Proceedings of the International Congress of Mathematicians
Hyderabad, India, 2010

Recent Advances in Structural Optimization

Yurii Nesterov[*]

Abstract

In this paper we present the main directions of research in Structural Convex Optimization. In this field, we use additional information on the structure of specific problem instances for accelerating standard Black-Box methods. We show that the proper use of problem structure can provably accelerate these methods by the order of magnitudes. As examples, we consider polynomial-time interior-point methods, smoothing technique, minimization of composite functions and some other approaches.

Mathematics Subject Classification (2010). Primary 90C25; Secondary 90C06.

Keywords. Convex optimization, structural optimization, complexity estimates, worst-case analysis, polynomial-time methods, interior-point methods, smoothing technique.

1. Introduction

Optimization problems are usually related to some models of the real-life situations. On the other hand, in order to develop a method for solving such problems, a numerical analyst starts from creating an appropriate model of a particular class of optimization problems. For this, there exist several reasons. Firstly, it is natural to use the developed scheme for solving many optimization problem with similar characteristics. Secondly, the model of the problem provides us with useful properties and inequalities helping to approach the optimal solution. Finally, fixation of the model allows us to perform a worst-case complexity analysis and to develop the optimal schemes.

The progress in performance of methods in Convex Optimization during the last three decades is closely related to evolution of the above models and

[*] Catholic University of Louvain (UCL), Department INMA/CORE
CORE, 34 voie du Roman Pays, 1348 Louvain-la-Neuve, Belgium.
E-mail: Yurii.Nesterov@uclouvain.be.

to better understanding of significance of the help we can offer to numerical schemes by opening an access to the structure of problem instances.

At the very first stage of the development of our field, the standard model of optimization problem was quite pour. It was not even clear that such a notion is necessary or useful. The tradition was to fix the analytic form of the problem and the classes of functional components. For example, for "unconstrained" convex optimization problem, the standard form was as follows:

$$\min_{x \in Q} f(x), \tag{1}$$

where $Q \subset R^n$ is a closed bounded convex set ($\|x\| \leq R$ for all $x \in Q$), and f is a closed convex function. If we assume that

$$\|\nabla f(x)\| \quad \leq \quad L \quad \forall x \in Q, \tag{2}$$

then we get the problem class $\mathcal{C}_1$, which is formed by the problems of unconstrained "nonsmooth" minimization. Assuming that

$$\|\nabla f(x) - \nabla f(y)\| \quad \leq \quad M\|x-y\| \quad \forall x, y \in Q, \tag{3}$$

we get the problem class $\mathcal{C}_2$ of smooth optimization problems. Thus, the model of the problem was represented as the set of useful properties and inequalities which can be somehow employed by optimization scheme. For example, if $f \in \mathcal{C}_1$, then by its convexity we know that

$$f(y) \quad \geq \quad f(x) + \langle \nabla f(x), y - x \rangle \quad \forall x, y \in Q. \tag{4}$$

Therefore, evaluating this function at points $\{x_i\}_{i=1}^k$, we can form its model as follows:

$$\mathcal{L}_k(x) \quad \stackrel{\mathrm{def}}{=} \quad \max_{1 \leq i \leq k} [f(x_i) + \langle \nabla f(x_i), x - x_i \rangle] \ \leq \ f(x), \quad x \in Q. \tag{5}$$

In this methodology, the elements of function $\mathcal{L}_k(x)$ and the bound (2) represent the *full available information* on our objective function. In other words, the designed numerical methods are obliged to work with these objects only. If $f \in \mathcal{C}_2$, then we have also

$$f(y) \quad \stackrel{(3)}{\leq} \quad f(x) + \langle \nabla f(x), y - x \rangle + \tfrac{1}{2} L \|y - x\|^2, \quad \forall x, y \in Q.$$

This inequality together with (5) helps to increase the rate of convergence of corresponding optimization schemes.

Thus, it was natural that the Complexity Theory of Convex Optimization, developed by A.Nemirovskii and D.Yudin [7] in the late 70's, was based on the *Black-Box Concept*. It was assumed that the only information the optimization methods can learn about the particular problem instance is the values and derivatives of these components at some test points. This data can be reported by a special unit called *oracle* which is *local.* This means that it is not changing if the function is modified far enough from the test point. At the time of its development, this concept fitted very well the existing computational practice, where the interface between the general optimization packages and the problem's data was established by Fortran subroutines created independently by the users.

Black-Box framework allows to speak about the lower performance bounds for different problem classes in terms of *informational complexity.* That is the lower estimate for the number of calls of oracle which is necessary for any optimization method in order to guarantee delivering an ϵ-solution to any problem from the problem class. In this performance measure we do not include at all the complexity of auxiliary computations of the scheme.

In the table below, the first column indicates the problem class, the second one gives an upper bound for allowed number of calls of oracle in the optimization scheme[1], and the last column gives the lower bound for analytical complexity of the problem class, which depends on the absolute accuracy ϵ and the class parameters.

Problem class	Limit for calls	Lower bound
$\mathcal{C}_1 : \ \|\nabla f(\cdot)\| \leq L$	$\leq O(n)$	$O\left(\frac{L^2R^2}{\epsilon^2}\right)$
$\mathcal{C}_2 : \ \|\nabla^2 f(\cdot)\| \leq M$	$\leq O(n)$	$O\left(\frac{M^{1/2}R}{\epsilon^{1/2}}\right)$
$\mathcal{C}_3 : \ \|\nabla f(\cdot)\| \leq L$	$\geq O(n)$	$O\left(n \ln \frac{LR}{\epsilon}\right)$

(6)

It is important that these bounds are *exact.* This means that there exist methods, which have efficiency estimates on corresponding problem classes proportional to the lower bounds. The corresponding *optimal methods* were developed in [6, 7, 16, 19, 20]. For further references, we present a simplified version of the optimal method [7] as applied to the problem (1) with $f \in \mathcal{C}_2$:[2]

Choose a starting point $y_0 \in Q$ *and set* $x_{-1} = y_0$. *For* $k \geq 0$ *iterate:*

$$\begin{aligned} x_k &= \arg\min_{x \in Q} \left[f(y_k) + \langle \nabla f(y_k), x - y_k \rangle + \tfrac{M}{2} \|x - y_k\|^2 \right], \\ y_{k+1} &= x_k + \tfrac{k}{k+3}(x_k - x_{k-1}). \end{aligned} \tag{7}$$

[1] If this upper bound is smaller than $O(n)$, then the dimension of the problem is really very big, and we cannot afford the method to perform this amount of calls.

[2] In method (11)-(13) from [7], we can set $a_k = 1 + k/2$ since in the proof we need only to ensure $a_{k+1}^2 - a_k^2 \leq a_{k+1}$.

As we see, the complexity of each iteration of this scheme is comparable with that of the simplest gradient method. However, the rate of convergence of method (7) is much faster.

After a certain period of time, it became clear that, despite to its mathematical excellence, Complexity Theory of Convex Optimization has a hidden drawback. Indeed, in order to apply convex optimization methods, we need to be *sure* that functional components of our problem are convex. However, we can check convexity only by analyzing the *structure* of these functions:[3] If our function is obtained from the *basic* convex functions by *convex* operations (summation, maximum, etc.), we conclude that it is convex. If not, then we have to apply general nonlinear optimization methods which usually do not have theoretical guarantees for the global performance.

Thus, the functional components of the problem are not in the black box in the moment we check their convexity and choose minimization scheme. However, we put them into the black box for numerical methods. This is the main conceptual contradiction of the standard Convex Optimization Theory.

As we have already mention, the progress in Convex Optimization was mainly related to discovering the different possibilities for opening the Back Box for numerical methods. In this paper we present some of these approaches and discuss the corresponding improvements of complexity estimates as compared to the standard Black-Box framework. The order of discussion of these approaches has certain logic. However, it does not reflect the chronology of the development.

2. Primal-dual Subgradient Methods

In our first approach, we do not accelerate the Black-Box methods. We just look inside the oracle and show how this information can be used for constructing an approximate solution to the dual problems [14].

Let the norm $\| \cdot \|$ be Euclidean. We can form a *linear model* of function $f \in \mathcal{C}_1$ as follows:

$$l_k(x) \quad = \quad \tfrac{1}{k+1} \sum_{i=0}^{k} [f(x_i) + \langle \nabla f(x_i), x - x_i \rangle].$$

This model can be used in the following optimization method [14]:

$$x_{k+1} \quad = \quad \arg\min_{x \in Q} \left[\, l_k(x) + \tfrac{1}{2}\gamma_k \|x - x_0\|^2 \right], \tag{8}$$

where $\gamma_k > 0$ are certain parameters. This is a usual Black-Box subgradient method for solving problem (1). If $\gamma_k = \frac{L}{R\sqrt{k+1}}$ and $\hat{x}_k = \frac{1}{k+1} \sum\limits_{i=0}^{k} x_i$, then

$$f(\hat{x}_k) - f(x^*) \quad \le \tfrac{2LR}{\sqrt{k+1}}, \quad k \ge 0, \tag{9}$$

[3] Numerical verification of convexity is an extremely difficult problem.

where x^* is the solution of problem (1). Thus, method (8) is optimal for our class of problems. In order to understand what is the dual problem, we need to look inside the oracle.

1. Discrete minimax. Consider the following variant of problem (1):

$$\min_{x \in Q} [f(x) = \max_{1 \leq j \leq m} f_j(x)], \tag{10}$$

where $f_j \in \mathcal{C}_1$, $j = 1, \ldots, m$. Denote $\Delta_m = \{y \in R^m_+ : \sum\limits_{j=1}^m y^{(j)} = 1\}$. Then

$$\begin{aligned} f^* &= \min_{x \in Q} \max_{1 \leq j \leq m} f_j(x) \;=\; \min_{x \in Q} \max_{y \in \Delta_m} \sum_{j=1}^m y^{(j)} f_j(x) \\ &= \max_{y \in \Delta_m} \left[\phi(y) \overset{\text{def}}{=} \min_{x \in Q} \sum_{j=1}^m y^{(j)} f_j(x) \right]. \end{aligned}$$

Thus, the dual problem is

$$f^* \;=\; \max_{y \in \Delta_m} \phi(y). \tag{11}$$

Note that the computation of the value of dual function $\phi(y)$ may be difficult since it requires to solve a nonlinear optimization problem.

Denote by e_j the jth coordinate vector in R^m. Let us look at the following variant of method (8) with $\gamma_k = \frac{L}{R\sqrt{k+1}}$.

```
**Initialization:** Set $l_0(x) \equiv 0$, $m_0 = 0 \in Z^m$.

**Iteration** ($k \geq 0$)**:**

  **1.** Choose any $j_k^* :\; f_{j_k^*}(x_k) = f(x_k)$.

  **2.** Set $l_{k+1}(x) = \frac{k}{k+1} l_k(x) + \frac{1}{k+1}[f(x_k) + \langle \nabla f_{j_k^*}(x_k), x - x_k \rangle]$.        (12)

  **3.** Compute $x_{k+1} = \arg\min_{x \in Q} \left\{ l_{k+1}(x) + \frac{1}{2}\gamma_k \|x - x_0\|^2 \right\}$.

  **4.** Update $m_{k+1} = m_k + e_{j_k^*}$.

**Output:** $\hat{x}_{k+1} = \frac{1}{k+1} \sum_{i=0}^k x_i, \quad \hat{y}_{k+1} = \frac{1}{k+1} m_{k+1}$.
```

Thus, the entries of vector $\hat{y}_k$ are the *frequencies* of appearing the corresponding functional components as the biggest ones of the objective function. For the output of this process we have the following guarantee:

$$f(\hat{x}_k) - \phi(\hat{y}_k) \quad \leq \tfrac{2LR}{\sqrt{k+1}}, \quad k \geq 0. \tag{13}$$

2. Primal-dual problem. Let f be a closed convex function defined on R^n. Consider the conjugate function:

$$f_*(s) \quad = \quad \sup_{x \in R^n} [\langle s, x \rangle - f(x)].$$

Then $f(x) = \max\limits_s [\langle s, x \rangle - f_*(s) : \ s \in \mathrm{dom} f_*]$, $x \in R^n$. Denote by $s(x)$ an optimal solution of the latter problem. Note that

$$\begin{aligned} f^* \quad &= \quad \min_{x \in Q} \max_{s \in \mathrm{dom} f_*} [\langle s, x \rangle - f_*(s)] \ = \ \max_{s \in \mathrm{dom} f_*} \min_{x \in Q} [\langle s, x \rangle - f_*(s)] \\ &= \quad \max_{s \in \mathrm{dom} f_*} \left[\ \psi(s) \overset{\mathrm{def}}{=} -\xi_Q(s) - f_*(s) \ \right], \end{aligned}$$

where $\xi_Q(u) = \max\limits_{x \in Q} \langle u, x \rangle$. Thus, the problem dual to (1) is as follows:

$$f^* \quad = \quad \max_{s \in \mathrm{dom} f_*} \psi(s). \tag{14}$$

It appears, that the method (8) is able to approximate the optimal solution to this problem. Indeed, let $\{x_k\}$ be formed by (8) with $\gamma_k = \frac{L}{R\sqrt{k+1}}$. Define $\hat{s}_k = \frac{1}{k+1} \sum\limits_{i=0}^{k} s(x_i)$. Then

$$f(\hat{x}_k) - \phi(\hat{y}_k) \quad \leq \tfrac{2LR}{\sqrt{k+1}}, \quad k \geq 0.$$

Again, we find an approximate solution to the dual problem without computing the values of the dual function (this may be difficult).

3. Polynomial-time Interior-point Methods

Thus, we have seen that a proper use of structure of the oracle can help in generating an approximate solution to the dual problem. Is it possible to use this structure for accelerating the Black-Box schemes? Intuitively we always hope that this is true. Unfortunately, structure is a very fuzzy notion, which is quite difficult to formalize. One possible way to describe the structure is to fix the *analytical type* of functional components. For example, we can consider the problems with linear constraints only. It can help, but this approach is very fragile: If we add just a single constraint of another type, then we get a new problem class, and all theory must be redone from scratch.

On the other hand, it is clear that having the structure at hand we can play a lot with the *analytical form* of the problem. We can rewrite the problem in many equivalent settings using non-trivial transformations of variables or constraints, introducing additional variables, etc. However, this would serve almost no purpose without fixing a clear final goal. So, let us try to understand what it could be.

As usual, it is better to look at classical examples. In many situations the sequential reformulations of the initial problem can be seen as a part of numerical scheme. We start from a complicated problem $\mathcal{P}$ and, step by step, change its structure towards to the moment we get a trivial problem (or, a problem which we know how to solve):

$$\mathcal{P} \longrightarrow \ldots \longrightarrow (f^*, x^*).$$

A good example of such a strategy is the standard approach for solving system of linear equations

$$Ax = b.$$

We can proceed as follows:

1. Check if A is symmetric and positive definite. Sometimes this is clear from the origin of the matrix.
2. Compute Cholesky factorization of this matrix:
   $$A = LL^T,$$
   where L is a lower-triangular matrix. Form two auxiliary systems
   $$Ly = b, \quad L^T x = y.$$
3. Solve these system by sequential exclusion of variables.

Imagine for a moment that we do not know how to solve the system of linear equations. In order to *discover* the above scheme we should apply the following

Golden Rules

1. Find a class of problems which can be solved very efficiently.[a]
2. Describe the transformation rules for converting the initial problem into desired form.
3. Describe the class of problems for which these transformation rules are applicable.

(15)

[a] In our example, it is the class of linear systems with triangular matrices.

In Convex Optimization, these rules were used already several times for breaking down the limitations of Complexity Theory.

Historically, the first example of that type was the theory of polynomial-time interior-point methods (IPM) based on *self-concordant barriers* [15]. The first step in the development of this theory was discovery of *unconstrained* minimization problems which can be solved efficiently by the Newton method. We say that a closed convex function f is *self-concordant* on its open domain $\operatorname{dom} f \subset R^n$ if

$$D^3 f(x)[h]^3 \quad \leq \quad 2D^2 f(x)[h]^2 \quad \forall x \in \operatorname{dom} f, \ h \in R^n,$$

where $D^k f(x)[h]^k$ is the kth differential of function f at x along direction h. It appears that the properties of these functions fit very well the Newton scheme.

Let us use the Hessian $\nabla^2 f(x)$ of such a function for defining a *local norm* around x:

$$\|h\|_x \quad = \quad \langle \nabla^2 f(x) h, h \rangle^{1/2}, \quad \|s\|_x^* \ = \ \langle s, [\nabla^2 f(x)]^{-1} s \rangle^{1/2}.$$

(It is possible to prove that if $\operatorname{dom} f$ is bounded, then the Hessian is nondegenerate at any point of the domain.) Then we can define the Dikin ellipsoid at x as follows:

$$W_r(x) \quad = \quad \{ y \in R^n : \ \|y - x\|_x \leq r \}.$$

It appears that for any $r < 1$ we have $W_r(x) \subset \operatorname{dom} f$ for any feasible x. Moreover, inside this ellipsiod we can predict very well the variation of the Hessian:

$$(1-r)^2 \nabla^2 f(x) \quad \preceq \quad \nabla^2 f(y) \ \preceq \ \tfrac{1}{(1-r)^2} \nabla^2 f(x), \quad \forall y \in W_r(x), \ x \in \operatorname{dom} f.$$

This property results in the following behavior of the Damped Newton Method:

$$x_{k+1} \quad = \quad x_k - \tfrac{[\nabla^2 f(x)]^{-1} \nabla f(x)}{1 + \|\nabla f(x)\|_x^*}.$$

If $\|\nabla f(x)\|_x \geq \beta$ for some $\beta \in (0,1)$, then $f(x_{k+1}) \leq f(x_k) - [\beta - \ln(1+\beta)]$, else

$$\|\nabla f(x_{k+1})\|_{x_{k+1}}^* \quad \leq \quad 2 \left(\|\nabla f(x_k)\|_{x_k}^* \right)^2. \tag{16}$$

Thus, we have now an affine-invariant description of the region of quadratic convergence of the Newton method:

$$\{ x \in \operatorname{dom} f : \ \|\nabla f(x)\|_x < \tfrac{1}{2} \}.$$

Now we can try to use our achievement for solving more complicated problems, the problems of *constrained* minimization.

Consider the following *standard* minimization problem:

$$\min_{x \in Q} \langle c, x \rangle, \tag{17}$$

where Q is a bounded closed convex set. Let us assume that $Q = \mathrm{Cl}(\mathrm{dom}\, f)$ for some self-concordant function f. Then we can try to solve (17) by a *path-following method*. Define the central path $x(t)$, $t > 0$, as follows:

$$tc + \nabla f(x(t)) \quad = \quad 0. \tag{18}$$

This is the first-order optimality condition for the unique minimum of self-concordant function

$$\psi_t(x) \quad = \quad t\langle c, x\rangle + f(x),$$

for which we already have a convenient description of the region of quadratic convergence of the Newton scheme. How quickly we can increase the penalty parameter keeping the possibility to come in a close neighborhood of the new point at the central path by a quadratically convergent Newton scheme? For that we need to ensure

$$\begin{aligned} \tfrac{1}{2} \quad &> \quad \|\nabla \psi_{t+\Delta}(x(t))\|^*_{x(t)} \;=\; \|(t+\Delta)c + \nabla f(x(t))\|^*_{x(t)} \\ &\overset{(18)}{=} \quad \Delta\|c\|^*_{x(t)} \;\overset{(18)}{=}\; \tfrac{\Delta}{t}\|\nabla f(x(t))\|^*_{x(t)}. \end{aligned} \tag{19}$$

Thus, in order to increase t in a linear rate, we need to assume uniform boundedness of the local norm of the gradient of f. This is the reason for working with the following barrier function.

Definition 1. Function f is called a ν-self-concordant barrier for convex set Q if it is self concordant on $\mathrm{int}\, Q$ and

$$\langle \nabla f(x), [\nabla^2 f(x)]^{-1} \nabla f(x)\rangle \quad \leq \quad \nu, \quad x \in \mathrm{dom}\, f.$$

The value ν is called the *parameter* of the barrier f.

Self-concordant barriers have many useful properties (see [15], [8]). One of them is related to asphericity of the set Q with respect to the point $x(0)$, which is called the *analytic center* of Q:

$$W_1(x(0)) \quad \subseteq \quad Q \;\subseteq\; W_{\nu+2\sqrt{\nu}}(x(0)). \tag{20}$$

Note that the value of the barrier parameter ν can be much smaller than the dimension of the space of variables.

As we can see from the reasoning (19), we can solve the standard minimization problems with complexity $O(\sqrt{\nu}\ln\frac{\nu}{\epsilon})$, where ϵ is the desired accuracy of the solution. How wide is the class of problems to which we can apply this machinery?

It appears that in principle we cover *all* convex optimization problems. Indeed, for any closed convex set Q we can define the following *universal barrier*:

$$f_Q(x) \quad = \quad \kappa \cdot \ln \operatorname{Vol} P(x), \quad P(x) = \{s : \langle s, y - x\rangle \leq 1 \; \forall y \in Q\}.$$

Then, for certain value of $\kappa > 0$, this function is $O(n)$-self-concordant barrier for Q. Hence, in principle, we can solve all convex optimization problems with complexity $O(\sqrt{n} \ln \frac{n}{\epsilon})$. Of course, in the framework of Back-Box methods this is just impossible. Hence, we conclude that something should violate the Black-Box assumptions. And indeed, this is the process of *creating* the self-concordant barriers. There exists a kind of calculus for doing this. However, it needs a direct access to the structure of the problem, possibility to introduce additional variables, etc. As a result, we are able to apply linearly convergent methods practically to all convex optimization problem with known structure. It is interesting that the standard classification of the problems in accordance to the level of smoothness of functional components is useless here. We need only a possibility to construct self-concordant barriers for their epigraphs. However, note that each iteration of the path-following schemes is quite heavy. This is the reason for development of the cheap gradient schemes, which we describe in the remaining sections.

4. Smoothing Technique

The second example of using the rules (15) needs more explanations. By certain circumstances, these results were discovered with a delay of twenty years. Perhaps they were too simple. Or maybe they are in a seemingly very sharp contradiction with the rigorously proved lower bounds of Complexity Theory.

Anyway, now everything looks almost evident. Indeed, in accordance to Rule 1 in (15), we need to find a class of very easy problems. And this class can be discovered *directly* in the table (6)! To see that, let us compare the complexity of the classes $\mathcal{C}_1$ and $\mathcal{C}_2$ for the accuracy of 1% ($\epsilon = 10^{-2}$). Note that in this case, the accuracy-dependent factors in the efficiency estimates vary from ten to ten thousands. So, the natural question is:

Can the easy problems *from $\mathcal{C}_2$ help us somehow in finding an approximate solution to the* difficult problems *from $\mathcal{C}_1$?*

And the evident answer is: Yes, of course! It is a simple exercise in Calculus to show that we can always approximate a Lipschitz-continuous nonsmooth convex function on a bounded convex set with a uniform accuracy $\epsilon > 0$ by a smooth convex function with Lipschitz-continuous gradient. We pay for the accuracy of approximation by a large Lipschitz constant M for the gradient, which should be of the order $O(\frac{1}{\epsilon})$. Putting this bound for M in the efficiency estimate of $\mathcal{C}_2$ in (6), we can see that in principle, it is possible to minimize nonsmooth convex functions by the oracle-based gradient methods with analytical complexity $O(\frac{1}{\epsilon})$. But what about the Complexity Theory? It seems that it was *proved* that such efficiency is just impossible.

It is interesting that in fact we do not get any contradiction. Indeed, in order to minimize a smooth approximation of nonsmooth function by an oracle-based scheme, we need to change the initial oracle. Therefore, from mathematical point of view, we violate the Black-Box assumption. On the other hand, in the majority of practical applications this change is not difficult. Usually we can work directly with the structure of our problem, at least in the cases when it is created by ourselves.

Thus, the basis of the *smoothing technique* [9, 10] is formed by two ingredients: the above observation, and a trivial but systematic way for approximating a nonsmooth function by a smooth one. This can be done for convex functions represented explicitly in a max-form:

$$f(x) \quad = \quad \max_{u \in Q_d} \{ \langle Ax - b, u \rangle - \phi(u) \},$$

where Q_d is a bounded and convex dual feasible set and $\phi(u)$ is a concave function. Then, choosing a nonnegative strongly convex function $d(u)$, we can define a smooth function

$$f_\mu(x) \quad = \quad \max_{u \in Q_d} \{ \langle Ax - b, u \rangle - \phi(u) - \mu \cdot d(u) \}, \quad \mu > 0, \tag{21}$$

which approximates the initial objective. Indeed, denoting $D_d = \max\limits_{u \in Q_d} d(u)$, we get

$$f(x) \quad \geq \quad f_\mu(x) \ \geq \ f(x) - \mu D_d.$$

At the same time, the gradient of function f_μ is Lipschitz-continuous with Lipschitz constant of the order of $O(\frac{1}{\mu})$ (see [9]) for details).

Thus, we can see that for an *implementable* definition (21), we get a possibility to solve problem (1) in $O(\frac{1}{\epsilon})$ iterations of the fast gradient method (7). In order to see the magnitude of improvement, let us look at the following example:

$$\min_{x \in \Delta_n} \left[f(x) \overset{\text{def}}{=} \max_{1 \leq j \leq m} \langle a_j, x \rangle \right], \tag{22}$$

where $\Delta_n \in R^n$ is a standard simplex. Then the properly implemented smoothing technique ensures the following rate of convergence:

$$f(x_N) - f^* \quad \leq \tfrac{4\sqrt{\ln n \cdot \ln m}}{N} \cdot \max_{i,j} |a_j^{(i)}|.$$

If we apply to problem (22) the standard subgradient methods (e.g. [14]), we can guarantee only

$$f(x_N) - f^* \quad \leq \tfrac{\sqrt{\ln n}}{\sqrt{N+1}} \cdot \max_{i,j} |a_j^{(i)}|.$$

Thus, up to a logarithmic factor, for obtaining the same accuracy, the methods based on smoothing technique need only a square root of iterations of the usual subgradient scheme. Taking into account, that usually the subgradient methods are allowed to run many thousands or even millions of iterations, the gain of the smoothing technique in computational time can be enormously big.[4]

It is interesting, that for problem (22) the computation of the smooth approximation is very cheap. Indeed, let us use for smoothing the *entropy function*:

$$d(u) \quad = \quad \ln m + \sum_{i=1}^{n} u^{(i)} \ln u^{(i)}, \quad u \in \Delta_m.$$

Then the smooth approximation (21) of the objective function in (22) has the following compact representation:

$$f_\mu(x) \quad = \quad \mu \ln \left[\frac{1}{m} \sum_{j=1}^{m} e^{\langle a_j, x\rangle/\mu} \right].$$

Thus, the complexity of the oracle for $f(x)$ and $f_\mu(x)$ is similar. Note that, as in the polynomial-time IPM theory, we apply the standard oracle-based method ((7) in this case) to a function which does not satisfy the Black-Box assumptions.

5. Conclusion

Let us briefly look at one more example of acceleration strategy in Structural Optimization.

Consider the problem of minimizing the *composite* objective function:

$$\min_{x \in R^n} [\, f(x) + \Psi(x) \,], \tag{23}$$

where the function f is a convex differentiable function on $\mathrm{dom}\,\Psi$ with Lipschitz-continuous gradient, and function Ψ is an *arbitrary* closed convex function. Since Ψ can be even discontinuous, in general this problem is very difficult. However, if we assume that function Ψ is *simple*, then the situation is changing. Indeed, suppose that for any $\bar{y} \in \mathrm{dom}\,\Psi$ we are able to solve explicitly the following auxiliary optimization problem:

$$\min_{x \in \mathrm{dom}\,\Psi} \left[f(\bar{y}) + \langle \nabla f(\bar{y}), x - \bar{y}\rangle + \tfrac{M}{2}\|x - \bar{y}\|^2 + \Psi(x) \right] \tag{24}$$

[4] It is easy to see that the standard subgradient methods for nonsmooth convex minimization need indeed $O(\frac{1}{\epsilon^2})$ operations to converge. Consider a univariate function $f(x) = |x|$, $x \in R$. Let us look at the subgradient process:

$$x_{k+1} = x_k - h_k f'(x_k), \quad x_0 = 1, \quad h_k = \tfrac{1}{\sqrt{k+1}} + \tfrac{1}{\sqrt{k+2}}, \quad k \geq 0.$$

It easy to see that $|x_k| = \frac{1}{\sqrt{k+1}}$. However, the step-size sequence is optimal [6].

(compare with (7)). Then it becomes possible to develop for problem (23) the fast gradient methods (similar to (7)), which have the rate of convergence of the order $O(\frac{1}{k^2})$ (see [11] for details; similar technique was developed in [3]). Note that the formulation (23) can be also seen as a part of Structural Optimization since we use the knowledge of the structure of its objective function directly in the optimization methods.

In this paper, we have considered several examples of significant acceleration of the usual oracle-based methods. Note that the achieved progress is visible only because of the supporting complexity analysis. It is interesting that all these methods have some prototypes proposed much earlier:

- Optimal method (7) is very similar to the *heavy point* method:

  $$x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1}),$$

  where α and β are some fixed positive coefficients (see [17] for historical details).

- Polynomial-time IPM are very similar to *some* variants of the classical barrier methods [4].

- The idea to apply smoothing for solving minimax problems is also not new (see [18] and the references therein).

At certain moments of time, these ideas were quite new and attractive. However, they did not result in a significant change in computational practice since they were not provided with a convincing complexity analysis. Indeed, many other schemes have similar theoretical justifications and it was not clear at all why these particular suggestions deserve more attention. Moreover, even now, when we know that the modified variants of some old methods give excellent complexity results, we cannot say too much about the theoretical efficiency of the original schemes.

Thus, we have seen that in Convex Optimization the complexity analysis plays an important role in *selecting* the promising optimization methods among hundreds of others. Of course, it is based on investigation of the worst-case situation. However, even this limited help is important for choosing the perspective directions for further research. This is true especially now, when the development of Structural Optimization makes the problem settings and corresponding efficiency estimates more and more interesting and diverse.

The size of this paper does not allow us to discuss other interesting setting of Structural Convex Optimization (e.g. optimization in relative scale [12, 13]). However, we hope that even the presented examples can help the reader to find new and interesting research directions in this promising field (see, for example, [1, 2, 5]).

References

[1] A. d'Aspremont, O. Banerjee, and L. El Ghaoui. First-Order Methods for Sparse Covariance Selection. *SIAM Journal on Matrix Analysis and its Applications*, **30**(1), 56–66, (2008).

[2] O. Banerjee, L. El Ghaoui, and A. d'Aspremont. Model Selection Through Sparse Maximum Likelihood Estimation. *Journal of Machine Learning Research*, **9**, 485–516 (2008).

[3] A. Beck and M. Teboulle. A Fast Iterative Shrinkage-Threshold Algorithm Linear Inverse Problems. Research Report, Technion (2008).

[4] A.V. Fiacco and G.P. McCormick. Nonlinear Programming: Sequential Unconstrained Minimization Technique. *John Wiley*, New York, 1968.

[5] S. Hoda, A. Gilpin, and J. Pena. Smoothing techniques for computing Nash equilibria of sequential games. Research Report. Carnegie Mellon University, (2008).

[6] Nemirovsky A, Yudin D. Problems complexity and method efficiency in Optimization. 1983. Willey-Interscience, New York.

[7] Yu. Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $O(\frac{1}{k^2})$. *Doklady AN SSSR* (translated as Soviet Mathematics Doklady), **269**(3), 543–547 (1983).

[8] Yu. Nesterov. Introductory Lectures on Convex Optimization. *Kluwer*, Boston, 2004.

[9] Yu. Nesterov. Smooth minimization of non-smooth functions. CORE Discussion Paper 2003/12 (2003). Published in *Mathematical Programming*, **103** (1), 127–152 (2005).

[10] Yu. Nesterov. Excessive gap technique in nonsmooth convex minimization. *SIAM Journal on Optimization*, **16** (1), 235–249 (2005).

[11] Yu. Nesterov. Gradient methods for minimizing composite objective function. *CORE Discussion Paper* 2007/76, (2007).

[12] Yu. Nesterov. Rounding of convex sets and efficient gradient methods for linear programming problems.*Optimization Methods and Software*, **23**(1), 109–135 (2008).

[13] Yu. Nesterov. Unconstrained convex minimization in relative scale. *Mathematics of Operations Research*, **34**(1), 180–193 (2009).

[14] Yu. Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, **120**(1), 261–283 (2009).

[15] Yu. Nesterov, A. Nemirovskii. Interior point polynomial methods in convex programming: Theory and Applications, SIAM, Philadelphia, 1994.

[16] B. Polyak. A general method of solving extremum problems. *Soviet Mat. Dokl.* **8**, 593–597 (1967)

[17] B. Polyak. Introduction to Optimization. *Optimization Software*, New York, 1987.

[18] R. Polyak. Smooth Optimization Methods for Minimax Problems. *SIAM J. Control and Optimization*, **26**(6), 1274–1286 (1988).

[19] N.Z. Shor. Minimization Methods for Nondifferentiable Functions. *Springer-Verlag*, Berlin, 1985.

[20] S.P. Tarasov, L.G. Khachiyan, and I.I. Erlikh. The Method of Inscribed Ellipsoids. *Soviet Mathematics Doklady*, **37**, 226–230 (1988).