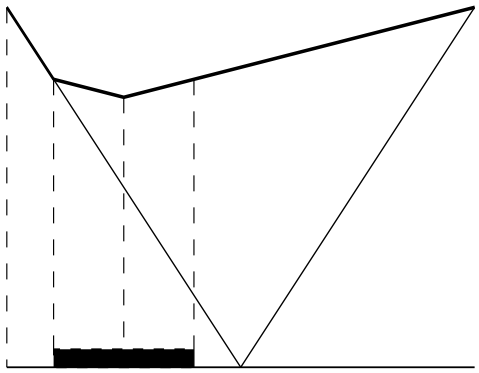


Figure 1.1: "Right" modification of f_i. The slopes for the *right* modification, from left to right, are $-2^{-3i}, -2^{-3(i+1)}, 2^{-3(i+1)}$, the breakpoints u and c_{i+1} are at the distances $|\Delta_i|/14$, $|\Delta_i|/4$ from the left endpoint of the active segment Δ_i of f_i. The bold segment on the axis is the active segment of the modified function (right modification).

In the case of the "left" modification, i.e., when $x_{i+1} \leq c_i$, we act in the "symmetric" manner, so that the breakpoints of the modified function are at the distances $\frac{3}{4}|\Delta_i|$ and $\frac{13}{14}|\Delta_i|$ from the left endpoint of Δ_i, and the slopes of the function, from left to right, are $-2^{-3(i+1)}$, $2^{-3(i+1)}$ and 2^{-3i}.

Let us verify that the modified function f_{i+1} satisfies the requirements imposed by the lemma. As we have mentioned, this is a convex continuous function; since we do not vary f_i outside the segment Δ_i and do not decrease it inside the segment, the modified function takes its values in $(0, 1)$ together with f_i. It suffices to verify that f_{i+1} satisfies (1_{i+1}) and (2_{i+1}).

(1_{i+1}) is evident by construction: the modified function indeed is modulus-like with required slopes in a segment of a required length. What should be proved is (2_{i+1}), the claim that the method $\mathcal{M}$ as applied to f_{i+1} during the first $i+1$ step does not visit the active segment of f_{i+1}. To prove this, it suffices to prove that the first $i+1$ search points generated by the method as applied to f_{i+1} are exactly the search point generated by it when minimizing f_i, i.e., they are the points $x_1, ..., x_{i+1}$. Indeed, these latter points for sure are outside the new active segment - the first i of them due to the fact that they even do not belong to the larger segment Δ_i, and the last point, x_{i+1} - by our construction, which ensures that the active segment of the modified function and x_{i+1} are separated by the midpoint c_i of the segment Δ_i.

Thus, we come to the necessity to prove that $x_1, ..., x_{i+1}$ are the first $i+1$ points generated by $\mathcal{M}$ as applied to f_{i+1}. This is evident: the points $x_1, ..., x_i$ are outside Δ_i, where f_i and f_{i+1} coincide; consequently, the information - the values and the subgradients - on the functions along the sequence $x_1, ..., x_i$ also is the same for both of the functions. Now, by definition of a method the information accumulated by it during the first i steps uniquely determines the first $i+1$ search points; since f_i and f_{i+1} are indistinguishable in a neighbourhood of the first i search points generated by $\mathcal{M}$ as applied to f_i, the initial $(i+1)$-point segments of the trajectories of $\mathcal{M}$ on f_i and on f_{i+1} coincide with each other, as claimed.

Thus, we have justified the inductive step and therefore have proved the lemma.

It remains to derive from the lemma the desired lower complexity bound. This is immediate.

According to the lemma, there exists function f_K in our family which is modulus-like in its active segment and is such that the method during its first K steps does not visit this active segment. But the K-th point x_K of the trajectory of $\mathcal{M}$ on f_K is exactly the result found by the method as applied to the function; since f_K is modulus-like in Δ_K and is convex everywhere, it attains its minimum f_K^* at the midpoint c_K of the segment Δ_K and outside Δ_K is greater than $f_K^* + 2^{-3K} \times 2^{-2k} = 2^{-5k}$ (the product here is half of the length of Δ_K times the slope of f_K). Thus,

$$f_K(\bar{x}(\mathcal{M}, f_K)) - f_K^* > 2^{-5K}.$$

On the other hand, $\mathcal{M}$, by its origin, solves all problems from the family to the accuracy ε, and we come to

$$2^{-5K} < \varepsilon,$$

i.e., to

$$K \equiv K' + 1 < \frac{1}{5}\log_2(\frac{1}{\varepsilon}).$$

as required in our lower complexity bound. ∎

1.2 Conclusion

The one-dimensional situation we have investigated is, of course, very simple; I spoke about it only to give you an impression of what we are going to do. In the main body of the course we shall consider much more general classes of convex optimization problems - multidimensional problems with functional constraints. Same as in our simple one-dimensional example, we shall ask ourselves what is the complexity of the classes and what are the corresponding optimal methods. Let me stress that these are optimal methods we mainly shall focus on - it is much more interesting issue than the complexity itself, both from mathematical and practical viewpoint. In this respect, one-dimensional situation is not typical - it is easy to guess that the bisection should be optimal and to establish its rate of convergence. In several dimensions situation is far from being so trivial and is incomparably more interesting.

1.3 Exercises: Brunn, Minkowski and convex pie

1.3.1 Prerequisites

Prerequisites for the exercises of this section are:

the Separation Theorem for convex sets:

> *if G is a convex subset of $\mathbf{R}^n$ with a nonempty interior and $x \in \mathbf{R}^n$ does not belong to the interior of G, then x can be separated from G by a linear functional, i.e., there exists a nonzero $e \in \mathbf{R}^n$ such that*
>
> $$e^T x \geq e^T y \quad \forall y \in G.$$

The Caratheodory Theorem. To formulate this theorem, let us start with the fundamental notion of *an extreme point* of a convex set $Q \subset \mathbf{R}^n$. Namely, a point $x \in G$ is called an extreme

point of G, if it cannot be represented as the midpoint of a nontrivial segment belonging to G, i.e., if the relation

$$x = \frac{1}{2}x' + \frac{1}{2}x''$$

with x' and x'' belonging to G is possible only if $x' = x'' = x$. For example, the extreme points of the segment $[0,1] \subset \mathbf{R}$ are 0 and 1; the extreme points of a triangle in the plane are the vertices of the triangle; the extreme points of a circle in a plane are the points belonging to the boundary circumference.

The Caratheodory Theorem is as follows:

> *Let G be a*
> *(a) nonempty,*
> *(b) closed,*
> *(c) bounded*
> *and*
> *(d) convex*
> *subset of $\mathbf{R}^n$. Then the set G_{ext} of extreme points of G is nonempty; moreover, every point $x \in G$ can be represented as a convex combination of at most $n+1$ extreme points of G: there exist $x_1, ..., x_{n+1} \in G_{\text{ext}}$ and nonnegative $\lambda_1, ..., \lambda_{n+1}$ with the unit sum such that*
>
> $$x = \sum_{i=1}^{n+1} \lambda_i x_i.$$

Corollary:

> *Let $x_1, ..., x_m$ be certain points of $\mathbf{R}^n$. Assume that a point $x \in \mathbf{R}^n$ can be represented as a convex combination of $x_1, ..., x_m$. Then x can be also represented as a convex combination of at most $n+1$ points from the set $\{x_1, ..., x_m\}$.*

Indeed, the *convex hull* G of $\{x_1, ..., x_m\}$, i.e., the set of all convex combinations of $x_1, ..., x_m$, is a nonempty, closed and bounded convex subset of $\mathbf{R}^n$ (why?), and the extreme points of this convex hull clearly belong to $\{x_1, ..., x_m\}$ (why?), and it suffices to apply to G the Caratheodory Theorem.

1.3.2 Brunn, Minkowski and Convex Pie

Consider the following game with two players - you and me. I am cooking a pie Q, which should be a closed and bounded convex domain in the space $\mathbf{R}^n$ (the word "domain" means that the interior of Q is nonempty). After the pie is cooked, I show it to you and offer to divide it in the following way: you choose a point $x \in \mathbf{R}^n$; given this point, I perform a *cut* passing through the point, i.e., choose a half-space Π with the boundary hyperplane passing through x:

$$\Pi = \{y \in \mathbf{R}^n \mid e^T y > e^T x\}$$

(e is a nonzero vector) and take, as my part, the intersection of Q with the half-space, and you take the remaining part of the pie, i.e., the set

$$Q_e = \{y \in Q \mid e^T y \leq e^T x\},$$

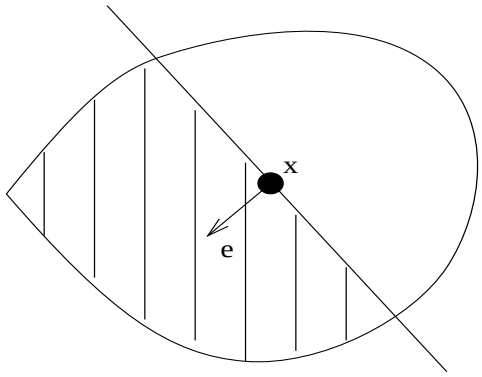


Figure 1.2: Dividing a pie: you choose x, I choose Π and take the dashed part of the pie

see Figure 1.2. How should you choose the point x in order to get a "substantial part" of the pie independently of how I have cooked the pie and perform the cut through the point chosen by you, i.e., in order to ensure "not too small" ratio

$$q_e = \mathrm{Vol}_n(Q_e)/\operatorname*{Vol}_n(Q)$$

(Vol_n stands for n-dimensional volume).

Some observations may be done immediately.

Exercise 1.3.1 $^{\#}$ *Prove that if $x \notin \mathrm{int}\, Q$, then I can give you "almost nothing", i.e., I can choose e resulting in $\mathrm{Vol}_n(Q_e) = 0$.*

Recall that a simplex in $\mathbf{R}^n$ is the convex hull of $n+1$ points (*vertices of the simplex*) which are affinely independent (the smallest affine subspace containing all the vertices is the whole $\mathbf{R}^n$)).

Exercise 1.3.2 $^{\#*}$ *Prove that if I have cooked a simplex Q, then, independently of how you choose x, I can leave you no more than a_n-th part of the pie, $a_n = (n/(n+1))^n \geq \exp\{-1\}$, i.e., given x, I can choose e in such a way that*

$$\mathrm{Vol}_n(Q_e) \leq a_n \mathrm{Vol}_n(Q).$$

Prove that if you choose x not as the barycenter (the arithmetic mean of the vertices) of the simplex, then I can make the above inequality strict.

A surprising fact is that the upper bound a_n on the worst-case value of your part is sharp:

> *given a closed and bounded convex domain $Q \subset \mathbf{R}^n$, you can choose $x \in \mathrm{int}\, Q$ in a way which ensures that*
>
> $$\mathrm{Vol}_n\left(\{y \in Q \mid e^T y \leq e^T x\}\right) \geq a_n \operatorname*{Vol}_n(Q) \tag{1.3.1}$$
>
> *for any nonzero e, where*
>
> $$a_n = \left(\frac{n}{n+1}\right)^n \geq \exp\{-1\}.$$

In order to ensure this inequality, it suffices to choose x as the center of gravity of Q, i.e., as

$$x = x^*(Q) = \mathrm{Vol}_n^{-1}(Q) \int_Q x dx.$$

This statement in its general (n-dimensional) form was first proved by B. Grunbaum, and we will refer to it as to the Grunbaum Theorem.

We are about to prove the Grunbaum Theorem. To this end let us start with a simple characterization of the center of gravity. Given a closed and bounded convex set $G \in \mathbf{R}^n$ and an affine functional $f(x)$ on $\mathbf{R}^n$, one can define the *momentum* of G with respect to f as

$$I_f(G) = \int_G f(x) dx;$$

in mechanical terms, $I_f(G)$ is the momentum of the unform mass distribution on G with respect to the hyperplane $\{f(x) = 0\}$.

Exercise 1.3.3 # *Prove that the center of gravity $x^*(Q)$ of a closed and bounded convex domain Q is uniquely defined by the relation*

$$I_f(Q) = f(x^*(Q))\mathrm{Vol}_n(Q)$$

for all affine functionals f.

In mechanical terms the statement of the latter exercise can be formulated as follows: the momentum of the uniform mass distribution on G with respect to any hyperplane is the same as if all the mass of G would be located at the center of gravity; in particular, the momentum of this distribution with respect to any hyperplane passing through the center of gravity is 0.

Exercise 1.3.4 #* *Prove that the center of gravity of a closed and bounded convex domain $Q \subset \mathbf{R}^n$ is an interior point of Q. Show that the center of gravity is affine invariant: if $x \mapsto \mathcal{A}(x) = Ax + b$ is an invertible affine mapping of $\mathbf{R}^n$ onto itself and $Q^+ = \mathcal{A}(Q)$ is the image of Q under this mapping, then*

$$x^*(Q^+) = \mathcal{A}(x^*(Q))$$

Exercise 1.3.5 #* *Prove that the center of gravity of an n-dimensional simplex is the barycenter of the simplex, i.e., the arithmetic mean of the $n+1$ vertices of the simplex. Is it true that the center of gravity of a polytope is the arithmetic mean of the vertices?*

The statement given by the latter exercise can be extended as follows. Let Π be an affine hyperplane in $\mathbf{R}^n$, let B be a closed and bounded convex set in Π with a nonempty $(n-1)$-dimensional interior, and let a be a point not belonging to the hyperplane. Consider the convex hull of a and B; this clearly is a closed and bounded convex domain in $\mathbf{R}^n$. We shall say that Q is a *conic set* with the base B and the vertex a, see Figure 1.3. E.g., a simplex is a conic set; one can take any facet of the simplex as its base and the vertex not belonging to the base as vertex of this conic set.

What can be said about the center of gravity of a conic set?

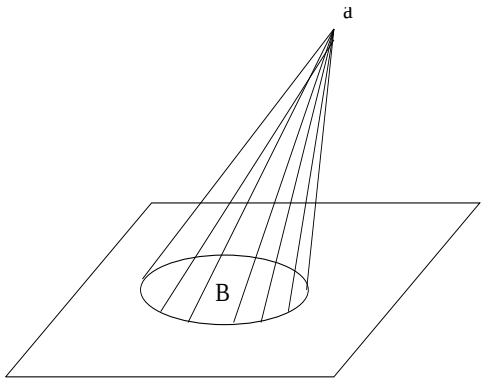


Figure 1.3: A conic set: B-base, a - vertex.

Exercise 1.3.6 $^{\#+}$ *Let Q be a conic set with base B and vertex a. The center of gravity of Q belongs to the segment Δ linking the vertex a and the center b of gravity of the base (regarded as a closed convex domain in $(n-1)$-dimensional space) and divides Δ in the ratio $n:1$:*

$$x^*(Q) = a/(n+1) + nb/(n+1).$$

To proceed with Grunbaum's theorem, we need general and extremely important theorem called the *Brunn-Minkowski Symmetrization Principle*.

Let Q be a closed and bounded convex domain in $\mathbf{R}^n$, and let l be a straight line in $\mathbf{R}^n$. Given Q, one can construct another set, Q^l, which is

(a) symmetric with respect to l (i.e., its cross-section by an affine hyperplane orthogonal to l either is empty, or is a $(n-1)$-dimensional Euclidean disk centered at the point where the hyperplane intersects l), and

(b) has the same $(n-1)$-volumes of cross-sections with orthogonal to l affine hyperplanes as those for Q, i.e., an affine hyperplane Γ which is orthogonal to l intersects Q if and only if it intersects Q^l, the $(n-1)$-dimensional volumes of the cross-sections of Q and Q^l by Γ being equal to each other.

Of course, properties (a) and (b) uniquely define Q^l; this set is called *symmetrization of Q with respect to the axis l*.

Now, the Brunn-Minkowski Symmetrization Principle claims that

the symmetrization of a closed convex domain Q with respect to an axis also is a closed convex domain.

Now let us derive the Grunbaum Theorem from the Symmetrization Principle. Let Q be a closed convex set, x^* be the center of gravity of Q (your choice in the problem of dividing the pie Q recommended by Grunbaum's theorem), and let e be a nonzero vector (my choice). Let $l = x^* + \mathbf{R}e$ be chosen as the symmetrization axis, and let Q^l be the symmetrization of Q with respect to l. When dividing Q according to our choices, I take the part Q_+ of Q to the right of the hyperplane Γ passing through x^* and orthogonal to l, and you take the part Q_- of Q to the left of Γ (l is oriented by the vector e). Let Q^l_+ and Q^l_- be similar parts of the "symmetrized pie" Q^l, see Figure 1.4.

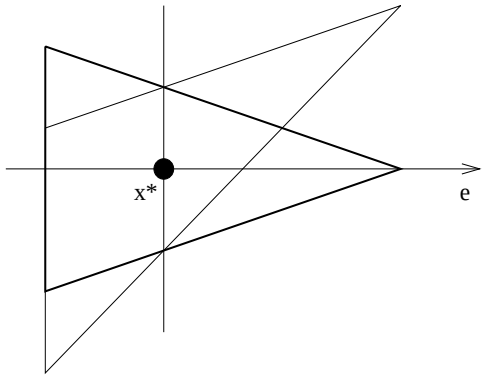


Figure 1.4: Partitioning of a pie and the induced partitioning of the symmetrized pie.

Exercise 1.3.7 # *Prove that x^* is the center of gravity of Q^l and that*

$$\mathrm{Vol}_n(Q_\pm) = \mathrm{Vol}_n(Q^l_\pm).$$

The latter exercise says that our parts in the pie Q given by the partitioning in question are the same as if we were dividing a symmetric with respect to certain axis l pie Q^l, x being chosen as the center of gravity of Q^l and e being the direction of the axis (let us call this *the special partitioning of* Q^l). Note that the symmetrized pie Q^l is convex - this crucial fact is given by the Symmetrization Principle. Thus, in order to establish the Grunbaum Theorem we may restrict ourselves with axial-symmetric closed and bounded convex domains and their special partitions. Of course, we are interested in the multi-dimensional case (since the Grunbaum theorem in the one-dimensional case is evident).

Let Q be an axial-symmetric closed convex and bounded convex domain in $\mathbf{R}^n$, $n > 1$. Without loss of generality we may assume that the symmetry axis of Q is the last coordinate axis and that the center of gravity of Q is at the origin. Denoting a point from $\mathbf{R}^n$ as $x = (u, t)$ (u is $(n-1)$-dimensional, t is scalar), we can represent Q as

$$Q = \{(u,t) \mid u \in \mathbf{R}^{n-1},\, t \in [-a, b],\, |u| \le \phi(t)\}, \tag{1.3.2}$$

where $a, b > 0$ and ϕ is a concave on $[a, b]$, positive on (a, b) (and, as it is easily seen, continuous) function on $[a, b]$.

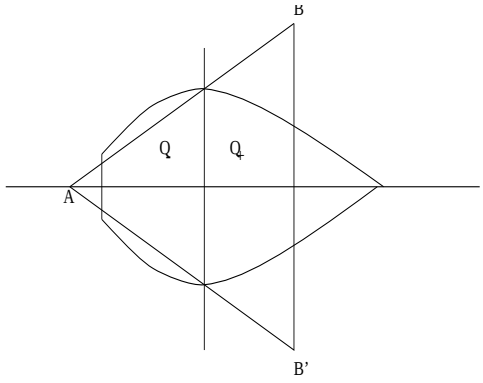
Exercise 1.3.8 # *Let Q be the given by (1.3.2) axial-symmetric closed and bounded convex domain with the center of gravity at the origin, and let Q_-, Q_+ be the parts of Q to the left and to the right of the hyperplane $\{t = 0\}$.*

1)* *Prove that the momentum $I_{-t}(Q_-)$ of Q_- with respect to the functional $(u, t) \mapsto -t$ is equal to the momentum $I_t(Q_+)$ of Q_+ with respect to the functional $(u, t) \mapsto t$.*

2)* *Prove that there exist positive α and β such that the cone*

$$Q^* = \{(u,t) \mid -\alpha \le t \le \beta,\, |u| \le \xi(t) \equiv \alpha^{-1}(t+\alpha)\phi(0)\}$$

(i.e., the cone with the same symmetry axis and the same cross-section by the hyperplane $\{t = 0\}$ as those of Q) has volumes of the parts Q^*_-, Q^*_+ to the left and to the right of the hyperplane



A and $Q^* = ABB'$

Figure 1.5: Cone Q^*: the volumes of the parts of the cone to the left and to the right of $\{t = 0\}$ are the same as the similar quantities for Q.

$\{t = 0\}$ *equal to the volumes of the similar parts* Q_-, Q_+ *of* Q*; in particular,*

$$\mathrm{Vol}_n(Q^*) = \mathrm{Vol}_n(Q).$$

3)* *Prove that the momentum* $I_{-t}(Q^*_-)$ *is not less than the momentum* $I_{-t}(Q_-)$*, while the momentum* $I_t(Q^*_+)$ *is not greater than the momentum* $I_t(Q_+)$*.*

4) *Conclude from 1) and 3) that* $I_{-t}(Q^*_-) \geq I_t(Q^*_+)$*.*

5) *Conclude from 4) that the center of gravity of the cone* Q^* *is to the left of the hyperplane* $\{t = 0\}$*.*

6)* *Conclude from 2) and 5) that*

$$\mathrm{Vol}_n(Q_-) = \mathrm{Vol}_n(Q^*_-) \geq (n/(n+1))^n \mathrm{Vol}_n(Q^*) = (n/(n+1))^n \mathrm{Vol}_n(Q),$$

so that the Grunbaum theorem is valid for special partitions of axial-symmetric closed and bounded convex domains and therefore for all partitions of all closed and bounded convex domains.

To understand the power of Grunbaum's theorem, let us think of whether the answer is "stable", i.e., what may happen if you choose a point not in the center of gravity. Consider the simple case when the pie is an n-dimensional Euclidean ball of the unit radius centered, say, at the origin. If you choose the origin, then, independently of the cut, you will get one half of the pie. Now, what happens when you choose a point x at a distance α, $0 < \alpha < 1$, from the origin, say, at the distance 0.5, and my cutting plane is orthogonal to x (and, of course, I take the larger part of the pie, i.e., that one which contains the origin)? The "spherical hat" you get "visually" looks large enough; I think you will be surprised that if n is large then you get "almost nothing", namely, exponentially small part of the pie.

Exercise 1.3.9 $^+$ *Prove that the volume of the "spherical hat"*

$$V_\alpha = \{x \in \mathbf{R}^n \mid |x| \leq 1, x_n \geq \alpha\}$$

($0 \le \alpha \le 1$) of the unit Euclidean ball $V = \{x \in \mathbf{R}^n \mid |x| \le 1\}$ satisfies the inequality as follows:

$$\mathrm{Vol}_n(V_\alpha)/\mathrm{Vol}_n(V) \le \kappa \exp\{-n\alpha^2/2\},$$

$\kappa > 0$ being a positive absolute constant. What is, numerically, the left hand side ratio when $\alpha = 1/2$ and $n = 64$?

The Grunbaum theorem says that the center of gravity of a closed and bounded convex domain Q is, in certain sense, a "far enough from the boundary" interior point of the domain. There are other statements of this type, e.g., the following one:

> Chord Partitioning Property:
> *let Q be a closed and bounded convex domain in $\mathbf{R}^n$, let $x^*(Q)$ be the center of gravity of Q and let δ be a chord of Q (i.e., a segment with the endpoints on the boundary of Q) passing through $x^*(Q)$. Then the ratio in which the center of gravity divides δ is at most $n : 1$ (i.e., the larger part is at most n times the smaller one).*

Exercise 1.3.10 * *Prove the Chord Partitioning Property.*

The Grunbaum theorem says that if you have a unit mass uniformly distributed over a closed and bounded convex domain, then there exists a point such that any half-space containing the point is of "a reasonable", not too small, mass. What can be said about an arbitrary distributed unit mass? The answer is given by another theorem of Grunbaum which is as follows.

> *let μ be a probability distribution of mass on $\mathbf{R}^n$. Then there exists a point x such that, for any closed half-space Π containing x, the mass of Π is at least $1/(n+1)$.*

It should be specified what is a "probability distribution of mass". Those familiar with the measure theory should think about a Borel probability measure on $\mathbf{R}^n$. A not so educated reader may think about a finite set of points assigned nonnegative masses with the sum of the masses equal to 1; the mass of an arbitrary subset A of $\mathbf{R}^n$ is simply the sum of masses of the points belonging to A.

Note that the constant $1/(n+1)$ is sharp, as it is seen from the example when μ is concentrated at vertices of a simplex, with the mass of each of the vertices equal to $1/(n+1)$.

The second theorem of Grunbaum, to the best of my knowledge, has no applications in Optimization, in contrast to the first one. I am citing it here because it gives me the possibility to introduce to you a very important theorem of Helley (same as the first theorem of Grunbaum gave us chance to become acquainted with the fundamental Symmetrization Principle). The Helley theorem is as follows:

> *let $\{Q_\alpha\}_{\alpha \in A}$ be a family of closed convex subsets of $\mathbf{R}^n$ such that any $n+1$ set from the family have a nonempty intersection. Assume, in addition, that either the family is finite, or the intersection of certain finite number of sets from the family is bounded. Then all the sets have a point in common.*

Exercise 1.3.11 * *Derive the second theorem of Grunbaum from the Helley Theorem.*

The actual challenge is, of course, to prove the Helley Theorem and the Symmetrization Principle, and this is the goal of the remaining exercises. We start with the Helley theorem (as with the simpler one).

Exercise 1.3.12 $^+$ *Prove the following fundamental fact:*

if $f_1(x), ..., f_m(x)$ are convex functions defined in a convex neighbourhood of a point $x_0 \in \mathbf{R}^n$ and

$$f(x) = \max_{i=1,...,m} f_i(x),$$

then the subgradient set $\partial f(x_0)$ can be found as follows: let us take those f_i which are active at x_0, i.e., $f_i(x_0) = f(x_0)$, and form the union U of the subgradient sets of these f_i at x_0; $\partial f(x_0)$ is exactly the convex hull of U.

Conclude that if f attains its minimum at x_0, then one can select no more than $n+1$ functions from the family $f_1, ..., f_m$ in such a way that the selected functions are active at x_0 and their maximum, same as f, attains its minimum at x_0.

Exercise 1.3.13 * *Let $\mathcal{F} = \{Q_1, ..., Q_k\}$ be a finite family of closed and bounded convex sets satisfying the premise of the Helley theorem. Consider the function*

$$r(x) = \max_{1 \le i \le k} \mathrm{dist}(x, Q_i),$$

where

$$\mathrm{dist}(x, Q) = \min_{y \in Q} |x - y|$$

is the distance from x to Q. Prove that $r(x)$ attains its minimum at certain x^* and verify that x^* belongs to all Q_i, $i = 1, ..., k$. Thus, the Helley theorem is valid for finite families of bounded sets.

Exercise 1.3.14 $^+$ *Conclude from the statement of the latter exercise that the Helley theorem is valid for finite families.*

Exercise 1.3.15 $^+$. *Derive from the statement of the latter exercise the complete Helley theorem.*

The Helley Theorem has a lot of applications in Convex Analysis. Let me present you the first which comes to my mind.

Let f be a continuous real-valued function defined on the finite segment $[a, b]$ of the real axis ($a < b$). There is a well-known problem of the finding the best polynomial approximation of f in the uniform norm:

given a positive integer n, find a polynomial p of the degree $< n$ which minimizes

$$\| f - p \|_\infty \equiv \max_{t \in [a,b]} |f(t) - p(t)|$$

over all polynomials of the degree $< n$.

The characterization of the solution is given by the famous Tschebyshev theorem:

a polynomial p of a degree $< n$ is the solution if and only if there exist $n+1$ points $t_0 < ... < t_n$ on the segment such that

$$\| f - p \|_\infty = |f(t_i) - p(t_i)|, i = 0, ..., n,$$

and neighbouring differences $f(t_i) - p(t_i)$ have opposite signs.

In fact this answer consists of two parts:

(1) the claim that there exists a $n+1$-point subset T of $[a,b]$ which "concentrates" the whole difficulties of uniform approximating f, i.e., is such that

$$\min_{p:\deg p<n} \| f-p \|_\infty = \min_{p:\deg p<n} \| f-p \|_{T,\infty}$$

(here $\| h \|_{T,\infty} = \max_{t\in T} |h(t)|$),
and

(2) characterization of the best uniform approximation of a function on a $(n+1)$-point set by a polynomial of a degree $< n$.

(2) heavily depends on the algebraic nature of polynomials; in contrast to this, (1) is a completely general fact.

Exercise 1.3.16 $^+$ *Prove the following generalization of (1):*

Let X be a compact set and f, $\phi_0, ..., \phi_{n-1}$ be continuous mappings from X into an m-dimensional linear space E provided with a norm $\|\cdot\|$. Consider the problem of finding the best uniform approximation of f by a linear combination of ϕ_i, i.e., of minimizing

$$\| f-p \|_{X,\infty} = \max_{t\in X} \| f(t)-p(t) \|$$

over p belonging to the linear span Φ of the functions ϕ_i, and let δ be the optimal value in the problem.

Prove that there exists a finite subset T in X of cardinality at most $n+1$ such that

$$\delta = \min_{p\in\Phi} \| f-p \|_{T,\infty} .$$

Note that (1) is the particular case of this general statement associated with $E=\mathbf{R}$, $X=[a,b]$ and $\phi_i(t)=t^i$.

Now let us come to the Symmetrization Principle. Thus, let Q be a closed and bounded convex domain in n dimensions; without loss of generality we may assume that the axis mentioned in the principle is the last coordinate axis. It is convenient to denote the coordinate along this axis by t and the vector comprised of the first $n-1$ coordinates by x. Let $[a,b]$ be the image of Q under the projection onto the axis, and let $Q_t = \{x \mid (x,t)\in Q\}$, $a\le t\le b$, be the inverse image of a point $(0,...,0,t)$ under the projection. The Symmetrization Principle claims exactly that the function

$$\phi(t) = \{\mathrm{Vol}_{n-1}(Q_t)\}^{1/(n-1)}$$

is concave and continuous on $[a,b]$ (this function is, up to a multiplicative constant, exactly the radius of the $(n-1)$-dimensional disk of the same as Q_t $(n-1)$-dimensional volume).

It is quite straightforward to derive from the convexity and closedness of Q that ϕ is continuous; let us ignore this easy and not interesting part of the job (I hope you can do it without my comments). The essence of the matter is, of course, to prove the concavity of ϕ. To this end let us reformulate the task as follows.

Given two compact subsets H, H' in $\mathbf{R}^k$, one can form their "linear combination" with real coefficients

$$\lambda H + \lambda' H' = \{x = \lambda x + \lambda' x' \mid x\in H, x'\in H'\},$$

same as the "product of H by a real λ":

$$\lambda H = \{x = \lambda x \mid x \in H\}.$$

The results of these operations are, of course, again compact subsets in $\mathbf{R}^k$ (as images of the compact sets $H \times H'$, H under continuous mappings). It is clear that the introduced operations share the following property of their arithmetic predecessors:

$$\lambda H + \lambda' H' = (\lambda H) + (\lambda' H'). \tag{1.3.3}$$

Now, given a compact set $H \subset \mathbf{R}^k$, one can define its "average radius" $r_k(H)$ as

$$r_k(H) = (\mathrm{Vol}_k(H))^{1/k}$$

(I use Vol_k as a synonym for the Lebesque measure; those not familiar with this notion, might think of convex polytopes H and use the Riemann definition of volume). Note that $r_k(H)$ is, up to a coefficient depending on k only, nothing but the radius of the Euclidean ball of the same volume as that one of H.

It is clear that

$$r_k(\lambda H) = \lambda r_k(H), \lambda \geq 0. \tag{1.3.4}$$

Now, our function $\phi(t)$ is nothing but $r_{n-1}(Q_t)$. To prove that the function is concave means to prove that, for any given $\lambda \in [0,1]$ and all $t', t'' \in [a,b]$ one has

$$r_{n-1}(Q_{\lambda t' + (1-\lambda)t''}) \geq \lambda r_{n-1}(Q_{t'}) + (1-\lambda) r_{n-1}(Q_{t''}). \tag{1.3.5}$$

Now, what do we know about $Q_{\lambda t' + (1-\lambda)t''}$ is that this set contains the set $\lambda Q_{t'} + (1-\lambda) Q_{t''}$; this is an immediate consequence of the convexity of Q. Therefore the left hand side in (1.3.5) is

$$r_{n-1}(Q_{\lambda t' + (1-\lambda)t''}) \geq r_{n-1}(Q' + Q''),\ Q' = \lambda Q_{t'},\ Q'' = (1-\lambda) Q_{t''}$$

(we have used (1.3.3) and evident monotonicity of $r_k(\cdot)$ with respect to inclusion), while the left hand side of (1.3.5) is nothing but $r_{n-1}(Q') + r_{n-1}(Q'')$ (look at (1.3.4)). Thus, to prove the concavity of ϕ, i.e., (1.3.5), it suffices to prove, for all k, the following statement:

(SA_k): *the function $r_k(H)$ is a super-additive function on the family of k-dimensional compact sets, i.e.,*

$$r_k(H + H') \geq r_k(H) + r_k(H'). \tag{1.3.6}$$

To prove the concavity of ϕ, it suffices to establish this statement only for convex compact sets, but in fact it is true independently of the convexity assumptions.

The goal of the remaining exercises is to give an inductive in k proof of (SA_k). Unfortunately, our scheme requires a little portion of the measure theory. Those familiar with this theory can omit the intuitive explanation of the required technique I start with.

Let $H \subset \mathbf{R}^k$, $k > 1$, be a closed and bounded set.

As above, let (x,t) be the coordinates in the space (x is $(k-1)$-dimensional, t is scalar), T be the projection of H onto the t-axis and H^t be the $(k-1)$-dimensional inverse image of the

point $(0, ..., 0, t)$ under this projection. Then the k-dimensional volume of the set "of course" can be computed as follows:

$$\mathrm{Vol}_k(H) = \int_T \mathrm{Vol}_{k-1}(H^t)dt = \int_T r_{k-1}^{k-1}(H^t)dt.$$

(the first equality, for the Lebesque integration theory, is the Fubini theorem; similar statement exists in the Riemann theory of multidimensional integration, up to the fact that H should be subject to additional restrictions which make all objects well-defined).

Now, for $l \geq 1$ the integral

$$J = \int_T f^l(t)dt$$

involving a bounded nonnegative function f defined on a bounded subset T of the axis can be computed according to the following reasoning. Consider the flat region given by

$$R = \{(t, s) \in \mathbf{R}^2 \mid t \in T, 0 \leq s \leq f(t)\}$$

and let

$$I = \int_R s^{l-1}dtds;$$

integrating first in s and then in t, we come to

$$I = l^{-1}J.$$

Integrating first in t and then in s, we come to

$$I = \int_0^\infty s^{l-1}\mu(s)ds, \quad \mu(s) = \mathrm{Vol}_1\{t \in T \mid f(t) \leq s\}$$

($\mu(s)$ is the measure of the set of those $t \in T$ for which $(t, s) \in R$). We took ∞ as the upper integration limit in order to avoid new notation; in fact $\mu(s)$ is identically zero for all large enough s, since R is bounded.

Thus, we come to the equality

$$\int_T f^l(t)dt = l\int_0^\infty s^{l-1}\mu(s)ds,\ \mu(s) = \mathrm{Vol}_1\{t \in T \mid f(t) \geq s\} \tag{1.3.7}$$

Of course, to make this reasoning rigorous, we need some additional assumptions, depending on what is the "power" of the integration theory we use. Those who know the Lebesque integration theory should realize that it for sure is sufficient to assume T and f to be measurable and bounded; in the case we are interested in (i.e., when T is the projection onto an axis of a closed and bounded set $H \in \mathbf{R}^k$ and $f(t) = r_{k-1}(H_t)$) these assumptions are satisfied automatically (since T is a compact set and $r_{k-1}(t)$, as it immediately follows from the compactness of H, is upper semicontinuous). Those who use a less general integration theory, are advised to trust in the fact that there exists a theory of integration which makes the above "evident" conclusions rigorous.

With (1.3.7) in mind, we are enough equipped to prove (SA_k) by induction in k. Let us start with the base:

Exercise 1.3.17 * *Prove (SA_1).*

The main effort, of course, is to make the inductive step.

Exercise 1.3.18 $^{*>}$ *Given (SA_{k-1}) for some $k > 1$ and (SA_1), prove (SA_k) and thus prove that (SA_k) is valid for all k, which completes the proof of the symmetrization principle.*

Lecture 2

Methods with linear convergence, I

In the introductory lecture we spoke about the complexity and optimal algorithm of minimizing a convex function of one variable over a segment. From now on we shall study general type multidimensional convex problems.

2.1 Class of general convex problems: description and complexity

A convex programming problem is

$$\textit{minimize} \;\; f(x) \;\; \textit{subject to} \;\; g_i(x) \leq 0,\, i = 1, ..., m,\; x \in G \subset \mathbf{R}^n. \tag{2.1.1}$$

Here the *domain* G of the problem is a closed convex set in $\mathbf{R}^n$ with a nonempty interior, the *objective* f and the *functional constraints* g_i, $i = 1, ..., m$, are convex continuous functions on G. Some more related terminology: any point $x \in G$ which satisfies the system

$$g_i(x) \leq 0,\, i = 1, ..., m$$

of functional constraints is called a *feasible solution* to the problem, and the problem is called *feasible*, if it admits feasible solutions. The set of all feasible solutions is called the *feasible set* of the problem.

The lower bound

$$f^* = \inf\{f(x) \mid x \in G,\, g_i(x) \leq 0,\, i = 1, ..., m\}$$

of the objective over the set of all feasible solutions is called the *optimal value*. For an infeasible problem, *by definition*, the optimal value is $+\infty$; this fits the standard convection that infimum of something over an empty set is $= \infty$ and sup of something over an empty set is $-\infty$. For a feasible problem, the optimal value is either a finite real (if the objective is below bounded on the set of feasible solutions), or $-\infty$.

A problem is called *solvable*, if it is feasible and the optimal value of the problem is attained at certain feasible solution; every feasible solution with this latter property is called an *optimal solution* to the problem, and the set of all optimal solutions in what follows will be denoted by X^*. The optimal value of a solvable problem is, of course, finite.

Following the same scheme as in the introductory lecture, let us fix a closed and *bounded* convex domain $G \subset \mathbf{R}^n$ and the number m of functional constraints, and let $\mathcal{P} = \mathcal{P}_m(G)$ be the family of all *feasible* convex problems with m functional constraints and the domain G. Note that since the domain G is bounded and all problems from the family are feasible, all of them are solvable, due to the standard compactness reasons.

In what follows we identify a problem instance from the family $\mathcal{P}_m(G)$ with a vector-valued function

$$p = (f, g_1, ..., g_m)$$

comprised of the objective and the functional constraints.

What we shall be interested in for a long time are the efficient methods for solving problems from the indicated very wide family. Similarly to the one-dimensional case, we assume that the methods have an access to the first order local oracle $\mathcal{O}$ which, given an input vector $x \in \text{int } G$, returns the values and some subgradients of the objective and the functional constraints at x, so that the oracle computes the mapping

$$x \mapsto \mathcal{O}(p, x) = (f(x)(, f'(x); g_1(x), g_1'(x); ...; g_m(x), g_m'(x)) : \text{int } G \to \mathbf{R}^{(m+1)\times(n+1)}.$$

The notions of a method and its complexity at a problem instance and on the whole family are introduced exactly as it was done in the one-dimensional case, as a set of rules for forming sequential search points, the moment of termination and the result as functions of the information on the problem; this information is comprised by the answers of the oracle obtained to the moment when a rule is to be applied.

The accuracy of a method at a problem and on the family is defined in the way similar to that one used in the one-dimensional case; the only question here is what to do with the functional constraints. A convenient way is to start with the *vector of residuals* of a point $x \in G$ regarded as an approximate solution to a problem instance p:

$$\text{Residual}(p, x) = (f(x) - f^*, (g_1(x))_+, ..., (g_m(x))_+)$$

which is comprised of the inaccuracy in the objective and the violations of functional constraints at x. In order to get a convenient scalar accuracy measure, it is reasonable to pass from this vector to the *relative accuracy*

$$\varepsilon(p, x) = \max\{\frac{f(x) - f^*}{\max_G f - f^*}, \frac{(g_1(x))_+}{(\max_G g_1)_+}, ..., \frac{(g_m(x))_+}{(\max_G g_m)_+}\};$$

to get the relative accuracy, we normalize each of the components of the vector of residuals by its maximal, over all $x \in G$, value and take the maximum of the resulting quantities. It is clear that the relative accuracy takes its values in $[0, 1]$ and is zero if and only if x is an optimal solution to p, as it should be for a reasonable accuracy measure.

After we have agreed how to measure accuracy of tentative approximate solutions, we define the accuracy of a method $\mathcal{M}$ at a problem instance as the accuracy of the approximate solution found by the method when applied to the instance:

$$\text{Accur}(\mathcal{M}, p) = \varepsilon(p, \bar{x}(p, \mathcal{M})).$$

The accuracy of the method on the family is its worse-case accuracy at the problems of the family:

$$\mathrm{Accur}(\mathcal{M}) = \sup_{p\in\mathcal{P}_m(G)} \mathrm{Accur}(\mathcal{M}, p).$$

Last, the complexity of the family is defined in the manner we already are acquainted with, namely, as the best complexity of a method solving all problems from the family to a given accuracy:

$$\mathrm{Compl}(\varepsilon) = \min\{\mathrm{Compl}(\mathcal{M}) \mid \mathrm{Accur}(\mathcal{M}) \le \varepsilon\}.$$

What we are about to do is to establish the following main result:

Theorem 2.1.1 *The complexity* $\mathrm{Compl}(\varepsilon)$ *of the family* $\mathcal{P}_m(G)$ *of general-type convex problems on an* m*-dimensional closed and bounded convex domain* G *satisfies the inequalities*

$$n\lfloor\frac{\ln(\frac{1}{\varepsilon})}{6\ln 2}\rfloor - 1 \le \mathrm{Compl}(\varepsilon) \le \lfloor 2.181\, n\, \ln(\frac{1}{\varepsilon})\rfloor. \tag{2.1.2}$$

Here the upper bound is valid for all $\varepsilon < 1$*. The lower bound is valid for all* $\varepsilon < \varepsilon(G)$*, where*

$$\varepsilon(G) \ge \frac{1}{n^3}$$

for all $G \subset \mathbf{R}^n$*; for an ellipsoid* G *one has*

$$\varepsilon(G) = \frac{1}{n},$$

and for a parallelotope G

$$\varepsilon(G) = 1.$$

Same as in the one-dimensional case, to prove the theorem means to establish the lower complexity bound and to present a method associated with the upper complexity bound (and thus optimal in complexity, up to an absolute constant factor, for small enough ε, namely, for $0 < \varepsilon < \varepsilon(G)$. We shall start with this latter task, i.e., with constructing an optimal method.

2.2 Cutting Plane scheme and Center of Gravity Method

The method we are about to present is based on a very natural extension of the bisection - the *cutting plane scheme*.

2.2.1 Case of problems without functional constraints

To explain the cutting plane scheme, let me start with the case when there are no functional constraints at all, so that $m = 0$ and the family is comprised of problems

$$\textit{minimize} \quad f(x) \quad \textit{s.t.} \quad x \in G \tag{2.2.1}$$

of minimizing convex continuous objectives over a given closed and bounded convex domain $G \subset \mathbf{R}^n$.

To solve such a problem, we can use the same basic idea as in the one-dimensional bisection. Namely, choosing somehow the first search point x_1, we get from the oracle the value $f(x_1)$ and a subgradient $f'(x_1)$ of f; thus, we obtain a linear function

$$f_1(x) = f(x_1) + (x - x_1)^T f'(x_1)$$

which, due to the convexity of f, underestimates f everywhere on G and coincides with f at x_1:

$$f_1(x) \le f(x),\ x \in G; \quad f_1(x_1) = f(x_1).$$

If the subgradient is zero, we are done - x_1 is an optimal solution. Otherwise we can point out a proper part of G which localizes the optimal set of the problem, namely, the set

$$G_1 = \{x \in G \mid (x - x_1)^T f'(x_1) \le 0\};$$

indeed, outside this new localizer our linear lower bound f_1 for the objective, and therefore the objective itself, is greater than the value of the objective at x_1.

Now, our new localizer of the optimal set, i.e., G_1, is, same as G, a closed and bounded convex domain, and we may iterate the process by choosing the second search point x_2 inside G_1 and generating the next localizer

$$G_2 = \{x \in G_1 \mid (x - x_2)^T f'(x_2) \le 0\},$$

and so on. We come to the following generic *cutting plane scheme*:

> *starting with the localizer $G_0 \equiv G$, choose x_i in the interior of the current localizer G_{i-1} and check whether $f'(x_i) = 0$; if it is the case, terminate, x_i being the result, otherwise define the new localizer*
>
> $$G_i = \{x \in G_{i-1} \mid (x - x_i)^T f'(x_i) \le 0\}$$
>
> *(see Figure 2.1) and loop.*
>
> *The approximate solution found after i steps of the routine is, by definition, the best point found so far, i.e., the point*
>
> $$\bar{x}_i \in \{x_1, ..., x_i\} \ \text{ such that } \ f(\bar{x}_i) = \min_{1 \le j \le i} f(x_j).$$

A cutting plane method, i.e., a method associated with the scheme, is governed by the rules for choosing the sequential search points in the localizers. In the one-dimensional case there is, basically, the only natural possibility for this choice - the midpoint of the current localizer (the localizer always is a segment). This choice results exactly in the bisection and enforces the lengths of the localizers to go to 0 at the rate 2^{-i}, i being the step number. In the multidimensional case the situation is not so simple. Of course, we would like to decrease a reasonably defined *size* of localizer at the highest possible rate; the problem is, anyhow, which size to choose and how to ensure its decreasing. When choosing a size, we should take care of two things

> (1) *we should have a possibility to conclude that if the size of a current localizer G_i is small, then the inaccuracy of the current approximate solution also is small;*
>
> (2) *we should be able to decrease at certain rate the size of sequential localizers by appropriate choice of the search points in the localizers.*

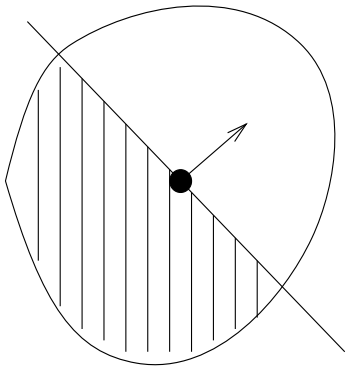


Figure 2.1: A step of a cutting plane method. Dashed region is the new localizer.

Let us start with a wide enough family of sizes which satisfy the first of our requirements.

Definition 2.2.1 *A real-valued function* $\mathrm{Size}(Q)$ *defined on the family* $\mathcal{Q}$ *of all closed and bounded convex subsets* $Q \subset \mathbf{R}^n$ *with a nonempty interior is called a size, if it possesses the following properties:*

(Size.1) Positivity: $\mathrm{Size}(Q) > 0$ *for any* $Q \in \mathcal{Q}$*;*

(Size.2) Monotonicity with respect to inclusion: $\mathrm{Size}(Q) \geq \mathrm{Size}(Q')$ *whenever* $Q' \subset Q$, Q, $Q' \in \mathcal{Q}$*;*

(Size.3) Homogeneity with respect to homotheties: *if* $Q \in \mathcal{Q}$, $\lambda > 0$, $a \in \mathbf{R}^n$ *and*

$$Q' = a + \lambda(Q - a) = \{a + \lambda(x - a) \mid x \in Q\}$$

is the image of Q *under the homothety with the center at the point* a *and the coefficient* λ*, then*

$$\mathrm{Size}(Q') = \lambda\,\mathrm{Size}(Q).$$

Example 1. The diameter

$$\mathrm{Diam}(Q) = \max\{|x - x'| \mid x, x' \in Q\}$$

is a size;

Example 2. The *average diameter*

$$\mathrm{AvDiam}(Q) = (\mathrm{Vol}_n(Q))^{1/n},$$

Vol_n being the usual n-dimensional volume, is a size.
To the moment these examples are sufficient for us.

Let us prove that any size of the indicated type satisfies requirement (1), i.e., if the size of a localizer is small, then the problem is "almost" solved.

Lemma 2.2.1 *Let* $\mathrm{Size}(\cdot)$ *be a size. Assume that we are solving a convex problem (2.2.1) by a cutting plane method, and let* G_i *and* $\bar{x}_i$ *be the localizers and approximate solutions generated by the method. Then we have for all* $i \geq 1$

$$\varepsilon(p, \bar{x}_i) \leq \varepsilon_i \equiv \frac{\mathrm{Size}(G_i)}{\mathrm{Size}(G)}, \tag{2.2.2}$$

p denoting the problem in question.

Proof looks completely similar to that one used for the bisection method. Indeed, let us fix $i \geq 1$. If $\varepsilon_i \geq 1$, then (2.2.2) is evident - recall that the relative accuracy always is ≤ 1. Now assume that $\varepsilon_i < 1$ for our i. Let us choose $\alpha \in (\varepsilon_i, 1]$, let x^* be a minimizer of our objective f over G and let

$$G^\alpha = x^* + \alpha(G - x^*).$$

Then

$$\mathrm{Size}(G^\alpha) = \alpha\,\mathrm{Size}(G) > \varepsilon_i\,\mathrm{Size}(G) = \mathrm{Size}(G_i)$$

(we have used the homogeneity of $\mathrm{Size}(\cdot)$ with respect to homotheties). Thus, the size of G^α is greater than that on of G_i, and therefore, due to the monotonicity of Size, G^α cannot be a subset of G_i. In other words, there exists a point

$$y \in G^\alpha \backslash G_i.$$

Since G^α clearly is contained in the domain of the problem and does not belong to the i-th localizer G_i, we have

$$f(y) > f(\bar{x}_i);$$

indeed, at each step j, $j \leq i$, of the method we remove from the previous localizer (which initially is the whole domain G of the problem) only those points where the objective is greater than at the current search point x_j and is therefore greater than at the best point $\bar{x}_i$ found during the first i steps; since y was removed at one of these steps, we conclude that $f(y) > f(\bar{x}_i)$, as claimed.

On the other hand, $y \in G^\alpha$, so that

$$y = (1-\alpha)x^* + \alpha z$$

with some $z \in G$. From convexity of f it follows that

$$f(y) \leq (1-\alpha)f(x^*) + \alpha f(z) = (1-\alpha)\min_G f + \alpha \max_G f,$$

whence

$$f(y) - \min_G f \leq \alpha(\max_G f - \min_G f).$$

As we know, $f(y) > f(\bar{x}_i)$, and we come to

$$f(x_i) - \min_G f \leq \alpha(\max_G f - \min_G f),$$

or, which is exactly the same, to

$$\varepsilon(p, \bar{x}_i) < \alpha.$$

Since α is an arbitrary real $> \varepsilon_i$, we conclude that (2.2.2) holds. ∎

Thus, we realize now what could be the sizes we are interested in, and the problem is how to ensure certain rate of their decreasing along the sequence of localizers generated by a cutting plane method. The difficulty here is that when choosing the next search point in the current localizer, we do not know what will be the next cutting plane; the only thing we know is that it will pass through the search point. Thus, we are interested in the choice of the search point which guarantees certain reasonable, not too close to 1, ratio of the size of the new localizer to

that one of the previous localizer independently of what will be the cutting plane. Whether such a choice of the search point is possible, it depends on the size we are using. For example, the diameter of a localizer, which is a very natural measure of it and which was successively used in the one-dimensional case, would be a very bad choice in the multidimensional case. To realize this, imagine that we are minimizing over the unit square on the two-dimensional plane, and our objective in fact depends on the first coordinate only. All our cutting planes (in our example they are lines) will be parallel to the second coordinate axis, and the localizers will be stripes of certain horizontal size (which we may enforce to tend to 0) and of some fixed vertical size (equal to 1). The diameters of the localizers here although decrease but do not tend to zero. Thus, the first of the particular sizes we have looked at does not fit the second requirement. In contrast to this, the second particular size, the *average diameter* AvDiam, is quite appropriate, due to the following geometric fact (which was presented in the problems accompanying the previous lecture):

Proposition 2.2.1 (Grunbaum) *Let Q be a closed and bounded convex domain in $\mathbf{R}^n$, let*

$$x^*(G) = \frac{1}{\mathrm{Vol}_n(G)} \int_G x dx$$

be the center of gravity of Q, and let Γ be an affine hyperplane passing through the center of gravity. Then the volumes of the parts Q', Q'' in which Q is partitioned by Γ satisfy the inequality

$$\mathrm{Vol}_n(Q'), \mathrm{Vol}_n(Q'') \le \{1 - \left(\frac{n}{n+1}\right)^{1/n}\}\mathrm{Vol}_n(Q) \le \exp\{-\kappa\}\mathrm{Vol}_n(Q),$$

$$\kappa = \ln(1 - 1/\mathrm{e}) = 0.45867...;$$

in other words,

$$\mathrm{AvDiam}(Q'), \mathrm{AvDiam}(Q'') \le \exp\{-\frac{\kappa}{n}\}\,\mathrm{AvDiam}(Q). \tag{2.2.3}$$

Note that the proposition states exactly that the smallest (in terms of the volume) fraction you can cut off a n-dimensional convex body by a hyperplane passing through the center of gravity of the body is the fraction you get when the body is a simplex, the plane passes parallel to a facet of the simplex and you cut off the part not containing the facet.

Corollary 2.2.1 *Consider the Center of Gravity method, i.e., the cutting plane method with the search points being the centers of gravity of the corresponding localizers:*

$$x_i = x^*(G_{i-1}) \equiv \frac{1}{\mathrm{Vol}_n(G_{i-1})} \int_{G_{i-1}} x dx.$$

For the method in question one has

$$\mathrm{AvDiam}(G_i) \le \exp\{-\frac{\kappa}{n}\}\,\mathrm{AvDiam}(G_{i-1}),\ i \ge 1;$$

consequently (see Lemma 2.2.2) the relative accuracy of i-th approximate solution generated by the method as applied to any problem p of minimizing a convex objective over G satisfies the inequality

$$\varepsilon(p, \bar{x}_i) \le \exp\{-\frac{\kappa}{n} i\},\ i \ge 1.$$

In particular, to solve the problem within relative accuracy $\varepsilon \in (0,1)$ it suffices to perform no more than

$$N = \rfloor \frac{1}{\kappa} n \ln\left(\frac{1}{\varepsilon}\right) \lfloor \leq \rfloor 2.181 n \ln\left(\frac{1}{\varepsilon}\right) \lfloor \tag{2.2.4}$$

steps of the method.

Remark 2.2.1 The Center of Gravity method for convex problems without functional constraints was invented in 1965 independently by A.Yu.Levin in Russia and J. Newman in the USA.

2.3 The general case: problems with functional constraints

The Center of Gravity method, as it was presented, results in the upper complexity bound stated in our Main Theorem, but only for problems without functional constraints. In order to establish the upper bound in full generality, we should modify the cutting plane scheme in a manner which enables us to deal with these constraints. The very first idea is to act as follows: after current search point is chosen and we have received the values and subgradients of the objective and the constraints at the point, let us check whether the point is feasible; if it is the case, then let us use the subgradient of the objective in order to cut off the part of the current localizer where the objective is greater that at the current search point, same as it was done in the case when there were no functional constraints. Now, if there is a functional constraint violated at the point, we can use its subgradient to cut off points where the constraint is for sure greater than it is at the search point; the removed points for sure are not feasible.

This straightforward approach cannot be used as it is, since the feasible set may have empty interior, and in this case our process, normally, will never find a feasible search point and, consequently, will never look at the objective. In this case the localizers will shrink to the feasible set of the problem, and this is fine, but if the set is not a singleton, we would not have any idea how to extract from our sequence of search points a point where the constraints are "almost satisfied" and at the same time the objective is close to the optimal value - recall that we simply did not look at the objective when solving the problem!

There is, anyhow, a simple way to overcome the difficulty - we should use for the cut the subgradient of the objective at the steps when the constraints are "almost satisfied" at the current search point, not only at the steps when the point is feasible. Namely, consider the method as follows:

> Cutting plane scheme for problems with functional constraints:
>
> *Given in advance the desired relative accuracy $\varepsilon \in (0,1)$, generate, starting with $G_0 = G$, the sequence of localizers G_i, as follows:*
>
> *given G_{i-1} (which is a closed and bounded convex subset of G with a nonempty interior), act as follows:*
>
> *1) choose i-th search point*
>
> $$x_i \in \text{int } G_{i-1}$$
>
> *and ask the oracle about the values*
>
> $$f(x_i), g_1(x_i), ..., g_m(x_i)$$

and subgradients

$$f'(x_i), g_1'(x_i), ..., g_m'(x_i)$$

of the objective and the constraints at x_i.

2) form the affine lower bounds

$$g_j^{(i)}(x) = g_j(x_i) + (x - x_i)^T g_j'(x_i)$$

for the functional constraints (these actually are lower bounds since the constraints are convex) and compute the quantities

$$g_j^{i,*} = \left(\max_G g_j^{(i)}(x) \right)_+, \ i = 1, ..., m.$$

3) If, for all $j = 1, ..., m$,

$$g_j(x_i) \leq \varepsilon g_j^{*,i}, \tag{2.3.1}$$

claim that i *is a productive step and go to 4.a), otherwise go to 4.b).*

4.a) [productive step]

If $f'(x_i) \neq 0$, *define the new localizer* G_i *as*

$$G_i = \{x \in G_{i-1} \mid (x - x_i)^T f'(x_i) \leq 0\}$$

and go to 5); otherwise terminate, x_i *being the result formed by the method.*

4.b) [non-productive step]

Choose j *such that*

$$g_j(x_i) > \varepsilon g_j^{*,i},$$

define the new localizer G_i *as*

$$G_i = \{x \in G_{i-1} \mid (x - x_i)^T g_j'(x_i) \leq 0\}$$

and go to 5).

5) Define i*-th approximate solution as the best (with the smallest value of the objective) of the search points* x_j *corresponding to the productive steps* $j \leq i$. *Replace* i *by* $i + 1$ *and loop.*

Note that the approximate solution $\bar{x}_i$ is defined only if a productive step already has been performed.

The rate of convergence of the above method is given by the following analogue of Lemma 2.2.2:

Proposition 2.3.1 *Let* Size *be a size, and let a cutting plane method be applied to a convex problem* p *with* m *functional constraints. Assume that for a given* N *the method either terminates in course of the first* N *steps, or this is not the case and the relation*

$$\frac{\text{Size}(G_N)}{\text{Size}(G)} < \varepsilon \tag{2.3.2}$$

is satisfied. In the first case the result $\bar{x}$ *found by the method is a solution to the problem of relative accuracy* $\leq \varepsilon$; *in the second case the* N*-th approximate solution is well-defined and solves the problem to the relative accuracy* ε.

Proof. Let us first note that for any i and j one has

$$g_j^{*,i} \le \left(\max_G g_j\right)_+; \tag{2.3.3}$$

this is an immediate consequence of the fact that $g_j^{(i)}(x)$ is a lower bound for $g_j(x)$ (an immediate consequence of the convexity of g_j), so that the maximum of this lower bound over $x \in G$, i.e., $g_j^{*,i}$, is at most the similar quantity for the constraint g_j itself.

Now, assume that the method terminates at certain step $i \le N$. According to the description of the method, it means that i is a productive step and 0 is a subgradient of the objective at x_i; the latter means that x_i is a minimizer of f over the whole G, so that

$$f(x_i) \le f^*.$$

Besides this, i is a productive step, so that

$$g_j(x_i) \le \varepsilon g_j^{*,i} \le \varepsilon \left(\max_G g_j\right)_+, \ j = 1, ..., m$$

(we have used (2.3.3)); these inequalities, combined with the definition of the relative accuracy, state exactly that x_i (i.e., the result obtained by the method in the case in question) solves the problem within the relative accuracy ε, as claimed.

Now assume that the method does not terminate in course of the first N steps. In view of our premise, here we have

$$\mathrm{Size}(G_N) < \varepsilon\, \mathrm{Size}(G). \tag{2.3.4}$$

Let x^* be an optimal solution to the problem, and let

$$G^\varepsilon = x^* + \varepsilon(G - x^*).$$

G^ε is a closed and bounded convex subset of G with a nonempty interior; due to homogeneity of Size with respect to homotheties, we have

$$\mathrm{Size}(G^\varepsilon) = \varepsilon\, \mathrm{Size}(G) > \mathrm{Size}(G^\varepsilon)$$

(the second inequality here is (2.3.4)). From this inequality and the monotonicity of the size it follows that G^ε cannot be a subset of G_N:

$$\exists y \in G^\varepsilon \backslash G_N.$$

Now, y is a point of G (since the whole G^ε is contained in G), and since it does not belong to G_N, it was cut off at some step of the method, i.e., there is an $i \le N$ such that

$$e_i^T(y - x_i) > 0, \tag{2.3.5}$$

e_i being the linear functional defining i-th cut.

Note also that since $y \in G^\varepsilon$, we have a representation

$$y = (1 - \varepsilon)x^* + \varepsilon z \tag{2.3.6}$$

with certain $z \in G$.

Let us prove that in fact i-th step is productive. Indeed, assume it is not the case. Then

$$e_i = g'_j(x_i) \tag{2.3.7}$$

for some j such that $g_j^{(i)}(x_i) > \varepsilon g_j^{*,i}$. From this latter inequality combined with (2.3.7) and (2.3.5) it follows that

$$g_j^{(i)}(y) > g_j(x_i) = g_j^{(i)}(x_i) > \varepsilon g_j^{*,i}. \tag{2.3.8}$$

On the other hand, we have

$$g_j^{(i)}(y) = (1-\varepsilon)g_j^{(i)}(x^*) + \varepsilon g_j^{(i)}(z) \leq \varepsilon g_j^{*,i}$$

(we have taken into account that $g_j^{(i)}$ is a lower bound for $g_j(\cdot)$ and therefore this bound is nonpositive at the optimal solution to the problem); the resulting inequality contradicts (2.3.8), and thus the step i indeed is productive.

Now, since i is a productive step, we have $e_i = f'(x_i)$, and (2.3.5) implies therefore that

$$f(y) > f(x_i);$$

from this latter inequality and (2.3.6), exactly as in the case of problems with no functional constraints, it follows that

$$f(x_i) - f^* < f(y) - f^* \leq \varepsilon(\max_G f - f^*). \tag{2.3.9}$$

Now let us summarize our considerations. We have proved that in the case in question (i.e., when the method does not terminate during first N steps and (2.3.2) is satisfied) there exist a productive step $i \leq N$ such that (2.3.9) holds. Since the N-th approximate solution is the best (in terms of the values of the objective) of the search points generated at the productive steps with step numbers $\leq N$, it follows that $\bar{x}_N$ is well-defined and

$$f(\bar{x}_N) - f^* \leq f(x_i) - f^* \leq \varepsilon(\max_G f - f^*); \tag{2.3.10}$$

since $\bar{x}_N$ is, by construction, the search point generated at certain productive step i', we have also

$$g_j(\bar{x}_N) = g_j^{(i')}(x_{i'} \leq \varepsilon g_j^{*,i'} \leq \varepsilon \left(\max_G g_j\right)_+, \; j = 1, ..., m;$$

these inequalities combined with (2.3.10) results in

$$\varepsilon(p, \bar{x}_N) \leq \varepsilon,$$

as claimed. ∎

Combining Proposition 2.3.1 and the Grunbaum Theorem, we come to the *Center of Gravity method for problems with functional constraints*. The method is obtained from our general cutting plane scheme for constrained problems by the following specifications:

first, we use, as a current search point, the center of gravity of the previous localizer:

$$x_i = \frac{1}{\mathrm{Vol}_n(Q_{i-1})} \int_{Q_{i-1}} x dx;$$

second, we terminate the method after N-th step, N being given by the relation

$$N = \lfloor 2.181 n \ln\left(\frac{1}{\varepsilon}\right) \rfloor.$$

With these specifications the average diameter of i-th localizer at every step, due to the Grunbaum Theorem, decreases with i at least as

$$\exp\{-\frac{\kappa}{n} i\} \operatorname{AvDiam}(G), \quad \kappa = 0.45867...,$$

and since $\frac{1}{\kappa} < 2.181$, we come to

$$\operatorname{AvDiam}(G_N) < \varepsilon \operatorname{AvDiam}(G);$$

this latter inequality, in view of Proposition 2.3.1, implies that the method does find an ε-solution to every problem from the family, thus justifying the upper complexity bound we are proving.

2.4 Exercises: Extremal Ellipsoids

There are two natural ways to define an ellipsoid W in $\mathbf{R}^n$. The first is to represent W as the set defined by a convex quadratic constraint, namely, as

$$W = \{x \in \mathbf{R}^n \mid (x-c)^T A (x-c) \leq 1\} \tag{2.4.1}$$

A being a symmetric positive definite $n \times n$ matrix and c being a point in $\mathbf{R}^n$ (the center of the ellipsoid).

The second way is to represent W as the image of the unit Euclidean ball under an affine invertible mapping, i.e., as

$$W = \{x = Bu + c \mid u^T u \leq 1\}, \tag{2.4.2}$$

where B is an $n \times n$ nonsingular matrix and c is a point from $\mathbf{R}^n$.

Exercise 2.4.1 # *Prove that the above definitions are equivalent: if $W \subset \mathbf{R}^n$ is given by (2.4.1), then W can be represented by (2.4.2) with B chosen according to*

$$A = (B^{-1})^T B^{-1}$$

(e.g., with B chosen as $A^{-1/2}$). Vice versa, if W is represented by (2.4.2), then W can be represented by (2.4.1), where one should set

$$A = (B^{-1})^T B^{-1}.$$

Note that the (positive definite symmetric) matrix A involved into (2.4.1) is uniquely defined by W (why?); in contrast to this, a nonsingular matrix B involved into (2.4.2) is defined by W up to a right orthogonal factor: the matrices B and B' define the same ellipsoid if and only if $B' = BU$ with an orthogonal $n \times n$ matrix U (why?)

From the second description of an ellipsoid it immediately follows that

if

$$W = \{x = Bu + c \mid u \in \mathbf{R}^n, u^T u \le 1\}$$

is an ellipsoid and

$$x \mapsto p + B'x$$

is an invertible affine transformation of $\mathbf{R}^n$ *(so that* B' *is a nonsingular* $n \times n$ *matrix), then the image of* W *under the transformation also is an ellipsoid.*

Indeed, the image is nothing but

$$W' = \{x = B'Bu + (p + B'c) \mid u \in \mathbf{R}^n, u^T u \le 1\},$$

the matrix $B'B$ being nonsingular along with B and B'. It is also worthy of note that

for any ellipsoid

$$W = \{x = Bu + c \mid u \in \mathbf{R}^n, u^T u \le 1\}$$

there exists an invertible affine transformation of $\mathbf{R}^n$*, e.g., the transformation*

$$x \mapsto B^{-1}x - B^{-1}c,$$

which transforms the ellipsoid exactly into the unit Euclidean ball

$$V = \{u \in \mathbf{R}^n \mid u^T u \le 1\}.$$

In what follows we mainly focus on various volume-related issues; to avoid complicated constant factors, it is convenient to take, as the volume unit, the volume of the unit Euclidean ball V in $\mathbf{R}^n$ rather than the volume of the unit cube. The volume of a body[1] Q measured in this unit, i.e., the ratio

$$\frac{\mathrm{Vol}_n(Q)}{\mathrm{Vol}_n(V)},$$

Vol_n being the usual Lebesque volume in $\mathbf{R}^n$, will be denoted $\mathrm{vol}_n(Q)$ (we omit the subscript n if the value of n is clear from the context).

Exercise 2.4.2 # *Prove that if* W *is an ellipsoid in* $\mathbf{R}^n$ *given by (2.4.2), then*

$$\mathrm{vol}\, W = |\,\mathrm{Det}\; B|, \tag{2.4.3}$$

and if W *is given by (2.4.1), then*

$$\mathrm{vol}\, W = |\,\mathrm{Det}\; A|^{-1/2}. \tag{2.4.4}$$

Our local goal is to prove the following statement:

Let Q *be a convex body in* $\mathbf{R}^n$ *(i.e., a closed and bounded convex set with a nonempty interior). Then there exist ellipsoids containing* Q*, and among these ellipsoids there is one with the smallest volume. This ellipsoid is unique; it is called the outer extremal ellipsoid associated*

[1] in what follows "body" means a set with a nonempty interior

with Q. Similarly, there exist ellipsoids contained in Q, and among these ellipsoids there is one with the largest volume. This ellipsoid is unique; it is called the inner extremal ellipsoid associated with Q.

In fact we are not too interested in the uniqueness of the extremal ellipsoids (and you may try to prove the uniqueness yourself); what actually is of interest is the existence and some important properties of the extremal ellipsoids.

Exercise 2.4.3 $^{\#}$. *Prove that if Q is a closed and bounded convex body in $\mathbf{R}^n$, then there exist ellipsoids containing Q and among these ellipsoids there is (at least) one with the smallest volume.*

Exercise 2.4.4 . *Prove that if Q is a closed and bounded convex body in $\mathbf{R}^n$, then there exist ellipsoids contained in Q and among these ellipsoids there is (at least) one with the largest volume.*

Note that extremal ellipsoids associated with a closed and bounded convex body Q "accompany Q under affine transformations": if $x \mapsto Ax + b$ is an invertible affine transformation and Q' is the image of Q under this transformation, then the image W' of an extremal outer ellipsoid W associated with Q (note the article: we has not proved the uniqueness!) is an extremal outer ellipsoid associated with Q', and similarly for (an) extremal inner ellipsoid.
The indicated property is, of course, an immediate consequence of the facts that affine images of ellipsoids are again ellipsoids and that the ratio of volumes remains invariant under an affine transformation of the space.

In what follows we focus on outer extremal ellipsoids. Useful information can be obtained from investigating these ellipsoids for "simple parts" of an Euclidean ball.

Exercise 2.4.5 $^{\#*}$ *Let $n > 1$,*

$$V = \{x \in \mathbf{R}^n \mid |x|_2 \equiv \left\{\sum_{i=1}^n x_i^2\right\}^{1/2} \leq 1\}$$

be the unit Euclidean ball, let e be a unit vector in $\mathbf{R}^n$ and let

$$V_\alpha = \{x \in V \mid e^T x \geq \alpha\},\ \alpha \in [-1, 1]$$

(V_α is what is called a "spherical hat").

Prove that if

$$-\frac{1}{n} < \alpha < 1,$$

then the set V_α can be covered by an ellipsoid W of the volume

$$\mathrm{vol}_n(W) \leq \{\frac{n^2}{n^2-1}\}^{n/2}\sqrt{\frac{n-1}{n+1}}(1-\alpha^2)^{(n-1)/2}(1-\alpha) < 1 = \mathrm{vol}_n(V);$$

W is defined as

$$W = \{x = \frac{1+n\alpha}{n+1}e + Bu \mid u^T u \leq 1\},$$

where

$$B = \left\{(1-\alpha^2)\frac{n^2}{n^2-1}\right\}^{1/2}\left(I - \beta e e^T\right), \quad \beta = 1 - \sqrt{\frac{(1-\alpha)(n-1)}{(1+\alpha)(n+1)}}$$

In fact the ellipsoid given by the latter exercise is the extremal outer ellipsoid associated with V_α.

Looking at the result stated by the latter exercise, one may make a number of useful conclusions.

1. When $\alpha = 0$, i.e., when the spherical hat V_α is a half-ball, we have

$$\mathrm{vol}_n(W) = \left\{1 + \frac{1}{n^2-1}\right\}^{n/2} \sqrt{1 - \frac{2}{n-1}} \leq$$

$$\leq \left\{\exp\{1/(n^2-1)\}\right\}^{n/2} \exp\{-1/(n-1)\} =$$

$$= \exp\{-\frac{n+2}{2(n^2-1)}\} < \exp\{-\frac{1}{2n-2}\} = \exp\{-\frac{1}{2n-2}\}\mathrm{vol}_n(V);$$

thus, for the case of $\alpha = 0$ (and, of course, for the case of $\alpha > 0$) we may cover V_α by an ellipsoid with the volume $1 - O(1/n)$ times less than that one of V. In fact the same conclusion (with another absolute constant factor $O(1)$) holds true when α is negative (so that the spherical hat is greater than half-ball), but "not too negative", say, when $\alpha \geq -\frac{1}{2n}$.

2. In order to cover V_α by an ellipsoid of absolute constant times less volume than that one of V we need α to be positive of order $O(n^{-1/2})$ or greater. This fits our observation that the volume of V_α itself is at least absolute constant times less than that one of V only if $\alpha \geq O(n^{-1/2})$ (exercise 1.9). Thus, whenever the volume of V_α is absolute constant times less than that one of V, we can cover V_α by an ellipsoid of the volume also absolute constant times less than that one of V; this covering is already given by the Euclidean ball of the radius $\sqrt{1-\alpha^2}$ centered at the point αe (which, anyhow, is *not* the optimal covering presented in exercise 2.4.5).

Exercise 2.4.6 *Let V be the unit Euclidean ball in $\mathbf{R}^n$, e be a unit vector and let $\alpha \in (0,1)$. Consider the "symmetric spherical stripe"*

$$V^\alpha = \{x \in V \mid -\alpha \leq e^T x \leq \alpha\}.$$

Prove that if $0 < \alpha < 1/\sqrt{n}$ then V^α can be covered by an ellipsoid W with the volume

$$\mathrm{vol}_n(W) \leq \alpha\sqrt{n} \left\{\frac{n(1-\alpha^2)}{n-1}\right\}^{(n-1)/2} < 1 = \mathrm{vol}_n(V).$$

Find an explicit representation of the ellipsoid.

We see that in order to cover a symmetric spherical stripe of the unit Euclidean ball V by an ellipsoid of volume less than that one of V, it suffices to have the "half-thickness" α of the stripe to be $< 1/\sqrt{n}$, which again fits our observation (exercise 1.9) that basically all volume of the unit n-dimensional Euclidean ball is concentrated in the $O(1/\sqrt{n})$ neighbourhood of its "equator" - the cross-section of the ball and a hyperplane passing through the center of the ball. A useful exercise is to realize when a non-symmetric spherical stripe

$$V^{\alpha,\beta} = \{x \in V \mid -\alpha \leq e^T x \leq \beta\}$$

of the (centered at the origin) unit Euclidean ball V can be covered by an ellipsoid of volume less than that one of V.

The results of exercises 2.4.5 and 2.4.6 imply a number of important geometrical consequences.

Exercise 2.4.7 *Prove the following theorem of Fritz John:*

Let Q be a closed and bounded convex body in $\mathbf{R}^n$. Then

(i) *Q can be covered by an ellipsoid W in such a way that the concentric to W n times smaller ellipsoid*

$$W' = (1 - \frac{1}{n})c + \frac{1}{n}W$$

(c is the center of W) is contained in Q. One can choose as W the extremal outer ellipsoid associated with Q.

(ii) *If, in addition, Q is central-symmetric with respect to certain point c, then the above result can be improved: Q can be covered by an ellipsoid W centered at c in such a way that the concentric to W $\sqrt{n}$ times smaller ellipsoid*

$$W'' = (1 - \frac{1}{\sqrt{n}})c + \frac{1}{\sqrt{n}}W$$

is contained in Q.

Note that the constants n and $\sqrt{n}$ in the Fritz John Theorem are sharp; an extremal example for (i) is a simplex, and for (ii) - a cube.

Here are several nice geometrical consequences of the Fritz John Theorem:

> Let Q be a closed and bounded convex body in $\mathbf{R}^n$. Then
>
> 1. *There exist a pair of concentric homothetic with respect to their common center parallelotopes p, P with homothety coefficient equal to $n^{-3/2}$ such that $p \subset Q \subset P$; in other words, there exists an invertible affine transformation of the space such that the image Q' of Q under this transformation satisfies the inclusions*
>
> $$\{x \in \mathbf{R}^n \mid \| x \|_\infty \leq \frac{1}{n^{3/2}}\} \subset Q' \subset \{x \in \mathbf{R}^n \mid \| x \|_\infty \leq 1\};$$
>
> *here*
>
> $$\| x \|_\infty = \max_{1 \leq i \leq n} |x_i|$$
>
> *is the uniform norm of x.*
>
> Indeed, from the Fritz John Theorem it follows that there exists an invertible affine transformation resulting in
>
> $$\{x \mid |x|_2 \leq 1/n\} \subset Q' \subset \{x \mid |x|_2 \leq 1\},$$
>
> Q' being the image of Q under the transformation (it suffices to transform the outer extremal ellipsoid associated with Q into the unit Euclidean ball centered at the origin). It remains to note that the smaller Euclidean ball in the above chain of inclusions contains the cube $\{x \mid \| x \|_\infty \leq n^{-3/2}\}$ and the larger one is contained in the unit cube.

2. *If Q is central-symmetric, then the parallelotopes mentioned in 1. can be chosen to have the same center, and the homothety coefficient can be improved to $1/n$; in other words, there exists an invertible affine transformation of the space which makes the image Q' of Q central symmetric with respect to the origin and ensures the inclusions*

$$\{x \mid \| x \|_\infty \leq \frac{1}{n}\} \subset Q' \subset \{x \mid \| x \|_\infty \leq 1\}.$$

The statement is given by the reasoning completely similar to that one used for 1., up to the fact that now we should refer to item (ii) of the Fritz John Theorem.

3. *Any norm $\| \cdot \|$ on $\mathbf{R}^n$ can be approximated, within factor $\sqrt{n}$, by a Euclidean norm: given $\| \cdot \|$, one can find a Euclidean norm*

$$|x|_A = (x^T A x)^{1/2},$$

A being a symmetric positive definite $n \times n$ matrix, in such a way that

$$\frac{1}{\sqrt{n}} |x|_A \leq \| x \| \leq |x|_A$$

for any $x \in \mathbf{R}^n$.

Indeed, let $\mathcal{B} = \{x \mid \| x \| \leq 1\}$ be the unit ball with respect to the norm $\| \cdot \|$; this is a closed and bounded convex body, which is central symmetric with respect to the origin. By item (ii) of the Fritz John Theorem, there exists a centered at the origin ellipsoid

$$W = \{x \mid x^T A x \leq n\}$$

(A is an $n \times n$ symmetric positive definite matrix) which contains $\mathcal{B}$, while the ellipsoid

$$\{x \mid x^T A x \leq 1\}$$

is contained in $\mathcal{B}$; this latter inclusion means exactly that

$$|x|_A \leq 1 \Rightarrow x \in \mathcal{B} \Leftrightarrow \| x \| \leq 1,$$

i.e., means that $|x|_A \geq \| x \|$. The inclusion $\mathcal{B} \subset W$, by similar reasons, implies that $\| x \| \geq n^{-1/2} |x|_A$.

Remark 2.4.1 The third of the indicated consequences says that any norm on $\mathbf{R}^n$ can be approximated, within constant factor $\sqrt{n}$, by an appropriately chosen Euclidean norm. It turns out that the quality of approximation can be done much better, if we would be satisfied by approximating the norm not at the whole space, but at a properly chosen subspace. Namely, there exists a marvelous and important theorem of Dvoretsky which is as follows:

there exists a function $m(n, \varepsilon)$ of positive integer n and positive real ε with the following properties:

first,

$$\lim_{n \to \infty} m(n, \varepsilon) = +\infty$$

and,

second, whenever $\|\cdot\|$ is a norm on $\mathbf{R}^n$, one can indicate a $m(n,\varepsilon)$-dimensional subspace $E \subset \mathbf{R}^n$ and a Euclidean norm $|\cdot|_A$ on $\mathbf{R}^n$ such that $|\cdot|_A$ approximates $\|\cdot\|$ on E within factor $1+\varepsilon$:

$$(1-\varepsilon)|x|_A \leq \| x \| \leq (1+\varepsilon)|x|_A, \quad x \in E.$$

In other words, the Euclidean norm is "marked by God": for any given integer k an arbitrary normed linear space contains an "almost Euclidean" k-dimensional subspace, provided that the dimension of the space is large enough.

2.4.1 Tschebyshev-type results for sums of random vectors

The series of exercises we are about to present deals with issues which, at the first glance, are far from extremal ellipsoids and optimization - namely, with the Tschebyshev-type results for sums of random vectors. These results (which are useful in various applications - non-parametric estimation, Stochastic Programming, etc.) are a surprising byproduct of our previous considerations.

Recall that the standard Tschebyshev equality states that if $\{\xi_i\}_{i=1}^\infty$ is a sequence of mutually independent random reals such that

$$\mathcal{E}\{\xi_i\} = 0, \, \mathcal{E}\{\xi_i^2\} < \infty, \; i = 1, 2, ...,$$

($\mathcal{E}$ stands for expectation), then

$$\mathcal{E}\{\left(\sum_{i=1}^N \xi_i\right)^2\} \leq \sum_{i=1}^N \mathcal{E}\{\xi_i^2\}$$

(in fact, of course, there is equality rather than inequality). The proof is immediate: open the parentheses and take the termwise expectation; the expectations of "cross-products" $\xi_i\xi_j$, $i \neq j$, will disappear, since ξ_i and ξ_j are independent with zero means.

The issue we are interested in is: what can be said when the sequence is formed of random vectors rather than random reals? Namely, let $\|\cdot\|$ be certain once for ever fixed norm on $\mathbf{R}^n$, and let $\{\xi_i\}_{i=1}^\infty$ be a sequence of mutually independent random vectors from $\mathbf{R}^n$ such that

$$\mathcal{E}\{\xi_i\} = 0, \, \mathcal{E}\{\| \xi_i \|^2\} \equiv s_i^2 < \infty, \, i = 1, 2, ...$$

Now, let

$$\Xi_N = \sum_{i=1}^N \xi_i.$$

What can be said about the quantities

$$S_N^2 = \mathcal{E}\{\| \Xi_N \|^2\} \; ?$$

First of all, if $\|\cdot\|$ is a Euclidean norm $|x|_A = (x^T A x)^{1/2}$ (A is an $n \times n$ symmetric positive definite matrix), then we have exactly the same inequality as in the scalar case:

$$S_N^2 \leq \sum_{i=1}^N s_i^2 \tag{2.4.5}$$

(in fact this is an equality). The proof remains the same:

$$|\Xi_N|_A^2 = \sum_{i,j=1}^N \xi_i^T A \xi_j;$$

when taking expectation, cross-product terms disappear, and we come to (2.4.5). As we shall see in a while, this is a "bad" reasoning: in heavily exploits the algebraic structure of the Euclidean norm and therefore cannot be extended onto more general norms.

The "Euclidean" version of the Tschebyshev inequality immediately leads to the following general statement: for *any* norm $\|\cdot\|$ on $\mathbf{R}^n$ one has

$$S_N^2 \le n \sum_{i=1}^N s_i^2. \tag{2.4.6}$$

Exercise 2.4.8 * *Prove (2.4.6).*

Exercise 2.4.9 * *Prove that the factor n in (2.4.6) is sharp, as far as the order of dependence on n is concerned: one can find a norm $\|\cdot\|$ on $\mathbf{R}^n$ and a sequence of iid (independent identically distributed) random n-dimensional vectors $\{\xi_i\}$ such that*

$$s_i^2 \equiv \mathcal{E}\{\| \xi_i \|^2\} = 1$$

and

$$\liminf_{N\to\infty} S_N^2/N \equiv \liminf_{N\to\infty} \frac{S_N^2}{\sum_{i=1}^N s_i^2} \ge \frac{2}{\pi} n.$$

Results stated in exercises 2.4.8 and 2.4.9 are quite satisfactory (and, let me say, not interesting at all) when the dimension n of our random vectors is regarded as something once for ever fixed; in all applications of the "vector Tschebyshev-type inequalities" known to me n is large, and what we actually are interested in is what is the (depending on the norm $\|\cdot\|$) constant factor $\gamma_{\|\cdot\|}$ in the estimate

$$S_N^2 \le \gamma_{\|\cdot\|} \sum_{i=1}^N s_i^2.$$

As we just have seen, this factor always does not exceed n, and sometimes (i.e., for some particular norms on $\mathbf{R}^n$ and some random sequences) this value n cannot be significantly improved. On the other hand, for some particular norms it can be improved (say, for a Euclidean norm it can be set to 1 independently of the dimension). What are those norms which allow to reduce the factor, as compared to its worst-case with respect to the norms value $O(n)$? The most important specification of this question is as follows: what happens in the case of the uniform norm

$$|x|_\infty = \max_{1\le i\le n} |x_i| \ ?$$

This norm arises in many applications, and, believe me, it is useful to know a "good" Tschebyshev-type inequality for this norm.

At the first glance it seems that the answer should be the same as that one for the 1-norm

$$|x|_1 = \sum_{i=1}^n |x_i|,$$

since the norms $|\cdot|_1$ and $|\cdot|_\infty$ are dual to each other, so that the norms are, roughly speaking, of the same level of similarity (better to say, dissimilarity) to the Euclidean norm (for this latter norm, $\gamma = 1$). Those who have solved exercise 2.4.9 know that for the 1-norm γ is as bad as it could be, i.e., it is of order of n; therefore one could expect the uniform norm also to be bad. Surprisingly, the latter norm is almost as good as the Euclidean one:

$$\gamma_{|\cdot|_\infty} = O(\ln n). \tag{2.4.7}$$

To establish (2.4.7), this is the goal of the remaining exercises.

It is reasonable to deal not with the uniform norm directly, but with the family of the standard norms "linking" the 1-norm and the uniform one, i.e., with the family of the l_p-norms on $\mathbf{R}^n$

$$|x|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}.$$

Here $1 \le p \le \infty$, and for $p = \infty$ the right hand side is, by definition, the uniform norm $\max_i |x_i|$. Note that $|\cdot|_2$ is the standard Euclidean norm.

Let us start with the following simple observation:

Exercise 2.4.10 $^+$ *One has*

$$n^{1/p'-1/p}|x|_p \ge |x|_{p'} \ge |x|_p, \quad x \in \mathbf{R}^n,\ \infty \ge p \ge p' \ge 1; \tag{2.4.8}$$

in particular, if $n \ge 3$, then

$$|x|_{2\ln n} \ge |x|_\infty \ge \exp\{-\frac{1}{2}\}|x|_{2\ln n}. \tag{2.4.9}$$

To get an idea of what to do, let us look once again at the proof of the Tschebyshev inequality for the case of the Euclidean norm $|\cdot|_2$, but now let us pass to an "equivalent" reasoning: rather than to express $|\Xi_N|_2^2$ in terms of pair inner products of the terms forming the sum Ξ_N, let us use the recurrence

$$|\Xi_{k+1}|_2^2 = |\Xi_k|_2^2 + 2\xi_{k+1}^T\Xi_k + |\xi_{k+1}|_2^2; \tag{2.4.10}$$

taking expectation, we come to

$$S_{k+1}^2 = S_k^2 + s_{k+1}^2 \tag{2.4.11}$$

(the expectation of the linear in ξ_{k+1} term $\xi_{k+1}^T\Xi_k$ equals to zero, since the second factor depends only on $\xi_1, ..., \xi_k$ and is therefore independent of ξ_{k+1}, while the expectation of the first factor is zero).

Now, the Cosine Theorem

$$|x+h|_2^2 = |x|_2^2 + 2h^Tx + |h|_2^2$$

underlying (2.4.10) is nothing but the second-order Taylor representation

$$f(x+h) = f(x) + h^T f'(x) + \frac{1}{2} h^T f''(x) h$$

of the quadratic function

$$f(x) = |x|_2^2;$$

for this particular function, the Taylor representation is an equality. Now note that if we, for a given f (which should not necessarily be $|x|_2^2$), start with an inequality of the type

$$f(x+h) \le f(x) + h^T f'(x) + r(h) \tag{2.4.12}$$

(i.e., with the first order Taylor expansion of f written as an inequality with an appropriate remainder $r(h)$ depending on h only and independent of x), we could evaluate from above the expectations

$$\sigma_N \equiv \mathcal{E}\{f(\Xi_N)\}$$

acting exactly in the manner we just have used for the quadratic f:

$$f(\Xi_{k+1}) \le f(\Xi_k) + \xi_{k+1}^T f'(\Xi_k) + r(\xi_{k+1}) \Rightarrow$$
$$\Rightarrow \mathcal{E}\{f(\Xi_{k+1})\} \le \mathcal{E}\{f(\Xi_k)\} + \mathcal{E}\{r(\xi_{k+1})\}$$

(the expectation of the linear in ξ_{k+1} term $\xi_{k+1}^T f'(\Xi_k)$ disappears since the first factor has zero mean and the second one is independent of the first). Thus, given (2.4.12), we may write the recurrence

$$\mathcal{E}\{f(\Xi_{k+1})\} \le \mathcal{E}\{f(\Xi_k)\} + \mathcal{E}\{r(\xi_{k+1})\}, \tag{2.4.13}$$

which allows to evaluate sequentially all the quantities $\mathcal{E}\{f(\Xi_N)\}$, $N = 1, 2, ...$, in terms of s_i, provided that we are able to bound from above, in terms of s_i, the expectations $\mathcal{E}\{r(\xi_i)\}$. A good news here is that our estimating abilities depend on the possibility to bound from above the remainder r in the first order Taylor expansion of f, i.e., on smoothness of f, not on the particular algebraic structure of f.

To summarize our considerations at this point, let us present the statement we just have proved in the following, although slightly restricted, but quite sufficient for us form:

let f be a continuously differentiable function on $\mathbf{R}^n$ such that, for certain constant $\omega_f \in [0, \infty)$ and all $x, h \in \mathbf{R}^n$ one has

$$f(x+h) \le f(x) + h^T f'(x) + \omega_f f(h). \tag{2.4.14}$$

Let also $\{\xi_i\}$ be a sequence of mutually independent random n-dimensional vectors with zero means and finite quantities

$$\sigma_i^{(f)} = \mathcal{E}\{f(\xi_i)\}.$$

Then, for all N, one has

$$\mathcal{E}\{f(\Xi_N)\} \le f(0) + \omega_f \sum_{i=1}^N \sigma_i^{(f)}, \tag{2.4.15}$$

where, as always,

$$\Xi_N = \sum_{i=1}^N \xi_i.$$

The proof is an immediate consequence of recurrence (2.4.13).

In order to obtain a Tschebyshev-type inequality for the uniform norm, we could try to use the latter statement, $f(x)$ being $|x|_\infty^2$. An immediate difficulty is that the function is *not* smooth, so that our reasoning cannot be applied to it directly. But this is a minor difficulty, since we, due to the result given by exercise 2.4.10, can approximate the uniform norm by a l_p-norm with finite $p > 2$, the latter norm being smooth.

Exercise 2.4.11 $^+$ *Let $p \in (2, \infty)$, and let*

$$f(x) = |x|_p^2 : \mathbf{R}^n \to \mathbf{R}.$$

Prove that this function satisfies (2.4.14) with

$$\omega_f = 2^{(2p-3)/(p-1)}(p-1) + 1 \le 4p - 3 \tag{2.4.16}$$

and, consequently,

$$\mathcal{E}\{|\Xi_N|_p^2\} \le (4p-3)\sum_{i=1}^N \mathcal{E}\{|\xi_i|_p^2\} \tag{2.4.17}$$

for every sequence $\{\xi_i\}$ of mutually independent random n-dimensional vectors with zero mean and finite variances $\mathcal{E}\{|\xi_i|_p^2\}$.

Exercise 2.4.12 * *Prove the following Tschebyshev-type inequality for the uniform norm:*

$$\mathcal{E}\{|\Xi_N|_\infty^2\} \le (8\ln n - 3)\exp\{1\}\sum_{i=1}^N \mathcal{E}\{|\xi_i|_\infty^2\} \tag{2.4.18}$$

for any N, $n \ge 3$ and any sequence $\{\xi_i\}$ of random mutually independent n-dimensional vectors with zero means and finite variances $\mathcal{E}\{|x_i|_\infty^2\}$.

Thus, the constant $\gamma_{|\cdot|_\infty}$ for the uniform norm on $\mathbf{R}^n$ can be chosen to be of order of $\ln n$, i.e., "almost independent of n". You may try to prove that the constant *should* be of order of $\ln n$, provided that no additional restrictions on $\{\xi_i\}$ are imposed.

Lecture 3

Methods with linear convergence, II

3.1 Lower complexity bound

To complete the proof of Theorem 2.1.1, it remains to establish the lower complexity bound. This is done, basically, in the same way as in the one-dimensional case. In order to avoid things difficult for verbal presentation, I shall restrict myself with demonstrating a slightly worse lower bound

$$\mathrm{Compl}(\varepsilon) \geq O(1)\frac{n\ln(\frac{1}{\varepsilon})}{\ln\left(n\ln(\frac{1}{\varepsilon})\right)}, \quad 0 < \varepsilon < \varepsilon^*(G), \tag{3.1.1}$$

$O(1)$ being a positive absolute constant and $\varepsilon^*(G)$ being certain positive quantity depending on G only (we shall see that this quantity measures how G differs from a parallelotope).

The "spoiled" bound (3.1.1) (which is by logarithmic denominator worse than the estimate announced in the Theorem) is more or less immediate consequence of our one-dimensional considerations. Of course, it is sufficient to establish the lower bound for the case of problems without functional constraints, since the constrained ones form a wider family (indeed, a problem without functional constraints can be thought of as a problem with a given number m of trivial, identically zero functional constraints). Thus, in what follows the number of constraints m is set to 0.

Let me start with the following simple observation. Let, for a given $\varepsilon > 0$ and a convex objective f, the set $G_\varepsilon(f)$ be comprised of all approximate solutions to f of relative accuracy not worse than ε:

$$G_\varepsilon(f) = \{x \in G \mid f(x) - \min_G f \leq \varepsilon(\max_G f - \min_G f)\}.$$

Assume that, for a given $\varepsilon > 0$, we are able to point out a finite set $\mathcal{F}$ of objectives with the following two properties:

(I) no different problems from $\mathcal{F}$ admit a common ε-solution:

$$G_\varepsilon(f) \cap G_\varepsilon(\bar{f}) = \emptyset$$

whenever $f, \bar{f} \in \mathcal{F}$ and $f \neq \bar{f}$;

(II) given in advance that the problem in question belongs to $\mathcal{F}$, one can compress an answer of the first order local oracle to be a $(\log_2 K)$-bit word. It means the following. For

certain positive integer K one can indicate a function $\mathcal{I}(f,x)$ taking values in a K-element set and a function $\mathcal{R}(i,x)$ such that

$$\mathcal{O}(f,x) = \mathcal{R}(\mathcal{I}(f,x),x), \quad f \in \mathcal{F}, x \in \text{int } G.$$

In other words, given in advance that the problem we are interested in belongs to $\mathcal{F}$, a method can imitate the first-order oracle $\mathcal{O}$ via another oracle $\mathcal{I}$ which returns $\log_2 K$ bits of information rather than infinitely many bits contained in the answer of the first order oracle; given the "compressed" answer $\mathcal{I}(f,x)$, a method can substitute this answer, along with x itself, into a universal (defined by $\mathcal{F}$ only) function in order to get the "complete" first-order information on the problem.

E.g., consider the family $\mathcal{F}_n$ comprised of 2^n convex functions

$$f(x) = \max_{i=1,\dots,n} \epsilon_i x_i,$$

where all ϵ_i are ± 1. At every point x a function from the family admits a subgradient of the form $\mathcal{I}(f,x) = \pm e_i$ (e_i are the orths of the axes), with i, same as the sign at e_i, depending on f and x. Assume that the first order oracle in question when asked about $f \in \mathcal{F}_n$ reports a subgradient of exactly this form. Since all functions from the family are homogeneous, given x and $\mathcal{I}(f,x)$ we know not only a subgradient of f at x, but also the value of f at the point:

$$f(x) = x^T \mathcal{I}(f,x).$$

Thus, our particular first-order oracle as restricted onto $\mathcal{F}_n$ can be compressed to $\log_2(2n)$ bits.

Now let us make the following observation:

(*):
under assumptions (I) and (II) the ε-complexity of the family $\mathcal{F}$, and therefore of every larger family, is at least

$$\frac{\log_2|\mathcal{F}|}{\log_2 K}.$$

Indeed, let $\mathcal{M}$ be a method which solves all problems from $\mathcal{F}$ within accuracy ε in no more than N steps. We may assume (since informationally this is the same) that the method uses the oracle $\mathcal{I}$, rather than the first-order oracle. Now, the behaviour of the method is uniquely defined by the sequence of answers of $\mathcal{I}$ in course of N steps; therefore there are at most K^N different sequences of answers and, consequently, no more than K^N different trajectories of $\mathcal{M}$. In particular, the set $\bar{X}$ formed by the results produced by $\mathcal{M}$ as applied to problems from $\mathcal{F}$ is comprised of at most K^N points. On the other hand, since $\mathcal{M}$ solves every of $|\mathcal{F}|$ problems of the family within accuracy ε, and no two different problems from the family admit a common ε-solution, $\bar{X}$ should contain at least $|\mathcal{F}|$ points. Thus,

$$K^N \geq |\mathcal{F}|,$$

as claimed.

As an immediate consequence of what was said, we come to the following result:

the complexity of minimizing a convex function over an n-dimensional parallelotope G within relative accuracy $\varepsilon < 1/2$ is at least $n/(1+\log_2 n)$.

Indeed, all our problem classes and complexity-related notions are affine invariant, so that we always may assume the parallelotope G mentioned in the assertion to be the unit cube

$$\{x \in \mathbf{R}^n \mid |x|_\infty \equiv \max_i |x_i| \leq 1\}.$$

For any $\varepsilon < \frac{1}{2}$ the aforementioned family

$$\mathcal{F}_n = \{f(x) = \max_i \epsilon_i x_i\}$$

clearly possesses property (I) and, as we have seen, at least for certain first-order oracle possesses also property (II) with $K = 2n$. We immediately conclude that the complexity of finding an ε-minimizer, $\varepsilon < 1/2$, of a convex function over an n-dimensional parallelotope is, at least for some first order oracle, no less than

$$\frac{\log_2 |\mathcal{F}|}{\log_2(2n)},$$

as claimed. In fact, of course, the complexity is at least n for *any* first order oracle, but to prove the latter statement it requires more detailed considerations.

Now let us use the above scheme to derive the lower bound (3.1.1). Recall that when studying the one-dimensional case, we have introduced certain family of univariate convex functions which was as follows. The functions of the family form a tree, with the root ("generation 0") being the function

$$f^{\rm root}(x) = |x|;$$

when subject to the left and to the right modifications, the function produces two "children", let them be called f_r and f_l; each of these functions, in turn, may be subject to the right and to the left modification, producing two new functions, so that at the level of "grandchildren" there are four functions f_{rr}, f_{rl}, f_{lr}, f_{ll}, and so on. Now, every of the functions $f_.$ of a "generation" $k > 0$ possesses its own active segment $\delta(f)$ of the length 2^{1-2k}, and at this segment the function is modulus-like:

$$f_.(x) = a^{(k)} + 8^{-k}|x - c(f)|,$$

$c(f)$ being the midpoint of $\delta(f)$. Note that $a^{(k)}$ depends only on the generation $f_.$ belongs to, not on the particular representative of the generation; note also that the active segments of the 2^k functions belonging to the generation k are mutually disjoint and that a function from our "population" coincides with its "parent" outside the active segment of the parent. In what follows it is convenient also to define the active segment of the root function $f^{\rm root}$ as the whole axis.

Now, let $\mathcal{F}_k$ be the set of 2^k functions comprising k-th generation in our population. Let us demonstrate that any first order oracle, restricted onto this family of functions, admits compression to $\log_2(2k)$ bits. Indeed, it is clear from our construction that in order to restore, given an x, $f(x)$ and a subgradient $f'(x)$, it suffices to trace the path of predecessors of f - its father, its grandfather, ... - and to find the "youngest" of them, let it be $\bar{f}$, such that x belongs to the active segment of $\bar{f}$ (let us call this predecessor the *active at* x predecessor of f). The active at x predecessor of f does exist, since the active segment of the "common predecessor" $f^{\rm root}$ is the whole axis. Now, f is obtained from $\bar{f}$ by a number of modifications; the first of them possibly varies $\bar{f}$ in a neighbourhood of x (x is in the active segment of $\bar{f}$), but the subsequent modifications do not, since x is outside the corresponding active segments. Thus, in a neighbourhood

of x f coincides with the function $\tilde{f}$ - the modification of $\bar{f}$ which leads from $\bar{f}$ to f. Now, to identify the local behaviour of $\tilde{f}$ (i.e., that one of f) at x, it suffices to indicate the "age" of $\tilde{f}$, i.e., the number of the generation it belongs to, and the type of the modification - left or right - which transforms $\bar{f}$ into $\tilde{f}$.
Indeed, given x and the age $\bar{k}$ of $\tilde{f}$, we may uniquely identify the active segment of $\bar{f}$ (since the segments for different members of the same generation $\bar{k}-1$ have no common points); given the age of $\bar{f}$, its active segment and the type of modification leading from $\bar{f}$ to $\tilde{f}$, we, of course, know $\tilde{f}$ in a neighbourhood of the active segment of $\bar{f}$ and consequently at a neighbourhood of x.

Thus, to identify the behaviour of f at x and therefore to imitate the answer of any given local oracle on the input x, it suffices to know the age $\bar{k}$ of the active at x predecessor of f and the type - left or right - of modification which "moves" the predecessor towards f, i.e., to know a point from certain $(2k)$-element set, as claimed.

Now let us act as follows. Let us start with the case when our domain G is a parallelotope; due to affine invariance of our considerations, we may assume G to be the unit n-dimensional cube:

$$G = \{x \in \mathbf{R}^n \mid |x|_\infty \leq 1\}.$$

Given a positive integer k, consider the family $\mathcal{F}^k$ comprised of the objectives

$$f_{i_1,...,i_n}(x) = \max\{f_{i_1}(x_1), ..., f_{i_n}(x_n)\}, \quad f_{i_s} \in \mathcal{F}_k, s = 1, ..., n.$$

This family contains $|\mathcal{F}|^n = 2^{nk}$ objectives, all of them clearly being convex and Lipschitz continuous with constant 1 with respect to the uniform norm $|\cdot|_\infty$. Let us demonstrate that there exists a first order oracle such that the family, equipped with this oracle, possesses properties (I), and (II), where one should set

$$\varepsilon = 2^{-5k}, \quad K = 2nk. \tag{3.1.2}$$

Indeed, a function $f_{i_1,...,i_n}$ attains its minimum $a^{(k)}$ exactly at the point $x_{i_1,...,i_n}$ with the coordinates comprised of the minimizers of $f_{i_s}(x_s)$. It is clear that within the cube

$$C_{z_1,...,i_n} = \{x \mid |x - x_{i_1,...,i_n}|_\infty \leq 2 \times 2^{-2k}\}$$

(i.e., within the direct product of the active segments of f_{i_s}, $s = 1, ..., n$) the function is simply

$$a^{(k)} + 8^{-k}|x - x_{i_1,...,i_n}|_\infty,$$

therefore outside this cube one has

$$f_{i_1,...,i_n}(x) - \min_G f_{i_1,...,i_n} > 2^{-5k} = \varepsilon.$$

Taking into account that all our functions $f_{i_1,...,i_n}$, being restricted onto the unit cube G, take their values in $[0, 1]$, so that for these functions absolute inaccuracy in terms of the objective is majorated by the relative accuracy, we come to

$$G_\varepsilon(f_{i_1,...,i_n}) \subset C_{i_1,...,i_n}.$$

It remains to note that the cubes $C.$ corresponding to various functions from the family are mutually disjoint (since the active segments of different elements of the generation $\mathcal{F}_k$ are disjoint). Thus, (I) is verified.

In order to establish (II), let us note that to find the value and a subgradient of $f_{i_1,...,i_n}$ at a point x it suffices to know the value and a subgradient at x_{i_s} of any function f_{i_s} which is "active" at x, i.e., is $\geq$ all other functions participating in the expression for $f_{i_1,...,i_n}$. In turn, as we know, to indicate the value and a subgradient of f_{i_s} it suffices to report a point from a $(2k)$-element set. Thus, one can imitate *certain* (not *any*) first order oracle for the family $\mathcal{F}^k$ via a "compressed" oracle reporting $\log_2(2nk)$-bit word (it suffices to indicate the number s, $1 \leq s \leq n$ of a component f_{i_s} active at x and a point of a $(2k)$-element set to identify f_{i_s} at x_{i_s}).

Thus, we may imitate certain first order oracle for the family $\mathcal{F}^k$ (comprised of 2^{kn} functions), given a "compressed" oracle with $K = 2nk$; it follows from (*) that the ε-complexity of $\mathcal{F}$ for $\varepsilon = 2^{-5k}$ (see (3.1.2)) is at least

$$\frac{\log_2(2^{nk})}{\log_2(2nk)};$$

expressing k in terms of ε, we come to

$$\mathrm{Compl}(\varepsilon) \geq \frac{n\log_2(\frac{1}{\varepsilon})}{5\log_2\left(n\log_2(\frac{2}{\varepsilon})\right)}, \quad \varepsilon = 2^{-5k}, k = 1, 2, ...;$$

taking into account that $\mathrm{Compl}(\varepsilon)$ is a non-increasing function, we come to an estimate of the type (3.1.1) with $\varepsilon^*(G) = \frac{1}{32}$ (recall that we were dealing with the unit cube G).

Let me stress what was in fact was proved. We have demonstrated that *it is possible to equip the family of all convex continuous functions on the unit cube with a first order oracle in such a way that the complexity of the family would satisfy lower bound (3.1.1), provided that* $\varepsilon < \frac{1}{32}$.

Now, what happens in the case when G is not a parallelotope? In this case we can find a pair of homothetic parallelotopes, p and P, such that $p \subset G \subset P$. Let us choose among these pairs p, P that one with the smallest possible similarity ratio, let it be called $\alpha(G)$. Those who have solved problems given at the previous lecture know, and other should believe that

$$\alpha(G) \leq n^{3/2}$$

for any closed and bounded convex body $G \subset \mathbf{R}^n$. After appropriate affine transformation of the space (the transformation does not influence the complexity - once again, all our complexity-related notions are affine invariant) we may assume that p is the unit cube. Thus, we come to the situation as follows:

$$C_n \equiv \{x \in \mathbf{R}^n \mid |x|_\infty \leq 1\} \subset G \subset \{x \in \mathbf{R}^n \mid |x|_\infty \leq \alpha(G)\}. \tag{3.1.3}$$

Now let us look at the family of problems

$$\textit{minimize} \;\; f(x) \;\; \textit{s.t.} \;\; x \in G$$

associated with $f \in \mathcal{F}^k$. It is easily seen that all the objectives from the family attain their global (i.e., over the whole space) minima within the unit cube and the sets of approximate

solutions of *absolute* accuracy 2^{-5k} to problems of the family are mutually disjoint (these sets are exactly the small cubes already known to us). Therefore our previous reasoning states that it is possible to equip the family with a first order oracle in such a way that the worst-case, with respect to the family $\mathcal{F}^k$, complexity of finding an approximate solution of *absolute* accuracy 2^{-5k} is at least $\frac{n\ln_2 k}{\ln_2(2nk)}$. On the other hand, the Lipschitz constant of every function from the family $\mathcal{F}^k$ taken with respect to the uniform norm is, as we know, at most 1, so that the variation

$$\max_G f - \min_G f$$

of such a function on the domain G is at most the diameter of G with respect to the uniform norm; the latter diameter, due to (3.1.3), is at most $2\alpha(G)$. It follows that any method which solves all problems from $\mathcal{F}^k$ within relative accuracy $2^{-5k-1}/\alpha(G)$ solves all these problems within absolute accuracy 2^{-5k} as well; thus, the complexity of minimizing convex function over G within relative accuracy $2^{-5k-1}/\alpha(G)$ is at least $\frac{n\log_2 k}{\log_2(2nk)}$:

$$\mathrm{Compl}(2^{-5k-1}/\alpha(G)) \geq \frac{n\log_2 k}{\log_2(nk)},\ k = 1, 2, ...$$

This lower bound immediately implies that

$$\mathrm{Compl}(\varepsilon) \geq O(1)\frac{n\log_2\left(\frac{1}{\alpha(G)\varepsilon}\right)}{\log_2\left(n\log_2(\frac{1}{\alpha(G)\varepsilon})\right)},\quad \alpha(G)\varepsilon < \frac{1}{64},$$

whence, in turn,

$$\mathrm{Compl}(\varepsilon) \geq O(1)\frac{n\ln(1/\varepsilon)}{\ln\left(n\ln(1/\varepsilon)\right)},\quad \varepsilon \leq \varepsilon^*(G) \equiv \frac{1}{64\alpha^2(G)}(\geq \frac{1}{64n^3});$$

this is exactly what is required in (3.1.1).

Note that our reasoning results in a lower bound which is worse than that one indicated in the Theorem not only by the logarithmic denominator, but also due to the fact that this is a lower bound for a *particular* first order oracle, not for an arbitrary one. In fact both these shortcomings - I mean the presence of the denominator and the "oracle-dependent" type of the lower bound - may be overcome by more careful reasoning, but this is what I am not going to do here.

3.2 The Ellipsoid method

We have presented the Center of Gravity method which is optimal in complexity, up to an absolute constant factor, at least when the required accuracy is small enough. The rate of convergence of the method, which is the best possible theoretically, looks fine also from the practical viewpoint. The upper complexity bound $O(1)n\ln(1/\varepsilon)$ associated with the method means that in order to improve inaccuracy by an absolute constant factor (say, reduce it by factor 10) it suffices to perform $O(1)n$ iterations more, which looks not too bad, taking into account how wide is the family of problems the method can be applied to. Thus, one may ask: what were the computational consequences of the method invented as early as in 1965? The

answer is: no consequences at all, since the method cannot be used in practice, provided that the dimension of the problem is, say > 4. The reason is that the auxiliary problems arising at the steps of the method, I mean those of finding centers of gravity, are computationally extremely hard; the source of the difficulty is that the localizers arising in the method may be almost arbitrary solids; indeed, all we can say is that if the domain G of the problem is a polytope given by k_0 linear inequalities, then the localizer G_i formed after i steps of the method is a polytope given by $k_0 + i$ linear inequalities; and all known deterministic methods for computing the center of gravity of a polytope in $\mathbf{R}^n$ given by $k > 2n$ linear inequalities take an exponential in n time to find the center. In our definition of complexity we ignore the computational effort required to implement the search rules; but in practice this effort, of course, should be taken into account, this is why the Center of Gravity method, for which this latter effort is tremendous, cannot be used at all.

The situation, nevertheless, is not as bad as one could think: to the moment we have not exploit all abilities of the cutting plane scheme.

Let us note that the cutting plane scheme can be "spoiled" as follows. Given previous localizer G_{i-1}, we apply our basic scheme to produce a new localizer, $\bar{G}_i$, but now this is something intermediate, not the localizer we are forwarding to the next step (this is why we denote by $\bar{G}_i$ the solid which in the basic scheme was designated G_i); the localizer we do use at the next step is certain larger solid $G_i \supset \bar{G}_i$. Thus, at a step of the modified cutting plane scheme we perform a cut, exactly as in the basic scheme, and enlarge the resulting localizer to obtain G_i.

At this point one could ask: what for should we add to an actual localizer something which for sure does not contain optimal solutions? The answer is: acting in this manner, we may stabilize geometry of our localizers and enforce them to be convenient for numerical implementation of the search rules. This is the idea underlying the *Ellipsoid method* I am about to present.

3.2.1 Ellipsoids

Recall that an *ellipsoid* in $\mathbf{R}^n$ is defined as a level set of a nondegenerate convex quadratic form, i.e., as a set of the type

$$W = \{x \in \mathbf{R}^n \mid (x - c)^T A(x - c) \leq 1\}, \tag{3.2.1}$$

where A is an $n \times n$ symmetric positive definite matrix and $c \in \mathbf{R}^n$ is the center of the ellipsoid. An equivalent definition, which will be more convenient for us, says that an ellipsoid is the image of the unit Euclidean ball under a one-to-one affine transformation:

$$W = W(B, c) = \{x = Bu + c \mid u^T u \leq 1\}, \tag{3.2.2}$$

where B is an $n \times n$ nonsingular matrix. It is immediately seen that one can pass from representation (3.2.2) to (3.2.1) by setting

$$A = (B^T)^{-1} B^{-1}; \tag{3.2.3}$$

since any symmetric positive definite matrix A admits a representation of the type (3.2.3) (e.g., with $B = A^{-1/2}$), the above definitions indeed are equivalent.

From (3.2.2) it follows immediately that

$$\mathrm{Vol}_n(W(B,c)) = |\,\mathrm{Det}\ B|\mathrm{Vol}_n(V), \tag{3.2.4}$$

where V denotes the unit Euclidean ball in $\mathbf{R}^n$.

Now, by compactness reasons it is easily seen that for any n-dimensional solid Q there exist ellipsoids containing Q and among these ellipsoids there is at least one with the smallest volume; in fact this *extremal outer ellipsoid* of Q is uniquely defined, but we are not interested in the uniqueness issues. The average diameter of this extremal outer ellipsoid is certain function of Q, let it be denoted by $\mathrm{EllOut}(Q)$ and called the *outer ellipsoidal size*:

$$\mathrm{EllOut}(Q) = \{\min\{\mathrm{Vol}_n(W) \mid W \text{ is an ellipsoid containing } Q\}\}^{1/n}.$$

It is immediately seen that the introduced function is a size, i.e., it is positive, monotone with respect to inclusions and homogeneous with respect to similarity transformations of the homogeneity degree 1.

We need the following simple lemma

Lemma 3.2.1 *Let $n > 1$, let*

$$W = \{x = Bu + c \mid u^T u \le 1\}$$

be an ellipsoid in $\mathbf{R}^n$, and let

$$\bar{W} = \{x \in W \mid (x - c)^T q \le 0\}$$

be a "half-ellipsoid" - the intersection of W and a half-space with the boundary hyperplane passing through the center of W (here $q \neq 0$). Then $\bar{W}$ can be covered by an ellipsoid W^+ of the volume

$$\mathrm{Vol}_n(W^+) = \kappa^n(n)\mathrm{Vol}_n(W),$$

$$\kappa^n(n) = \frac{n^2}{n^2-1}\sqrt{\frac{n-1}{n+1}} \le \exp\{-\frac{1}{2(n-1)}\}; \tag{3.2.5}$$

in particular,

$$\mathrm{EllOut}(W^+) \le \kappa(n)\,\mathrm{EllOut}(W) \le \exp\{-\frac{1}{2n(n-1)}\}\,\mathrm{EllOut}(W). \tag{3.2.6}$$

The ellipsoid W^+ is given by

$$W^+ = \{x = B^+u + c^+ \mid u^T u \le 1\},$$

$$B^+ = \alpha(n)B - \gamma(n)(Bp)p^T, \quad c^+ = c - \frac{1}{n+1}Bp,$$

where

$$\alpha(n) = \left\{\frac{n^2}{n^2-1}\right\}^{1/4}, \quad \gamma(n) = \alpha(n)\sqrt{\frac{n-1}{n+1}}, \quad p = \frac{B^T q}{\sqrt{q^T BB^T q}}.$$

To prove the lemma, it suffices to reduce the situation to the similar one with W being the unit Euclidean ball V; indeed, since W is the image of V under the affine transformation $u \mapsto Bu+c$, the half-ellipsoid $\bar{W}$ is the image, under this transformation, of the half-ball

$$\bar{V} = \{u \in V \mid (B^T q)^T u \leq 0\} = \{u \in V \mid p^T u \leq 0\}.$$

Now, it is quite straightforward to verify that a half-ball indeed can be covered by an ellipsoid V^+ with the volume being the required fraction of the volume of V; to verify this, it was one of the exercises distributed last time, and in the formulation of the exercise you were given the explicit representation of V^+. It remains to note that the image of V^+ under the affine transformation which maps the unit ball V onto the ellipsoid W is an ellipsoid which clearly contains the half-ellipsoid $\bar{W}$ and is in the same ratio of volumes with respect to W as V^+ is with respect to the unit ball V (since the ratio of volumes remains invariant under affine transformations). The ellipsoid W^+ given in formulation of the lemma is nothing but the image of V^+ under our affine transformation.

3.2.2 The Ellipsoid method

Bearing in mind our "spoiled" cutting plane scheme, we may interpret the statement of Lemma 3.2.1 as follows: assume that at certain step of a cutting plane method the localizer G_{i-1} to be updated is an ellipsoid. Let us choose the current search point as the center of the ellipsoid; then the cut will result in the intermediate localizer $\bar{G}_i$ which is a half-ellipsoid. Let us cover this intermediate localizer by an ellipsoid given by our lemma and choose the latter ellipsoid as the new localizer G_i. Now we are in the same position as we were - the new localizer is an ellipsoid, and we may proceed in the same manner. And due to our lemma, we do decrease in certain ratio certain size of localizers - namely, the outer ellipsoidal size EllOut.

Two issues should be thought of. First, how to initialize our procedure - i.e., how to enforce the initial localizer to be an ellipsoid (recall that in our basic scheme the initial localizer was the domain G of the problem). The answer is immediate - let us take as G_0 an arbitrary ellipsoid containing G. The second difficulty is as follows: in our "spoiled" cutting plane scheme the localizers, generally speaking, are not subsets of the domain of the problem; it may, consequently, happen that the center of a localizer is outside the domain; how to perform a cut in the latter case? Here again the difficulty can be immediately overcome. If the center x_i of the current localizer G_{i-1} is outside the interior of G, then, due to the Separation Theorem for convex sets, we may find a nonzero linear functional $e^T x$ which separates x_i and int G:

$$(x - x_i)^T e \leq 0,\ x \in G.$$

Using e for the cut, i.e., setting

$$\bar{G}_i = \{x \in G_{i-1} \mid (x - x_i)^T e \leq 0\},$$

we remove from the previous localizer G_{i-1} only those points which do not belong to the domain of the problem, so that $\bar{G}_i$ indeed can be thought of as a new intermediate localizer.

Thus, we come to the *Ellipsoid method* which, as applied to a convex programming problem

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

works as follows:

Initialization. *Choose $n \times n$ nonsingular matrix B_0 and a point x_1 such that the ellipsoid*

$$G_0 = \{x = B_0 u + x_1 \mid u^T u \leq 1\}$$

contains G. Choose $\beta > 0$ such that

$$\beta \leq \frac{\mathrm{EllOut}(G)}{\mathrm{EllOut}(G_0)}.$$

i-th step, $i \geq 1$. Given B_{i-1}, x_i, act as follows:

1) *Check whether $x_i \in \mathrm{int}\, G$. If it is not the case, then call step i non-productive, find a nonzero e_i such that*

$$(x - x_i)^T e_i \leq 0 \ \forall x \in G$$

and go to 3), otherwise go to 2).

2) *Call the oracle to compute the quantities*

$$f(x_i), f'(x_i), g_1(x_i), g_1'(x_i), ..., g_m(x_i), g_m'(x_i).$$

If one of the inequalities

$$g_j(x_i) \leq \varepsilon \left(\max_G \{g_j(x_i) + (x - x_i)^T g_j'(x_i)\} \right)_+, \ j = 1, ..., m \tag{3.2.7}$$

is violated, say, k-th of them, call i-th step non-productive, set

$$e_i = g_k'(x_i)$$

and go to 3).

If all inequalities (3.2.7) are satisfied, call i-th step productive and set

$$e_i = f'(x_i).$$

If $e_i = 0$, terminate, x_i being the result found by the method, otherwise go to 3).

3) *Set*

$$p = \frac{B_{i-1}^T e_i}{\sqrt{e_i^T B_{i-1} B_{i-1}^T e_i}}, \ B_i = \alpha(n) B_{i-1} - \gamma(n)(B_{i-1} p) p^T, \ x_{i+1} = x_i - \frac{1}{n+1} B_{i-1} p, \tag{3.2.8}$$

$\alpha(n)$ and $\gamma(n)$ being the quantities from Lemma 3.2.1.

If

$$\kappa^i(n) < \varepsilon\beta, \tag{3.2.9}$$

terminate with the result $\bar{x}$ being the best, with respect to the objective, of the search points associated with the productive steps:

$$\bar{x} \in \mathrm{Argmin}\{f(x) \mid \ x \text{ is one of } x_j \text{ with productive } j \leq i\}$$

otherwise go to the next step.

The main result on the Ellipsoid method is as follows:

Theorem 3.2.1 *The associated with a given relative accuracy $\varepsilon \in (0,1)$ Ellipsoid method $Ell(\varepsilon)$, as applied to a problem instance $p \in \mathcal{P}_m(G)$, terminates in no more than*

$$\mathrm{Compl}(Ell(\varepsilon)) = \rfloor \frac{\ln\left(\frac{1}{\beta\varepsilon}\right)}{\ln(1/\kappa(n))} \lfloor \leq \rfloor 2n(n-1)\ln\left(\frac{1}{\beta\varepsilon}\right) \lfloor$$

steps and solves p within relative accuracy ε: the result $\bar{x}$ is well defined and

$$\varepsilon(p,\bar{x}) \leq \varepsilon.$$

Given the direction e_i defining i-th cut, it takes $O(n^2)$ arithmetic operations to update (B_{i-1}, x_i) into (B_i, x_{i+1}).

Proof. The complexity bound is an immediate corollary of the termination test (3.2.9). To prove that the method solves p within relative accuracy ε, note that from Lemma 3.2.1 it follows that

$$\mathrm{EllOut}(G_i) \leq \kappa^i(n)\,\mathrm{EllOut}(G_0) \leq \kappa^i(n)\beta^{-1}\,\mathrm{EllOut}(G)$$

(the latter inequality comes from the origin of β). It follows that if the method terminates at a step N due to (3.2.9), then

$$\mathrm{EllOut}(G_N) < \varepsilon\,\mathrm{EllOut}(G).$$

Due to this latter inequality, we immediately obtain the accuracy estimate as a corollary of our general convergence statement on the cutting plane scheme (Proposition 2.3.1). Although the latter statement was formulated and proved for the basic cutting plane scheme rather than for the "spoiled" one, the reasoning can be literally repeated in the case of the "spoiled" scheme. ∎

Note that the complexity of the Ellipsoid method depends on β, i.e., on how good is the initial ellipsoidal localizer we start with. Theoretically, we could choose as G_0 the ellipsoid of the smallest volume containing the domain G of the problem, thus ensuring $\beta = 1$; for "simple" domains, like a box, a simplex or a Euclidean ball, we may start with this optimal ellipsoid not only in theory, but also in practice. Even with this good start, the Ellipsoid method has $O(n)$ times worse theoretical complexity than the Center of Gravity method (here it takes $O(n^2)$ steps to improve inaccuracy by an absolute constant factor). As a compensation of this theoretical drawback, the Ellipsoid method is not only of theoretical interest, it can be used for practical computations as well. Indeed, if G is a simple domain from the above list, then all actions prescribed by rules 1)-3) cost only $O(n(m+n))$ arithmetic operations. Here the term mn comes from the necessity to check whether the current search point is in the interior of G and, if it is not the case, to separate the point from G, and also from the necessity to maximize the linear approximations of the constraints over G; the term n^2 reflects complexity of updating $B_{i-1} \mapsto B_i$ after e_i is found. Thus, the arithmetic cost of a step is quite moderate, incomparably to the tremendous one for the Center of Gravity method.

3.3 Exercises: The Center of Gravity and the Ellipsoid methods

3.3.1 Is it actually difficult to find the center of gravity?

I cannot say that it is *proved* that, given a polytope in $\mathbf{R}^n$ defined by m linear inequalities, one cannot find the center of gravity of the polytope at a "moderate" (polynomial in n and m) # of

arithmetic operations. What can be said is that efficient, in the indicated sense, *deterministic* algorithms are unknown, and there are strong reasons to doubt whether such an algorithm exists. Randomized algorithms, this is another story: it was proved recently that it is possible to approximate the center of gravity of a polytope (within accuracy sufficient for our goals) by certain Monte Carlo algorithm at a polynomial in n, m arithmetic cost. Nevertheless, I have not heard much about applications of these new tools in optimization, since the theoretical complexity of the algorithms, although polynomial, seems to be inappropriately large. There is, anyhow, hope that practical behaviour of the Monte Carlo routine will be much better than that one prescribed by the theoretical analysis, so that the approach seems to be worthy of testing and you are welcome to do it. The idea of the Monte Carlo routine will be explained in a while, but let me start with the deterministic approach.

The most natural (and actually the only one which comes to my mind) way to compute the center of gravity of a polytope $P \subset \mathbf{R}^n$ is to partition P into simplexes S_i, $i = 1, ..., n$, with no simplex in the partitioning covering an interior point of another simplex, and to compute the centers of gravity x_i and the volumes v_i of the simplexes; as we already know (Exercise 1.3.5), the center of gravity of a simplex

$$S = \mathrm{Conv}\{w_0, ..., w_n\}$$

is the arithmetic mean of the vertices $w_0, ..., w_n$, while the volume of S is

$$\frac{1}{n!}|\operatorname{Det}(w_1 - w_0, w_2 - w_0, ..., w_n - w_0)|$$

(why?) Thus, given the partitioning, we can easily compute x_i and v_i and then set

$$x^*(P) = \{\sum_i v_i\}^{-1} \sum_i v_i x_i. \tag{3.3.1}$$

Exercise 3.3.1 $^{\#}$ *Prove (3.3.1).*

Now, there is no theoretical difficulty to partition a polytope into simplexes. Indeed, let us use induction in dimension: after all $(n-1)$-dimensional facets of the polytope are partitioned into $(n-1)$-dimensional simplexes, let us choose a vertex w in the interior of P; the conic sets with the bases being $(n-1)$-dimensional simplexes coming from the facets and the vertex at w clearly form a partitioning of P. Of course, this is a tremendous work (think of partitioning in this manner an n-dimensional cube; you will get $2^n n!$ simplexes!). But may be this is a bad way to partition a polytope and there are much better ones? The answer is negative:

Exercise 3.3.2 $^{\#*}$ *Let $P \subset \mathbf{R}^n$ be a polytope contained in the unit cube $\{|x|_\infty \leq 1\}$ and containing, for some $\alpha > 0$, a cube of the type $\{|x - a|_\infty \leq \alpha\}$. Prove that any partitioning of P into non-overlapping simplexes contains at least*

$$(1 + o(1)) \left(\frac{\sqrt{n}\alpha}{\mathrm{e}}\right)^n \sqrt{2\pi n}$$

elements, with $o(1) \to 0$ as $n \to \infty$. Thus, if $\alpha\sqrt{n}$ is not small, say, is ≥ 10, then the partitioning contains at least $\exp\{n\}$ simplexes.

I have no idea whether the lower bound indicated in the latter exercise is "tight", but already this bound says that simplicial partitions are of no use for computing the center of gravity of a polytope.

Now let me present you the Monte Carlo scheme. To compute the center of gravity means, by definition, to compute certain n-dimensional integral. It is quite traditional to compute an integral

$$I = \frac{1}{\mathrm{Vol}\, G} \int_G \phi(x) dx$$

via the Monte Carlo simulation: given a generator of a random vector ξ uniformly distributed in G, one can generate a sample $\xi_1, ..., \xi_N$ of independent vectors with the indicated distribution and approximate the integral (which is nothing but the expectation of the random variable $\phi(\xi)$) by the empirical average

$$I_N = \frac{1}{N} \sum_{i=1}^{N} \phi(\xi_i).$$

In the case we are interested in $\phi(x) = x$ is a vector-valued function; due to the Tschebyshev equality for random vectors, the expected squared $|\cdot|_2$-error of the approximation is given by

$$\sigma_N^2 \equiv \mathcal{E}\{|I - I_N|_2^2\} = \frac{\sigma^2}{N},$$

where

$$\sigma^2 = \mathcal{E}\{|I - I_1|^2\} = \frac{1}{\mathrm{Vol}\, G} \int_G |x - x^*(G)|_2^2 dx.$$

The rate of convergence seems to be quite reasonable: it takes $O(\sigma^2/\varepsilon^2)$ calls for the generator to approximate $x^*(G)$ within accuracy of order of ε. Of course, σ might be large (imagine that G is "long and thin"), but this is an illusion: an n-dimensional solid in $\mathbf{R}^n$ can be "large and thin" only with respect to a badly chosen Euclidean metric; for a properly chosen metric, as we know from Exercise 2.4.7, the "asphericity" of G, i.e., the ratio of the radii of a pair of concentric balls, with the smaller one contained in G and the larger one containing G, does not exceed n. It is easily seen that with respect to this metric one has $\sigma^2 \leq n^2$, and, by the way, this is exactly the metric in which we actually should measure inaccuracy when approximating the center of gravity. Of course, the "proper" metric is not given in advance, but we are not interested in it: the Monte Carlo procedure is independent of any metric, the latter is involved only in the convergence analysis.

Thus, everything looks fine - everything but one point: how to generate the uniformly distributed in G random vector. An immediate idea is: let us embed G into a box G^+. There is no difficulty in generating a random vector uniformly distributed in a box (you should call n times the standard generator of a uniformly distributed in $[0, 1]$ random variable and make the outputs of the generator, after appropriate scalings, the coordinates of the random vector we need). Now, when asked to generate a random vector from G, let us generate sequentially random vectors uniformly distributed in G^+ until a vector belonging to G is met; this vector will be the output of our new generator and will, of course, be distributed uniformly in G, as required. Are we done? No! Indeed, let us look how long it takes to meet a vector from G when generating a sample of vectors uniformly distributed in the box G^+. The answer is immediate: the expected # of elements in the sample is $\mathrm{Vol}(G^+)/\mathrm{Vol}(G)$. In other words, if the linear

dimensions of G are something like 10% of those of G^+ (which is not so small ratio), then the complexity of generating a point from G (# of calls for the inner generator) will be something like 10^n - and for $n = 20$ we shall for sure die before our generator produces a single answer.

Thus, we did not overcome what is called "the damn of the dimension" - we simply have transformed our initial question into that one of how to generate, at a moderate cost, a vector uniformly distributed in G. The transformed question, anyhow, admits an affirmative answer as follows.

Consider a discrete mesh formed by vectors with the coordinates being multiples of certain small positive ν, and let Γ be the part of the mesh comprised of vectors belonging to G. Let $x_0 \in \Gamma$ be once for ever fixed. Consider the random walk along Γ described as follows: the moving point, initially placed at x_0, at an integer moment t chooses randomly one of the $2n$ directions $\pm e_i$, $i = 1, ..., n$ (e_i are the standard orths) and tries to make a step of the length ν in the chosen direction. If the step keeps the point within G, then it actually is performed; otherwise the point does not move at the moment t and repeats the attempt to move at the next moment, $t + 1$, again choosing randomly the direction of the step.

It is clear that the described process (which can be easily simulated) moves the point along the connectedness component, Γ_0, of the point x_0 in the mesh Γ. This random walk is certain Markov's chain, and since the probability to come in a finitely many steps N (N is equal to the largest length of a simple path in Γ_0) from any given point of Γ_0 to any other point of Γ_0 is positive, there exists a unique steady-state probability distribution π of positions of the point, an the distribution π_t of the position of the point at a moment t converges, as $t \to \infty$, to π.

Exercise 3.3.3 * *Prove that π is nothing but the uniform distribution on Γ_0*

Thus, the steady-state distribution of the position of our walking point is uniform on Γ_0, and the distributions π_t, which we actually can simulate, for large t are almost uniform (since they converge to π as $t \to \infty$). Thus, choosing in advance certain "large" T and simulating T-step trajectory of the point, we can generate an "almost uniformly distributed in Γ_0" (and therefore "almost uniformly distributed in G", provided that ν is small) vector.

Of course, the main issue here is how large should be our "large" T - why could not we again come to something like $\exp\{O(n)\}$? It turns out, anyhow, that the things are not so bad: it suffices to take T as a *polynomial* of n and $1/\nu$ .This is a relatively recent and actually brilliant result of Cobham, Dyer, Frieze; the proof heavily exploits, via the isoperimeter inequalities, the convexity of G (as it should be: the same approach to generating vectors uniformly distributed in a "slightly nonconvex" G immediately results in an exponential in n value of T). It remains to note that it is not quite straightforward to apply this scheme to the problem of approximating the center of gravity of a polytope; to this end one should relate the construction to a properly chosen Euclidean metric (since now a "long and thin" G do causes troubles - it enforces ν to be very small in order for Γ_0 to be a "good" discretization of G) and to choose small (although small "polynomially in n", not exponentially) ν. As a result, T becomes a high-degree polynomial of n, not speaking on preliminary computational effort required to choose a "proper" Euclidean metric underlying the mesh. Let me repeat, anyhow, that from the practical viewpoint these difficulties might be partly imaginary and coming from rough estimates used in the proofs. The scheme of the generator is simple, and it is worthy to test the approach numerically, not bothering much about "bad" theoretical values of T and ν.

3.3.2 Some extensions of the Cutting Plane scheme

To the moment we have applied the Cutting Plane scheme to convex optimization problems in the standard form

$$\textit{minimize} \;\; f(x) \;\; \textit{s.t.} \quad g_j(x) \leq 0,\, j = 1, ..., m,\, x \in G \subset \mathbf{R}^n \tag{3.3.2}$$

(G is a *solid*, i.e., a closed and bounded convex set with a nonempty interior, f and g_j are convex and continuous on G). In fact the scheme has a wider field of applications. Namely, consider a "generic" problem as follows:

$$(f): \quad \textit{minimize} \;\; f(x) \;\; \textit{s.t.} \;\; x \in G_f \subset \mathbf{R}^n; \tag{3.3.3}$$

here G_f is certain (specific for the problem instance) solid and f is a function taking values in the extended real axis $\mathbf{R} \cup \{-\infty\} \cup \{+\infty\}$ and finite on the interior of G.

Let us make the following assumption on our abilities to get information on (f):
(A): *we have an access to an oracle $\mathcal{A}$ which, given on input a point $x \in \mathbf{R}^n$, informs us whether x belongs to the interior of G_f; if it is not the case, the oracle reports a nonzero functional e_x which separates x and G_f, i.e., is such that*

$$(y - x)^T e_x \leq 0, \;\; y \in G_f;$$

if $x \in \text{int}\, G_f$, then the oracle reports $f(x)$ and a functional e_x such that the Lebesque set

$$L_f^*(x) = \{y \in G_f \mid \; f(y) < f(x)\}$$

is contained in the open half-space

$$\{y \in \mathbf{R}^n \mid \; (y - x)^T e_x < 0\}.$$

In the mean time we shall see that under assumption (A) we can efficiently solve (f) by cutting plane methods; but before coming to this main issue let me indicate some interesting particular cases of the "generic" problem (f).

Example 1. Convex optimization in the standard setting. Consider the usual convex problem (3.3.2) and assume that the feasible set of the problem has a nonempty interior where the constraints are negative. We can rewrite (3.3.2) as (3.3.3) by setting

$$G_f = \{x \in G \mid \; g_j(x) \leq 0,\, j = 1, ..., m\}.$$

Now, given in advance the domain G of the problem (3.3.2) and being allowed to use a first-order oracle $\mathcal{O}$ for (3.3.2), we clearly can imitate the oracle $\mathcal{A}$ required by (A).

Example 1 is not so interesting - we simply have expressed something we already know in a slightly different way. The next examples are less trivial.

Example 2. Quasiconvex optimization. Assume that the objective f involved into (3.3.2) and the functional constraints g_j are quasiconvex rather than convex; recall that a function h defined on a convex set Q is called *quasiconvex*, if the sets

$$L_h(x) = \{y \in Q \mid \; h(y) \leq h(x)\}$$

are convex whenever $x \in Q$ (note the difference between $L_h(x)$ and $L_h^*(x)$; the latter set is defined via strict inequality $<$, the former one via $\leq$). Besides quasiconvexity, assume that the functions f, g_j, $j = 1, ..., m$, are *regular*, i.e., continuous on G, differentiable on the interior of G with a nonzero gradient at any point which is not a minimizer of the function over G. Last, assume that the feasible set G_f of the problem possesses a nonempty interior and that the constraints are negative on the interior of G_f.

If h is a regular quasiconvex function on G and $x \in \text{int } G$, then the set $L_h^*(x) = \{y \in G \mid h(y) < h(x)\}$ belongs to the half-space

$$\Pi_h(x) = \{y \in G \mid (y - x)^T h'(x) < 0\}.$$

Exercise 3.3.4 $^{\#+}$ *Prove the latter statement.*

Thus, in the case in question the sets $\{y \in G \mid f(y) < f(x)\}$, $\{y \in G \mid g_j(y) < g_j(x)\}$ are contained in the half-spaces $\{y \mid (y - x)^T f'(x) < 0\}$, $\{y \mid (y - x)^T g_j'(x) < 0\}$, respectively. It follows that in the case in question, same as in the convex case, we, given an access to a first-order oracle for (3.3.2), can imitate a required by (A) oracle $\mathcal{A}$ for the induced problem (3.3.3).

Example 3. Linear-fractional programming. Consider the problem

$$\textit{minimize } f(x) = \max_{\omega \in \Omega} \left\{ \frac{a_\omega(x)}{b_\omega(x)} \right\} \textit{ s.t. } g_j(x) \leq 0,\ j = 1, ..., m,\ b_\omega(x) > 0,\ \omega \in \Omega,\ x \in G; \quad (3.3.4)$$

here G is a solid in $\mathbf{R}^n$ and Ω is a finite set of indices, a_ω and b_ω are affine functions and g_j are, say, convex and continuous on G. The problem is, as we see, to minimize the maximum of ratios of given linear forms over the convex set defined by the inclusion $x \in G$, convex functional constraints $g_j(x) \leq 0$ and additional linear constraints expressing positivity of the denominators.

Let us set

$$G_f = \{x \in G \mid g_j(x) \leq 0, b_\omega(x) \geq 0, \omega \in \Omega\};$$

we assume that the (closed and convex) set G_f possesses a nonempty interior and that the functions g_j are negative, while b_ω are positive on the interior of G_f.

By setting

$$f(x) = \begin{cases} \max_\omega \{a_\omega(x)/b_\omega(x)\} & x \in \text{int } G_f \\ +\infty, & \textit{otherwise} \end{cases} \quad (3.3.5)$$

we can rewrite our problem as

$$\textit{minimize } \ f(x) \ \textit{ s.t. } \ x \in G_f. \quad (3.3.6)$$

Now, assume that we are given G in advance and have an access to a first-order oracle $\mathcal{O}$ which, given on input a point $x \in \text{int } G$, reports the values and subgradients of functional constraints at x, same as reports all $a_\omega(\cdot)$, $b_\omega(\cdot)$.

Under this assumptions we can imitate for (3.3.6) the oracle $\mathcal{A}$ required by the assumption (A). Indeed, given $x \in \mathbf{R}^n$, we first check whether $x \in \text{int } G$, and if it is not the case, find a nonzero functional e_x which separates x and G (we can do it, since G is known in advance); of course, this functional separates also x and G_f, as required in (A). Now, if $x \in \text{int } G$, we ask the

first-order oracle $\mathcal{O}$ about the values and subgradients of g_j, a_ω and b_ω at x and check whether all g_j are negative at x and all $b_\omega(x)$ are positive. If it is not the case and, say, $g_k(x) \geq 0$, we claim that $x \notin \text{int } G_f$ and set e_x equal to $g_k'(x)$; this functional is nonzero (since otherwise g_k would attain a nonnegative minimum at x, which contradicts our assumptions about the problem) and clearly separates x and G_f (due to the convexity of g_k). Similarly, if one of the denominators $b_{\bar\omega}$ is nonpositive at x, we claim that $x \notin \text{int } G_f$ and set

$$e_x = \nabla b_{\bar\omega};$$

here again e_x is nonzero and separates x and G_f (why?)

Last, if $x \in \text{int } G$, all g_j are negative at x and all b_ω are positive at the point, we claim that $x \in \text{int } G_f$ (as it actually is the case), find the index $\omega = \omega(x)$ associated with the largest at x of the fractions $a_\omega(\cdot)/b_\omega(\cdot)$, compute the corresponding fraction at x (this is nothing but $f(x)$) and set

$$e_x = \nabla_y a_{\omega(x)}(y) - f(x)\nabla_y b_\omega(x)(y).$$

Since in the latter case we have

$$L_f^*(x) \subset \{y \in G_f \mid \frac{a_{\omega(x)}(y)}{b_{\omega(x)}(y)} < f(x)\} \subset \{y \in G_f \mid (y-x)^T e_x < 0\},$$

we do fit the requirements imposed by (A).

Note that the problem of the type (3.3.4) arises, e.g., in the famous *von Neumann Model of Economy Growth* which is as follows. Consider an economy where m kinds of goods are circulating. The economy is described by a pair of $m \times n$ matrices A and B with positive entries, where row index i stands for goods and column index j stands for "processes". A process j takes, as input, a_{ij} units of good i and produces, as output, b_{ij} units of the same good, per year. Now, let x^t be an n-dimensional vector with coordinates x_j^t being the "intensities" by which we let j-th process work in year t. Then the amount of goods consumed by all the processes run in year t is represented by the vector Ax^t, and the amount of goods produced in the same year is given by the vector Bx^t. If we have no external sources of goods, then the "trajectory" of our economy should satisfy the inequalities

$$Ax^{t+1} \leq Bx^t, \ t = 0, 1, ...$$

(for $t = 0$ the right hand side should be replaced by a positive vector representing the "starting" amount of goods). Now, in the von Neumann Economic Growth problem it is asked what is the largest growth factor, γ^*, for which there exists a "semi-stationary growth trajectory", i.e., a trajectory of the type $x^t = \gamma^t x^0$. In other words, we should solve the problem

$$\textit{maximize } \gamma \ \textit{ s.t. } \gamma Ax \leq Bx \ \textit{ for some nonzero } x \geq 0.$$

Without loss of generality, x in the above formulation can be taken as a point form the standard simplex

$$G = \{x \in \mathbf{R}^n \mid x \geq 0, \sum_j x_j = 1\}$$

(which should be regarded as a solid in its affine hull). It is clearly seen that the problem in question can be rewritten as follows:

$$\textit{minimize} \quad \max_{i=1,...,m} \frac{\sum_j a_{ij}x_j}{\sum_j b_{ij}x_j} \quad s.t. \quad x \in G; \tag{3.3.7}$$

this is a problem of the type (3.3.6).

It is worthy to note that the von Neumann growth factor γ^* describes, in a sense, the highest rate of growth of our economy (this is far from being clear in advance: why the "Soviet" proportional growth is the best one? Why could we not get something better along an oscillating trajectory?) The exact statement on optimality of the von Neumann semi-stationary trajectory (or, better to say, the simplest one of these statements) is as follows:

Proposition 3.3.1 *Let $\{x^t\}_{t=1}^T$ be a trajectory of our economy, so that x^t are nonnegative, $x^0 \neq 0$ and*

$$Ax^{t+1} \leq Bx^t,\ t = 0, 1, ..., T-1.$$

Assume that $x^T \geq \lambda^T x^0$ for some positive λ (so that our trajectory results, for some T, in growth of the amount of goods in λ^T times in T years). Then $\lambda \leq \gamma^*$.

Note that following the von Neumann trajectory

$$x^t = (\gamma^*)^T x^0,$$

x^0 being the x-component of an optimal solution to (3.3.7), does ensure growth by factor $(\gamma^*)^T$ each T years.

Exercise 3.3.5 * *Prove Proposition 3.3.1.*

Example 4. Generalized Eigenvalue problem. Let us again look at problem (3.3.4). What happens if the index set Ω becomes infinite? To avoid minor elaborations, let us assume (as it actually is the case in all applications known to me) that Ω is a compact set and that the functions $a_\omega(x)$, $b_\omega(x)$ are continuous in (x, ω) (and, as above, are affine in x). As far as reformulation (3.3.5) - (3.3.6) is concerned, no difficulties occur. The possibility to imitate the oracle $\mathcal{A}$, this is another story (it hardly would be realistic to ask $\mathcal{O}$ to report infinitely many numerators and denominators). Note, anyhow, that from the discussion accompanying the previous example it is immediately seen that at a given x we are not interested in *all* a_ω, b_ω; what in fact we are interested in are the *active* at x numerator and denominator, i.e., either (any) $b_\omega(\cdot)$ such that $b_\omega(x) \leq 0$, if such a denominator exists, or those $a_\omega(\cdot)$, $b_\omega(\cdot)$ with the largest ratio at the point x, if all the denominators at x are positive. Assume therefore that we know G in advance and have an access to an oracle $\mathcal{O}$ which, given on input $x \in \text{int } G$, reports the values and subgradients of the functional constraints g_j at x and, besides this, tells us (at least in the case when all g_j are negative at x) whether all the denominators $b_\omega(x)$, $\omega \in \Omega$, are positive; if it is the case, then the oracle returns the numerator and the denominator with the largest ratio at x, otherwise returns the denominator which is nonpositive at x. Looking at the construction of $\mathcal{A}$ given in the previous example we immediately conclude that in our now situation we again can imitate a compatible with (A) oracle $\mathcal{A}$ for problem (3.3.6).

In fact the "semiinfinite fractional problem" we are discussing possesses interesting applications; let me introduce one of them which is of extreme importance for modern Control Theory

- the *Generalized Eigenvalue problem*. The problem is as follows: given two affine functions, $A(x)$ and $B(x)$, taking values in the space of symmetric $m \times m$ matrices ("affine" means that the entries of the matrices are affine functions of x), minimize, with respect to x, the Rayleigh ratio

$$\max_{\omega \in \mathbf{R}^m \setminus \{0\}} \frac{\omega^T A(x)\omega}{\omega^T B(x)\omega}$$

of the quadratic forms associated with these matrices under the constraints that $B(x)$ is positive definite (and, possibly, under additional convex constraints on x). In other words, we are looking for a pair (x, λ) satisfying the constraints

$$B(x) \ \textit{is positive definite} \ , \lambda B(x) - A(x) \ \textit{is positive semidefinite}$$

and additional constraints

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

(g_j are convex and continuous on the solid G) and are interested in the pair of this type with the smallest possible λ.

The Generalized Eigenvalue problem (the origin of the name is that in the particular case when $B(x) \equiv I$ is the unit matrix we come to the problem of minimizing, with respect to x, the largest eigenvalue of $A(x)$) can be immediately written down as a "semiinfinite fractional problem"

$$\textit{minimize} \ \max_{\omega \in \Omega} \frac{\omega^T A(x)\omega}{\omega^T B(x)\omega} \ \textit{s.t.}\ g_j(x) \le 0,\ j = 1, ..., m,\ \omega^T B(x)\omega > 0,\ \omega \in \Omega,\ x \in G; \qquad (3.3.8)$$

here Ω is the unit sphere in $\mathbf{R}^m$. Note that the numerators and denominators in our "objective fractions" are affine in x, as required by our general assumptions on fractional problems.

Assume that we are given G in advance, same as the data identifying the affine in x matrix-valued functions $A(x)$ and $B(x)$, and let we have an access to a first-order oracle providing us with local information on the "general type" convex constraints g_j. Then it is not difficult to decide, for a given x, whether $B(x)$ is positive definite, and if it is not the case, to find $\omega \in \Omega$ such that the denominator $\omega^T B(x)\omega$ is nonpositive at x. Indeed, it suffices to compute $B(x)$ and to subject the matrix to the Choleski factorization (I hope you know what it means). If factorization is successful, we find a lower-triangular matrix Q with nonzero diagonal such that

$$B(x) = QQ^T,$$

and $B(x)$ is positive definite; if the factorization fails, then in course of it we automatically meet a unit vector ω which "proves" that $B(x)$ is not a positive semidefinite, i.e., is such that $\omega^T B(x)\omega \le 0$. Now, if $B(x)$, for a given x, is positive semidefinite, then to find ω associated with the largest at x of the fractions

$$\frac{\omega^T A(x)\omega}{\omega^T B(x)\omega}$$

is the same as to find the eigenvector of the (symmetric) matrix $Q^{-1}A(x)(Q^T)^{-1}$ associated with the largest eigenvalue of the matrix, Q being the above Choleski factor of $B(x)$ (why?); to find this eigenvector, this is a standard Linear Algebra routine.

Thus, any technique which allows to solve (f) under assumption (A) immediately implies a numerical method for solving the Generalized Eigenvalue problem.

It is worthy to explain what is the "control source" of Generalized Eigenvalue problems. Let me start with the well-known issue - stability of a linear differential equation

$$z'(t) = \alpha z(t)$$

($z \in \mathbf{R}^s$). As you for sure know, the maximal "growth" of the trajectories of the equation as $t \to \infty$ is predetermined by the eigenvalue of α with the largest real part, let this part be λ; namely, all the trajectories admit, for any $\epsilon > 0$, the estimate

$$|z(t)|_2 \le C_\epsilon \exp\{(\lambda + \epsilon)t\}|z(0)|_2,$$

and vice versa: from the fact that all the trajectories admit an estimate

$$|z(t)|_2 \le C \exp\{at\}|z(0)|_2 \tag{3.3.9}$$

it follows that $a \ge \lambda$.

There are different ways to prove the above fundamental Lyapunov Theorem, and one of the simplest is via quadratic Lyapunov functions. Let us say that a quadratic function $z^T Lz$ (L is a symmetric positive definite $s \times s$ matrix) *proves that the decay rate of the trajectories is at most* a, if for *any* trajectory of the equation one has

$$z^T(t)Lz'(t) \le az^T(t)Lz(t),\ t \ge 0; \tag{3.3.10}$$

if it is the case, then of course

$$\frac{d}{dt} \ln \left(z^T(t)Lz(t) \right) \le 2a$$

and, consequently,

$$\left(z^T(t)Lz(t) \right)^{1/2} \le \exp\{at\} \left(z^T(0)Lz(0) \right)^{1/2},$$

which immediately results in an estimate of the type (3.3.9). Thus, any positive definite symmetric matrix L which satisfies, for some a, relation (3.3.10) implies an upper bound (3.3.9) on the trajectories of the equation, the upper bound involving just this a. Now, what does it mean that L satisfies (3.3.10)? Since $z'(t) = \alpha z(t)$, it means exactly that

$$z^T(t)L\alpha z(T) \equiv \frac{1}{2} z^T(t)(L\alpha + \alpha^T L)z(t) \le az^T(t)Lz(t)$$

for all t and all trajectories of the equation; since $z(t)$ can be an arbitrary vector of $\mathbf{R}^s$, the latter inequality means that

$$2aL - (\alpha^T L + L\alpha) \ \ \textit{is positive semidefinite.} \tag{3.3.11}$$

Thus, any pair comprised of a real a and a positive definite symmetric L satisfying (3.3.11) results in upper bound (3.3.9); the best (with the smallest possible a) bound (3.3.9) which can be obtained on this way is given by the solution to the problem

$$\textit{minimize } a \textit{ s.t. } L \textit{ is positive definite, } 2aL - (\alpha^T L - L\alpha) \textit{ is positive definite};$$

this is nothing but the Generalized Eigenvalue problem with $B(L) = 2L$, $A(L) = \alpha^T L + L\alpha$ and no additional constraints on $x \equiv L$. And it can be proved that the best a given by this construction is nothing but the largest of the real parts of the eigenvalues of α, so that in the case in question the approach based on quadratic Lyapunov functions and Generalized Eigenvalue problems results in complete description of the behaviour of the trajectories as $t \to \infty$.

In fact, of course, what was said is of no "literal" significance: what for should we solve a Generalized Eigenvalue problem in order to find something which can be found by a direct computation of the eigenvalues of α? The indicated approach becomes meaningful when we come from our simple case of a linear differential equation with constant coefficients to a much more difficult (and more important for practice) case of a *differential inclusion*. Namely, assume that we are given a multivalued mapping $z \mapsto Q(z) \subset \mathbf{R}^s$ and are interested in bounding the trajectories of the differential inclusion

$$z'(t) \in Q(z(t)),\ t \geq 0; \tag{3.3.12}$$

such an inclusion may model, e.g., a time-varying dynamic system

$$z'(t) = \alpha(t)z(t)$$

with certain unknown $\alpha(\cdot)$. Assume that we know finitely many matrices α_1,...,α_M such that

$$Q(z) = \mathrm{Conv}\{\alpha_1 z, ..., \alpha_M z\}$$

(e.g., we know bounds on entries of $\alpha(t)$ in the above time-varying system). In order to obtain an estimate of the type (3.3.9), we again may use a quadratic Lyapunov function $z^T Lz$: if for all trajectories of the inclusion one has

$$z(t)Lz'(t) \leq az^T(t)Lz(t),$$

or, which is the same, if

$$z^T Lz' \leq az^T Lz\ \forall (z \in \mathbf{R}^s, z' \in Q(z)) \tag{3.3.13}$$

then, same as above,

$$(z^T(t)Lz(t))^{1/2} \leq \exp\{at\}(z^T(0)Lz(0))^{1/2}$$

for all trajectories.

Now, the left hand side in (3.3.13) is linear in z', and in order for the inequality in (3.3.13) to be satisfied for all $z' \in Q(z)$ it is necessary and sufficient to have $z^T L\alpha_i z \equiv \frac{1}{2}\left(z^T(\alpha_i^T L + L\alpha_i)z\right) \leq z^T Lz$ for all z and all i, i.e., to ensure positive semidefiniteness of the matrices $2aL - (\alpha_i^T L + L\alpha_i)$, $i = 1, ..., M$. By setting

$$B(L) = \mathrm{Diag}(2L, ..., 2L),\ A(L) = \mathrm{Diag}(\alpha_1^T L + L\alpha_1, ..., \alpha_M^T L + L\alpha_M)$$

we convert the problem of finding the best quadratic Lyapunov function (i.e., that one with the best associated decay rate a) into the Generalized Eigenvalue problem

$$\textit{minimize } a \textit{ s.t. } B(L) \textit{ positive definite, } aB(L) - A(L) \textit{ positive semidefinite}$$

with no additional constraints on $x \equiv L$.

Note that in the case of differential inclusions (in contrast to that one of equations with constant coefficients) the best decay rate which can be demonstrated by a quadratic Lyapunov function is not necessarily the actually best decay rate possessed by the trajectories of the inclusion; the trajectories may behave themselves better than it can be demonstrated by a *quadratic* Lyapunov function. This shortcoming, anyhow, is compensated by the fact that the indicated scheme is quite tractable computationally; this is why it becomes now one of the standard tools for stability analysis and synthesis.

It is worthy to add that the von Neumann problem is a very specific case of a Generalized Eigenvalue problem (make the entries of Ax and Bx the diagonal entries of *diagonal* matrices).

I have indicated a number of important applications of (f); it is time now to think how to solve the problem. There is no difficulty in applying to the problem the cutting plane scheme.

> The Cutting Plane scheme for problem (f):
> Initialization. *Choose a solid G_0 which covers G_f.*
> i-th step, $i \geq 1$. *Given a solid G_{i-1} (the previous localizer), choose*
>
> $$x_i \in \text{int } G_i$$
>
> *and call the oracle $\mathcal{A}$, x_i being the input.*
>
> *Given the answer $e_i \equiv e_{x_i}$ of the oracle,*
> *- call step i productive if the oracle says that $x_i \in \text{int } G_f$, and call the step non-productive otherwise;*
> *- check whether $e_i = 0$ (due to (A), this may happen only at a productive step); if it is the case, terminate, x_i being the result found by the method. Otherwise*
>
> *- define i-th approximate solution to (f) as the best, in terms of the values of f, of the search points x_j associated with the productive steps $j \leq i$;*
>
> *- set*
>
> $$\bar{G}_i = \{x \in G_{i-1} \mid (x - x_i)^T e_i \leq 0\};$$
>
> *- embed the intermediate localizer $\bar{G}_i$ into a solid G_i and loop.*

The presented scheme defines, of course, a family of methods rather than a single method. The basic implementation issues, as always, are how to choose x_i in the interior of G_{i-1} and how to extend $\bar{G}_i$ to G_i; here one may use the same tactics as in the Center of Gravity or in the Ellipsoid methods. An additional problem is how to start the process (i.e., how to choose G_0); this issue heavily depends on a priori information on the problem, and here we hardly could do any universal recommendations.

Now, what can be said about the rate of convergence of the method? First of all, we should say how we measure inaccuracy. A convenient general approach here is as follows.

Let $x \in G_f$ and let, for a given $\varepsilon \in (0, 1)$,

$$G_f^\varepsilon = x + \varepsilon(G_f - x) = \{y = (1 - \varepsilon)x + z \mid z \in G_f\}$$

be the image of G_f under the similarity transformation which shrinks G_f to x in $1/\varepsilon$ times. Let

$$f_x(\varepsilon) = \sup_{y \in G_f^\varepsilon} f(y).$$

The introduced quantity depends on x; let us take the infimum of it over $x \in G_f$:

$$f^*(\varepsilon) = \inf_{x \in G_f} f_x(\varepsilon).$$

By definition, an ε-solution to f is any point $x \in G_f$ such that

$$f(x) \leq f^*(\varepsilon).$$

Let us motivate the introduced notion. The actual motivation is, of course, that the notion works, but let us start with a kind of speculation. Assume for a moment that the problem is solvable, and let x^* be an optimal solution to it. One hardly could argue that a point $\bar{x} \in G_f$ which is at the distance of order of ε of x^* is a natural candidate on the role of an ε-solution; since all points from $G_{x^*}^{\varepsilon}$ are at the distance at most $\varepsilon \,\mathrm{Diam}(G_f)$ from x^*, all these points can be regarded as ε-solutions, in particular, the worst of them (i.e., with the largest value of f) point $x^*(\varepsilon)$. Now, what we actually are interested in are the values of the objective; if we agree to think of $x^*(\varepsilon)$ as of an ε-solution, we should agree that any point $x \in G_f$ with $f(x) \leq f(x^*(\varepsilon))$ also is an ε-solution. But this latter property is shared by *any* point which is an ε-solution in the sense of the above definition (look, $f(x^*(\varepsilon))$ is nothing but $f_{x^*}(\varepsilon)$), and we are done - our definition is justified!

Of course, this is nothing but a speculation. What might, and what might not be called a good approximate solution, this cannot be decided in advance; the definition should come from the real-world interpretation of the problem, not from inside the Optimization Theory. What could happen with our definition in the case of a "bad" problem, it can be seen from the following example:

$$\textit{minimize} \quad \frac{x}{10^{-20} + x}, \quad x \in G_f = [0, 1].$$

Here in order to find a solution with the value of the objective better than, say, $1/2$ (note that the optimal value is 0) we should be at the distance of order 10^{-20} of the exact solution $x = 0$. For our toy problem it is immediate, of course, to indicate the solution exactly, but think what happens if the same effect is met in the case of a multidimensional and nonpolyhedral G_f. We should note, anyhow, that the problems like that one just presented are "intrinsically bad"; in "good" situations our definition does work:

Exercise 3.3.6 # *Prove that*

1) if the function f involved into (f) is convex and continuous on G_f, then any ε-solution x to the problem satisfies the inequality

$$f(x) - \min_{G_f} f \leq \varepsilon(\max_{G_f} f - \min_{G_f} f),$$

i.e., solves the problem within relative accuracy ε;

2) if the function f involved into (f) is Lipschitz continuous on G_f with respect to certain norm $|\cdot|$, L_f being the corresponding Lipschitz constant, and if x is a ε-solution to (f), then

$$f(x) - \min_{G_f} f \leq \mathrm{Diam}(G_f) L_f \varepsilon,$$

where Diam *is the diameter of G_f with respect to the norm in question.*

Now - the end of the story.

Exercise 3.3.7 $^{\#*}$ *Prove the following statement:*

let a cutting plane method be applied to (f)*, and let* $\mathrm{Size}(\cdot)$ *be a size. Assume that for certain* N *the method either terminates in course of the first* N *steps, or this is not the case, but* $\mathrm{Size}(G_N)$ *is smaller than* $\mathrm{Size}(G_f)$*. In the first of the cases the result produced by the method is a minimizer of* f *over* G_f*, and in the second case the* N*-th approximate solution* $\bar{x}_N$ *is well-defined an is an* ε'*-solution to* (f) *for any*

$$\varepsilon' > \frac{\mathrm{Size}(G_N)}{\mathrm{Size}(G_f)}.$$

Exercise 3.3.8 $^{\#}$ *Write a code implementing the Ellipsoid version of the Cutting Plane scheme for* (f)*. Use the code to find the best decay rate for the differential inclusion*

$$z'(t) \in Q(z) \subset \mathbf{R}^3,$$

where

$$Q(z) = \mathrm{Conv}\{\alpha_1 z, ..., \alpha_M z\}$$

and α_i*,* $i = 1, 2, ..., M = 2^6 = 64$*, are the vertices of the polytope*

$$P = \{ \begin{pmatrix} 1 & p_{12} & p_{13} \\ p_{21} & 1 & p_{23} \\ p_{31} & p_{32} & 1 \end{pmatrix} \mid |p_{ij}| \leq 0.1 \}.$$

Lecture 4

Polynomial solvability of Linear Programming

This lecture is devoted to one of the most important theoretical applications of the Ellipsoid method, namely, to the proof of polynomial solvability of Linear Programming. Let me first explain what is the problem we are going to discuss.

4.1 Classes P and NP

Consider the following pair of problems:

"Connectedness":

Given a graph Γ, detect whether there exists a path leading from a given vertex a to another given vertex b.

"Hamiltonian cycle":

Given a graph Γ, detect whether there is a path leading from a given vertex a to another given vertex b and visiting each of the remaining vertices exactly once.

Both of the problems belong to a very important world which is called the *class of* NP (Non-deterministic Polynomial) *problems*. The characteristic features of the citizens of this world are as follows:

1) *The data identifying a problem instance (in our example these are the graph Γ along with the pair of vertices marked as a and b) are "completely finite" and can be naturally represented by a finite binary word x.*

Indeed, the data form a finite sequence of integers - you should specify the number n of vertices in the graph, the zero-one entries of the incidence matrix

$$A_{ij} = \begin{cases} 1, & \textit{vertices } i \textit{ and } j \textit{ are linked by an arc} \\ 0, & \textit{otherwise} \end{cases}$$

to indicate what the arcs are, and the numbers i_a and i_b enumerating the vertices a and b. And a finite sequence of integers can, of course, be coded by a finite binary word; it suffices to code binary digits in the integers as $0 \mapsto 00$, $1 \mapsto 11$ and to use, say 01 to separate the subsequent integers in the sequence.

2) *The problem itself is to detect whether in a given set of "candidate" solutions there exists one - the "actual solution" - which satisfies a given condition $\mathcal{P}$. Candidate solutions also admit natural coding by finite binary words y.*
Our problems "Connectedness" and "Hamiltonian cycle" are exactly of this type; here candidate solutions are paths in a graph, i.e., finite sequences of integers, and such sequences, as we know, can be naturally coded by finite binary words.

3) *Given a binary word x specifying the data, i.e., the problem instance, and a binary word y representing a candidate solution, one can easily check whether y actually is a solution to the problem instance x.* More precisely that means that there exists an algorithm M which, given x and y, verifies efficiently whether y indeed satisfies the required condition $\mathcal{P}$ with respect to the data x, i.e., computes the function

$$\mathcal{P}(x,y) = \begin{cases} 0, & y \text{ is an actual solution to } x \\ 1, & \text{otherwise} \end{cases}$$

in no more than polynomial $p(l(x)+l(y))$ number of elementary steps, where $l(u)$ is the binary length of a finite word u; a predicate $P(x,y)$ possessing this property is called *polynomially computable*.
In the above definition one can use any known from Mathematical Logic algorithmic models; all these models lead to the same set of polynomially computable predicates.

Last characteristic property of problems from the class NP is as follows:

4) *When looking for a solution y to a problem instance x, one can in advance restrict himself with "not too long" candidate solutions: the binary length $l(y)$ of any actual solution y to a problem instance x is bounded from above by certain polynomial $q(l(x))$ of the binary length $l(x)$ of the word x coding the instance.*

You can easily verify that the above two problems - "Connectedness" and "Hamiltonian cycle" - indeed belong to the class NP. This class contains also a lot of other extremely important discrete problems, say, numerous problems involving graphs, Boolean programming, etc.

There is a straightforward way to solve a problem from the class NP. Indeed, given a problem instance x, let us look through all binary words y with $l(y) \le q(l(x))$ - by definition, all actual solutions to the problem, if any exists, belong to this finite set of binary words. After a candidate solution y is chosen, let us test whether it is an actual solution; to this end it suffices to compute $P(x,y)$, and we know that it is possible to do in polynomial in $l(x)+l(y)$ time. If y solves x, we are done; if in course of our straightforward search no solution is found, we are in our right to say that the problem instance is unsolvable. Since the set of candidate solutions is finite, the complete procedure will take finitely many elementary steps, and the problem instance will be solved in finite time. A bad news, of course, is that this "finite time" for our brut force procedure is extremely large, since the number of candidate solutions is exponential in the length $l(x)$ of the input data. And what we are interested in is a *polynomial time* solution algorithm, i.e., that one with the running time bounded from above by a polynomial of the length $l(x)$ of the input data. In theoretical Computer Science polynomiality of an algorithm is regarded as a criterion of its effectiveness, and a problem is regarded as tractable if it admits a polynomial time solution algorithm. The problems with this latter property are called *polynomially solvable*, and the part of the class NP formed by these "tractable" problems is called P.

Thus, there exists a wide class of problems NP with a subclass of "simple", tractable, polynomially solvable problems P. I believe that everyone of you realizes that the first of our two

problems, "Connectedness", belongs to this latter subclass, since one immediately can invent an algorithm which, as applied to an n-vertex graph, solves the problem in $O(n^3)$ elementary steps (this bound can be significantly improved, but for our now purposes it does not matter). In contrast to the "Connectedness" problem, nobody knows a polynomial algorithm for the problem of detecting whether there exists a Hamiltonian cycle in a graph; all we know is that this is the most complicated, a so called NP-*complete* problem in the class NP. In fact the discovery of NP-complete problems somewhere 25 years ago was a great event in Discrete and Combinatorial Mathematics. Before this the situation with discrete problems was as follows. There were some of them which were known to be "tractable" (as we say now, polynomially solvable), e.g., the "Connectedness" problem or a close to it problem of finding shortest paths in a graph with the arcs assigned positive integer lengths. There were numerous extremely important for applications problems, like Boolean Programming or the problem of finding a Hamiltonian cycle, for which efficient algorithms were unknown; every of these "hard" problems had its own theory. Along with introducing the notion of the class NP it was proved that there are "most complicated" problems in the class, the NP-*complete* ones. The definition of a NP-complete problem is as follows. Let $\mathcal{P}(\cdot,\cdot)$ and $\mathcal{P}'(\cdot,\cdot)$ be two problems from NP (it is very natural to identify problems from NP with the related functions $\mathcal{P}(x,y)$). We say that $\mathcal{P}$ is *polynomially reducible* to $\mathcal{P}'$, if there exists an algorithm N with the following properties:

(1) given on input a binary word x representing an instance of the problem $\mathcal{P}$, it transforms it into a binary word $N(x)$ which represents an instance of the problem $\mathcal{P}'$, and the answer in this new problem is "yes" if and only if the answer in the original problem x is "yes"; thus, instead of solving an instance x of the problem $\mathcal{P}$, it suffices to solve the instance $N(x)$ of the problem $\mathcal{P}'$;

(2) the algorithm N computes $N(x)$ in time polynomial in $l(x)$.

It is clear that if $\mathcal{P}$ is polynomially reducible to $\mathcal{P}'$ and the latter problem is polynomially solvable, then the former one also polynomially solvable; to get a polynomial time algorithm for $\mathcal{P}$, it suffices to act on its data x by the algorithm N, which takes polynomial in $l(x)$ time and, consequently, produces the result $N(x)$ of the binary length polynomial in $l(x)$ (since the length of the output cannot be essentially larger that the running time of N - you should at least write down all binary digits of $N(x)$, and it, for any model of algorithm, takes $O(l(N(x)))$ elementary steps). After $N(x)$ is computed, let us apply to it the polynomial time algorithm for $\mathcal{P}'$, which, by assumption, do exists; the resulting answer "yes/no" will be exactly the same as for the original problem $\mathcal{P}$ (definition of reducibility), and the computation will take polynomial in $l(N(x))$ (and, consequently, in $l(x)$) time. Thus, the total time of solving x is bounded from above by a sum of two polynomials of $l(x)$i, and is therefore a polynomial in $l(x)$.

Now, a problem $\mathcal{P}$ from NP is called NP-complete, if *any* other problem from NP is polynomially reducible to $\mathcal{P}$. From the above discussion it is clear that polynomial solvability of any of NP-complete problems would imply polynomial solvability of *all* problems from the class NP.

The latter conclusion is nothing but a tautology. The actual breakthrough was to prove that NP-complete problems do exist, which is by no means evident in advance. Moreover, the discovery of NP-complete problems was not a pure existence theorem; several NP-complete problems were indicated explicitly, e.g., the Boolean Programming problem or that one on Hamiltonian cycle. Soon the situation became as follows: basically all important problems for which polynomial time algorithms remained unknown turned out to be NP-complete. This was actually a great deed, since it became clear that when looking for theoretically efficient

algorithms, you need not develop its own theory for every difficult problem; in a sense, we are dealing with numerous equivalent formulations of a *single* difficult problem. Let me stress that we still do not know whether the problem actually is difficult, whether the class NP is indeed larger than the subclass of polynomially solvable problems P. The conjecture is that NP≠P - this conjecture simply summarizes the failure of at least 40-year intensive attempts to find efficient algorithms for Boolean Programming, Travelling Salesman problem and tens of others. Nevertheless, the conjecture NP≠P, after 25 years, remains unproved; it is called the "major open question in Computer Science".

I have said that soon after the notions of the class NP and of a NP-complete problem were discovered in mid-sixties, it was found that almost all important problems for which no polynomial time algorithms were known are NP-complete. There were, anyhow, two remarkable exceptions: the Graph Isomorphism problem, the complexity status of this problem still being unknown, and Linear Programming.

4.2 Linear Programming

A Linear Programming problem

$$\textit{minimize}\ \ c^T u\ \ \textit{s.t.}\ \ Pu \leq q, x \in \mathbf{R}^k \tag{4.2.1}$$

(here A is an $l \times k$ matrix) is, in its nature, a problem of continuous optimization. Nevertheless, it can be easily included into the realm of problems from the class NP - it suffices to assume the coefficients to be integer. Let me stress that we do not pass to Integer Programming - the coefficients are integer, but the solution we are looking for may be real. To include the LP problem with integer coefficients (for the time being, no other LP problems are considered) into the class NP, it is convenient to write the problem down along with its dual, thus reducing the problem to a system of linear equalities and inequalities with unknowns $u \in \mathbf{R}^k$ and $v \in \mathbf{R}^l$:

$$\begin{cases} Pu \leq q, \\ v \leq 0 \\ P^T v = c, \\ c^T u - q^T v = 0 \end{cases} ; \tag{4.2.2}$$

each linear equality can be represented by a pair of linear inequalities, and we conclude that LP problems with integer data can be easily reduced to systems of linear inequalities, again with integer data. Thus, we may speak about the generic problem of solving the system

$$Ax \leq b, \tag{4.2.3}$$

A being an $m \times n$ matrix, b being an m-dimensional vector, all entries in A and b being integer. The data in the problem form a finite sequence of integers and can be therefore coded by a finite binary word. The length of this word is something like twice the total number L of bits in the binary representation of all coefficients of the problem:

$$L = \sum_{i=1}^{m}\sum_{j=1}^{n}(\lfloor \log_2(|A_{ij}|+1)\rfloor+1) + \sum_{i=1}^{m}(\lfloor \log_2(|b_i|+1)\rfloor+1);$$

this quantity L is called the *input length* of the problem instance.

Now, the problem of *solving* a system of linear (or any other) inequalities is *not* of the type allowed for NP problems; in these latter problems the question is whether a solution exists, not the solution itself. Thus, for the time being let us speak about the reduced version of an LP problem, namely, the *LP Feasibility problem* as follows:

LPF: *given system of linear inequalities (4.2.3) with integer coefficients, detect whether the system is solvable*

In this form the problem is of the type required by the definition of the class NP, which, anyhow, does not mean automatically that it is in NP - to verify this inclusion, we should verify the remaining items of the definition of a NP problem. What about the second property of a problem from the class NP - possibility to represent candidate solutions by finite binary words? Although we are looking for solutions with real entries, it is well-known (and we shall prove it in a while) that a solvable system of linear inequalities with integer coefficients always admits a rational solution with the entries

$$x_j^* = \frac{p_i}{q_i},\ |p_i|, |q_i| \leq 2^{O(L)};$$

it follows that without loss of generality we may look for a rational solution, i.e., for something which can be represented by a finite sequence of integers and therefore can be coded by a finite binary word. Moreover, from the above bounds on the magnitudes of the numerators and the denominators in a solution it follows that the length of the word representing a solution may be restricted to $O(Ln)$. Since clearly $n \leq L$, this upper bound on the output length is polynomial in L, as required by the definition of the class NP. And, of course, given the data and a candidate solution, we can check, simply by substitution, whether the candidate is a solution; this verification, of course, takes time polynomial in L and the binary length of the candidate solution. Thus, all restrictions imposed by the definition of a problem from the class NP are satisfied, and we see that the LP feasibility problem FLP with integer coefficients is in NP.

Now, what is the complexity status of the problem? Is it polynomially solvable or not? This crucial question (which can be posed, of course, not only for the feasibility problem FLP, but also for the standard Linear Programming problem with integer coefficients) remained open for something like 20 years. Of course, an LP program, even with real coefficients, can be solved in finitely many arithmetic operations by the Simplex method, and this method was and is extremely efficient in practice. Anyhow, is the Simplex method polynomial? The answer is "no". In early sixties Klee and Minty have discovered a family of LP problems $\mathcal{P}_n$ with n variables and $2n$ linear inequality constraints, $n = 1, 2, ...$, as follows: the input length of $\mathcal{P}_n$ is $O(n)$, and for certain particular version of the Simplex method it takes 2^n iterations to reach the solution. Thus, at least this particular version of the method is *not* polynomial; similar "bad" examples were found also for all other basic versions of the method. Thus, the Simplex method, in spite of its excellent practical performance, is not polynomial, and therefore it does not allow to understand whether LP is polynomially solvable.

Polynomial solvability of LP was established only in 1979 by Leonid Khachiyan, and the tool he used came from nonsmooth convex optimization - it was the Ellipsoid method invented 3 years earlier. Now we come to the Khachiyan Theorem on polynomial solvability of LP.

4.2.1 Polynomial solvability of FLP

The key point of Khachiyan's reasoning was to prove that the LP Feasibility problem is polynomially solvable. To this end Khachiyan acts as follows.

The first step. We may assume that the columns of the matrix A are linearly independent. Indeed, we are asked to find a linear combination of the columns which does not exceed the right hand side vector; since the set of linear combinations of all columns is exactly the same as the set of linear combinations of any column basis, we may from the very beginning look for a combination of this latter type, thus setting to zero the variables associated with the non-basis columns. In other words, we loose nothing when eliminate from our matrix A all "redundant" columns which are linear combinations of the remaining ones. Of course, to perform such an elimination, it requires to find a column basis, but this is a polynomial time linear algebra routine. Thus, from now on let us assume that the columns in A are linearly independent.

The second step. Let us prove that if the system is solvable, then it admits a solution x^* with not too large coordinates:

$$x^* \in B = \{x \in \mathbf{R}^n \mid |x|_\infty \leq 2^L\}.$$

We do not say that *all* solutions are inside the cube, we say that the solution set, if nonempty, intersects the cube.

The proof is immediate. The set X^* of solutions, if nonempty, is a polyhedral set which does not contain lines. Indeed, if it contains a line, that means that our linear inequalities are satisfied along a line, which is possible, of course, only if the left hand sides of these inequalities are zero at the direction of the line. But this means linear dependence of columns in A, and we just have excluded this case. From the theory of Linear Programming it is known that any polyhedral set defined by a system of linear inequalities and not containing lines has an extreme point x^*, and this extreme point, as substituted into the inequalities, makes equalities certain n *linearly independent* inequalities of the system (and, possibly, some inequalities more). In other words, x^* is a solution to a $n \times n$ linear subsystem

$$A'x = b'$$

of our initial system, with the inequality signs replaced by the equality one, and $n \times n$ matrix A' here is nonsingular. Due to the Kramer rule, the entries of x^* are ratios of the Kramer determinants:

$$(x^*)_j = \frac{\Delta_j}{\Delta}.$$

Now, the determinants here are the determinants of certain $n \times n$ submatrices of the integer matrix $[A, b]$. According to the Hadamard inequality, the modulus of the determinant of a $n \times n$ matrix with columns $p_1, ..., p_n$ does not exceed the product of the Euclidean norms of the columns:

$$|\operatorname{Det}(p_1, ..., p_n)| \leq \prod_{i=1}^{n} |p_i|_2$$

("the volume of a parallelotope does not exceed the product of lengths of the vectors the parallelotope is spanned by"). We immediately conclude that

$$\log_2 |\operatorname{Det}(p_1, ..., p_n)| \leq \sum_{i,j=1}^{n} \log_2(1 + |(p_i)_j|)$$

(indeed, $\prod_{j=1}^n(1+|a_j|)^2 \geq \sum_{j=1}^n |a_j|^2$). In our case, when Δ_j and Δ come from an integer matrix with the input length L, we, by definition of the length, conclude that

$$|\Delta_j|, |\Delta| \leq 2^L. \tag{4.2.4}$$

Thus, the numerators in the Kramer formulae are of absolute value at most 2^L; the denominator is a nonzero integer, and we conclude that

$$|x^*|_j \leq 2^L,\ j = 1, ..., n,$$

as claimed.

Third step. Now let us reformulate our linear feasibility problem as a problem of nonsmooth convex programming. To this end let us introduce the *residual function*

$$f(x) = \max_{i=1,...,m} \{a_i^T x - b_i\},$$

a_i^T being the rows of A. This is a piecewise linear convex function; it is nonpositive exactly on the set of solutions to our system of inequalities. Thus, to detect whether the system has a solution in the box B (as we already know, this is exactly the same as to detect whether the system is solvable) is the same as to detect whether the optimal value f^* in the problem

$$\textit{minimize}\ \ f(x)\ \ \textit{s.t.}\ \ x \in B \tag{4.2.5}$$

is nonpositive or positive. The crucial point here is the following simple statement:

Lemma 4.2.1 *Assume that the optimal value f^* in problem (4.2.5) is positive (so that system (4.2.3) is unfeasible). Then f^* is not too small:*

$$f^* \geq 2^{-3L}. \tag{4.2.6}$$

Proof. The optimal value f^* clearly is not less than the optimal value in the following Linear Programming problem:

$$(P): \quad \textit{minimize}\ \ t\ \ \textit{s.t.}\ \ Ax \leq b + te,$$

e being the m-dimensional vector of ones. The problem is, of course, feasible; if f^* is positive, then, as we know, the system $Ax \leq b$ is unsolvable, and consequently the optimal value t^* in our auxiliary LP problem is positive. Now, the rows in the matrix $[A, -e]$ of problem (P) are linearly independent; indeed, the columns of A are independent by assumption, and if e were a linear combination of these rows:

$$e = A\bar{x},$$

then the feasible set of (P) would contain, together with any point (x, t), the line $(x + s\bar{x}, t + s)$, $s \in \mathbf{R}$, and the objective in (P) would be below unbounded on the feasible set of the problem; as we know, is not the case. Thus, the columns of $[A, -e]$ are linearly independent, so that the feasible set of (P) does not contain lines. From the theory of Linear Programming it now follows that among optimal solutions to (P) there is one, (x^+, t^*), which corresponds to an extreme point of the feasible set, and this solutions makes equalities certain $(n + 1)$ linearly independent inequality constraints in (P). Same as above, we conclude that the coordinates of

this solutions are ratios of certain determinants coming from $(n+1)\times(n+1)$ submatrices of $[A,-e]$. In particular,

$$t^* = \frac{\Delta^*}{\Delta}$$

is the ratio of this type. The denominator here, exactly as above, does not exceed in absolute value $2^{(2m+L)}$ (the binary length of the matrix $[A,-e]$ is at most $L+2m$), and the numerator is an integer. We conclude that if t^* is positive (and we know that it is), then $t^* \geq 2^{-2m-L} \geq 2^{-3L}$, so that $f^* \geq t^* \geq 2^{-3L}$, as claimed. ∎

Now we see that to find out whether (4.2.3) is solvable is the same as to evaluate, within accuracy, say $2^{-3L}/3$, the optimal value to the nonsmooth convex optimization program (4.2.5). If the approximate optimal value is less than $2^{-3L}/2$, we can say that the initial system is solvable, if it is not the case, the system is unsolvable. To solve the nonsmooth problem, one can use the Ellipsoid method tuned to the relative accuracy like

$$\varepsilon = \frac{2^{-3L}}{3(2n2^L \times 2^L)} \leq 2^{-6L-3}$$

(the quantity in the parentheses in the denominator is a rough upper bound for the variation of the objective on the cube B; note also that $n \leq L$). To solve the problem to the desired accuracy, it takes $O(1)n^2\ln(1/\varepsilon) = O(1)n^2L$ steps of the method, with $O(mn+n^2)$ arithmetic operations per step (it takes $O(mn)$ operations to compute the value and a subgradient of the objective at a point and $O(n^2)$ operations to update the current ellipsoid). Thus, the overall effort is

$$N = O(1)n^2(n^2+mn)L \leq O(L^5)$$

arithmetic operations, which is a polynomial of L. Of course, in our basic representation of the Ellipsoid method we dealt with operations of precise real arithmetic, which is not allowed now: we should use completely finite model of computations and count bit operations. A straightforward analysis demonstrates, anyhow, that we may replace the operations of the precise real arithmetic with inexact operations which keep $O(nL)$ digits of the result before and after the dot, and this "inexact" implementation of the method still evaluates the optimal value within the desired accuracy. Thus, we may decide whether the initial system of inequalities is solvable or not in no more than $O(L^5)$ arithmetic operations with $O(nL)$-bit operands, which takes a polynomial in L number of elementary steps. We conclude that the feasibility problem is polynomially solvable.

4.2.2 From detecting feasibility to solving linear programs

Recall that we have reduced the Linear Programming problem to that one of solving a system of linear inequalities, but we still did not point out a polynomial time algorithm for this latter problem; we know how to decide, in polynomial time, whether the system of inequalities is feasible, but do not know how to find a solution if it exists. There are many ways to reduce solving a system of linear inequalities to a "small" series of feasibility problems, and the simplest for explanation is as follows. Let us first check whether the system in question is solvable (we already know how to do it in a polynomial time). If infeasibility is detected, we are done. Now assume that we have found the system to be feasible. Let us prove that in this case it is possible

to delete some of inequalities of the system and to set the remaining inequalities to equalities in such a way that the resulting system of linear equations would be solvable and any solution to it would also be a solution to the initial system of inequalities. To this end consider the following construction. We start with the initial system of inequalities, let it be called $\mathcal{L}_0$, and at i-th step of the construction have another system, $\mathcal{L}_{i-1}$, which is obtained from the initial one by deleting some inequalities and setting some inequalities to equalities in such a way that, first,

the number of inequalities in the updated system $\mathcal{L}_{i-1}$ is at most $m+1-i$,

$\mathcal{L}_{i-1}$ is solvable, and all solutions to the system are also solutions to the initial system $\mathcal{L}_0$.

Note that for $i=1$ these requirements are satisfied by trivial reasons.

Now assume that $\mathcal{L}_{i-1}$ includes at least one inequality. Let us choose one of the inequalities of the system, let us make it the equality and test whether the resulting system $\mathcal{L}'_i$ is solvable. If it is the case, then the system $\mathcal{L}'_i$ clearly can be chosen as $\mathcal{L}_i$: it contains one inequality less than $\mathcal{L}_{i-1}$, is solvable and all solutions to the system clearly are solutions to $\mathcal{L}_{i-1}$ and therefore, by our inductive hypothesis, to $\mathcal{L}_0$. Now assume that the system $\mathcal{L}'_i$ is unfeasible. Since $\mathcal{L}_{i-1}$ is solvable and $\mathcal{L}'_i$ is not, we conclude that the inequality we have made equality is non-active at the set of solutions to $\mathcal{L}_{i-1}$, i.e., is strict at any solution to the latter system. But this means exactly that eliminating this inequality from the system $\mathcal{L}_{i-1}$ we do not extend the set of solutions, so that we can take as $\mathcal{L}_i$ the system which is obtained from $\mathcal{L}_{i-1}$ by deleting the inequality in question; the resulting system is solvable and has the same set of solutions as that one of $\mathcal{L}_{i-1}$; consequently, the set of solutions to $\mathcal{L}_i$ is contained in that one of the initial system. And, of course, we have reduced by 1 the number of inequalities in our current system.

In no more than m steps of the indicated construction we come to a system $\mathcal{L}_m$ which does not contain inequalities, i.e., is a system of linear equations; our construction ensures that the system is solvable and that any solution to it solves the initial system of inequalities. Now, to find a solution to a system of linear equations it takes, of course, polynomial time. To pass from the initial system $\mathcal{L}_0$ to $\mathcal{L}_m$ it also takes polynomial time: we should solve m feasibility problems of the same input length L as that one of the initial system. Thus, Linear Programming admits a polynomial time solution algorithm and is therefore polynomially solvable.

To conclude, let me say that although the Ellipsoid method, in the worst case sense, is incomparably better than the Simplex method, in practical applications everything looks opposite: the Ellipsoid method works more or less according to its theoretical efficiency bound, i.e., requires something like $O(n^4+n^3m)$ arithmetic operations per accuracy digit, while the Simplex method in real-world applications *never* behaves itself too bad, God knows why; the observed # of steps in the case of a Linear Programming problem with n variables and m inequality constraints normally is proportional, with a very moderate coefficient, to n, which is much-much better than the complexity of the Ellipsoid method. Numerous experiments with the Ellipsoid method in the beginning of eighties definitely confirm practical superiority of the Simplex method. But this is not the end of the story: in 1984, Narendra Karmarkar proposed a completely new polynomial time algorithm for Linear Programming. The algorithm of Karmarkar turned out to be very efficient in practice and, in contrast to the Ellipsoid method, indeed became a serious practical alternative to the Simplex method.

4.3 Exercises: Around the Simplex method and other Simplices

4.3.1 Example of Klee and Minty

In early sixties, Klee and Minty demonstrated that the worst-case behaviour of the Simplex method is very bad. Let us reproduce their famous example of a "hard" family of LP problems.

Consider an LP program $\mathcal{P}$ of the form

$$\textit{maximize} \;\; x_n \;\; \textit{s.t.} \;\; x \in \mathbf{R}^n,\, Ax \leq b$$

with m inequality constraints and a nonempty and bounded feasible set Q. Let us call a *simplex path* associated with $\mathcal{P}$ a sequence $v^1, ..., v^N$ of vertices of the polytope Q such that

the objective increases along the sequence:

$$(v^1)_n < (v^2)_n < ... < (v^N)_n;$$

the vertices v^k, v^{k+1} neighbouring in the sequence are neighbouring in the polytope Q, i.e., they are linked by an edge of the polytope.

If $v^1, ..., v^N$ is a simplex path, then the Simplex method as applied to the problem may, at least for certain pivoting rules, generate a sequence of iterates which includes the simplex path as a fragment. Klee and Minty prove that, given n, one can point out a problem $\mathcal{P}$ of the aforementioned type with $m = 2n$ in such a way that the problem would admit a simplex path of the length $N = 2^n$, so that the Simplex method, at least with certain "bad" pivoting rules, would solve the problem in 2^n steps. Later on, the result was strengthen: it was shown that not only "bad" pivoting rules, but also all standard rules may result, on some problem instances, in an exponential in the number m of constraints and n of variables amount of pivots. It still remains unknown whether one can invent pivoting rules free of this drawback. This seems to be an extremely difficult question, due to the following reasons. If there exists a pivoting rule which results in a polynomial in n and m number of pivots, then for any polytope $Q \subset \mathbf{R}^n$ defined by m linear inequalities and for any pair v, v' of vertices of the polytope there exists a path of vertices of Q which starts at v, ends at v', contains a polynomial in m, n number of vertices and is such that any pair of vertices which are neighbours in the path are also neighbours in Q (indeed, it suffices to choose the objective in the LP problem $c^T x \to \max \mid x \in Q$ in a way which makes v' the optimal solution and to run the Simplex method, v being the starting vertex). On the other hand, there is a well-known almost 40-years old conjecture of M. Hirsh (1957) that any pair of vertices of a polytope of dimensions (m, n) can be linked by a path of the aforementioned type comprised of no more than $m - n + 1$ vertices; in numerous attempts to prove the conjecture, nobody was lucky to prove even a weaker statement - to obtain a polynomial in m, n upper bound on the length of a path. And to find "polynomial" pivoting rules for the Simplex method is, of course, more difficult than to prove the weak version of the Hirsh conjecture.

Let us come back to the example of Klee and Minty.

Exercise 4.3.1 *Consider an LP problem $\mathcal{P}_n$:*

$$x_n \to \max \mid \sum_{j=1}^{n} a_{ij} x_j \leq b_i,\; i = 1, ..., 2n,$$

the feasible polytope Q_n of the problem being contained in the cube

$$C_n = \{x \in \mathbf{R}^n \mid 0 \le x_i \le 1, i = 1, ..., n\}.$$

Let the LP problem $\mathcal{P}_n^+$ be defined as

$$x_{n+1} \to \max \mid \sum_{j=1}^{n} a_{ij} x_j \le b_j,\ j = 1, ..., 2n, \frac{1}{3} x_n \le x_{n+1} \le 1 - \frac{1}{3} x_n.$$

Prove that

1) the set of vertices of the $(n+1)$-dimensional polytope Q_n^+ of feasible solutions to $\mathcal{P}_n^+$ is as follows: and each vertex $v = (x_1, ..., x_n)$ of Q_n "gives birth" a pair of vertices v_+, v_- in Q_n^+, namely,

$$v_+ = (x_1, ..., x_n, 1 - \frac{1}{3} x_n), \quad v_- = (x_1, ..., x_n, \frac{1}{3} x_n);$$

2) the "children" v_+ and v_- of a vertex v of Q_n are neighbouring in Q_n^+;

3) if $v^1, ..., v^N$ is a simplex path for $\mathcal{P}_n$, then the sequence

$$(v^1)_-, (v^2)_-, ..., (v^N)_-, (v^N)_+, (v^{N-1})_+, ..., (v^1)_+$$

is a simplex path for the problem $\mathcal{P}_n^+$.

Exercise 4.3.2 (the Klee and Minty "hard" LP problem) *Prove that the problem*

$$x_n \to \max \mid 0 \le x_1 \le 1, \frac{1}{3} x_1 \le x_2 \le 1 - \frac{1}{3} x_1, ..., \frac{1}{3} x_{n-1} \le x_n \le 1 - \frac{1}{3} x_{n-1}$$

with n variables and $2n$ inequality constraints admits a simplex path with 2^n vertices. What is the input length of the problem?

4.3.2 The method of outer simplex

Looking at the Ellipsoid method, one may ask: why should we restrict ourselves to ellipsoids and not to use localizers of some other "simple" shape? This is a reasonable question. In fact all properties of the "ellipsoidal" shape which were important in the Ellipsoid version of the cutting plane scheme were as follows:

(a) *all ellipsoids in $\mathbf{R}^n$ are affine equivalent to each other, and, in particular, are affine equivalent to certain "standard" ellipsoid, the unit Euclidean ball; therefore in order to look what happens after a cut passing through the center of an ellipsoid, it suffices to study the simplest case when the ellipsoid is the unit Euclidean ball;*

(b) *the part of the unit Euclidean ball which is cut off the ball by a hyperplane passing through the center of the ball can be covered by an ellipsoid of volume less than that one of the ball.*

Now, can we point out another "simple shape" which satisfies the above requirements? The natural candidate is, of course, a simplex. All n-dimensional simplexes are affine equivalent to each other, so that we have no problems with (a); the only bottleneck for simplexes could be (b). The below exercises demonstrate that everything is ok with (b) as well.

Let us start with considering the standard simplex (which now plays the role of the unit Euclidean ball).

Exercise 4.3.3 $^+$ *Let*

$$\Delta = \{x \in \mathbf{R}^n \mid x \geq 0, \sum_{i=1}^n x_1 \leq 1\}$$

be the standard n-dimensional simplex in $\mathbf{R}^n$, let

$$c = (\frac{1}{n+1}, ..., \frac{1}{n+1})^T$$

be the barycenter ($\equiv$ the center of gravity, see exercise 1.5) of the simplex, let

$$g = (\gamma_1, ..., \gamma_n)^T \geq 0$$

be a nonzero nonnegative vector such that

$$g^T c = 1,$$

and let

$$\bar{\Delta} = \{x \in \Delta \mid g^T x \leq g^T c = 1\}.$$

Prove that if $\vartheta \in [0, 1]$, then the simplex Δ_ϑ with the vertices $v_0 = 0$,

$$v_i = (1 - \vartheta + \vartheta\gamma_i)^{-1} e_i,\ i = 1, ..., n,$$

(e_i are the standard orths of the axes) contains $\bar{\Delta}$.

Prove that under appropriate choice of ϑ one can ensure that

$$\mathrm{Vol}_n(\Delta_\vartheta) \leq \chi^n(n)\mathrm{Vol}_n(\Delta), \quad \chi^n(n) = \left(1 + \frac{1}{n^2-1}\right)^{n-1}\left(1 - \frac{1}{n+1}\right) \leq 1 - O(n^{-2}). \quad (4.3.1)$$

Exercise 4.3.4 *Let D be a simplex in $\mathbf{R}^n$, $v_0, ..., v_n$ be the vertices of D,*

$$w = \frac{1}{n+1}(v_0 + ... + v_n)$$

be the barycenter of D and g be a nonzero linear functional.

Prove that the set

$$\bar{D} = \{x \in D \mid (x - w)^T g \leq 0\}$$

can be covered by a simplex D' such that

$$\mathrm{Vol}_n(D') \leq \chi^n(n)\mathrm{Vol}_n(D),$$

with $\chi(n)$ given by (4.3.1). Based on this observation, construct a cutting plane method for convex problems with functional constraints where all localizers are simplexes.
What should be the associated size?
What is the complexity of the method?

Note that the progress in volumes of the subsequent localizers in the method of outer simplex (i.e., the quantity $\chi^n(n) = 1 - O(n^{-2})$) is worse than that one $\kappa^n(n) = 1 - O(n^{-1})$ in the Ellipsoid method. It does not, anyhow, mean that the former method is for sure worse than the latter one: in the Ellipsoid method, the actual progress in volumes *always* equals to $\kappa^n(n)$, while in the method of outer simplex the progress depends on what are the cutting planes; the quantity $\chi^n(n)$ is nothing but the worst case bound for the progress, and the latter, for a given problem, may happen to be more significant.

Lecture 5

Linearly converging methods for games

The cutting plane scheme and the methods based on it, e.g., the Center-of-Gravity and the Ellipsoid one, can be applied not only to the convex optimization problems, but to other problems with convex structure, I mean solving convex-concave games and variational inequalities with monotone operators. This is our now topic; today I shall speak about convex-concave games.

5.1 Convex-concave games

The problem of solving a convex-concave game (it is also called the *saddle point* problem for convex-concave functions) is as follows. We are given a pair of solids $P \subset \mathbf{R}^p$ and $Q \subset \mathbf{R}^q$ and a continuous function

$$f(x,y) : G = P \times Q \to \mathbf{R}$$

which is convex in $x \in P$ for every fixed $y \in Q$ and concave in $y \in Q$ for every fixed $x \in P$. The problem is to approximate a *saddle point* (x^*, y^*) of the function, i.e., a point where f attains its minimum in x and its maximum in y:

$$f(x,y^*) \geq f(x^*,y^*) \geq f(x^*,y), \;\; (x,y) \in P \times Q.$$

Let me recall you the standard interpretation of the saddle point problem. There are two players, let them be you and me. My possibility is to choose a point $x \in P$, and yours - to choose $y \in Q$; after we have chosen x and y, I pay to you the sum $f(x,y)$. Of course, my goal is to decrease my payment, and yours - to increase it. A saddle point (x^*, y^*) is an equilibrium: if you keep the choice y^*, I cannot decrease my payment by passing from x^* to any other choice, same as if I keep the choice x^*, you cannot enforce me to pay more by passing from y^* to any other choice. Note that the issue of existence of the equilibrium is far from being trivial; in the case in question (f is continuous and convex-concave, P and Q are solids) the existence is given by the well-known von Neumann Lemma.

The indicated interpretation explains the following construction: let

$$\bar{f}(x) = \max_{y \in Q} f(x,y), \quad \underline{f}(y) = \min_{x \in P} f(x,y).$$

$\bar{f}(x)$ is my loss function: if I choose x, you cannot enforce me to pay more than $\bar{f}(x)$, and can enforce me to pay this sum (if you know x). Similarly, $\underline{f}(y)$ is your profit function: if you choose y, I cannot pay you less than $\underline{f}(y)$ and, if I know y, can pay exactly this sum. Thus, if I choose as x any minimizer of $\bar{f}$, I may be sure that I shall not pay more than $\min_{x\in P}\bar{f}(x)$; similarly, if you choose as y any maximizer of $\underline{f}(y)$, you may be sure that your profit will be at least $\max_{y\in Q}\underline{f}(y)$. The existence of a saddle point implies that our worst-case expectations are equal:

$$\min_{x\in P}\bar{f}(x)=\min_{x\in P}\max_{y\in Q}f(x,y)=\max_{y\in Q}\min_{x\in P}f(x,y)=\max_{y\in Q}\underline{f}(y);$$

and it is easily seen that the set $S^*(f)$ of all saddle points of f is nothing but the direct product of the optimal sets of our cost functions:

$$S^*(f)=\operatorname*{Argmin}_P\bar{f}\times\operatorname*{Argmax}_Q\underline{f}.$$

Thus, to solve the game is exactly the same as to solve the pair of convex optimization problems

$$\textit{minimize}\ \ \bar{f}(x),\ x\in P;\quad \textit{maximize}\ \ \underline{f}(y),\ y\in Q$$

(note that $\bar{f}$ clearly is convex and continuous on P, while $\underline{f}$ is concave and continuous on Q), the optimal values in the problems being equal to each other. The difficulty, anyhow, is that the objectives in the problems are defined implicitly, this is why we need specific tools to find saddle points.

Now I can easily motivate the notion of an ε-saddle point. Let $\varepsilon\in(0,1)$ be the required relative accuracy. We say that a pair $(x,y)\in P\times Q$ is an ε-saddle point, if

$$\bar{f}(x)-\underline{f}(y)\le\varepsilon\left(\max_{P\times Q}f-\min_{P\times Q}f\right)\equiv\varepsilon V(f);$$

note that the left hand side is nothing but

$$(\bar{f}(x)-\min_P\bar{f})+(\max_Q\underline{f}-\underline{f}(y))$$

(since the optimal values $\min_P\bar{f}$ and $\max_Q\underline{f}$ are equal to each other), so that the left hand side is the sum of the absolute inaccuracy of x regarded as an approximate minimizer to my loss function and that one of y regarded as an approximate maximizer of your profit function. This sum has a very natural interpretation in the equilibrium terms: if you keep your choice to be y and I pass from x to any other choice $x'\in P$, I cannot save more than $\bar{f}(x)-\underline{f}(y)$, and similarly for you.

In what follows we assume that we have an access to the first-order oracle which, given on input a pair $(x,y)\in\operatorname{int}(P\times Q)$, returns on output the value $f(x,y)$ of the cost function f, its subgradient $f'_x(x,y)$ in x and its supergradient $f'_y(x,y)$ in y. It is convenient to assemble the "gradient part" of the answer in a single vector

$$g(x,y)=(f'_x(x,y),-f'_y(x,y))\in\mathbf{R}^p\times\mathbf{R}^q.$$

The crucial property of this vector is as follows:

Lemma 5.1.1 *One has*

$$(z'-z)^T g(z') \geq f(x',y) - f(x,y'),\ z'=(x',y') \in \text{int } G,\ z=(x,y)\in G,\ G=P\times Q. \quad (5.1.1)$$

In particular, if $z=(x,y)$ is a saddle point of f, then

$$(z'-z)^T g(z') \geq 0,\ z' \in \text{int } G. \quad (5.1.2)$$

The proof is immediate:

$$f(x,y') - f(x',y') \geq (x-x')^T f'_x(x',y')$$

and

$$f(x',y) - f(x',y') \leq (y-y')^T f'_y(x',y'),$$

whence

$$f(x',y) - f(x,y') \leq (z'-z)^T g(z');$$

thus, we come to (5.1.1). To prove (5.1.2), it suffices to note that if (x,y) is a saddle point, then

$$f(x',y) \geq f(x,y) \geq f(x,y')$$

and therefore the right hand side of (5.1.1) is nonnegative.

5.2 Cutting plane scheme for games: updating localizers

Now assume that, given an access to the first-order oracle, we are asked to find an ε-saddle point of f. Looking at relation (5.1.2), we immediately get the idea to use the cutting plane scheme: having chosen a search point $z=(x,y)\in \text{int}\,(P\times Q)$ and given the answer $g(z)$ of the oracle at the input z, we can point out a half-space, namely

$$\Pi_z = \{w\in \mathbf{R}^{p+q} \mid (w-z)^T g(z) \leq 0\},$$

which for sure contains the saddle set of f; the boundary hyperplane of the half-space passes through the search point, and we may forget about all points of $P\times Q$ which belong to the complementary half-space, thus reducing the current localizer. To be more exact, we should take into account that Π_z is not necessarily a half-space; if $g(z)$ is zero, this is the whole space. But this is the minor difficulty: if $g(z)$ is zero, that means that $f(x',y)$ attains its minimum in $x'\in P$ at $x'=x$, while $f(x,y')$ attains its maximum in $y'\in Q$ at $y'=y$ (recall that the components of g are sub- and minus supergradient of f with respect to x and y), so that z is a saddle point, and we are done.

Thus, we come to the following *cutting plane scheme* for saddle point problems:

Initialization. Choose the initial localizer - a solid $G_0\supset G$.
i-th step. Given the previous localizer - a solid G_{i-1} -

1) choose i-th search point $z_i=(x_i,y_i)\in \text{int}\, G_{i-1}$;

2) check whether $z_i\in \text{int}\, G$. If it is not the case, call the step i *non-productive*, choose a nonzero vector e_i which separates z_i and G:

$$(z-z_i)^T e_i \leq 0\ \ \forall z\in G$$

and go to 3), otherwise call the oracle, z_i being the input. If the returned vector $e_i \equiv g(z_i)$ is zero, terminate and claim that z_i is a saddle point, otherwise call the step i *productive* and go to 3).

3) Define i-th intermediate localizer $\bar{G}_i$ as

$$\bar{G}_i = \{z \in G_{i-1} \mid (z - z_i)^T e_i \leq 0\}.$$

Embed $\bar{G}_i$ into a solid G_i, the updated localizer, and loop.

I have presented the scheme in the form which covers both implementations known to us - the Center-of-Gravity version and the Ellipsoid one. We see that the situation is completely similar to that one arising in convex minimization without functional constraints. In particular, we know how to decrease at a problem-independent rate an appropriately chosen size of the localisers. Namely, in the Center-of-Gravity method we set

$$G_0 = G, \quad z_i = \frac{1}{\mathrm{Vol}\, G_{i-1}} \int_{G_{i-1}} z dz, \quad G_i = \bar{G}_i,$$

which results in

$$\mathrm{AvDiam}(G_i) \leq \exp\{-O(1)\frac{i}{p+q}\}\, \mathrm{AvDiam}(G).$$

In the Ellipsoid method all G_i are ellipsoids, z_i is the center of G_{i-1} and G_i is the ellipsoid of the smallest volume containing the half-ellipsoid $\bar{G}_i$; this strategy results in

$$\mathrm{EllOut}(G_i) \leq \exp\{-O(1)\frac{i}{(p+q)^2}\}\, \mathrm{EllOut}(G_0).$$

In both the cases $O(1)$ is certain positive absolute constant.

5.3 Cutting plane scheme for games: generating solutions

You might think that we are done, but this is too early to think so – we did not say how to generate approximate solutions to the problem. In the minimization case we simply take as i-th approximate solution the best, in terms of the objective, of the search points generated at the productive steps. Now we also have an objective, our cost function f, but we are not interested in minimizing it, we are interested in something else, so that there is no reason to look at the values of f. In fact the situation is not so simple, as it is seen from the following example:

$$f(x, y) = xy : [-1, 3] \times [-1, 1] \to \mathbf{R}.$$

This simple cost function is, of course, convex-concave. The corresponding loss and profit functions are

$$\bar{f}(x) = \max_{y \in [-1,1]} xy = |x|, \quad \underline{f}(y) = \min_{x \in [-1,3]} xy = \begin{cases} -y, & y \geq 0 \\ 3y, & y \leq 0 \end{cases},$$

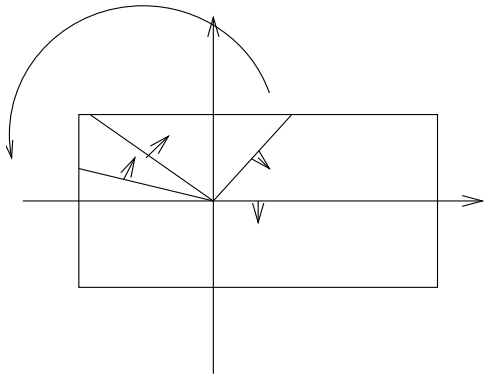
the saddle point is unique and this is the origin. Now let us look what happens when we apply to the problem, say, the Center-of-Gravity method. Note that

$$g(x, y) = (f'_x(x, y), -f'_y(x, y))^T = (y, -x)^T;$$

this vector is orthogonal to (x, y), so that all our cutting lines

$$\{z \mid z^T g(z_i) = 0\}$$

pass through the origin. Now, the center of gravity of our initial localizer, i.e., of the rectangle $[-1, 3] \times [-1, 1]$, is $(1, 0)$, and the first cut results in the rectangular localizer $[-1, 3] \times [0, 1]$. What happens later, it is clear from the picture:



After several steps the localizers G_i become triangles of the type shown on the picture; as $i \to \infty$, these triangles shrink to the segment $[-1, 0]$ of the x-axis, and their centers of gravity, i.e., our search points, converge to the point $(-2/3, 0)$. Thus, the method converges, but this is a bad news, since it does not converge to the saddle point!

It turns out, anyhow, that the situation is not hopeless. As we shall see in a while, it is possible to pass from the search points to certain convex combinations of these points which do converge to the saddle set.

The trick we are about to demonstrate is as follows. Assume that we have performed N steps, and let I_N be the set of all indices $i \leq N$ associated with productive steps; assume that this set is nonempty. Consider the auxiliary concave function

$$g^N(z) = \min_{i \in I_n} (z_i - z)^T g(z_i) \equiv \min_{i \in I_n} g_i(z).$$

I claim that if the maximum of the function over G is small, then we can form a good approximation to the saddle point set. Indeed, the maximum of g^N over G is the same as the maximum over the set of certain convex combination of the linear components of the function:

$$\exists \lambda^* = \{\lambda_i^*\}_{i \in I_N} \in \Delta = \{\lambda \mid \lambda_i \geq 0, \sum_{i \in I_N} \lambda_i = 1\} : \quad \max_{z \in G} g^N(z) = \max_{z \in G} \{\sum_{i \in I_N} \lambda_i g_i(z)\}.$$

This is the standard duality statement, and the proof is immediate:

$$\max_{z \in G} g^N(z) \equiv \max_{z \in G} \min_{i \in I_N} g_i(z) = \max_{z \in G} \min_{\lambda \in \Delta} \sum_{i \in I_N} \lambda_i g_i(z) = \min_{\lambda \in \Delta} \max_{z \in G} \sum_{i \in I_N} \lambda_i g_i(z)$$

(when interchanging $\min_\lambda$ and $\max_z$, we have applied the von Neumann Lemma to the bilinear in z and λ and therefore concave in z and convex in λ Lagrange function $\sum \lambda_i g_i(z)$). I shall call the indicated weights λ_i^* the *optimal Lagrange multipliers* associated with N-th step of the cutting plane scheme.

Now let us prove the following lemma:

Lemma 5.3.1 *Let*

$$z^N = (x^N, y^N) = \sum_{i \in I_N} \lambda_i^* z_i$$

be the weighted sum of the search points z_i associated with productive steps $i \le N$, the weights being the optimal Lagrange multipliers. Then z^N is feasible, i.e., $z^N \in G$, and the absolute inaccuracy of z^N as an approximate saddle point of f admits the following upper bound:

$$\bar{f}(x^N) - \underline{f}(y^N) \le \max_{z \in G} g^N(z). \tag{5.3.1}$$

Proof. Inclusion $z^N \in G$ is evident, since $z_i \in G$ when $i \in I_N$ and G is convex. Let $z = (x, y) \in G$. From Lemma 5.1.1 it follows that

$$g_i(z) \equiv (z_i - z)^T g(z_i) \ge f(x_i, y) - f(x, y_i),\ i \in I_N.$$

Taking weighted sum of these inequalities, we come to

$$\sum_{i \in I_N} \lambda_i^* g_i(z) \ge \sum_{i \in I_N} \lambda_i^* f(x_i, y) - \sum_{i \in I_N} \lambda_i^* f(x, y_i) \ge f(x^N, y) - f(x, y^N)$$

(the concluding inequality follows from the convexity of f in the first argument and the concavity in the second). Taking maxima of both sides with respect to $z = (x, y) \in G$ and using the origin of the optimal Lagrange weights, we come to

$$\max_{z \in G} g^N(z) \ge \bar{f}(x^N) - \underline{f}(y^N),$$

as claimed. ∎

Thus, given in advance that the maximum of the function g^N is small, we can, looking at the function, aggregate the associated search points in a good approximation to the saddle set of f. And it turns out that our cutting plane scheme enforces the maximum of g^N to go to zero as $N \to \infty$:

Lemma 5.3.2 *Let a convex-concave saddle point problem be solved by a cutting plane method, and let* Size *be a size. Assume that at certain step N we have*

$$\varepsilon_N \equiv \frac{\mathrm{Size}(G_N)}{\mathrm{Size}(G)} < 1.$$

Then the set I_N of those indices $i \le N$ which are associated with productive steps is nonempty and

$$\max_{z \in G} g^N(z) \le \frac{\varepsilon_N}{1 - \varepsilon_N} V(f).$$

Proof. To prove that I_N is nonempty, note that by construction of the vector e_i associated with a non-productive step the set

$$G_{i-1} \backslash G_i$$

does not intersect G (at such a step, we remove from the previous localizer only those points which do not belong to G). Consequently, if we have performed N steps and all of them were

non-productive, then G_N must contain G and therefore the size of G_N, due to the monotonicity of the size, should be at least the size of G, which in fact is not the case.

Thus, I_N is nonempty, so that g^N is well-defined. Let z^* be a maximizer of the function over G, let $\alpha \in (\varepsilon_N, 1)$ and let

$$G^\alpha = z^* + \alpha(G - z^*) \equiv \{(1-\alpha)z^* + \alpha z \mid z \in G\}.$$

The size $\mathrm{Size}(G^\alpha)$ is exactly $\alpha\,\mathrm{Size}(G)$, due to the homogeneity of the size with respect to similarity transformations. Thus,

$$\mathrm{Size}(G^\alpha) > \mathrm{Size}(G_N)$$

and therefore, due to the monotonicity of the size with respect to inclusions, G^α cannot be covered by G_N:

$$\exists w = (1-\alpha)z^* + \alpha z, z \in G: \quad w \notin G_N.$$

Since w does not belong to the N-th localizer, it was cut off at one of our steps:

$$(w - z_i)^T e_i > 0$$

for some $i \le N$. I claim that the step i was productive; indeed, otherwise all points of G, and, in particular, the point w would satisfy the inequality

$$(z - z_i)^T e_i \le 0,$$

which, as we know, is not the case.

Thus, $i \in I_N$, so that $e_i = g(z_i)$. We come to

$$(w - z_i)^T g(z_i) > 0,$$

whence

$$g_i(w) = (z_i - w)^T g(z_i) < 0.$$

Since $w = (1-\alpha)z^* + \alpha z$, we come to

$$(1-\alpha)g_i(z^*) \le -\alpha g_i(z) \equiv -\alpha(z_i - w)^T g(z_i).$$

In view of our initial lemma, the quantity $(z_i - w)^T g(z_i)$ is bounded from below by the difference of two values of f at certain points of G, i.e., the quantity is $\ge$ than minus the variation of f over G:

$$(z_i - w)^T g(z_i) \ge -V(f),$$

and we come to

$$(1-\alpha)g_i(z^*) \le \alpha V(f).$$

The left hand side in the latter inequality is $\ge g^N(z^*) = \max_z g^N$ (since g^N is the minimum of all g_i, $i \in I_N$), and we come to

$$\max_G g^N \le \frac{\alpha}{1-\alpha} V(f).$$

Since $\alpha \in (\varepsilon_i, 1)$ is arbitrary, we are done. ∎

Combining the results of Lemmas 5.3.1 and 5.3.2, we come to the conclusion as follows.

Proposition 5.3.1 *Assume that we are solving a convex-concave saddle point problem by a cutting plane method and let* Size *be a size. If a step N is such that*

$$\varepsilon_N \equiv \frac{\text{Size}(G_N)}{\text{Size}(G)} < 1,$$

then the set I_N of indices $i \le N$ of the productive steps is nonempty, and one can point out a feasible approximate saddle point which solves the game within relative inaccuracy $\varepsilon_N/(1-\varepsilon_N)$. To form z^N, it suffices to solve the auxiliary piesewise linear convex optimization problem

$$\text{maximize} \quad \min_{i\in I_N}(z_i - z)^T g(z_i) \quad \text{over} \quad z \in G,$$

find the corresponding optimal Lagrange multipliers λ_i^*, $i \in I_N$ and set

$$z^N = \sum_{i\in I_N} \lambda_i^* z_i.$$

As a corollary, we obtain the following statement:

Proposition 5.3.2 *The complexity of solving a convex-concave game within relative accuracy $\varepsilon \in (0,1)$ by the Center-of-Gravity version of the cutting plane scheme does not exceed*

$$N_{CG}(\varepsilon) = \rfloor O(1)(p+q)\ln(\frac{1+\varepsilon}{\varepsilon})\lfloor;$$

for the Ellipsoid method, the complexity does not exceed

$$N_{Ell}(\varepsilon) = \rfloor O(1)(p+q)^2 \ln\left(\frac{1+\varepsilon}{\beta\varepsilon}\right)\lfloor, \quad \beta = \frac{\text{EllOut}(G)}{\text{EllOut}(G_0)}.$$

In both the estimates, $O(1)$ are appropriately chosen absolute constants.

We see that to solve convex-concave saddle point problems is *at most* as difficult as to minimize convex functions. At the same time, to solve saddle point problems is *at least* as difficult as to minimize a convex function over a solid (indeed, the latter problem can be regarded as a saddle point one with the cost function independent of the choice of one of the players). Thus, in our now situation we may make the same conclusions on the level of optimality of the methods as those for the minimization problems. Namely, the Center-of-Gravity method is optimal in complexity, up to an absolute constant factor, provided that the required accuracy is small enough; under the latter assumption, the Ellipsoid method is optimal up to a factor of order of the dimension of the problem.

5.4 Concluding remarks

As far as actual computations are concerned, the disadvantage of the outlined scheme is the necessity to solve auxiliary piesewise-linear convex optimization problems. This is not so hard as one can think. One should not solve these problems at each step, since their solutions are not used in the search process; it suffices to solve the auxiliary problem exactly once, at the termination step. It also is worthy of note that if P and Q are polytopes, then the auxiliary

problems are Linear Programming programs, and one can use the corresponding very efficient pivoting technique.

Let me also note that the outlined scheme is a rather unusual example of an iterative process: the influence of the initial search points does not disappear with time. Indeed, look at our two-dimensional example: we cannot allow ourselves to forget about the very first search point even after thousands of steps, since after this point is removed the convex hull of the search points becomes far from the actual saddle point.

5.5 Exercises: Maximal Inscribed Ellipsoid

The below series of exercises deals with the following geometrical problem which is important for efficient implementation of the cutting plane scheme:

($\mathcal{P}$): *Given a solid $G \subset \mathbf{R}^n$, find among all ellipsoids*

$$W = W(B,c) = \{x = Bu + c \mid u^T u \leq 1\} \tag{5.5.1}$$

that one of the maximal volume.

Let us start with the existence issues.

Exercise 5.5.1 $^{\#}$ *Prove that problem $\mathcal{P}$ is solvable.*

Our next task is to demonstrate that $\mathcal{P}$ is a convex optimization problem. The first step is easy:

Exercise 5.5.2 $^{\#*}$ *Prove that if W is an ellipsoid, then one can find a representation (5.5.1) of W involving a positive semidefinite symmetric B.*

According to Exercise 5.5.2, in what follows we may restrict ourselves with PD-representations of ellipsoids, i.e., with representations of the form (5.5.1) with symmetric positive definite B. Thus, from now on all matrices in question, if the opposite is not explicitly stated, are symmetric positive definite.

Now let us prove that the set $\mathcal{W}(Q)$ of pairs (B,c) corresponding to ellipsoids $W(B,c)$ contained in a closed convex domain $Q \subset \mathbf{R}^n$, int $Q \neq \emptyset$, is convex.

Exercise 5.5.3 $^{\#}$

1) Prove that if Q is a half-space

$$Q = \{x \in \mathbf{R}^n \mid a^T x \leq b\},$$

then $\mathcal{W}(Q)$ is open convex set given by the constraints

B is symmetric positive definite;

$|Ba|_2 + a^T c \leq b$.

Derive from this observation that $\mathcal{W}(Q)$ is convex whenever Q is convex with a nonempty interior.

2) What are the constraints defining $\mathcal{W}(Q)$ in the case when Q is a polyhedral set $\{x \in \mathbf{R}^n \mid a_i^T x \leq b_i,\ i = 1, ..., n\}$?

We have demonstrated that the feasible set of problem $\mathcal{P}$ is convex. Now let us look at the objective. Let

$$\mathrm{vol}\, W(B,c) = \mathrm{Det}\ B$$

(this is nothing but the ratio of volumes of $W(B,c)$ and the unit Euclidean ball), and let

$$F(B) \equiv -\ln \mathrm{Det}\ B = -\ln \mathrm{vol}\, W(B,c).$$

Note that $\mathcal{P}$ can be equivalently rewritten as

($\mathcal{P}'$) : *given a solid $G \subset \mathbf{R}^n$, find $(B,c) \in \mathcal{W}(G)$ which minimizes the function $F(B)$.* We already know that the feasible domain of $\mathcal{P}'$ is convex. Now let us prove that the objective $F(\cdot)$ also is convex, so that $\mathcal{P}'$ is a convex optimization problem.

Exercise 5.5.4 $^{\#+}$ *Prove that the function $F(B) = -\ln \mathrm{Det}\ B$ is convex on the (open) convex cone $\mathcal{S}_n^+$ comprised of $n \times n$ positive semidefinite matrices.*

Note that the function

$$\mathrm{EllIn}(G) = \max\{\mathrm{vol}^{1/n} W(B,c) \mid W(B,c) \in G\} = \exp\{F^*/n\},$$

F^* being the optimal value in $\mathcal{P}'$, clearly is a size. The goal of the remaining exercises is to establish the following theorem of L. Khachiyan:

Let (B^*, c^*) be an optimal solution to $\mathcal{P}'$ and let e be a nonzero vector. Consider the set

$$\bar{G} = \{x \in G \mid (x - c^*)^T e \le 0\}.$$

Then

$$\mathrm{EllIn}(\bar{G}) \le \exp\{-\frac{\omega}{n}\}\mathrm{EllIn}(G), \tag{5.5.2}$$

ω being certain positive absolute constant.

I believe you understand the meaning of this theorem: it says that in the cutting plane scheme it is possible to decrease certain size, namely, EllIn$(\cdot)$, of current localizers at, basically, the same rate as that one at which the Center-of-Gravity method reduces the size AvDiam$(\cdot)$. To decrease the size EllIn at the indicated rate, it suffices to choose the current search point as the center of the maximal in volume ellipsoid inscribed into the previous localizer. Since the rate of convergence of a cutting plane method can be expressed in terms of the rate at which the method decreases (some) size of the localizers, the *Inscribed Ellipsoid method* which implements the indicate tactics solves convex problems within relative accuracy ε in $O(1)n\ln(1/\varepsilon)$ steps, $O(1)$ being an absolute constant; thus, the method is optimal in complexity, up to an absolute constant factor, provided than ε is small. Now, what is the advantage of the Inscribed Ellipsoid Method as compared to the Center-of-Gravity method which also is optimal? The advantage is that the auxiliary problems arising in the Inscribed Ellipsoid method, i.e., those of finding the extremal inscribed ellipsoids, are, as we just have seen, convex optimization problems and are therefore computationally tractable, in contrast to computationally untractable problem of finding the center of gravity of a solid.

Of course, one cannot implement the outlined "inscribed ellipsoid tactics" literally (how to find the *exact* extremal inscribed ellipsoid in finite time?), but this is not a great difficulty, since the statement of Khachiyan's theorem is stable: the conclusion remains true if $W(B^*, c^*)$ is an

"almost optimal" solution to $\mathcal{P}'$, namely, with absolute inaccuracy in terms of the objective $F(\cdot)$ less than certain once for ever fixed positive quantity (one can take as the inaccuracy something like 0.01). And to find an almost optimal, in the indicated sense, inscribed ellipsoid, this is a problem which can be solved in finite time, say, by the known to us Ellipsoid method (which now would be better to call the "Circumscribed Ellipsoid" algorithm). It turns out, that if G is a polytope in $\mathbf{R}^n$ defined by m linear inequality constraints, then to find an "almost optimal" inscribed ellipsoid it takes somewhere $O(m^8)$ arithmetic operations. Although this is too much for practice, theoretically it is an important result: we see that it is possible to solve convex optimization problems with n variables in the theoretically optimal number $O(n\ln(1/\varepsilon)$ oracle calls, ε being the required relative accuracy, and with polynomial in n and $\ln(1/\varepsilon)$ arithmetic cost of processing the answers of the oracle. It should be also mentioned that the particular convex optimization problem $\mathcal{P}'$ we are interested in can be solved by much more efficient technique than that one given by the "universal" Circumscribed Ellipsoid method; the latest results state that in order to solve, within relative accuracy ε, a convex optimization problem on an n-dimensional box, it suffices to perform $O(n\ln(1/\varepsilon))$ steps of the slightly modified Inscribed Ellipsoid method, the arithmetic cost of a step being, up to a logarithmic in n factor, $O(n^{3.5})$ operations.

The goal of the remaining exercises is to prove Khachiyan's theorem. We are not going to deal with the version related to "almost optimal" inscribed ellipsoids; from the below proof one can easily see that in fact the reasoning is "stable" and can be word by word extended onto the case when $W(B^*, c^*)$ is "almost optimal" rather than optimal.

Let us first understand what should be proved. Our goal is to demonstrate that if $W^* = W(B^*, c^*)$ is the extremal inscribed ellipsoid for a solid G, $\bar{G}$ is the part of G bounded by a hyperplane passing through the center c^* of W^* and W is an arbitrary ellipsoid contained in $\bar{G}$, then

$$\mathrm{vol}\, W \le (1-\kappa)\,\mathrm{vol}\, W^*$$

for certain absolute constant $\kappa > 0$. This statement is affine invariant, so that when proving it, it suffices to consider the case when W^* is the unit Euclidean ball centered at the origin. Now, what we for sure know about W is that the interior of the latter ellipsoid does not contain center of W^*, i.e., the origin (indeed, W is contained in a half-space with the boundary passing through the center of W^*). Last, we know that W^* is the largest ellipsoid contained in the convex hull G^* of W^* and W.

Thus, we come to the task as follows.

We are given an ellipsoid W with the interior not containing the origin and the unit Euclidean ball W^* centered at the origin. It is known that W^* is the ellipsoid of the maximal volume contained in the convex hull G^* of W^* and W. We should prove that then

$$\mathrm{vol}\, W \le (1-\kappa)\,\mathrm{vol}\, W^* \equiv (1-\kappa)$$

for certain absolute constant $\kappa > 0$.

Our plan is: let us fix $\kappa \in (0, 0.01]$ and assume that the statement in question is *not* valid for this κ, so that the data (i.e., W) in the above statement satisfy the inequality

$$\mathrm{vol}\, W > (1-\kappa) \tag{5.5.3}$$

opposite to that one we should prove. And we shall prove that under this assumption κ cannot be small, thus coming to the desired result.

In what follows $O(1)$ denotes various absolute constants, c is the center of W and $(1-\lambda_i)$, $i=1,...,n$, are the half-axes of the ellipsoid, so that

$$W = W(B,c),$$

B being a symmetric positive definite matrix with the eigenvalues $(1-\lambda_i)$, $i=1,...,n$.

Exercise 5.5.5 $^>$ *Prove that*

$$0 \le \sum_i \lambda_i \le O(1)\kappa \tag{5.5.4}$$

and

$$\sum_i \lambda_i^2 \le O(1)\kappa. \tag{5.5.5}$$

Besides this,

$$|c|_2 \ge 1 - O(1)\sqrt{\kappa}. \tag{5.5.6}$$

Now let V be the $(n-1)$-dimensional ellipsoid which is the cross-section of W with the affine hyperplane $\Pi_c = \{x \mid (x-c)^T c = 0\}$ passing through c orthogonal to c, and let V^* be the intersection of W^* with the parallel to Π_c hyperplane $\Pi_0 = \{x \mid x^T c = 0\}$. Let $1-\mu_i$, $i=1,...,n-1$, be the half-axes of the ellipsoid V. It is convenient to choose the coordinate axes in such a way that the first $(n-1)$ would be parallel to the principal axes of the ellipsoid V and the last, n-th axis would be collinear to c. Let us denote by s the collection of the first $(n-1)$ coordinates of a point with respect to these axes and by t the last coordinate. In the indicated coordinates the situation is as follows:

A. The set G^* contains the "disk"

$$V^* = \{(s,0) \mid \sum_{i=1}^{n-1} s_i^2 \le 1\}$$

and the "ellipsoidal disk"

$$V = \{(s,\tau) \mid s_i = (1+\mu_i)u_i, \sum_{i=1}^{n-1} u_i^2 \le 1\}, \quad \tau = |c|_2 \ge (1 - O(1)\sqrt{\kappa})$$

B. G^* contains also the unit ball

$$W^* = \{(s,t) \mid t^2 + \sum_{i=1}^{n-1} s_i^2 \le 1\}$$

and the ball

$$U = \{(s,t) \mid (t-\tau)^2 + \sum_{i=1}^{n-1} s_i^2 \le (1 - O(1)\sqrt{\kappa})^2\}$$

(this latter fact holds true since U, under appropriate choice of the corresponding $O(1)$, is contained already in W in view of (5.5.5)). Let us add to these two facts the third as follows:

Exercise 5.5.6 * *One has*

$$-O(1)\sqrt{\kappa} \le \sum_{i=1}^{n-1} \mu_i \le O(1)\sqrt{\kappa} \tag{5.5.7}$$

and

$$\sum_{i=1}^{n-1} \mu_i^2 \le O(1)\kappa. \tag{5.5.8}$$

Now we are prepared to finish our job. To this end let us look at the situation: we know that the unit ball W^* is the maximum volume ellipsoid contained in G^*; on the other hand, G^* contains the convex hull Q of V^* and V, which is "almost" the cylinder $C = \{(s,t) \mid |s|_2 \le 1, 0 \le t \le \tau\}$, and two half-balls, the first of them being

$$W_-^* = \{(s,t) \mid |s|_2^2 + t^2 \le 1, t \le 0\},$$

and the second being

$$W_+ = \{(s,t) \mid |s|_2^2 + (t-\tau)^2 \le (1 - O(1)\sqrt{\kappa})^2\}.$$

Thus, G^* is "almost" the set G^+ comprised of the cylinder C and two half-balls covering the bases of the cylinder; of course, G^+ contains an ellipsoid of the volume "significantly larger" than that one of W^* (we may make the first $(n-1)$ half-axes equal to 1 and the last one equal to $\tau/2+1$, which would result in almost 50% growth of the volume as compared to that one of the unit ball W^*). The difficulty is that G^* is not exactly G^+, and we should look carefully whether or not the difference has serious consequences. The idea, anyhow, is clear: we should construct an ellipsoid $E \in G^*$ centered at the point $c/2 = (0, \tau/2)$ and with the principal axes parallel to the coordinate ones; the n-th half-axis should be "large", say, $w = 1.1$, and the preceding half-axis $1 - \omega_i$ should be as large as possible under the restriction that $E \subset G^*$. If we were lucky to demonstrate that, for small enough κ, it is possible to find such a E with

$$\prod_{i=1}^{n-1} (1 - \omega_i) > (1.1)^{-1}, \tag{5.5.9}$$

we would be done: the latter inequality means that $\mathrm{vol}\, E > \mathrm{vol}\, W^*$, and this is impossible since E is an ellipsoid contained in G^*, and W^* is the maximal in volume ellipsoid contained in the latter set. Thus, we would prove that assumption (5.5.3) with small κ leads to a contradiction, as we desire.

Let us start with the simple part of the task.

Exercise 5.5.7 > *Let κ be less than an appropriate absolute constant κ_1, and let the quantities ω_i identifying the above ellipsoid E be less in absolute values than* 0.1. *Then the parts of the ellipsoid E outside the stripe*

$$S = \{(s,t) \mid 0 \le t \le \tau\}$$

are covered by the half-balls W_-^* and W_+ and therefore are contained in G^*.

Now the concluding, a slightly more difficult, part of the reasoning:

Exercise 5.5.8 $^>$ *Prove that if $\kappa \le \kappa_2$ with an appropriately chosen absolute constant κ_2 and if ω_i are given by*

$$1-\omega_i = \min_{|r|\tau<1} \phi_i(r),\ \phi_i(r) = \frac{1-0.5\mu_i - 1.1r\mu_i}{\sqrt{1-\tau^2 r^2}}, \tag{5.5.10}$$

i.e.,

$$1-\omega_i = \phi_i(r_i),\ r_i = \frac{2.2\mu_i}{\tau^2(2-\mu_i)}, \tag{5.5.11}$$

then $|\omega_i| \le 0.1$ and, besides this, the part of the ellipsoid E in the stripe S is contained in G^*, so that the whole ellipsoid E is contained in G^* (see the previous exercise).

Demonstrate that

$$|\omega_i - \frac{1}{2}\mu_i| \le O(1)\mu_i^2. \tag{5.5.12}$$

Exercise 5.5.9 $^>$ *Conclude from the previous exercises that the ellipsoid E for small enough positive absolute constant κ satisfies (5.5.9) and thus complete the proof of Khachiyan's theorem.*

Lecture 6

Variational inequalities with monotone operators

We continue considerations related to non-optimization problems with convex structure, and this lecture is devoted to variational inequalities with monotone operators.

6.1 Variational inequalities with monotone operators

Let $F(x) : \mathbf{R}^n \mapsto \mathbf{R}^n$ be a multi-valued mapping with the domain $\mathrm{Dom}\, F \subset \mathbf{R}^n$. These latter words mean that the image $F(x)$ of $x \in \mathrm{Dom}\, F$ is a nonempty set $F(x) \subset \mathbf{R}^n$. We shall be interested in the so called *monotone* operators, i.e., such that

$$(y - y')^T (x - x') \geq 0 \quad \forall x, x' \in \mathrm{Dom}\, F \quad \forall y \in F(x), y' \in F(x').$$

In a while we shall look at important examples of these mappings.

Let G be a closed convex subset in $\mathbf{R}^n$ with a nonempty interior, and let F be a monotone mapping with the domain containing the interior of G. The pair (G, F) defines a *variational inequality*

$$\textit{find} \quad x_* \in G \cap \mathrm{Dom}\, F \quad \textit{and} \quad y_* \in F(x_*) \quad \textit{s.t.} \quad y_*^T (x - x_*) \geq 0 \ \forall x \in G. \tag{6.1.1}$$

The corresponding x_*'s will be called *strong solutions* to the variational inequality. Note that if x_* is a strong solution and $x \in \mathrm{int}\, G$, $y \in F(x)$, then

$$y^T (x - x_*) \geq y_*^T (x - x_*) \geq 0;$$

thus, if x^* is a strong solution to the inequality, then

$$y^T (x - x_*) \geq 0 \ \forall x \in \mathrm{int}\, G \ \forall y \in F(x). \tag{6.1.2}$$

The points $x^* \in G$ satisfying this latter relation are called *weak solutions* to the variational inequality. As we just have seen, each strong solution is a weak solution as well; the inverse statement, as we shall see in a while, generally speaking, is not true; it becomes true under very mild restrictions on the operator. By some reasons (which soon will become clear) it is more

reasonable to look for weak solutions, and this is why in what follows I call these weak solutions simply the *solutions* to the variational inequality given by G and F.

Now it is time to present examples.

1. *Subrgadient of a convex function.* Let $f(x)$ be a finite continuous convex function defined on G ($f(x) = +\infty$ outside G), and let

$$F(x) = \partial f(x)$$

be defined as the set of all subgradients of f at x. Since the subgradients do exist at all interior points of G, the domain of this multi-valued mapping contains the interior of G, so that the pair (G, F) defines a variational inequality. Note that the operator in question is monotone: if $x, x' \in G$ and $f'(x) \in \partial f(x)$, $f'(x') \in \partial f(x')$, then

$$(f'(x))^T(x - x') \geq f(x) - f(x') \geq (f'(x'))^T(x - x').$$

What are the strong solutions to the inequality? Let us demonstrate that these points are nothing but the minimizers of f on G, and that in the case in question weak solutions are exactly the same as strong ones. Indeed, if x_* is a weak solution, then, for any $x \in \text{int } G$ one has

$$y^T(x - x_*) \geq 0, \ y \in \partial f(x);$$

it follows that the convex continuous function $f_x(t) = f(x_* + t(x - x_*))$ on the segment $0 \leq t \leq 1$ possesses nonnegative derivative at any point $t > 0$ where it is differentiable, and therefore the function is monotone, so that $f_x(1) = f(x) \geq f_x(0) = f(x_*)$, and since x is an arbitrary point of int G, x_* is a minimizer of f. Thus, any weak solution to the inequality is a minimizer of f. Now, if x_* is a minimizer of f on G, then 0 is a subgradient of f at x^* (recall that we are speaking about subgradients on G), and x_* clearly is a strong solution to the inequality. Thus, weak solutions are minimizers of f, and minimizers of f are strong solutions; since strong solutions always are weak ones, we conclude that all three notions - a weak/strong solution to the variational inequality and a minimizer of f - in the case in question coincide with each other.

Now note that if we "shrink" the subgradient mapping F, i.e., pass from it to another mapping $x \mapsto \bar{F}(x) \subset F(x)$ such that $\bar{F}(x)$ is nonempty when x is in the interior of G, then the set of weak solutions may only increase - indeed, if $y^T(x - x_*) \geq 0$ for all $y \in F(x)$ and all $x \in \text{int } G$, then for sure $y^T(x - x_*) \geq 0$ for all $y \in \bar{F}(x) \subset F(x)$ for all $x \in \text{int } G$; at the same time, our proof that a weak solution is a minimizer of f still will be valid, so that the set of weak solutions, which initially - before we have shrunk F - was the set of all minimizers of f, after passing from F to $\bar{F}$ will still be the set of all minimizers of f. On the other hand, the set of strong solutions after shrinking the mapping may become smaller, since when passing from $F(x_*)$ to $\bar{F}(x_*)$ we may loose the y which previously demonstrated that x_* is a strong solution. In fact the set of strong solutions may even become empty, as is clear from the following example:

$$G = \mathbf{R}, \ f(x) = |x|.$$

The mapping F here is

$$F(x) = \begin{cases} \{-1\}, & x < 0 \\ [-1, 1], & x = 0 \\ \{1\}, & x > 0 \end{cases}$$

and the strong solution is $x = 0$. If passing from F to $\bar{F}$ we exclude 0 from the set of values of the mapping at $x = 0$, we loose this unique strong solution, but it still remains a weak one.

This basic example is very instructive. We see that the weak solutions are more stable than the strong ones; their existence does not depend on whether we took enough care of $F(x)$ to be not too "small". In fact there is a very strong existence theorem for weak solutions:

Proposition 6.1.1 *Let G be a solid and F be a monotone operator with the domain containing the interior of F. Then the set of weak solutions to the variational inequality defined by (G, F) is a nonempty closed convex subset of G.*

Proof. The set X_* of all weak solutions is given by an infinite system of linear inequalities with respect to $w \in G$:

$$\mathcal{F} = \{y^T(x - w) \geq 0 \mid x \in \text{int } G, y \in F(x)\},$$

so that X_* clearly is closed and convex subset of G. The only thing which should be proved is that X_* is nonempty. To this end let us demonstrate that any *finite* subsystem of the indicated system of inequalities is solvable; this would imply the nonemptiness of X_* in view of the standard compactness reasons ("a nested family of compacts has a point in common").

Thus, consider a finite subsystem

$$y_i^T(x_i - w) \geq 0,\ i = 1, ..., N$$

of $\mathcal{F}$, so that $y_i \in F(x_i)$. Assume that the system has no solutions in G; then, due to the von Neumann Lemma, certain convex combination of the inequalities from the system, the inequality

$$I(w) \equiv \sum_{i=1}^{M} \lambda_i y_i^T(x_i - w) \geq 0$$

also has no solutions in G (cf. similar reasoning for games). On the other hand, the point

$$w = \sum_{i=1}^{M} \lambda_i x_i$$

belongs to int G. Let $y \in F(w)$, from the monotonicity of F it follows that

$$\sum_{i=1}^{M} \lambda_i y_i^T(x_i - w) \geq \sum_{i=1}^{M} \lambda_i y^T(x_i - w) = y^T(\sum_{i=1}^{M} \lambda_i(x_i - w)) = y^T 0 = 0,$$

and w turns out to be a solution to the inequality $I(w) \geq 0$, which is the desired contradiction. ∎

Let us come back to important examples of variational inequalities with monotone operators. The inequality associated with a convex function gives an instructive example, but it does not motivate the importance of the notion of a variational inequality - we know direct ways to deal with a problem of minimizing a convex function, so that what for should we reduce it to a variational inequality? The below examples give the required motivation.

2. *Kuhn-Tucker point of a constrained optimization problem.* Consider an optimization problem

$$\textit{minimize}\ \ f(u)\ \ \textit{s.t.}\quad g_i(u) \leq 0,\ i = 1, ..., m,\ u \in \mathbf{R}^k,$$

and assume for the sake of simplicity that the constraints g_i, $i = 1, ..., m$, same as the objective f are finite and continuously differentiable. A Kuhn-Tucker point of a problem is, by definition, a point $x_* \equiv (u^*, \lambda^*) \in \mathbf{R}^n \equiv \mathbf{R}^k \times \mathbf{R}^m$ such that

$$f'(u_*) + \sum_{i=1}^{m} \lambda_i^* g_i'(u_*) = 0,$$

$\lambda_i^* \geq 0$, $i = 1, ..., m$, $g_i(u^*) \leq 0$, $i = 1, ..., m$, and

$$\lambda_i^* g_i(u_*) = 0.$$

Let us write down a variational inequality which identifies the Kuhn-Tucker points. To this end set

$$F(x) \equiv F(u, \lambda) = (f'(u) + \sum_{i=1}^{m} \lambda_i g_i'(u), -g_1(u), ..., -g_m(u)) : \mathbf{R}^n \to \mathbf{R}^n,$$

$$G = \{x = (u, \lambda) \in \mathbf{R}^n \mid \lambda \geq 0\}.$$

Thus, we have defined a single-valued continuous operator on $\mathbf{R}^n$ and a closed convex domain in the latter space; let us prove that the strong solutions to the associated variational inequality are exactly the Kuhn-Tucker points of the initial constrained optimization problem.

Indeed, strong solutions to our inequality are exactly the pairs $x_* = (u_*, \lambda_*)$ with nonnegative λ^* such that

$$l(v, h) = v^T (f'(u_*) + \sum_i \lambda_i^* g_i'(u_*)) - \sum_i h_i g_i(v_*) \geq 0$$

whenever $v \in \mathbf{R}^k$ and $h \in \mathbf{R}^m$ are such that $h_i \geq 0$ for those i with $\lambda_*^i = 0$. We immediately conclude that

$$f(u_*) + \sum_i \lambda_i^* g_i'(u_*) = 0$$

(otherwise we could make $l(v, 0)$ negative by an appropriate choice of v),

$$g_i(u^*) \leq 0, \; i = 1, ..., m,$$

(otherwise we could make $l(0, h)$ negative by an appropriate choice of $h > 0$) and

$$\lambda_*^i > 0 \Rightarrow g_i(u^*) = 0$$

(otherwise we could make $l(0, h)$ negative by an appropriate choice of h, since the components of h associated with positive λ_*^i have no restrictions on their signs). Thus, any strong solution is a Kuhn-Tucker point.

Now, if $x^* = (u_*, \lambda_*)$ is a Kuhn-Tucker point, then the vector $F(x^*)$ is

$$(0, -g_1(u_*), ..., -g_m(u_*));$$

if $x = (u, \lambda) \in G$, then

$$(F(x^*))^T (x - x^*) = -\sum_i (\lambda^i - \lambda_*^i) g_i(u_*) = -\sum_i \lambda^i g_i(u_*) \geq 0$$

(the second equality comes from the complementary slackness $\lambda_*^i g_i(u_*) = 0$, and the concluding inequality - from the fact that u_* is feasible for the optimization problem and therefore $g_i(u_*) \leq 0$). Thus, Kuhn-Tucker points are strong solutions to the inequality, and the summary of our observations is that in the case in question strong solutions are exactly the same as the Kuhn-Tucker points of the initial problem.

Of course, the operator F should not necessarily be monotone; anyhow, if the problem in question is convex (f and all g_i are convex), then the operator is monotone on G:

$$(u-v)^T(f'(u)+\sum_i \lambda_i g_i'(u) - f'(v) - \sum_i \mu_i g_i'(v)) - \sum_i (\lambda_i - \mu_i)(g_i(u) - g_i(v)) =$$

$$= (u-v)^T(f'(u) - f'(v)) + \sum_i \lambda_i (u-v)^T(g_i'(u) - g_i'(v)) +$$

$$+ \sum_i \Big(\lambda_i - \mu_i)(g_i(u) - g_i(v) - (u-v)^T g_i'(v)\Big) \geq 0.$$

It is easily seen that in the case of an operator coming from a convex problem weak solutions are strong ones, since the operator is single-valued and continuous; and if F is a single-valued and continuous operator on a closed convex domain G, then, passing to limit in the inequality

$$F^T(x_* + t(x - x_*))(x - x_*) \geq 0$$

(x_* is a weak solution to the variational inequality and $x \in \text{int } G$) as $t \to 0$, we come to $F^T(x_*)(x - x^*) \geq 0$, $x \in \text{int } G$, and consequently $F^T(x_*)(x - x^*) \geq 0$, $x \in G$, so that x^* is a strong solution.

Thus, solving a convex optimization problem can be reduced to finding a weak solution to a variational inequality with monotone operator. This also is not a good motivation, since we know direct ways to deal with constrained convex problems. My next (and last) example is free of this shortcoming.

3. *Nash equilibrium.* Consider a game with k participants; i-th of the participants chooses a point $x_i \in G_i \subset \mathbf{R}^{n_i}$, G_i being certain solid; the payment of i-th player is a function

$$f_i(x^1, ..., x^k)$$

of his choice and choices of all other players.

A point $z_* = (x_*^1, ..., x_*^k) \in G = \prod_{i=1}^k G_i$ is called *Nash equilibrium*, if nobody can improve his position by his separate actions:

$$x_*^i \in \operatorname*{Argmin}_{x^i \in G_i} f_i(x_*^1, ..., x_*^{i-1}, x^i, x_*^{i+1}, ..., x_*^k), \ i = 1, ..., k.$$

This is a very natural extension of the notion of a saddle point in a game of two players we have considered last time; the saddle points in the 2-player game with the cost function $f(x, y)$ are nothing but Nash equilibriums in the game with

$$f_1(x, y) = f(x, y), \ f_2(x, y) = -f(x, y)$$

(this is an equivalent reformulation of the initial game which is symmetric with respect to both of the players; previously we were saying that I wish to decrease f and you wish to increase

it, but this is the same as to say that I wish to decrease f_1 by controlling x and you wish to decrease f_2 by controlling y). Note that the Nash game coming from the usual 2-player game is a game with zero sum:

$$f_1(x,y) + f_2(x,y) \equiv 0.$$

This is not the case in a general Nash game: the sum

$$s(z) = \sum_{i=1}^{k} f_i(z)$$

should not be zero. Note also that the notion of Nash equilibrium reflects a very specific understanding of the goals of the game: each player thinks only of himself, there is no cooperation. For games with nonzero sum (or even for games with zero sum and more than two players) this formalization of the motivations of the players might be not so reasonable, so that there is a deep theory of cooperative games where the definition of equilibrium differs from that one given by Nash, but this is not the issue we are interested in now. My goal is to demonstrate that the problem of finding the Nash equilibrium, under reasonable convexity assumptions, can be reduced to solving a variational inequality with a monotone operator. The related assumptions are as follows:

(i) The cost function f_i of i-th player is Lipschitz continuous on G, convex in the variable $x^i \in G_i$ controlled by the player and is concave with respect to the collection of variables controlled by the remaining players (note: with respect to the *collection* $\{x^j, j \neq i\}$, not to x^j's separately), $i = 1, ..., k$;

(ii) The sum $s(z)$ of the cost functions $f_i(z)$ is convex on G.

Given a game possessing the indicated properties, let us associate with the game a mapping $F(z) : G \to \mathbf{R}^{n_1} \times ... \times \mathbf{R}^{n_k}$ defined as follows:

$$F(x^1, ..., x^k) = \partial f_1(\cdot, x^2, x^3, ..., x^k)(x^1) \times$$

$$\partial f_2(x^1, \cdot, x^3, ..., x^k)(x^2) \times ... \times \partial f_k(x^1, ..., x^{k-1}, \cdot)(x^k),$$

where $\partial f(\cdot, y)(x)$ denotes the subgradient set of a convex with respect to the corresponding argument function at the value x of the argument. Note that for the case of 2-player game $F(z)$ is exactly the set of all possible values of the vector $g(z)$ which we used last time.

Let us make the following observation:

Proposition 6.1.2 *Under assumptions (i), (ii) the mapping $F : G \to \mathbf{R}^n$, $n = n_1 + ... + n_k$, is monotone, and each weak solution to the variational inequality defined by (G, F) is a Nash equilibrium. Besides this, F is semibounded on G:*

$$\mathcal{V}(F) = \sup_{z \in \text{int } G} \sup_{z' \in G} \sup_{y \in F(z)} (z' - z)^T y \leq \mathcal{V} \equiv \sum_i (\max_G f_i - \min_G f_i). \tag{6.1.3}$$

Proof. First of all, let us prove monotonicity. Let $u = (u^1, ..., u^k), v = (v^1, ..., v^k) \in G$ and let $U = (U_1, ..., U_k) \in F(u)$, $V = (V_1, ..., V_k) \in F(v)$. Let us set

$$w^i = \frac{u^i + v^i}{2}, \ \delta^i = \frac{u^i - v^i}{2}.$$

Due to the origin of F, we have

$$f_i(u^1,...,u^{i-1},w^i,u^{i+1},...,u^k) \geq f_i(u) - (\delta^i)^T U_i,$$

$$f_i(v^1,...,v^{i-1},w^i,v^{i+1},...,v^k) \geq f_i(v) + (\delta^i)^T V_i,$$

whence

$$f_i(u^1,...,u^{i-1},w^i,u^{i+1},...,u^k) + f_i(v^1,...,v^{i-1},w^i,v^{i+1},...,v^k) \geq f_i(u) + f_i(v) - (\delta_i)^T(U_i - V_i).$$

Since f_i is concave with respect to the collection of all arguments excluding the i-th of them, the left hand side in this inequality is $\leq 2f_i(w)$, and we come to

$$2f_i(w) \geq f_i(u) + f_i(v) - (\delta_i)^T(U_i - V_i).$$

Taking sum in i, we come to

$$2s(w) \geq s(u) + s(v) - \frac{1}{2}(u-v)^T(U-V),$$

or

$$\frac{1}{2}(u-v)^T(U-V) \geq s(u) + s(v) - 2s(w) \geq 0$$

(the latter inequality follows from the convexity of s), as required.

Now, if $x_* = (x_*^1,...,x_*^k)$ is a weak solution to the variational inequality defined by (G,F), then for any i, any $x^i \in \operatorname{int} G_i$ and any $g_i \in \partial f_i(x_*^1,...,x_*^{i-1},\cdot,x_*^{i+1},...,x_*^k)(x^i)$ there exists $g \in F(\bar{x}_i)$, $\bar{x}_i = F(x_*^1,...,x_*^{i-1},x^i,x_*^{i+1},...,x_*^k)$, such that g_i is the i-th component of g; from the fact that x_* is a weak solution and from Lipschitz continuity of f_i it can be easily derived that

$$g^T(\bar{x}_i - x_*) = g_i^T(x^i - x_*^i) \geq 0,$$

so that x_*^i is a weak solution to the variational inequality associated with the monotone operator

$$x^i \mapsto \partial f_i(x_*^1,...,x_*^{i-1},\cdot,x_*^{i+1},...,x_*^k)(x^i);$$

as we already know, this means that x_*^i is a minimizer of $f_i(x_*^1,...,x_*^{i-1},x^i,x_*^{i+1},...,x_*^k)$ over G_i, and therefore x_* indeed is a Nash equilibrium.

It remains to verify (6.1.3), but this is immediate: if $u = (u^1,...,u^k) \in \operatorname{int} G$, $v = (v^1,...,v^k) \in G$ and $U = (u_1,...,u_k) \in F(u)$, then

$$(v^i - u^i)^T U_i \leq f_i(u^1,...,u^{i-1},v^i,u^{i+1},...,u^k) - f_i(u) \leq \max_G f_i - \min_G f_i,$$

whence

$$(v-u)^T U \leq \mathcal{V},$$

as claimed. ∎

6.2 Cutting plane scheme for variational inequalities

I think that the presented examples demonstrate that variational inequalities with monotone operators form an important class of problems; in a sense, this is the widest class of problems with convex structure. What we are interested in is, of course, how to solve these general problems. As always, we restrict ourselves with a "bounded" case, i.e., that one of a variational inequality

$$\textit{find} \quad x_* \in G \ \ \textit{such that} \ \ y^T(x - x_*) \geq 0 \ \forall x \in \text{int } G \ \forall y \in F(x)$$

with G being a solid and F being monotone and semibounded:

$$\mathcal{V}(F) \equiv \sup\{y^T(u - x) \mid x \in \text{int } G, y \in F(x), u \in G\} < \infty.$$

As always, we should start with choosing a good accuracy measure for candidate solutions. To this end let us note that a weak solution to variational inequality is, by definition, a point x_* such that all linear forms $y^T(x - x_*)$ coming from the operator are nonnegative at x_* or, which is the same, all linear forms $y^T(x_* - x)$ are nonpositive at x^*. Thus, the weak solutions are exactly the points of G where the convex function

$$\nu(u) = \sup\{y^T(u - x) \mid x \in \text{int } G, y \in F(x)\}$$

is nonpositive; note that due to semiboundedness this function if finite on G (and $\mathcal{V}(F)$ is exactly the upper bound of this function over $u \in G$). Thus, it is natural to regard this function as the measure of absolute accuracy of candidate solutions and to define a solution of relative inaccuracy $\varepsilon \in (0, 1)$ as any point $x \in G$ such that

$$\nu(x) \leq \varepsilon \mathcal{V}(F).$$

In fact we wish to minimize the convex function $\nu(x)$, but, same as in the case of games, the difficulty is that the function is not given explicitly; moreover, in contrast to the case of games even to compute the value of ν at a point is an untractable problem, since the maximization problem involved into the definition of ν is, generally speaking, non-convex.

Nevertheless, the problem is quite tractable. Assume that we are given ε and have an access to an oracle which, given on input $x \in \text{int } G$, reports on output a vector $g(x) \in F(x)$, and our goal is to solve the variational inequality within relative accuracy ε. How should we act, it is clear: by definition, the solution set is comprised of points x satisfying, for all $z \in \text{int } G$, linear inequalities $(x - z)^T g(z) \leq 0$, and we may use these inequalities for cuts. Thus, we come to the following generic scheme which does not differ from that one for the case of games and is as follows:

Initialization. Choose the initial localizer - a solid $G_0 \supset G$.
i-th step. Given the previous localizer - a solid G_{i-1} -

1) choose i-th search point $x_i \in \text{int } G_{i-1}$;

2) check whether $x_i \in \text{int } G$. If it is not the case, call the step i *non-productive*, choose a nonzero vector e_i which separates x_i and G:

$$(x - x_i)^T e_i \leq 0 \ \ \forall x \in G$$

and go to 3), otherwise call the oracle, x_i being the input. If the returned vector $e_i \equiv g(x_i)$ is zero, terminate and claim that x_i is an exact (and even strong) solution, otherwise call the step i *productive* and go to 3).

3) Define i-th intermediate localizer $\bar{G}_i$ as

$$\bar{G}_i = \{x \in G_{i-1} \mid (x - x_i)^T e_i \leq 0\}.$$

Embed $\bar{G}_i$ into a solid G_i, the updated localizer, and loop.

As always, with the Center-of-Gravity or the Ellipsoid implementation we may ensure certain rate of decreasing appropriately chosen sizes of the sequential localizers. The point, same as in the case of games, is how to aggregate search points x_i into approximate solutions. You should not be surprised by the fact that this question can be resolved exactly as in the case of games.

Assume that we have performed N steps of a cutting plane method and among these steps there were productive ones; let I_N be the set of indices of these productive steps. Same as in the case of games, we form the function

$$g^N(x) = \min_{i \in I_N} g_i(x) \equiv \min_{i \in I_N} (x_i - x)^T g(x_i),$$

find the optimal Lagrange multipliers $\lambda_i^* \geq 0$, $i \in I_N$, with the unit sum:

$$\max_{x \in G} g^N(x) = \max_{x \in G} \sum_{i \in I_N} \lambda_i^* g_i(x)$$

and define N-th approximate solution as the corresponding convex combination of the search points:

$$\bar{x}_N = \sum_{i \in I_N} \lambda_i^* x_i.$$

In the case of games our reasoning was as follows: first, we demonstrated that this aggregate solves the game within absolute inaccuracy $\max_G g^N$, and, second, proved that the cutting plane scheme automatically decreases the latter maximum at the same rate at which it decreases certain size of localizers. Now we shall act completely similarly.

Lemma 6.2.1 *One has*

$$\nu(\bar{x}_N) \leq \max_G g^N. \tag{6.2.1}$$

Proof. Let x be an arbitrary point of int G, and let $y \in F(x)$. From monotonicity of F it follows that

$$y^T(x_i - x) \leq g^T(x_i)(x_i - x) = g_i(x),$$

whence

$$y^T(\bar{x}_N - x) \leq \sum_{i \in I_N} \lambda_i^* g_i(x) \leq \max_G g^N$$

(the concluding inequality follows from the origin of λ_i^*). Since the resulting inequality is valid for all x and all $y \in F(x)$, we come to

$$\nu(\bar{x}_N) \leq \max_G g^N,$$

as claimed. ∎

The fact that the cutting plane scheme decreases $\max_G g^N$ at the same rate at which it decreases certain size of localizers can be formulated as follows:

Lemma 6.2.2 *Let a variational inequality with a semibounded monotone operator be solved by a cutting plane method, and let* Size *be a size. Assume that at certain step N we have*

$$\varepsilon_N \equiv \frac{\text{Size}(G_N)}{\text{Size}(G)} < 1.$$

Then the set I_N of those indices $i \leq N$ which are associated with productive steps is nonempty and

$$\max_{x \in G} g^N(x) \leq \frac{\varepsilon_N}{1-\varepsilon_N} \mathcal{V}(F). \tag{6.2.2}$$

There is actually nothing to prove here, since one can word by word repeat the reasoning used for the saddle point case. Indeed, the latter reasoning was completely independent of what is the origin of the linear forms g_i, only the fact that these are exactly the forms used at the productive steps of the method and that all these forms are uniformly bounded from below; minus this lower bound is exactly the constant factor which appears in the right hand side of (6.2.2). In our case the lower bound for $g_i(x) = (x_i - x)^T g(x_i)$ can be taken equal to $-\mathcal{V}(F)$, simply by definition of the latter quantity, and this is the origin of (6.2.2).

Now we can summarize our considerations as follows:

Proposition 6.2.1 *Assume that we are solving a variational inequality with a monotone semibounded operator by a cutting plane method and let* Size *be a size. If a step N is such that*

$$\varepsilon_N \equiv \frac{\text{Size}(G_N)}{\text{Size}(G)} < 1,$$

then the set I_N of indices $i \leq N$ of the productive steps is nonempty, and one can point out a feasible point $\bar{x}_N$ which solves the inequality within relative inaccuracy $\varepsilon_N/(1-\varepsilon_N)$. To form $\bar{x}_N$, it suffices to solve the auxiliary piecewise linear convex optimization problem

$$\textit{maximize} \quad \min_{i \in I_N} (x_i - x)^T g(x_i) \quad \textit{over} \quad x \in G,$$

find the corresponding optimal Lagrange multipliers λ_i^*, $i \in I_N$ and set

$$\bar{x}_N = \sum_{i \in I_N} \lambda_i^* x_i.$$

As a corollary, we obtain the following statement:

Proposition 6.2.2 *The complexity of solving an n-dimensional variational inequality within relative accuracy $\varepsilon \in (0,1)$ by the Center-of-Gravity version of the cutting plane scheme does not exceed*

$$N_{CG}(\varepsilon) = \lfloor O(1) n \ln(\frac{1+\varepsilon}{\varepsilon}) \rfloor;$$

for the Ellipsoid method, the complexity does not exceed

$$N_{Ell}(\varepsilon) = \lfloor O(1) n^2 \ln\left(\frac{1+\varepsilon}{\beta\varepsilon}\right) \rfloor, \quad \beta = \frac{\text{EllOut}(G)}{\text{EllOut}(G_0)}.$$

In both the estimates, $O(1)$ are appropriately chosen absolute constants.

Of course, the methods possess the same optimality properties as their "optimization" parents.

To conclude, let me explain why we separately studied the case of saddle point problems; now it is completely clear that this was a particular case of variational inequalities. By the same reason one might ask what for should we be interested in convex minimization problems, which also can be reduced to solving variational inequalities with monotone operators. The answer is as follows: as far as the cutting plane scheme is concerned, there actually is no difference between the saddle point problems and variational inequalities, and there is almost no difference between the case of these inequalities and optimization problems without functional constraints (the only difference is that in the optimization case we should not solve auxiliary piecewise linear problems). Nevertheless, the conclusions we make are not completely identical, since they relate to different inaccuracy measures. It is easily seen that our now inaccuracy measure, being restricted to saddle point problems, is dominated by the inaccuracy in terms of the corresponding cost function. In other words, an approximate saddle point of certain absolute inaccuracy solves, within the same absolute inaccuracy, the variational inequality associated with the saddle point problem, but the inverse statement is not true. Thus, last time we dealt with a particular case of the variational problem and have obtained for this case stronger results - although they look exactly as our today ones - because these former results dealt with a more "strict" inaccuracy measure.

Now we say "good bye" to the cutting plane scheme and methods with linear instance-independent convergence for convex problems. This is not because there is nothing to say more - there are other implementations of the scheme, e.g., the *Inscribed Ellipsoid method* which has the same optimal information-based complexity as the Center-of-Gravity method and at the same time, similarly to the Ellipsoid method, results in reasonable auxiliary problems which can be solved in polynomial time; those who have looked through the problems distributed last time know what the method is. There are also other "optimal" in complexity implementations of the scheme, but I shall not speak about these more sophisticated constructions. What should be stressed is that the cutting plane scheme, although extremely simple in its basic ideas, demonstrates ability to solve, with theoretically quite reasonable complexity, very general problems of convex structure. This is due to the intrinsic simplicity of convex optimization, simplicity which is not seen so clear in more traditional optimization schemes.

6.3 Exercises: Around monotone operators

The below problems are motivated by the following natural question: when any weak solution to a variational inequality with a monotone operator is a strong one? This question is not too interesting, but it gives us the possibility to get acquainted with the cornerstone notion of a *maximal* monotone operator.

To motivate this notion, let us look at the example of variational inequality given by the data

$$G = \mathbf{R},\ F_*(x) = \partial|x|;$$

here we are able to "kill" the (unique) strong solution $x = 0$ by deleting from the set of values of F_* at $x = 0$ the required value 0. The resulting operator F still is monotone, but it is "bad" - we can extend the graph of the operator, i.e., the set

$$\mathcal{G}(F) = \{(x, y) \mid x \in \mathrm{Dom}\, F, y \in F(x)\}$$

to a larger set which still is a graph of a monotone operator.

Note, by the way, that an operator $F : \mathbf{R}^n \to \mathbf{R}^n$ is monotone if and only if its graph is a monotone set in $\mathbf{R}^n \times \mathbf{R}^n$, i.e.,

$$(x, y), (x', y') \in \mathcal{G}(F) \Rightarrow (y - y')^T (x - x') \geq 0.$$

In particular, an operator *inverse* to a monotone one (i.e., with the domain formed by the projection of $\mathcal{G}(F)$ onto the y-space and naturally defined sets of values) also is monotone, since both the operators have "symmetric" graphs (or, better to say, the same graph), and the property of the monotonicity of a graph is kept under swapping x and y.

Thus, one of the reasons for a weak solution not to be a strong one is "non-maximality" of the corresponding operator. The fundamental notion of a *maximal monotone* operator is defined as follows: a monotone operator $F : \mathbf{R}^n \to \mathbf{R}^n$ with a domain $\mathrm{Dom}\, F$ is called *maximal monotone*, if its graph is a maximal monotone set in $\mathbf{R}^n \times \mathbf{R}^n$, i.e., if it is impossible to point out a pair (x, y) not belonging to $\mathcal{G}(F)$ such that the set

$$\mathcal{G}(F) \cup \{(x, y)\}$$

also would be monotone. Note that a monotone operator is maximal if and only if its inverse is maximal.

By standard sophistic (Zorn's Lemma) any monotone operator can be extended (generally speaking, not uniquely) to a maximal monotone one, i.e., the graph of the initial operator can be embedded into the graph of a maximal monotone operator.

Exercise 6.3.1 $^{\#*}$ *Let $F : \mathbf{R}^n \to \mathbf{R}^n$ be a maximal monotone operator. Detect which of the following statements are for sure true and which are, generally speaking, false:*

1) $\mathrm{Dom}\, F$ *is a closed convex set*
2) $\mathrm{Dom}\, F$ *is closed*
3) $\mathrm{Dom}\, F$ *is convex*
4) $\mathcal{G}(F)$ *is a closed convex set*
5) $\mathcal{G}(F)$ *is closed*
6) $\mathcal{G}(F)$ *is convex*
7) $F(x)$ *is a closed convex set for any* $x \in \mathrm{Dom}\, F$
8) $F(x)$ *is a closed set for any* $x \in \mathrm{Dom}\, F$
9) $F(x)$ *is a convex set for any* $x \in \mathrm{Dom}\, F$

The domain of a maximal monotone operator not necessarily is convex, but its non-convexity is nothing but an illusion:

Exercise 6.3.2 $^{\#*}$ *Let $F : \mathbf{R}^n \to \mathbf{R}^n$ be a maximal monotone operator, and let G be the closed convex hull of* $\mathrm{Dom}\, F$. *Prove that* $\mathrm{int}\, G \subset \mathrm{Dom}\, F$.

A basic example of a monotone operator is as follows: let $f : \mathbf{R}^n \to \mathbf{R}^+ \equiv \mathbf{R} \cup \{+\infty\}$ be a *convex* function, and let $\mathrm{Dom}\, f$ be the set of those x with finite $f(x)$. Of course, $\mathrm{Dom}\, f$ is a convex set. Assume that this set possesses a nonempty interior; in this case let us say that f is a *proper* convex function of n variables. Let $x \mapsto f'(x) = \partial f(x)$ be the mapping which maps x

into the set of all subgradients of f at x (this set, by definition, is empty for $x \notin \mathrm{Dom}\, f$; it can be also empty if $x \in \mathrm{Dom}\, f$, but it for sure is nonempty when $x \in \mathrm{int\ Dom}\, f$); thus,

$$\mathrm{int\ Dom}\, f \subset \mathrm{Dom}\, f' \subset \mathrm{Dom}\, f.$$

Exercise 6.3.3 $^{\#*}$ *Prove that if f is a proper convex function of n variables, then f' is a monotone operator. Is this operator maximal monotone?*

Exercise 6.3.4 $^{\#*}$ *Let f be a proper convex function of n variables. Consider the following properties of f:*

1) f' is maximal monotone;

2) f is lower semicontinuous;

3) the graph $\mathcal{G}(f) = \{(x,t) \mid x \in \mathrm{Dom}\, f,\ t \geq f(x)\}$ is a closed set;

4) for all $x \in \mathbf{R}^n$ one has

$$f(x) = \sup_{(u,v)\in\mathcal{G}(f')} \{f(u) + v^T(x-u)\}.$$

Prove that 2)⇔3)⇔4)⇒1) (concluding ⇒ is not a misprint).

The latter exercise immediately provides us with a number of interesting examples of maximal monotone operators, e.g., with the following one: let G be a closed convex domain in $\mathbf{R}^n$ (as always, "domain" means nonemptyness of the interior). Given a finite continuous convex function f on G, we may extend it outside G by the value $+\infty$, and the resulting function, of course, will be convex and lower semicontinuous (and proper). The subdifferential mapping f' of the resulting function is, as we now know, a maximal monotone operator.

Exercise 6.3.5 *Let us apply the aforementioned construction to the function $f \equiv 0$ (recall that this defines f only on G; outside G $f = +\infty$). Prove that the resulting monotone mapping f' is as follows: the domain of it is G; $f'(x) = \{0\}$ for $x \in \mathrm{int}\, G$, and at the boundary of G there is a "crown": if $x \in \partial G$, then $f'(x)$ is comprised of all functionals e such that $e^T(z-x) \leq 0$ for all $z \in G$; nonzero functionals e with this property are called normal to G at x, and the set of these normal functionals is nonempty at any $x \in \partial G$.*

The maximal monotone operator associated with a closed convex domain G in the way presented in the latter exercise will be denoted $\mathcal{N}_G$.

Given a pair of monotone operators $F_1, F_2 : \mathbf{R}^n \to \mathbf{R}^n$ with $\mathrm{Dom}\, F_1 \cap \mathrm{Dom}\, F_2 \neq \emptyset$, one can define the *sum* $F_1 + F_2$ of the operators as follows:

$$\mathrm{Dom}(F_1 + F_2) = \mathrm{Dom}\, F_1 \cap \mathrm{Dom}\, F_2;$$

$$(F_1 + F_2)(x) = \{y_1 + y_2 \mid y_1 \in F_1(x), y_2 \in F_2(x)\},\ x \in \mathrm{Dom}(F_1 + F_2).$$

Exercise 6.3.6 $^{\#}$ *Prove that the sum of monotone operators also is monotone.*

One of the basic results in the theory of variational inequalities is given by the following Theorem of Rockafellar:

let F_1 and F_2 be maximal monotone, and let int Dom F_1 *intersect* int Dom F_2. *Then the operator $F_1 + F_2$ also is maximal monotone.*

With the Theorem of Rockafellar in hands, we can resolve the question we started with.

Exercise 6.3.7 $^{\#*}$ *Prove that if G is a closed convex domain in $\mathbf{R}^n$ and F is a maximal monotone operator such that* Dom $F \supset$ int G, *and F is semibounded on* int G:

$$\sup\{(u-x)^T y \mid u \in G, (x,y) \in \mathcal{G}(F)\} < \infty,$$

then any weak solution to the variational inequality associated with (G, F) is a strong solution to the inequality.

The latter exercise gives rather mild sufficient conditions for a weak solution to be a strong one. The concluding exercise deals with stronger conditions; as a compensation for the weaker result, the proof is much simpler and does not use any serious tools.

Exercise 6.3.8 $^{\#}$ *Let G be a closed convex domain in $\mathbf{R}^n$ and $F : G \to \mathbf{R}^n$ be a single-valued and continuous monotone operator. Prove that then every weak solution to the variational inequality given by (G, F) is a strong solution to the inequality.*

Lecture 7

Large-scale optimization problems

7.1 Goals and motivations

Today we start a new topic - complexity and efficient methods of large-scale convex optimization. Thus, we come back to problems of the type

$$(p) \quad \textit{minimize } f(x) \textit{ s.t. } g_i(x) \leq 0,\ i = 1, ..., m,\ x \in G,$$

where G is a given solid in $\mathbf{R}^n$ and $f, g_1, ..., g_m$ are convex continuous on G functions. The family of all *consistent* problems of the indicated type was denoted by $\mathcal{P}_m(G)$, and we are interested in finding ε-solution to a problem instance from the family, i.e., a point $x \in G$ such that

$$f(x) - \min_G f \leq \varepsilon(\max_G f - \min_G f),\ g_i(x) \leq \varepsilon(\max_G g_i)_+,\ i = 1, ..., m.$$

We have shown that the complexity of the family in question satisfies the inequalities

$$\lfloor O(1) n \ln(1/\varepsilon) \rfloor \leq \mathrm{Compl}(\varepsilon) \leq \rfloor 2.182 n \ln(1/\varepsilon) \lfloor, \tag{7.1.1}$$

where $O(1)$ is an appropriate positive absolute constant; what should be stressed that the upper complexity bound holds true for all $\varepsilon \in (0, 1)$, while the lower one is valid only for not too large varepsilon, namely, for

$$\varepsilon < \varepsilon(G).$$

The "critical value" $\varepsilon(G)$ depends, as we remember, on affine properties of G; for the box it is $\frac{1}{2}$, and for *any* n-dimensional solid G one has

$$\varepsilon(G) \geq \frac{1}{2n^3}.$$

Thus, our complexity bounds identify complexity, up to an absolute constant factor, only for small enough values of ε; there is an initial interval of values of the relative accuracy

$$\Delta(G) = [\varepsilon(G), 1)$$

where we to the moment have only an upper bound on the complexity and have no lower bound. Should we be bothered by this incompleteness of our knowledge? I think we should. Indeed,

what is the initial segment, it depends on G; if G is a box, then this segment is once for ever fixed, so that there, basically, is nothing to worry about - one hardly might be interested in solving optimization problems within relative inaccuracy $\geq 1/2$, and for smaller ε we know the complexity. But if G is a more general set than a box, then there is something to think about: all we can say about an arbitrary n-dimensional G is that $\varepsilon(G) \geq 1/(2n^3)$; this lower bound tends to 0 as the dimension n of the problem increases, so that for large n $\Delta(G)$ can in fact "almost cover" the interval $(0, 1)$ of all possible values of ε. On the other hand, when solving large-scale problems of real world origin, we often are not interested in too high accuracy, and it may happen that the value of ε we actually are interested in is exactly in $\Delta(G)$, where we do not know what is the complexity and what are the optimal methods. Thus, we have reasons, both of theoretical and practical origin, to be interested in the pre-asymptotic behaviour of the complexity.

The difficulty in investigating the behaviour of the complexity in the initial range of values of the accuracy is that this behaviour depends on affine properties of the domain G, and this is something too diffuse for quantitative description. This is why it is reasonable to restrict ourselves with certain "standard" domains G. We already know what happens when G is a parallelotope, or, which is the same, a box - in this case there, basically, is no initial segment. And, of course, the next interesting case is when G is an ellipsoid, or, which is the same, an Euclidean ball (all our notions are affine invariant, so to speak about Euclidean balls is the same as to speak about arbitrary ellipsoids). This is the case we shall focus on. In fact we shall assume G to be something like a ball rather than a ball exactly. Namely, let us fix a real $\alpha \geq 1$ and assume that the *asphericity* of G is at most α, i.e., there is a pair of concentric Euclidean balls $V_{\rm in}$ and $V_{\rm out}$ with the ratio of radii not exceeding α and such that the smaller ball is inside G, and the larger one contains G:

$$V_{\rm in} \subset G \subset V_{\rm out}.$$

7.2 The main result

My main goal is to establish the following

Theorem 7.2.1 *The complexity of the family $\mathcal{P}_m(G)$ of convex problems with m functional constraints on an n-dimensional domain G of asphericity α satisfies the bounds*

$$\min\{n, \lfloor \frac{1}{(2\alpha\varepsilon)^2} - 0 \rfloor\} \leq \mathrm{Compl}(\varepsilon) \leq \lfloor 4\frac{\alpha^2}{\varepsilon^2} \rfloor + 1, \ 0 < \varepsilon < 1; \tag{7.2.1}$$

here $\lfloor a - 0 \rfloor$ is the largest integer which is smaller than a real a.

Comment. Before proving the theorem, let us think what the theorem says. First, it says that the complexity of convex minimization on a domain similar to an Euclidean ball *is bounded from above uniformly in the dimension* by a function $O(1)\alpha^2\varepsilon^{-2}$; the asphericity α is responsible for the level of similarity between the domain and a ball. Second, we see that in the large-scale case, when the dimension of the domain is large enough for given α and ε, or, which is the same, when the inaccuracy ε is large enough for a given dimension (and asphericity), namely, when

$$\varepsilon \geq \frac{1}{2\alpha\sqrt{n}}, \tag{7.2.2}$$

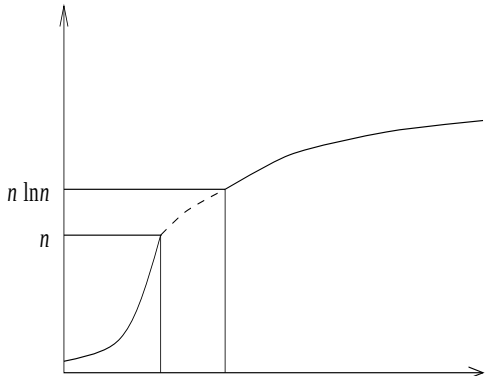


Figure 7.1: Complexity of convex minimization over an n-dimensional Euclidean ball. The whole range $[1, \infty)$ of values of $1/\varepsilon$ can be partitioned into three segments:
• the *initial segment* $[1, 2\sqrt{n}]$; within this segment the complexity, up to an absolute constant factor, is ε^{-2}; at the right endpoint of the segment the complexity is equal to n; in this initial segment the complexity is independent on the dimension and is in fact defined by the affine geometry of G;
• the *final segment* $[\frac{1}{\varepsilon(G)}, \infty) = [2n, \infty)$; here the complexity, up to an absolute constant factor, is $n \ln(1/\varepsilon)$, this is the standard asymptotics known to us; in this final segment the complexity forgets everything about the geometry of G;
• the *intermediate segment* $[2\sqrt{n}, 2n]$; at the left endpoint of this segment the complexity is $O(n)$, at the right endpoint it is $O(n \ln n)$; within this segment we know complexity up to a factor of order of its logarithm rather than up to an absolute constant factor.

then the complexity admits a lower bound $O(1)\alpha^{-2}\varepsilon^{-2}$ which differs from the aforementioned upper bound by factor $O(1)\alpha^{-4}$ which depends on asphericity only. Thus, in the large-scale case (7.2.2) our upper complexity bound coincides with the complexity up to a factor depending on asphericity only; if G is an Euclidean ball ($\alpha = 1$), then this factor does not exceed 16.

Now, our new complexity results combined with the initial results related to the case of small inaccuracies gives us basically complete description of the complexity in the case when G is an Euclidean ball. The behaviour of complexity in this case is as shown on Figure 7.1.

Now let me prove the theorem.

7.3 Upper complexity bound: the Gradient Descent

The upper complexity bound is associated with one of the most traditional methods for nonsmooth convex minimization - the *short-step Subgradient Descent method*. The generic scheme of the method is as follows:

given a starting point $x_1 \in \mathrm{int}\, G$ *and the required relative accuracy* $\varepsilon \in (0, 1)$, *form the sequence*

$$x_{i+1} = x_i - \gamma_i R e_i, \tag{7.3.1}$$

where $\gamma_i > 0$ *is certain stepsize,* R *is the radius of the Euclidean ball* V_{out} *covering* G *and the unit vector* e_i *is defined as follows:*

(i) *if $x_i \notin \mathrm{int}\, G$, then e_i is an arbitrary unit vector which separates x_i and G:*

$$(x - x_i)^T e_i \le 0,\ x \in G;$$

(ii) *if $x_i \in \mathrm{int}\, G$, but there is a constraint g_j which is "ε-violated" at x_i, i.e., is such that*

$$g_j(x_i) > \varepsilon \left(\max_{x \in G} \{ g_j(x_i) + (x - x_i)^T g_j'(x_i) \} \right)_+, \tag{7.3.2}$$

then

$$e_i = \frac{1}{|g_j'(x_i)|_2} g_j'(x_i);$$

(iii) *if $x_i \in \mathrm{int}\, G$ and no constraint is ε-violated at x_i, i.e., no inequality (7.3.2) is satisfied, then*

$$e_i = \frac{1}{|f'(x_i)|_2} f'(x_i).$$

Note that the last formula makes sense only if $f'(x_i) \neq 0$; if in the case of (iii) we meet with $f'(x_i) = 0$, then we simply terminate and claim that x_i is the result of our activity.

Same as in the cutting plane scheme, let us say that search point x_i is *productive*, if at i-th step we meet the case (iii), and *non-productive* otherwise, and let us define i-th approximate solution $\bar{x}_i$ as the best (with the smallest value of the objective) of the productive search points generated in course of the first i iterations (if no productive search point is generated, $\bar{x}_i$ is undefined).

The efficiency of the method is given by the following.

Proposition 7.3.1 *Let a problem (p) from the family $\mathcal{P}_m(G)$ be solved by the short-step Subgradient Descent method associated with accuracy ε, and let N be a positive integer such that*

$$\frac{2 + \frac{1}{2} \sum_{j=1}^N \gamma_j^2}{\sum_{j=1}^N \gamma_j} < \frac{\varepsilon}{\alpha}. \tag{7.3.3}$$

Then either the method terminates in course of N steps with the result being an ε-solution to (p), or $\bar{x}_N$ is well-defined and is an ε-solution to (p).

In particular, if

$$\gamma_i \equiv \varepsilon/\alpha,$$

then (7.3.3) is satisfied by

$$N \equiv N(\varepsilon) = \lfloor 4 \frac{\alpha^2}{\varepsilon^2} \rfloor + 1,$$

and with the indicated choice of the stepsizes we can terminate the method after the N-th step; the resulting method solves any problem from the class within relative accuracy ε with the complexity $N(\varepsilon)$, which is exactly the upper complexity bound stated in Theorem 7.2.1.

Proof. Let me make the following crucial observation: let us associate with the method the localizers

$$G_i = \{ x \in G \mid (x - x_j)^T e_j \le 0,\ 1 \le j \le i \}. \tag{7.3.4}$$

Then the presented method fits our generic cutting plane scheme for problems with functional constraints, up to the fact that G_i now should not necessarily be solids (they may possess empty

interior or even be themselves empty) and x_i should not necessarily be an interior point of G_{i-1}. But all these particularities were not used in the proof of the general proposition on the rate of convergence of the scheme (Lecture 2, Proposition 2.3.1), and in fact there we have proved the following:

Proposition 7.3.2 *Assume that we are generating a sequence of search points $x_i \in \mathbf{R}^n$ and associate with these points vectors e_i and approximate solutions $\bar{x}_i$ in accordance to (i)-(iii). Let the sets G_i be defined by the pairs (x_i, e_i) according to (7.3.4), and let* Size *be a size. Assume that in course of N steps we either terminate due to vanishing the subgradient of the objective at a productive search point, or this is not the case, but*

$$\mathrm{Size}(G_N) < \varepsilon\,\mathrm{Size}(G)$$

(if G_N is not a solid, then, by definition, $\mathrm{Size}(G_N) = 0$*). In the first case the result formed at the termination is an ε-solution to the problem; in the second case such a solution is $\bar{x}_N$ (which is for sure well-defined).*

Now let us apply this latter proposition to our short-step Subgradient Descent method and to the size

$$\mathrm{InnerRad}(Q) = \max\{r \mid Q \text{ contains an Euclidean ball of radius } r\}.$$

We know in advance that G contains an Euclidean ball V_{in} of the radius R/α, so that

$$\mathrm{InnerRad}(G) \geq R/\alpha. \tag{7.3.5}$$

Now let us estimate from above the size of i-th localizer G_i, provided that the localizer is well-defined (i.e., that the method did not terminate in course of the first i steps due to vanishing the subgradient of the objective at a productive search point). Assume that G_i contains an Euclidean ball of certain radius $r > 0$, and let x^+ be the center of the ball. Since V is contained in G_i, we have

$$(x - x_j)^T e_j \leq 0,\ x \in V, 1 \leq j \leq i,$$

whence

$$(x^+ - x_j)^T e_j + h^T e_j \leq 0,\ |h|_2 \leq r,\ 1 \leq j \leq i,$$

and since e_j is a unit vector, we come to

$$(x^+ - x_j)^T e_j \leq -r,\ 1 \leq j \leq i. \tag{7.3.6}$$

Now let us write down the cosine theorem:

$$|x_{j+1} - x^+|_2^2 = |x_j - x^+|_2^2 + 2(x_j - x^+)^T(x_{j+1} - x_j) + |x_{j+1} - x_j|_2^2 =$$

$$= |x_j - x^+|_2^2 + 2\gamma_j R(x^+ - x_j)^T e_j + \gamma_j^2 R^2 \leq |x_j - x^+|_2^2 - 2\gamma_j R r + \gamma_j^2 R^2.$$

We see that the squared distance from x_j to x^+ is decreased with j at least by the quantity $2\gamma_j R - \gamma_j^2 R^2$ at each step; since the squared distance cannot become negative, we come to

$$R(2r\sum_{j=1}^{i}\gamma_j - \sum_{j=1}^{i}\gamma_j^2 R) \leq |x_1 - x^+|_2^2 \leq 4R^2$$

(we have used the fact that G is contained in the Euclidean ball V_{out} of the radius R). Thus, we come to the estimate

$$r \leq \frac{2 + \frac{1}{2}\sum_{j=1}^{i} \gamma_j^2}{\sum_{j=1}^{i} \gamma_j} R$$

This bound acts for the radius r of an arbitrary Euclidean ball contained in G_i, and we come to

$$\text{InnerRad}(G_i) \leq \frac{2 + \frac{1}{2}\sum_{j=1}^{i} \gamma_j^2}{\sum_{j=1}^{i} \gamma_j} R. \tag{7.3.7}$$

Combining this inequality with (7.3.5), we come to

$$\frac{\text{InnerRad}(G_i)}{\text{InnerRad}(G)} \leq \frac{2 + \frac{1}{2}\sum_{j=1}^{i} \gamma_j^2}{\sum_{j=1}^{i} \gamma_j} \alpha, \tag{7.3.8}$$

and due to the definition of N, we come to

$$\frac{\text{InnerRad}(G_N)}{\text{InnerRad}(G)} < \varepsilon,$$

Thus, the conclusion of the Theorem follows from Proposition 7.3.2. ∎

7.4 The lower bound

The lower bound in (7.2.1) is given by a simple reasoning which is in fact already known to us. Due to similarity reasons, we without loss of generality may assume that G is contained in the Euclidean ball of the radius $R = \frac{1}{2}$ centered at the origin and contains the ball of the radius $r = \frac{1}{2\alpha}$ with the same center. It, of course, suffices to establish the lower bound for the case of problems without functional constraints. Besides this, due to monotonicity of the complexity in ε, it suffices to prove that if $\varepsilon \in (0, 1)$ is such that

$$M = \lfloor \frac{1}{(2\alpha\varepsilon)^2} - 0 \rfloor \leq n,$$

then the complexity $\text{Compl}(\varepsilon)$ is at least M. Assume that this is not the case, so that there exists a method $\mathcal{M}$ which solves all problems from the family in question in no more than $M-1$ step. We may assume that $\mathcal{M}$ solves any problem exactly in M steps, and the result always is the last search point. Let us set

$$\delta = \frac{1}{2\alpha\sqrt{M}} - \varepsilon,$$

so that $\delta > 0$ by definition of M. Now consider the family $\mathcal{F}_0$ comprised of functions

$$f(x) = \max_{1 \leq i \leq M} (\xi_i x^i + d_i)$$

where $\xi_i = \pm 1$ and $0 < d_i < \delta$. Note that these functions are well-defined, since $M \leq n$ and therefore we have enough coordinates in x.

Now consider the following M-step construction.

The first step:
let x_1 be the first search point generated by $\mathcal{M}$; this point is instance-independent. Let i_1 be sign of the largest in absolute value of the coordinates of x_1, $\xi^*_{i_1}$ be the index of the coordinate and let $d^*_{i_1} = \delta/2$. Let $\mathcal{F}_1$ be comprised of all functions from $\mathcal{F}$ with $\xi_{i_1} = \xi^*_{i_1}$, $d_{i_1} = d^*_{i_1}$ and $d_i \leq \delta/4$ for all $i \neq i_1$. It is clear that all the functions of the family $\mathcal{F}_1$ possess the same local behaviour at x_1 and are positive at this point.

The second step:
let x_2 be the second search point generated by $\mathcal{M}$ as applied to a problem from the family $\mathcal{F}_1$; this point does not depend on the representative of the family, since all these representatives have the same local behaviour at the first search point x_1. Let i_2 be the index of the largest in absolute value of the coordinates of x_2 *with indices different from* i_1, let $\xi^*_{i_2}$ be the sign of the coordinate, and let $d^*_{i_2} = \delta/4$. Let $\mathcal{F}_2$ be comprised of all functions from $\mathcal{F}_1$ such that $\xi_{i_2} = \xi^*_{i_2}$, $d_{i_2} = d^*_{i_2}$ and $d_i \leq \delta/8$ for all i different from i_1 and i_2. Note that all functions from the family coincide with each other in a neighbourhood of the two-point set $\{x_1, x_2\}$ and are positive at this set.

Now it is clear how to proceed. After k steps of the construction we have a family $\mathcal{F}_k$ comprised of all functions from $\mathcal{F}$ with the parameters ξ_i and d_i being set to certain fixed values for k values $i_1, ..., i_k$ of the index i and all d_i for the remaining i being $\leq 2^{-(k+1)}\delta$; the family satisfies the following predicate

$\mathcal{P}_k$: *the first k points $x_1, ..., x_k$ of the trajectory of $\mathcal{M}$ as applied to any function from the family do not depend on the function, and all the functions from the family coincide with each other in certain neighbourhood of the k-point set $\{x_1, ..., x_k\}$ and are positive at this set.*
From $\mathcal{P}_k$ it follows that the $(k+1)$-th search point x_{k+1} generated by $\mathcal{M}$ as applied to a function from the family $\mathcal{F}_k$ is independent of the function. At the step $k+1$ we

find the index i_{k+1} of the largest in absolute value of the coordinates of x_{k+1} with indices different from $i_1, ..., i_k$,

define $\xi^*_{i_{k+1}}$ as the sign of the coordinate,

set $d^*_{i_{k+1}} = 2^{-(k+1)}\delta$,

and

define $\mathcal{F}_{k+1}$ as the set of those functions from $\mathcal{F}_k$ for which $\xi_{i_{k+1}} = \xi^*_{i_{k+1}}$, $d_{i_{k+1}} = d^*_{i_{k+1}}$ and $d_i \leq 2^{-(k+2)}$ for i different from $i_1, ..., i_{k+1}$.

It is immediately seen that the resulting family satisfies the predicate $\mathcal{P}_{k+1}$, and we may proceed in the same manner.

Now let us look what will be found after M step of the construction. We will end with a family $\mathcal{F}_M$ which consists of exactly one function

$$f = \max_{1 \leq i \leq M} (\xi^*_i x_i + d^*_i)$$

such that f is positive along the sequence $x_1, ..., x_M$ of search points generated by $\mathcal{M}$ as applied to the function. On the other hand, G contains the ball of the radius $r = 1/(2\alpha)$ centered at the origin, and, consequently, contains the point

$$x^* = -\sum_{i=1}^{M} \frac{\xi^*_i}{2\alpha\sqrt{M}} e_i,$$

e_i being the basic orths in $\mathbf{R}^n$. We clearly have

$$f^* \equiv \min_{x\in G} f(x) \le f(x^*) < -\frac{1}{2\alpha\sqrt{M}} + \delta \le -\varepsilon$$

(the concluding inequality follows from the definition of δ). On the other hand, f clearly is Lipschitz continuous with constant 1 on G, and G is contained in the Euclidean ball of the radius 1/2, so that the variation $(\max_G f - \min_G f)$ of f over G is ≤ 1. Thus, we have

$$f(x_M) - f^* > 0 - (-\varepsilon) = \varepsilon \ge \varepsilon(\max_G f - \min_G f);$$

since, by construction, x_M is the result obtained by $\mathcal{M}$ as applied to f, we conclude that $\mathcal{M}$ does not solve the problem f within relative accuracy ε, which is the desired contradiction with the origin of M.

7.5 Exercises: Around Subgradient Descent

The short-step version of the Subgradient Descent presented in Lecture 7 (I hope you are familiar with the lecture) is quite appropriate for proving the upper complexity bound; as a computational scheme, it is not too attractive. The most unpleasant property of the scheme is that it actually is a short-step procedure: one should from the very beginning tune it to the desired accuracy ε, and the stepsizes $\gamma_i = \varepsilon/\alpha$ associated with the upper complexity bound stated in Theorem 7.2.1 should be of order of ε. For the sake of simplicity, let us assume for a moment that G is an Euclidean ball of radius R, so that $\alpha = 1$. With the choice of stepsizes $\gamma_i = \varepsilon/\alpha = \varepsilon$, the method will *for sure* be very slow, since to pass from the starting point x_1 to a reasonable neighbourhood of the optimal solution it normally requires to cover a distance of order of the diameter of the ball G, i.e., of order of R, and it will take *at least* $M = O(1/\varepsilon)$ steps (since a single step moves the point by $\gamma_i R = \varepsilon R$) even in the ideal case when all directions $-e_i$ look directly to the optimal solution. The indicated observation does not contradict the theoretical optimality of the method in the large-scale case, where the worst-case complexity, as we know, is at least is $O(\varepsilon^{-2})$, which is much worse that the above M, namely, something like M^2. Note, anyhow, that right now we were comparing the *best possible worst-case behaviour of a method*, the complexity of the family, and the *best possible ideal-case* behaviour of the short-step Subgradient Descent. And there is nothing good in the fact that even in the ideal case the method is slow.

There are, anyhow, more or less evident possibilities to make the method computationally more reasonable. The idea is not to tune the method to the prescribed accuracy in advance, thus making the stepsizes small from the very beginning, but to start with "large" stepsizes and then decrease them at a reasonable rate. To implement the idea, we need an auxiliary tool (which is important an interesting in its own right), namely, *projections*.

Let Q be a closed and nonempty convex subset in $\mathbf{R}^n$. The *projection* $\pi_Q(x)$ of a point $x \in \mathbf{R}^n$ onto Q is defined as the closest, with respect to the usual Euclidean norm, to x point of Q, i.e., as the solution to the following optimization problem

$$(P_x): \qquad \textit{minimize } |x-y|_2^2 \textit{ over } y \in Q.$$

Exercise 7.5.1 # *Prove that $\pi_Q(x)$ does exist and is unique.*

Exercise 7.5.2 # *Prove that a point $y \in Q$ is a solution to (P_x) if and only if the vector $x - y$ is such that*

$$(u - y)^T(x - y) \leq 0 \ \forall u \in Q. \tag{7.5.1}$$

Derive from this observation the following important property:

$$|\pi_Q(x) - u|_2^2 \leq |x - u|_2^2 - |x - \pi_Q(x)|_2^2. \tag{7.5.2}$$

Thus, when we project a point onto a convex set, the point becomes closer to any point u of the set, namely, the squared distance to u is decreased at least by the squared distance from x to Q.

Derive from (7.5.1) that the mappings $x \mapsto \pi_Q(x)$ and $x \mapsto x - \pi_Q(x)$ are Lipschitz continuous with Lipschitz constant 1.

Now consider the following modification of the Subgradient Descent scheme:
given a solid Q which covers G, a starting point $x_1 \in \text{int } Q$ and a pair of sequences $\{\gamma_i > 0\}$ (stepsizes) and $\{\varepsilon_i \in (0, 1)\}$ (tolerances), form the sequence

$$x_{i+1} = \pi_Q(x_i - \gamma_i \rho e_i), \tag{7.5.3}$$

where 2ρ is the Euclidean diameter of Q and the unit vector e_i is defined as follows:

(i) *if $x_i \notin \text{int } G$, then e_i is an arbitrary unit vector which separates x_i and G:*

$$(x - x_i)^T e_i \leq 0, \ x \in G;$$

(ii) *if $x_i \in \text{int } G$, but there is a constraint g_j which is "ε_i-violated" at x_i, i.e., is such that*

$$g_j(x_i) > \varepsilon_i \left(\max_{x \in G} \{ g_j(x_i) + (x - x_i)^T g_j'(x_i) \} \right)_+, \tag{7.5.4}$$

then

$$e_i = \frac{1}{|g_j'(x_i)|_2} g_j'(x_i);$$

(iii) *if $x_i \in \text{int } G$ and no constraint is ε_i-violated at x_i, i.e., no inequality (7.5.4) is satisfied, then*

$$e_i = \frac{1}{|f'(x_i)|_2} f'(x_i).$$

If in the case of (iii) $f'(x_i) = 0$, then e_i is an arbitrary unit vector.

After e_i is chosen, loop.

The modification, as we see, is in the following:

1) We add projection onto a covering G solid Q into the rule defining the updating $x_i \mapsto x_{i+1}$ and use the half of the diameter of Q as the scale factor in the steplength (in the basic version of the method, there was no projection, and the scale factor was the radius of the ball V_{out});

2) We use time-dependent tactics to distinguish between search points which "almost satisfy" the functional constraints and those which "significantly violate" a constraint.

3) If we meet with a productive point x_i with vanishing subgradient of the objective, we choose as e_i an arbitrary unit vector and continue the process. Note that in the initial version of the method in the case in question we terminate and claim that x_i is an ε-solution; now we also could terminate and claim that x_i is an ε_i-solution, which in fact is the case, but ε_i could be large, not the accuracy we actually are interested in.

Exercise 7.5.3 # *Prove the following modification of Proposition 7.3.1 (Lecture 7):*

Let a problem (p) from the family $\mathcal{P}_m(G)$ be solved by the aforementioned Subgradient Descent method with nonincreasing sequence of tolerances $\{\varepsilon_i\}$. Assume that for a pair of positive integers $N > N'$ one has

$$\Gamma(N' : N) \equiv \frac{2 + \frac{1}{2}\sum_{j=N'}^{N} \gamma_j^2}{\sum_{j=N'}^{N} \gamma_j} < \varepsilon_N \frac{r_{\rm in}}{\rho}, \tag{7.5.5}$$

where $r_{\rm in}$ is the maximal of radii of Euclidean balls contained in G. Then among the search points $x_{N'}, x_{N'+1}, ..., x_N$ there were productive ones, and the best of them (i.e., that one with the smallest value of the objective) point $\bar{x}_{N',N}$ is an $\varepsilon_{N'}$-solution to (p).

Derive from this result that in the case of problems without functional constraints ($m = 0$), where ε_i do not influence the process at all, the relation

$$\varepsilon^*(N) \equiv \min_{N \geq M \geq M' \geq 1} \{\rho \Gamma(M' : M)/r_{\rm in}\} < 1 \tag{7.5.6}$$

implies that the best of the productive search points found in course of the first N steps is well-defined and is an $\varepsilon^*(N)$-solution to (p).

Looking at the statement given by Exercise 7.5.3, we may ask ourselves what could be a reasonable way to choose the stepsizes γ_i and the tolerances ε_i. Let us start with the case of problems without functional constraints, where we can forget about the tolerances - they do not influence the process. What we are interested in is to minimize over stepsizes the quantities $\varepsilon^*(N)$. For a given pair of positive integers $M \geq M'$ the minimum of the quantity

$$\Gamma(N' : N) = \frac{2 + \frac{1}{2}\sum_{j=M'}^{M} \gamma_j^2}{\sum_{j=M'}^{M} \gamma_j}$$

over positive γ_j is attained when $\gamma_j = \frac{2}{\sqrt{M - M' + 1}}$, $M' \leq j \leq M$, and is equal to $\frac{2}{\sqrt{M-M'+1}}$; thus, to minimize $\varepsilon^*(N)$ for a given i, one should set $\gamma_j = \frac{2}{\sqrt{N}}$, $j = 1, ..., N$, which would result in

$$\varepsilon^*(N) = 2N^{-1/2} \frac{\rho}{r_{\rm in}}.$$

This is, basically, the choice of stepsizes we used in the short-step version of the Subgradient Descent; an unpleasant property of this choice is that it is "tied" to N, and we would like to avoid necessity to fix in advance the number of steps allowed for the method. A natural idea is to use the recommendation $\gamma_j = 2N^{-1/2}$ in the "sliding" way, i.e., to set

$$\gamma_j = 2j^{-1/2}, \ j = 1, 2, ... \tag{7.5.7}$$

Let us look what will be the quantities $\varepsilon^*(N)$ for the stepsizes (7.5.7).

Exercise 7.5.4 # *Prove that for the stepsizes (7.5.7) one has*

$$\varepsilon^*(N) \leq \Gamma(\rfloor N/2 \lfloor : N) \frac{\rho}{r_{\rm in}} \leq \kappa N^{-1/2} \frac{\rho}{r_{\rm in}}$$

with certain absolute constant κ. Compute the constant.

We see that the stepsizes (7.5.7) result in optimal, up to an absolute constant factor, rate of convergence of the quantities $\varepsilon^*(N)$ to 0 as $N \to \infty$. Thus, when solving problems without functional constraints, it is reasonable to use the aforementioned Subgradient Descent with stepsizes (7.5.7); according to the second statement of Exercise 7.5.3 and Exercise 7.5.4, for all N such that

$$\varepsilon(N) \equiv \kappa N^{-1/2} \frac{\rho}{r_{\rm in}} < 1$$

the best of the productive search points found in course of the first N steps is well-defined and solves the problem within relative accuracy $\varepsilon(N)$.

Now let us look at problems with functional constraints. It is natural to use here the same rule (7.5.7); the only question now is how to choose the tolerances ε_i. A reasonable policy would be something like

$$\varepsilon_i = \min\{0.9999, 1.01\kappa i^{-1/2} \frac{\rho}{r_{\rm in}}\}, \tag{7.5.8}$$

Exercise 7.5.5 # *Prove that the Subgradient Descent with stepsizes (7.5.7) and tolerances (7.5.8), as applied to a problem (p) from the family $\mathcal{P}_m(G)$, possesses the following convergence properties: for all N such that*

$$\kappa N^{-1/2} \frac{\rho}{r_{\rm in}} < 0.99$$

among the search points $x_{]N/2[}, x_{]N/2[+1}, ..., x_N$ there are productive ones, and the best (with the smallest value of the objective) of these points solves (p) within relative inaccuracy not exceeding

$$\varepsilon_{]N/2[} \le \chi N^{-1/2} \frac{\rho}{r_{\rm in}},$$

χ being an absolute constant.

Note that if one chooses $Q = V_{\rm out}$ (i.e., $\rho = R$, so that $\rho/r_{\rm ini} = \alpha$ is the asphericity of G), then the indicated rate of convergence results in the same (up to an absolute constant factor) as for the basic short-step Subgradient Descent complexity of solving problems from the family within relative accuracy ε.

Lecture 8

Subgradient Descent and Bundle methods

This lecture, same as the previous one, is devoted to large-scale nonsmooth convex problems. For the sake of simplicity I restrict myself to problems without functional constraints:

$$(f) \qquad \textit{minimize} \;\; f(x) \;\; \textit{s.t.} \;\; x \in G \subset \mathbf{R}^n.$$

From now on we assume that G is a closed and bounded convex subset in $\mathbf{R}^n$, possibly, with empty interior, and that the objective is convex and Lipschitz continuous on G:

$$|f(x) - f(y)| \leq L(f)|x - y|, \;\; x, y \in G,$$

where $L(f) < \infty$ and $|\cdot|$ is the usual Euclidean norm in $\mathbf{R}^n$. Note that the subgradient set of f at any point from G is nonempty and contains subgradients of norms not exceeding $L(f)$; from now on we assume that the oracle in question reports such a subgradient at any input point $x \in G$.

We would like to solve the problem within *absolute* inaccuracy $\leq \varepsilon$, i.e., to find $x \in G$ such that

$$f(x) - f^* \equiv f(x) - \min_G f \leq \varepsilon.$$

8.1 Subgradient Descent method

The simplest way to solve the problem is to apply the standard *Subgradient Descent* method which generates the sequence of search points $\{x_i\}_{i=1}^{\infty}$ according to the rule

$$x_{i+1} = \pi_G(x_i - \gamma_i g(x_i)), \;\; g(x) = f'(x)/|f'(x)|, \tag{8.1.1}$$

where $x_1 \in G$ is certain starting point, $\gamma_i > 0$ are positive stepsizes and

$$\pi_G(x) = \mathrm{argmin}\{|x - y| \mid y \in G\}$$

is the standard *projector* onto G. Of course, if we meet a point with $f'(x) = 0$, we terminate with optimal solution at hands; from now on I ignore this trivial case.

As always, i-th approximate solution $\bar{x}_i$ found by the method is the best - with the smallest value of f - of the search points $x_1, ..., x_i$; note that all these points belong to G.

It is easy to investigate the rage of convergence of the aforementioned routine. To this end let x^* be the closest to x_1 optimal solution to the problem, and let

$$d_i = |x_i - x^*|.$$

We are going to see how d_i vary. To this end let us start with the following simple and important observation:

Lemma 8.1.1 *Let $x \in \mathbf{R}^n$, and let G be a closed convex subset in $\mathbf{R}^n$. Under projection onto G, x becomes closer to any point u of G, namely, the squared distance from x to u decreases at least by the squared distance from x to G:*

$$|\pi_G(x) - u|^2 \leq |x - u|^2 - |x - \pi_G(x)|^2. \tag{8.1.2}$$

Proof is immediate. There is nothing to do if $x \in G$, since then $\pi_G(x) = x$. Now, if $x \notin G$, then from the optimality conditions for the problem

$$|x - y| \to \min,\ y \in G$$

it follows that $(u - \pi_G(x))^T(x - \pi_G(x)) \leq 0$ (it is also evident geometrically). Thus, in the triangle $(u\, \pi_G(x)\, x)$ the angle opposite to the side $[ux]$ is acute, and therefore the square of this side is $\geq$ the sum of squares of two other sides. ∎

From Lemma 8.1.1 it follows that

$$\begin{aligned} d_{i+1}^2 \equiv |x_{i+1} - x^*|^2 = |\pi_G(x_i - \gamma_i g(x_i)) - x^*|^2 \leq |x_i - \gamma g(x_i) - x^*|^2 = \\ = |x_i - x^*|^2 - 2\gamma_i (x_i - x^*)^T f'(x_i)/|f'(x_i)| + \gamma_i^2 \leq \\ \leq d_i^2 - 2\gamma_i (f(x_i) - f^*)/|f'(x_i)| + \gamma_i^2 \end{aligned}$$

(the concluding inequality is due to the convexity of f). Thus, we come to the recurrence

$$d_{i+1}^2 \leq d_i^2 - 2\gamma_i (f(x_i) - f^*)/|f'(x_i)| + \gamma_i^2; \tag{8.1.3}$$

in view of the evident inequality

$$f(x_i) - f^* \geq f(\bar{x}_i) - f^* \equiv \varepsilon_i,$$

and since $|f'(x)| \leq L(f)$, the recurrence implies that

$$d_{i+1}^2 \leq d_i^2 - 2\gamma_i \frac{\varepsilon_i}{L(f)} + \gamma_i^2. \tag{8.1.4}$$

The latter inequality allows to make several immediate conclusions.

1) From (8.1.4) it follows that

$$2\sum_{i=1}^{N} \gamma_i \varepsilon_i \leq L(f) \left(d_1^2 + \sum_{i=1}^{N} \gamma_i^2 \right);$$

since ε_i clearly do not increase with i, we come to

$$\varepsilon_N \le \nu_N \equiv \frac{L(f)}{2}\left[\frac{|x_1 - x^*|^2 + \sum_{i=1}^N \gamma_i^2}{\sum_{i=1}^N \gamma_i}\right]. \tag{8.1.5}$$

The right hand side in this inequality clearly tends to 0 as $N \to \infty$, provided that

$$\sum_{i=1}^{\infty} \gamma_i = \infty,\ \gamma_i \to 0, i \to \infty$$

(why?), which gives us certain general statement on convergence of the method as applied to a Lipschitz continuous convex function; note that we did not use the fact that G is bounded.

Of course, we would like to choose the stepsizes resulting in the best possible estimate (8.1.5). The simplest way to do it is to minimize the right hand side in (8.1.5) in γ_i for some given N. This optimization results in

$$\gamma_i = \gamma^* \equiv \frac{|x_1 - x^*|}{\sqrt{N}},\ i = 1, ..., N; \nu_N \equiv \nu_N^* = L(f)\frac{|x_1 - x^*|}{\sqrt{N}};$$

this is the $1/\sqrt{N}$-rate of convergence already known to us. Of course, we cannot use the above stepsizes literally, since they depend on the unknown initial distance $|x_1 - x^*|$ to the optimal set. What we could use are the stepsizes

$$\gamma_i = DN^{-1/2},\ i = 1, ..., N, \tag{8.1.6}$$

D being an a priori upper bound on the diameter of G; for these stepsizes we obtain from (8.1.5) the following upper bound on the absolute accuracy of N-th approximate solution:

$$\varepsilon_N \le L(f)DN^{-1/2}. \tag{8.1.7}$$

A disadvantage of stepsizes (8.1.6) is that we should tune them to certain chosen in advance number of steps N, and for large N enforce the stepsizes to be small from the very beginning. A reasonable way to overcome this disadvantage is to use rule (8.1.6) in a "sliding" mode, i.e., to set

$$\gamma_i = Di^{-1/2},\ i = 1, 2, ... \tag{8.1.8}$$

For these stepsizes, (8.1.5) results in an estimate of the type

$$\varepsilon_i \le \nu_N = O(1)L(f)D\frac{\ln N}{\sqrt{N}}, \tag{8.1.9}$$

which is by logarithmic factor worse than (8.1.8). This logarithmic factor comes from the fact that the series $\sum_{i=1}^N \gamma_i^2$ in our case is of order $D^2 \ln N$. It turns out, anyhow, that the logarithmic factor comes from bad reasoning, not from the fact that the stepsizes (8.1.7) actually are bad. Indeed, note that our basic recurrence (8.1.3) implies that for any $N \ge M \ge 1$ one has

$$2\varepsilon_N \sum_{i=M}^N \gamma_i \le L(f)\left(d_M^2 + \sum_{i=M}^N \gamma_i^2\right) \le L(f)\left(D^2 + \sum_{i=m}^N \gamma_i^2\right),$$

whence

$$\varepsilon_N \le \frac{L(f)}{2}\left[\frac{D^2+\sum_{i=M}^N \gamma_i^2}{\sum_{i=N}^M \gamma_i}\right];$$

with $M = \lfloor N/2 \rfloor$ and $\gamma_i = Di^{-1/2}$ the right hand side in the latter inequality does not exceed $O(1)DN^{-1/2}$, and we come to the optimal, up to an absolute constant factor, estimate

$$\varepsilon_N \le O(1)\frac{L(f)D}{\sqrt{N}},\ N = 1, 2, ... \tag{8.1.10}$$

($O(1)$ is an easily computable absolute constant). I call this rate *optimal*, since the lower complexity bound we have obtained at the previous lecture it fact says that if G is an Euclidean ball of diameter D in $\mathbf{R}^n$ and L is a given constant, then the complexity at which one can minimize over G, within *absolute* accuracy ε, an arbitrary Lipschitz continuous with constant L convex function f is at least

$$\min\left\{n; O(1)\left(\frac{LD}{\varepsilon}\right)^2\right\},$$

so that in the large-scale case, when

$$n \ge \left(\frac{LD}{\varepsilon}\right)^2,$$

the lower complexity bound coincides, within absolute constant factor, with the upper bound given by (8.1.10).

Thus, we can choose the stepsizes γ_i according to (8.1.7) and obtain dimension-independent rate of convergence (8.1.10); this rate of convergence does not admit "significant" uniform in the dimension improvement, provided that G is an Euclidean ball.

2) The stepsizes (8.1.7) are theoretically optimal and more or less reasonable form the practical viewpoint, provided that you deal with a domain G of reasonable diameter, i.e., the diameter of the same order of magnitude as the distance from the starting point to the optimal set. If the latter assumption is not satisfied (as it often is the case), the stepsizes should be chosen more carefully. A reasonable idea here is as follows. Our rate-of-convergence proof in fact was based on a very simple relation

$$d_{i+1}^2 \le d_i^2 - 2\gamma_i(f(x_i) - f^*)/|f'(x_i)| + \gamma_i^2;$$

let us choose as γ_i the quantity which results in the strongest possible inequality of this type, namely, that one which minimizes the right hand side:

$$\gamma_i = \frac{f(x_i) - f^*}{|f'(x_i)|}. \tag{8.1.11}$$

Of course, this choice is possible only when we know the optimal value f^*. Sometimes this is not a problem, e.g., when we reduce a system of convex inequalities

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

to the minimization of

$$f(x) = \max_i f_i(x);$$

here we can take $f^* = 0$. In more complicated cases people use some on-line estimates of f^*; I would not like to go in details, so that I assume that f^* is known in advance. With the stepsizes (8.1.11) (proposed many years ago by B.T. Polyak) our recurrence becomes

$$d_{i+1}^2 \le d_i^2 - (f(x_i) - f^*)^2 |f'(x_i)|^{-2} \le d_i^2 - \varepsilon_i^2 L^{-2}(f),$$

whence $\sum_{i=1}^N \varepsilon_i^2 \le L^2(f) d_1^2$, and we immediately come to the estimate

$$\varepsilon_N \le L(f)|x_1 - x^*| N^{-1/2}. \tag{8.1.12}$$

This estimate seems the best one, since it involves the actual distance $|x_1 - x^*|$ to the optimal set rather than the diameter of G; in fact G might be even unbounded. I believe that whenever one can use the Polyak stepsizes, this is the best possible tactics for the Subgradient Descent method.

Summary. Let me make certain summary. We see that the Subgradient Descent, which we last lecture were exploiting in order to obtain an optimal method for large scale convex minimization over Euclidean ball, can be applied to minimization of a convex Lipschitz continuous function over an arbitrary n-dimensional closed convex domain G; if G is bounded, then, under appropriate choice of stepsizes, one can ensure the inequalities

$$\varepsilon_N \equiv \min_{1 \le i \le N} f(x_i) - f^* \le O(1) L(f) D(G) N^{-1/2}, \tag{8.1.13}$$

where $O(1)$ is a moderate absolute constant, $L(f)$ is the Lipschitz constant of f and $D(G)$ is the diameter of G. If the optimal value of the problem is known, then one can use stepsizes which allow to replace $D(G)$ by the distance $|x_1 - x^*|$ from the starting point to the optimal set; in this latter case, G should not necessarily be bounded. And the rate of convergence is optimal, I mean, it cannot be improved by more than an absolute constant factor, provided that G is an n-dimensional Euclidean ball and $n > N$.

Note also that if G is a "simple" set, say, an Euclidean ball, or a box, or the standard simplex

$$\{x \in \mathbf{R}^n_+ \mid \sum_{i=1}^n x_i = 1\},$$

then the method is computationally very cheap - a step costs only $O(n)$ operations in addition to those spent by the oracle. Theoretically all it looks perfect. It is not a problem to speak about an upper accuracy bound $O(N^{-1/2})$ and about optimality of this bound in the large scale case; but in practice such a rate of convergence would result in tens of thousands of steps, which is too much for the majority of applications. Note that in practice we are interested in "typical" behaviour of a method rather than in its worst case behaviour and worst case optimality. And from this practical viewpoint the Subgradient Descent is far from being optimal: there are other methods with the same worst case theoretical complexity bound, but with significantly better "typical" performance; needless to say that these methods are more preferable in actual computations. What we are about to do is to look at a certain family of methods of this latter type.

8.2 Bundle methods

Common sense says to us that the weak point in the Subgradient Descent is that when running the method, we almost loose previous information on the objective; the whole "prehistory" is compressed to the current iterate (this was not the case with the cutting plane methods, where the prehistory was memorized in the current localizer). Generally speaking, what we actually know about the objective after we have formed a sequence of search points $x_j \in G$, $j = 1, ..., i$? All we know is the *bundle* - the sequence of affine forms

$$f(x_j) + (x - x_j)^T f'(x_j)$$

reported by the oracle; we know that every form from the sequence underestimates the objective and coincides with it at the corresponding search point. All these affine forms can be assembled into a single piecewise linear convex function - *i-th model of the objective*

$$f_i(x) = \max_{1 \le j \le i} \{f(x_j) + (x - x_j)^T f'(x_j)\}.$$

This model underestimates the objective:

$$f_i(x) \le f(x), \ x \in G, \tag{8.2.1}$$

and is exact at the points $x_1, ..., x_i$:

$$f_i(x_j) = f(x_j), \ j = 1, ..., i. \tag{8.2.2}$$

And once again - the model accumulates all our knowledge obtained so far; e.g., the information we possess does not contradict the hypothesis that the model is exact everywhere. Since the model accumulates the whole prehistory, it is reasonable to formulate the search rules for a method in terms of the model. The most natural and optimistic idea is to trust in the model completely and to take, as the next search point, the minimizer of the model:

$$x_{i+1} \in \operatorname*{Argmin}_G f_i.$$

This is the Kelley cutting plane method - the very first method proposed for nonsmooth convex optimization. The idea is very simple - if we are lucky and the model is good everywhere, not only along the previous search points, we would improve significantly the best found so far value of the objective; and if the model is bad, then it will be corrected at the "right" place. From compactness of G one can immediately derive that the method does converge and is even finite if the objective is piecewise linear. Unfortunately, it turns out that the rate of convergence of the method is a disaster; one can demonstrate that the worst-case number of steps required by the Kelley method to solve a problem f within absolute inaccuracy ε (G is the unit n-dimensional ball, $L(f) = 1$) is at least

$$O(1) \left(\frac{1}{\varepsilon}\right)^{(n-1)/2}.$$

We see how dangerous is to be too optimistic, and it is clear why: even in the case of smooth objective the model is close to the objective only in a neighbourhood of the search points; until the number of these points becomes very-very large, this neighbourhood forms a "negligible"

part of the domain G, so that the global characteristic of the model - its minimizer - is very unstable and until the termination phase has small in common with the actual optimal set. Let me say that the Kelley method in practice is much better than one could think looking at its worst-case complexity (a method with practical behaviour like this estimate simply could not be used even in the dimension 10), but the qualitative conclusions from the estimate are more or less valid also in practice - the Kelley method sometimes is too slow.

A natural way to improve the Kelley method is as follows. All we may hope on is that the model approximates the objective in a neighbourhood of the search points. Therefore it is reasonable to enforce the next search point to be not too far from the previous ones, more exactly, from the "most perspective", the best of them, since the latter, as the method goes on, hopefully will become close to the optimal set. To forbid the new iterate to move far away, let us choose x_{i+1} as the minimizer of the *penalized model*:

$$x_{i+1} = \operatorname*{argmin}_G \{f_i(x) + \frac{d_i}{2}|x - x_i^+|^2\},$$

where x_i^+ is what is called the *current prox center*, and the *prox coefficient* $d_i > 0$ is certain parameter. When d_i is large, we enforce x_{i+1} to be close to the prox center, and when it is small, we act almost as in the Kelley method. What is displayed, is the generic form of the *bundle methods*; to specify a method from this family, one need to indicate the policies of updating the prox centers and the prox coefficients. There is a number of reasonable policies of this type, and among these policies there are those resulting in methods with very good practical performance. I would not like to go in details here; let me say only that, first, the best theoretical complexity estimate for the traditional bundle methods is something like $O(\varepsilon^{-3})$; although non-optimal, this upper bound is incomparably better than the *lower* complexity bound for the method of Kelley. Second, there is more or less unique reasonable policy of updating the prox center, in contrast to the policy for updating the prox coefficient. Practical performance of a bundle algorithm heavily depends on this latter policy, and sensitivity to the prox coefficient is, in a sense, the weak point of the bundle methods. Indeed, even without addressing to computational experience we can guess in advance that the scheme should be sensitive to d_i - since in the limiting cases of zero and infinite prox coefficient we get, respectively, the Kelley method, which can be slow, and the "method" which simply does not move from the initial point. Thus, both small and large prox coefficients are forbidden; and it is unclear how to choose the "golden middle" - our information has nothing in common with any quadratic terms in the model, these terms are invented by us.

8.2.1 The Level method

What I am going to do is to present to you the Level method, a new method from the bundle family; this method, in a sense, is free from the aforementioned shortcomings of the traditional bundle scheme. Namely, the method possesses the optimal complexity bound $O(\varepsilon^{-2})$, and the difficulty with tuning the prox coefficient in the method is resolved in a funny way - this problem does not occur at all.

To describe the method, let me introduce several simple quantities. Given i-th model $f_i(\cdot)$, we can compute its optimum, same as in the Kelley method; but now we are interested not in the point where the optimum is attained, but in the optimal value

$$f_i^- = \min_G f_i$$

of the model. Since the model underestimates the objective, the quantity f_i^- is a lower bound for the actual optimal value; and since the models clearly increase with i at every point, their minima also increase, so that

$$f_1^- \le f_2^- \le ... \le f^*. \tag{8.2.3}$$

On the other hand, let f_i^+ be the best found so far value of the objective:

$$f_i^+ = \min_{1\le j\le i} f(x_j) = f(\bar{x}_i), \tag{8.2.4}$$

where $\bar{x}_i$ is the best (with the smallest value of the objective) of the search point generated so far. The quantities f_i^+ clearly decrease with i and overestimate the actual optimal value:

$$f_1^+ \ge f_2^+ \ge ... \ge f^*. \tag{8.2.5}$$

It follows that the *gaps*

$$\Delta_i = f_i^+ - f_i^-$$

are nonnegative and nonincreasing and bound from above the inaccuracy of the best found so far approximate solutions:

$$f(\bar{x}_i) - f^* \le \Delta_i,\ \Delta_1 \ge \Delta_2 \ge ... \ge 0. \tag{8.2.6}$$

Now let me describe the method. Its i-th step is as follows:

1) *solve the piecewise linear problem*

$$\textit{minimize}\ \ f_i(x)\ \ \textit{s.t.}\ \ x \in G$$

to get the minimum f_i^- of the i-th model;

2) *form the level*

$$l_i = (1-\lambda) f_i^- + \lambda f_i^+ \equiv f_i^- + \lambda\Delta_i,$$

$\lambda \in (0,1)$ being the parameter of the method (normally, $\lambda = 1/2$), and define the new iterate x_{i+1} as the projection of the previous one x_i onto the level set

$$Q_i = \{x \in G \mid\ f_i(x) \le l_i\}$$

of the i-th model, the level set being associated with l_i:

$$x_{i+1} = \pi_{Q_i}(x_i). \tag{8.2.7}$$

Computationally, the method requires solving two auxiliary problems at each iteration. The first is to minimize the model in order to compute f_i^-; this problem arises in the Kelley method and does not arise in the bundle ones. The second auxiliary problem is to project x_i onto Q_i; this is, basically, the same quadratic problem which arises in bundle methods and does not arise in the Kelley one. If G is a polytope, which normally is the case, the first of these auxiliary problems is a linear program, and the second is a convex linearly constrained quadratic program; to solve them, one can use the traditional efficient simplex-type technique.

Let me note that the method actually belongs to the bundle family, and that for this method the prox center always is the last iterate. To see this, let us look at the solution

$$x(d) = \operatorname*{argmin}_G\{f_i(x) + \frac{d}{2}|x - x_i|^2\}$$

of the auxiliary problem arising in the bundle scheme as at a function of the prox coefficient d. It is clear that $x(d)$ is the closest to x point in the set $\{x \in G \mid f_i(x) \le f_i(x(d))\}$, so that $x(d)$ is the projection of x_i onto the level set

$$\{x \in G \mid f_i(x) \le l_i(d)\}$$

of the i-th model associated with the level $l_i(d) = f_i(x(d))$ (this latter relation gives us certain equation which relates d and $l_i(d)$). As d varies from 0 to ∞, $x(d)$ moves along certain path which starts at the closest to x_i point in the optimal set of the i-th model and ends at the prox center x_i; consequently, the level $l_i(d)$ varies from f_i^- to $f(x_i) \ge f_i^+$ and therefore, for certain value d_i of the prox coefficient, we have $l_i(d_i) = l_i$ and, consequently, $x(d_i) = x_{i+1}$. Note that the only goal of this reasoning was to demonstrate that the Level method does belong to the bundle scheme and corresponds to certain implicit control of the prox coefficient; this control exists, but is completely uninteresting for us, since the method does not require knowledge of d_i.

Now let me formulate and prove the main result on the method.

Theorem 8.2.1 *Let the Level method be applied to a convex problem (f) with Lipschitz continuous, with constant $L(f)$, objective f and with a closed and bounded convex domain G of diameter $D(G)$. Then the gaps Δ_i converge to 0; namely, for any positive ε one has*

$$i > c(\lambda)\left(\frac{L(f)D(G)}{\varepsilon}\right)^2 \Rightarrow \Delta_i \le \varepsilon, \tag{8.2.8}$$

where

$$c(\lambda) = \frac{1}{(1-\lambda)^2\lambda(2-\lambda)}.$$

In particular,

$$i > c(\lambda)\left(\frac{L(f)D(G)}{\varepsilon}\right)^2 \Rightarrow f(\bar{x}_i) - f^* \le \varepsilon.$$

Proof. The "in particular" part is an immediate consequence of (8.2.8) and (8.2.6), so that all we need is to verify (8.2.8). To this end assume that N is such that $\Delta_N > \varepsilon$, and let us bound N from above.

1^0. Let us partition the set $I = \{1, 2, ..., N\}$ of iteration indices in groups $I_1, ..., I_k$ as follows. The first group ends with the index $i(1) \equiv N$ and contains all indices $i \le i(1)$ such that

$$\Delta_i \le (1-\lambda)^{-1}\Delta_N \equiv (1-\lambda)^{-1}\Delta_{i(1)};$$

since, as we know, the gaps never increase, I_1 is certain final segment of I. If it differs from the whole I, we define $i(2)$ as the largest of those $i \in I$ which are not in I_1, and define I_2 as the set of all indices $i \le i(2)$ such that

$$\Delta_i \le (1-\lambda)^{-1}\Delta_{i(2)}.$$

I_2 is certain preceding I_1 segment in I. If the union of I_1 and I_2 is less than I, we define $i(3)$ as the largest of those indices in I which are not in $I_1 \cup I_2$, and define I_3 as the set of those indices $i \le i(3)$ for which

$$\Delta_i \le (1-\lambda)^{-1}\Delta_{i(3)},$$

and so on.

With this process, we partition the set I of iteration indices into sequential segments I_1,..,I_k (I_s follows I_{s+1} in I). The last index in I_s is $i(s)$, and we have

$$\Delta_{i(s+1)} > (1-\lambda)^{-1}\Delta_{i(s)}, s = 1, ..., k-1 \tag{8.2.9}$$

(indeed, if the opposite inequality would hold, then $i(s+1)$ would be included into the group I_s, which is not the case).

2^0. My main (and very simple) observation is as follows:

the level sets Q_i of the models corresponding to certain group of iterations I_s have a point in common, namely, the minimizer, u_s, of the last, $i(s)$-th, model from the group.

Indeed, since the models increase with i, and the best found so far values of the objective decrease with i, for all $i \in I_s$ one has

$$f_i(u_s) \le f_{i(s)}(u_s) = f^-_{i(s)} = f^+_{i(s)} - \Delta_{i(s)} \le f^+_i - \Delta_{i(s)} \le f^+_i - (1-\lambda)\Delta_i \equiv l_i$$

(the concluding $\le$ in the chain follows from the fact that $i \in I_s$, so that $\Delta_i \le (1-\lambda)^{-1}\Delta_{i(s)}$).

3^0. The above observation allows to estimate from above the number N_s of iterations in the group I_s. Indeed, since x_{i+1} is the projection of x_i onto Q_i and $u_s \in Q_i$ for $i \in I_s$, we conclude from Lemma 8.1.1 that

$$|x_{i+1} - u_s|^2 \le |x_i - u_s|^2 - |x_i - x_{i+1}|^2, \ i \in I_s,$$

whence, denoting by $j(s)$ the first element in I_s,

$$\sum_{i \in I_s} |x_i - x_{i+1}|^2 \le |x_{j(s)} - u_s|^2 \le D^2(G). \tag{8.2.10}$$

Now let us estimate from below the steplengths $|x_i - x_{i+1}|$. At the point x_i the i-th model f_i equals to $f(x_i)$ and is therefore $\ge f^+_i$, and at the point x_{i+1} the i-th model is, by construction of x_{i+1}, less or equal (in fact is equal) to $l_i = f^+_i - (1-\lambda)\Delta_i$. Thus, when passing from x_i to x_{i+1}, the i-th model varies at least by the quantity $(1-\lambda)\Delta_i$, which is, in turn, at least $(1-\lambda)\Delta_{i(s)}$ (the gaps may decrease only!). On the other hand, f_i clearly is Lipschitz continuous with the same constant $L(f)$ as the objective (recall that, according to our assumption, the oracle reports subgradients of f of the norms not exceeding $L(f)$). Thus, at the segment $[x_i, x_{i+1}]$ the Lipschitz continuous with constant $L(f)$ function f_i varies at least by $(1-\lambda)\Delta_{i(s)}$, whence

$$|x_i - x_{i+1}| \ge (1-\lambda)\Delta_{i(s)}L^{-1}(f).$$

From this inequality and (8.2.10) we conclude that the number N_s of iterations in the group I_s satisfies the estimate

$$N_s \le (1-\lambda)^{-2}L^2(f)D^2(G)\Delta^{-2}_{i(s)}.$$

4^0. We have $\Delta_{i(1)} > \varepsilon$ (the origin of N) and $\Delta_{i(s)} > (1-\lambda)^{-s+1}\Delta_{i(1)}$ (see (8.2.9)), so that the above estimate of N_s results in

$$N_s \le (1-\lambda)^{-2}L^2(f)D^2(G)(1-\lambda)^{2(s-1)}\varepsilon^{-2},$$

whence

$$N = \sum_s N_s \le c(\lambda)L^2(f)D^2(G)\varepsilon^{-2},$$

as claimed. ∎

8.2.2 Concluding remarks

The theorem we have proved says that the level method is optimal in complexity, provided that G is an Euclidean ball of a large enough dimension. In fact there exists "computational evidence", based on many numerical tests, that the method is also optimal in complexity in a fixed dimension. Namely, as applied to a problem of minimization of a convex function over an n-dimensional solid G, the method finds an approximate solution of *absolute* accuracy ε in no more than

$$c(f)\, n\, \ln\left(\frac{V(f)}{\varepsilon}\right)$$

iterations, where $V(f) = \max_G f - \min_G f$ is the variation of the objective over G and $c(f)$ is certain problem-dependent constant *which never is greater than 1* and typically is something around 0.2 - 0.5. Let me stress that this is an experimental fact, not a theorem; I have strong doubts that there exists a theorem of this type, but empirically this "law" is supported by hundreds of tests.

To illustrate this point, let me present numerical results related to one of the standard test problems called MAXQUAD. This is a small, although difficult, problem of maximizing the maximum of 5 convex quadratic forms of 10 variables. In the below table you see the results obtained on this problem by the Subgradient Descent and by the Level methods. In the Subgradient Descent the Polyak stepsizes were used (to this end, the method was equipped with the exact optimal value, so that the experiment was in favour of the Subgradient Descent). The results are as follows:

Subgradient Descent: 100,000 steps, best found value -0.8413414 (absolute inaccuracy 0.0007), running time $54''$;

Level: 103 steps, best found value -0.8414077 (absolute inaccuracy ¡0.0000001), running time $2''$, "complexity index" $c(f) = 0.47$.

Runs:

Subgradient Descent *Level*

i	f_i^+	i	f_i^+
1	5337.066429	1	5337.066429
2	98.595071		
		6	98.586465
8	6.295622	16	7.278115
31	0.198568		
		39	−0.674044
41	−0.221810		
		54	−0.811759
		73	−0.841058
		81	−0.841232
		103	−0.841408
201	−0.801369		
4001	$^*-0.839771$		
5001	−0.840100		
17001	−0.841021		
25001	−0.841144		
50001	−0.841276		
75001	−0.841319		
100000	−0.481341		

* marks the result obtained by the Subgradient Descent after $2''$ (the total CPU time of the Level method); to that moment the Subgradient Descent has performed 4,000 iterations, but has restored the solution within 2 accuracy digits rather then 6 digits given by Level.

The Subgradient Descent with the "default" stepsizes $\gamma_i = O(1)i^{-1/2}$ in the same 100,000 iterations was unable to achieve value of the objective less than -0.837594, i.e., found the solution within a single accuracy digit.

8.3 Exercises: Mirror Descent

Looking at the 3-line convergence proof for the standard Subgradient Descent:

$$x_{i+1} = \pi_G(x_i - \gamma_i f'(x_i)/|f'(x_i)|) \Rightarrow |x_{i+1} - x^*|^2 \leq |x_i - x^*|^2 - 2\gamma_i (x_i - x^*)^T f'(x_i)/|f'(x_i)| + \gamma_i^2$$

$$\Rightarrow |x_{i+1} - x^*|^2 \leq |x_i - x^*|^2 - 2\gamma_i (f(x_i) - f^*)/|f'(x_i)| + \gamma_i^2$$

$$\Rightarrow \min_{i \leq N} f(x_i) - f^* \leq \frac{|x_1 - x^*|^2 + \sum_{i=1}^N \gamma_i^2}{2\sum_{i=1}^N \gamma_i} \blacksquare$$

one should be surprised. Indeed, all of us know the origin of the gradient descent: if f is smooth, a step in the antigradient direction decreases the first-order expansion of f and therefore, for a reasonably chosen stepsize, increases f itself. Note that this standard reasoning has nothing in common with the above one: we deal with a *nonsmooth* f, and it should not decrease in the direction of *an* anti-subgradient independently of how small is the stepsize; there is a subgradient in the subgradient set which actually possesses the desired property, but this is not necessarily the subgradient used in the method, and even with the "good" subgradient you could say nothing about the amount the objective can be decreased by. The "correct" reasoning deals

with algebraic structure of the Euclidean norm rather than with local behaviour of the objective, which is very surprising; it is a kind of miracle. But we are interested in understanding, not in miracles. Let us try to understand what is behind the phenomenon we have met.

First of all, what is a subgradient? Is it actually a vector? The answer, of course, is "no". Given a convex function f defined on an n-dimensional vector space E and an interior point x of the domain of f, you can define a nonempty set of *support functionals* - linear forms $f'(x)[h]$ of $h \in E$ which are *support* to f at x, i.e., such that

$$f(y) \geq f(x) + f'(x)[y-x], y \in \mathrm{Dom}\, f;$$

these forms are intrinsically associated with f and x. Now, *having chosen somehow an Euclidean structure* $(\cdot,\cdot)$ *on* E, you may associate with linear forms $f'(x)[h]$ vectors $f'(x)$ from E in such a way that

$$f'(x)[h] = (f'(x), h),\ h \in \mathbf{R}^n,$$

thus coming from support functionals to subgradients-vectors. The crucial point is that these vectors are *not* defined by f and x only; they also depend on what is the Euclidean structure on E we use. Of course, normally we think of an n-dimensional space as of the coordinate space $\mathbf{R}^n$ with once for ever fixed Euclidean structure, but this habit sometimes is dangerous; the problems we are interested in are defined in affine terms, not in the metric ones, so why should we always look at the problems via certain once for ever fixed Euclidean structure which has nothing in common with the problem? Developing systematically this evident observation, one may come to the most advanced and recent convex optimization methods like the polynomial time interior point ones. Our now goal is much more modest, but we also shall get profit from the aforementioned observation. Thus, once more: *the "correct" objects associated with f and x are not vectors from E, but elements of the dual to E space E^* of linear forms on E.* Of course, E^* is of the same dimension as E and therefore it can be identified with E; but there are many ways to identify these spaces, and no one of them is "natural", more preferable than others.

Since the support functionals $f'(x)[h]$ "live" in the dual space, the Gradient Descent cannot avoid the necessity to identify somehow the initial - *primal* - and the dual space, and this is done via the Euclidean structure the method is related to - as it was already explained, this is what allows to associate with a support functional - something which "actually exists", but belongs to the dual space - a subgradient, a vector belonging to the primal space; in a sense, this vector is a phantom - it depends on the Euclidean structure on E. Now, is a Euclidean structure the only way to identify the dual and the primal spaces? Of course, no, there are many other ways. What we are about to do is to consider certain family of "identifications" of E and E^* which includes, as particular cases, all identifications given by Euclidean structures. This family is as follows. Let $V(\phi)$ be a smooth (say, continuously differentiable) convex function on E^*; its support functional $V'(\phi)[\cdot]$ at a point ϕ is a linear functional on the dual space. Due to the well-known fact of Linear Algebra, every linear form $L[\eta]$ on the dual space is defined by a vector l from the primal space, in the sense that

$$L[\eta] = \eta[l]$$

for all $\eta \in E^*$; this vector is uniquely defined, and the mapping $L \mapsto l$ is a linear isomorphism between the space $(E^*)^*$ dual to E^* and the primal space E. This isomorphism is "canonical" -

it does not depend on any additional structures on the spaces in question, like inner products, coordinates, etc., it is given by intrinsic nature of the spaces. In other words, in the quantity $\phi[x]$ we may think of ϕ being fixed and x varying over E, which gives us a linear form on E (this is the origin of the quantity); but we also can think of x as being fixed and ϕ varying over E^*, which gives us a linear form on E^*; and Linear Algebra says to us that every linear form on E^* can be obtained in this manner from certain uniquely defined $x \in E$. Bearing in mind this symmetry, let us from now on rename the quantity $\phi[x]$ as $\langle \phi, x \rangle$, thus using more "symmetric" notation.

Thus, every linear form on E^* corresponds to certain $x \in E$; in particular, the linear form $V'(\phi)[\cdot]$ on E^* corresponds to certain vector $V'(\phi) \in E$:

$$V'(\phi)[\eta] \equiv \langle \eta, V'(\phi) \rangle \, , \eta \in E^*.$$

We come to certain mapping

$$\phi \mapsto V'(\phi) : E^* \to E;$$

this mapping, under some mild assumptions, is a continuous one-to-one mapping with continuous inverse, i.e., is certain identification (*not necessarily linear*) of the dual and the primal spaces.

Exercise 8.3.1 *Let $(\cdot, \cdot)$ be certain Euclidean structure on the dual space, and let*

$$V(\phi) = \frac{1}{2}(\phi, \phi).$$

The aforementioned construction associates with V the mapping $\phi \mapsto V'(\phi) : E \to E^*$; on the other hand, the Euclidean structure in question itself defines certain identification of the dual and the primal space, the identification $\mathcal{I}$ given by the identity

$$\phi[x] = (\phi, \mathcal{I}^{-1}x), \ \phi \in E^*, \, x \in E.$$

Prove that $\mathcal{I} = V'$.

Assume that $\{e_i\}_{i=1}^n$ is a basis in E and $\{e_i^*\}$ is the biorthonormal basis in E^* (so that $e_i^*[e_j] = \delta_{ij}$), and let A be the matrix which represents the inner product in the coordinates of E^* related to the basis $\{e_i^*\}$, i.e., $A_{ij} = (e_i^*, e_j^*)$. What is the matrix of the associated mapping V' taken with respect to the $\{e_i^*\}$-coordinates in E^* and the $\{e_i\}$-coordinates in E?

We see that all "standard" identifications of the primal and the dual spaces, i.e., those given by Euclidean structures, are covered by our mappings $\phi \mapsto V'(\phi)$; the corresponding V's are, up to the factor $1/2$, squared Euclidean norms. A natural question is what are the mappings associated with *other* squared norms.

Exercise 8.3.2 *Let $\| \cdot \|$ be a norm on E, let*

$$\| \phi \|_* = \max\{\phi[x] \mid x \in E, \, \| x \| \leq 1\}$$

be the conjugate norm on and assume that the function

$$V_{\|\cdot\|_*}(\phi) = \frac{1}{2} \| \phi \|_*^2$$

is continuously differentiable. Then the mapping $\phi \mapsto V'(\phi)$ is as follows: you take a linear form $\phi \neq 0$ on E, maximize it on the $\|\cdot\|$-unit ball of E; the maximizer is unique and is exactly $V'(\phi)$; and, of course, $V'(0) = 0$. In other words, $V'(\phi)$ is nothing but the "direction of the fastest growth" of the functional ϕ, where, of course, the rate of growth in a direction is defined as the "progress in ϕ per $\|\cdot\|$-unit step in the direction".

Prove that the mapping $\phi \mapsto \phi'$ is a continuous mapping form E onto E'. Prove that this mapping is a continuous one-to-one correspondence between E and E' if and only if the function

$$W_{\|\cdot\|}(x) = \frac{1}{2} \| x \|^2 : E \to \mathbf{R}$$

is continuously differentiable, and in this case the mapping

$$x \mapsto W'_{\|\cdot\|}(x)$$

is nothing but the inverse to the mapping given by $V'_{\|\cdot\|_*}$.

Now, every Euclidean norm $\|\cdot\|$ on E induces, as we know, a Subgradient Descent method for minimization of convex functions over closed convex domains in E. Let us write down this method in terms of the corresponding function $V_{\|\cdot\|_*}$. For the sake of simplicity let us ignore for the moment the projector π_G, thus looking at the method for minimizing over the whole E. The method, as it is easily seen, would become

$$\phi_{i+1} = \phi_i - \gamma_i f'(x_i) / \| f'(x_i) \|_*, \; x_i = V'(\phi_i), \; V \equiv V_{\|\cdot\|_*}. \tag{8.3.1}$$

Now, in the presented form of the Subgradient Descent there is nothing from the fact that $\|\cdot\|$ is a Euclidean norm; the only property of the norm which we actually need is the differentiability of the associated function V. Thus, given a norm $\|\cdot\|$ on E which induces a differentiable outside 0 conjugate norm on the conjugate space, we can write down certain method for minimizing convex functions over E. How could we analyse the convergence properties of the method? In the case of the usual Subgradient Descent the proof of convergence was based on the fact that the anti-gradient direction $f'(x)$ is a descent direction for certain *Lyapunov function*, namely, $|x - x^*|^2$, x^* being a minimizer of f. In fact our reasoning was as follows: since f is convex, we have

$$\langle f'(x), x^* - x \rangle \leq f(x) - f(x^*) \leq 0, \tag{8.3.2}$$

and the quantity $\langle f'(x), x - x^* \rangle$ is, up to the constant factor 2, the derivative of the function $|x - x^*|^2$ in the direction $f'(x)$. Could we say something similar in the general case, where, according to (8.3.1), we should deal with the situation $x = V'(\phi)$? With this substitution, the left hand side of (8.3.2) becomes

$$\langle f'(x), x^* - V'(\phi) \rangle = \frac{d}{dt} |_{t=0} \, V^+(\phi - t f'(x)), \; V^+(\psi) = V(\psi) - \langle \psi, x^* \rangle .$$

Thus, we can associate with (8.3.1) the function

$$V^+(\phi) \equiv V^+_{\|\cdot\|_*}(\phi) = V_{\|\cdot\|_*}(\phi) - \langle \phi, x^* \rangle , \tag{8.3.3}$$

x^* being a minimizer of f, and the derivative of this function in the direction $-f'(V'(\phi))$ of the trajectory (8.3.1) is nonpositive:

$$\langle -f'(V'(\phi)), (V^+)'(\phi)\rangle \le f(x^*) - f(x) \le 0,\ \phi \in E^*. \tag{8.3.4}$$

Now we may try to reproduce the reasoning which leads to the rate-of-convergence estimate for the Subgradient Descent for our now situation, where we speak about process (8.3.1) associated with an arbitrary norm on E (the norm should result, of course, in a continuously differentiable V).

For the sake of simplicity, let us restrict ourselves with the simple case when V possesses a Lipschitz continuous derivative. Thus, from now on let $\|\cdot\|$ be a norm on E such that the mapping

$$\mathcal{V}(\phi) \equiv V'_{\|\cdot\|_*}(\phi) : E^* \to E$$

is Lipschitz continuous, and let

$$\mathcal{L} \equiv L_{\|\cdot\|} = \sup\{\frac{\| \mathcal{V}(\phi') - \mathcal{V}(\phi'') \|}{\| \phi' - \phi'' \|_*} \mid \phi' \ne \phi'',\ \phi', \phi'' \in E^*\}.$$

For the sake of brevity, from now on we write V instead of $V_{\|\cdot\|_*}$.

Exercise 8.3.3 *Prove that*

$$V(\phi + \eta) \le V(\phi) + \langle \eta, \mathcal{V}(\phi)\rangle + \mathcal{L}V(\eta),\ \phi, \eta \in E^*. \tag{8.3.5}$$

Now let us investigate process (8.3.1).

Exercise 8.3.4 * *Let $f : E \to \mathbf{R}$ be a Lipschitz continuous convex function which attains its minimum on E at certain point x^*. Consider process (cf. (8.3.1))*

$$\phi_{i+1} = \phi_i - \gamma_i f'(x_i)/|f'(x_i)|_*,\ x_i = V'(\phi_i),\ \phi_1 = 0, \tag{8.3.6}$$

and let $\bar{x}_i$ be the best (with the smallest value of f) of the points $x_1, ..., x_i$ and let $\varepsilon_i = f(\bar{x}_i) - \min_E f$. Prove that then

$$\varepsilon_N \le L_{\|\cdot\|}(f)\frac{|x^*|^2 + \mathcal{L}\sum_{i=1}^N \gamma_i^2}{2\sum_{i=1}^N \gamma_i},\ N = 1, 2, ... \tag{8.3.7}$$

where $L_{\|\cdot\|}(f)$ is the Lipschitz constant of f with respect to the norm $\|\cdot\|$. In particular, the method converges, provided that

$$\sum_i \gamma_i = \infty,\ \gamma_i \to 0,\ i \to \infty.$$

Note that the basic result explains what is the origin of the "Subgradient Descent miracle" which motivated our considerations; as we see, this miracle comes not from the very specific algebraic structure of the Euclidean norm, but from certain "robust" analytic property of the norm (the Lipschitz continuity of the derivative of the conjugate norm), and we can fabricate similar miracles for arbitrary norms which share the indicated property. In fact you could use the outlined *Mirror Descent* scheme, with necessary (and more or less straightforward) modifications, in order to extend everything what we know about the usual - "Euclidean" - Subgradient Descent (I mean, the versions for optimization over a domain rather than over the whole space and for optimization over solids under functional constraints) onto the general "non-Euclidean" case, but we skip here these issues.

Lecture 9

Large-scale games and variational inequalities

In this lecture we extend convex minimization methods with dimension-independent complexity on saddle point problems and variational inequalities with monotone operators.

In what follows we will be interested in finding a weak solution to a variational inequality with monotone operator, i.e., in solving a problem as follows:

$$\text{find } z^* \in G \text{ such that } \xi^T(w - z^*) \geq 0 \text{ for all } w \in G,\ \xi \in F(w), \tag{9.0.1}$$

where

$G \subset \mathbf{R}^n$ is closed and bounded convex set;

$w \mapsto F(w)$ is a multi-valued monotone operator defined on G. Recall that it means that for any $w \in G$ $F(w)$ is a nonempty subset of $\mathbf{R}^n$ and

$$(\xi - \xi')^T(w - w') \geq 0 \quad \forall w, w' \in G\ \forall \xi \in F(w), \xi' \in F(w').$$

From now on we assume that F is *bounded*:

$$L(F) = \sup\{|\xi| \mid \xi \in F(w), w \in G\} < \infty;$$

here and further $|\cdot|$ is the standard Euclidean norm on $\mathbf{R}^n$. Thus, compared to our first investigation of variational inequalities in Lecture 6, now we do not require the domain G of the inequality to possess a nonempty interior; as a compensation, we assume that the operator is well-defined on the whole domain, not on the interior of the domain only, and assume the operator to be bounded rather than semibounded.

Our goal is to solve the inequality within prescribed *absolute* accuracy ε, i.e., to find $z \in G$ such that

$$\nu(z) \equiv \sup\{\xi^T(z - w) \mid w \in G, \xi \in F(w)\} \leq \varepsilon.$$

Note that $\nu(z)$ is exactly the inaccuracy measure we dealt with in Lecture 6; the only difference is that there we were looking for a solution of a given *relative* accuracy ε, i.e., such that

$$\nu(z) \leq \varepsilon \sup_G \nu.$$

Recall that we were especially interested in a particular class of variational inequalities with monotone operators, i.e., those coming from saddle point problems. Namely, let $G = P \times Q$ be a direct product of two closed and bounded convex subsets, and let $f(x, y)$ be a Lipschitz continuous function on G which is convex in $x \in P$ and concave in $y \in Q$; in Lecture 5 we have associated with this function the *saddle point problem*

$$\textit{find } z^* = (x^*, y^*) \in G \textit{ s.t. } f(x, y^*) \geq f(x^*, y^*) \geq f(x^*, y), \ (x, y) \in G; \tag{9.0.2}$$

we have seen that the solutions to the saddle point problem are exactly the pairs of optimal solutions to the following two convex programs:

$$\textit{minimize } \bar{f}(x) = \max_{y \in Q} f(x, y) \textit{ over } x \in P,$$

$$\textit{maximize } \underline{f}(x) = \min_{x \in P} f(x, y) \textit{ over } y \in Q,$$

the optimal values of the problems being equal to each other. In the case of games we used a specific accuracy measure

$$\nu_g(x, y) = \bar{f}(x) - \underline{f}(y) = \{\bar{f}(x) - \min_P \bar{f}\} + \{\max_Q \underline{f} - \underline{f}(y)\},$$

the sum of inaccuracies of x and y regarded as approximate solutions to the problems of minimizing $\bar{f}$ and of maximizing $\underline{f}$. Recall that we have associated with the saddle point problem in question the variational inequality with monotone operator

$$F_f(z) = \{(f'_x, -f'_y) \mid f'_x \in \partial_x f(z), \ f'_y \in \partial_y f(z)\};$$

here the sub- and supergradient sets of f are comprised of sub- and supergradients of norms not exceeding the corresponding Lipschitz constants of f, so that the operator F_f is bounded. From Lecture 6 (Proposition 6.1.2) we know that this operator is monotone and that the weak solutions to the variational inequality associated with the operator are exactly the saddle points of f.

My goal is to demonstrate that the methods of the subgradient and the bundle types, i.e., methods for convex optimization with dimension-independent rate of convergence, can be naturally extended onto large-scale variational inequalities. Let me start with the Subgradient Descent method.

9.1 Subrgadient Descent method for variational inequalities

The Subgradient Descent method, as applied to variational inequality (9.0.1), generates search points z_i according to

$$z_{i+1} = \pi_G(z_i - \gamma_i \xi_i / |\xi_i|), \ \xi_i \in F(z_i); \tag{9.1.1}$$

here the starting point z_1 is a point of G, γ_i are positive stepsizes and π_G is the projector onto G:

$$\pi_G(z) = \operatorname{argmin}\{|z - w| \mid w \in G\}.$$

As always, (9.1.1) makes sense only if $\xi_i \neq 0$; if we meet with the situation $\xi_i = 0$, then z_i is an exact solution to the inequality and we can terminate. In what follows I ignore this trivial case.

The indicated scheme is very old; for saddle point problems it was proposed by Arrow, Hurvits and Uzava somewhere in mid-fifties. It is easily seen that trajectory (9.1.1), generally speaking, does not converge to the solution set, independently of how one chooses the stepsizes. Indeed, consider the variational inequality associated with the simplest 2-player game

$$f(x,y) = xy,\ P = Q = [-1,1].$$

We have considered this particular game in Lecture 5 and know that the saddle point here is unique and is at the origin. On the other hand, here

$$F(x,y) = (y,-x)^T$$

is a single-valued mapping with the value at a point (x,y) orthogonal to the vector (x,y); it follows that

$$|(z_i - \gamma_i \xi_i/|\xi_i|)|^2 = |z_i|^2 + \gamma_i^2,$$

so that if there were no projector, the trajectory would go further and further from the saddle point. The presence of projector prevents the trajectory of going too far, but, as it is easily seen, does not enforce its convergence to the origin, i.e., to the unique solution to the variational inequality in question. There were many papers on how to ensure convergence of the method; sometimes people imposed additional restrictions on the operator in question, sometimes modified the algorithm. What we shall see in a while is that the divergence of the algorithm is an illusion: the method converges, but in the so called *ergodic*, or, better to say, in the Cezari sense. Namely, it turns out that one can say something quite reasonable about the *weighted search points*

$$\bar{z}_M^N = \left(\sum_{i=M}^N \gamma_i |\xi_i|^{-1}\right)^{-1} \sum_{i=M}^N \gamma_i |\xi_i|^{-1} z_i.$$

Proposition 9.1.1 . *Let a variational inequality with bounded monotone operator be solved by a Subgradient Descent method with positive stepsizes γ_i. Then for all integer M, N, $M \le N$, one has*

$$\nu(\bar{z}_M^N) \le L(F) \frac{D^2(G) + \sum_{i=M}^N \gamma_i^2}{2\sum_{i=M}^N \gamma_i}, \tag{9.1.2}$$

$D(G)$ being the diameter of G. If the variational inequality comes from a 2-player game with zero sum, then the same estimate is valid for the quantity $\nu_g(\bar{z}_M^N)$.

Proof. Let z be an arbitrary point of G, and let $\xi \in F(z)$. Let us look at the quantities $|z_i - z|^2$:

$$\begin{aligned} |z_{i+1} - z|^2 = |\pi_G(z_i - \gamma_i |\xi_i|^{-1}\xi_i) - z|^2 \le |z_i - \gamma_i |\xi_i|^{-1}\xi_i - z|^2 = \\ = |z_i - z|^2 - 2\gamma_i |\xi_i|^{-1} \xi_i^T (z_i - z) + \gamma_i^2 \le \\ \le |z_i - z|^2 + \gamma_i^2 - 2\gamma_i |\xi_i|^{-1} \xi^T (z_i - z) \end{aligned}$$

(we have sequentially used the basic property of the projection nd the monotonicity of the operator). Thus, we come to

$$\xi^T \left(\sum_{i=M}^N \gamma_i |\xi_i|^{-1} (z_i - z)\right) \le |z_M - z|^2 - |z_{N+1} - z|^2 + \sum_{i=M}^N \gamma_i^2 \le D^2(G) + \sum_{i=M}^N \gamma_i^2.$$

The resulting relation can be rewritten as

$$\xi^T(\bar{z}_M^N - z) \le \frac{D^2(G) + \sum_{i=M}^N \gamma_i^2}{2\sum_{i=M}^N \gamma_i |\xi_i|^{-1}} \le L(F)\frac{D^2(G) + \sum_{i=M}^N \gamma_i^2}{2\sum_{i=M}^N \gamma_i};$$

this inequality is valid for all $z \in G$ and all $\xi \in F(z)$, and taking maximum over z, ξ, we come to (9.1.2).

Now consider the case when the variational inequality in question comes from a 2-player zero sum game with the cost function f. Here, as in the general case, we have

$$|z_{i+1} - z|^2 \le |z_i - z|^2 + \gamma_i^2 - 2\gamma_i |\xi_i|^{-1} \xi_i^T (z_i - z).$$

As we know from Lemma 5.1.1 (Lecture 5), for the monotone operator associated with a convex-concave saddle point problem one has

$$\xi_i^T (z_i - z) \ge f(x_i, y) - f(x, y_i);$$

here $z = (x, y)$, $z_i = (x_i, y_i)$. Now, same as above, we came to the inequality

$$2\sum_{i=M}^N \gamma_i |\xi_i|^{-1} \left(f(x_i, y) - f(x, y_i)\right) \le D^2(G) + \sum_{i=M}^N \gamma_i^2.$$

Due to the Jensen inequality, the left hand side here is greater or equal to

$$\left(2\sum_{i=M}^N \gamma_i |\xi_i|^{-1}\right)\left(f(\bar{x}_M^N, y) - f(x, \bar{y}_M^N)\right),$$

and we come to

$$f(\bar{x}_M^N, y) - f(x, \bar{y}_M^N) \le L(F)\frac{D^2(G) + \sum_{i=M}^N \gamma_i^2}{2\sum_{i=M}^N \gamma_i}.$$

The concluding inequality holds true for all $z = (x, y) \in G = P \times Q$, and taking maximum in $(x, y) \in G$, we come to

$$\nu_g(\bar{z}_M^N) \le L(F)\frac{D^2(G) + \sum_{i=M}^N \gamma_i^2}{2\sum_{i=M}^N \gamma_i},$$

as claimed. ∎

We see that the accuracy estimate for the Subgradient Descent as applied to a variational inequality with a monotone operator or to a saddle point problem, looks exactly similar to the estimate for the case of minimizing a convex Lipschitz continuous function. Therefore we already know what is the optimal rate of convergence and what are the corresponding stepsizes. The latter are

$$\gamma_i = D(G)/\sqrt{i},$$

and the associated accuracy estimate (related to $M = [N/2]$) is

$$\nu(\bar{x}_{[N/2]}^N) \le O(1)L(F)D(G)N^{-1/2}, \ N = 1, 2, ...,$$

$O(1)$ being an absolute constant. Thus, in the large-scale case we have basically the same possibilities to solve variational inequalities with bounded monotone operators (or to find saddle points of Lipschitz continuous convex-concave functions) as those to minimize Lipschitz continuous convex functions; recall that the same situation was with linearly converging methods in a fixed dimension.

9.2 Level method for variational inequalities and games

As applied to variational inequalities with monotone operators or to games, the Subgradient Descent possesses optimal worst-case behaviour in the large-scale case, but the "typical" behaviour of the method, same as in the case of convex minimization problems, is far from being appropriate. And same as in the case of convex minimization, to improve the practical behaviour of the method, it is natural to use the bundle scheme, where we use the whole information collected so far. Let us start with the saddle point problem.

9.2.1 Level method for games

Consider the saddle point problem (9.0.2) with Lipschitz continuous with constant $L(f)$ function $f(x,y)$, and assume that we have already generated search points $z_1,...,z_i \in G$ and know the sub- and supergradients $f'_x(z_j)$, $f'_y(z_j)$, $j=1,...,i$; then we can form the models

$$\bar{f}_i(x) = \max_{1\le j\le i}\{f(z_j) + (f'_x(z_j))^T(x-x_j)\}$$

and

$$\underline{f}_i(y) = \min_{1\le j\le i}\{f(z_j) + (f'_y(z_j))^T(y-y_j)\}$$

of the cost functions

$$\bar{f}(x) = \max_{y\in Q} f(x,y),\ \underline{f}(y) = \min_{x\in P} f(x,y)$$

of the players; from convexity-concavity of f it follows that the function $\bar{f}_i$ underestimates $\bar{f}$, while $\underline{f}_i$ overestimates $\underline{f}$. It follows that the function

$$\nu_i(x,y) = \bar{f}_i(x) - \underline{f}_i(y)$$

underestimates the function

$$\nu_g(x,y) = \bar{f}(x) - \underline{f}(y).$$

Note that the optimal set of the latter function on G is exactly the saddle set of f, and the optimal value of ν_g is 0. Since ν_i underestimates ν_g, the optimal value

$$-\Delta_i \equiv \min_G \nu_i = \min_P \bar{f}_i - \max_Q \underline{f}_i$$

is nonpositive; we shall call Δ_i *i-th gap*. It is clear from our construction that

$$\nu_1(\cdot,\cdot) \le \nu_2(\cdot,\cdot) \le ... \le \nu_g(\cdot,\cdot), \tag{9.2.1}$$

$$\Delta_1 \ge \Delta_2 \ge ... \ge 0. \tag{9.2.2}$$

Note also that $\bar{f}_i(x_i) \ge f(z_i)$, $\underline{f}_i(y_i) \le f(z_i)$, and therefore

$$\nu_i(z_i) \ge 0. \tag{9.2.3}$$

The Level method for saddle point problems is as follows:

i-th step: given $z_1, ..., z_i$, form the models $\bar{f}_i$, $\underline{f}_i$ and ν_i and compute the optimal value $-\Delta_i$ of the model ν_i. Define the next iterate z_{i+1} as

$$z_{i+1} = \pi_{G_i}(z_i),\ G_i = \{z \in G \mid \nu_i(z) \leq -(1-\lambda)\Delta_i\}$$

and loop.
The quantity λ in the updating formula is a once for ever fixed quantity from $(0, 1)$.

Note that an iteration of the method, same as in the minimization case, requires to compute $-\Delta_i$, i.e., to solve a pair of minimization problems with piecewise linear objectives $\bar{f}_i$ (this function should be minimized over P) and $-\underline{f}_i$ (this function should be minimized over Q). Besides this, one should compute the projection of z_i onto the convex set G_i. If P and Q, as it normally is the case, are polytopes, these problems are, respectively, a Linear Programming and a linearly constrained Quadratic Programming programs.

Note that in course of minimizing a piecewise linear function $\bar{f}_i$, one can find the corresponding Lagrange multipliers - nonnegative λ_j, $j = 1, ..., i$, with unit sum such that

$$\min_{x\in P} \bar{f}_i(x) = \min_{x\in P} \sum_{j=1}^{i} \lambda_j \left(f(z_j) + (f'_x(z_j))^T (x - x_j) \right);$$

similarly, when maximizing $\underline{f}_i$ over Q, one can find the corresponding Lagrange multipliers $\mu_j \geq 0$, $j = 1, ..., i$, with unit sum:

$$\max_{y\in Q} \underline{f}_i(y) = \max_{y\in Q} \sum_{j=1}^{i} \mu_j \left(f(z_j) + (f'_y(z_j))^T (y - y_j) \right).$$

We define i-th approximate solution generated by the method as

$$\bar{z}_i = (\bar{x}_i = \sum_{j=1}^{i} \mu_j x_j, \bar{y}_i = \sum_{j=1}^{i} \lambda_j y_j).$$

Thus, we use the Lagrange multipliers related to $\underline{f}_i$ to aggregate the x-components of the search points, while the Lagrange multipliers related to $\bar{f}_i$ are the weights in aggregating the y-components.

The convergence properties of the method are given by the following

Proposition 9.2.1 *Let the Level method be applied to the saddle point problem associated with a convex-concave Lipschitz continuous function $f(x, y) : P \times Q \to \mathbf{R}$. Then for any i one has*

$$\nu_g(\bar{z}_i) \leq \Delta_i, \tag{9.2.4}$$

and the right hand side of the latter inequality is $\leq \varepsilon$ whenever

$$i > c(\lambda) \left(\frac{L(f)D(G)}{\varepsilon} \right)^2, \tag{9.2.5}$$

where $L(f)$ is the Lipschitz constant of f, $D(G)$ is the diameter of G and

$$c(\lambda) = \frac{2}{(1-\lambda)^2\lambda(2-\lambda)}.$$

Proof. Let us start with proving (9.2.4). To this end let us fix $(x, y) \in G$. We have

$$f(x_j, y) \le f(z_j) + (f'_y(z_j))^T (y - y_j),$$

$$f(x, y_j) \ge f(z_j) + (f'_x(z_j))^T (x - x_j),$$

and taking weighted sums of these inequalities and subtracting the resulting inequalities from each other, we come to

$$\sum_{j=1}^{i} \{\lambda_j f(x, y_j) - \mu_j f(x_j, y)\} \ge$$

$$\ge \sum_{j=1}^{i} \lambda_j \{f(z_j) + (f'_x(z_j))^T (x - x_j)\} - \sum_{j=1}^{i} \mu_j \{f(z_j) + (f'_y(z_j))^T (y - y_j)\}. \tag{9.2.6}$$

The left hand side in this inequality, due to convexity-concavity of f, is less than or equal to $f(x, \bar{y}_i) - f(\bar{x}_i, y)$, and we come to the inequality

$$f(x, \bar{y}_i) - f(\bar{x}_i, y) \ge \sum_{j=1}^{i} \lambda_j \{f(z_j) + (f'_x(z_j))^T (x - x_j)\} - \sum_{j=1}^{i} \mu_j \{f(z_j) + (f'_y(z_j))^T (y - y_j)\}.$$

Taking minimum over $(x, y) \in G$ and taking into account the origin of the Lagrange multipliers, we come to

$$-\nu_g(\bar{z}_i) \ge \min_P \bar{f}_i - \max_Q \underline{f}_i \equiv -\Delta_i,$$

as required in (9.2.4).

It remains to demonstrate that Δ_i for sure becomes $\le \varepsilon$ when i satisfies (9.2.5). Indeed, let N be an integer such that $\Delta_N > \varepsilon$; let us prove that then N does not exceed the right hand side of (9.2.5). Same as in the proof presented at the previous lecture, we can partition the set $I = \{1, ..., N\}$ of iteration numbers $\le N$ into groups $I_1, ..., I_k$ as follows: the group I_1 ends with the iteration $i_1 = N$ and is comprised of the iterations i such that

$$\Delta_i \le (1 - \lambda)^{-1} \Delta_{i_1};$$

since, as we know, the gaps do not increase with i, this is a segment of the set I. If this segment does not coincide with I, let i_2 be the last iteration in I which does not belong to I_1, and let I_2 be comprised of all indices $i \le i_2$ such that

$$\Delta_i \le (1 - \lambda)^{-1} \Delta_{i_2}.$$

If $I_1 \cup I_2$ does not coincide with I, let i_3 be the last index in I which does not belong neither to I_1, nor to I_2, and let I_3 be comprised of all indices $i \le i_3$ such that

$$\Delta_i \le (1 - \lambda)^{-1} \Delta_{i_3},$$

and so on. With this construction, I will be partitioned into finitely many segments $I_1, ..., I_k$ (the segments are counted from right to left) in such a way that

$$\Delta_i \le (1 - \lambda)^{-1} \Delta_{i_s}, \ i \in I_s, \tag{9.2.7}$$

and

$$\Delta_{i_{s+1}} > (1-\lambda)^{-1}\Delta_{i_s}.$$

My claim is that the sets G_i associated with iterations i from a group I_s have a point in common, namely, the minimizer w_s of the last, i_s-th, model ν_{i_s} of the group. Indeed, since the models grow with i, we have

$$\nu_i(w_s) \le \nu_{i_s}(w_s) = -\Delta_{i_s} \le -(1-\lambda)\Delta_i$$

(we have used (9.2.7)), so that $w_s \in G_i$, as claimed.

Now we can complete the proof in the same way as it was done for the optimization version of the method. As we know from (9.2.3), $\nu_i(z_i) \ge 0$, so that $z_i \notin G_i$; it follows that $\nu_i(z_{i+1}) = -(1-\lambda)\Delta_i$ and

$$\nu_i(z_i) - \nu_i(z_{i+1}) \ge (1-\lambda)\Delta_i \ge (1-\lambda)\Delta_{i_s}. \tag{9.2.8}$$

Since, as it is immediately seen, ν_i is Lipschitz continuous with constant $L(f)$ with respect to x and to y, it is Lipschitz continuous with constant $\sqrt{2}L(f)$ with respect to z, and we conclude from (9.2.8) that

$$|z_i - z_{i+1}| \ge (1-\lambda)\Delta_{i_s}L^{-1}(f)/\sqrt{2}. \tag{9.2.9}$$

At the same time, since $w_s \in G_i$, $i \in I_s$, we have

$$|z_{i+1} - w_s|^2 = |\pi_{G_i}(z_i) - w_s|^2 \le |z_i - w_s|^2 - |z_i - z_{i+1}|^2 \le |z_i - w_s|^2 - \frac{1}{2L^2(f)}(1-\lambda)^2\Delta_{i_s}^2,$$

and we conclude from this recurrence that the number N_s of indices in I_s satisfies the estimate

$$N_s \le 2D^2(G)L^2(f)(1-\lambda)^{-2}\Delta_{i_s}^{-2}.$$

From (9.2.7) it follows that

$$\Delta_{i_s} > (1-\lambda)^{-(s-1)}\Delta_{i_1} > (1-\lambda)^{-(s-1)}\varepsilon,$$

and we come to

$$N_s \le 2D^2(G)L^2(f)(1-\lambda)^{-2}(1-\lambda)^{2(s-1)}\varepsilon^{-2},$$

so that

$$N = \sum_s N_s \le \frac{2}{(1-\lambda)^2\lambda(2-\lambda)}D^2(G)L^2(f)\varepsilon^{-2},$$

as claimed. ∎

9.2.2 Level method for variational inequalities

As we know, to solve a saddle point problem is the same as to minimize an implicitly defined convex function ν_g. Similarly, to find a weak solution to a variational inequality with monotone operator (9.0.1) is the same as to minimize the implicitly defined convex function

$$\nu(z) = \sup\{\xi^T(z-w) \mid w \in G,\, \xi \in F(w)\}; \tag{9.2.10}$$

indeed, by definition of a weak solution, this function is nonpositive at any weak solution, and by setting $w = z$ in the formula defining ν we see that the function is nonnegative everywhere;

thus, the minimum value of the function is 0, and weak solutions to inequality are exactly the minimizers of ν. We see that we should minimize an implicitly defined convex function with zero optimal value; same as in the case of games, given $w \in G$ and $\xi \in F(w)$, we know an affine function $\xi^T(z-w)$ of z which underestimates ν, and therefore can use the Level scheme, namely, as follows. At i-th step of the method we already have generated the search points $z_1, ..., z_i \in G$ and know values $\xi_j \in F(z_j)$ of the operator at these points; this information allows us to form the *i-th model*

$$\nu_i(z) = \max_{1 \le j \le i} \xi_j^T (z - z_j)$$

of the function ν; in view of (9.2.10) this function underestimates ν, and since the optimal value of ν is 0, the optimal value of the model is nonpositive:

$$-\Delta_i \equiv \min_{z \in G} \nu_i(z) \le 0. \tag{9.2.11}$$

It is clear that the models grow with i:

$$\nu_1(\cdot) \le \nu_2(\cdot) \le ... \le \nu(\cdot), \tag{9.2.12}$$

and, consequently,

$$\Delta_1 \ge \Delta_2 \ge ... \ge 0. \tag{9.2.13}$$

The Level method for variational inequalities is as follows:

i-th step: given $z_1, ..., z_i$, form the model ν_i and compute the optimal value $-\Delta_i$ of the model. Define the next iterate z_{i+1} as

$$z_{i+1} = \pi_{G_i}(z_i), \ G_i = \{z \in G \mid \nu_i(z) \le -(1-\lambda)\Delta_i\}$$

and loop.

The quantity λ in the updating formula is a once for ever fixed quantity from $(0, 1)$.

Note that in course of minimizing a piecewise linear function ν_i, one can find the corresponding Lagrange multipliers - nonnegative λ_j, $j = 1, ..., i$, with unit sum such that

$$\min_{z \in G} \nu_i(z) = \min_{z \in G} \sum_{j=1}^{i} \lambda_j \xi_j^T (z - z_j).$$

We define i-th approximate solution generated by the method as

$$\bar{z}_i = \sum_{j=1}^{i} \lambda_j z_j;$$

note that this is exactly the way of aggregating search points we have used in the cutting plane scheme for variational inequalities.

The convergence properties of the method are given by the following

Proposition 9.2.2 *Let the Level method be applied to variational inequality (9.0.1) associated with a monotone and bounded operator F. Then for any i one has*

$$\nu(\bar{z}_i) \le \Delta_i, \tag{9.2.14}$$

and the right hand side of the latter inequality is $\leq \varepsilon$ *whenever*

$$i > c(\lambda)\left(\frac{L(F)D(G)}{\varepsilon}\right)^2, \tag{9.2.15}$$

where $L(F) = \sup\{|\xi| \mid \xi \in F(z), z \in G\}$, $D(G)$ *is the diameter of* G *and*

$$c(\lambda) = \frac{1}{(1-\lambda)^2\lambda(2-\lambda)}.$$

Proof. Relation (9.2.14) is, up to notation, the statement of Lemma 6.2.1 (Lecture 6). The implication

$$i > c(\lambda)L^2(F)D^2(G)\varepsilon^{-2} \Rightarrow \Delta_i \leq \varepsilon$$

is given by word-by-word repeating the proof of Proposition 9.2.1. ∎

9.3 Exercises: Around Level

In Lecture 8 we became acquainted with the family of bundle methods and a particular member of the family - the Level method. The below exercises present useful additional information on the Level.

9.3.1 "Prox-Level"

Consider a convex optimization problem

$$\textit{minimize} \;\; f(x) \textit{ over } x \in G, \tag{9.3.1}$$

where f is a Lipschitz continuous with constant $L(f)$ convex function and G is a bounded and closed convex subset of $\mathbf{R}^n$. The Level method, in its basic version, solves the problem as follows: after $i-1$ steps of the procedure we already have generated the search points $x_1, ..., x_i \in G$ and know the values $f(x_j)$ and subgradients $f'(x_j)$, $|f'(x_j)| \leq L(f)$, of the objective at the points x_j, $j = 1, ..., i$. At i-th step we

1) form the *i-th model* of the objective

$$f_i(x) = \max_{1\leq j\leq i}\{f(x_j) + (x - x_j)^T f'(x_j)\},$$

and compute its minimal value over G, let it be called f_i^-;

2) compute the best found so far value of the objective

$$f_i^+ = \min_{1\leq j\leq i} f(x_j);$$

3) define i-th *level*

$$l_i = f_i^- + \lambda\Delta_i, \;\; \Delta_i = f_i^+ - f_i^-$$

($\lambda \in (0,1)$ is the parameter of the method)
and, finally,

4) project the current iterate x_i onto the level set

$$Q_i = \{x \in G \mid f_i(x) \leq l_i\}$$

of the i-th model to get the new iterate:

$$x_{i+1} = \pi_{Q_i}(x_i).$$

This scheme can be modified in various directions; e.g., intuitively it is strange that in order to obtain the next iterate x_{i+1} we project onto the level set of the model the previous iterate x_i independently of whether x_i is good or bad, I mean, independently of whether $f(x_i)$ is the best found so far value of the objective f_i^+ or $f(x_i)$ is greater (even significantly greater) than f_i^+. It seems to be more reasonable to move from the best of the iterates - in this case our chances to improve the best found so far value of the objective look better. It turns out that the strategy "to obtain x_{i+1}, project onto Q_i the best of the search points" also is appropriate - it results in basically the same complexity bound as that one for the initial version of the method. To demonstrate this, we need the following preliminary observation:

Exercise 9.3.1 $^{\#+}$ *Let $P_1 \supset P_2 \supset P_3 \supset ...$ be a sequence of closed nonempty convex subsets of $\mathbf{R}^n$, and let $y_0 \in \mathbf{R}^n$. Let*

$$y_i = \pi_{P_i}(y_0)$$

be the projections of y_0 onto P_i. Then

$$|y_0 - y_i|^2 \geq \sum_{j=1}^{i} |y_{j-1} - y_j|^2. \tag{9.3.2}$$

Now we are able to present and justify the modified version of Level. Let

$$\Delta_i = f_i^+ - f_i^-$$

be the same *gaps* as in the basic version. When running the method, we compute these gaps; as it is immediately seen, these gaps never increase, independently on how we update the search points (provided that the latter always are chosen in G). Let us define the "regeneration steps" $j_1, j_2, ...$ as follows: the first step is a regeneration one:

$$j_1 = 1.$$

Now, let $i \geq 1$ and let j be the last regeneration step preceding the step i. It is possible that

$$\Delta_i < (1 - \lambda)\Delta_j;$$

if it is the case, we claim that i also is a regeneration step, otherwise i is not a regeneration step.

Now, the modified version of the method differs from the basic one by replacing rules 3) - 4) with the following ones:

3*) Check whether i is a regeneration step. If it is the case, set i-th level l_i, same as in the basic version of the method, equal to

$$f_i^- + \lambda\Delta_i;$$

if i is not a regeneration step, then set the level l_i equal to

$$\min\{l_{i-1}, f_i^- + \lambda\Delta_i\}.$$

Note that the difference between the new rule for computing the level and the initial one is that now we enforce the level l_i *at a non-regeneration step* to be not greater than the level used at the previous step; if this requirement is satisfied by the "usual" level $f_i^- + \lambda\Delta_i$, we choose as l_i this latter quantity, otherwise we set $l_i = l_{i-1}$.

4*) Define the new iterate x_{i+1} as the projection of the best found so far (i.e., with the smallest vale of the objective) search point $\bar{x}_i$ onto the level set $Q_i = \{x \in G \mid f_i(x) \le l_i\}$:

$$x_{i+1} = \pi_{Q_i}(\bar{x}_i).$$

If there are several candidates on the role of $\bar{x}_i$ (i.e., the smallest value of f on the set $\{x_1, ..., x_i\}$ is attained at more than one point), $\bar{x}_i$ is defined as the *latest* of the candidates.

Exercise 9.3.2 $^{\#+}$ *Let I_k is the sequence of iterations which starts with k-th regeneration step j_k and ends with the step $i_k \equiv j_{k+1} - 1$ (so that the sequence contains exactly one regeneration step, which starts the sequence).*

1) *Prove that the level sets Q_i, $i \in I_k$, are nonempty and decrease:*

$$Q_{j_k} \supset Q_{j_k+1} \supset ... \supset Q_{i_k};$$

in particular, all level sets are nonempty, so that the modified method is well-defined;

2) *Prove that the number N_k of iterations in the group I_k satisfies the estimate*

$$N_k \le D^2(G)L^2(f)(1-\lambda)^{-4}\Delta_{j_k}^{-2},$$

$D(G)$ being the diameter of G;

3) *Derive from 2) that for any positive ε and any N such that*

$$N > c(\lambda)\left(L(f)D(G)/\varepsilon\right)^2, \; c(\lambda) = (1-\lambda)^{-4}(2-\lambda)^{-1}\lambda^{-1},$$

one has

$$f(\bar{x}_N) - \min_G f \le \Delta_N \le \varepsilon,$$

so that the efficiency estimate for the modified method is, up to the value of $c(\lambda)$, the same as for the initial version of the method.

9.3.2 Level for constrained optimization

To the moment we know how to apply the Level method to convex problems without functional constraints, to variational inequalities and games, but do not know how to use the method for a convex problem with functional constraints:

$$\textit{minimize } f(x) \textit{ s.t. } f_i(x) \le 0, \, i = 1, ..., m, \; x \in G.$$

The goal of the below exercises is to introduce to you a version of the method for this latter problem.

By setting $g(x) = \max_i f_i(x)$, we may reduce the situation to the case of a single functional constraint, i.e., to a problem

$$\textit{minimize } f(x) \textit{ s.t. } g(x) \le 0,\ x \in G; \tag{9.3.3}$$

from now on we assume that f and g are Lipschitz continuous with constant L and G is a closed and bounded convex set in $\mathbf{R}^n$ of a diameter D. Of course, we assume that the problem is feasible (and therefore is solvable). Besides this, it is convenient to assume that g is not completely redundant, i.e., that $g(x) > 0$ although somewhere on G.

Assume that we already have generated search points $x_1, ..., x_i \in G$ and know from the oracle the values $f(x_j)$, $g(x_j)$ and subgradients $f'(x_j)$, $g'(x_j)$ (of norms not exceeding L) of the objective and the constraints. This information can be assembled into the piecewise linear *models* of the objective and the constraint:

$$f_i(x) = \max_{1\le j\le i}\{f(x_j) + (x - x_j)^T f'(x_j)\},$$

$$g_j(x) = \max_{1\le j\le i}\{g(x_j) + (x - x_j)^T g'(x_j)\}.$$

Besides this, we can form the *model problem*

$$(P_i) \qquad \textit{minimize } f_i(x) \textit{ s.t. } g_i(x) \le 0,\ x \in G.$$

Exercise 9.3.3 *Prove that (P_i) is solvable, and that the optimal value f_i^* of the problem is a non-increasing function of i which is $\le$ the actual optimal value f^* of problem (9.3.3).*

Let T_i be the set of all pairs $(f(x_j), g(x_j))$, $j = 1, ..., i$, and let

$$C(i) = \operatorname{Conv} T_i + \mathbf{R}^2_+$$

be the set of all pairs (u, v) on the plane $\mathbf{R}^2$ which are minorized by convex combinations of the pairs from T_i.

Exercise 9.3.4 * *Let $(u, v) \in C(i)$. Prove that one can easily (i.e., without oracle calls or complicated computations) find a point x in G where the pair $(f(x), g(x))$ of values of the objective and the constraint is "not worse" than (u, v):*

$$f(x) \le u,\ g(x) \le v.$$

Now we come to the key construction - to *support function* h_i. This is the function of $\alpha \in [0, 1]$ defined as

$$h_i(\alpha) = \min_{1\le j\le i}\{\alpha(f(x_j) - f_i^*) + (1 - \alpha)g(x_j)\}.$$

Exercise 9.3.5 *Prove that h_i is a concave piecewise linear function with Lipschitz constant $2LD$ on the segment $[0, 1]$ and that*

$$h_i(\alpha) = \min\{\alpha(u - f_i^*) + (1 - \alpha)v \mid (u, v) \in C(i)\}.$$

Besides this,

$$h_1(\cdot) \ge h_2(\cdot) \ge ...$$

From now on we set

$$\rho(u,v) = \max\{(u)_+, (v)_+\}.$$

Exercise 9.3.6 * *Prove that*

$$\Delta_i \equiv \max_{0\le\alpha\le 1} h_i(\alpha) = \min_{(u,v)\in C(i)} \rho(u,v).$$

Note that h_i is defined in terms of the information on the problem accumulated in course of the first i steps, so that after these steps we can easily compute the *i-th gap* Δ_i. According to exercise 9.3.6, this gap is nonnegative, and we can point out in the set $C(i)$ (this set also is defined in terms of the information accumulated during the first i steps) a pair (u_i, v_i) such that

$$u_i - f_i^* \le \Delta_i,\ v_i \le \Delta_i.$$

In view of exercise 9.3.4, we can easily associate with the latter pair a point $\bar{x}_i \in G$ such that

$$f(\bar{x}_i) - f_i^* \le \Delta_i,\ g(\bar{x}_i) \le \Delta_i.$$

Since f_i^* is $\le$ the actual optimal value of the problem (exercise 9.3.3), we see that

the information collected at the first i steps allows us to form an approximate solution $\bar{x}_i$ to problem (9.3.3) with absolute residuals $f(\bar{x}_i) - f^*$ and $(g(\bar{x}_i))_+$ in the values of the objective and the constraint not exceeding the i-th gap Δ_i.

Thus, *all we interested in is to enforce the gaps Δ_i to converge to 0 at the highest possible rate.*

The gap is the maximum of the support function over $\alpha \in [0,1]$, so that if we would be able to enforce the gaps to tend to 0, we for sure will be able to enforce the values of h_i at a particular point α to tend to something nonpositive (this values do converge as $i \to \infty$, since the sequence $\{h_i\}$ is monotone nonincreasing sequence of uniformly Lipschitz continuous functions (exercise 9.3.5) with nonnegative maximums over $[0,1]$ (exercise 9.3.6)). On the other hand, to enforce

$$\lim_{i\to\infty} h_i(\alpha) \le 0$$

for *a given* $\alpha \in [0,1]$ seems to be easier than to ensure

$$\lim_{i\to\infty} \max_{[0,1]} h_i \equiv \lim_{i\to\infty} \Delta_i = 0,$$

so it makes sense to start with the easier problem. In fact we know how to solve it:

Exercise 9.3.7 * *Given $\alpha \in [0,1]$, let us use the basic Level method to minimize the function*

$$f^\alpha(x) = \alpha f(x) + (1-\alpha) g(x)$$

over G. Prove that after N steps of the method we for sure will have

$$h_i(\alpha) \le O(1) D L N^{-1/2},$$

where the constant $O(1)$ depends on the parameter λ of the method only.

Now we are in the situation as follows: at i-th step we are given certain piecewise linear concave and Lipschitz continuous with the constant $2LD$ function $h_i(\alpha)$ on the segment $[0,1]$; this function depends on our behaviour at the previous steps. It is for sure known that as i grows, the functions do not increase at every point of the segment. Choosing at certain moment an $\alpha \in [0,1]$, we know how to behave ourselves in future in a way which, after an arbitrary number N of steps (counted starting from the moment when α was chosen), would make the value of $h_\cdot(\alpha)$ less than $\leq O(1)LDN^{-1/2}$; and our goal is to enforce convergence to 0 of the quantities $\Delta_i \equiv \max_{\alpha\in[0,1]} h_i(\alpha)$. Given this description of our goals and abilities, could you guess how to achieve the goal? Please try to do it before reading the concluding exercises. The expected rate of convergence of Δ_i to 0 should be as follows: to ensure $\Delta_i \leq \varepsilon$, it should take no more than

$$N(\varepsilon) = O(1)\left(\frac{LD}{\varepsilon}\right)^2 \ln(\frac{LD}{\varepsilon})$$

steps.

The way I know is as follows. The function h_i possibly is negative somewhere on $[0,1]$; consider the segment

$$\delta_i = \{\alpha \in [0,1] \mid h_i(\alpha) \geq 0\}.$$

Since h_i decrease with i, these segments are embedded into each other: $\delta_{i+1} \subset \delta_i$. Let us fix a $\mu \in (0,1/2)$ and call a point $\alpha \in [0,1]$ *μ-centered* with respect to δ_i, if the point belongs to δ_i and the smaller of two parts into which the point splits the segment δ_i is at least μ times the segment itself. Now consider the following policy. We start with

$$\alpha^0 = 1/2$$

and use the given by exercise 9.3.7 tactics to decrease $h_i(\alpha^0)$. After each i-th step we check whether α^0 is μ-centered with respect to δ_i. Immediately after this condition turns out to be violated, we replace α^0 by the midpoint α^1 of the current segment δ_i and switch to decreasing $h_\cdot(\alpha^1)$, using the same tactics given by Exercise 9.3.7. This stage is terminated immediately after we discover that α^1 is no more μ-centered with respect to the current segment δ_i; when it happens, we set our "reference value of α" to the midpoint α^2 of the current segment δ_i and switch to decreasing $h_\cdot(\alpha^2)$, and so on.

Exercise 9.3.8 * *Prove that the indicated policy for any positive ε ensures that $\Delta_i \leq \varepsilon$ whenever*

$$i \geq O(1)(LD/\varepsilon)^2 \ln(LD/\varepsilon),$$

with $O(1)$ depending on the parameters λ and μ of the method (recall that λ is the parameter used by our "working horse" - the basic Level method).

Now let us summarize our construction. To solve (9.3.3), we sequentially apply the basic Level method to auxiliary problems of the type

$$\textit{minimize } \alpha f(x) + (1-\alpha)g(x),\ x \in G,$$

from time to time varying α. Needless to say that in practical implementation there is no necessity to solve the auxiliary problems from scratch; you may use the bundles for f and g collected so far. The complexity of the method, as we have seen, is $O(1)(LD/\varepsilon)^2 \ln(LD/\varepsilon)$,

which is by logarithmic factor worse than the optimal complexity $O(1)(LD/\varepsilon)^2$ given by the Subgradient Descent. In spite of this theoretical shortcoming, the practical behaviour of the method is significantly better than that one of the Subgradient Descent.

Note also that in principle we could reduce the constrained problem to the saddle point problem for the Lagrange function

$$\mathcal{L}(x, y) = f(x) + yg(x),$$

which could be solved by the saddle point version of Level. This straightforward approach has a severe drawback: we do not know in advance how large is the optimal Lagrange multiplier (and even whether it exists - the problem should not necessarily satisfy the Slater condition). Since all our efficiency estimates depend on the diameters of the domains involved into the problem, this Lagrange approach would result in bad efficiency estimates. What we in fact use is *Fritz John* duality: an optimal solution to a convex problem (9.3.3) is, for some $\alpha \in [0, 1]$, a solution to the problem

$$\textit{minimize} \;\; \alpha f(x) + (1 - \alpha)g(x), \;\; x \in G$$

(why?) The advantage of this latter duality is that it does not require any qualification of constraints and, consequently, is insensitive to the Slater condition.

Lecture 10

Smooth convex minimization problems

To the moment we have more or less complete impression of what is the complexity of solving general nonsmooth convex optimization problems and what are the corresponding optimal methods. In this lecture we start a new topic: optimal methods for smooth convex minimization. Today I shall deal with the simplest case of unconstrained problems with smooth convex objective. Thus, we shall be interested in methods for solving the problem

$$(f) \qquad \textit{minimize } f(x) \textit{ over } \; x \in \mathbf{R}^n,$$

where f is a convex function of the smoothness type $\mathrm{C}^{1,1}$, i.e., a continuously differentiable with Lipschitz continuous gradient:

$$|f'(x) - f'(y)| \leq \mathcal{L}(f)|x - y| \;\; \forall x, y \in \mathbf{R}^n;$$

from now on $|\cdot|$ denotes the standard Euclidean norm. We assume also that the problem is solvable (i.e., $\mathrm{Argmin}\, f \neq \emptyset$) and denote by $R(f)$ the distance from the origin (which will be the starting point for the below methods) to the optimal set of (f). Thus, we have defined certain family $\mathcal{S}_n$ of convex minimization problems; the family is comprised of all programs (f) associated with $\mathrm{C}^{1,1}$-smooth convex functions f on $\mathbf{R}^n$ which are below bounded and attain their minimum. As usual, we provide the family with the first-order oracle $\mathcal{O}$ which, given on input a point $x \in \mathbf{R}^n$, reports the value $f(x)$ and the gradient $f'(x)$ of the objective at the point. The accuracy measure we are interested in is the absolute inaccuracy in terms of the objective:

$$\varepsilon(f, x) = f(x) - \min_{\mathbf{R}^n} f.$$

It turns out that complexity of solving a problem from the above family heavily depends on the Lipschitz constant $\mathcal{L}(f)$ of the objective and the distance $R(f)$ from the starting point to the optimal set; this is why in order to analyse complexity it makes sense to speak not on the whole family $\mathcal{S}_n$, but on its subfamilies

$$\mathcal{S}_n(L, R) = \{f \in \mathcal{S} \mid \; \mathcal{L}(f) \leq L, R(f) \leq R\},$$

where L and R are positive parameters identifying, along with n, the subfamily.

10.1 Traditional methods

Unconstrained minimization of $C^{1,1}$-smooth convex functions is, in a sense, the most traditional area in Nonlinear Programming. There are many methods for these problems: the simplest Gradient Descent method with constant stepsize, Gradient Descent with complete relaxation, numerous Conjugate Gradient routines, etc. Note that I did not mention the Newton method, since the method, in its basic form, requires twice continuously differentiable objectives and second-order information, and we are speaking about the first-order information and objectives which are not necessarily twice continuously differentiable. One might think that since the field is so well-studied, among the traditional methods there for sure are optimal ones; surprisingly, this is not the case, as we shall see in a while.

To get a kind of a reference point, let me start with the simplest method for smooth convex minimization - the *Gradient Descent*. This method, as applied to (f), generates the sequence of search points

$$x_{i+1} = x_i - \gamma_i f'(x_i),\ x_0 = 0. \tag{10.1.1}$$

Here γ_i are positive stepsizes; various versions of the Gradient Descent differ from each other by the rules for choosing these stepsizes. In the simplest version - *Gradient Descent with constant stepsize* - one sets

$$\gamma_i \equiv \gamma. \tag{10.1.2}$$

With this rule for stepsizes the convergence of the method is ensured if

$$0 < \gamma < \frac{2}{\mathcal{L}(f)}. \tag{10.1.3}$$

Proposition 10.1.1 *Under assumption (10.1.3) process (10.1.1) - (10.1.2) converges at the rate $O(1/N)$: for all $N \geq 1$ one has*

$$f(x_N) - \min f \leq \frac{R^2(f)}{\gamma(2 - \gamma\mathcal{L}(f))}\, N^{-1}. \tag{10.1.4}$$

In particular, with the optimal choice of γ, i.e., $\gamma = \mathcal{L}^{-1}(f)$, one has

$$f(x_N) - \min f \leq \frac{\mathcal{L}(f)R^2(f)}{N}. \tag{10.1.5}$$

Proof. Let x^* be the minimizer of f of the norm $R(f)$. We have

$$f(x) + h^T f'(x) \leq f(x+h) \leq f(x) + h^T f'(x) + \frac{\mathcal{L}(f)}{2}|h|^2 \tag{10.1.6}$$

(we have used the convexity of f and the Lipschitz continuity of the gradient of f). From the second inequality it follows that

$$f(x_{i+1}) \leq f(x_i) - \gamma|f'(x_i)|^2 + \frac{\mathcal{L}(f)}{2}\gamma^2|f'(x_i)|^2 =$$

$$= f(x_i) - \omega|f'(x_i)|^2,\ \omega = \gamma(1 - \gamma\mathcal{L}(f)/2) > 0, \tag{10.1.7}$$

whence

$$|f'(x_i)|^2 \leq \omega^{-1}(f(x_i) - f(x_{i+1})); \tag{10.1.8}$$

we see that the method is monotone and that

$$\sum_i |f'(x_i)|^2 \leq \omega^{-1}(f(x_1) - f(x^*)) = \omega^{-1}(f(0) - f(x^*)) \leq \omega^{-1}\frac{\mathcal{L}(f)R^2(f)}{2} \tag{10.1.9}$$

(the concluding inequality follows from the second inequality in (10.1.6), where one should set $x = x^*$, $h = -x^*$).

We have

$$|x_{i+1} - x^*|^2 = |x_i - x^*|^2 - 2\gamma(x_i - x^*)^T f'(x_i) + \gamma^2|f'(x_i)|^2 \leq$$

$$\leq |x_i - x^*|^2 - 2\gamma(f(x_i) - f(x^*)) + \gamma^2|f'(x_i)|^2$$

(we have used convexity of f, i.e., have applied the first inequality in (10.1.6) with $x = x_i$, $h = x^* - x_i$). Taking sum of these inequalities over $i = 1, ..., N$ and taking into account (10.1.9), we come to

$$\sum_{i=1}^{N}(f(x_i) - f(x^*)) \leq (2\gamma)^{-1}R^2(f) + \gamma\omega^{-1}\mathcal{L}(f)R^2(f)/4 = \frac{R^2(f)}{\gamma(2 - \gamma\mathcal{L}(f))}.$$

Since the absolute inaccuracies $\varepsilon_i = f(x_i) - f(x^*)$, as we already know, decrease with i, we obtain from the latter inequality

$$\varepsilon_N \leq \frac{R^2(f)}{\gamma(2 - \gamma\mathcal{L}(f))}\, N^{-1},$$

as required. ∎

In fact our upper bound on the accuracy of the Gradient Descent is sharp: it is easily seen that, given $L > 0$, $R > 0$, a positive integer N and a stepsize $\gamma > 0$, one can point out a quadratic form $f(x)$ on the plane such that $\mathcal{L}(f) = L$, $R(f) = R$ and for the N-th point generated by the Gradient Descent as applied to f one has

$$f(x_N) - \min f \geq O(1)\frac{LR^2}{N}.$$

10.2 Complexity of classes $\mathcal{S}_n(L,R)$

The rate of convergence of the Gradient Descent $O(N^{-1})$ given by Proposition 10.1.1 is dimension-independent and is twice better in order than the known to us optimal in the nonsmooth large scale case rate $O(N^{-1/2})$. There is no surprise in the progress: we have passed from a very wide family of nonsmooth convex problems to the much more narrow family of smooth convex problems. What actually is a surprise is that the rate $O(N^{-1})$ still is not optimal; the optimal dimension-independent rate of convergence in smooth convex minimization turns out to be $O(N^{-2})$.

Theorem 10.2.1 *The complexity of the family $\mathcal{S}_n(L,R)$ satisfies the bounds*

$$O(1)\min\left\{n, \sqrt{\frac{LR^2}{\varepsilon}}\right\} \leq \mathrm{Compl}(\varepsilon) \leq \lfloor\sqrt{\frac{4LR^2}{\varepsilon}}\rfloor \tag{10.2.1}$$

with positive absolute constant $O(1)$.

Note that in the large scale case, namely, when

$$n^2 \geq \frac{LR^2}{\varepsilon},$$

the upper complexity bound only by absolute constant factor differs from the lower one. When dimension is not so large, namely, when

$$n^2 << \frac{LR^2}{\varepsilon},$$

both lower and upper complexity bounds are not sharp; the lower bound does not say anything reasonable, and the upper is valid, but significantly overestimates the complexity. In fact in this "low scale" case the complexity turns out to be $O(1)n\ln(LR^2/\varepsilon)$, i.e., is basically the same as for nonsmooth convex problems. Roughly speaking, in a fixed dimension the advantages of smoothness can be exploited only when the required accuracy is not too high.

Let us prove the theorem. As always, we should point out a method associated with the upper complexity bound and to prove somehow the lower bound.

10.2.1 Upper complexity bound: Nesterov's method

The upper complexity bound is associated with a relatively new construction developed by Yurii Nesterov in 1983. This construction is a very nice example which demonstrates the meaning of the complexity approach; the method with the rate of convergence $O(N^{-2})$ was found mainly because the investigating of complexity enforced to believe that such a method should exist.

The method, as applied to problem (f), generates the sequence of search points x_i, $i = 0, 1, ...$, and the sequence of approximate solutions y_i, $i = 1, 2, ...$, along with auxiliary sequences of vectors q_i, positive reals L_i and reals $t_i \geq 1$, according to the following rules:

Initialization: Set

$$x_0 = y_1 = q_0 = 0;\ t_0 = 1;$$

choose as L_0 an arbitrary positive initial estimate of $\mathcal{L}(f)$.

i-th step, $i \geq 1$: Given x_{i-1}, y_i, q_{i-1}, L_{i-1}, t_{i-1}, act as follows:

1) *set*

$$x_i = y_i - \frac{1}{t_{i-1}}(y_i + q_{i-1}) \tag{10.2.2}$$

and compute $f(x_i)$, $f'(x_i)$;

2) *Testing sequentially the values $l = 2^j L_{i-1}$, $j = 0, 1, ...$, find the first value of l such that*

$$f(x_i - \frac{1}{l}f'(x_i)) \leq f(x_i) - \frac{1}{2l}|f'(x_i)|^2; \tag{10.2.3}$$

set L_i equal to the resulting value of l;

3) *Set*

$$y_{i+1} = x_i - \frac{1}{L_i} f'(x_i),$$

$$q_i = q_{i-1} + \frac{t_{i-1}}{L_i} f'(x_i),$$

define t_i as the larger root of the equation

$$t^2 - t = t_{i-1}^2$$

and loop.

Note that the method is as simple as the Gradient Descent; the only complication is a kind of line search in rule 2). In fact the same line search (similar to what is called the *Armiljo-Goldstein rule*) should be used in the Gradient Descent, if the Lipschitz constant of the gradient of the objective is not given in advance (of course, in practice it is never given in advance).

Theorem 10.2.2 *Let $f \in \mathcal{S}_n$ be minimized by the aforementioned method. Then for any N one has*

$$f(y_{N+1}) - \min f \le \frac{2\widehat{\mathcal{L}}(f)R^2(f)}{(N+1)^2}, \;\; \widehat{\mathcal{L}}(f) = \max\{2\mathcal{L}(f), L_0\}. \tag{10.2.4}$$

Besides this, first N steps of the method require N evaluations of f' and no more than $2N + \log_2(\widehat{\mathcal{L}}(f)/L_0)$ evaluations of f.

Proof. 1^0. Let us start with the following simple statement:

A. *If $l \ge \mathcal{L}(f)$, then*

$$f(x - \frac{1}{l} f'(x)) \le f(x) - \frac{1}{2l}|f'(x)|^2.$$

This is an immediate consequence of the inequality

$$f(x - tf'(x)) \le f(x) - t|f'(x)|^2 + \frac{\mathcal{L}(f)}{2} t^2 |f'(x)|^2,$$

which in turn follows from (10.1.6).

We see that if $L_0 \ge \mathcal{L}(f)$, then all L_i are equal to L_0, since $l = L_0$ for sure passes all tests (10.2.3). Further, if $L_0 < \mathcal{L}(f)$, then the quantities L_i never become $\ge 2\mathcal{L}(f)$. Indeed, in the case in question the starting value L_0 is $\le \mathcal{L}(f)$. When updating L_{i-1} into L_i, we sequentially subject to test (10.2.3) the quantities L_{i-1}, $2L_{i-1}$, $4L_{i-1}$, ..., and choose as L_i the first of these quantities which passes the test. It follows that if we come to $L_i \ge 2\mathcal{L}(f)$, then the quantity $L_i/2 \ge \mathcal{L}(f)$ did not pass one of our tests, which is impossible by **A**. Thus,

$$L_i \le \widehat{\mathcal{L}}(f), \;\; i = 1, 2, ...$$

Now, the total number of evaluations of f' in course of the first N steps clearly equals to N, while the total number of evaluations of f is N plus the total number, K, of tests (10.2.3) performed during these steps. Among K tests there are N successful (when the tested value of l satisfies (10.2.3)), and each of the remaining tests increases value of $L.$ by the factor 2. Since $L.$, as we know, is $\le 2\mathcal{L}(f)$, the number of these remaining tests is $\le \log_2(\widehat{\mathcal{L}}(f)/L_0)$, so that

the total number of evaluations of f in course of the first N steps is $\leq 2N + \log_2(\widehat{\mathcal{L}}(f)/L_0)$, as claimed.

2^0. It remains to prove (10.2.4). To this end let us add to **A.** the following observation

B. *For any $i \geq 1$ and any $y \in \mathbf{R}^n$ one has*

$$f(y) \geq f(y_{i+1}) + (y - x_i)^T f'(x_i) + \frac{1}{2L_i}|f'(x_i)|^2. \tag{10.2.5}$$

Indeed, we have

$$f(y) \geq (y - x_i)^T f'(x_i) + f(x_i) \geq (y - x_i)^T f'(x_i) + \left(f(y_{i+1}) + \frac{1}{2L_i}|f'(x_i)|^2\right)$$

(the concluding inequality follows from the definition of L_i and y_{i+1}).

3^0. Let x^* be the minimizer of f of the norm $R(f)$, and let

$$\varepsilon_i = f(y_i) - f(x^*)$$

be the absolute accuracy of i-th approximate solution generated by the method.

Applying (10.2.5) to $y = x^*$, we come to inequality

$$0 \geq \varepsilon_{i+1} + (x^* - x_i)^T f'(x_i) + \frac{1}{2L_i}|f'(x_i)|^2. \tag{10.2.6}$$

Now, the relation (10.2.2) can be rewritten as

$$x_i = (t_{i-1} - 1)(y_i - x_i) - q_{i-1},$$

and with this substitution (10.2.6) becomes

$$0 \geq \varepsilon_{i+1} + (x^*)^T f'(x_i) + (t_{i-1} - 1)(x_i - y_i)^T f'(x_i) + q_{i-1}^T f'(x_i) + \frac{1}{2L_i}|f'(x_i)|^2. \tag{10.2.7}$$

At the same time, (10.2.5) as applied to $y = y_i$ results in

$$0 \geq -\delta_i \equiv f(y_{i+1}) - f(y_i) + (y_i - x_i)^T f'(x_i) + \frac{1}{2L_i}|f'(x_i)|^2,$$

which implies an upper bound on $(y_i - x_i)^T f'(x_i)$, or, which is the same, a lower bound on the quantity $(x_i - y_i)^T f'(x_i)$, namely,

$$(x_i - y_i)^T f'(x_i) \geq \varepsilon_{i+1} - \varepsilon_i + \frac{1}{2L_i}|f'(x_i)|^2 + \delta_i.$$

Substituting this estimate into (10.2.7), we come to

$$0 \geq t_{i-1}\varepsilon_{i+1} - (t_{i-1} - 1)\varepsilon_i + (x^*)^T f'(x_i) + q_{i-1}^T f'(x_i) + (t_{i-1} - 1)\delta_i + \frac{t_{i-1}}{2L_i}|f'(x_i)|^2. \tag{10.2.8}$$

Multiplying this inequality by t_{i-1}/L_i, we come to

$$0 \geq \frac{t_{i-1}^2 - t_{i-1}}{L_i}\delta_i + \frac{t_{i-1}^2}{L_i}\varepsilon_{i+1} - \frac{t_{i-1}^2 - t_{i-1}}{L_i}\varepsilon_i + (x^*)^T\left(\frac{t_{i-1}}{L_i}f'(x_i)\right) + q_{i-1}^T\frac{t_{i-1}}{L_i}f'(x_i) + \frac{t_{i-1}^2}{2L_i^2}|f'(x_i)|^2. \tag{10.2.9}$$

We have $(t_{i-1}/L_i)f'(x_i) = q_i - q_{i-1}$ and $t_{i-1}^2 - t_{i-1} = t_{i-2}^2$ (here $t_{-1} = 0$); thus, (10.2.9) can be rewritten as

$$0 \geq \frac{t_{i-1}^2 - t_{i-1}}{L_i}\delta_i + \frac{t_{i-1}^2}{L_i}\varepsilon_{i+1} - \frac{t_{i-2}^2}{L_i}\varepsilon_i + (x^*)^T(q_i - q_{i-1}) + q_{i-1}^T(q_i - q_{i-1}) + \frac{1}{2}|q_i - q_{i-1}|^2 =$$

$$= \frac{t_{i-1}^2 - t_{i-1}}{L_i}\delta_i + \frac{t_{i-1}^2}{L_i}\varepsilon_{i+1} - \frac{t_{i-2}^2}{L_i}\varepsilon_i + (x^*)^T(q_i - q_{i-1}) + \frac{1}{2}|q_i|^2 - \frac{1}{2}|q_{i-1}|^2.$$

Since the quantities L_i do not decrease with i and ε_{i+1} is nonnegative, we shall only strengthen the resulting inequality by replacing the coefficient t_{i-1}^2/L_i at ε_{i+1} by t_{i-1}^2/L_{i+1}; thus, we come to the inequality

$$0 \geq \frac{t_{i-1}^2 - t_{i-1}}{L_i}\delta_i + \frac{t_{i-1}^2}{L_{i+1}}\varepsilon_{i+1} - \frac{t_{i-2}^2}{L_i}\varepsilon_i + (x^*)^T(q_i - q_{i-1}) + \frac{1}{2}|q_i|^2 - \frac{1}{2}|q_{i-1}|^2.$$

Taking sum of these inequalities over $i = 1, ..., N$, we come to

$$\sum_{i=1}^{N} \frac{t_{i-1}^2 - t_{i-1}}{L_i}\delta_i + \frac{t_{N-1}^2}{L_{N+1}}\varepsilon_{N+1} \leq -(x^*)^T q_N - \frac{1}{2}|q_N|^2 \leq \frac{1}{2}|x^*|^2 \equiv R^2(f)/2; \tag{10.2.10}$$

thus,

$$\varepsilon_{N+1} \leq \frac{L_{N+1}}{t_{N-1}^2}\left[\frac{R^2(f)}{2} - \sum_{i=1}^{N} \frac{t_{i-1}^2 - t_{i-1}}{L_i}\delta_i\right]. \tag{10.2.11}$$

As we know, $L_{N+1} \leq \widehat{\mathcal{L}}(f)$, and from the recurrence

$$t_i = \frac{1}{2}(1 + \sqrt{1 + 4t_{i-1}^2}),\ t_0 = 1$$

it immediately follows that $t_i \geq (i+1)/2$; therefore (10.2.11) implies (10.2.4). ∎

Theorem 10.2.2 implies the upper complexity bound announced in (10.2.1); to see this, it suffices to apply to a problem from $\mathcal{S}_n(L,R)$ the Nesterov method with $L_0 = L$; in this case, as we know from **A.**, all L_i are equal to L_0, so that we may skip the line search 2); thus, to run the method it will be sufficient to evaluate f and f' at the search points only, and in view of (10.2.4) to solve the problem within absolute accuracy ε it is sufficient to perform the indicated in (10.2.1) number of oracle calls.

Let me add some words about the Nesterov method. As it is, it looks as an analytical trick; I would be happy to present to you a geometrical explanation, but I do not know it. Historically, there were several predecessors of the method with more clear geometry, but these methods had slightly worse efficiency estimates; the progress in these estimates made the methods less and less geometrical, and the result is as you just have seen. By the way, I have said that I do not know geometrical *explanation* of the method, but I can immediately point out geometrical *presentation* of the method. It is easily seen that the vectors q_i can be eliminated from the equations describing the method, and the resulting description is simply

$$x_i = (1 + \alpha_i)y_i - \alpha_i y_{i-1},\ \alpha_i = \frac{t_{i-2} - 1}{t_{i-1}},\ i \geq 1,\ y_0 = y_1 = 0,$$

$$y_{i+1} = x_i - \frac{1}{L_i} f'(x_i)$$

(L_i are given by rule 2)). Thus, i-th search point is a kind of forecast - it belongs to the line passing through the pair of the last approximate solutions y_i, y_{i-1}, and for large i y_i is approximately the middle of the segment $[y_{i-1}, x_i]$; the new approximate solution is obtained from this forecast x_i by the standard Gradient Descent step with Armiljo-Goldstein rule for the steplength.

Let me make one more remark. As we just have seen, the simplest among the traditional methods for smooth convex minimization - the Gradient Descent - is far from being optimal; its worst-case order of convergence can be improved by factor 2. What can be said about optimality of other traditional methods, like the Conjugate Gradients? Surprisingly, the answer is as follows: no one of the traditional methods is proved to be optimal, i.e., with the rate of convergence $O(N^{-2})$. For some versions of the Conjugate Gradient family it is known that their worst-case behaviour is not better, and sometimes is even worse, than that one of the Gradient Descent, and no one of the remaining traditional methods is known to be better than the Gradient Descent.

10.2.2 Lower bound

Surprisingly, the "most difficult" smooth convex minimization problems are the quadratic ones; the lower complexity bound in (10.2.1) comes exactly from investigating the complexity of large scale unconstrained convex quadratic minimization with respect to the *first order* minimization methods (I stress this "first order"; of course, the second order methods, like the Newton one, minimize a quadratic function in one step).

Let

$$\bar{\sigma} = \{\sigma_1 < \sigma_2 < ... < \sigma_n\}$$

be a set of n points from the half-interval $(0, L]$ and

$$\bar{\mu} = \{\mu_1, ..., \mu_n\}$$

be a sequence of nonnegative reals such that

$$\sum_{i=1}^{n} \mu_i = R^2.$$

Let $e_1, ..., e_n$ be the standard orths in $\mathbf{R}^n$, and let

$$A = \mathrm{Diag}\{\sigma_1, ..., \sigma_n\},$$

$$b = \sum_{i=1}^{n} \sqrt{\mu_i}\sigma_i e_i.$$

Consider the quadratic form

$$f(x) \equiv f_{\bar{\sigma},\bar{\mu}}(x) = \frac{1}{2} x^T A x - x^T b;$$

this function clearly is smooth, with the gradient $f'(x) = Ax$ being Lipschitz continuous with constant L; the minimizer of the function is the vector

$$\sum_{i=1}^{n} \sqrt{\mu_i} e_i, \tag{10.2.12}$$

and the norm of this vector is exactly R. Thus, the function f belongs to $\mathcal{S}_n(L,R)$. Now, let $\mathcal{U}$ be the family of all rotations of $\mathbf{R}^n$ which remain the vector b invariant, i.e., the family of all orthogonal $n \times n$ matrices U such that $Ub = b$. Since $\mathcal{S}_n(L,R)$ contains the function f, it contains also the family $\mathcal{F}(\bar{\sigma}, \bar{\mu})$ comprised by all rotations

$$f_U(x) = f(Ux)$$

of the function f associated with $U \in \mathcal{U}$. Note that f_U is the quadratic form

$$\frac{1}{2} x^T A_U x - x^T b, \;\; A_U = U^T A U.$$

Thus, the family $\mathcal{F}(\bar{\sigma}, \bar{\mu})$ is contained in $\mathcal{S}_n(L,R)$, and therefore the complexity of this family underestimates the complexity of $\mathcal{S}_n(L,R)$. What I am going to do is to bound from below the complexity of $\mathcal{F}(\bar{\sigma}, \bar{\mu})$ and then choose the worst "parameters" $\bar{\sigma}$ and $\bar{\mu}$ to get the best possible, within the bounds of this approach, lower bound for the complexity of $\mathcal{S}_n(L,R)$.

Consider a convex quadratic form

$$g(x) = \frac{1}{2} x^T Q x - x^T b;$$

here Q is a positive semidefinite symmetric $n \times n$ matrix. Let us associate with this form the *Krylov subspaces*

$$\mathcal{E}_i(g) = \mathbf{R}b + \mathbf{R}Qb + ... + \mathbf{R}Q^{i-1}b$$

and the quantities

$$\varepsilon_i^*(g) = \min_{x \in \mathcal{E}_i(Q,b)} g(x) - \min_{x \in \mathbf{R}^n} g(x).$$

Note that these quantities clearly remain invariant under rotations

$$(Q,b) \mapsto (U^T Q U, U^T b)$$

associated with $n \times n$ orthogonal matrices U. In particular, the quantities $\varepsilon_i^*(f_U)$ associated with $f_U \in \mathcal{F}(\bar{\sigma}, \bar{\mu})$ in fact do not depend on U:

$$\varepsilon_i^*(f_U) \equiv \varepsilon_i^*(f). \tag{10.2.13}$$

The key to our complexity estimate is given by the following:

Proposition 10.2.1 *Let $\mathcal{M}$ be a method for minimizing quadratic functions from $\mathcal{F}(\bar{\sigma}, \bar{\mu})$ and M be the complexity of the method on the latter family. Then there is a problem f_U in the family such that the result $\bar{x}$ of the method as applied to f_U belongs to the space $\mathcal{E}_{2M+1}(f_U)$; in particular, the inaccuracy of the method on the family in question is at least $\varepsilon_{2M+1}^*(f_U) \equiv \varepsilon_{2M+1}^*(f)$.*

The proposition will be proved in the concluding part of the lecture.

Let us derive from Proposition 10.2.1 the desired lower complexity bound. Let $\varepsilon > 0$ be fixed, let M be the complexity $\mathrm{Compl}(\varepsilon)$ of the family $\mathcal{S}_n(L,R)$ associated with this ε. Then there exists a method which solves in no more than M steps within accuracy ε all problems from $\mathcal{S}_n(L,R)$, and, consequently, all problems from the smaller families $\mathcal{F}(\bar\sigma,\bar\mu)$, for all $\bar\sigma$ and $\bar\mu$. According to Proposition 10.2.1, this latter fact implies that

$$\varepsilon^*_{2M+1}(f) \le \varepsilon \tag{10.2.14}$$

for any $f = f_{\bar\sigma,\bar\mu}$. Now, the Krylov space $\mathcal{E}_N(f)$ is, by definition, comprised of linear combinations of the vectors $b, Ab, ..., A^{N-1}b$, or, which is the same, of the vectors of the form $q(A)b$, where q runs over the space $\mathcal{P}_N$ of polynomials of degrees $\le N$. Now, the coordinates of the vector b are $\sigma_i\sqrt{\mu_i}$, so that the coordinates of $q(A)b$ are $q(\sigma_i)\sigma_i\sqrt{\mu_i}$. Therefore

$$f(q(A)b) = \sum_{i=1}^n \{\frac{1}{2}\sigma_i^3 q^2(\sigma_i)\mu_i - \sigma_i q(\sigma_i)\mu_i\}.$$

As we know, the minimizer of f over $\mathbf{R}^n$ is the vector with the coordinates $\sqrt{\mu_i}$, and the minimal value of f, as it is immediately seen, is

$$-\frac{1}{2}\sum_{i=1}^n \sigma_i\mu_i.$$

We come to

$$2\varepsilon^*_N(f) = \min_{q\in\mathcal{P}_{N-1}} \sum_{i=1}^n \{\sigma_i^3 q^2(\sigma_i)\mu_i - 2\sigma_i^2 q(\sigma_i)\mu_i\} + \sum_{i=1}^n \sigma_i\mu_i =$$

$$= \min_{q\in\mathcal{P}_{N-1}} \sum_{i=1}^n \sigma_i(1-\sigma_i q(\sigma_i))^2\mu_i.$$

Substituting $N = 2M+1$ and taking in account (10.2.14), we come to

$$2\varepsilon \ge \min_{q\in\mathcal{P}_{2M}} \sum_{i=1}^n \sigma_i(1-\sigma_i q(\sigma_i))^2\mu_i. \tag{10.2.15}$$

This inequality is valid for all sets $\bar\sigma = \{\sigma_i\}_{i=1}^n$ comprised of n different points from $(0,L]$ and all sequences $\bar\mu = \{\mu_i\}_{i=1}^n$ of nonnegative reals with the sum equal to R^2; it follows that

$$2\varepsilon \ge \sup_{\bar\sigma}\max_{\bar\mu}\min_{p\in\mathcal{P}_{2M}} \sum_{i=1}^n \sigma_i(1-\sigma_i q(\sigma_i))^2\mu_i; \tag{10.2.16}$$

due to the von Neumann Lemma, one can interchange here the minimization in q and the maximization in $\bar\mu$, which immediately results in

$$2\varepsilon \ge R^2 \sup_{\bar\sigma}\min_{q\in\mathcal{P}_{2M}}\max_{\sigma\in\bar\sigma}\sigma(1-\sigma q(\sigma))^2. \tag{10.2.17}$$

Now, the quantity

$$\nu^2(\bar\sigma) = \min_{q\in\mathcal{P}_{2M}}\max_{\sigma\in\bar\sigma}\sigma(1-\sigma q(\sigma))^2$$

is nothing but the squared best quality of the approximation, in the uniform (Tschebyshev's) norm on the set $\bar{\sigma}$, of the function $\sqrt{\sigma}$ by a linear combination of the $2M$ functions

$$\phi_i(\sigma) = \sigma^{i+1/2},\ i = 0, ..., 2M-1$$

From Exercise 1.3.16 we know that

the quality of the best uniform on the segment $[0, L]$ approximation of the function $\sqrt{\sigma}$ by a linear combination of $2M$ functions $\phi_i(\sigma)$ is the same as the quality of the best uniform approximation of the function by such a linear combination on certain $(2M+1)$-point subset of $[0, L]$.

We see that if $2M + 1 \le n$, then

$$2\varepsilon \ge R^2 \sup_{\bar{\sigma}} \nu^2(\bar{\sigma}) = R^2 \min_{q \in \mathcal{P}_{2M}} \max_{\sigma \in [0,L]} \sigma(1 - \sigma q(\sigma))^2.$$

The concluding quantity in this chain can be computed explicitly, see Exercise 12.6.5; it is equal to $(4M+1)^{-2}LR^2$. Thus, we have established the implication

$$2M + 1 \le n \Rightarrow (4M+1)^{-2}LR^2 \le 2\varepsilon,$$

which immediately results in

$$M \ge \min\left\{\frac{n-1}{2}, \frac{1}{4}\sqrt{\frac{LR^2}{2\varepsilon}} - \frac{1}{4}\right\};$$

this is nothing but the lower bound announced in (10.2.1).

10.2.3 Appendix: proof of Proposition 10.2.1

Let $\mathcal{M}$ be a method for solving problems from the family $\mathcal{F}(\bar{\sigma}, \bar{\mu})$, and let M be the complexity of the method at the family. Without loss of generality one may assume that $\mathcal{M}$ always performs exactly $N = M + 1$ steps, and the result of the method as applied to any problem from the family always is the last, N-th point of the trajectory. Note that the statement in question is evident if $\varepsilon^*_{2M+1}(f) = 0$; thus, to prove the statement, it suffices to consider the case when $\varepsilon^*_{2M+1}(f) > 0$. In this latter case the optimal solution $A^{-1}b$ to the problem does not belong to the Krylov space $\mathcal{E}_{2M+1}(f)$ of the problem. On the other hand, the Krylov subspaces evidently form an increasing sequence,

$$\mathcal{E}_1(f) \subset \mathcal{E}_2(f) \subset ...,$$

and it is immediately seen that if one of these inclusions is equality, then all subsequent inclusions also are equalities. Thus, the sequence of Krylov spaces is as follows: the subspaces in the sequence are strictly enclosed each into the next one, up to some moment, when the sequence stabilizes. The subspace at which the sequence stabilizes clearly is invariant for A, and since it contains b and is invariant for $A = A^T$, it contains $A^{-1}b$ (why?). We know that in the case in question $\mathcal{E}_{2M+1}(f)$ does not contain $A^{-1}b$, and therefore all inclusions

$$\mathcal{E}_1(f) \subset \mathcal{E}_2(f) \subset ... \subset \mathcal{E}_{2M+1}(f)$$

are strict. Since it is the case with f, it is also the case with every f_U (since there is a rotation of the space which maps $\mathcal{E}_i(f)$ onto $\mathcal{E}_i(f_U)$ simultaneously for all i). With this preliminary observation in mind we can immediately derive the statement of the Proposition from the following

Lemma 10.2.1 *For any $k \leq N$ there is a problem f_{U_k} in the family such that the first k points of the trajectory of $\mathcal{M}$ as applied to the problem belong to the Krylov space $\mathcal{E}_{2k-1}(f_{U_k})$.*

Proof is given by induction on k.

base $k = 0$ is trivial.

step $k \mapsto k+1$, $k < N$: let U_k be given by the inductive hypothesis, and let $x_1, ..., x_{k+1}$ be the first $k+1$ points of the trajectory of $\mathcal{M}$ on f_{U_k}. By the inductive hypothesis, we know that $x_1, ..., x_k$ belong to $\mathcal{E}_{2k-1}(f_{U_k})$. Besides this, the inclusions

$$\mathcal{E}_{2k-1}(f_{U_k}) \subset \mathcal{E}_{2k}(f_{U_k}) \subset \mathcal{E}_{2k+1}(f_{U_k})$$

are strict (this is given by our preliminary observation; note that $k < N$). In particular, there exists a rotation V of $\mathbf{R}^n$ which is identical on $\mathcal{E}_{2k}(f_{U_k})$ and is such that

$$x_{k+1} \in V^T \mathcal{E}_{2k+1}(f_{U_k}).$$

Let us set

$$U_{k+1} = U_k V$$

and verify that this is the rotation required by the statement of the Lemma for the new value of k. First of all, this rotation remains b invariant, since $b \in \mathcal{E}_1(f_{U_k}) \subset \mathcal{E}_{2k}(f_{U_k})$ and V is identical on the latter subspace and therefore maps b onto itself; since U_k also remains b invariant, so does U_{k+1}. Thus, $f_{U_{k+1}} \in \mathcal{F}(\bar{\sigma}, \bar{\mu})$, as required.

It remains to verify that the first $k+1$ points of the trajectory of $\mathcal{M}$ of $f_{U_{k+1}}$ belong to $\mathcal{E}_{2(k+1)-1}(f_{U_{k+1}}) = \mathcal{E}_{2k+1}(f_{U_{k+1}})$. Since $U_{k+1} = U_k V$, it is immediately seen that

$$\mathcal{E}_i(f_{U_{k+1}}) = V^T \mathcal{E}_i(f_{U_k}),\ i = 1, 2, ...;$$

since V (and therefore $V^T = V^{-1}$ is identical on $\mathcal{E}_{2k}(f_{U_k})$, it follows that

$$\mathcal{E}_i(f_{U_{k+1}}) = \mathcal{E}_i(f_{U_k}),\ i = 1, ..., 2k. \tag{10.2.18}$$

Besides this, from $Vy = y$, $y \in \mathcal{E}_{2k}(f_{U_k})$, it follows immediately that the functions f_{U_k} and $f_{U_{k+1}}$, same as their gradients, coincide with each other on the subspace $\mathcal{E}_{2k-1}(f_{U_k})$. This patter subspace contains the first k search points generated by $\mathcal{M}$ as applied to f_{U_k}, this is our inductive hypothesis; thus, the functions f_{U_k} and $f_{U_{k+1}}$ are "informationally indistinguishable" along the first k steps of the trajectory of the method as applied to one of the functions, and, consequently, $x_1, ..., x_{k+1}$, which, by construction, are the first $k+1$ points of the trajectory of $\mathcal{M}$ on f_{U_k}, are as well the first $k+1$ search points of the trajectory of $\mathcal{M}$ on $f_{U_{k+1}}$. By construction, the first k of these points belong to $\mathcal{E}_{2k-1}(f_{U_k}) = \mathcal{E}_{2k-1}(f_{U_{k+1}})$ (see (10.2.18)), while x_{k+1} belongs to $V^T \mathcal{E}_{2k+1}(f_{U_k}) = \mathcal{E}_{2k+1}(f_{U_{k+1}})$, so that $x_1, ..., x_{k+1}$ do belong to $\mathcal{E}_{2k+1}(f_{U_{k+1}})$. ∎

Lecture 11

Constrained smooth and strongly convex problems

In this lecture we continue investigation of optimal methods for smooth large scale convex problems; what we are interested in now are constrained problems.

In the nonsmooth case, as we know, it is easy to pass from methods for problems without functional constraints to those with constraints. It is not the case with smooth convex programs; we do not know a completely satisfactory way to extend the Nesterov method onto problems with smooth functional constraints. This is why it I shall restrict myself with something intermediate, namely, with *composite* minimization.

11.1 Composite problem

Consider a problem as follows:

$$\textit{minimize } f(x) = F(\phi(x)) \textit{ s.t. } x \in G; \tag{11.1.1}$$

Here G is a closed convex subset in $\mathbf{R}^n$ which contains a given starting point $\bar{x}$. Now, the inner function

$$\phi(x) = (\phi_1(x), ..., \phi_k(x)) : \mathbf{R}^n \to \mathbf{R}^k$$

is a k-dimensional vector-function with convex continuously differentiable components possessing Lipschitz continuous gradients:

$$|\phi_i'(x) - \phi_i'(x')| \leq L_i |x - x'| \ \ \forall x, x',$$

$i = 1, ..., k$. The outer function

$$F : \mathbf{R}^k \to \mathbf{R}$$

is assumed to be Lipschitz continuous, convex and monotone (the latter means that $F(u) \geq F(u')$ whenever $u \geq u'$). Under these assumptions f is a continuous convex function on $\mathbf{R}^n$, so that (11.1.1) is a convex programming program; we assume that the program is solvable, and denote by $R(f)$ the distance from the $\bar{x}$ to the optimal set of the problem.

When solving (11.1.1), we assume that *the outer function F is given in advance*, and the inner function is observable via the standard first-order oracle, i.e., we may ask the oracle to compute,

at any desired $x \in \mathbf{R}^n$, the value $\phi(x)$ and the collection of the gradients $\phi_i'(x)$, $i = 1, ..., k$, of the function.

Note that many convex programming problems can be written down in the generic form (11.1.1). Let me list the pair of the most important examples.

Example 1. Minimization without functional constraints. Here $k = 1$ and $F(u) \equiv u$, so that (11.1.1) becomes a problem of minimizing a $C^{1,1}$-smooth convex objective over a given convex set. It is a new problem for us, since to the moment we know only how to minimize a smooth convex function over the whole space.

Example 2. Minimax problems. Let

$$F(u) = \max\{u_1, ..., u_k\};$$

with this outer function, (11.1.1) becomes the minimax problem - the problem of minimizing the maximum of finitely many smooth convex functions over a convex set. The minimax problems arise in many applications. In particular, in order to solve a system of smooth convex inequalities $\phi_i(x) \le 0$, $i = 1, ..., k$, one can minimize the residual function $f(x) = \max_i \phi_i(x)$.

Let me stress the following. The objective f in a composite problem (11.1.1) should not necessarily be smooth (look what happens in the minimax case). Nevertheless, I speak about the composite problems in the part of the course which deals with smooth minimization, and we shall see that the complexity of composite problems with smooth inner functions is the same as the complexity of unconstrained minimization of a smooth convex function. What is the reason for this phenomenon? The answer is as follows: although f may be non-smooth, we know in advance the source of nonsmoothness, i.e, we know that f is combined from smooth functions and how it is combined. In other words, when solving a nonsmooth optimization problem of a general type, all our knowledge comes from the answers of a local oracle. In contrast to this, when solving a composite problem, we from the very beginning possess certain global knowledge on the structure of the objective, namely, we know the outer function F, and only part of the required information, that one on the inner function, comes from a local oracle.

11.2 Gradient mapping

Let me start with certain key construction - the construction of a gradient mapping associated with a composite problem. Let $A > 0$ and $x \in \mathbf{R}^n$. Consider the function

$$f_{A,x}(y) = F(\phi(x) + \phi'(x)[y - x]) + \frac{A}{2}|y - x|^2.$$

This is a strictly convex continuous function of y which tends to ∞ as $|y| \to \infty$, and therefore it attains its minimum on G at a unique point

$$T(A, x) = \operatorname*{argmin}_{y \in G} f_{A,x}(y).$$

We shall say that A is *appropriate* for x, if

$$f(T(A, x)) \le f_{A,x}(T(A, x)), \tag{11.2.1}$$

and if it is the case, we call the vector

$$p(A,x) = A(x - T(A,x))$$

the *A-gradient of f at x*.

Let me present some examples which motivate the introduced notion.

Gradient mapping for a smooth convex function on the whole space. Let $k = 1$, $F(u) \equiv u$, $G = \mathbf{R}^n$, so that our composite problem is simply an unconstrained problem with $\mathrm{C}^{1,1}$-smooth convex objective. In this case

$$f_{A,x}(y) = f(x) + (f'(x))^T(y-x) + \frac{A}{2}|y-x|^2,$$

and, consequently,

$$T(A,x) = x - \frac{1}{A}f'(x),\ f_{A,x}(T(A,x)) = f(x) - \frac{1}{2A}|f'(x)|^2;$$

as we remember from the previous lecture, this latter quantity is $\le f(T(A,x))$ whenever $A \ge \mathcal{L}(f)$ (item **A.** of the proof of Theorem 10.2.2), so that all $A \ge \mathcal{L}(f)$ for sure are appropriate for x, and for these A the vector

$$p(A,x) = A(x - T(A,x)) = f'(x)$$

is the usual gradient of f.

Gradient mapping for a smooth convex function on a convex set. Let, same as above, $k = 1$ and $F(u) \equiv u$, but now let G be a proper convex subset of $\mathbf{R}^n$. We still have

$$f_{A,x}(y) = f(x) + (f'(x))^T(y-x) + \frac{A}{2}|y-x|^2,$$

or, which is the same,

$$f_{A,x}(y) = [f(x) - \frac{1}{2A}|f'(x)|^2] + \frac{A}{2}|y-\bar{x}|^2,\ \bar{x} = x - \frac{1}{A}f'(x).$$

We see that the minimizer of $f_{A,x}$ over G is the projection onto G of the global minimizer $\bar{x}$ of the function:

$$T(A,x) = \pi_G(x - \frac{1}{A}f'(x)).$$

As we shall see in a while, here again all $A \ge \mathcal{L}(f)$ are appropriate for any x. Note that for a "simple" G (with an easily computed projection mapping) there is no difficulty in computing an A-gradient.

Gradient mapping for the minimax composite function. Let

$$F(u) = \max\{u_1, ..., u_k\}.$$

Here

$$f_{A,x}(y) = \max_{1\le i\le k}\{\phi_i(x) + (\phi_i'(x))^T(y-x)\} + \frac{A}{2}|y-x|^2,$$

and to compute $T(A,x)$ means to solve the auxiliary optimization problem

$$T(A,x) = \operatorname*{argmin}_G f_{A,x}(y);$$

we already met auxiliary problems of this type in connection with the bundle scheme.

The main properties of the gradient mapping are given by the following

Proposition 11.2.1 (i) *Let*

$$A(f) = \sup_{u\in\mathbf{R}^k, t>0} \frac{F(u+tL(f))-F(u)}{t}, \;\; L(f) = (L_1,...,L_k).$$

Then every $A \geq A(f)$ is appropriate for any $x \in \mathbf{R}^n$;

(ii) *Let $x \in \mathbf{R}^n$, let A be appropriate for x and let $p(A,x)$ be the A-gradient of f at x. Then*

$$f(y) \geq f(T(A,x)) + p^T(A,x)(y-x) + \frac{1}{2A}|p(A,x)|^2 \tag{11.2.2}$$

for all $y \in G$.

Proof. (i) is immediate: as we remember from the previous lecture,

$$\phi_i(y) \leq \phi_i(x) + (\phi_i'(x))^T(y-x) + \frac{L_i}{2}|y-x|^2,$$

and since F is monotone, we have

$$F(\phi(y)) \leq F(\phi(x) + \phi'(x)[y-x] + \frac{1}{2}|y-x|^2 L(f)) \leq$$

$$\leq F(\phi(x) + \phi'(x)[y-x]) + \frac{A(f)}{2}|y-x|^2 \equiv f_{A(f),x}(y)$$

(we have used the definition of $A(f)$). We see that if $A \geq A(f)$, then

$$f(y) \leq f_{A,x}(y)$$

for all y and, in particular, for $y = T(A,x)$, so that A is appropriate for x.

(ii): since $f_{A,x}(\cdot)$ attains its minimum on G at the point $z \equiv T(A,x)$, there is a subgradient p of the function at z such that

$$p^T(y-z) \geq 0, \;\; y \in G. \tag{11.2.3}$$

Now, since p is a subgradient of $f_{A,x}$ at z, we have

$$p = (\phi'(x))^T\zeta + A(z-x) = (\phi'(x))^T\zeta - p(A,x) \tag{11.2.4}$$

for some $\zeta \in \partial F(u)$, $u = \phi(x) + \phi'(x)[z-x]$. Since F is convex and monotone, one has

$$f(y) \geq F(\phi(x) + \phi'(x)[y-x]) \geq F(u) + \zeta^T(\phi(x) + \phi'(x)[y-x] - u) =$$

$$= F(u) + \zeta^T\phi'(x)[y-z] = F(u) + (y-z)^T(p + p(A,x))$$

(the concluding inequality follows from (11.2.4)). Now, if $y \in G$, then, according to (11.2.3), $(y-z)^T p \geq 0$, and we come to

$$y \in G \Rightarrow f(y) \geq F(u) + (y-z)^T p(A,x) = F(\phi(x) + \phi'(x)[z-x]) + (y-z)^T p(A,x) =$$

$$= f_{A,x}(z) - \frac{A}{2}|z-x|^2 + (y-x)^T p(A,x) + (x-z)^T p(A,x) =$$

[note that $z - x = \frac{1}{A} p(A,x)$ by definition of $p(A,x)$]

$$= f_{A,x}(z) + (y-x)^T p(A,x) - \frac{1}{2A}|p(A,x)|^2 + \frac{1}{A}|p(A,x)|^2 =$$

$$= f_{A,x}(z) + (y-x)^T p(A,x) + \frac{1}{2A}|p(A,x)|^2 \geq f(T(A,x)) + (y-x)^T p(A,x) + \frac{1}{2A}|p(A,x)|^2$$

(the concluding inequality follows from the fact that A is appropriate for x), as claimed. ∎

11.3 Nesterov's method for composite problems

Now let me describe the Nesterov method for composite problems. It is a straightforward generalization of the basic version of the method; the only difference is that we replace the usual gradients by A-gradients.

The method, as applied to problem (11.1.1), generates the sequence of search points x_i, $i = 0, 1, ...$, and the sequence of approximate solutions y_i, $i = 1, 2, ...$, along with auxiliary sequences of vectors q_i, positive reals A_i and reals $t_i \geq 1$, according to the following rules:

Initialization: Set

$$x_0 = y_1 = \bar{x};\ q_0 = 0;\ t_0 = 1;$$

choose as A_0 an arbitrary positive initial estimate of $A(f)$.

i-th step, $i \geq 1$: given x_{i-1}, y_i, q_{i-1}, A_{i-1}, t_{i-1}, act as follows:

1) *set*

$$x_i = y_i - \frac{1}{t_{i-1}}(y_i - \bar{x} + q_{i-1}) \tag{11.3.1}$$

and compute $\phi(x_i)$, $\phi'(x_i)$;

2) *Testing sequentially the values $A = 2^j A_{i-1}$, $j = 0, 1, ...$, find the first value of A which is appropriate for $x = x_i$, i.e., for A to be tested compute $T(A, x_i)$ and check whether*

$$f(T(A,x_i)) \leq f_{A,x_i}(T(A,x_i)); \tag{11.3.2}$$

set A_i equal to the resulting value of A;

3) *Set*

$$y_{i+1} = x_i - \frac{1}{A_i} p(A_i, x_i),$$

$$q_i = q_{i-1} + \frac{t_{i-1}}{A_i} p(A_i, x_i),$$

define t_i as the larger root of the equation

$$t^2 - t = t_{i-1}^2$$

and loop.

Note that the method requires solving auxiliary problems of computing $T(A,x)$, i.e., problems of the type

$$\textit{minimize } F(\phi(x)+\phi'(x)(y-x))+\frac{A}{2}|y-x|^2 \textit{ s.t. } y\in G;$$

sometimes these problems are very simple (e.g., when (11.1.1) is the problem of minimizing a $C^{1,1}$-smooth convex function over a simple set), but this is not always the case. If (11.1.1) is a minimax problem, then the indicated auxiliary problems are of the same structure as those arising in the bundle methods, with the only simplification that now the "bundle" - the # of linear components in the piecewise-linear term of the auxiliary objective - is of a once for ever fixed cardinality k; if k is a small integer and G is a simple set, the auxiliary problems still are not too difficult. In more general cases they may become too difficult, and this is the main practical obstacle for implementation of the method.

The convergence properties of the method are given by the following

Theorem 11.3.1 *Let problem (11.1.1) be solved by the aforementioned method. Then for any N one has*

$$f(y_{N+1})-\min_G f\le\frac{2\widehat{\mathcal{A}}(f)R^2(f)}{(N+1)^2},\ \ \widehat{\mathcal{A}}(f)=\max\{2\mathcal{A}(f),A_0\}. \tag{11.3.3}$$

Besides this, first N steps of the method require N evaluations of ϕ' and no more than $M=2N+\log_2(\widehat{\mathcal{A}}(f)/A_0)$ evaluations of ϕ (and solving no more than M auxiliary problems of computing $T(A,x)$).

Proof. The proof is given by word-by-word reproducing the proof of Theorem 10.2.2, with the statements (i) and (ii) of Proposition 11.2.1 playing the role of the statements **A**, **B** of the original proof. Of course, we should replace L_i by A_i and $f'(x_i)$ by $p(A_i,x_i)$. To make the presentation completely similar to that one given in the previous lecture, assume without loss of generality that the starting point $\bar{x}$ of the method is the origin.

1^0. In view of Proposition 11.2.1.(i), from $A_0\ge\mathcal{A}(f)$ it follows that all A_i are equal to A_0, since A_0 for sure passes all tests (11.3.2). Further, if $A_0<\mathcal{A}(f)$, then the quantities A_i never become $\ge 2\mathcal{A}(f)$. Indeed, in the case in question the starting value A_0 is $\le\mathcal{A}(f)$. When updating A_{i-1} into A_i, we sequentially subject to the test (11.3.2) the quantities A_{i-1}, $2A_{i-1}$, $4A_{i-1},...$ and choose as A_i the first of these quantities which passes the test. It follows that if we come to $A_i\ge 2\mathcal{A}(f)$, then the quantity $A_i/2\ge\mathcal{A}(f)$ did not pass one of our tests, which is impossible. Thus,

$$A_i\le\widehat{\mathcal{A}}(f),\ \ i=1,2,...$$

Now, the total number of evaluations of ϕ' in course of the first N steps clearly equals to N, while the total number of evaluations of ϕ is N plus the total number, K, of tests (11.3.2) performed during these steps. Among K tests there are N successful (when the tested value of A satisfies (11.3.2)), and each of the remaining tests increases value of $A_.$ by the factor 2. Since $A_.$, as we know, is $\le 2\mathcal{A}(f)$, the number of these remaining tests is $\le\log_2(\widehat{\mathcal{A}}(f)/A_0)$, so that the total number of evaluations of ϕ in course of the first N steps is $\le 2N+\log_2(\widehat{\mathcal{A}}(f)/A_0)$, as claimed.

2^0. It remains to prove (11.3.3). Let x^* be the minimizer of f on G of the norm $R(f)$, and let

$$\varepsilon_i = f(y_i) - f(x^*)$$

be the absolute accuracy of i-th approximate solution generated by the method.

Applying (11.2.2) to $y = x^*$, we come to inequality

$$0 \geq \varepsilon_{i+1} + (x^* - x_i)^T p(A_i, x_i) + \frac{1}{2A_i}|p(A_i, x_i)|^2. \tag{11.3.4}$$

Now, relation (11.3.1) can be rewritten as

$$x_i = (t_{i-1} - 1)(y_i - x_i) - q_{i-1},$$

(recall that $\bar{x} = 0$), and with this substitution (11.3.4) becomes

$$0 \geq \varepsilon_{i+1} + (x^*)^T p(A_i, x_i) + (t_{i-1} - 1)(x_i - y_i)^T p(A_i, x_i) + q_{i-1}^T p(A_i, x_i) + \frac{1}{2A_i}|p(A_i, x_i)|^2. \tag{11.3.5}$$

At the same time, (11.2.2) as applied to $y = y_i$ results in

$$0 \geq f(y_{i+1}) - f(y_i) + (y_i - x_i)^T p(A_i, x_i) + \frac{1}{2A_i}|p(A_i, x_i)|^2,$$

which implies an upper bound on $(y_i - x_i)^T p(A_i, x_i)$, or, which is the same, a lower bound on the quantity $(x_i - y_i)^T p(A_i, x_i)$, namely,

$$(x_i - y_i)^T p(A_i, x_i) \geq \varepsilon_{i+1} - \varepsilon_i + \frac{1}{2A_i}|p(A_i, x_i)|^2.$$

Substituting this estimate into (11.3.4), we come to

$$0 \geq t_{i-1}\varepsilon_{i+1} - (t_{i-1} - 1)\varepsilon_i + (x^*)^T p(A_i, x_i) + q_{i-1}^T p(A_i, x_i) + t_{i-1}\frac{1}{2A_i}|p(A_i, x_i)|^2. \tag{11.3.6}$$

Multiplying this inequality by t_{i-1}/A_i, we come to

$$0 \geq \frac{t_{i-1}^2}{A_i}\varepsilon_{i+1} - \frac{t_{i-1}^2 - t_{i-1}}{A_i}\varepsilon_i + (x^*)^T(t_{i-1}/A_i)p(A_i, x_i) + q_{i-1}^T\frac{t_{i-1}}{A_i}p(A_i, x_i) + \frac{t_{i-1}^2}{2A_i^2}|p(A_i, x_i)|^2. \tag{11.3.7}$$

We have $(t_{i-1}/A_i)p(A_i, x_i) = q_i - q_{i-1}$ and $t_{i-1}^2 - t_{i-1} = t_{i-2}^2$ (here $t_{-1} = 0$); thus, (11.3.7) can be rewritten as

$$0 \geq \frac{t_{i-1}^2}{A_i}\varepsilon_{i+1} - \frac{t_{i-2}^2}{A_i}\varepsilon_i + (x^*)^T(q_i - q_{i-1}) + q_{i-1}^T(q_i - q_{i-1}) + \frac{1}{2}|q_i - q_{i-1}|^2 =$$

$$= \frac{t_{i-1}^2}{A_i}\varepsilon_{i+1} - \frac{t_{i-2}^2}{A_i}\varepsilon_i + (x^*)^T(q_i - q_i - 1) + \frac{1}{2}|q_i|^2 - \frac{1}{2}|q_{i-1}|^2.$$

Since the quantities A_i do not decrease with i and ε_{i+1} is nonnegative, we shall only strengthen the resulting inequality by replacing the coefficient t_{i-1}^2/A_i at ε_{i+1} by t_{i-1}^2/A_{i+1}; thus, we come to the inequality

$$0 \geq \frac{t_{i-1}^2}{A_{i+1}}\varepsilon_{i+1} - \frac{t_{i-2}^2}{A_i}\varepsilon_i + (x^*)^T(q_i - q_{i-1}) + \frac{1}{2}|q_i|^2 - \frac{1}{2}|q_{i-1}|^2;$$

taking sum of these inequalities over $i = 1, ..., N$, we come to

$$\frac{t_{N-1}^2}{A_{N+1}}\varepsilon_{N+1} \leq -(x^*)^T q_N - \frac{1}{2}|q_N|^2 \leq \frac{1}{2}|x^*|^2 \equiv R^2(f)/2; \tag{11.3.8}$$

thus,

$$\varepsilon_{N+1} \leq \frac{A_{N+1}R^2(f)}{2t_{N-1}^2}. \tag{11.3.9}$$

As we know, $A_{N+1} \leq 2\widehat{\mathcal{A}}(f)$ and $t_i \geq (i+1)/2$; therefore (11.3.9) implies (11.3.3). ∎

11.4 Smooth strongly convex problems

In convex optimization, when speaking about a "simple" nonlinear problem, traditionally it is meant that the problem is unconstrained and the objective is smooth and strongly convex, so that the problem is

$$(f) \qquad \textit{minimize } f(x) \textit{ over } x \in \mathbf{R}^n,$$

and the objective f is continuously differentiable and such that for two positive constants $l < L$ and all x, x' one has

$$l|x - x'|^2 \leq (f'(x) - f'(x'))^T(x - x') \leq L|x - x'|^2.$$

It is a straightforward exercise in Calculus to demonstrate that this property is equivalent to

$$f(x) + (f'(x))^T h + \frac{l}{2}|h|^2 \leq f(x + h) \leq f(x) + (f'(x))^T h + \frac{L}{2}|h|^2 \tag{11.4.1}$$

for all x and h. A function satisfying the latter inequality is called (l, L)-strongly convex, and the quantity

$$Q = \frac{L}{l}$$

is called the *condition number* of the function. If f is twice differentiable, then the property of (l, L)-strong convexity is equivalent to the relation

$$lI \leq f''(x) \leq LI, \ \forall x,$$

where the inequalities should be understood in the operator sense. Note that convexity + $C^{1,1}$-smoothness of f is equivalent to (11.4.1) with $l = 0$ and some finite L (L is an upper bound for the Lipschitz constant of the gradient of f).

We already know how one can minimize efficiently $C^{1,1}$-smooth convex objectives; what happens if we know in advance that the objective is strongly convex? The answer is given by the following statement.

Theorem 11.4.1 *Let $0 < l \leq L$ be a pair of reals with $Q = L/l \geq 2$, and let $\mathcal{SC}_n(l, L)$ be the family of all problems (f) with (l, L)-strongly convex objectives. Let us provide this family by the relative accuracy measure*

$$\nu(x, f) = \frac{f(x) - \min f}{f(\bar{x}) - \min f}$$

(here $\bar{x}$ is a once for ever fixed starting point) and the standard first-order oracle. The complexity of the resulting class of problems satisfies the inequalities as follows:

$$O(1)\min\{n, Q^{1/2}\ln(\frac{1}{2\varepsilon})\} \le \mathrm{Compl}(\varepsilon) \le O(1)Q^{1/2}\ln(\frac{2}{\varepsilon}), \tag{11.4.2}$$

where $O(1)$ are positive absolute constants.

Today I shall not speak about the lower complexity bound - it will be obtained in the mean time as a byproduct of our considerations related to quadratic problems. What we shall focus on is the upper bound. This bound is dimension-independent and in the large scale case is sharp - it coincides with the lower bound within an absolute constant factor. In fact we can say something reasonable about the lower complexity bound in the case of a fixed dimension as well; namely, if $n \ge Q^{1/2}$, then, besides the lower bound (11.4.2) (which does not say anything reasonable for small ε), the complexity admits also the lower bound

$$\mathrm{Compl}(\varepsilon) \ge O(1)\frac{\sqrt{Q}}{\ln Q}\ln(\frac{1}{2\varepsilon})$$

which is valid for all $\varepsilon \in (0,1)$.

Let me stress that what is important in the above complexity results is not the logarithmic dependence on the accuracy, but the dependence on the condition number Q. A linear convergence on the family in question is not a great deal; it is possessed by more or less all traditional methods. E.g., the Gradient Descent with constant step $\gamma = \frac{1}{L}$ solves problems from the family $\mathcal{SC}_n(l,L)$ within relative accuracy ε in $O(1)Q\ln(1/\varepsilon)$ steps. It does not mean that the Gradient Descent as good as a method might be: in actual computations the factor $\ln(1/\varepsilon)$ is a quite moderate integer, something less than 20, while you can easily meet with condition number Q of order of thousands and tens of thousands. In order to compare a pair of methods with efficiency estimates of the type $A\ln(1/\varepsilon)$ it is reasonable to look at the value of the factor denoted by A (usually this value is a function of some characteristics of the problem, like its dimension, condition number, etc.), since this factor, not $\log(1/\varepsilon)$, is responsible for the efficiency. And from this viewpoint the Gradient Descent is bad - its efficiency is too sensitive to the condition number of the problem, it is proportional to Q rather than to $Q^{1/2}$, as it should be according to our complexity bounds. Let me add that no traditional method, even among the conjugate gradient and the quasi-Newton ones, is known to possess the "proper" efficiency estimate $O(Q^{1/2}\ln(1/\varepsilon))$.

The upper complexity bound $O(1)Q^{1/2}\ln(2/\varepsilon)$ can be easily obtained by applying to the problem the Nesterov method with properly chosen restarts. Namely, consider the problem

$$(f_G) \qquad \textit{minimize } f(x) \textit{ s.t. } x \in G,$$

associated with an (l,L)-strongly convex objective and a closed convex subset $G \subset \mathbf{R}^n$ which contains our starting point $\bar{x}$; thus, we are interesting in something more general than the simplest unconstrained problem (f) and less general than the composite problem (11.1.1). Assume that we are given in advance the parameters (l,L) of strong convexity of f (this assumption, although not too realistic, allows to avoid sophisticated technicalities). Given these parameters, let us set

$$N = 4\sqrt{\frac{L}{l}}$$

and let us apply to (f_G) N steps of the Nesterov method, starting with the point $y^0 \equiv \bar{x}$ and setting $A_0 = L$. Note that with this choice of A_0, as it is immediately seen from the proof of Theorem 11.3.1, we have $A_i \equiv A_0$ and may therefore skip tests (11.3.2). Let y^1 be the approximate solution to the problem found after N steps of the Nesterov method (in the initial notation it was called y_{N+1}). After y^1 is found, we restart the method, choosing y^1 as the new starting point. After N steps of the restarted method we restart it again, with the starting point y^2 being the result found so far, etc. Let us prove that this scheme with restarts ensures that

$$f(y^i) - \min_G f \leq 2^{-i}[f(\bar{x}) - \min_G f],\ i = 1, 2, ... \tag{11.4.3}$$

Since to pass from y^i to y^{i+1} it requires $N = O(\sqrt{Q})$ oracle calls, (11.4.3) implies the upper bound in (11.4.2).

Thus, we should prove (11.4.3). This is immediate. Indeed, let $z \in G$ be a starting point, and z^+ be the result obtained after N steps of the Nesterov method started at z. From Theorem 11.3.1 it follows that

$$f(z^+) - \min_G f \leq \frac{4LR^2(z)}{(N+1)^2}, \tag{11.4.4}$$

where $R(z)$ is the distance between z and the minimizer x^* of f over G. On the other hand, from (11.4.1) it follows that

$$f(z) \geq f(x^*) + (z - x^*)^T f'(x^*) + \frac{l}{2}R^2(z). \tag{11.4.5}$$

Since x^* is the minimizer of f over G and $z \in G$, the quantity $(z - x^*)^T f'(x^*)$ is nonnegative, and we come to

$$f(z) - \min_G f \geq \frac{lR^2(G)}{2}$$

This inequality, combined with (11.4.4), implies that

$$\frac{f(z^+) - \min_G f}{f(z) - \min_G f} \leq \frac{8L}{l(N+1)^2} \leq \frac{1}{2}$$

(the concluding inequality follows from the definition of N), and (11.4.3) follows.

Lecture 12

Unconstrained quadratic optimization

We have studied the complexity and efficient algorithms for nonsmooth and smooth convex minimization. Of course, the presentation was uncomplete; e.g., we have passed from the general case, where the objectives might be simultaneously nonsmooth and degenerate - with "flat graphs" - directly to the case of "very good" objectives, i.e., with Lipschitz continuous gradient, which is the strongest smoothness assumption; last lecture we spoke also about strongly convex optimization, thus imposing on the objective simultaneously the strongest smoothness and the strongest nondegeneracy assumptions. The classes of problems studied so far can be regarded as extreme points in the world of nonlinear convex optimization; there are also "interior points" which are obtained by imposing on the objective some intermediate smoothness restrictions, like Holder continuity of the gradient, and these restrictions can be combined with various hypotheses on the degree of nondegeneracy. For the variety of these "intermediate" problem classes we also know the complexity and the optimal methods, but I decided to restrict the course with the basic results related to the "extreme classes" of problems. This program is more or less completed; the remaining topics are rather specific, but I think they are of definite interest.

12.1 Complexity of quadratic problems: motivation

Our today topic is *unconstrained quadratic optimization*, i.e., methods for minimizing convex quadratic forms

$$f(x) = f_{A,b}(x) = x^T Ax - 2b^T x : \mathbf{R}^n \to \mathbf{R}$$

over the whole $\mathbf{R}^n$. Everybody knows that above quadratic form is a convex function if and only if the symmetric matrix A is positive semidefinite; such a convex function attains its minimum on $\mathbf{R}^n$ if and only if it is below bounded, and if and only if the linear equation

$$Ax = b \tag{12.1.1}$$

is solvable; the solutions to this equations are exactly the minimizers of f. Thus, the problem of minimizing a quadratic form is nothing but a problem of solving a linear $n \times n$ equation with a symmetric positive semidefinite matrix. It is one of the standard problems of Linear Algebra,

and you may, of course, ask why should we deal with this Linear Algebra problem in our course devoted to convex minimization and what news in this more than traditional area can result from our complexity approach. The answer is as follows. When considering convex optimization problems, we normally spoke about the *first-order methods*, i.e., about iterative methods which use the first-order information on the objective – its values and its first-order derivatives. In the case of a quadratic objective these methods, consequently, need computing the matrix-vector products Ax and do not require neither a direct access to the matrix A, nor updating this matrix. In other words, a first-order method for quadratic minimization is an iterative procedure which solves a system of linear equations via matrix-vector multiplications, without direct access to the matrix of the system. For these iterative methods, A is not an explicit part of the input; it is given implicitly, via an *oracle* which, given on input a vector x, reports on output the product Ax (as far as the right hand side vector b is concerned, it is reasonable to assume that it is an explicit part of the input - it is given to the method in advance). Thus, what we are about to do is to investigate the complexity of solving linear systems of equations (with symmetric positive semidefinite matrices) by iterative methods which use only matrix-vector multiplications. Note that most of the usual Linear Algebra routines, like Gauss elimination, Choleski decomposition, etc. - are not iterative, I mean, they need a direct access to the matrix and update it when solving the system.

Now, you can ask: in practice one hardly can meet with a situation when a method indeed has no access to the matrix of a linear system, so why it is reasonable to investigate the complexity of linear equations with respect to iterative methods of the indicated, rather specific, type? The answer is as follows: direct methods of Linear Algebra, like Colossi decomposition, spend something $O(n^3)$ arithmetic operations in order to solve the problem; this estimate is sharp for the case of dense matrices, same as sparse, but "unstructured" matrices; better estimates are known only for sparse matrices of a good structure. And the situation is as follows: before you pay $O(n^3)$ arithmetic operations, you have nothing - no approximate solution, bad or good, at all; after you have paid this cost, you get all - I mean, the exact solution (I am considering the idealized case of exact real arithmetic, just ignoring the rounding errors). In contrast to this, in an iterative method you have certain approximate solution at each step. Let $K(\nu)$ be the number of steps of the iterative method in question which is sufficient to solve all problems from a given family within a desired accuracy ν; for many interesting families of problems (say, for those given by bounds on the condition number) this quantity is independent of the size n of the problem. Whenever we deal with such a family of problems, we know in advance that in the large-scale case it would be better to use the iterative method: to solve the problem within the desired accuracy, it suffices to perform $K(\nu)$ matrix-vector multiplications, i.e., to perform $MK(\nu)$ arithmetic operations, where M is the cost of a single matrix-vector multiplication. Even for dense matrices, $M = n^2$; thus, the arithmetic cost of an ν-solution for the iterative method in question will be at most $O(n^2K(\nu))$, and for a direct method $O(n^3)$. Thus, the iterative method will be more efficient than the direct one whenever $n >> K(\nu)$. This advantage of iterative methods becomes more visible if we deal with large-scale sparse matrices, since here $M << n^2$; in the case of "unstructured" sparsity the arithmetic cost of a direct method still is $O(n^3)$, and we get an additional factor M/n^2 in favour of the iterative method.

I believe that the above reasoning explains why investigating complexity of linear operator equations with respect to iterative methods based on matrix-vector multiplications is of some interest for Linear Algebra; and it is for sure interesting for convex optimization. Roughly

speaking, it turns out that classes of smooth convex and strongly convex optimization problems are as difficult as their subclasses formed by quadratic problems, and these subclasses are the main source of lower complexity bounds for related classes of non-quadratic problems.

12.2 Families of source-representable quadratic problems

Thus, we are going to investigate the complexity of linear operator equations with respect to iterative methods based on matrix-vector multiplications. Let me start with the following convention. If A is a symmetric positive semidefinite matrix and r is a nonnegative real, then the operator A^r is defined in the natural way an is a well-defined operator on the whole $\mathbf{R}^n$. If $r < 0$, we define A^r as the operator on the image space of A which is inverse to A^{-r}. In other words, if e_i, λ_i are the eigenvectors and the eigenvalues of A, then, for $r \geq 0$, the operator A^r has e_i as eigenvectors with the eigenvalues λ_i^r ($0^0 = 1$). If $r < 0$, then A^r is defined on the linear span of those e_i which are associated with nonzero λ_i, and the eigenvalues of A^r associated with these e_i are λ_i^r.

Now let me describe the families of equations in question. Let Σ be a closed compact subset of the nonnegative ray $[0, \infty)$, τ be a real and R be a positive real. Consider the family $\mathcal{U}_n(\Sigma, \tau, R)$ comprised of all equations

$$Ax = b \tag{12.2.1}$$

with a symmetric $n \times n$ positive semidefinite matrix A, such that

(i) the spectrum of A belongs to Σ;

(ii) the equation (12.2.1) is solvable, and its normal (i.e., of the minimal Euclidean norm) solution $x^*(A, b)$ can be represented as

$$x^*(A, b) = A^\tau u \tag{12.2.2}$$

with some $u \in \mathbf{R}^n$ such that $|u| \leq R$.

Relation (12.2.2), by reasons coming from certain physical interpretations, is called the "source-representability" of the solution. The meaning of this assumption is very clear: the matrix A generates the scale of norms $|A^r x|$ on $\mathbf{R}^n$ (more exactly, if A is degenerate and $r > 0$, these are seminorms on $\mathbf{R}^n$, not norms, and is A is degenerate and $r < 0$, these are norms not on $\mathbf{R}^n$, but on the image space of A); assumption (ii) says simply that the norm of the solution in a fixed norm of this family should be bounded from above by a given quantity.

Let us provide the family with the accuracy measure

$$\nu_\omega(x; A, b) = |A^\omega(x - x^*(A, b))|,$$

if the right hand side is well-defined; otherwise $\nu_\omega(x; A, b)$ is $+\infty$. Here ω is a real parameter.

Consider typical examples.

(I) The family $\mathcal{U}_n(\Sigma, 0, R)$ is comprised of all solvable linear operator equation of a given size with the eigenvalues of the (symmetric) matrix of the equation belonging to a given compact set and the solution of a norm not exceeding R;

(Ia) The accuracy ν_0 is exactly the distance between the approximate and the exact solutions.

(II) The family $\mathcal{U}_n(\Sigma, -1, R)$ is comprised of all solvable linear operator equations of a given size with the eigenvalues of the (symmetric) matrix of the equation belonging to Σ and the norm of the right hand side vector not exceeding R

(indeed, to say that $x^*(A,b) = A^{-1}u$ with $|u| \le R$ is exactly the same as to say that $b = Ax^*(A,b)$ is of the norm $\le R$);

(IIa) The accuracy ν_1 is exactly the norm $|Ax - b|$ of the residual $Ax - b$.

(III) The family $\mathcal{U}_n(\Sigma, -\frac{1}{2}, R)$ is comprised of all solvable linear operator equations of a given size with the eigenvalues of the (symmetric) matrix of the equation belonging to Σ and such that the initial residual in terms of the associated with the equation quadratic minimization problem, i.e., the quantity

$$f_{A,b}(0) - \min_x f_{A,b}(x),$$

is $\le R^2$

(indeed, we have $f_{A,b}(0) - \min_x f_{A,b}(x) = (x^*(A,b))^T Ax^*(A,b) = |A^{1/2}x^*(A,b)|^2$, so that to say that $x^*(A,b) = A^{-1/2}u$ with $|u| \le R$ is the same as to say that $|A^{1/2}x^*(A,b)| \le R$, i.e., that $(x^*(A,b))^T Ax^*(A,b) \le R^2$);

(IIIa) The accuracy $\nu_{1/2}$ is the norm of $A^{1/2}(x - x^*(A,b))$, so that $\nu^2_{1/2}(x; A, b)$ is exactly the residual $f_{A,b}(x) - \min f_{A,b}$ in terms of the quadratic objective associated with the equation.

After we have defined the families of problems we are interested in and have equipped the families with accuracy measures, we may pose the question of what is the complexity of a family with respect to iterative methods and what are the associated optimal methods. We actually know a significant part of the answer.

12.3 Lower complexity bounds

Recall that we already know the following important fact (Proposition 10.2.1, Lecture 10):

Let $f_{A,b}$ be a convex quadratic form on $\mathbf{R}^n$ and $\mathcal{F}$ be the family of rotations of this form, i.e., the family of all quadratic forms of the type $f_{U^TAU,U^Tb}(x) \equiv f(Ux)$ associated with orthogonal linear operators U which remain b invariant. Let also $\mathcal{M}$ be an iterative first order method of complexity M for minimizing forms from $\mathcal{F}$. Then there is a problem $f = f_{A',b'} \in \mathcal{F}$ such that the result $\bar{x}$ formed by the method $\mathcal{M}$ as applied to f belongs to the $(2M+1)$-st Krylov subspace

$$\mathcal{E}_{2M+1}(A', b') = \mathrm{Lin}\{b', A'b', (A')^2b', ..., (A')^{2M}b'\}$$

of the problem $f_{A',b'}$; in other words,

$$\bar{x} = p(A')b'$$

for certain polynomial p of a degree $\le 2M$.

Now note that a first-order method for minimizing quadratic forms is nothing but an iterative method, based on matrix-vector multiplications, for the associated linear operator equations. Consider a family of linear operator equations which is invariant with respect to rotations (so that (U^TAU, U^Tb) belongs to $\mathcal{F}$ whenever U is on orthogonal matrix and $(A,b) \in \mathcal{F}$). The above statement says that if $\mathcal{M}$ is an iterative method for solving linear operator equations from $\mathcal{F}$ with the complexity equal to M, then, for any problem (A,b) from the family, there exists a rotation $(A',b') = (U^TAU, U^Tb)$ of the problem such that the result $\bar{x}(A',b')$ of $\mathcal{M}$ as applied to (A',b') is of the form $p(A')b'$ for certain polynomial p of a degree $\le 2M$. In particular, if $\nu(x; A, b)$ is the accuracy measure in question, then the error $\nu(\mathcal{M}, (A',b'))$ of the method on the problem (A',b') is at least the quantity

$$\nu^*(2M, A', b') = \min_{p:\deg p \le 2M} \nu(p(A')b'; A', b'). \tag{12.3.1}$$

Since $\mathcal{F}$ is invariant with respect to rotations, (A', b') belongs to $\mathcal{F}$ whenever (A, b) does, so that the left hand side of (12.3.1) is less than or equal to the inaccuracy of the method on the whole family. If, in addition, the accuracy measure also is invariant with respect to rotations:

$$\nu(U^T x; U^T A U, U^T b) = \nu(x; A, b)$$

for any $(A, b) \in \mathcal{F}$, any x and any rotation U, then the right hand side in (12.3.1) is the same as if it were $A' = A$, $b' = b$. Thus, we come to the following

Proposition 12.3.1 *Let $\mathcal{F}$ be a family of linear operator equations which is invariant with respect to rotations and $\nu(x; A, b)$ be an accuracy measure for the family, which also is invariant with respect to rotations. Then, for any method $\mathcal{M}$, based on matrix-vector multiplications, of complexity M on the family the worst-case accuracy of $\mathcal{M}$ on the family satisfies the inequality as follows:*

$$\nu(\mathcal{M}) \geq \sup_{(A,b)\in\mathcal{F}} \min_{p:\deg p \leq 2M} \nu(p(A)b; A, b). \tag{12.3.2}$$

This proposition can be directly applied to the families $\mathcal{U}_n(\Sigma, \tau, R)$ and the accuracy measures $\nu_\omega(x; A, b)$ which clearly possess the required invariance properties. Moreover, for these families and accuracy measures, the right hand side in (12.3.2) can be computed more or less explicitly:

Proposition 12.3.2 *Consider the case of $\mathcal{F} = \mathcal{U}(\Sigma; \tau, R)$, $\nu(x; A, b) = \nu_\omega(x; A, b)$. Assume that 0 is not an isolated point of Σ (i.e., either $0 \notin \Sigma$, or 0 is a limiting point of Σ); if $0 \in \Sigma$, assume, in addition, that $\omega + \tau > 0$.Finally assume that that $2M + 1 < n$. Then the right hand side in (12.3.2) equals to*

$$R\nu^*(2M; \Sigma; \tau + \omega) \equiv R \min_{p:\deg p \leq 2M} \max_{t\in\Sigma} |t|^{\tau+\omega} |1 - tp(t)|.$$

In particular, the ν_ω-inaccuracy of any method $\mathcal{M}$ of the complexity M on the family $\mathcal{U}_n(\Sigma, \tau, R)$ satisfies the lower bound

$$\nu(\mathcal{M}) \geq R\nu^*(2M, \Sigma; \tau + \omega). \tag{12.3.3}$$

Proof. Let $(A, b) \in \mathcal{U}(\Sigma, \tau, R)$, and let $x^*(A, b) = A^\tau u$, $|u| \leq R$, be the corresponding "source representation" of the normal solution to the problem. Of course, in this representation u can be chosen from the image space of A:

$$u = \sum_{i=1}^{k} u_i e_i,$$

where $k \leq n$, e_i are certain mutually orthogonal unit eigenvectors of A with mutually disjoint positive eigenvalues λ_i. Let

$$\mu_i = u_i^2,$$

so that

$$\sum_{i=1}^{k} \mu_i \leq R^2. \tag{12.3.4}$$

We have

$$x^*(A, b) = \sum_{i=1}^{k} u_i \lambda_i^\tau e_i, \; b = Ax^*(A, b) = \sum_{i=1}^{k} u_i \lambda_i^{1+\tau} e_i,$$

whence for a polynomial p one has

$$p(A)b = \sum_{i=1}^{k} u_i p(\lambda_i)\lambda_i^{1+\tau} e_i$$

and

$$\nu_\omega^2(p(A)b; A, b) = |A^\omega(p(A)b - x^*(A,b))|^2 = \sum_{i=1}^{k} \lambda_i^{2\tau+2\omega}(1-\lambda_i p(\lambda_i))^2\mu_i. \tag{12.3.5}$$

Since $\lambda_i \in \Sigma$, this relation combined with (12.3.4) implies that

$$\nu_\omega^2(p(A)b; A, b) \le R^2 \max_{t\in\Sigma} t^{2\tau+2\omega}(1-tp(t))^2,$$

whence

$$\nu_\omega(p(A)b; A, b) \le R \max_{t\in\Sigma} t^{\tau+\omega}|1-tp(t)|. \tag{12.3.6}$$

Choosing as p the polynomial p_{2M}^* of the degree $\le 2M$ which solves the saddle point problem

$$\mathcal{S}: \qquad \min_{p:\deg p\le 2M} \max_{t\in\Sigma} t^{\tau+\omega}|1-tp(t)|$$

(the problem clearly is solvable), we make the right hand side of (12.3.6) equal to $R\nu^*(2M, \Sigma, \tau+\omega)$; therefore (12.3.6) implies that the right hand side of (12.3.2) is $\le R\nu^*(2M, \Sigma, \tau+\omega)$.

To prove the inverse inequality, let us act as follows. The quantity $\nu^*(2M, \Sigma, \tau+\omega)$, by construction, is the quality of the best, in the uniform on Σ norm, approximation of the function

$$f(t) = t^{\tau+\omega}$$

by a linear combination of the $2M+1$ functions

$$f_i(t) = t^{i+\tau+\omega},\ i = 1, ..., 2M+1;$$

since $\tau+\omega \ge 0$, all aforementioned functions are continuous on the (compact) set Σ. From Exercise 1.3.16 we know that there is a finite subset Σ' of Σ with at most $2M+2$ points such that it is as difficult to approximate f by a linear combination of f_i on Σ' as it is on the whole Σ:

$$\nu^*(2M, \Sigma, \tau+\omega) = \min_{p:\deg p\le 2M} \max_{t\in\Sigma'} t^{\tau+\omega}|1-tp(t)|. \tag{12.3.7}$$

Let $\Sigma' = \{\lambda_1 < \lambda_2 < ... < \lambda_k\}$, with $k \le 2M+2$; note that the cardinality k of the set Σ' is $\le n$ in view of our premise. Note also that we may assume that $\lambda_1 > 0$. Indeed, all λ_i belong to Σ, so that the inequality in question is evident if $0 \notin \Sigma$. If $0 \in \Sigma$, then, in view of our premise, $\tau+\omega > 0$, so that the right hand side of (12.3.7) will remain unchanged if we delete from Σ' the point 0 (if initially it belonged to the set).

Further, (12.3.7) can be rewritten as

$$[\nu^*(2M, \Sigma, \tau+\omega)]^2 = \min_{p:\deg p\le 2M} \max_{t\in\Sigma'} t^{2\tau+2\omega}(1-tp(t))^2;$$

due to the standard duality reasons, there exists a probability distribution of masses $\pi_1, ..., \pi_k$ on Σ' such that the right hand side in the latter equality is

$$\min_{p:\deg p\le 2M} \sum_{i=1}^{k} \pi_i \lambda_i^{2\tau+2\omega}(1-\lambda_i p(\lambda_i))^2,$$

so that

$$\nu^*(2M,\Sigma,\tau+\omega) = \min_{p:\deg p\le 2M} \sum_{i=1}^{k} \pi_i \lambda_i^{2\tau+2\omega}(1-\lambda_i p(\lambda_i))^2. \tag{12.3.8}$$

Now consider the following problem $(A,b) \in \mathcal{U}_n(\Sigma,\tau,R)$: the first k eigenvalues of A are λ_i, $i=1,...,k$ (the remaining, if any, are arbitrary points of Σ); the right hand side b is given by

$$b = R\sum_{i=1}^{k} \lambda_i^{1+\tau}\sqrt{\pi_i}\, e_i,$$

where $e_1, ..., e_k$ are mutually orthogonal unit eigenvectors of A associated with the eigenvalues $\lambda_1, ..., \lambda_k$. It is immediately seen that the problem does belong to the family and that for this particular problem the quantity

$$\min_{p:\deg p\le 2M} \nu_\omega^2(p(A)b; A, b)$$

coincides with the right hand side of (12.3.8) (cf. (12.3.5)); thus, (12.3.8) implies that the right hand side of (12.3.2) is $\ge R\nu^*(2M,\Sigma,\tau+\omega)$. The inverse inequality was already proved. ∎

12.4 Complexity of linear operator equations

Proposition 12.3.2 gives us certain lower bound for the complexity of the classes $\mathcal{U}(\Sigma,\tau,R)$; these bounds, anyhow, are not in our usual form: they are lower bounds for inaccuracy of a method of a given complexity, not bounds for the complexity of a method of a given inaccuracy. Of course, there is no problem in reformulating the bounds in our standard form. To this end it suffices to pass from the function

$$\nu^*(N,\Sigma,\gamma) = \min_{p:\deg p\le N} \max_{t\in\Sigma} t^\gamma |1-tp(t)|$$

to the inverse function

$$N^*(\nu,\Sigma,\gamma) = \min\{N \mid \nu^*(N,\Sigma,\gamma) \le \nu\}.$$

With this definition, Proposition 12.3.2 provides us with the lower bound in the following complexity result:

Theorem 12.4.1 *Let us equip the family $\mathcal{U}_n(\Sigma,\tau,R)$ of linear operator equations with the accuracy measure $\nu_\omega(x;A,b)$. Assume that 0 is not an isolated point of Σ; in the case of $0\in\Sigma$ assume also that $\tau+\omega>0$. Under these assumptions, the complexity of the family satisfies the inequalities as follows:*

$$\min\{\frac{n-1}{2}; \frac{N^*(\nu/R,\Sigma,\tau+\omega)}{2}\} \le \mathrm{Compl}(\nu) \le \min\{n; N^*(\nu/R,\Sigma,\tau+\omega)\}. \tag{12.4.1}$$

Proof. The lower bound is an immediate reformulation of Proposition 12.3.2. Indeed, if M is the complexity of a method $\mathcal{M}$ which solves all problems from the family within inaccuracy ν, then, according to Proposition 12.3.2, either $2M+1 \geq n$, or $\nu^*(2M, \Sigma, \tau+\omega) \leq \nu/R$; in the latter case $2M \geq N^*(\nu/R, \Sigma, \tau+\omega)$ by definition of this latter quantity, and we come to the lower bound in (12.4.1).

To prove the upper bound, let us act as follows. The complexity of any family of quadratic problems is $\leq n$, since in n matrix-vector multiplications one can identify the matrix of the equation and then solve the equation without further calls to the oracle. It remains to demonstrate that the complexity is also bounded from above by the quantity $N = N^*(\nu/R, \Sigma, \tau+\omega)$. Indeed, by definition of this quantity there exists a polynomial p_N^* of the degree N such that

$$\max_{t\in\Sigma} t^{\tau+\omega}|1 - tp_N^*(t)| \leq \nu/R.$$

Consider the following *Tschebyshev method* for finding an approximate solution $\bar{x}(A,b)$ to a problem (A,b):

$$\bar{x}(A,b) = p_N^*(A)b.$$

To find $\bar{x}$, it requires exactly N multiplications of vectors by A (look at the Horner scheme). From inequality (12.3.6) it follows that the inaccuracy of the method on any problem instance from the family is $\leq \nu$, as claimed. ∎

We have obtained tight (coinciding with each other within factor 2) upper and lower bounds for the complexity of the families of linear operator equations in question, which is fine. As a byproduct, we have pointed out optimal methods - the Tschebyshev ones. As these methods appeared in our reasoning, they were tuned to the required accuracy; of course, this is not so reasonable, and a reasonable version of the method generates the sequential approximate solutions as

$$x_k(A,b) = p_k^*(A)b,$$

where p_k^* is the optimal solution to the optimization problem

$$\textit{find the polynomial of the degree } k \textit{ which minimizes the functional } \max_{t\in\Sigma} t^{\tau+\omega}|1 - tp(t)|,$$

which is the classical problem of the best uniform approximation.

Unfortunately, to use the Tschebyshev methods in practice is not so easy, not only because one should solve in advance certain nontrivial problems of the best uniform approximation, but mainly because normally you have no data even to pose these problems: to this end, you should know in advance the localizer Σ of the spectrum of the operator, same as the parameter τ of the family in question. Besides this, you should decide in advance what is the accuracy measure you are interested in, since even for a fixed family of problems different accuracy measures result in different Tschebyshev algorithms. Thus, Tschebyshev approach is not so attractive from the practical viewpoint; we shall see in the next lecture that there is a *single* method, namely the Conjugate Gradient one, which behaves itself at *any* family in question, with respect to *any* of our accuracy measures, in the optimal way, so that to get an optimal method, there in fact is no necessity to use the Tschebyshev approach. The actual meaning of our complexity result, in contrast to our usual practice, is in indicating tight complexity bounds rather than optimal methods. To the moment these bounds are represented in a rather sophisticated form of optimal values in certain problems of the best uniform approximation. In more or less all

interesting for applications cases these optimal values admit good explicit estimates (or even explicit representation). Let us look at several typical cases.

Example 1: nondegenerate problems. Let $\Sigma = [l, L]$ with certain positive $l < L$; the assumption that the spectrum of a symmetric positive semidefinite matrix A belongs to $[l, L]$ means exactly that any quadratic form

$$f_{A,b}(x) = x^T Ax - 2b^T x + c$$

is strongly convex with the parameters $2l, 2L$. In order to evaluate the function

$$\nu^*(N, [l, L], \gamma) = \min_{p:\deg p \le N} \max_{l \le t \le L} t^\gamma |1 - tp(t)|$$

it is reasonable to start with the case of $\gamma = 0$, where the function can be computed explicitly; the optimal polynomial is given by

$$1 - tp_N^*(t) = T_{N+1}^{-1}(\frac{Q+1}{Q-1}) T_{N+1}(\frac{2t - l - L}{L - l}),\ Q = \frac{L}{l},$$

where

$$T_k(z) = \cos(k \arccos z)$$

are the Tschebyshev polynomials (I shall not prove here the statements of this type; this is the subject of exercises accompanying the lecture). Further,

$$\left(\frac{\sqrt{Q}-1}{\sqrt{Q}+1}\right)^{N+1} \ge \nu^*(N, [l, L], 0) = T_{N+1}^{-1}(\frac{Q+1}{Q-1}) \ge \frac{1}{2}\left(\frac{\sqrt{Q}-1}{\sqrt{Q}+1}\right)^{N+1}. \tag{12.4.2}$$

This is an explicit representation of $\nu^*(\cdot, [l, L], \gamma)$ for the case of $\gamma = 0$; for the case of an arbitrary γ, one can use simple estimates

$$L^\gamma \nu^*(N, [l, L], 0) \ge \nu^*(N, [l, L], \gamma) \ge l^\gamma \nu^*(N, [l, L], 0),\ \gamma > 0; \tag{12.4.3}$$

for the case of negative γ, you need to invert the inequality signs in this chain. The estimates (12.4.3) identify ν^* within the factor $Q^{|\gamma|}$.

The bounds on $\nu^*(\cdot, [l, L], \gamma)$ immediately imply bounds on the inverse function $N^*(\cdot, \gamma)$ involved in our complexity result; e.g., from (12.4.2) it follows that

$$c\ \ln^{-1}\left(\frac{\sqrt{Q}+1}{\sqrt{Q}-1}\right)\ln(\frac{R}{\nu}) \le N^*(\nu/R, [l, L], 0) \le C\ \ln^{-1}\left(\frac{\sqrt{Q}+1}{\sqrt{Q}-1}\right)\ln(\frac{R}{\nu}),\ \nu < R/2; \tag{12.4.4}$$

Consider, for example, the family of quadratic problems $\mathcal{U}_n([l, L], -\frac{1}{2}, R)$; this class is comprised of the $[2l, 2L]$-strongly convex quadratic forms $f(x)$ on $\mathbf{R}^n$ such that

$$f(0) - \min f \le R^2 \tag{12.4.5}$$

(see (III)). Let $\omega = 1/2$, which, as we know from (IIIa), corresponds to the accuracy measure

$$\sqrt{f(x) - \min f}.$$

From (12.4.1) and (12.4.4) it follows that the complexity of the family in question (associated with $\omega = 1/2$) is, up to an absolute multiplicative constant, equal to

$$\min\{n, Q^{1/2}\ln(R/\nu)\}$$

(I assume that $Q \geq 2$ and $\nu < R/2$). On the other hand, consider the family $\mathcal{SC}_n(2l, 2L)$ of all, not necessary quadratic, strongly convex with parameters $2l, 2L$ unconstrained problems on $\mathbf{R}^n$ (we have dealt with this family in the previous lecture). This family clearly contains $\mathcal{U}_n([l, L], -\frac{1}{2}, R)$, and a method which solves all problems from $\mathcal{SC}_n$ with certain relative accuracy ε for sure solves all problems from $\mathcal{U}_n$ within absolute inaccuracy (measured by $\nu_{\frac{1}{2}}(\cdot)$) $R\sqrt{\varepsilon}$; consequently, the complexity of the method on $\mathcal{U}_n$ (and therefore on the larger class $\mathcal{SC}_n$) is at least

$$O(1)\min\{n, Q^{1/2}\ln(1/\varepsilon)\};$$

since this lower bound is valid for the complexity of an arbitrary method which solves the problems from $\mathcal{SC}_n(2l, 2L)$ within relative inaccuracy ε, we come to the lower bound

$$O(1)\min\{n, Q^{1/2}\ln(1/\varepsilon)\}$$

on the ε-complexity of the class $\mathcal{SC}_n(2l, 2L)$ of strongly convex optimization problems; this is the lower complexity bound announced in the previous lecture.

Example 2: degenerate case. Now let $\Sigma = [0, L]$; in other words, we deal with below bounded convex quadratic forms with Lipschitz continuous, with constant $2L$, gradient, but do not impose on the forms any nondegeneracy assumptions, thus including into our considerations ill-conditioned linear operator equations. Here

$$\nu^*(N, [0, L], \gamma) = \min_{p:\deg p \leq N} \max_{0 \leq t \leq L} t^\gamma |1 - tp(t)|.$$

The exact values of ν^* are known for $\gamma = 1/2$; the optimal polynomial p is given by

$$1 - tp_N^*(t) = (-1)^N(2N+3)^{-1}\sqrt{L/t}\,T_{2N+3}(\sqrt{t/L}),$$

where T_k again are the Tschebyshev polynomials, and

$$\nu^*(N, [0, L], \frac{1}{2}) = L^{1/2}(2N+3)^{-1}. \tag{12.4.6}$$

For positive γ not equal to $1/2$ one can use the following estimates:

$$\frac{1}{2}L^\gamma(1 + (N+1)/\gamma)^{-2\gamma} \leq \nu^*(N, [0, L], \gamma) \leq L^\gamma(1 + 2(N+1)/(2\gamma+1))^{-2\gamma}, \tag{12.4.7}$$

so that the function N^* involved into the complexity bounds can be estimated as

$$c\gamma\left(\frac{RL^\gamma}{\nu}\right)^{\frac{1}{2\gamma}} \leq N^*(\nu, [0, L], \gamma) \leq C(\gamma+1)\left(\frac{RL^\gamma}{\nu}\right)^{\frac{1}{2\gamma}} \tag{12.4.8}$$

with certain positive absolute constants c and C.

Consider a couple of instructive particular cases, where the questions and the answers can be "translated to the non-quadratic language".

(A): $\tau = 0$, $\omega = 1/2$. The assumption that the convex quadratic form in question belongs to $\mathcal{U}_n([0, L], 0, R)$ means exactly that the form possesses Lipschitz continuous with constant $2L$ gradient and that we are interested in the accuracy measure which is the square root of the residual in terms of the objective. The above relations result in the complexity

$$O(1)\min\{n, \frac{R\sqrt{L}}{\nu}\};$$

in terms of the residual ε in the objective rather than the square root of this residual ν, we come to the already known to us $\varepsilon^{-1/2}$-complexity

$$O(1)\min\{n, \left(\frac{LR^2}{\varepsilon}\right)^{1/2}\},$$

which leads to the "proper" lower bound for the complexity of smooth non-quadratic convex minimization; this bound already was established in Lecture 10 (by the way, via the same proof).

(B): $\tau = 0$, $\omega = 1$. We make the same assumptions on the quadratic form as in (A), but measure the accuracy by the norm $|Ax - b|$ of the residual in associated equation, i.e. (up to factor $\frac{1}{2}$) by the norm of the gradient of the form at an approximate solution. The complexity is

$$O(1)\min\{n, \left(\frac{LR}{\nu}\right)^{1/2}\}.$$

Note that the accuracy measure can be extended on the class of smooth non-quadratic convex problems, and the corresponding complexity problem is as follows:

given a convex objective f with Lipschitz continuous, with constant L, gradient and such that the distance from the origin to the optimal set does not exceed R, what is the complexity, N, of finding an approximate minimizer x of f such that $|f'(x)| \le \nu$?

Our complexity bound for the quadratic case says that

$$N \ge O(1)\min\{n, \left(\frac{LR}{\nu}\right)^{1/2}\};$$

this lower bound gives us a correct guess of what is N: it turns out that

$$N \le O(1)\left(\frac{LR}{\nu}\right)^{1/2}\ln\left(\frac{LR}{\nu}\right).$$

We could also consider the case of $\tau = -\frac{1}{2}$ (the assumption on the quadratic form f is that it is convex and possesses Lipschitz continuous with constant $2L$ gradient and the initial residual $f(0) - \min f$ is $\le R^2$ for some given R) and $\omega = 1$ (we measure the inaccuracy by the norm of the gradient at an approximate solution); the associated complexity is

$$O(1)\min\{n, \frac{R\sqrt{L}}{\nu}\},$$

and this result, same as in the above cases, admits natural "non-quadratic" extension.

Example 3: "completely continuous operators". Let $\Sigma = \{\sigma_i\}_{i=1}^{\infty} \cup \{0\}$ be a monotonically decreasing converging to 0 sequence completed by the limit 0 of the sequence. It is easily seen that in this case

$$\nu^*(2N, \Sigma, \gamma) \leq \sigma_N^{\gamma}(1 + 2(N+1)/(2\gamma+1))^{-2\gamma}, \ \gamma > 0.$$

In particular, in the case of $\sigma_i = O(i^{-\alpha})$ we come to

$$N^*(\nu, \Sigma, \gamma) \leq const\, \nu^{-1/(\gamma(2+\alpha))};$$

in the case of $\sigma_i = ci^{-\alpha}$ this upper bound is sharp in order.

12.5 Ill-posed problems

Consider the case when 0 is a limiting point of Σ, i.e., the case of families containing ill-conditioned equations. Our complexity results say that there is certain nontrivial rate of convergence if $\tau + \omega > 0$, i.e., when the exact solution to the problem is bounded in the norm $|A^{-\tau}x|$ which is "stronger" than the norm $|A^{\omega}x|$ in which we measure the inaccuracy. Here "stronger" means that the first norm dominates, with a constant factor depending on the norm of A and $\tau + \omega$ only, the second one:

$$|A^{\omega}x| \leq \| A \|^{\tau+\omega} |A^{-\tau}x|, \ x \in \mathbf{R}^n.$$

The fact that in the case of $\tau + \omega > 0$ we have certain nontrivial rate of convergence in the problem of approximating the solution to an equation from $\mathcal{U}_n(\Sigma, \tau, R)$ in the norm $|A^{\omega}x|$ means that the problem is well-posed, in spite of possible ill-conditioness of the matrix of the equation. What happens if $\tau + \sigma \leq 0$? The answer is that here the problem is ill-posed; the complexity of the family $\mathcal{U}_n(\Sigma, \tau, R)$ with respect to the accuracy measure $\nu_{\omega}(\cdot)$ associated with $\omega \leq -\tau$ turns out to be at least $(n-3)/3$ for all $\nu < R$. This means that in the case in question the iterative methods have no advantages as compared to the direct methods of Linear Algebra: in order to obtain a progress in accuracy, one should perform no less than $O(n)$ steps, which basically means that he is enforced to identify the matrix of the equation.

12.6 Exercises: Around quadratic forms

The goal of the forthcoming exercises is to prove the statements on the best uniform approximation which were used without proofs in Lecture 12. The problem we are interested in is as follows:

We are given a nonnegative integer N, a compact subset Σ of the nonnegative ray and a real γ. It is assumed that 0 is not an isolated point of Σ (it may be a limiting point of the set); in the case of $0 \in \Sigma$ it is also assumed that $\gamma > 0$. The problem is to find the polynomial p^* of degree $\leq N$ which minimizes the functional

$$l(p) = \max_{t \in \Sigma} t^{\gamma} |1 - tp(t)| \qquad (P)$$

over the space $\mathcal{P}_N$ of all polynomials of degrees $\leq N$.

In what follows we always assume (not indicating it each time explicitly) that the cardinality of Σ is $\geq N+2$, since otherwise the optimal value in (P) is zero, and the optimal p simply extrapolates the function $1/t$ from Σ on the whole axis.

Exercise 12.6.1 *Prove that (P) is solvable.*

The following exercise gives the Tschebyshev type characterization of the optimal solution to (P):

Exercise 12.6.2 $^+$ *Let p be a polynomial of a degree $\leq N$.*

$$q(t) = t^\gamma(1-tp(t)), \;\; \delta = \max_{t\in\Sigma} |q(t)|.$$

Prove that p is an optimal solution to (P) if and only if p possesses the "alternance property": there exist $N+2$ points $\sigma_0 < \sigma_1 < ... < \sigma_{N+1}$ in Σ (the points of Tschebyshev's alternance) such that

$$q(\sigma_i) = (-1)^i\delta.$$

Exercise 12.6.3 *Let p^* be an optimal solution to (P) (in fact one can write "the" instead of "an", since the optimal set in (P) is a singleton). Prove that the polynomial*

$$s(t) = 1 - tp(t)$$

has exactly $N+1$ positive roots $t_1 < t_2 < ... < t_{N+1}$ (which separate the alternance points σ_i, $i = 0, ..., N+1$) and that

$$s(t) = \prod_{i=1}^{i}(1 - t/t_i).$$

The next exercise justifies the statements announced in Example 1, Lecture 12:

Exercise 12.6.4 * *Let $\Sigma = [l, L]$, $0 < l < L$; then the optimal solution to problem (P) associated with $\gamma = 0$ is given by the polynomial*

$$p_N^*(t) = T_{N+1}^{-1}(\frac{L+l}{L-l})T_{N+1}(\frac{2t-l-L}{L-l}),$$

and the optimal value is

$$T_{N+1}^{-1}(\frac{L+l}{L-l}),$$

T_k being the Tschebyshev polynomial of the degree k. Derive from this the estimates (161) and (162) announced in Lecture 12.

Note that problem (P) associated with $\Sigma = [l, L]$ and $\gamma = 0$ is in fact the following problem:

(P') : *given a segment Δ on the axis and a point $\bar{t}$ outside the segment, find among the polynomials of a given degree $N+1$ with the absolute value on Δ not exceeding 1 the polynomial with the largest possible absolute value at $\bar{t}$*

(in the initial formulation, $\Delta = [l, L]$ and $\bar{t} = 0$; instead of looking for the largest at $\bar{t}$ polynomial among those with the uniform norm 1 on Δ, we were looking for the smallest in the uniform

norm on Δ polynomial among the polynomials with the value 1 at $\bar{t}$; of course, these are two equivalent formulations of the same problem).

When solving (P'), we can restrict ourselves with $\Delta = [-1, 1]$; Exercise 12.6.4 says that the solution to this latter version of (P') is given by the Tschebyshev polynomial of the degree given in the formulation of the problem (*Markov's Theorem*). Surprisingly, the answer does not depend on the location of the point $\bar{t}$! This is a rare case when the optimal solution to a natural extremal problem involving parameter is independent of the value of the parameter.

Exercise 12.6.5 * *Let $\Sigma = [0, L]$ with some $L > 0$, and let $\gamma = 1/2$. Prove that the solution to (P) is given by*

$$1 - tp_N^*(t) = (-1)^N (2N + 3)\sqrt{L/t}T_{2N+3}(\sqrt{t/L}),$$

$T_k(t)$ being the Tschebyshev polynomial of the degree k.

Exercise 12.6.6 * *Prove the upper bound on $\nu^*(\cdot)$ given in Example 3 of Lecture 12.*

Lecture 13

Optimality of the Conjugate Gradient method

In the previous lecture we have established tight complexity bounds for the families of linear operator equations $\mathcal{U}_n(\Sigma, \tau, R)$. Recall that a family $\mathcal{U}_n(\Sigma, \tau, R)$ is comprised of all linear systems

$$Ax = b$$

with a given number n of variables and equations, such that

the spectrum of the symmetric positive semidefinite matrix A belongs to a given compact set $\Sigma \subset \mathbf{R}_+$;

equation $Ax = b$ is solvable,

and

the normal (i.e., of the minimal Euclidean norm) solution $x^*(A, b)$ to the equation admits "source representation"

$$x^*(A, b) = A^\tau u$$

with some u such that $|u| \leq R$.

We have equipped the aforementioned family of problems with the accuracy measure

$$\nu_\omega(x; A, b) = |A^\omega(x - x^*(A, b))|$$

and have proved that the corresponding complexity of the family (taken with respect to the methods based on matrix-vector multiplications) satisfies the bounds

$$\min\left\{\frac{n-1}{2}, \frac{N^*(\nu/R, \Sigma, \omega+\tau)}{2}\right\} \leq \mathrm{Compl}(\nu) \leq \min\left\{n, N^*(\nu/R, \Sigma, \omega+\tau)\right\}, \qquad (13.0.1)$$

where function $N^*(\varepsilon, \Sigma, \gamma)$ is defined as the smallest integer N such that there exists a polynomial $p(t)$ of degree N with the property

$$\max_{t\in\Sigma} t^\gamma |1 - tp(t)| \leq \varepsilon.$$

Here and in what follows we assume that 0 is not an isolated point of Σ, and that $\tau + \omega > 0$ in the case when $0 \in \Sigma$.

We have seen that upper complexity bound in (13.0.1) is associated with the Tschebyshev method

$$x_k = p_k^*(A)b$$

where p_k^* are the optimal solutions to the following best unform approximation problems

$\mathcal{P}_k$: *find polynomial* p *of degree* $\leq k$ *which minimizes the deviation* $\max_{t\in\Sigma} t^{\omega+\tau}|1-tp(t)|$.

As it was explained in the previous lecture, Tschebyshev's methods, in spite of their optimality, are not too attractive from the computational viewpoint, since to specify the corresponding polynomials p_k^*, one needs a lot of a priori information on the problem: he should know the localizer Σ of the spectrum of A, the parameters τ, R of source representation of the solution, etc.

The goal of this lecture is to demonstrate that there is a single and well-known method - the *Conjugate Gradient* - which behaves itself in the optimal way on *every* of the classes of linear operator equations with source-representable solutions with respect to *every* inaccuracy measure ν_ω.

13.1 The Conjugate Gradient method

Let me recall to you what is the Conjugate Gradient method (CGM). As applied to linear operator equation

$$Ax = b$$

with symmetric positive semidefinite matrix A, the method generates the sequence of iterates

$$x_k = x_k(A,b) \equiv p_k(A)b,\ k = 1,2,...$$

in such a way that x_k minimizes the quadratic potential

$$f(x) \equiv f_{A,b}(x) = x^TAx - 2b^Tx$$

of the equation over the $(k-1)$st Krylov subspace of the pair (A,b):

$$x_k \in \operatorname*{Argmin}_{x\in E_{k-1}(A,b)} f_{A,b}(x). \tag{13.1.1}$$

Recall that the Krylov subspaces of the pair (A,b) are given by

$$E_i(A,b) = \mathrm{Lin}\{b, Ab, ..., A^ib\} = \{p(A)b \mid \deg p \leq i\}.$$

The aforementioned description of the method says *what* the method does, but does not say *how* CGM does it. The standard description of CGM as an iterative routine is given by the recurrence

$$p_k = r_k - \frac{r_k^Tq_{k-1}}{p_{k-1}^Tq_{k-1}}p_{k-1} \tag{13.1.2}$$

$$q_k = Ar_k - \frac{r_k^Tq_{k-1}}{p_{k-1}^Tq_{k-1}}q_{k-1} \tag{13.1.3}$$

$$x_{k+1} = x_k - \frac{r_k^T p_k}{p_k^T q_k} p_k \tag{13.1.4}$$

$$r_{k+1} = r_k - \frac{r_k^T p_k}{p_k^T q_k} q_k \tag{13.1.5}$$

$$\tag{13.1.6}$$

initialized as

$$x_0 = p_0 = q_0 = 0;\ r_0 = -b.$$

In (13.1.2), fractions with zero denominators by definition are zeros.

Note that there are many equivalent descriptions of the CGM; I choose that one which stresses that at a step of the method we perform exactly one matrix-vector multiplication and add to this multiplication $O(n)$ arithmetic operations to update our iterates x_i, $r_i = Ax_i - b$, p_i and $q_i = Ap_i$.

Of course, from the presented iterative description of CGM it is not seen why the iterate x_k is a minimizer of the quadratic potential $f_{A,b}(\cdot)$ over the Krylov subspace $E_{k-1}(A, b)$; this is a theorem which can be found in any textbook where the CGM is presented. In what follows we never use the presentation of the CGM as an iterative routine; I mentioned it only to demonstrate that CGM indeed is a method based on matrix-vector multiplications, and that the arithmetic cost of a step in the method, modulo the matrix-vector multiplication, is $O(n)$.

Let me add that the Conjugate Gradient Method is in fact finite: it finds a solution to any solvable linear operator equation with n variables in no more than n steps. This property of the method also is proved in any Linear Algebra textbook; in fact it is a more or less immediate consequence of the "non-iterative" description of the method, see exercises accompanying this lecture.

13.2 Main result

My goal is to establish the following

Theorem 13.2.1 *Consider the family of problems $\mathcal{U}_n(\Sigma, \tau, R)$ equipped with the accuracy measure $\nu_\omega(x; A, b)$. Assume, as usual, that 0 is not an isolated point of Σ, and, besides this, that*

$$0 < \omega + \tau \tag{13.2.1}$$

and

$$\omega \le 1/2. \tag{13.2.2}$$

Under these assumptions, the number of steps $N_{CG}(\nu)$ in which the Conjugate Gradient Method solves any problem instance from the family within accuracy ν satisfies the estimate

$$N_{CG}(\nu) \le \min\left\{n; \frac{1+2\tau}{2(\omega+\tau)}\left[N^*(\nu/R, \Sigma, \omega+\tau) + 1\right]\right\}. \tag{13.2.3}$$

Let me make some comments.

1) In view of complexity bounds (13.0.1) the theorem demonstrates that the complexity of the CGM, up to the factor $2\rfloor\frac{1+2\tau}{2(\omega+\tau)}\lfloor$, coincides with the complexity of the class of problems in

question. This is a very remarkable property, since to run the method, you need no information on the particular class the problem belongs to. We see that the CGM possesses very strong "adaptation abilities".

2) The theorem states that the CGM demonstrates the aforementioned "adaptive abilities" in certain restricted range of values of τ and ω. The restrictions defining the "admissible" values of these parameters are formed by the inequality

$$\omega + \tau > 0$$

and the inequality

$$\omega \leq 1/2.$$

The first of these inequalities is crucial, provided that $0 \in \Sigma$; in this latter case this is our standard "well-posedness" assumption commented in the previous lecture. In the "non-degenerate" case $0 \notin \Sigma$ the restriction $\omega + \tau > 0$ can in fact be eliminated at the cost of slight modification of the conclusion.

As far as the restriction $\omega \leq 1/2$ is concerned, it relates to the particular version of the Conjugate Gradient we deal with. When passing from the CGM to a slightly more general method CGM_m which, as applied to a problem instance (A, b), generates the sequence of approximate solutions x_k according to the rule

$$x_k \in \operatorname*{Argmin}_{x \in E_{k-1}(A,b)} \{x^T A^m x - 2b^T A^{m-1} x\}$$

($m > 0$ is integer parameter of the method), one has a complexity bound completely similar to (13.2.3), but now in the larger range of values of ω and τ, namely, that one given by the relations

$$\tau + \omega > 0, \ \omega \leq m/2.$$

13.3 Proof of the main result

13.3.1 CGM and orthogonal polynomials

From the initial - "non-iterative" - description of CGM it follows that the method as applied to a problem instance (A, b), generates the sequence of iterates

$$x_k = p_k(A)b,$$

p_k being certain polynomial of degree not exceeding $k - 1$. Let us realize what are the polynomials. Assume that $(A, b) \in \mathcal{U}_n(\Sigma, \tau, R)$, and let

$$x^* = A^\tau u, \ |u| \leq R, \tag{13.3.1}$$

be the source representation of the normal solution to the equation. Since A is symmetric, its image space is orthogonal to the kernel of A, and when we project a point x onto the image space of A, the vector Ax remains unchanged, since the projection varies only the component of x along the kernel of A. Thus, the projection of a solution to the equation onto the image space of A also is a solution; in particular, the projection of the normal solution x^* onto the image

space of A is a solution; since the projection decreases the norm of vector whenever it varies it and x^* is the solution of the minimal norm, we conclude that the normal solution belongs to the image space of A. Now, looking at the source representation (13.3.1) of x^*, we see that we always can replace in this representation u by its projection u' onto the image space of A (indeed, if $\tau = 0$, then $u = x^*$ coincides with its projection onto the image space of A, and if $\tau \neq 0$, then, by definition of the operator A^τ, $A^\tau u = A^\tau u'$, and of course $|u'| \leq |u| \leq R$). Thus, we may assume that u belongs to the image space of A, so that

$$u = \sum_{i=1}^{m} u_i e_i, \tag{13.3.2}$$

where e_i, $i = 1, ..., m$, are mutually orthogonal eigenvectors of A with positive eigenvalues λ_i. Consequently,

$$x^* = A^\tau u = \sum_{i=1}^{m} \lambda_i^\tau u_i e_i, \tag{13.3.3}$$

$$b = Ax^* = \sum_{i=1}^{m} \lambda_i^{1+\tau} u_i e_i \tag{13.3.4}$$

and for any polynomial p

$$p(A)b = \sum_{i=1}^{m} \lambda_i^{1+\tau} p(\lambda_i) u_i e_i, \tag{13.3.5}$$

$$Ap(A)b = \sum_{i=1}^{m} \lambda_i^{1+\tau} \lambda_i p(\lambda_i) u_i e_i. \tag{13.3.6}$$

Now, from the initial description of CGM it follows that kth approximate solution found by the method is the minimizer of $f_{A,b}(x)$ over $x = p(A)b$, where p runs over the space P_{k-1} of polynomials of degree $\leq k-1$. Let us look what is $f_{A,b}(p(A)b)$. We have

$$f_{A,b}(p(A)b) = [p(A)b]^T [Ap(A)b] - 2b^T [p(A)b] =$$

[see (13.3.4), (13.3.5) and (13.3.6) and take into account that $\{e_i\}$ form an orthonormal system]

$$= \sum_{i=1}^{m} \left[\lambda_i^{3+2\tau} p^2(\lambda_i) - 2\lambda_i^{2+2\tau} p(\lambda_i) \right] u_i^2 =$$

$$= \sum_{i=1}^{m} \lambda_i^{1+2\tau} \left(1 - \lambda_i p(\lambda_i)\right)^2 u_i^2 - \sum_{i=1}^{m} \lambda_i^{1+2\tau} u_i^2.$$

The second sum here is independent of p; as it is immediately seen, this sum, together with the sign -, is nothing but the optimal value of $f_{A,b}$:

$$f^* \equiv \min_x f_{A,b}(x) = -\sum_{i=1}^{m} \lambda_i^{1+2\tau} u_i^2.$$

Thus, we come to the relation

$$d_{A,b}(p(A)b) \equiv f_{A,b}(p(A)b) - \min_x f_{A,b}(x) = \sum_{i=1}^{m} \lambda_i^{1+2\tau} (1 - \lambda_i p(\lambda_i))^2 u_i^2. \tag{13.3.7}$$

Note that $d_{A,b}$ also is a quadratic potential of our equation given by the additional requirement that the minimal value of the potential is 0.

To simplify notation, let us denote by μ the discrete measure on Σ which is concentrated on the set $\{\lambda_i\}_{i=1}^m$ and relates to the point λ_i the mass u_i^2; then (13.3.7) can be rewritten as

$$d_{A,b}(p(A)b) = \int_\Sigma t^{1+2\tau}(1-tp(t))^2 d\mu(t). \tag{13.3.8}$$

Note that the measure $\mu(\Sigma)$ of the whole Σ is nothing but $\sum_i u_i^2 = |u|^2 \le R^2$.

Now, as we remember, the iterates $x_k = p_k(A)b$ of the CGM correspond to the polynomials p_k which minimize the potential $f_{A,b}$, or, which is the same, the potential $d_{A,b}$ over all x's of the form $p(A)b$, p being a polynomial of degree $< k$. Combining this description of the method and (13.3.8), we come to the conclusion as follows:

Proposition 13.3.1 *Assume that a problem $(A,b) \in \mathcal{U}_n(\Sigma,\tau,R)$ is solved by the Conjugate Gradient Method, and let*

$$x_k = p_k(A)b$$

be the iterates generated by the method. Then there exists a measure μ concentrated on a finite subset of the set $\Sigma\backslash\{0\}$ such that

$$\int_\Sigma d\mu(t) \le R^2 \tag{13.3.9}$$

and p_k, for every k, is an optimal solution to the following quadratic problem

$$\text{minimize } \int_\Sigma t^{1+2\tau}(1-tp(t))^2 d\mu(t) \text{ s.t. } p \in P_{k-1}, \tag{13.3.10}$$

P_{k-1} being the space of all polynomials of degree $< k$.

As an immediate consequence of this observation, we come to

Corollary 13.3.1 *The polynomials*

$$q_k(t) = 1 - tp_k(t)$$

for every k are optimal solutions to the quadratic optimization problem

$$\mathcal{Q}_k\text{: minimize } \int_\Sigma t^{1+2\tau} q^2(t) d\mu(t) \text{ s.t. } q \in P_k^0, \tag{13.3.11}$$

P_k^0 being the space of all polynomials of degree $\le k$ equal to 1 at the point $t = 0$.

The polynomials q_k satisfy the orthogonality conditions

$$\int_\Sigma t^{2+2\tau} p(t) q_k(t) d\mu(t) = 0, \ \deg p < k. \tag{13.3.12}$$

The proof is immediate: to say that $p(t) \in P_{k-1}$ is the same as to say that $q(t) = 1 - tp(t)$ is in P_k^0; (13.3.12) is exactly the optimality condition for the quadratic optimization problem (13.3.10).

Corollary claims that $\{q_k\}$ is the system of orthogonal polynomials with respect to the measure $d\zeta(t) = t^{2+2\tau} d\mu(t)$ on Σ; these polynomials are normalized by the conditions

$$\deg q_k \le k; \ q_k(0) = 1.$$

13.3.2 Expression for inaccuracy

After we have found certain "functional-analytic" representation of the trajectory of the CGM, let us find similar representation for the inaccuracy

$$\nu_\omega(k) \equiv \nu_\omega(p_k(A)b; A, b)$$

of k-th approximate solution found by the method. We have

$$x^* = \sum_{i=1}^{k} \lambda_i^\tau u_i e_i,$$

$$p_k(A)b = \sum_{i=1}^{k} \lambda_i^{1+\tau} p_k(\lambda_i) u_i e_i,$$

whence

$$A^\omega(x_k - x^*) = A^\omega[p_k(A)b - x^*] = \sum_{i=1}^{m} \lambda_i^\omega \lambda_i^\tau (\lambda_i p_k(\lambda_i) - 1) u_i e_i$$

and

$$\nu_\omega^2(k) \equiv |A^\omega(x_k - x^*)|^2 = \sum_{i=1}^{m} \lambda_i^{2\omega+2\tau} (1 - \lambda_i p_k(\lambda_i))^2 u_i^2,$$

or, which is the same,

$$\nu_\omega^2(k) = \int_\Sigma t^{2(\omega+\tau)} q_k^2(t) d\mu(t). \tag{13.3.13}$$

13.3.3 Momentum inequality

We find ourselves in the situation as follows. We know something about the polynomials q_k, namely, we know that these polynomials are optimal solutions to the quadratic optimization problems $\mathcal{Q}_k$; and our goal is to derive from this observation something reasonable about the quantities $\nu_\omega(k)$ given by (13.3.13). To this end we intend to use the following general statement

Proposition 13.3.2 [The Momentum Inequality] *Let ξ be a measure with the compact support set belonging to the positive ray. Then the function*

$$\mathrm{Mom}_\xi(s) = \int t^s d\xi(t)$$

is logarithmically convex in s: for any triple $s \le s' \le s''$ with $s'' > s$ one has

$$\mathrm{Mom}_\xi(s') \le [\mathrm{Mom}_\xi(s)]^{\frac{s''-s'}{s''-s}} [\mathrm{Mom}_\xi(s'')]^{\frac{s'-s}{s''-s}}. \tag{13.3.14}$$

This proposition is nothing but a reformulation of the Holder inequality (see the exercises accompanying this lecture).

Let us fix k and take as ξ the measure

$$d\xi(t) = q_k^2(t) d\mu(t);$$

since μ, in view of its origin, is a discrete measure concentrated on a finite subset of the positive ray, ξ fits the requirements of the Momentum Inequality. Now let us set

$$s = 0,\ s' = 2(\omega + \tau),\ s'' = 1 + 2\tau;$$

under the premise of the theorem we are proving, we have

$$0 < 2(\omega + \tau) \leq 1 + 2\tau \tag{13.3.15}$$

(see (13.2.1) and (13.2.2)), so that these quantities satisfy the requirements $s \leq s' \leq s''$, $s'' > s$. With this choice of s, s', s'' we have

$$\mathrm{Mom}_\xi(s) \equiv \mathrm{Mom}_\xi(0) = \int_\Sigma q_k^2(t) d\mu(t); \tag{13.3.16}$$

$$\mathrm{Mom}_\xi(s') \equiv \mathrm{Mom}_\xi(2(\omega + \tau)) = \int_\Sigma t^{2(\omega+\tau)} q_k^2(t) d\mu(t) \equiv \nu_\omega^2(k) \tag{13.3.17}$$

(see (13.3.13)) and

$$\mathrm{Mom}_\xi(s'') \equiv \mathrm{Mom}_\xi(1 + 2\tau) = \int_\Sigma t^{1+2\tau} q_k^2(t) d\mu(t). \tag{13.3.18}$$

Applying the Momentum Inequality, we come to certain upper bound on the quantity $\mathrm{Mom}_\xi(2(\omega + \tau))$ we are interested in:

$$\nu_\omega^2(k) = \mathrm{Mom}_\xi(2(\tau + \omega)) \leq [\mathrm{Mom}_\xi(0)]^\alpha [\mathrm{Mom}_\xi(1 + 2\tau)]^\beta =$$

$$= [\mathrm{Mom}_\xi(0)]^\alpha \left\{ \int_\Sigma t^{1+2\tau} q_k^2(t) d\mu(t) \right\}^\beta,$$

$$\alpha = \frac{1 - 2\omega}{1 + 2\tau},\ \beta = 1 - \alpha = \frac{2(\omega + \tau)}{1 + 2\tau}. \tag{13.3.19}$$

Our further actions will be as follows. Frits, we shall prove that

$$\mathrm{Mom}_\xi(0) \leq \int_\Sigma d\mu(t) \quad (\leq R^2), \tag{13.3.20}$$

which is the central point of the whole reasoning. Second, we shall establish certain good upper bound on $\mathrm{Mom}_\xi(1 + 2\tau)$, systematically using the fact that this quantity, by construction of the CGM, is the optimal value in problem $\mathcal{Q}_k$ (see Corollary 13.3.1).

13.3.4 Proof of (13.3.20)

It is possible that $q_k = 0$ at the support set of μ; in this case the left hand side of (13.3.20), in view of (13.3.16), is 0, and (13.3.20) is, of course, valid. Now assume that q_k is not identically zero at the support set of the measure μ. Then this set (which, by construction, is a finite subset of the positive ray) is comprised of at least $k + 1$ points with positive μ-masses (if the number of these points would be less than $k + 1$, we could find a polynomial $q \in P_k^0$ vanishing at these points and the optimal value in $\mathcal{Q}_k$ would be zero, and we just have assumed that it is not the

case). Corollary 13.3.1 says that $q_1, ..., q_k$ form an orthogonal system of polynomials of degrees not exceeding $1, 2, ..., k$, respectively, in the space $L_2(\zeta)$,

$$d\zeta(t) = t^{2+2\tau} d\mu(t).$$

In the case in question the support set of the measure ζ (which coincides with the support set of the measure μ) is comprised of at least $k+1$ points, and from the standard properties of orthogonal polynomials (see exercises accompanying this lecture) it follows that q_k is a polynomial of the degree exactly k with all roots $t_1, ..., t_k$ being positive and distinct. Without loss of generality we may assume that $t_1 < ... < t_k$. Since $q_k(0) = 1$, we have

$$q_k(t) = \prod_{i=1}^{k} (1 - t/t_i).$$

Now let

$$p(t) = \prod_{i=2}^{k} (1 - t/t_i) = q(t)(1 - t/t_1)^{-1}.$$

In view of the orthogonality relation (13.3.12) we have

$$\int_\Sigma t^{2+2\tau} p(t) q_k(t) d\mu(t) = 0,$$

which, in view of

$$q_k(t) = p(t)(1 - t/t_1),$$

can be rewritten as

$$\int_0^{t_1} t^{2+2\tau} p^2(t)(1 - t/t_1) d\mu(t) = \int_{t_1}^{\infty} t^{2+2\tau} p^2(t)(t/t_1 - 1) d\mu(t). \tag{13.3.21}$$

We have

$$\mathrm{Mom}_\xi(0) = \int_0^\infty q_k^2(t) d\mu(t) = \int_0^\infty p^2(t)(1 - t/t_1)^2 d\mu(t) =$$

$$= \int_0^{t_1} p^2(t)(1 - t/t_1)^2 d\mu(t) + \int_{t_1}^\infty p^2(t)(1 - t/t_1)^2 d\mu(t) \leq$$

[since $(t/t_1 - 1)^2 \leq t_1^{-2-2\tau} t^{2+2\tau} [t/t_1 - 1]$ whenever $t \geq t_1$ in view of $1 + 2\tau > 0$, see (13.3.15)]

$$\leq \int_0^{t_1} p^2(t)(1 - t/t_1)^2 d\mu(t) + t_1^{-2-2\tau} \int_{t_1}^\infty t^{2+2\tau} p^2(t)(t/t_1 - 1) d\mu(t) =$$

[see (13.3.21)]

$$= \int_0^{t_1} p^2(t)(1 - t/t_1) \left[(1 - t/t_1) + t^{2+2\tau} t_1^{-2-2\tau} \right] d\mu(t) \leq$$

[since $0 \leq 1 - t/t_1 \leq 1$ and $0 \leq 1 - t/t_1 + t^{2+2\tau} t_1^{-2-2\tau} \leq 1$, $0 \leq t \leq t_1$; recall that $1 + 2\tau > 0$]

$$\leq \int_0^{t_1} p^2(t) d\mu(t) \leq$$

[since $0 \leq p(t) \leq 1$ as $0 \leq t \leq t_1$ by construction of $p(t)$]

$$\leq \int_0^{t_1} d\mu(t),$$

as required in (13.3.20).

13.3.5 Concluding the proof of Theorem 13.2.1

Now we are enough equipped to proof Theorem 13.2.1. Let us fix $\nu > 0$. We should prove that the CGM with

$$k = \min\{n; \kappa[N^*(\nu/R, \Sigma, \tau + \omega) + 1]\},\ \kappa = \rfloor \frac{1 + 2\tau}{2(\omega + \tau)} \lfloor, \tag{13.3.22}$$

steps solves every problem (A, b) from the family $\mathcal{U}_n(\Sigma, \tau, R)$ within inaccuracy ν. Without loss of generality we may assume that the minimum in the right hand side of (13.3.22) corresponds to the second quantity in the brackets, since otherwise the statement in question is evident: the Conjugate Gradient Method with n steps finds, as we know, exact solution to any solvable linear equation. Thus, in what follows we may assume that k given by (13.3.22) is nothing but

$$k = \kappa(N + 1),\ N = N^*(\nu/R, \Sigma, \omega + \tau). \tag{13.3.23}$$

Let (A, b) be a problem from $\mathcal{U}_n(\Sigma, \tau, R)$ and $\{x_s\}$ be the trajectory of the CGM as applied to this problem and $\mu(\cdot)$ be the corresponding measure given by Proposition 13.3.1. In view of (13.3.19) and (13.3.20) we have

$$\nu_\omega^2(x_s; A, b) \le R^{2\alpha} \left\{ \int_\Sigma t^{1+2\tau} q_k^2(t) d\mu(t) \right\}^\beta,\ \beta = 1 - \alpha = \frac{2(\omega + \tau)}{1 + 2\tau}.$$

Thus, to prove the statement in question it suffices to demonstrate that for k given by (13.3.23) one has

$$R^{2\alpha} \left\{ \int_\Sigma t^{1+2\tau} q_k^2(t) d\mu(t) \right\}^\beta \equiv R^{2\alpha} \left\{ \min_{q \in P_k^0} \int_\Sigma t^{1+\tau} q^2(t) d\mu(t) \right\}^\beta \le \nu^2 \tag{13.3.24}$$

(the equivalence in this relation follows from Corollary 13.3.1).

The proof of (13.3.24) is as follows.

By definition of N there exists a polynomial $q_* \in P_{N+1}^0$ such that

$$\nu/R \ge \varepsilon \equiv \min_{q \in P_{N+1}^0} \max_{t \in \Sigma} t^{\tau+\omega} |q(t)| = \max_{t \in \Sigma} t^{\tau+\omega} |q_*(t)|. \tag{13.3.25}$$

It is possible, first, that $q_* = 0$ μ-almost everywhere, or, which is the same, that $\varepsilon = 0$. Since $q_* \in P_{N+1}^0 \subset P_k^0$, the left hand side in (13.3.24) in the case in question is 0, and (13.3.24) is, of course, true.

Now assume that $\varepsilon > 0$. From the standard Tschebyshev-type results (see Exercise 12.6.2) it follows that there exist $N + 2$ points

$$\rho \equiv r_0 < r_1 < ... < r_{N+1}$$

in Σ - the points of Tschebyshev alternance - such that

$$r_i^{\tau+\omega} q_*(r_i) = (-1)^i \varepsilon. \tag{13.3.26}$$

From this "oscillation" property it follows, in turn, that q_* has $N + 1$ distinct positive roots

$$t_1 < ... < t_{N+1},$$

which separate the points $r_0, ..., r_{N+1}$; in particular,

$$\rho < t_1. \tag{13.3.27}$$

Since q_* is a polynomial of degree at most $N+1$ with the roots $t_1, ..., t_{N+1}$ and $q_*(0) = 1$, we have

$$q_*(t) = \prod_{i=1}^{N+1} (1 - t/t_i). \tag{13.3.28}$$

Now let us set

$$q(t) = q_*^\kappa(t).$$

We have $q(0) = 1$ and $\deg q = \kappa(N+1) = k$, so that $q \in P_k^0$. Thus, q belongs to the set over which the minimum in (13.3.24) is taken; therefore to prove (13.3.24) it suffices to demonstrate that

$$\max_{t\in\Sigma} t^{1+2\tau} q^2(t) \le \varepsilon^{\frac{1+2\tau}{(\omega+\tau)}} \equiv \varepsilon^{2/\beta}. \tag{13.3.29}$$

Indeed, this inequality combined with (13.3.9) would result in

$$R^{2\alpha}\{\int_\Sigma t^{1+2\tau} q^2(t) d\mu(t)\}^\beta \le R^{2\alpha}\{\varepsilon^{2/\beta} R^2\}^\beta = R^2\varepsilon^2 \le \nu^2$$

(the concluding inequality follows from (13.3.25)), as required in (13.3.24).

To prove (13.3.29), let us start with the observation that

$$\rho \ge \varepsilon^{1/(\omega+\tau)}. \tag{13.3.30}$$

Indeed, (13.3.26) says that $\rho = (\varepsilon/q_*(\rho))^{1/(\omega+\tau)}$, while (13.3.28) and (13.3.27) imply that

$$0 \le t \le \rho \Rightarrow 0 < q_*(t) < 1. \tag{13.3.31}$$

Now let $t \in \Sigma$. Assume, first, that $t \ge \rho$. From (13.3.25) it follows that

$$|q_*(t)| \le \varepsilon t^{-\omega-\tau},$$

whence

$$t^{1+2\tau} q^2(t) = t^{1+2\tau} q_*^{2\kappa}(t) \le t^{1+2\tau-2\kappa(\omega+\tau)} \varepsilon^{2\kappa} \le$$

[since $1 + 2\tau - 2\kappa(\omega+\tau) \le 0$ by definition of κ and $t \ge \rho$ by assumption on t]

$$\le \rho^{1+2\tau-2\kappa(\omega+\tau)} \varepsilon^{2\kappa} \le$$

[again since $1 + 2\tau - 2\kappa(\omega+\tau) \le 0$ and in view of (13.3.30)]

$$\le \varepsilon^{(1+2\tau-2\kappa(\omega+\tau))/(\omega+\tau)} \varepsilon^{2\kappa} = \varepsilon^{(1+2\tau)/(\omega+\tau)}.$$

Thus,

$$\{t \in \Sigma,\, t \ge \rho\} \Rightarrow t^{1+2\tau} q^2(t) \le \varepsilon^{(1+2\tau)/(\omega+\tau)}. \tag{13.3.32}$$

Now let $t \in \Sigma$ and $t \le \rho$. From (13.3.31) and (13.3.25) it follows that

$$0 < q_*(t) \le \min\{1; \varepsilon t^{-\omega-\tau}\},$$

whence

$$t^{1+2\tau}q^2(t) \le t^{1+2\tau}\min\{1;\varepsilon^{2\kappa}t^{-2\kappa(\omega+\tau}\} =$$

$$= \min\{t^{1+2\tau};\varepsilon^{2\kappa}t^{1+2\tau-2\kappa(\omega+\tau)}\} \le \max_{s>0}\min\{s^{1+2\tau};\varepsilon^{2\kappa}s^{1+2\tau-2\kappa(\omega+\tau)}\} =$$

[recall that $1+2\tau>0$, $1+2\tau-2(\omega+\tau)\le 0$]

$$= \varepsilon^{\frac{1+2\tau}{\omega+\tau}},$$

and we come to the relation

$$\{t\in\Sigma, t\le\rho\} \Rightarrow t^{1+2\tau}q^2(t) \le \varepsilon^{\frac{1+2\tau}{\omega+\tau}}, \tag{13.3.33}$$

which combined with (13.3.32) proves (13.3.29). ∎

13.4 Exercises: Around Conjugate Gradient Method

Let us start with a series of simple exercises which state some useful facts about the Conjugate Gradient Method as applied to a linear operator equation

$$Ax = b$$

with symmetric $n\times n$ positive semidefinite matrix A. In what follows the equation is assumed to be solvable; we exclude the trivial case $b=0$.

Exercise 13.4.1 $^{\#+}$ *Let $e_1,...,e_m$ be the minimal, with respect to the number of elements, collection of eigenvectors of A such that b can be represented as a linear combination of the vectors from the collection. Prove that the corresponding eigenvalues λ_i, $i=1,...,m$, are distinct and positive and that m is exactly the number of steps in which the CGM finds the exact solution to the equation. In particular, the number of steps before the exact solution is found never exceeds the maximal number of different positive eigenvalues of A (and, consequently, never exceeds n).*

Exercise 13.4.2 $^{\#}$ *Assume that Σ is a compact subset of the nonnegative half-ray, 0 is not an isolated point of Σ and ω and τ are such that $\omega+\tau>0$ and $\omega\le 1/2$, as required by the premise of the Theorem on optimality of the CGM (Lecture 13). Assume that problem (A,b) belongs to the family $\mathcal{U}_n(\Sigma^*,\tau,R)$, where Σ^* is the union of Σ and a set Σ' comprised of at most k points which are to the right of Σ (note this "to the right"!). Prove that for any $\nu>0$ the CGM solves (A,b) within ν_ω-inaccuracy $\le\nu$ in no more than*

$$k+\rfloor\frac{1+2\tau}{2(\omega+\tau)}\lfloor[N^*(\mu/R,\Sigma,\omega+\tau)+1]$$

steps (compare with the statement of Theorem 13.2.1). Thus, everything looks as if the CGM spends k first steps "to kill" the eigenvalues of A outside Σ and then switches to solving the problem with the "truncated" spectrum.

Observation given by Exercise 13.4.1 underlies an important trick in solving linear systems: the *prescaled conjugate gradient scheme*. Assume we are interested in solving the "perturbed" equation

$$Ax = b,\ A = A_0 + A_1,$$

where the "perturbation" A_1 is symmetric and the "unperturbed" matrix A_0 is symmetric positive definite and is "simpler" than A (in the sense that we can invert A_0 significantly cheaper than A).

Let us act as follows. Imagine that we have passed from the initial system to the equivalent one by substitution $x = A_0^{-1/2}u$ and termwise multiplication of the system by $A_0^{-1/2}$. The resulting system with respect to u would be

$$(I + B_1)u = b_1,\ B_1 = A_0^{-1/2}A_1A_0^{-1/2}, b_1 = A_0^{-1/2}b.$$

Let u_k be the trajectory of the usual CGM as applied to the transformed system and $x_k = A_0^{-1/2}u_k$ be the image of this trajectory in the x-variables. This latter sequence is, as it is easily seen, governed by a simple recurrence which does not involve B_1, b_1 or $A_0^{1/2}$; each step of this recurrence requires nothing but a single multiplication of a vector by A and a single multiplication of another vector by A_0^{-1}, plus $O(n)$ arithmetic operations. This recurrence is called the Conjugate Gradient Method *prescaled* by the matrix A_0. Note that the potentials

$$f_{A,b}(x)$$

of the initial system and

$$f_{A',b'}(u) \quad [A' = A_0^{-1/2}AA_0^{-1/2} = I + B_1,\ b' = A_0^{-1/2}b]$$

of the transformed one are related to each other:

$$f_{A,b}(A_0^{-1/2}u) = f_{A',b'}(u),$$

so that the rates of convergence of both the sequences $\{x_k\}$ and $\{u_k\}$ to the solutions of, respectively, the initial and the transformed problems *measured in terms of the residuals in the corresponding potentials* are identical to each other.

The recurrent form of the CGM prescaled by A_0 is

$$p_k = A_0^{-1}r_k - \frac{q_{k-1}^T A_0^{-1}r_k}{p_{k-1}^T q_{k-1}}p_{k-1} \tag{13.4.1}$$

$$q_k = AA_0^{-1}r_k - \frac{q_{k-1}^T A_0^{-1}r_k}{p_{k-1}^T q_{k-1}}q_{k-1} \tag{13.4.2}$$

$$x_{k-1} = x_k - \frac{p_{k-1}^T r_k}{p_k^T q_k}p_k \tag{13.4.3}$$

$$r_{k+1} = r_k - \frac{p_{k-1}^T r_k}{p_k^T q_k}q_k \tag{13.4.4}$$

(a fraction with zero denominator by definition is zero) initialized as

$$x_0 = p_0 = q_0 = 0,\ r_0 = -b.$$

Exercise 13.4.3 # *Prove that the CGM prescaled by a positive definite symmetric matrix A_0 indeed is given by the above recurrence. What is the CGM prescaled by the unit matrix?*

Now, are there any advantages of the prescaled CGM as compared to the standard one? Of course, we may expect something reasonable only when the perturbation A_1 is "small". The first - and the standard - interpretation of the notion of a "small" perturbation is that one when $r \equiv \mathrm{Rank}\, A_1$ is relatively small ($r << n$). A good example here is an "almost" block-diagonal matrix A, i.e., the matrix of the form of the form

$$\begin{pmatrix} q_1 & p_2^T & p_3^T & ... & p_k^T \\ p_2 & q_2 & & & \\ p_3 & & q_3 & & \\ ... & ... & ... & ... & ... \\ p_k & & & & q_k \end{pmatrix}$$

(blank space corresponds to zero blocks) which is nothing but a block-diagonal matrix A_0 with the diagonal blocks $q_1, ..., q_k$ "spoiled" by several first dense rows and columns; if l is the number of these dense rows (the row size of q_1), then one can decompose A as $A_0 + A_1$ with $\mathrm{Rank}\, A_1 \leq 2l$. In the case in question the matrix $I + B_1$ of the transformed system differs from the unit matrix by the matrix $B_1 = A_0^{-1/2} A_1 A_0^{-1/2}$ of the rank r and therefore $I + B_1$ has at most r eigenvalues which differ from 1. It follows that $I + B_1$ has at most $r+1$ different eigenvalues and therefore the standard CGM as applied to the transformed system solves it *exactly* in no more than $r+1 << n$ steps; since the trajectory of the prescaled CGM reproduces the same behaviour in x-variables, the prescaled CGM finds the exact solution to the initial system in the same $r + 1 << n$ steps.

The discussed notion of a "small" perturbation A_1 is, in a sense, structural: the entries of A_1 as compared to those of A_0 should not be small, what is small is the rank of the perturbation as compared to that one of A_0. There is another way to define a "relatively small" perturbation, namely, to say that

$$(\rho^{-1} - 1)A_0 \geq A_1 \geq (\rho - 1)A_0$$

for some $\rho \in (0, 1)$ ($\geq$ is understood in the operator sense, i.e., as positive semidefiniteness of the corresponding differences). Here, of course, the rank of A_1 should not be small (it may coincide with n), but the quadratic form $x^T Ax$ associated with the perturbed matrix A is "compatible" with that one associated with the unperturbed matrix A_0:

$$\rho \leq \frac{x^T Ax}{x^T A_0 x} \leq \rho^{-1} \ \forall x \neq 0. \tag{13.4.5}$$

Exercise 13.4.4 * *Prove that in the case of (13.4.5) the CGM prescaled by A_0 solves the system $Ax = b$ within the relative inaccuracy in terms of the potential*

$$\frac{f_{A,b}(x) - \min f_{A,b}}{f_{A,b}(0) - \min f_{A,b}}$$

not exceeding $\nu \in (0, 1/2)$ in no more than

$$N = O(1)\rho^{-1} \ln(1/\nu)$$

steps, $O(1)$ being an absolute constant.

To comment the result of the latter exercise from the computational viewpoint, let me note that, given a unique "unstructured" linear system, you hardly can decompose the matrix A of the system in an "unperturbed" part A_0 with known or easily computable inverse and a perturbation A_1 which satisfies (13.4.5). In some applications, anyhow, e.g., in the Newton-type methods for smooth minimization, one is required to solve *a series* of linear systems with matrices "slightly" varying from system to system. In this case one can try the strategy as follows. The first system $A^{(1)}x = b^{(1)}$ is solved via computing the matrix $[A^{(1)}]^{-1}$ by any standard direct method of Linear Algebra. When solving the second system, $A^{(2)}x = b^{(2)}$, we try first the CGM prescaled by the matrix $A^{(1)}$. If the method is able to solve the system at a "reasonable" arithmetic cost (here "reasonable" means, say, 10% of the CPU time spent to invert $A^{(1)}$), we pass to the next system; otherwise we stop the CGM and solve the second system by direct computation of $[A^{(2)}]^{-1}$. When solving the third system, we again start with the CGM prescaled by the last matrix for which we know the exact inverse (i.e., prescaled by $A^{(1)}$, if the CGM was successful when solving the second system, or by $A^{(2)}$, if it failed); if the method was successful, we go to the next system, if not, solve the third system via direct computation of $[A^{(3)}]^{-1}$ and use $A^{(3)}$ as our new prescaling matrix for the fourth system, an so on. It is clear that even in the worst case (no one of our attempts to save time by running the CGM is successful) this strategy increases the total CPU time, as compared to that one without trying the CGM, by at most 10%; at the same time if the matrices of the subsequent systems indeed are close to each other and, consequently, the runs of the CGM will "normally" be successful (say, will solve 90% of our systems in 10% of the time required by the direct Linear Algebra), we will reduce the total CPU time by factor of 5.

The remaining exercises deal with some auxiliary facts used in Lecture 13 without proof. We start with the Momentum Inequality:

Exercise 13.4.5 $^{\#*}$ *Prove Proposition 13.3.2 (Lecture 13).*

Now let us prove the (quite standard) property of the roots of orthogonal polynomials. Let μ be a (finite) measure with the support set contained in some segment $[a, b]$ of the real axis; for the sake of simplicity assume that $-\infty < a < b < \infty$, and let the number of points in the support set of μ be at least $N + 1$. Those not too familiar with the Measure Theory may assume, in addition, that the measure is discrete, i.e., there are $M > N$ points x_i of the segment $[a, b]$ to which the measure assignees positive weights μ_i, and for any $A \subset \mathbf{R}$ the measure $\mu(A)$ of the set is simply the total weight of the points which are in A. Recall that in the latter case the integral $\int f(t)d\mu(t)$ is nothing but $\sum_{i=1}^{M} f(x_i)\mu_i$, and $L_2(\mu)$ is the Euclidean space comprised of the real-valued functions on the finite set $\{x_1, ..., x_M\}$ with the inner product

$$(f, g) = \int f(t)g(t)d\mu(t) \equiv \sum_{i=1}^{M} f(x_i)g(x_i)\mu_i.$$

Exercise 13.4.6 $^{\#*}$

1) Prove that there exists a $(N+1)$-term μ-orthogonal system of polynomials, i.e., a sequence of polynomials q_0,...,q_N of the degrees $0, 1, ..., N$, respectively, such that
a) q_j is a nonzero element $L_2(\mu)$;

b) the polynomials are μ-orthogonal to each other, i.e.,

$$(q_i, q_j) = 0,\ 0 \le i < j \le N.$$

2) Prove that the sequence $\{q_i\}$ is uniquely defined by the aforementioned properties, up to multiplication of its terms by nonzero reals.

3) Prove that any polynomial q of degree $k \le N$ can be uniquely represented as a linear combination of $q_0, ..., q_k$. Derive from this that q_k is μ-orthogonal to any polynomial p of degree $< k$, i.e., for such a p one has $(q_k, p) = 0$.

4) Prove that if p is a nonzero polynomial of degree $\le N$, then $(p, p) > 0$. Derive from this observation that the natural mapping of the space P_N of polynomials of degree $\le N$ into $L_2(\mu)$ (a polynomial p is mapped into itself regarded as an element of $L_2(\mu)$) is an embedding: different elements of P_N are mapped into different elements of $L_2(\mu)$. Is it possible to replace here N with $N + 1$?

Let us fix an $(N + 1)$-term μ-orthogonal sequence of polynomials $q_0, ..., q_N$. Let E_k, $k \le N$, be the linear span of the polynomials $q_0, ..., q_{k-1}$ regarded as elements of $L_2(\mu)$, so that E_k is a subspace of $L_2(\mu)$. Let also T be the operator in $L_2(\mu)$ which maps an element of $L_2(\mu)$ represented by a function $f(t)$ into the element represented by the function $tf(t)$. We denote by T_k the restriction of T onto E_k, i.e., the operator on E_k defined as follows: for $f \in E_k$ $T_k f$ is the orthogonal projection of the vector Tf onto E_k.

Exercise 13.4.7 $^{\#}$ *Let $1 \le k \le N$. Prove that*

1) the operator T and the operators T_k are symmetric:

$$(Tf, g) = (f, Tg) = \int tf(t)g(t)d\mu(t),\ f, g \in L_2(\mu);$$

$$(T_k f, g) = (f, T_k g) = (Tf, g),\ f, g \in E_k.$$

2)$^{+}$ the spectrum of the operator T_k coincides with the set of real roots of the polynomial q_k: if λ is an eigenvalue of T_k, then $q_k(\lambda) = 0$, and vice versa.

3)$^{+}$ prove that all eigenvalues of T_k belong to the interval (a, b) and are of multiplicity 1. Derive from this observation and 2) that all k roots of q_k are real, distinct and belong to the open interval (a, b). Prove that q_{k-1} and q_k have no common roots.

Exercise 13.4.8 $^{\#*}$ *Prove the Separation Property: if $t_{i,1} < t_{i,2} < ... < t_{i,i}$ are the roots of q_i (we know from the previous exercise that q_i has i distinct roots and that they belong to (a, b)), then the roots of q_{i+1} separate those of q_i:*

$$a < t_{i+1,1} < t_{i,1} < t_{i+1,2} < t_{i,2} < ... < t_{i+1,i} < t_{i,i} < t_{i+1,i+1} < b.$$

Lecture 14

Convex Stochastic Programming

In this final lecture we shall speak about a problem which is quite different from those considered so far – about *stochastic optimization*. A general *single-stage Stochastic Programming* program is as follows:

$$\textit{minimize}\ \ f_0(x) = \int_\Omega F_0(x,\omega)dP(\omega)\ \ \text{s.t.}\ \ x \in G,\ f_j(x) = \int_\Omega F_j(x,\omega)dP(\omega) \le 0,\ j = 1,...,m. \tag{14.0.1}$$

Here

- $F_j(x,\omega)$ are functions of the design vector $x \in \mathbf{R}^n$ and parameter $w \in \Omega$
- G is a subset in $\mathbf{R}^n$
- P is a probability distribution on the set Ω.

Thus, a stochastic program is a Mathematical Programming program where the objective and the constraints are expectations of certain functions depending both of the design vector x and random parameter ω.

The main source of programs of the indicated type is optimization in stochastic systems, like Queuing Networks, where the processes depend not only on the design parameters, like performances and numbers of serving devices of different types), but also on random factors. As a result, the characteristics of such a system (e.g., time of serving a client, cost of service, etc.) are random variables depending, as on parameters, on the design parameters of the system. It is reasonable to measure the quality of the system by the expected values of the indicated random variables (in the dynamic systems, we should speak about the steady-state expectations). These expected values have the form of the objective and the constraints in (14.0.1), so that to optimize a system in question over its design parameters is a program of the type (14.0.1).

Looking at stochastic program (14.0.1), you can immediately ask what are, if any, the specific features of these problems and why these problems need specific treatment. (14.0.1) is, finally, nothing but the usual Mathematical Programming problem; we know how to solve tractable (convex) Mathematical Programming programs, and to the moment we never took into account what is the origin of the objective and the constraints, are they given by explicit simple formulae, or are they integrals, as in (14.0.1), or solutions to differential equations, or whatever else. All what was important for us were the properties of the objective and the constraints – convexity, smoothness, etc., but not their origin.

This "indifference to the origin of the problem" indeed was the feature of our approach, but it was its weak point, not the strong one. In actual computations, the performance of an algorithm heavily depends not only on the quality of the method we apply to solve the problem, but also on the computational effort needed to provide the algorithm by the information on the problem instance – the job which in our model of optimization was the task of the oracle. We did not think how to help the oracle to solve his task and took care only of total number of oracle calls – we did our best to reduce this number. In the most general case, when we have no idea of what is the internal structure of the problem, this is the only possible approach. But the more we know about the structure of the problem, the more should we think of how to simplify the task of the oracle in order to reduce the overall computational expenses.

Stochastic Programming is an example when we know something about the structure of the problem in question. Namely, let us look at a typical stochastic system, like Queuing Network. Normally the functions $F_j(x,\omega)$ associated with the system are "algorithmically simple" – given x and ω, we can more or less easily compute the quantities $F_j(x,\omega)$ and even their derivatives with respect to x; to this end it suffices to create a simulation model of the system and run it at the given by ω realization of the random parameters (arrival and service times, etc.); even for a rather sophisticated system, a single simulation of this type is relatively fast. Thus, typically we have no difficulties with simulating *realizations* of the random quantities $F_j(x,\cdot)$. On the other hand, even for relatively simple systems it is, as a rule, impossible to compute the *expected values* of the indicated quantities in a "closed analytic form", and the only way to evaluate these expected values if to use a kind of the Monte-Carlo method: to run not a single, but many simulations, for a fixed value of the design parameters, and to take the empiric average of the observed random quantities as an estimate of their expected values. According to the well-known results on the rate of convergence of the Monte-Carlo method, to estimate the expected values within inaccuracy ϵ it requires $O(1/\epsilon^2)$ simulations, and this is just to get the estimate of the objective and the constraints of problem (14.0.1) at a single point! Now imagine that we are going to treat (14.0.1) as a usual "black-box represented" optimization problem and intend to imitate the usual first-order oracle for it via the aforementioned Monte-Carlo estimator. In order to get an ϵ-solution to the problem, we, even in good cases, need to estimate within accuracy $O(\epsilon)$ the objective and the constraints along the search points. It means that the method will require much more simulations than the aforementioned $O(1/\epsilon^2)$: this quantity should be multiplied by the information-based complexity of the optimization method we are going to use. As a result, the indicated approach in most of the cases results in inappropriately long computations.

An extremely surprising thing is that there exists another way to solve the problem. This way, under reasonable convexity assumptions, results in *overall number of* $O(1/\epsilon^2)$ *computations only* – as if there were no optimization at all and the only goal were to estimate the objective and the constraints at a given point. The subject of our today lecture is this other way – *Stochastic Approximation*.

To get a convenient framework for presenting Stochastic Approximation, it is worthy to modify a little the way we are looking at our problem. Assume that when solving it, we are allowed to generate a random sample ω_1, ω_2,... of "random factors" involved into the problem; the elements of the sample are assumed to be mutually independent and distributed according to P. Assume also that given x and ω, we are able to compute the values $F_j(x,\omega)$ and the

gradients $\nabla_x F_j(x,\omega)$ of the integrants in (14.0.1). Note that under mild regularity assumptions the differentiation with respect to x and taking expectation are interchangeable:

$$f_j(x) = \int_\Omega F_j(x,\omega)dP(\omega),\ \nabla f_j(x) = \int_\Omega \nabla_x F_j(x,\omega)dP(\omega). \tag{14.0.2}$$

It means that the situation is covered by the following model of an optimization method solving (14.0.1):

At a step i, we (the method) form i-th search point x_i and forward it to the oracle which we have in our disposal. The oracle returns the quantities

$$F_j(x_i,\omega_i),\ \nabla_x F_j(x_i,\omega_i),\ j = 0,...,m,$$

(in our previous interpretation it means that a single simulation of the stochastic system in question is performed), and this answer is the portion of information on the problem we get on the step in question. Using the information accumulated so far, we generate the new search point x_{i+1}, again forward it to the oracle, enrich our accumulated information by its answer, and so on.

The presented scheme is a very natural definition of a method based on *stochastic first order oracle* capable to provide the method *random unbiased* (see (14.0.2)) *estimates* of the values and the gradients of the objective and the constraints of (14.0.1). Note that the estimates are not only unbiased, but also form a kind of Markov chain: the distribution of the answers of the oracle at a point depends only on the point, not on the previous answers (recall that $\{\omega_i\}$ are assumed to be independent).

Now, for our further considerations it is completely unimportant that the random estimates reported by the oracle come from certain integrants in such a way that the observations of $f_j(x)$ come from the values, and observations of $\nabla f_j(x)$ come from the gradients of the integrants. It suffices to postulate the following:

- **The goal** is to solve the Convex Programming program[1]

$$f_0(x) \to \min \mid\ f_j(x) \le 0,\ j = 1,...,m; x \in G \subset \mathbf{R}^n \tag{14.0.3}$$

  [G is closed and convex, f_i, $i = 1,...,m$, are convex and Lipschitz continuous on G]

- **The information** obtained by an optimization method at i-th step, $i = 1,2,...$, comes form a *stochastic oracle*, i.e., is a collection

$$\phi_j(x_i,\omega_i) \in \mathbf{R}, \psi_j(x_i,\omega_i) \in \mathbf{R}^n,\ j = 0,...,m,$$

  of reals and vectors, where

  - $\{\omega_i\}$ is sequence of independent identically distributed, according to certain probability distribution P, random parameters taking values in certain space Ω;

[1] Of course, the model we are about to present makes sense not only for convex programs; but the methods we are interested in will, as always, work well only in convex case, so that we loose nothing when imposing the convexity assumption from the very beginning

– x_i is i-th search point generated by the method; this point may be an arbitrary deterministic function of the information obtained by the method so far (i.e., at the steps $1, ..., i-1$)

It is assumed (and it is crucial) that *the information obtained by the method is unbiased:*

$$\mathcal{E}\phi_j(x,\omega) = f_j(x),\ f_j'(x) \equiv \mathcal{E}\psi_j(x,\omega) \in \partial f_j(x),\ \ j = 0, ..., m,\ x \in G; \tag{14.0.4}$$

here $\mathcal{E}$ is the expectation with respect to the distribution of the random parameters in question.

We have described the generic scheme of the optimization methods we are interested in now. Let us look what are the abilities of these methods.

14.1 Stochastic Approximation: simple case

Let us start with the simplest (and perhaps the most important for applications) case when there are no functional constraints in the problem, so that it is simply the problem

$$f(x) \to \min \mid\ x \in G \subset \mathbf{R}^n \tag{14.1.1}$$

(we write f instead of f_0 and, similarly, drop the index in our "observation functions" $\phi(x,\omega)$ (the estimate of $f(x)$) and $\psi(x,\omega)$ (the estimate of $f'(x)$).

14.1.1 Assumptions

To get reasonable complexity results, we need to bound somehow the magnitude of the random noise in the process in the stochastic oracle (as it is always done in all statistical considerations). Mathematically, the most convenient way to do it is as follows: let

$$L = \sup_{x \in G} \left\{ \mathcal{E}|\psi(x,f)|^2 \right\}^{1/2}. \tag{14.1.2}$$

From now on we assume that the oracle is such that $L < \infty$. The quantity L will be called the *intensity* of the oracle at the problem in question; in what follows it plays the same role as the Lipschitz constant of the objective in large-scale minimization of Lipschitz continuous convex functions (Lecture 8).

14.1.2 The Stochastic Approximation method

The method we are interested in is completely similar to the Subgradient Descent method from Section 8.1. Namely, the method generates search points according to the recurrence (cf. (8.1.1))

$$x_{i+1} = \pi_G(x_i - \gamma_i \psi(x_i, \omega_i)),\ i = 1, 2, ..., \tag{14.1.3}$$

where

- $x_1 \in G$ is an arbitrary starting point;

- γ_i are deterministic positive stepsizes;
- $\pi_G(x) = \mathrm{argmin}_{y\in G}|x-y|$ is the standard projector on G.

The only difference with (8.1.1) is that now we replace the direction $g(x_i) = f'(x_i)/|f'(x_i)|$ of the subgrade of f at x_i by the random estimate $\psi(x_i,\omega_i)$ of the subgradient.

Note that the search trajectory of the method is governed by the random variables ω_i and is therefore random:

$$x_i = x_i(\omega^{i-1}),\ \omega^s = (\omega_1,...,\omega_s).$$

Recurrence (14.1.3) defines the sequence of the search points, not that one of the approximate solutions. In the deterministic case (Section 8.1), we extracted from the search points the approximate solutions by choosing the best (with the smallest value of the objective) of the search points generated so far. The same could work in the stochastic case as well, but here we meet with the following obstacle: we "do not see" the values of the objective, and therefore cannot say which of the search point is better. To resolve the difficulty, we use the same trick as in the case of games and variational inequalities (Section 9.1) where we dealt with implicitly defined objectives, namely, define i-th approximate solution as the sliding average of the search points:

$$x^i \equiv x^i(\omega^{i-1}) = \left[\sum_{i/2\le t\le i}\gamma_i\right]^{-1}\sum_{i/2\le t\le i}\gamma_t x_t. \tag{14.1.4}$$

The efficiency of the resulting method is given by the following

Theorem 14.1.1 *For the aforementioned method one has, for all positive integer N,*

$$\varepsilon_N \equiv \mathcal{E}[f(x^N) - \min_{x\in G} f(x)] \le \frac{D^2 + L^2\sum_{N/2\le i\le N}\gamma_i^2}{2\sum_{N/2\le i\le N}\gamma_i}, \tag{14.1.5}$$

G being the diameter of G.

In particular, for γ_i chosen according to

$$\gamma_i = \frac{D}{L\sqrt{i}} \tag{14.1.6}$$

one has

$$\varepsilon_N \le O(1)\frac{LD}{\sqrt{N}} \tag{14.1.7}$$

with appropriately chosen absolute constant $O(1)$.

Proof is completely similar to that one used in Section 8.1 to get the estimate (8.1.5). Namely, from (14.1.3) and Lemma 8.1 it follows that if x^* is an optimal solution to (14.1.1), then

$$\delta_{i+1}^2 \equiv |x_{i+1}(\omega^i) - x^*|^2 \le \delta_i^2 - 2\gamma_i(x_i(\omega^{i-1}) - x^*)^T\psi(x_i(\omega^{i-1}),\omega_i) + \gamma_i^2|\psi(x_i(\omega^{i-1}),\omega_i)|^2.$$

Taking expectations of both sides, we get

$$d_{i+1}^2 \equiv \mathcal{E}\delta_{i+1}^2 \le d_i^2 - 2\gamma_i\mathcal{E}\left[(x_i(\omega^{i-1}) - x^*)^T\psi(x_i(\omega^{i-1}),\omega_i)\right] + \mathcal{E}\left[|\psi(x_i(\omega^{i-1}),\omega_i)|^2\right]. \tag{14.1.8}$$

The expectation of the second term in the right hand side can be computed as follows: first we can take it with respect to ω_i, and then take the expectation of the result with respect to ω^{i-1}. In the product to which we should apply the inner expectation (the one over ω_i) the first factor is independent of this variable, so that all we should do to get the inner expectation is to replace $\psi(x_i(\omega^{i-1}),\omega_i)$ with $\mathcal{E}_{\omega_i}\psi(x_i(\omega^{i-1}),\omega_i)$; the latter quantity, due to (14.0.4), is exactly $f'(x_i(\omega^{i-1}))\in\partial f(x_i(\omega^{i-1}))$, so that

$$\mathcal{E}\left[(x_i-x^*)^T\phi(x_i,\omega_i)\right]=\mathcal{E}\left[(x_i-x^*)^Tf'(x_i)\right]\geq\mathcal{E}[f(x_i)-f(x^*)]$$

(the concluding inequality follows from the convexity of f).

Similar reasoning applied to the third term in the right hand side of (14.1.8) (with (14.1.2) playing the role of (14.0.4), demonstrates that this term does not exceed $\gamma_i^2L^2$. Thus, we get from (14.1.8)

$$d_{i+1}^2\leq d_i^2-2\gamma_i\mathcal{E}\left[f(x_i)-f(x^*)\right]+\gamma_i^2L^2,$$

whence, due to $d_i\leq D$,

$$\mathcal{E}\left[\sum_{N/2\leq i\leq N}2\gamma_i(f(x_i)-f(x^*))\right]=2\sum_{N/2\leq i\leq N}\gamma_i\mathcal{E}\left[f(x_i)-f(x^*)\right]\leq D^2+L^2\sum_{N/2\leq i\leq N}\gamma_i^2.$$

The left hand side in this relation, due to Jensen's inequality, is not less than

$$2[\sum_{N/2\leq i\leq N}\gamma_i]\mathcal{E}\left[f(x^N)-f(x^*)\right],$$

and we come to (14.1.5). (14.1.7) is an immediate consequence of (14.1.5) - (14.1.6). ∎

14.1.3 Comments

The statement and the proof of Theorem 14.1.1 are completely similar to the related "deterministic" considerations of Section 8.1. The only difference is that now we are estimating from above the expected inaccuracy of N-th approximate solution; this is quite natural, since the stochastic nature of the process makes it impossible to say something reasonable about the quality of *every* realization of the random vector $x^N=x^N(\omega^{N-1})$

It turns out that the rate of convergence established in (14.1.7) is, in certain sense, unimprovable. Namely, it is not difficult to prove the following statement.

Proposition 14.1.1 *For every $L>0$, any $D>0$, any positive integer N and any stochastic-oracle-based N-step method $\mathcal{M}$ of minimizing univariate convex functions over the segment $G=[0,D]$ on the axis there exists a linear function f and a stochastic oracle with intensity L on the function such that*

$$\mathcal{E}[f(x^N)-\min_G f]\geq O(1)\frac{LD}{\sqrt{N}},$$

x^N being the result formed by the method as applied to the problem f. Here $O(1)$ is properly chosen positive absolute constant.

Note that in the deterministic case the rate of convergence $O(1/\sqrt{N})$ was unimprovable only in the large-scale case; in contrast to this, in the stochastic case this rate becomes optimal already when we are minimizing univariate linear functions.

Convergence rate $O(1/\sqrt{N})$ can be improved only if the objective is *strongly convex*. The simplest and the most important result here is as follows (the proof is completely similar to that one of Theorem 14.1.1):

Proposition 14.1.2 *Assume that convex function f on $\mathbf{R}^n$ attains its minimum on G at a unique point x^* and is such that*

$$f(x^*) + \frac{\theta}{2}|x - x^*|^2 \le f(x) \le f(x^*) + \frac{\Theta}{2}|x - x^*|^2,\ x \in G, \tag{14.1.9}$$

with certain positive θ and Θ. Consider process (14.1.3) with the stepsizes

$$\gamma_i = \frac{\gamma}{i}, \tag{14.1.10}$$

γ being a positive scale factor satisfying the relation

$$\gamma\theta > 1. \tag{14.1.11}$$

Then

$$\varepsilon_N \equiv \mathcal{E}(f(x_N) - \min_G f) \le c(\gamma\theta)\Theta\frac{D^2 + \gamma^2 L^2}{N}, \tag{14.1.12}$$

where D is the diameter of G, L is the intensity of the oracle at the problem in question and $c(\cdot)$ is certain problem-independent function on $(1, \infty)$.

The algorithm (14.1.3) with the stepsizes (14.1.10) and the approximate solutions identical to the search points is called the *classical Stochastic Approximation*; it originates from Kiefer and Wolfovitz. A good news about the algorithm is its rate of convergence: $O(1/N)$ instead of $O(1/\sqrt{N})$. A bad news is that this better rate is ensured only in the case of problems satisfying (14.1.9) and that the rate of convergence is very sensitive to the choice of the scale factor γ in the stepsize formula (14.1.10): if this scale factor does not satisfy (14.1.11), the rate of convergence may become *worse in order*. To see this, consider the following simple example: the problem is

$$f(x) = \frac{1}{2}x^2,\ [\theta = \Theta = 1]; \quad G = [-1, 1]; \quad x_1 = 1,$$

the observations are given by

$$\psi(x, \omega) = x + \omega \quad [= f'(x) + \omega],$$

and ω_i are the standard Gaussian random variables ($\mathcal{E}\omega_i = 0$, $\mathcal{E}\omega_i^2 = 1$); we do not specify $\phi(x, \omega)$, since it is not used in the algorithm. In this example, the best choice of γ is $\gamma = 1$ (in this case one can make (14.1.11) an equality rather than strict inequality due to the extreme simplicity of the objective). For this choice of γ one has

$$\mathcal{E}f(x_{N+1}) \equiv \mathcal{E}[f(x_{N+1}) - \min_x f(x)] \le \frac{1}{2N},\ N \ge 1.$$

In particular, it takes no more than 50 steps to reach expected inaccuracy not exceeding 0.01.

Now assume that when solving the problem we overestimate the quantity θ and choose stepsizes according to (14.1.10) with $\gamma = 0.1$. How many steps do we need in this case to reach the same expected inaccuracy 0.01 – 500, 5000, or what? The answer is astonishing: approximately 1,602,000 steps. And with $\gamma = 0.05$ (20 times less than the optimal value of the parameter) the same accuracy costs more than 5.2×10^{14} steps!

We see how dangerous the "classical" rule (14.1.10) for the stepsizes is: underestimating of γ ($\equiv$ overestimating of θ) may kill the procedure completely. And where from, in more or less complicated cases, could we take a reasonable estimate of θ? It should be said that there exist stable versions of the classical Stochastic Approximation (they, same as our version of the routine, use "large", as compared to $O(1/i)$, stepsizes and take, as approximate solutions, certain averages of the search points). These stable version of the method are capable to reach (under assumptions similar to (14.1.9)) the $O(1/N)$-rate of convergence, even with the optimal coefficient at $1/N$. Note, anyhow, that the nondegeneracy assumption (14.1.9) is crucial for the $O(1/N)$-rate of convergence; if it is removed, the best possible rate, as we know from Proposition 14.1.1, becomes $O(1/\sqrt{N})$, and this is the rate given by our "robust" Stochastic Approximation with "large" steps and averaging.

14.2 MinMax Stochastic Programming problems

Now let us consider problems with constraints. In the case of Stochastic Programming (similarly to what happens in the case of smooth convex optimization) straightforward extension of the results and methods from the case without functional constraints to that one when they are present causes slight difficulties; this is why we restrict ourselves with the *MinMax Stochastic program*

$$\textit{minimize} \quad \bar{f}(x) = \max_{j=1,...,m} f_j(x) \quad \textit{s.t.} \quad x \in G \in \mathbf{R}^n. \tag{14.2.1}$$

Here, as always, G is a closed and bounded convex set and the functions f_j are convex and Lipschitz continuous on G. We assume that the functions f_j are observed via stochastic oracle with observation functions

$$\phi_j(x,\omega), \psi_j(x,\omega) \quad [\mathcal{E}\phi_j(x,\omega) = f_j(x),\ \mathcal{E}\psi_j(x,\omega) \in \partial f_j(x),\ \forall x \in G]\,.$$

Note that although (14.2.1) is, formally, a convex optimization problem without functional constraints, our previous approach cannot be used here, since we have unbiased observations for f_j only, not for $\bar{f}$, and the transformation $(f_1, ..., f_m) \mapsto \bar{f}$ is nonlinear and therefore does not preserve unbiasedness.

The right way to solve the problem is to convert it into the saddle point program

$$\textit{find saddle point of} \quad L(x,y) = \sum_{i=1}^{m} y_j f_j(x) \quad \textit{on} \quad G \times Y,\ Y = \{y \geq 0 \mid \sum_j y_j = D\}, \tag{14.2.2}$$

D being the diameter of the set G.

Since the loss function

$$\bar{L}(x) \equiv \max_{y \in Y} L(x,y)$$

of the first player in the game with the cost function L (see Lecture 5) is exactly $D\bar{f}$, the x-components of saddle points of L are exactly the minimizers of $\bar{f}$ over G, so that in order to approximate a minimizer of $\bar{f}$ it suffices to approximate a saddle point of L.

The advantage of the reduction (14.2.1) $\Rightarrow$ (14.2.2) is that the answers of the oracle allow to form *unbiased* observations of partial derivatives (more exactly, sub- and superdifferentials) of L with respect to x and to y; these observations are, respectively,

$$\psi_x(x,y,\omega)=\sum_{j=1}^m y_i\psi_j(x,\omega),\ \ \psi_y(x,y,\omega)=(\phi_1(x,\omega),...,\phi_m(x,\omega))^T.$$

Now we can approximate the saddle set of L by the stochastic version of the Subgradient Descent method for games (Section 9.1) as follows:

$$\begin{pmatrix} x_{i+1}\\ y_{i+1}\end{pmatrix}=\pi_{G\times Y}\begin{pmatrix} x_i-\gamma_i\psi_x(x_i,y_i,\omega_i)\\ y_i+\gamma_i\psi_y(x_i,y_i,\omega_i)\end{pmatrix},\ i=1,2,..., \tag{14.2.3}$$

where the starting point (x_1,y_1) is an arbitrary deterministic point of $G\times Y$, and γ_i are positive deterministic stepsizes.

Same as in Section 9.1, the approximate solution to (14.2.1) generated after N steps of the method is the sliding average

$$x^N=\left[\sum_{N/2\le i\le N}\gamma_i\right]^{-1}\sum_{N/2\le i\le N}\gamma_i x_i$$

of the search points x_i.

The results on the convergence of the proposed algorithm are given by the following statement (we omit its proof):

Theorem 14.2.1 *Let problem (14.2.1) be solved by the method (14.2.3), and assume that the intensity*

$$L=D^{-1}\sup_{x\in G}\left[\sum_{j=1}^m\mathcal{E}|\phi_j(x,\omega)|^2\right]^{1/2}+\sup_{x\in G}\left[\sum_{j=1}^m\mathcal{E}|\psi_j(x,\omega)|^2\right]^{1/2}$$

of the stochastic oracle is finite. Then, for every positive integer N, the expected inaccuracy of the N-th approximate solution to the problem can be estimated from above as

$$\varepsilon_N\equiv\mathcal{E}\left[\bar{f}(x^{N+1})-\min_G\bar{f}\right]\le 4\frac{D^2+L^2\sum_{N/2\le i\le N}\gamma_i^2}{\sum_{N/2\le i\le N}\gamma_i}. \tag{14.2.4}$$

In particular, with the stepsizes given by

$$\gamma_i=\frac{D}{L\sqrt{i}} \tag{14.2.5}$$

one has

$$\varepsilon_N\le O(1)\frac{LD}{\sqrt{N}},\ N=1,2,..., \tag{14.2.6}$$

with appropriately chosen absolute constant $O(1)$.

Hints to Exercises

Hints to Section 1.3

Exercise 1.3.4$^+$: to prove that $x^*(Q) \in \text{int } Q$, note that if x is not in the interior of Q, then, due to the Separation Theorem for convex sets, there exists an affine functional f such that $f(x) > f(y)$ for all $y \in \text{int } Q$, and use the result of exercise 1.3.3.

Exercise 1.3.5$^+$: note that there exists an invertible affine mapping of the space which performs a given permutation of the vertices of the simplex, and use the result of exercise 1.3.4.

Exercise 1.3.8:

1): this is an immediate consequence of the characterization of the center of gravity given by exercise 1.3.3

2): Given a positive α, consider the cone

$$Q_\alpha = \{(x,t) \mid |x| \le \alpha^{-1}(t+\alpha)\phi(0),\ -\alpha \le t \le \beta(\alpha)\},$$

where $\beta(\alpha) > 0$ is defined by the requirement that the volume of the part of Q_α to the right of the hyperplane $\{t = 0\}$ is equal to the volume of the similar part of Q (of course, such a β does exist). Prove that for an appropriately chosen $\alpha > 0$ the volume of the part of Q_α to the left of the hyperplane $\{t = 0\}$ is equal to the volume of the similar part of Q.

3)$^+$: make rigorous the following "mechanical arguments": since ϕ is concave, the t-coordinate of any point from $G = Q_+^* \backslash Q_+$ is $\ge$ than the t-coordinate of any point from $H = Q_+ \backslash Q_+^*$. Further, G and H are of the same volumes (by construction), so that to obtain Q_+^* from Q_+ we should move certain masses to the hyperplane $\{t = 0\}$, which decreases the momentum of the functional t. Apply similar reasoning to Q_-^*, Q_-.

6): the center of gravity of the cone Q^* can be computed directly:

$$x^*(Q^*) = (0, -\alpha + n(\beta+\alpha)/(n+1)) \equiv (0, \tau);$$

it is immediately seen that the part Q' of Q^* to the left of the hyperplane $\{t = \tau\}$ is of the volume $(n/(n+1))^n \,\text{Vol}_n(Q^*) = (n/(n+1))^n \,\text{Vol}_n(Q)$ (the latter equality follows from 2)). Since 5) says that $\tau \le 0$, the volume of Q' does not exceed that one of Q_-^*, Q.E.D.

Exercise 1.3.10: let x^* be the center of gravity of Q, and let $\delta = [a,b]$ be a chord passing through x^*, $[a,x^*]$ being the larger, and $[x^*,b]$ being the smaller part of the chord. Choose the support to Q hyperplane Π passing through b, and consider the hyperplane Π_0 passing through x^* parallel to Π. It is convenient to think of Π as being horizontal with a being above the hyperplane.

Now, imagine that we place a lamp at a and make Π_0 not transparent, excluding the transparent hole formed by the intersection of Q and Π_0. Let B be the lightened part of Π; consider the conic set Q^+ with the vertex at a and the base B (i.e., the convex hull of a and B). Prove that the center of gravity of this set cannot be above Π_0, and conclude from this that the ratio of $[a, x^*]$ to $[x^*, b]$ is at most n.

Exercise 1.3.11$^+$: consider the family of all closed half-spaces of μ-measure $> n/(n+1)$. Prove that the family satisfies the premise of Helley's theorem and verify that any point which belongs to the intersection of the sets from the family satisfies the conclusion of the second theorem of Grunbaum.

Exercise 1.3.13$^+$: use the result stated in exercise 1.3.12

Exercise 1.3.16$^+$: the statement in question is exactly the following one:

let x^{n+1} denote the ordered collection of $n+1$ points $x_1, ..., x_{n+1}$ of X, and let

$$r(x^{n+1}) = \min_{p\in\Phi} \max_{i=1,...,n+1} \| f(x_i) - p(x_i) \| .$$

Then

$$\delta' \equiv \max_{x^{n+1}} r(x^{n+1}) = \delta.$$

To prove the statement, verify that the maximum in the latter statement is achieved (so that δ' is well-defined) and is $\le \delta$. To prove the inverse inequality, apply the Helley theorem to the family of sets

$$\{H_x = \{p \in \Phi \mid \| f(x) - p(x) \| \le \delta'\}\}_{x\in X}.$$

Exercise 1.3.17$^+$: note that the sum $H_1 + H_2$ of one-dimensional closed and nonempty sets contains the translations H_1', H_2' of H_1, H_2, respectively, with the intersection comprised of a single point.

Exercise 1.3.18$^+$: we should prove that

$$\int_{T_+} r_{k-1}^{k-1}(H_+^t)dt \equiv (\mathrm{Vol}_k(H_1 + H_2))^{1/k} \ge$$

$$\ge (\mathrm{Vol}_k(H_1))^{1/k} + (\mathrm{Vol}_k(H_2))^{1/k} \equiv \int_{T_1} r_{k-1}^{k-1}(H_1^t)dt + \int_{T_2} r_{k-1}^{k-1}(H_2^t)dt,$$

or, which is the same in view of (1.3.7), that

$$\left(\int_0^\infty s^{k-1}\mu_+(s)ds\right)^{1/k} \ge \left(\int_0^\infty s^{k-1}\mu_1(s)ds\right)^{1/k} + \left(\int_0^\infty s^{k-1}\mu_2(s)ds\right)^{1/k},$$

where

$$\mu_i(s) = \mathrm{Vol}_1(\{t \in T_i \mid r_{k-1}(H_i^t) \ge s\}), i = 1, 2,$$

((x, t) are the coordinates in $\mathbf{R}^k$, x being $(k-1)$-dimensional and t being scalar, T_i are the images of H_i under the projection onto the t-axis and H_i^t are the inverse images of the point $(0, ..., 0, t)$ in H_i under this projection),
and

$$\mu_+(s) = \mathrm{Vol}_1(\{t \in T_+ \mid r_{k-1}(H_+^t) \ge s\}),$$

T_+, H_+^t being similar objects for $H_+ = H_1 + H_2$.

Prove that T_i is a nonempty compact set and the function $r_{k-1}(H_i^t)$ is upper semicontinuous on T_i, so that it attains its maximum ρ_i over T_i, $i = 1, 2$.

Reduce the situation to the case of positive ρ_i and prove and use the fact that if

$$\alpha_i = \rho_i/(\rho_1 + \rho_2),\ i = 1, 2,$$

then for all nonnegative s one has

$$\mu_+(s)ds \geq \mu_1(\alpha_1 s) + \mu_2(\alpha_2 s).$$

Hints to Section 2.4

Exercise 2.4.5$^+$: note that V_α is contained in the set of solutions to the system of the following pair of quadratic inequalities:

$$x^T x \leq 1; \tag{14.2.7}$$

$$(e^T x - \alpha)(e^T x - 1) \leq 0; \tag{14.2.8}$$

Consequently, V_α satisfies any convex quadratic inequality which can be represented as a weighted sum of (14.2.7) and (14.2.8), the weights being positive. Every inequality of this type defines an ellipsoid; compute the volume of the ellipsoid as the function of the weights and optimize it with respect to the weights.

Exercise 2.4.6$^+$: use the same construction as that one for exercise 2.4.5

Exercise 2.4.7$^+$: use the results given by exercises 2.4.5 and 2.4.6

Exercise 2.4.8$^+$: use item 3. of the list of consequences of the Fritz John Theorem

Exercise 2.4.9$^+$: set

$$\| x \| = |x|_1 \equiv \sum_{i=1}^{n} |x_i|$$

and consider the sequence of random vectors defined as follows: to generate ξ_i, choose randomly, according to the uniform distribution on $\{1, ..., n\}$, an integer ν between 1 and n and set all entries of ξ_i with the indices different from ν to zero; the ν-th entry is set to ± 1, with a randomly chosen sign (this choice is independent of that one for ν and both the signs have equal probabilities to be chosen).

Exercise 2.4.12$^+$: combine (2.4.9) and (2.4.17)

Hints to Section 3.3

Exercise 3.3.2: prove that the volume of every simplex in the partitioning is at most

$$\frac{1}{n!}(2\sqrt{n})^n.$$

Exercise 3.3.3: note that if Γ_0 is not a singleton, then

$$\pi(x) = \frac{1}{|I(x)|}\sum_{x' \in I(x)} \pi(x'),$$

here $I(x)$ is the set of all vertices in $\Gamma_0 \backslash \{x\}$ from which the moving point can reach x in a single step and $|I(x)|$ is the # of elements in $I(x)$. Think of where the function $\pi(x)$, $x \in \Gamma_0$, might achieve its maximum (it is a discrete version of the Maximum Modulus principle for harmonic functions, is not it?).

Exercise 3.3.5$^+$: given the trajectory $\{x^t\}_{t=0}^T$ with $x^T \geq \lambda^T x^0$, consider the trajectory $\{\bar{x}^t = \lambda^{-t}\}_{t=0}^T$ and verify that it satisfies the inequalities

$$\bar{A}\bar{x}^{t+1} \leq B\bar{x}^t,\ t = 0, ..., T-1$$

(here $\bar{A} = \lambda A$) and the relation

$$\bar{x}^T \geq \bar{x}^0.$$

Look what $\bar{A}$ and B do with the nonnegative and nonzero vector

$$\bar{x} = \frac{1}{T}\sum_{t=0}^{T-1} \bar{x}^t.$$

Exercise 3.3.7: follow the proof of the similar statement for the case of convex problems without functional constraints (Lecture 2).

Hints to Section 4.3

Exercise 4.3.4: without loss of generality we may assume that the linear form $g^T x$ attains its minimum over D at the vertex v_0 and that $g^T(w - v_0) = 1$. Choosing v_0 as our new origin and $v_1 - v_0, ..., v_n - v_0$ as the orths of our new coordinate axes, we come to the situation studied in exercise 4.3.3.

Hints to Section 5.5

Exercise 5.5.2: note that a nonsingular $n \times n$ matrix B can be represented as

$$B = B'U,$$

where B' is positive semidefinite symmetric and U is orthogonal; this is called the *polar decomposition* of B. Note that by evident reasons

$$W(B, c) = W(B', c).$$

If the existence of the polar decomposition is a news for you, here is the proof: the matrix $B' = (BB^T)^{1/2}$ is well-defined symmetric positive definite; if $U = (B')^{-1}B$, then

$$UU^T = (B')^{-1}(BB^T)(B')^{-1} = (B')^{-1}(B')^2(B')^{-1} = I,$$

so that U is orthogonal. By construction, $B = B'U$. ∎

Exercise 5.5.6: use (5.5.4), (5.5.5) and the following basic fact:

let $\nu_1 \geq \nu_2 \geq ... \geq \nu_n$ be half-axes of an n-dimensional ellipsoid E and let $\nu'_1 \geq \nu'_2 \geq ... \geq \nu'_{n-1}$ be the half-axes of the $(n-1)$-dimensional ellipsoid obtained by intersecting E by a hyperplane passing through the center of E. Then ν'_i separate ν_i:

$$\nu_1 \geq \nu'_1 \geq \nu_2 \geq \nu'_2 \geq ... \geq \nu_{n-1} \geq \nu'_{n-1} \geq \nu_n.$$

For those non-acquainted with this important statement I would add that this is an immediate geometrical consequence of the following fundamental characterization of the eigenvalues $\nu_1 \geq \nu_2 \geq ... \geq \nu_n$ of a symmetric $n \times n$ matrix A (Fisher-Courant):

$$\nu_i = \min_{F \in \mathcal{F}_i} \max_{x \in F, |x|_2 = 1} x^T A x,$$

where $\mathcal{F}_i$ is the family of all linear subspaces in $\mathbf{R}^n$ of codimension $n - i + 1$.

If these facts are new for you, it would be a good idea to find the proofs yourself (it is not difficult).

Hints to Section 6.3

Exercise 6.3.1: the true statements are 5), 7), 8) and 9)

Exercise 6.3.2$^+$: the statement is trivial if int $G = \emptyset$; thus, we may assume that int $G \neq \emptyset$. It suffices to lead to a contradiction the assumption that a point $x \in \text{int } G$ does not belong to $\text{Dom } F$; without loss of generality, we may assume that the point in question is 0. What we should do is demonstrate that one can find y in such a way that $x^T(\xi - y) \geq 0$ whenever $x \in \text{Dom } F$ and $\xi \in F(x)$ (this would mean that one can add to the graph of F the pair $(0, y)$ without violating monotonicity, thus yielding the desired contradiction with the maximal monotonicity of F). Thus, we should prove that certain infinite system of linear inequalities is solvable. To this end prove, first, that each finite subsystem of the system is solvable and, second, that there exists a finite subsystem of the system with a bounded solution set.

Exercise 6.3.3: f' is not necessarily maximal monotone; counter-example: $f(x) = 0$, $-1 < x < 1$, $f(-1) = f(1) = 1$, $f(x) = +\infty$ when $|x| > 1$.

Exercise 6.3.7$^+$: use the following observation: a weak solution to the variational inequality defined by G and a *maximal* monotone operator H with $\text{int } G \subset \text{Dom } H$ is a strong solution to the inequality, provided that H is semibounded on int G. Apply this observation to the operator $H = F + \mathcal{N}_G$

Hints to Section 7.5

Exercise 7.5.3: follow the line of argument of the original proof of Proposition 7.3.1. Namely, apply the proof to the "shifted" process which starts at $x_{N'}$ and uses at its i-th iteration, $i \geq 1$, the stepsize $\gamma_{i+N'-1}$ and the tolerance $\varepsilon_{i+N'-1}$. This process differs from that one considered in the lecture in two issues:

(1) presence of time-varying tolerance in detecting productivity and an "arbitrary" step, instead of termination, when a productive search point with vanishing subgradient of the objective is met;

(2) exploiting the projection onto $Q \supset G$ when updating the search points.

To handle (1), prove the following version of Proposition 2.3.1 (Lecture 2):

Assume that we are generating a sequence of search points $x_i \in \mathbf{R}^n$ and associate with these points vectors e_i and approximate solutions $\bar{x}_i$ in accordance to (i)-(iii). Let

$$G_i = \{x \in G \mid (x - x_j)^T e_j \leq 0,\ 1 \leq j \leq i\},$$

and let Size *be a size. Assume that for some M*

$$\mathrm{Size}(G_M) < \varepsilon_M \,\mathrm{Size}(G)$$

(if G_M is not a solid, then, by definition, $\mathrm{Size}(G_M) = 0$*). Then among the search points $x_1, ..., x_M$ there were productive ones, and the best (with the smallest value of the objective) of these productive points is a ε_1-solution to the problem.*

To handle (2), note that when estimating $\mathrm{InnerRad}(G_N)$, we used the equalities

$$|x_{j+1} - x^+|_2^2 = |x_j - x^+|_2^2 + ...$$

and would be quite satisfied if $=$ in these inequalities would be replaced with $\leq$; in view of Exercise 7.5.2, this replacement is exactly what the projection does.

Hints to Section 8.3

Exercise 8.3.4*: use the result of exercise 8.3.3 and (8.3.4) to demonstrate that

$$V^+(\phi_{i+1}) \le V^+(\phi_i) - 2\gamma_i(f(x_i) - f^*)/|f'(x_i)|_* + \frac{\mathcal{L}}{2}\gamma_i^2,\ V^+(\phi) = V(\phi) - \langle \phi, x^* \rangle\,,$$

and then act exactly as in the case of the usual Subgradient Descent.

Hints to Section 9.3

Exercise 9.3.4: since $(u, v) \in C(i)$, we have

$$(u, v) \geq \sum_{j=1}^{i} \lambda_j (f(x_j), g(x_j))$$

for some nonnegative λ_j with unit sum; set

$$x = \sum_j \lambda_j x_j.$$

Exercise 9.3.6: note that if $\max\{u, v\} \geq 0$, then

$$\rho(u, v) = \max_{0 \leq \alpha \leq 1} \{\alpha u + (1 - \alpha)v\}$$

and take into account that if $(u, v) \in C(i)$, then $\max\{u - f_i^*, v\} \geq \min_G \max\{f_i(x) - f_i^*, g_i(x)\} \geq 0$ (why?). With these observations, apply the von Neumann Lemma to the convex-concave function

$$F(\alpha, (u, v)) = \alpha(u - f_i^*) + (1 - \alpha)v$$

regarded as a function on the direct product $[0, 1] \times C(i)$.

Exercise 9.3.7: prove that $h_i(\alpha)$, for every i, is majorated by the corresponding gap in the Level process, and use the efficiency estimate for the usual Level.

Exercise 9.3.8: formulate exactly and use the following simple facts:

1) if h is a concave nonnegative function on a segment δ, α is μ-centered with respect to δ and $h(\alpha)$ is small, then $\max_\delta h$ also is small; thus, if in our method we were lucky not to vary the current value of α for a long time, we make the gap small;

2) if α is the midpoint of a segment δ and $\delta' \subset \delta$ is such that α is *not* μ-centered with respect to δ', then δ' is "significantly" (at least by a factor $1 - O(1)\mu$) less than δ; thus, each step where we are enforced to vary the current value of α is accompanied by a significant reduction in the length of the current segment δ_i as compared to the length of the similar segment at the step where the "old" value of α was set.

3) if the length of δ_i becomes small, then Δ_i also becomes small.

Hints to Section 12.6

Exercise 12.6.4: apply the criterion given by exercise 12.6.2. To compute the optimal value, use the representation

$$T_k(t) = \mathrm{ch}(k \operatorname{arcch}(t)),\ t \geq 1$$

of the Tschebyshev polynomial in the region $t \geq 1$.

Exercise 12.6.5: apply the criterion given by exercise 12.6.2. To prove (166), look at the value of the objective in (P) at polynomials p given by

$$(1 - tp(t)) = (1 - tp_m^*(t))^k$$

with $(m+1)k \leq N+1$, $p_m^*(t)$ being the optimal solution of problem (P) for the case of $N = m$ and $\gamma = 1/2$.

Exercise 12.6.6: to get a polynomial $s(t) = 1 - tp(t)$ of degree $2N+1$ such that $t^\gamma|1 - tp(t)|$ is small along the sequence $\{\sigma_i\}$, take it in the form

$$1 - tp(t) = (1 - tp_N(t)) \prod_{i=1}^{N} (1 - t/\sigma_i),$$

where $p_N(t)$ is the associated with (166) polynomial of the degree N corresponding to the case $L = \sigma_{N+1}$.

Hints to Section 13.4

Exercise 13.4.4: show that the spectrum of the matrix $I + B_1$ of the transformed system belongs to the segment $\Sigma = [\rho, 1/\rho]$, so that the condition number of this matrix does not exceed ρ^{-2}. Apply subsequently the upper complexity bound for the family of problems $\mathcal{U}_n(\Sigma, -\frac{1}{2}, R)$, $R^2 = f_{A,b}(0) - \min f_{A,b}$, given in Lecture 12 and the result on optimality of the usual CGM given in Lecture 13.

Exercise 13.4.5: use the Holder inequality. Those more experienced in Complex Analysis might prefer to obtain the statement in question as an immediate consequence of the Hadamard Theorem:

if a non identically zero function $f(z)$ is analytic and bounded in the stripe $a < \Re z < b$, then the function

$$M(x) = \ln \sup_{y \in \mathbf{R}} |f(x + iy)|$$

is convex in $x \in (a, b)$.

Exercise 13.4.6: Prove that the functions $1, t, t^2, ..., t^N$ are linearly independent in $L_2(\mu)$; apply to this sequence the Gramm-Schmidt orthogonalization process.

Exercise 13.4.8: combine the result stated in exercise 13.4.7.3) with the following fundamental fact:

let A be a symmetric operator in n-dimensional Euclidean space E, and let $\lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n$ be the eigenvalues of A taken with their multiplicities. Now, let E' be a subspace in E of codimension l, let A' be the restriction of A onto E' (i.e., the operator on E' such that $A'x$ is the orthogonal projection onto E' of the vector Ax) and let $\mu_1 \leq \mu_2 \leq ...\mu_{n-l}$ be the eigenvalues of A' taken with their multiplicities. Then

$$\lambda_i \leq \mu_i \leq \lambda_{i+l}, \ i = 1, ..., n - l.$$

Those for whom this general Separation Theorem for eigenvalues is new are welcomed to obtain it from the *Courant-Fisher characterization of the eigenvalues of a symmetric operator:*

let A be a symmetric operator on n-dimensional Euclidean space E and let $\lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n$ be the eigenvalues of A taken with their multiplicities. Then

$$\lambda_i = \max_{L \in \mathcal{E}_i} \min_{x \in L, x^T x = 1} x^T A x,$$

where $\mathcal{E}_i$ denotes the set of all subspaces of E of codimension $i - 1$.

Those not familiar with the latter extremely important statement, are strongly recommended to prove it (it is not difficult).

Solutions to exercises

Solutions to Section 1.3

Exercise 1.3.2: in view of exercise 1.3.1, I can give you almost nothing if you choose x outside the interior of Q. Now assume that you have chosen $x \in \text{int } Q$, so that

$$x = \sum_{i=1}^{n+1} \lambda_i u_i$$

for certain positive λ_i with the unit sum. Look what happens if I leave you the part of the pie given, in the barycentric coordinates[2], as

$$Q' = \{y \in Q \mid \lambda_{i^*}(y) \geq \lambda_{i^*}(x)\}$$

where i^* is the index of the largest of λ_i. In other words, look what happens if my cutting hyperplane is parallel to a properly chosen $(n-1)$-dimensional facet of the simplex (and I take the part neighbouring to the facet).

Exercise 1.3.4: to prove that the center of gravity is the interior point of the domain, assume, on contrary, that $x^* \equiv x^*(Q) \notin \text{int } Q$. Then, due to the Separation Theorem for convex sets, there exists an affine functional f such that $f(x^*) = 0$ and $0 \geq f(y)$ for all $y \in Q$, the inequality being strict for $y \in \text{int } Q$. In view of exercise 1.3.3, we have

$$0 = f(x^*) = \text{Vol}_n(Q)^{-1} \int_Q f(y)dy;$$

the right hand side in this equality clearly is strictly negative, which gives the desired contradiction.

The proof of the affine invariance is as follows:

$$\int_{Q^+} ydy = \int_Q \mathcal{A}(x)|\text{ Det } A|dx$$

[2] given $n+1$ affinely independent points u_i in $\mathbf{R}^n$, one can represent any $y \in \mathbf{R}^n$ as the linear combination

$$y = \lambda_1(y)u_1 + ... + \lambda_{n+1}(y)u_{n+1},$$

with the unit sum of coefficients (why?); these coefficients (which, under this latter requirement, are uniquely defined) are called the *barycentric coordinates* of y with respect to $\{u_i\}$. The barycentric coordinates are affine functions of y (why?)

(substitution of variable in an n-dimensional integral). Further,

$$\int_Q \mathcal{A}(x)|\,\mathrm{Det}\ A|dx = \int_Q (Ax+b)|\,\mathrm{Det}\ A|dx = |\,\mathrm{Det}\ A|\mathrm{Vol}_n(Q)b + |\,\mathrm{Det}\ A|A(\int_Q xdx) =$$

$$= |\det A|\mathrm{Vol}_n(Q)b + |\,\mathrm{Det}\ A|A(\mathrm{Vol}_n(Q)x^*(Q)) =$$

$$= (|\,\mathrm{Det}\ A|\mathrm{Vol}_n(Q))\mathcal{A}(x^*(Q)) = \mathrm{Vol}_n(Q^+)\mathcal{A}(x^*(Q)).$$

Thus,

$$\int_{Q^+} ydy = \mathrm{Vol}_n(Q^+)\mathcal{A}(x^*(Q)),$$

whence

$$x^*(Q^+) \equiv \mathrm{Vol}_n^{-1}(Q^+)\int_{Q^+} ydy = \mathcal{A}(x^*(Q)).$$

Exercise 1.3.5: let $v_1, ..., v_{n+1}$ be the vertices of the simplex, and let $i \mapsto \sigma(i)$ be a permutation of the indices $1, ..., n+1$. There exists a unique affine transformation T_σ of the space which, for all i, maps v_i to $v_{\sigma(i)}$; of course, this transformation maps the simplex onto itself. If

$$x = \lambda_1 v_1 + ... + \lambda_{n+1} v_{n+1}$$

is the representation of a point in the simplex as a convex combination of the vertices, then

$$T_\sigma(x) = \sum_{i=1}^{n+1} \lambda_i v_{\sigma(i)} = \sum_{i=1}^{n+1} \lambda_{\sigma^{-1}(i)} v_i.$$

Now, if x is the center of gravity of the simplex, then, in view of the statement given by exercise 1.3.4, we have $T_\sigma x = x$ (recall that T_σ maps the simplex onto itself). Since the barycentric coordinates of a point are uniquely defined by the point, we come to

$$\lambda_{\sigma^{-1}(i)} = \lambda_i$$

for all i and all permutations σ, whence $\lambda_1 = ... = \lambda_{n+1} = 1/(n+1)$, Q.E.D.

Exercise 1.3.6: due to affine invariance of the statement in question and affine invariance on the center of gravity, it suffices to consider the following special situation: b is at the origin, the base B of Q belongs to the hyperplane $\{x_n = 0\}$, $a = (0, ..., 0, 1)$. It is clear that the last coordinate of a point in Q varies from 0 to 1, so that we have

$$\mathrm{Vol}_n(Q) = \int_Q dx = \int_0^1 \mathrm{Vol}_{n-1}(Q_t)dt,$$

where $Q_t = \{x' \in \mathbf{R}^{n-1} \mid (x', t) \in Q\}$. It is clear that

$$Q_t = (1-t)B \equiv \{(1-t)x \mid x \in B\}$$

(indeed, $(x', t) \in Q$ if and only if $(x', t) = \alpha(0, ..., 0, 1) + (1-\alpha)(x_1, ..., x_{n-1}, 0)$ with some $\alpha \in [0, 1]$ and with $x'' \equiv (x_1, ..., x_{n-1}) \in B$. It follows that $\alpha = t$, so that $(x', t) \in Q$ if and only if $x' = (1-t)x''$ for some $x'' \in B$).

Since $Q_t = (1-t)B$, we have $\mathrm{Vol}_{n-1}(Q_t) = (1-t)^{n-1}\,\mathrm{Vol}_{n-1}(B)$. Thus,

$$\underset{n}{\mathrm{Vol}}(Q) = \int_0^1 (1-t)^{n-1}\,\underset{n-1}{\mathrm{Vol}}(B)dt = n^{-1}\,\underset{n-1}{\mathrm{Vol}}(B).$$

Similarly,

$$\mathrm{Vol}_n(Q)(x^*(Q))_n = \int_Q x_n dx = \int_0^1 t\mathrm{Vol}_{n-1}(Q_t)dt = \int_0^1 t(1-t)^{n-1}\mathrm{Vol}_{n-1}(B)dt =$$

$$= \mathrm{Vol}_{n-1}(B)\int_0^1 (1-\tau)\tau^{n-1}d\tau = \mathrm{Vol}_{n-1}(B)(1/n - 1/(n+1)) = \mathrm{Vol}_{n-1}(B)n^{-1}(n+1)^{-1},$$

whence, due to $\mathrm{Vol}_n(Q) = n^{-1}\,\mathrm{Vol}_{n-1}(B)$,

$$(x^*(Q))_n = 1/(n+1).$$

Now, if $i < n$, then

$$\mathrm{Vol}_n(Q)(x^*(Q))_i = \int_Q x_i dx = \int_0^1 \{\int_{Q_t} (x')_i dx'\}dt.$$

As we know, $Q_t = (1-t)B$; the substitution $x' = (1-t)y$ in the inner integral results in

$$\int_{Q_t} (x')_i dx' = (1-t)^n \int_B y_i dy = 0$$

(the latter equality follows from the assumption that the $(n-1)$-dimensional center of gravity of B is 0), and we come to

$$\mathrm{Vol}_n(Q)(x^*(Q))_i = 0, i = 1, ..., n-1.$$

Thus,

$$x^*(Q) = (0, ..., 0, 1/(n+1)) = (0, ...0, 1)/(n+1) + n0/(n+1) = a/(n+1) + nb/(n+1),$$

as claimed.

Exercise 1.3.8 3): from the concavity of ϕ it follows that $\alpha \geq a$ (otherwise Q_-^* would be a proper part of Q_- and therefore would be of less volume than the latter set). On the other hand, the slope ξ' of ξ is $\geq \phi'(-0)$ (since otherwise Q_- would be a proper part of Q_-^* and therefore would be of less volume that this latter set). Since ϕ is concave, it follows that $\phi(t) \geq \xi(t)$ for small in absolute values negative t and $\phi(t) \leq \xi(t)$ for all positive t. From this latter observation it follows that $\beta \leq b$ (otherwise Q_+ would be a proper part of Q_+^* and these sets could not be of equal volumes).

Now consider the function

$$g_+(t) = \begin{cases} \xi^{n-1}(t) - \phi^{n-1}(t), & 0 \leq t \leq \beta \\ -\phi^{n-1}(t), & \beta < t \leq b \end{cases}.$$

As we just have seen, this function is nonnegative for $0 \leq t \leq \beta$ and is nonnegative when $\beta < t \leq b$. We clearly have

$$0 = \mathrm{Vol}_n(Q_+^*) - \mathrm{Vol}_n(Q_+) = c_n \int_0^b g_+(t)dt$$

(c_n is the $(n-1)$-dimensional volume of the unit Euclidean ball in $\mathbf{R}^{n-1}$), which combined with the fact that g_+ is nonnegative to the left of β and is nonpositive to the right of this point implies that

$$0 \geq c_n \int_0^b t g_+(t)dt = I_t(Q_+^*) - I_t(Q_+);$$

this is the first inequality required in 3).

Now, since $\alpha \geq a$, ϕ is a nonnegative concave function on $[-a, 0]$, ξ is a nonnegative linear function on $[-\alpha, 0]$ and $\xi(0) = \phi(0)$, $\xi' \geq \phi'(-0)$, the function

$$g_-(t) = \begin{cases} \xi^{n-1}(t) - \phi^{n-1}(t), & -a \leq t \leq 0 \\ \xi^{n-1}(t), & -\alpha \leq t < -a \end{cases}$$

is as follows: there is certain $\tau \in [-\alpha, 0]$ such that $g_-(t)$ is nonnegative when $-\alpha \leq t < \tau$ and is nonpositive when $\tau < t \leq 0$. Besides this, from $\mathrm{Vol}_n(Q_-^*) = \mathrm{Vol}_n(Q_-)$ it follows that

$$\int_{-\alpha}^0 g_-(t)dt = 0.$$

It follows that

$$\int_{-\alpha}^0 (-t) g_-(t)dt \geq 0,$$

or, which is the same, that

$$I_{-t}(Q_-^*) \geq I_{-t}(Q_-),$$

which completes the proof of 3).

> When transforming Q into Q^*, we move masses from $Q \backslash Q^*$ to the left, thus increasing the momentum I_{-t} of the "left" part of the body and decreasing the momentum I_t of the "right" part of it. As a result, the momentum I_{-t} of the left part of the updated body becomes larger than the momentum I_t of the right part of it, since before updating the parts were balanced.

Exercise 1.3.9: the left hand side ratio clearly is equal to

$$\frac{\int_\alpha^1 (1-s^2)^{(n-1)/2} ds}{2\int_0^1 (1-s^2)^{(n-1)/2} ds}$$

(to get the n-dimensional volumes, integrate the $(n-1)$-dimensional volumes of the cross-sections of the body by the hyperplanes $\{x_n = s\}$ over s).

If $n > 1$, then the numerator can be bounded from above as follows:

$$\int_\alpha^1 (1-s^2)^{(n-1)/2} ds \leq \int_\alpha^1 \exp\{-\frac{(n-1)s^2}{2}\} ds \leq \int_\alpha^1 \exp\{-\frac{(n-1)\alpha s}{2}\} ds <$$

$$< \frac{2}{(n-1)\alpha} \exp\{-\frac{(n-1)\alpha^2}{2}\}.$$

If $n > 3$, then the denominator can be bounded from below as follows:

$$2\int_0^1 (1-s^2)^{(n-1)/2} ds > 2\int_0^{\sqrt{2/(n-1)}} (1-s^2)^{(n-1)/2} ds >$$

$$2\int_0^{\sqrt{2/(n-1)}}(1-\frac{2}{n-1})^{(n-1)/2}ds > \omega\sqrt{2/(n-1)}$$

with certain absolute constant ω.

Thus, in the case of $n > 3$ the ratio in question does not exceed

$$\kappa'\frac{1}{\alpha\sqrt{n}}\exp\{-\frac{n\alpha^2}{2}\},$$

κ' being an absolute constant. We conclude that if $\alpha\sqrt{n} \geq 1$ then

$$\mathrm{Vol}_n(V_\alpha)/\mathrm{Vol}_n(V) \leq \kappa'\exp\{-\frac{n\alpha^2}{2}\};$$

in the opposite case we clearly have

$$\mathrm{Vol}_n(V_\alpha)/\mathrm{Vol}_n(V) \leq 1 \leq \kappa''\exp\{-\frac{n\alpha^2}{2}\},$$

where $\kappa'' = \exp\{1/2\}$. It remains to set $\kappa = \max\{\kappa', \kappa''\}$.

Exercise 1.3.10: let $[a, b]$ be a chord of a closed convex domain Q passing through the center of gravity of Q; without loss of generality, let the latter be at the origin. Assume that $[a, 0]$ is the larger, and $[0, b]$ is the smaller part of the chord. Now, from Separation Theorem for convex sets it follows that there exists a nonconstant linear functional which attains its minimum over Q at b; under appropriately chosen coordinates we may assume that this functional is simply the last coordinate, x_n; let $-\beta$ be the value of the functional at b and α be its value at a, so that $\alpha > \beta > 0$, and what should be proved is that

$$\alpha \leq n\beta.$$

Let Q_0 be the cross-section of Q by the hyperplane $\Pi_0 = \{x_n = 0\}$, and K be the cone with the vertex at a which has Q_0 as its intersection with Π_0:

$$K = \{x = a + \lambda(y - a) \mid y \in Q_0, \lambda \geq 0\}.$$

Last, let Q^* be the part of K which is above the hyperplane $\Pi = \{x_n = -\beta\}$:

$$Q^* = \{x \in K \mid x_n \geq -\beta\}.$$

Let Q^*_-, Q^*_+ be the parts of Q^* in the half-spaces $\{x_n \leq 0\}$, $\{x_n \geq 0\}$, respectively, and let Q_-, Q_+ be similarly defined parts of Q. We claim that Q_- is contained in Q^*_-, while Q_+ contains Q^*_+.

The claim is "geometrically evident", and the rigorous proof is as follows. The first statement: if $x \in Q$ is a point with nonpositive last coordinate, then the segment $[a, x]$ intersects Π_0; the intersection point x' (which belongs to Q since a and x belongs to Q and Q is convex) gives us the ray $[a, x')$ which is contained in the cone K and contains x, so that x belongs to K. Since $x \in Q$ and the whole Q is above the hyperplane Π (i.e., $x_n \geq -\beta$), x belongs to Q^* and consequently to Q^*_- (recall that $x_n \leq 0$ due to $x \in Q_-$).

Now let us verify that Q_+^* is contained in Q_+. Indeed, if $x = a + \lambda(x' - a) \in Q_+^*$, where $\lambda \geq 0$ and $x' \in Q_0$, then

$$0 \leq x_n = a_n + \lambda(x'_n - a_n) = a_n - \lambda a_n = (1 - \lambda)\alpha,$$

so that $\lambda \leq 1$. Therefore $x = (1 - \lambda)a + \lambda x'$ is a convex combination of two points from Q and therefore belongs to Q, and since $x_n \geq 0$, x belongs to Q_+.

Now, since the center of gravity of Q is at the origin, the momentum of Q_+ with respect to the functional x_n is equal to the momentum of Q_- with respect to the functional $-x_n$. The momentum of Q_+^* with respect to x_n does not exceed that one of Q_+ (since Q_+^* is contained in Q_+), and the momentum of Q_-^* with respect to the functional $-x_n$ is less than that one of Q_-. We conclude that

$$I_{x_n}(Q_+^*) \leq I_{-x_n}(Q_-^*),$$

and therefore the center of gravity of Q^* is below the hyperplane $\{x_n = 0\}$ or on the hyperplane. But the set Q^* is conic with the vertex at the hyperplane $\{x_n = \alpha\}$E and the base in the hyperplane $\{x_n = -\beta\}$, and therefore the n-th coordinate of the center of gravity is $\alpha/(n + 1) - nb/(n + 1)$ (exercise 1.3.6), and since the coordinate is nonpositive, we come to $\alpha \leq n\beta$, as required.

Exercise 1.3.11: let $\mathcal{F}$ be the family comprised of all closed half-spaces with μ-measure $> n/(n + 1)$. Let us prove that the family satisfies the premise of the Helley theorem. Let us verify, first, that every $n+1$ sets of the family have a point in common. Indeed, otherwise there would be $n + 1$ half-spaces $\Pi_1, ..., \Pi_{n+1}$ from the family with empty intersection, so that the unit of the $n + 1$ open half-spaces $\mathbf{R}^n \backslash \Pi_i$ would be the whole space. Since the μ-measure of each of these open half-spaces is $< 1/(n+1)$ (the definition of the family), we conclude that the measure of the whole space is < 1, which is a contradiction.

It remains to verify that there exist finitely many sets from $\mathcal{F}$ with bounded intersection. Indeed, for a given unit vector e and large enough α the half-space $\{x \mid e^T x \leq \alpha\}$ belongs to the family; choosing $e = \pm e_i$, $e_1, ..., e_n$ be the basis orths in $\mathbf{R}^n$, we find that the intersection of the corresponding $2n$ half-spaces from the family is a bounded parallelotope, as required.

Thus, $\mathcal{F}$ satisfies the premise of Helley's theorem, and due to this theorem all the sets from the family have a point in common, let it be x. Let us prove that this is the point required by the Grunbaum theorem, i.e., that any closed half-space Π containing x is of μ-measure at least $1/(n + 1)$. Assume that it is not the case and there exists a closed half-space Π containing x of the mass $< 1/(n + 1)$. Translating the boundary hyperplane outside Π we obtain a half-space Π' which contains x in its interior; a small enough translation results in the mass of Π' being still less than $1/(n+1)$. It follows that the closed half-space $\mathbf{R}^n \backslash (\text{int } \Pi')$ is of mass $> n/(n+1)$ and therefore belongs to $\mathcal{F}$; at the same time, this half-space does not contain x, which is a contradiction. ∎

Exercise 1.3.12: without loss of generality we may assume that the functions f_i active at x_0 are $f_1, ..., f_k$. Let $V_i = \partial f_i(x_0)$, $i = 1, ..., k$; then V_i are closed and bounded convex sets. Therefore the convex hull of the union U of $V_1, ..., V_k$ is nothing but the set of all convex combinations of k vectors, i-th of the vectors belonging to U_i:

$$V = \{v = \lambda_1 v_1 + ... + \lambda_k v_k \mid \lambda_i \geq 0, \sum_i \lambda_i = 1, v_i \in U_i\}.$$

We should prove that V is $\partial f(x_0)$. It is clear that $V \subset \partial f(x_0)$; indeed, if

$$v = \sum_{i=1}^{k} \lambda_i v_i$$

is a convex combination of subgradients of $f_1, ..., f_k$ at x_0, then

$$f_i(x) \geq f_i(x_0) + (x - x_0)^T v_i = f(x_0) + (x - x_0)^T v_i$$

(we have used the fact that f_i are active at x_0), whence

$$f(x) \geq \max_{i=1,...,k} f_i(x) \geq \sum_{i=1}^{k} \lambda_i f_i(x) \geq \sum_{i=1}^{k} \lambda_i (f(x_0) + (x - x_0)^T v_i) = f(x_0) + (x - x_0)^T v,$$

so that v is a subgradient of f at x_0.

It remains to verify that every subgradient g of f at x_0 belongs to V. To this end let us first prove that V is closed. Indeed, V is the image of a closed and bounded (and therefore compact) set $V_1 \times V_2 \times ... \times V_k \times \Delta$ (Δ is the simplex $\{(\lambda_1, ..., \lambda_k) \mid \lambda_i \geq 0, \sum_i \lambda_i = 1\}$) under the continuous mapping

$$(v_1, ..., v_k, \lambda_1, ..., \lambda_k) \mapsto \sum_{i=1}^{k} \lambda_i v_i$$

and therefore is a compact and, consequently, a closed set.

Assume that a subgradient g of f at x_0 does not belong to the closed and convex set V and let us lead this assumption to a contradiction. Using the Separation Theorem for convex sets, we may find a nonzero h such that

$$g^T h > \max_{v \in V} g^T v,$$

whence, due to the origin of V,

$$g^T h > \max_{i=1,...,k} \max_{v \in \partial f_i(x_0)} v^T h = \max_{i=1,...,k} \lim_{t \to +0} t^{-1}(f_i(x_0 + th) - f_i(x_0)) =$$

$$= \max_{i=1,...,k} \lim_{t \to +0} t^{-1}(f_i(x_0 + th) - f(x_0)) \tag{14.2.9}$$

(we have taken into account that $f_i(x_0) = f(x_0), i = 1, ..., k$).

Now note that in a neighbourhood of x_0 f coincides with $\max_{i=1,...,k} f_i$ (recall that all f_i, $i = 1, ..., m$, are convex in a neighbourhood of x_0 and therefore continuous at x_0 and that $f_1, ..., f_k$ are exactly those of f_i which are active at x_0), so that g, being a subgradient of f at x_0, is also a subgradient of the function $\bar{f}(x) = \max_{i=1,...,k} f_i(x)$. It follows that

$$g^T h \leq \lim_{t \to +0} t^{-1}(\bar{f}(x_0 + th) - \bar{f}(x_0)) = \lim_{t \to +0} t^{-1}(\max_{i=1,...,k} f_i(x_0 + th) - f(x_0)) =$$

$$\lim_{t \to +0} \max_{i=1,...,k} t^{-1}(f_i(x_0 + th) - f(x_0)) = \max_{i=1,...,k} \lim_{t \to +0} t^{-1}(f_i(x_0) - f(x_0)),$$

which is a contradiction to (14.2.9).

Now assume that x_0 is a minimizer of f. Then 0 is a subgradient of f at x_0; according to already proved statement, this subgradient is a convex combination of certain subgradients

$v_1, ..., v_k$ of $f_1, ..., f_k$ at x_0; in view of the Caratheodory theorem, we may represent 0 as a convex combination of $l \leq n+1$ of these subgradients, let them be $v_{i_1}, ..., v_{i_l}$. Again using our main statement, we see that 0 is a subgradient at x_0 of the maximum of properly chosen $l \leq n+1$ functions of the family $f_1, ..., f_k$, as required by the second statement of the exercise.

Exercise 1.3.13: $r(x)$ clearly is a convex function of x; since all Q_i are bounded, $r(x) \to \infty$ as $|x| \to \infty$, so that a minimizer x^* of r does exist. According to the second statement of exercise 1.3.12, we may select $l \leq n+1$ of the functions $\mathrm{dist}(x, Q_i)$ in such a way that their maximum $\bar{r}(x)$ attains its minimum at x^*, the value of this maximum at x^* being exactly that one of r. But the minimal value of $\bar{r}$ clearly is 0 (since the $l \leq n+1$ sets which participate in forming $\bar{r}$ have a point in common); thus, $0 = \bar{r}(x^*) = r(x^*)$, so that x^* is a common point of all Q_i, as claimed.

Exercise 1.3.14: given a finite family of closed convex sets satisfying the premise of the Helley theorem, we clearly can find a large enough ball such that the family comprised of the intersections of the initial sets with the ball also satisfies the premise (it suffices to choose a point in each of intersections of $n+1$ members of the family and take the ball which covers the resulting finite set). Due to the statement given in exercise 1.3.13, the "truncated" sets (and therefore the initial ones) have a point in common.

Exercise 1.3.15: in the case of a finite family the conclusion is already proved (exercise 1.3.14). It remains to consider the case when the family is infinite; in this case, by assumption given in the premise of the theorem, there exists a finite subfamily $\mathcal{F}$ of our family such that the intersection V of the sets from the subfamily is bounded. Let α denote a finite subfamily of the initial family containing $\mathcal{F}$, and let F_α be the intersection of all sets from the subfamily α. Then F_α is a nonempty (due to already proved version of the theorem for finite families) and closed subset of V and is therefore a nonempty compact set. The sets F_α corresponding to all possible α clearly form a nested family of nonempty compacts ("nested" means that the intersection of finitely many sets of the family is again a set from the family); from the well-known theorem on compacts the members of such a family have a nonempty intersection, Q.E.D.

Exercise 1.3.16: without loss of generality we may assume that ϕ_i are linearly independent (no nontrivial linear combination of ϕ_i is identically zero), since otherwise we could replace the family $\{\phi_i\}$ by its maximal linearly independent subfamily without violating the linear span of the functions. Thus, we may assume that Φ is a n-dimensional linear space. Let $k = n+1$, $X^k = X \times X \times ... \times X$ (k direct factors), and let $x^k = (x_1, ..., x_k)$ denote a point from X^k. Set

$$r(x^k) = \min_{p \in \Phi} \| f - p \|_{\{x_1, ..., x_k\}, \infty}$$

be the minimal deviation of f from Φ on the set comprised of the points from the collection x^k. $r(x^k)$ clearly is the minimum of a family of continuous functions of x^k and therefore is upper semicontinuous function on the compact set X^k; therefore it attains its maximum, let it be δ', at certain point x^k_*.

The statement we should prove means exactly that $\delta' = \delta$; the inequality $\delta \geq \delta'$ is evident (if f can be approximated by an element of Φ, within the uniform error δ, on the whole X, it, of course, can be approximated within the same error at any k-point subset of X). What should be proved is the inverse inequality

$$\delta' \geq \delta.$$

To prove the inverse inequality, let us act as follows.

For any $x \in X$ let $H_x = \{p \in \Phi \mid \| f(x) - p(x) \| \leq \delta'\}$ be the set of all elements of Φ which approximate f within accuracy δ' at x. It is clear that H_x is a closed and convex subset of the $(k-1)$-dimensional linear space Φ. To prove that $\delta \geq \delta'$ it suffices to prove that all sets of the family $\{H_x\}_{x \in X}$ have a point in common, and to prove this latter statement we, of course, should use the Helley theorem. Let us verify the premise of the theorem. From the origin of δ' it follows that any k of the sets of the family $\{H_x\}_{x \in X}$ have a point in common. Thus, the only thing we need is to verify that if X is infinite, then there is a finitely many sets from the family with a bounded intersection. To this end let us note that any finite subset Y of X such that the restrictions of ϕ_i onto Y form a linearly independent system of functions gives us a finite subfamily $\{H_x\}_{x \in Y}$ of the initial family with a bounded intersection. Indeed, to say that the restrictions of ϕ_i onto Y are linearly independent is the same as to say that the mapping which maps an element from Φ into its restriction onto Y is a linear embedding; under this embedding the intersection of the sets H_x, $x \in Y$, is mapped into a bounded subset of a finite-dimensional space (comprised of all functions on Y taking values in E), and therefore the intersection itself also should be bounded.

Thus, we come to the necessity to verify the following fact (which is important in its own right):

given k linearly independent functions $\phi_i : X \to E$ taking values in a linear space E, one can find a finite subset Y in X such that the restrictions of ϕ_i onto Y are linearly independent; in fact this set can be chosen to have at most k points.

The proof is immediate (induction in k): for $k = 1$ the statement is evident (a linearly independent system comprised of a single function, i.e., a function which is not identically zero, has a linearly independent restriction on a properly chosen single point set, namely, on any point where the function is nonzero). Now the inductive step: given $k+1$ linearly independent on X functions $\phi_1, ..., \phi_{k+1}$, we, by the inductive hypothesis, can find a finite subset T (comprised of at most k points) such that the restrictions of $\phi_1, ..., \phi_k$ onto T are linearly independent. It is possible that the restrictions of $\phi_1, ..., \phi_{k+1}$ onto T also are linearly independent, and then we are done. Now assume that these $k+1$ restrictions are linearly dependent; since the first k of them are linearly independent, we have

$$\phi_{k+1}(x) = \lambda_1 \phi_1(x) + ... + \lambda_k \phi_k(x),\ x \in T.$$

This latter equality is violated at certain $x' \in X$ (since otherwise $\phi_1, ..., \phi_{k+1}$ would be linearly dependent); let us add x' to T. The extended set T' is comprised of at most $k+1$ points and the restrictions of ϕ_i, $i = 1, ..., k+1$ onto T' are linearly independent. Indeed, any nontrivial linear combination of ϕ_i which could be identically zero on T' could be rewritten as a representation of ϕ_{k+1} as a linear combination of ϕ_i, $i \leq k$ (since $T' \supset T$ and the restrictions of ϕ_i, $i \leq k$, onto T are linearly independent); this linear combination should be exactly the aforementioned one (again due to linear independence of ϕ_i, $i \leq k$, on T); but this latter equality is violated at $x' \in T'$, which gives the desired contradiction. ∎

Exercise 1.3.17: without loss of generality we may assume that the smallest points in H_1 and H_2 coincide with zero (since translating any of the sets we simply translate the sum of the sets), and let b_i denote the largest point of H_i, $i = 1, 2$.

Since H_1 contains 0, $H_1 + H_2$ contains H_2, and since H_2 contains b_2, $H_1 + H_2$ contains $b_2 + H_1$. Now, the set H_2 is to the left of b_2, and $b_2 + H_1$ is to the right of b_2, so that we have

found two subsets in $H_1 + H_2$ which are congruent to H_1, H_2, respectively, and have a single point in common, so that

$$\mathrm{Vol}_1(H_1 + H_2) \geq \mathrm{Vol}_1(H_1) + \mathrm{Vol}_1(H_2),$$

as required in (SA_1).

Exercise 1.3.18: we should prove that

$$(\mathrm{Vol}_k(H_1 + H_2))^{1/k} \geq (\mathrm{Vol}_k(H_1))^{1/k} + (\mathrm{Vol}_k(H_2))^{1/k}, \tag{14.2.10}$$

or, which is the same in view of (1.3.7), that

$$\left(\int_0^\infty s^{k-2}\mu_+(s)ds\right)^{1/k} \geq \left(\int_0^\infty s^{k-2}\mu_1(s)ds\right)^{1/k} + \left(\int_0^\infty s^{k-2}\mu_2(s)ds\right)^{1/k}, \tag{14.2.11}$$

where

$$\mu_i(s) = \mathrm{Vol}_1 T_i(s),\ T_i(s) = \{t \in T_i \mid r_{k-1}(H_i^t) \geq s\}), i = 1, 2,$$

((x, t) are the coordinates in $\mathbf{R}^k$, x being $(k-1)$-dimensional and t being scalar, T_i are the images of H_i under the projection onto the t-axis and H_i^t are the inverse images of the point $(0, ..., 0, t)$ in H_i under this projection),
and

$$\mu_+(s) = \underset{1}{\mathrm{Vol}}(\{t \in T_+ \mid r_{k-1}(H_+^t) \geq s\}),$$

T_+, H_+^t being similar objects for $H_+ = H_1 + H_2$.

(14.2.10) is evident in the case when $\mathrm{Vol}_k(H_1)$ or $\mathrm{Vol}_k(H_2)$ is zero, since H_+ contains translations of both H_1 and H_2 and therefore the volume of H_+ is not less than those of H_1 and H_2. Thus, it suffices to consider the case when both H_1 and H_2 are of positive volumes.

Note that from compactness and nonemptiness of H_i it follows immediately that T_i is a nonempty compact set and $r_{k-1}(H_i^t)$ is upper semicontinuous on T_i, $i = 1, 2$. Let s_i be the maximum of $r_{k-1}(H_i^t)$ over $t \in T_i$, $i = 1, 2$.

Let us first note that whenever $t' \in T_1, t'' \in T_2$, one has $t' + t'' \in T_+$ and

$$H_+^{t'+t''} \supset H_1^{t'} + H_2^{t''}, \tag{14.2.12}$$

whence, from the inductive hypothesis (SA_{k-1}),

$$r_{k-1}(H_1^{t'}) + r_{k-1}(H_2^{t''}) \geq r_{k-1}(H_+^{t'+t''}). \tag{14.2.13}$$

Let

$$\rho_i = \max_{t \in T_i} r_{n-1}(H_i^t),\ i = 1, 2,$$

and let

$$\alpha_i = \rho_i/(\rho_1 + \rho_2),$$

so that α_i are positive and $\alpha_1 + \alpha_2 = 1$.

Our next observation is that whenever $0 \le s \le \rho_1 + \rho_2$, one has

$$\mu_+(s) \ge \mu_1(\alpha_1 s) + \mu_2(\alpha_2 s) \tag{14.2.14}$$

Indeed, by definition

$$\mu_+(\sigma) = \mathrm{Vol}_1 T_+(\sigma),\ T_+(\sigma) = \{t \in T_+ \mid r_{n-1}(H_+^t) \ge \sigma\},$$

$$\mu_i(\sigma) = \mathrm{Vol}_1 T_i(\sigma),\ T_i(\sigma) = \{t \in T_i \mid r_{n-1}(H_i^t) \ge \sigma\},\ i = 1, 2.$$

Note that the sets $T_i(\cdot)$ are compact, since $r_{k-1}(H_i^t)$ are, as it was already mentioned, upper semicontinuous functions on the compact sets T_i, $i = 1, 2$.

To prove (14.2.14), let us fix $s \in [0, \rho_1 + \rho_2]$ and note that then $s_i = \alpha_i s \le \rho_i$, $i = 1, 2$, so that the sets $T_i(s_i)$ are nonempty. In view of (14.2.12)-(14.2.13) the sum of these sets is contained in $T_+(s_1 + s_2) = T_+(s)$; in view of already proved (SA_1) (exercise 1.3.17), we conclude that

$$\mu_+(s) \equiv \mathrm{Vol}_1(T_+(s)) \ge \mathrm{Vol}_1(T_1(\alpha_1 s) + T_2(\alpha_2 s)) \ge$$

$$\ge \mathrm{Vol}_1(T_1(\alpha_1 s)) + \mathrm{Vol}_1(T_2(\alpha_2 s)) = \mu_1(\alpha_1 s) + \mu_2(\alpha_2 s),$$

as required by (14.2.14).

From (1.3.7) we have

$$r_k^k(H_i) = (k-1)\int_0^\infty s^{k-2}\mu_i(s)ds \equiv (k-1)\int_0^{\rho_i} s^{k-2}\mu_i(s)ds,\ i = 1, 2,$$

(note that $\mu_i(s)$ clearly vanishes when $s > \rho_i$), whence

$$r_k^k(H_i) = (k-1)\alpha_i^{k-1}\int_0^{\rho_i/\alpha_i} s^{k-2}\mu_i(\alpha_i s)ds = (k-1)\alpha_i^{k-1}\int_0^{\rho_1+\rho_2} s^{k-2}\mu_i(\alpha_i s)ds.$$

Thus,

$$\int_0^{\rho_1+\rho_2} s^{k-2}\mu_i(\alpha_i s)ds = (k-1)^{-1} r_k^k(H_i)\alpha_i^{1-k}. \tag{14.2.15}$$

Besides this, from (1.3.7) it follows that

$$\int_0^{\rho_1+\rho_2} s^{k-2}\mu_+(s)ds \le (k-1)^{-1} r_k^k(H_+). \tag{14.2.16}$$

Taking sum of equalities (14.2.15) over $i = 1, 2$ and taking into account (14.2.14) and (14.2.16), we come to

$$r_k^k(H_1 + H_2) = (k-1)\int_0^\infty s^{k-2}\mu_+(s)ds \ge \sum_{i=1}^2 r_k^k(H_i)\alpha_i^{1-k},$$

and to prove the required relation

$$r_k(H_1 + H_2) \ge r_k(H_1) + r_k(H_2)$$

it suffices to use the elementary inequality

$$r^k/\alpha^{k-1} + s^k/\beta^{k-1} \ge (r+s)^k$$

whenever positive reals r, s, α, β are such that $\alpha + \beta = 1$ and $k \ge 2$. ∎

Remark 14.2.1 The elementary inequality mentioned in the last proof is a particular case of the general inequality

$$\sum_{i=1}^{m} |r_i|^k \alpha_i^{-k+1} \geq |\sum_{i=1}^{m} r_i|^k (\sum_{i=1}^{m} \alpha_i)^{-k+1}$$

(α_i are positive, r_i are real and $k \geq 2$).

You are welcome to regard the proof of this general inequality as an additional exercise.

Solutions to Section 2.4

Exercise 2.4.5: when taking a weighted sum of the inequalities (14.2.7) and (14.2.8), we may take the first weight being equal to 1 (this is equivalent to certain normalization of the resulting inequality, and the normalization, of course, does not influence the volume of the ellipsoid defined by the inequality). Let $2\lambda \geq 0$ be the weight for the second inequality; the "combined" inequality associated with this weight clearly is

$$x^T(I + 2\lambda ee^T)x - 2\lambda(1+\alpha)e^T x + 2\lambda\alpha \leq 1;$$

the center c of the ellipsoid is given by

$$(I + 2\lambda ee^T)c = \lambda(1+\alpha)e,$$

whence

$$c = \frac{\lambda(1+\alpha)}{1+2\lambda}e.$$

The inequality rewritten with respect to the center is

$$(x-c)^T(I+2\lambda ee^T)(x-c) \leq 1 - 2\lambda\alpha + c^T(I+2\lambda ee^T)c =$$

$$= \frac{(1-2\lambda\alpha)(1+2\lambda)+\lambda^2(1+\alpha)^2}{1+2\lambda} \equiv b(\lambda) \equiv$$

$$\equiv \frac{s^2(\lambda)}{1+2\lambda}, \quad s(\lambda) = 1 + (1-\alpha)\lambda.$$

The corresponding ellipsoid is given by

$$W_\lambda = \{x \mid (x-c)^T A_\lambda (x-c) \leq 1\}, \quad A_\lambda = b^{-1}(\lambda)(I + 2\lambda ee^T);$$

recall that this ellipsoid contains V_α, since the inequality defining the ellipsoid is a weighted sum of the inequalities (14.2.7) and (14.2.8), the weights being 1 and 2λ.

Due to exercise 2.4.2, the volume of the ellipsoid W_λ is given by

$$\mathrm{vol}(W_\lambda) = \mathrm{Det}\left(A_\lambda^{-1/2}\right) = \left(\frac{b^n(\lambda)}{1+2\lambda}\right)^{1/2} =$$

$$= \left\{\frac{s^{2n}(\lambda)}{(1+2\lambda)^{n+1}}\right\}^{1/2} \equiv v^{1/2}(\lambda).$$

Let us minimize the function $v(\lambda)$; we should start with finding its critical points given by the equation

$$v'(\lambda) = 0,$$

i.e., by

$$2ns^{2n-1}(\lambda)s'(\lambda)(1+2\lambda)^{n+1} - 2(n+1)(1+2\lambda)^n s^{2n}(\lambda) = 0;$$

after immediate simplifications we come to a linear equation which gives

$$\lambda = \frac{1+n\alpha}{(n-1)(1-\alpha)}.$$

In the range $(-1/n, 1)$ of values of α the right hand side turns out to be the minimizer of $v(\lambda)$ over $\lambda \geq 0$.

For the optimal value of λ the center of the ellipsoid becomes

$$c = \frac{1+n\alpha}{1+n},$$

and the volume of it is

$$\mathrm{vol}(W_\lambda) = v^{1/2}(\lambda) = \frac{s^n(\lambda)}{(1+2\lambda)^{(n+1)/2}} = \left\{\frac{n^2}{n^2-1}\right\}^{n/2} \sqrt{\frac{n-1}{n+1}}(1-\alpha^2)^{(n-1)/2}(1-\alpha).$$

It is easily seen that the right hand side in the latter equality is a decreasing function of $\alpha \in [-1/n, 1]$ and that the function is equal to 1 at the point $\alpha = -1/n$. Thus, the ellipsoid given by the above reasoning is, in the case of $1 > \alpha > -1/n$, of smaller volume than the unit Euclidean ball, as claimed.

The fact that the ellipsoid W_λ can be represented as the image of the unit Euclidean ball under the affine mapping indicated in exercise 2.4.5 is quite straightforward. ∎

Exercise 2.4.6: All points of V^α clearly satisfy the following pair of quadratic inequalities:

$$x^T x \leq 1;$$

$$x^T e e^T x \leq \alpha^2.$$

Taking weighted sum of the inequalities with the weights 1 and $\lambda \geq 0$, we observe that V^α is, for every $\lambda \geq 0$, covered by the ellipsoid

$$W^\lambda = \{x \mid x^T A(\lambda) x \leq 1\}, \quad A(\lambda) = (1+\lambda\alpha^2)^{-1}(I + \lambda e e^T).$$

Let us find the best covering of this type, i.e., let us minimize in $\lambda \geq 0$ the quantity

$$\mathrm{vol}(W^\lambda) = \mathrm{Det}^{-1/2} A(\lambda) = \left((1+\lambda\alpha^2)^n/(1+\lambda)\right)^{1/2}.$$

Am immediate calculation results in

$$\lambda = \frac{1-n\alpha^2}{(n-1)\alpha^2}$$

(note that in our range $|\alpha| < 1/\sqrt{n}$ of values of α λ is positive). The indicated choice of λ results in

$$\mathrm{vol}(W^\lambda) = \alpha\sqrt{n}\left\{\frac{n(1-\alpha^2)}{n-1}\right\}^{(n-1)/2};$$

the derivative of the right hand side in α is positive when $0 \leq \alpha < 1/\sqrt{n}$ and the right hand side is equal to 1 when $\alpha = 1/\sqrt{n}$, so that the right hand side is < 1 for $0 \leq \alpha < 1/\sqrt{n}$. ∎

Exercise 2.4.7:

(i): Let W be (an) extremal outer ellipsoid associated with Q. As we know, there exists an invertible affine transformation of the space which maps this ellipsoid onto the centered at 0 unit Euclidean ball V, and since the extremal ellipsoids "accompany" the corresponding bodies

when the latter are subject to invertible affine transformations, V is (an) extremal outer ellipsoid associated with the image Q' of Q under this transformation. What we should prove is that the concentric to V Euclidean ball V' of the radius $1/n$ is contained in Q'. Assume that it is not the case; then Q' does not contain the interior of V' (indeed, Q' is closed, and if containing the interior of V', it would contain V' as well). Thus, there exists a point u such that $|u|_2 < 1/n$ and $u \notin Q'$. Due to the Separation Theorem for convex sets, there exists a linear functional

$$e^T x,$$

e being a unit vector, such that

$$e^T u \leq e^T z, z \in Q'.$$

Besides this, we know that Q' is contained in the unit Euclidean ball V. Thus, Q' is contained in the spherical hat

$$V_\alpha = \{x \in V \mid e^T x \geq \alpha \equiv e^T u\};$$

since e is a unit vectors and $|u|_2 < 1/n$, we have $\alpha > -1/n$. Due to this latter observation, V_α can be covered by an ellipsoid of volume strictly less than that on of V (exercise 2.4.5), and this ellipsoid, of course, covers $Q' \subset V_\alpha$ as well. This is the desired contradiction, since V is an extremal outer ellipsoid for Q'.

(ii) Consider all ellipsoids centered at c and containing Q; it is immediately seen that this family of ellipsoids is nonempty and contains at least one ellipsoid W of the smallest volume. (Of course, if we knew that the extremal outer ellipsoid of Q is unique, we could use the symmetry reasons to prove that W is nothing but the extremal outer ellipsoid associated with Q.) Same as above, let us perform an invertible affine transformation which maps W onto the centered at 0 unit Euclidean ball V, so that V is (an) ellipsoid of minimal volume among those centered at the origin ellipsoids which contain the image, Q', of Q under the above transformation. Let us prove that the centered at the origin Euclidean ball V'' of the radius $1/\sqrt{n}$ is contained in Q'; this is exactly what we need.

Assume, on contrary, that V'' is not contained in Q'. Same as above, it means that there exists a point u, $|u|_2 < 1/\sqrt{n}$, in the space which does not belong to Q', and, consequently, there is a linear functional

$$e^T x,$$

e being a unit vector, such that

$$e^T u \leq e^T z, z \in Q'.$$

Since Q' is central symmetric with respect to the origin (recall that Q was central symmetric with respect to the center of W, by definition of W), the latter inequality implies that

$$-e^T u \geq e^T z, z \in Q';$$

since also $Q' \subset V$, we conclude that Q' is contained in the symmetric spherical stripe

$$V^\alpha = \{x \in V \mid |e^T x| \leq \alpha \equiv |e^T u|\}.$$

Since $|u|_2 < 1/\sqrt{n}$ and $|e|_2 = 1$, we have $\alpha < 1/\sqrt{n}$, and the statement given by exercise 2.4.6 says to us that V^α (and therefore $Q' \subset V^\alpha$) can be covered by a centered at the origin ellipsoid of volume less than that one of V, which is the desired contradiction. ∎

Exercise 2.4.8: according to item 3. of the list of consequences of the Fritz John Theorem, there exists a Euclidean norm $|\cdot|_A$ such that

$$n^{-1/2}|x|_A \le \| x \| \le |x|_A,\ x \in \mathbf{R}^n. \tag{14.2.17}$$

From the left inequality it follows that

$$\mathcal{E}\{|\xi_i|_A^2\} \le n s_i^2,$$

whence, due to the Tschebyshev inequality for a Euclidean norm,

$$\mathcal{E}\{|\Xi_N|_A^2\} \le n \sum_{i=1}^N s_i^2;$$

the left hand side in this inequality, due to the right inequality in (14.2.17), is $\ge S_N^2$, and we are done.

Exercise 2.4.9: let us look what is going on in the situation described in the hint to the exercise. Here, by construction, we have $s_i = 1$ (the $|\cdot|_1$ - norm of all ξ_i is 1). Now, let us look at k-th entry Ξ_N^k in the sum Ξ_N. This is a sum of iid random reals with zero mean and the dispersion $\sigma^2 = n^{-1}$; due to the Central Limit Theorem, the distribution of the normalized entry

$$\zeta_N = \Xi_N^k \sqrt{n/N}$$

converges, as $N \to \infty$, to the standard Gaussian distribution; it is easily seen that

$$\lim_{N\to\infty} \mathcal{E}\{|\zeta_N|\} \frac{2}{\sqrt{2\pi}} \int_0^\infty t \exp\{-t^2/2\} dt = \frac{2}{\sqrt{2\pi}}.$$

Consequently,

$$\mathcal{E}\{|\Xi_N^k|\} = \sqrt{N/n} \frac{2}{\sqrt{2\pi}} (1 + o(1)),\ N \to \infty,$$

whence, due to the choice of the norm,

$$\mathcal{E}\{|\Xi_N|_1\} = \sqrt{Nn} \frac{2}{\sqrt{2\pi}} (1 + o(1)),\ N \to \infty;$$

it remains to note that, due to the Cauchy inequality,

$$\mathcal{E}\{|\Xi_N|_1^2\} \ge \left(\mathcal{E}\{|\Xi_N|_1\}\right)^2.$$

Exercise 2.4.10: it is clearly seen that

$$|x|_\infty = \lim_{p\to\infty} |x|_p;$$

therefore it suffices to demonstrate (2.4.8) for the case of finite $p \ge p' > 1$. Due to the homogeneity reasons, we may restrict ourselves with the case of $|x|_{p'} = 1$. Let $\alpha_i = |x_i|^{p'}$, so that $\sum_i \alpha_i = 1$ and $\alpha_i \ge 0$. The quantity $|x|_p$ is nothing but

$$\left(\sum_{i=1}^n \alpha_i^{p/p'} \right)^{1/p}.$$

The function

$$f(\alpha) = \sum_{i=1}^{n} \alpha_i^{p/p'}$$

is convex on the simplex $\{\alpha \geq 0, \sum_i \alpha_i = 1\}$, and due to its symmetry it attains its minimum over the simplex at its barycenter (i.e., when $\alpha_i = 1/n$, $i = 1, ..., n$), and attains its maximum at a vertex of the simplex. Thus,

$$1 \geq f(\alpha) \geq n(1/n)^{p/p'}$$

and consequently

$$1 \geq |x|_p \geq n^{1/p-1/p'}$$

whenever $|x|_{p'} = 1$, as claimed.

Exercise 2.4.11: it is convenient to denote, for a given $p \in (2, \infty)$,

$$\pi(x) = |x|_p,\ x \in \mathbf{R}^n,\ f(x) = \pi^2(x),\ q = \frac{p}{p-1}$$

(so that $1/p + 1/q = 1$).

The function $f(x)$ clearly is continuously differentiable, with the derivative at a point $x \neq 0$ given by

$$f'(x) = 2\pi(x)\zeta(x) \equiv 2|x|_p\zeta(x), \tag{14.2.18}$$

where the vector $\zeta(x)$ is defined as follows:

$$\zeta_j(x) = \mathrm{sign}(\bar{x}_j)|\bar{x}_j|^{p-1}, \quad \bar{x} = \frac{x}{|x|_p}.$$

1^0. Let us first prove that if u and v are vectors of the unit $|\cdot|_p$-norm, then

$$|\zeta(u) - \zeta(v)|_q \leq (p-1)2^{(p-2)/(p-1)}|u - v|_p. \tag{14.2.19}$$

Indeed, let

$$\eta_j = \max\{|u|_j, |v|_j\};$$

then one clearly has

$$|\eta|_p \leq 2^{1/p} \tag{14.2.20}$$

(recall that u and v are of the unit l_p-norm). On the other hand, since u and v are of the unit l_p-norm, we have

$$\zeta_j(u) = |u_j|^{p-1}\,\mathrm{sign}\,u_j,\ \zeta_j(v) = |v_j|^{p-1}\,\mathrm{sign}\,v_j,$$

and since the absolute value of the derivative of the function $|t|^{p-1}\,\mathrm{sign}\,t$ on the segment $-\eta_j \leq t \leq \eta_j$ (which covers both u_j and v_j) does not exceed $(p-1)\eta_j^{p-2}$ (recall that $p \geq 2$), we come to

$$|\zeta_j(u) - \zeta_j(v)| \leq (p-1)\eta_j^{p-2}|u_j - v_j|.$$

It follows that if

$$\beta = p/q \equiv p - 1 \geq 1,\ \alpha = \beta/(\beta - 1) = \frac{p-1}{p-2} = \frac{p}{(p-2)q},$$

then

$$(p-1)^{-q}\sum_{j=1}^{n}|\zeta_j(u)-\zeta_j(v)|^q \le \sum_{j=1}^{n}\eta_j^{(p-2)q}|u_j-v_j|^q \le$$

$$\le \left(\sum_{j=1}^{n}\eta_j^{(p-2)q\alpha}\right)^{1/\alpha}\left(\sum_{j=1}^{n}|u_j-v_j|^{q\beta}\right)^{1/\beta} = |\eta|_p^{p(p-2)/(p-1)}|u-v|_p^{p/(p-1)}$$

(we have used the Holder inequality $\sum_j a_j b_j \le (\sum_j a_j^\alpha)^{1/\alpha}(\sum_j b_j^\beta)^{1/\beta}$ which holds true for all nonnegative a_j, b_j and positive α, β such that $1/\alpha + 1/\beta = 1$). We conclude form the above chain of inequalities that

$$|\zeta(u)-\zeta(v)|_q \le (p-1)|\eta|_p^{p-2}|u-v|_p,$$

which combined with (14.2.20) results in (14.2.19).

2^0. Now let x and y be a pair of nonzero vectors from $\mathbf{R}^n$ and let $\bar{x}$, $\bar{y}$ be their normalizations of the unit $|\cdot|_p$-norm:

$$\bar{x} = \frac{x}{\pi(x)}, \quad \bar{y} = \frac{y}{\pi(y)}.$$

We have

$$|\bar{x}-\bar{y}|_p = |\frac{\pi(y)-\pi(x)}{\pi(x)\pi(y)}x + \frac{1}{\pi(y)}(x-y)|_p \le \frac{|\pi(y)-\pi(x)|}{\pi(x)\pi(y)}\pi(x) + \frac{1}{\pi(y)}|x-y|_p \le 2\frac{|x-y|_p}{|y|_p},$$

whence, in view of (14.2.19),

$$|\zeta(x)-\zeta(y)|_q \le (p-1)2^{(2p-3)/(p-1)}|x-y|_p/|y|_p$$

(we have taken into account that, by definition of $\zeta(\cdot)$, $\zeta(x) = \zeta(\bar{x})$, $\zeta(y) = \zeta(\bar{y})$). The resulting inequality combined with (14.2.18) results in

$$|f'(x)-f'(y)|_q = 2||x|_p\zeta(x) - |y|_p\zeta(y)|_q \le 2\left(|y|_p|\zeta(x)-\zeta(y)|_q + ||x|_p - |y|_p||\zeta(x)|_q\right) \le$$

$$\le 2\left((p-1)2^{(2p-3)/(p-1)}|x-y|_p + |x-y|_p|\zeta(x)|_q\right) = 2((p-1)2^{(2p-3)/(p-1)}+1)|x-y|_p$$

(we have taken into account that $|\zeta(x)|_q = 1$, which is an immediate consequence of the definition of $\zeta(\cdot)$). Thus,

$$|f'(x)-f'(y)|_q \le 2\omega_f|x-y|_p, \quad \omega_f = (p-1)2^{(2p-3)/(p-1)}+1. \tag{14.2.21}$$

When deriving this inequality, x and y were assumed to be nonzero, but actually the inequality is valid for all x and y, due to the continuity of both the sides in x, y.

3^0. It remains to note that

$$|f(x+h)-f(x)-h^Tf'(x)| = |\int_0^1 (f'(x+th)-f'(x))^T h dt| \le \int_0^1 |f(x+th)-f'(x)|_q|h|_p dt$$

(we have used the Holder inequality), whence. in view of (14.2.21),

$$|f(x+h)-f(x)-h^Tf'(x)| \le \omega_f|h|_p^2 = \omega_f f(h).$$

as claimed.

Exercise 2.4.12: combining (2.4.9) and (2.4.17) (where one should set $p = 2 \ln n$), we come to

$$\mathcal{E}\{|\Xi_N|_\infty^2\} \leq \mathcal{E}\{|\Xi_N|_p^2\} \leq (8 \ln n - 3) \sum_{i=1}^{n} \mathcal{E}\{|\xi_i|_p^2\} \leq$$

$$(8 \ln n - 3) \exp\{1\} \sum_{i=1}^{N} \mathcal{E}\{|\xi_i|_\infty^2\},$$

Q.E.D.

Solutions to Section 3.3

Exercise 3.3.4: the statement in question is evident in the case of $h'(x) = 0$, since then, due to the regularity assumption, x is a minimizer of h over G and $L_h^*(x) = \emptyset$. Now let $h'(x) \neq 0$. It suffices to prove that $L_h(x) \subset \operatorname{cl} \Pi_h(x)$ (since $L_h^*(x)$ is an open in G subset of the set $L_h(x)$ and the boundary hyperplane of Π_h passes through an interior point of G). The set $L_h(x)$ is convex and contains x; consequently, if there would be a point y in $L_h(x)$ which is not contained in $\operatorname{cl} \Pi_h(x)$, then the whole segment $[x, y]$ would belong to $L_h(x)$, and h would be $\leq h(x)$ on the segment; this is impossible, since the derivative of h at x in the direction $y - x$ is positive.

Solutions to Section 4.3

Exercise 4.3.3: The point w_i where the hyperplane

$$\Pi = \{x \in \mathbf{R}^n \mid g^T x = 1\}$$

intersects i-th coordinate axis is nothing but $\gamma_i^{-1} e_i$ (if $\gamma_i = 0$, then Π does not intersect the i-th axis). Let I be the set of indices of those points w_i which are well-defined and are inside Δ, and let J be the set of remaining indices from $\{1, ..., n\}$; thus,

$$I = \{i : 1 \le i \le n, \gamma_i > 1\}.$$

Now, to prove that the simplex Δ_ϑ contains $\bar{\Delta}$ is the same as to verify that the simplex contains all vertices of the latter polytope. It is easily seen that the vertices of $\bar{\Delta}$ are those vertices of Δ which are inside $\bar{\Delta}$ and also the points where Π intersects the edges of Δ (why?) Thus, the vertices of $\bar{\Delta}$ are as follows:

first, the origin is a vertex; it clearly belongs to Δ_ϑ;

second, the vertices e_j of the initial simplex Δ with indices $j \in J$; since for $j \in J$ one has $\gamma_j \le 1$ and, consequently, $1/(1 - \vartheta + \vartheta\gamma_j) \ge 1$, all these vertices belong to Δ_ϑ;

third, any pair of indices $i \in I$, $j \in J$, produces a vertex in $\bar{\Delta}$, namely, the point w_{ij} where the hyperplane Π intersects the edge $\Gamma_{ij} = \{x_i + x_j = 1, x \ge 0\}$ of the simplex Δ. The i-th and the j-th coordinates of w_{ij} are given by

$$\xi_i = \frac{1 - \gamma_j}{\gamma_i - \gamma_j}, \quad \xi_j = \frac{\gamma_i - 1}{\gamma_i - \gamma_j}, \tag{14.2.22}$$

and the remaining coordinates of w_{ij} are zeros. In order to prove that w_{ij} belongs to Δ_ϑ it suffices to verify that

$$(1 - \vartheta + \vartheta\gamma_i)\xi_i + (1 - \vartheta + \vartheta\gamma_j)\xi_j \le 1 \tag{14.2.23}$$

(indeed, the intersection of Δ_ϑ with the coordinate plane spanned by the axes i and j is nothing but the triangle

$$x_i \ge 0;\ x_j \ge 0;\ (1 - \vartheta + \vartheta\gamma_i)x_i + (1 - \vartheta + \vartheta\gamma_j)x_j \le 1$$

in the plane). Inequality (14.2.23) is an immediate consequence of (14.2.22). Thus, Δ_ϑ does contain $\bar{\Delta}$.

It remails to evaluate the ratio of volumes of Δ and Δ_ϑ. Since both the simplexes differ only in lengths of the edges outgoing from their common vertex 0, not in the directions of the edges, we have

$$\psi(\vartheta) \equiv \frac{\mathrm{Vol}(\Delta)}{\mathrm{Vol}(\Delta_\vartheta)} = \prod_{i=1}^{n}(1 - \vartheta + \vartheta\gamma_i) \equiv \exp\{\phi(\vartheta, g)\},$$

where

$$\phi(\vartheta, t) = \sum_{i=1}^{n} \ln(1 - \vartheta + \vartheta t_i),$$

$t = (t_1, ..., t_n) \ge 0$. Since the function $\phi(\vartheta, t)$ is concave in $t = (t_1, ..., t_n) \in \Gamma = \{t \ge 0 \mid \sum_i t_i = n + 1\}$, it attains its minimum over Γ at an extreme point of this set, i.e., when one of t_i is $n + 1$

and the remaining are zeros; the minimum is equal to $(n-1)\ln(1-\vartheta)+\ln(1+n\vartheta)$. Since $g \in \Gamma$ (due to $g \geq 0$, $g^T c = 1$), we conclude that

$$\psi(\vartheta) \geq (1-\vartheta)^{n-1}(1+n\vartheta).$$

Maximizing the right hand side of the latter relation in $\vartheta \in (0,1)$, we come to $\vartheta = n^{-2}$. With this value of ϑ one has

$$\psi(\vartheta) \geq \left(1-\frac{1}{n^2}\right)^{n-1}\left(1+\frac{1}{n}\right) \equiv \chi^{-n}(n) > 1 + O(\frac{1}{n^2}).$$

Solutions to Section 5.5

Exercise 5.5.4: Let B' and B'' be two positive definite symmetric $n \times n$ matrices and let $\alpha \in (0,1)$. We should prove that

$$-\ln \mathrm{Det}\,(\alpha B' + (1-\alpha)B'') \le \alpha[-\ln \mathrm{Det}\ B'] + (1-\alpha)[-\ln \mathrm{Det}\ B'']. \tag{14.2.24}$$

The linear transformation $X \mapsto X' = (B')^{-1/2} X (B')^{-1/2}$ maps a symmetric positive definite X into the matrix X' of the same type and varies $F(X)$ by an additive constant (equal to $-F(B')$); this mapping transforms B' into the unit matrix I, B'' into certain matrix C and $B''' \equiv \alpha B' + (1-\alpha)B''$ into $C''' = \alpha I + (1-\alpha)C$. When passing from (B', B'', B''') to (I, C, C'''), both sides of (14.2.24) varies by the same constant, so that to establish (14.2.24) is the same as to check that

$$-\ln \mathrm{Det}\,(\alpha I + (1-\alpha)C) \le -\alpha \ln \mathrm{Det}\ I - (1-\alpha)\ln \mathrm{Det}\ C. \tag{14.2.25}$$

Let λ_i, $i = 1, ..., n$, be the eigenvalues of C taken with their multiplicities; the left hand side of (14.2.25) is $-\sum_1^n \ln(\alpha + (1-\alpha)\lambda_i)$, the right hand side is $-\alpha \sum_1^n \ln(1) - (1-\alpha)\sum_1^n \ln((1-\alpha)\lambda_i$, and (14.2.25) is an immediate consequence of the fact that the function $-\ln s$ is convex on the positive ray. ∎

Solutions to Section 6.3

Exercise 6.3.2: we should prove that the system of linear inequalities

$$x^T(\xi - y) \geq 0,\ (x,\xi) \in \mathcal{G}(F) \tag{14.2.26}$$

with unknown y has a solution, provided that $0 \in \text{int Dom}\, F$. To this end let us first prove that any finite subsystem

$$x_i^T(\xi_i - y) \geq 0, i = 1, ..., N,$$

of the system in question has a solution. On contrary, let the indicated finite subsystem be unsolvable. Then there exist Lagrange multipliers λ_i, nonnegative with the unit sum, such that the linear form

$$\sum_i \lambda_i x_i^T(\xi_i - y)$$

is negative everywhere, which means that

$$\sum_i \lambda_i x_i = 0,\ \sum_i \lambda_i x_i^T \xi_i < 0.$$

On the other hand,

$$(x_i - x_j)^T(\xi_i - \xi_j) \geq 0,\ i, j = 1, ..., N,$$

and taking average of these inequalities with the weights $\lambda_i\lambda_j$, we come to

$$2\sum_i \lambda_i x_i^T \xi_i \geq 0$$

(we have used the fact that $\sum_i \lambda_i x_i = 0$), which is the desired contradiction.

It remains to prove that one of the finite subsystems of (14.2.26) has a bounded solution set (after it is proved, we may refer to the general fact: a family of closed sets in $\mathbf{R}^n$ with any finite subfamily possessing a nonempty intersection also possesses a nonempty intersection, provided that there is a finite subfamily with compact intersection; to prove the indicated fact, you may immediately reduce the situation to that one with a family comprised of compact sets; and a nested family of compact sets does possess a nonempty intersection). Since $0 \in \text{int Dom}\, F$, we can find a finite set $\{x_i \in \text{Dom}\, F\}_{i=1,...,N}$ such that 0 belongs to the interior of the convex hull of x_i, $i = 1, ..., N$; choosing $\xi_i \in F(x_i)$, we come to a finite subsystem of (14.2.26) which says that the inner products of y and all x_i are bounded from above by given constants; consequently, $y^T x \leq a$ for some finite a and all x from a convex hull of x_i, i.e., for all x from a small neighbourhood of 0; of course, the solution set of the subsystem is bounded. ∎

Exercise 6.3.4:

2) $\Leftrightarrow$ 3): this is the standard fact of Analysis which has nothing in common with convexity: a function $f : X \mapsto \mathbf{R}^+$ which is not identically $+\infty$ (X is a metric space) is lower semicontinuous if and only if its graph $\{(t,x) \in \mathbf{R} \times X \mid x \in \text{Dom}\, f, t \geq f(x)\}$ is closed, and the proof is given by a quite straightforward verification of definitions.

3) $\Rightarrow$ 4): the right hand side $\bar{f}(x)$ of the relation in question clearly is a convex lower semicontinuous function (the latter fact follows from already proved equivalence between 2) and 3): the graph of f clearly is closed, since it is an intersection of closed half-spaces); due to the basic theorem on existence of a subgradient of a convex function everywhere at the interior of its

domain, we have $f(x) = \bar{f}(x)$, $x \in \operatorname{int} \operatorname{Dom} f$, and of course $f(x) \geq \bar{f}(x)$ everywhere (due to the definition of a subgradient). Now assume that x is a boundary point of $\operatorname{Dom} f$ and let x' be an interior point of the domain. Then all points of the half-segment $[x', x)$ are the interior points of $\operatorname{Dom} f$ (since this set is convex), and both f and $\bar{f}$ are convex and lower semicontinuous on $[x', x]$ and coincide with each other and are finite on $[x', x)$. A convex function on a segment (taking values in $\mathbf{R}^+$ and finite on the interior of the segment) clearly is upper semicontinuous on the segment (why?), so that f and $\bar{f}$ are both lower- and upper-semicontinuous on $[x', x]$ and therefore are continuous on the segment; since they coincide on its interior, they coincide at the whole segment and, in particular, at x.

Thus, f and $\bar{f}$ coincide on the closure of $\operatorname{Dom} f$. To verify that the functions coincide everywhere, it remains to prove that if $x \notin \operatorname{cl} \operatorname{Dom} f$, then $\bar{f}(x) = \infty$. Indeed, let it not be the case, and let $\bar{x}$ be an interior point of $\operatorname{Dom} f$, and let x^+ be the point where the segment $[\bar{x}, x]$ intersects the boundary of $\operatorname{Dom} f$. The function $\bar{f}$ is, as we know, lower semicontinuous and finite on $[\bar{x}, x]$ and therefore, as we just have seen, is continuous on the segment, and, in particular, is bounded on it. Now let V be a ball centered at $\bar{x}$ and contained in $\operatorname{int} \operatorname{Dom} f$ and W be the conic hull of V and x. Let also $x_i \in [\bar{x}, x^+)$ be a sequence converging to x^+ and $g_i(u) = f(x_i) + \xi_i^T(x - x_i)$, $\xi_i \in \partial f(x_i)$. The affine functions g_i are bounded from above at x (since they are $\leq \bar{f}$) and are uniformly bounded from above on V (since f is bounded from above on this ball). It follows that the functions are uniformly bounded from above on W, and since the sequence $g_i(x_i)$ is bounded (we already know that f is finite and continuous on $[\bar{x}, x^+]$) and x_i converge to an interior point of W, we immediately conclude that $g_i(\cdot)$ are uniformly bounded on W (why?). This observation implies that the sequence ∇g_i is bounded, and passing to subsequences we may assume that $\nabla g_i \to e$ as $i \to \infty$. Besides this we already know that $g_i(x_i) = f(x_i) \to f(x^+)$, $i \to \infty$; since $f(u) \geq g_i^T(x_i)(u - x_i) + g_i(x_i)$, $u \in \mathbf{R}^n$, we come to

$$f(u) \geq e^T(u - x^+) + f(x^+).$$

Thus, $x^+ \in \operatorname{Dom} f$ and $\partial f(x^+) \neq \emptyset$. Now, from the Separation Theorem for convex sets it follows that there exists a nonzero functional h such that

$$h^T(u - x^+) \leq 0, \ u \in \operatorname{Dom} f;$$

one clearly has

$$e + th \in \partial f(x^+), t \geq 0.$$

Since x is an interior point of $\operatorname{Dom} f$ and h is nonzero, we have

$$h^T(x - x^+) < 0$$

and consequently $h^T(x - x^+) > 0$; we come to

$$+\infty > \bar{f}(x) \geq f(x^+) + (e + th)^T(x - x^+)$$

for all $t > 0$ (the concluding inequality follows from definition of $\bar{f}$), which clearly is a contradiction to $h^T(x - x^+) > 0$. Thus, the implication 3) $\Rightarrow$ 4) is proved.

4) $\Rightarrow$ 2): the right hand side in 4) clearly is lower semicontinuous on $\mathbf{R}^n$.
Thus, we have proved the equivalence of 2), 3), 4).

2) $\Rightarrow$ 1): assume that f is lower semicontinuous, but f' is not maximal monotone, so that there exists $(x,y) \notin \mathcal{G}(f')$: $(x-x')^T(y-y') \geq 0$ for all $(x',y') \in \mathcal{G}(f')$. Passing from $f(u)$ to $f(u)-y^Tu$, we can reduce the situation to the case when $y=0$; by translation we may also assume that $0 \in \operatorname{int}\operatorname{Dom} f$. Let us look what happens with f at the segment $[0,x]$. Certain part Δ of this segment (of the type $[0,x^+)$ or $[0,x^+]$) is contained in $\operatorname{Dom} f$, and the remaining part (if any) is outside $\operatorname{Dom} f$. If $x' \in [0,x^+)$, then $(x-x')^Ty' \leq 0$ for all $y' \in f'(x')$, so that f does not increase on $[0,x^+)$; consequently, $x^+ \in \operatorname{Dom} f$ and f is finite at x^+ due to lower semicontinuity of f on $\mathbf{R}^n$; since f is convex, finite and lower semicontinuous on $[0,x^+]$, it is a continuous finite function on $[0,x^+]$. Let us prove that $x^+=x$. Indeed, let $x_i \in [0,x^+)$ converge to x^+, and let $g_i(u)=f(x_i)+\xi_i(u-x_i)$, $\xi_i \in f'(x_i)$. The affine forms g_i decrease in the direction x (we already know that f does not increase on $[0,x^+)$) and are bounded from above at 0, and therefore they are bounded from above at x. Besides this, these forms are uniformly bounded from above in a ball W centered at 0 and $g_i(x_i)$ possesses finite limit as $i \to \infty$. From these observations, exactly as when proving the implication 3)$\Rightarrow$ 4), we may conclude that ∇g_i form a bounded sequence and, consequently, that $f'(x^+)$ is nonempty. Since x^+ is a boundary point of $\operatorname{Dom} f$, $f'(x^+)$ contains functionals with arbitrary large derivative in the direction x, or, which is the same, in the direction $x-x^+$ (see the reasoning used to prove the implication 3) $\Rightarrow$ 4)), which is impossible due to $(x-x')^Ty' \leq 0$ for all $x' \in \operatorname{Dom} f'$ and all $y' \in f'(x')$ (look what happens when $x'=x^+$).

Thus, we come to $x \in \operatorname{Dom} f$, and, as we remember, $(x-x')^Ty' \leq 0$ for all $x' \in \operatorname{Dom} f'$ and all $y' \in f'(x')$. We conclude that if $x' \in \operatorname{int}\operatorname{Dom} f$, then the convex function f is finite on the segment $[x',x]$ and its derivative along the direction $x-x'$ of the segment is nonpositive at any point from $[x',x)$; due to lower semicontinuity of f we conclude that it is continuous on the segment and $f(x) \leq f(x')$. Thus, $f(x) \leq f(x')$ for any $x' \in \operatorname{int}\operatorname{Dom} f$. Let us prove that actually $f(x) \leq f(x')$ for *any* x'. The inequality is evident if $x' \notin \operatorname{Dom} f$, since then $f(x')=+\infty$. Now, if $x' \in \operatorname{Dom} f$, then, as we have seen, f is continuous and finite on the segment $[0,x']$; since we already know that $f(x) \leq f(u)$ for $u \in [0,x')$, we conclude that $f(x) \leq f(x')$, as claimed. Thus, $x \in \operatorname{Dom} f$ is a minimizer of f; but then $f'(x)$ clearly is nonempty and contains 0. We see that $(x,0)$ belongs to the graph of f', and at the very beginning we assumed the opposite. Thus, the implication 2)$\Rightarrow$ 1) is proved.

Exercise 6.3.7: let $\bar{x}$ be a weak solution to the variational inequality defined by G and F, and let $H=F+\mathcal{N}_G$. In view of the description of $\mathcal{N}_G$ given in exercise 6.3.5, every point of G, and, in particular, $\bar{x}$, is a weak solution to the variational inequality defined by G and $\mathcal{N}_G$; consequently, $\bar{x}$ is a weak solution to the variational inequality defined by (G,H). Besides this, $\mathcal{V}(\mathcal{N}_G)=0$ (why?), so that H is semibounded on int G. Note also that H is maximal monotone in view of the Rockafellar Theorem.

Now let us prove the following:

let $\bar{x}$ be a weak solution to a variational inequality (G,R) involving a monotone semibounded on int G *operator R,* int $G \subset \operatorname{Dom} R \subset G$. *Then adding $(\bar{x},0)$ to the graph of R we preserve monotonicity of the operator. In particular, if R is maximal monotone, then $\bar{x} \in \operatorname{Dom} R$ and $0 \in R(\bar{x})$.*

Let us prove the latter statement. Without loss of generality we may assume that $\bar{x}=0$; what we should prove is that $y^Tx \geq 0$ for all $x \in \operatorname{Dom} R$ and all $y \in R(x)$. By definition of a weak solution, the desired statement does hold true for $x \in \operatorname{int} G$; now let $x \in \partial G$ and $y \in R(x)$. Assume that $y^Tx<0$, and let us lead the assumption to a contradiction. To this end let z be

an interior point of G, and let $x_i = (1 - 1/i)(\frac{1}{2}x) + (1/i)z$. Then $x_i \in \operatorname{int} G$, so that one can find $y_i \in R(x_i)$. We know that

$$(x_i - x)^T y_i \geq (x_i - x)^T y \to_{i\to\infty} -\frac{1}{2}x^T y > 0,$$

so that

$$(x_i - x)^T y_i \geq \alpha > 0$$

for all large enough values of i. On the other hand,

$$2x_i - x = \frac{2}{i-1}(z - x_i),$$

and we come to

$$\frac{2}{i-1}y_i^T(z - x_i) = y_i^T(2x_i - x) \geq y_i^T(x_i - x) \geq \alpha > 0$$

(recall that $x_i \in \operatorname{int} G$ and $y_i \in R(x_i)$, so that $y_i^T x_i = y_i^T(x_i - \bar{x}) \geq 0$); the resulting inequality contradicts the assumption that R is semibounded on $\operatorname{int} G$, and we are done.

Applying the statement we just have proved to the operator H, we see that $\bar{x} \in \operatorname{Dom} H \subset \operatorname{Dom} F$ and $0 \in H(\bar{x})$, i.e., there exists $y \in F(\bar{x})$ such that $-y \in \mathcal{N}_G(\bar{x})$. By construction of $\mathcal{N}_G$ we have $y^T(x - \bar{x}) \geq 0$, $x \in G$, so that $\bar{x}$ is a strong solution to the variational inequality given by (G, F). ∎

Solutions to Section 8.3

Exercise 8.3.4: From exercise 8.3.3 we know that

$$V(\phi+\eta) \le V(\phi) + \langle \eta, V'(\phi)\rangle + \mathcal{L}V(\eta);$$

since V^+ differs from V by a linear term, we have

$$V^+(\phi+\eta) \le V^+(\phi) + \langle \phi, (V^+)'(\phi)\rangle + \mathcal{L}V(\eta);$$

substituting $\phi = \phi_i$ and $\eta = -\gamma_i f'(x_i)/|f'(x_i)|_*$, we come to the recurrence

$$D_{i+1} \equiv V^+(\phi_{i+1}) \le D_i - \gamma_i \left\langle f'(x_i), V'(\phi_i) - x^* \right\rangle / |f'(x_i)|_* + \frac{\mathcal{L}}{2}\gamma_i^2,$$

and since $V'(\phi_i) = x_i$, we have

$$\left\langle f'(x_i), V'(\phi_i) - x^* \right\rangle = \left\langle f'(x_i), x_i - x^* \right\rangle \ge f(x_i) - f^*;$$

thus,

$$D_{i+1} \le D_i - \gamma_i (f(x_i) - f^*)/|f'(x_i)|_* + \frac{\mathcal{L}}{2}\gamma_i^2. \tag{14.2.27}$$

Since also $|f'(x_i)|_* \le L(f) \equiv L_{\|\cdot\|}(f)$ (why?), we immediately come to

$$\varepsilon_N \le L(f)\frac{D_1 - D_{N+1} + \frac{\mathcal{L}}{2}\sum_{i=1}^N \gamma_i^2}{\sum_{i=1}^N \gamma_i}. \tag{14.2.28}$$

It remains to note that, since $\phi_1 = 0$, we have

$$\delta = D_1 - D_{N+1} = V(\phi_1) - V(\phi_{N+1}) - \langle \phi_1 - \phi_{N+1}, x^*\rangle =$$

$$= -V(\phi_{N+1}) - \langle \phi_{N+1}, x^*\rangle \le -\frac{1}{2}|\phi_{N+1}|_*^2 + |\phi_{n+1}|_*|x^*| \le \frac{1}{2}|x^*|^2.$$

Solutions to Section 9.3

Exercise 9.3.1: induction on i. For $i = 1$ the statement is evident. Now, since y_i is the projection of y_0 onto P_i and $y_{i+1} \in P_{i+1} \subset P_i$, we have

$$(y_0 - y_i)^T (y_{i+1} - y_i) \leq 0,$$

whence

$$|y_0 - y_{i+1}|^2 = |y_0 - y_i|^2 + 2(y_0 - y_i)^T (y_i - y_{i+1}) + |y_i - y_{i+1}|^2 \geq |y_0 - y_i|^2 + |y_i - y_{i+1}|^2,$$

which clearly justifies the inductive step.

Exercise 9.3.2:

1): let us verify that the sets Q_i, $i \in I_k$, are nonempty, or, which is the same, that

$$f_i^- \leq l_i, \ i \in I_k.$$

One clearly has

$$l_i = f_j^- + \lambda \Delta_j \tag{14.2.29}$$

for certain j which is $\geq j_k$ and is $\leq i$. Since f_l^- do not decrease with l and f_l^+ do not increase with l, we have

$$\Delta_i \leq \Delta_j - (f_i^- - f_j^-),$$

whence

$$f_i^- \leq \Delta_j - \Delta_i + f_j^-. \tag{14.2.30}$$

Since both i and j belong to I_k, we have

$$\Delta_i \geq (1 - \lambda)\Delta_{j_k} \geq (1 - \lambda)\Delta_j \tag{14.2.31}$$

(we have taken into account that the gaps never increase). Combining (14.2.30) and (14.2.31), we come to

$$f_i^- \leq f_j^- + \lambda \Delta_j,$$

which, in view of (14.2.29), means that $f_i^- \leq l_i$, as claimed.

The fact that $Q_i \supset Q_{i+1}$, $i, i+1 \in I_k$, is evident: $l_i \geq l_{i+1}$ by construction, while $f_i \leq f_{i+1}$ by the basic properties of the models.

2): according to rule 4*), the iterations from I_k can be partitioned into sequential fragments $J_1, ..., J_s$ as follows: during the steps $i \in J_l$ the best found so far search point $\bar{x}_i$ remains constant, let it be denoted by x^l; at the final step from the group J_l, $l < s$, we form the new best search point x^{l+1}. According to the description of the method, we have

$$x_{i+1} = \pi_{Q_i}(x^l), \ i \in J_l. \tag{14.2.32}$$

The points x^l are defined for $l = 1, 2, ..., s$; let us define x^{s+1} as the search point generated at the last iteration of the last group J_s. Then for all $l \leq s$ we have

$$x^{l+1} = \pi_{Q_{q_l}}(x^l), \tag{14.2.33}$$

where q_l denotes the last iteration from the group J_l.

Now consider group $J_l = \{p_l, p_l + 1, ..., q_l\}$. From 1) we already know that $Q_i \in Q_{i+1}$, $i, i+1 \in I_k$, so that from exercise 9.3.1 and (14.2.32), (14.2.33) it follows that

$$|x^l - x^{l+1}|^2 \geq \sum_{i \in J_l} |y_{i-1} - y_i|^2, \tag{14.2.34}$$

where $y_i = x_i$, $q_l \geq i \geq p_l$, and $y_{p_l - 1} = x^l$. On the other hand, in view of the standard properties of projection and the inclusion $x^{s+1} \in Q_{q_s} \subset Q_{q_l}$ from (14.2.33) it follows that

$$|x^1 - x^{s+1}|^2 \geq \sum_{l=1}^{s} |x^l - x^{l+1}|^2;$$

combining this observation with (14.2.34), we come to

$$|x^1 - x^{s+1}|^2 \geq \sum_{i \in I_k} |x_i - x_{i+1}|^2 \tag{14.2.35}$$

(note that $y_{p_l - 1} = x_{q_{l-1}+1}$).

Now we can use the same reasoning as for the basic version of the method. Namely, $f_i(x_i) = f(x_i) \geq f_i^+$, while $l_i \leq f_i^- + \lambda \Delta_i$, so that $f_i(x_i) \geq l_i$. It follows that $f_i(x_{i+1}) = l_i$, so that when passing from x_i to x_{i+1}, we decrease f_i at least by $(1 - \lambda)\Delta_i \geq (1 - \lambda)^2 \Delta_{j_k}$. Since f_i clearly is Lipschitz continuous with constant $L(f)$, we come to $|x_i - x_{i+1}| \geq (1 - \lambda)^2 \Delta_{j_k} L^{-1}(f)$. Thus, the right hand side in (14.2.34) is at least $N_k L^{-2}(f)(1 - \lambda)^4 \Delta_{j_k}^2$, while the right hand side is $\leq D^2(G)$, and 2) follows.

3) The inequality $\Delta_i \geq f(\bar{x}_i) - \min_G f$ immediately follows from $f_i \leq f$. Now assume that N is such that $\Delta_N > \varepsilon$. Let k be the first integer such that $I_1 \cup ... \cup I_k \supset \{1, ..., N\}$; then the first index in I_k is $\leq N$, and consequently $\Delta_{j_k} \geq \Delta_N > \varepsilon$. We have, by construction of the groups I_l,

$$\Delta_{j_{l+1}} < (1 - \lambda)\Delta_{j_l},$$

so that

$$\Delta_{j_l} \geq (1 - \lambda)^{-(k-l)} \Delta_{j_l} \geq (1 - \lambda)^{-(k-l)} \varepsilon, \ 1 \leq l \leq k.$$

From 2) it now follows

$$N \leq \sum_{l=1}^{k} N_l \leq \sum_{l=1}^{k} (1 - \lambda)^{-4} D^2(G) L^2(f) \Delta_l^{-2} \leq$$

$$\leq (1 - \lambda)^{-4} D^2(G) L^2(f) \sum_{p=0}^{\infty} \varepsilon^{-2} (1 - \lambda)^{2p} \equiv D^2(G) L^2(f) \varepsilon^{-2} c(\lambda),$$

as required.

Solutions to Section 12.6

Exercise 12.6.2:

1) Let p be an optimal solution to (P); let us prove that then p possesses the "alternance property". Since the cardinality of Σ is $> N+1$, $\delta > 0$ (otherwise the polynomial $1 - tp(t)$ would possess more than $N+1$ zeros and would be zero identically, which is impossible due to the structure of the polynomial). Let Σ' be the subset of Σ formed by the points $t \in \Sigma$ where $|q(t)| = \delta$. Note that Σ' is a closed subset of Σ which does not contain 0 (the latter is, of course, true if $0 \notin \Sigma$; if $0 \in \Sigma$, then, by assumption, $\gamma > 0$, so that q is a continuous on $[0, \infty)$ function with $q(0) = 0$, and since $q(0) = 0$ and $\delta > 0$, we have $0 \notin \Sigma'$). We may partition the set Σ' into sequential parts Σ_0, Σ_1,... (with Σ_i being a closed set and Σ_{i+1} being to the right of Σ_i) in such a way that $q(t) = \kappa(-1)^i\delta$ for $t \in \Sigma_i$, where κ (the sign of q on Σ_0) is either $+1$, or -1. The number K of sets Σ_i in the partitioning clearly is finite (why?). The "alternance property" means exactly that the number of these sets is at least $N+2$ and that $\kappa = +1$, and this is what should be proved.

Assume, first, that $K < N+2$. One can find $K \leq N+1$ points $(0 <)t_0 < ... < t_{K-2}$ such that $\Sigma_i \subset (t_i, t_{i+1})$, where $t_{K-1} = \infty$. Let

$$\pi(t) = \prod_{i=1}^{K-1} (t - t_i);$$

since $K - 1 \leq N$, π is a polynomial of degree $\leq N$. There clearly exists a neighbourhood Σ'' of the set Σ' such that

$$\sup_{t \in \Sigma''} t^\gamma |1 - t[p(t) - r\kappa\pi(t)]| = \sup_{t \in \Sigma''} t^\gamma |s(t) - r\kappa\pi(t)| < \delta, \ s(t) = 1 - tp(t),$$

for all small enough values of $r > 0$ (why?), and consequently

$$\max_{t \in \Sigma} t^\gamma |s(t) - r\kappa\pi(t)| < \delta$$

for all small enough positive r (why?), which contradicts the optimality of p.

Thus, $K \geq N+2$; it remains to verify that $\kappa = 1$. Since the polynomial $s(t)$ is positive at $t = 0$ and its sign on Σ_i is $(-1)^i\kappa$, $i = 0, ..., K$, in the case of $\kappa < 0$ $s(t)$ would have at least $K \geq N+2$ changes of the sign on the axis; since the degree of the polynomial is $\leq N+1$, it would be zero identically, which is not the case.

2) Now assume that p possesses the "alternance property"; let us prove that then p is optimal. Indeed, otherwise there would exist a polynomial π of degree $\leq N$ such that $t^\gamma|1 - t\pi(t)| < t^\gamma|1 - tp(t)|$ when $t = \sigma_i$, (σ_i, $i = 0, ..., N+1$, are the points of alternance). Consequently, the polynomial

$$\Delta(t) = (1 - tp(t)) - (1 - t\pi(t)) = t(\pi(t) - p(t))$$

at the points σ_i, $i = 0, ..., N+1$, would be of the same sign as q, i.e., of the signs $(-1)^i$. It would follow that $\Delta(t)$ would have at least $N+1$ positive zeros, and since this polynomial is also 0 at $t = 0$, it would be zero identically, which is the desired contradiction.

Solutions to Section 13.4

Exercise 13.4.1 : The normal solution x^* to a solvable operator equation with symmetric matrix A belongs, as we know, to the image space E of A; this subspace is invariant for A, and the restriction of A onto E is a symmetric positive definite operator on E. The right hand side vector b also belongs to E (since $b = Ax^*$, $x^* \in E$ and E is invariant for A); it follows that all Krylov subspaces of (A, b) belong to E and, consequently, the trajectory of the CGM on (A, b) coincides with the trajectory of the method as applied to the natural "restriction" of the equation onto E. Thus, replacing, if necessary, $\mathbf{R}^n$ with E, we may assume without loss of generality that A is nonsingular.

Now, equation $Ax = b$ with symmetric positive definite matrix A clearly belongs to the family $\mathcal{U}_n(\Sigma, 0, R)$, where Σ is the finite set comprised of eigenvalues of A (all of them are positive) and R is the norm of the solution to the equation. Bearing in mind this observation and taking into account Proposition 13.3.1 (Lecture 13), we come to the following expression for the

"inaccuracy in terms of the objective"

$$d_{A,b}(x_k) \equiv f_{A,b}(x_k) - \min_x f_{A,b}$$

of k-th approximate solution found by the CGM is given by

$$d_{A,b}(x_k) = \min_{q \in P_k^0} \int t q^2(t) d\mu(t), \tag{14.2.36}$$

where μ is the measure on Σ given by the Proposition; from the construction given in Lecture 13 it follows that the cardinality of the support set Σ' of this measure is exactly m.

Now, Σ' is an m-point subset of the positive ray, and there exists a polynomial $q_m(t)$ of the degree m, $q_m(0) = 1$, which vanishes on Σ'; it follows that the right hand side of (14.2.36) corresponding to $k = m$ is zero, so that $d_{A,b}(x_m) = 0$ and x_m is the exact solution to the equation.

Vice versa, if x_k is an exact solution to the equation, then $d_{A,b}(x_k) = 0$, and from (14.2.36) it follows that there exists a polynomial $q \in P_k^0$ which vanishes on Σ' and therefore possessing $m = \operatorname{card} \Sigma'$ distinct roots; it is possible only if $k \geq \deg q \geq m$, so that $k \geq m$. ∎

Exercise 13.4.7:

2): if λ is an eigenvalue of T_k, then λ is real (since T_k is symmetric, see 1)) and there exist a nonzero $p \in E_k$ such that $(T_k - \lambda I)p = 0$. By definition of T_k it means that the nonzero polynomial p of degree $\leq k - 1$ is such that the polynomial $q(t) = (t - \lambda)p(t)$ is μ-orthogonal to E_k. Since q is of the degree at most k and is nonzero, we conclude from the uniqueness statement of Exercise 13.4.6 that q is proportional, with a nonzero coefficient, to q_k, so that λ is a root of q_k. Vice versa, if λ is a real root of q_k, then the polynomial $p(t) = q_k(t)/(t - \lambda)$ is well defined and is a nonzero element of E_k; we have $(t - \lambda I)p = q_k$, and since q_k is orthogonal to E_k, $(T_k - \lambda I)p = 0$ by definition of T_k, so that λ is an eigenvalue of T_k.

3): as we know from 1), the operator T_k is symmetric; besides this, one clearly has

$$a(p, p) \leq (Tp, p) = (T_k p, p) \leq b(p, p), \ p \in E_k,$$

so that

a) the spectrum of T_k belongs to the segment $[a, b]$.

Let us prove that

b) neither a nor b belong to the spectrum of T_k;

c) every eigenvalue of T_k is of multiplicity 1.

To prove b), assume that a belongs to the spectrum of T_k; then there exists a nonzero polynomial p of degree $\le k-1$ such that

$$a(p,p) = (T_k p, p).$$

The latter quantity is equal to (Tp,p) by construction of T_k, so that $((T-aI)p,p) = 0$, or, which is the same,

$$\int (t-a)p^2(t)d\mu(t) = 0.$$

Since the support set of μ belongs to $[a,b]$, the latter relation is possible only if this support set is covered by the set of roots of the polynomial $(t-a)p^2(t)$; since p is nonzero, the latter set is comprised of at most $k \le N$ points (recall that $\deg p \le k-1$), and from the very beginning we have assumed that the cardinality of the support set of μ is at least $N+1$. Thus, a cannot belong to the spectrum of T_k. A completely similar reasoning demonstrates that b also does not belong to this spectrum, as claimed in b).

To prove c), assume, on contrary, that certain eigenvalue λ of T_k is of multiplicity at least 2. Then there exist two linearly independent polynomials p', p'' of degrees $\le k-1$ such that the nonzero polynomials $(t-\lambda)p'(t)$ and $(t-\lambda)p''(t)$ of degrees at most k are μ-orthogonal to E_k; same as in the proof of 2), we conclude that these polynomials are proportional, with nonzero coefficients, to q_k. Consequently, there exist a nonzero ω such that $(t-\lambda)p'(t) \equiv \omega(t-\lambda)p''(t)$, or, which is the same, $(t-\lambda)(p'(t)-\omega p''(t)) \equiv 0$. The latter relation implies that $p' = \omega p''$, which is a contradiction (p' and p'' were assumed to be linearly independent). Thus, c) is proved.

Since the dimension of E_k is exactly k (exercise 13.4.6.4)) and, as we know, T_k is symmetric, T_k has k real eigenvalues (counted with multiplicities); from c) it follows that all these k eigenvalues are distinct; and from a) and b) we know that these eigenvalues belong to (a,b). Thus, the spectrum of T_k is comprised of k distinct points from (a,b). As we know from 2), q_k vanishes at the spectrum of T_k, so that q_k has k distinct roots in (a,b).

To complete the proof of 3), it suffices to demonstrate that q_k and q_{k-1} cannot have a root in common. This is evident when $k=1$ (since q_0 is a nonzero constant). Now assume that $1 < k \le N$ and that q_k and q_{k-1} have a common root λ. It means that $q_k(t) = (t-\lambda)p(t)$ and $q_{k-1}(t) = (1-\lambda)r(t)$ with certain polynomials p and r of degrees exactly $k-1$ and $k-2$, respectively. Let

$$p = \sum_{i=0}^{k-1} c_i q_i$$

be the representation of p in the basis $q_0, ..., q_{k-1}$; then $c_{k-1} \ne 0$, since p is of the degree $k-1$. We come to

$$q_k(t) = (t-\lambda)p(t) = \sum_{i=0}^{k-1} c_i(t-\lambda)q_i(t) = (t-\lambda)^2 r(t) + \sum_{i=1}^{k-2} c_i(t-\lambda)q_i(t).$$

Let us take here the inner product with the polynomial r. We come to

$$0 = (q_k, r) = \int r(t)q_k(t)d\mu(t) = c_{k-1}\int (t-\lambda)^2 r^2(t)d\mu(t) + \sum_{i=0}^{k-2}\int (t-\lambda)r(t)q_i(t)d\mu(t) =$$

$$= c_{k-1} \int (t-\lambda)^2 r^2(t) d\mu(t) + \sum_{i=0}^{k-2} (q_{k-1}, q_i) = c_{k-1} \int (t-\lambda)^2 r^2(t) d\mu(t).$$

Since $c_{k-1} \neq 0$, the resulting relation says that $\int (t-\lambda)^2 r^2(t) d\mu(t) = 0$. We conclude that the support set of μ is covered by the set comprised of λ and the roots of r; this latter set is of cardinality at most $k-1 < N$, so that the support set of μ is of cardinality $< N$, which is a contradiction.