ECORE
INTERNATIONAL ASSOCIATION FOR RESEARCH AND TEACHING
Economics, Finance, Operations Research, Econometrics and Statistics

++ research ++ teaching ++

ECORE DISCUSSION PAPER

2011/16

Random Gradient-Free Minimization of Convex Functions

Yurii NESTEROV

CORE
Université catholique de Louvain

ECARES
Université Libre de Bruxelles

CORE DISCUSSION PAPER
2011/1

Random gradient-free minimization of convex functions

Yu. NESTEROV[1]

January 2011

Abstract

In this paper, we prove the complexity bounds for methods of Convex Optimization based only on computation of the function value. The search directions of our schemes are normally distributed random Gaussian vectors. It appears that such methods usually need at most n times more iterations than the standard gradient methods, where n is the dimension of the space of variables. This conclusion is true both for nonsmooth and smooth problems. For the later class, we present also an accelerated scheme with the expected rate of convergence $O(n^2/k^2)$, where k is the iteration counter. For Stochastic Optimization, we propose a zero-order scheme and justify its expected rate of convergence $O(n/k^{1/2})$. We give also some bounds for the rate of convergence of the random gradient-free methods to stationary points of nonconvex functions, both for smooth and nonsmooth cases. Our theoretical results are supported by preliminary computational experiments.

Keywords: convex optimization, stochastic optimization, derivative-free methods, random methods, complexity bounds.

---

[1] Université catholique de Louvain, CORE, B-1348 Louvain-la-Neuve, Belgium. E-mail: yurii.nesterov@uclouvain.be. This author is also member of ECORE, the association between CORE and ECARES.

The research results presented in this paper have been supported by a grant "Action de recherche concertées ARC 04/09-315" from the "Direction de la recherche scientifique – Communauté française de Belgique".

This paper presents research results of the Belgian Program on Interuniversity Poles of Attraction initiated by the Belgian State, Prime Minister's Office, Science Policy Programming. The scientific responsibility is assumed by the author.

1 Introduction

Motivation. Derivative-free optimization methods were among the first schemes suggested in the early days of the development of Optimization Theory (e.g. [4]). These methods have an evident advantage of a simple preparatory stage (the program of computation of the function value is always much simpler than the program for computing the vector of the gradient). However, very soon it was realized that these methods are much more difficult for theoretical investigation. For example, the famous method by Nelder and Mead [5] has only an empirical justification up to now. Moreover, the possible rate of convergence of the derivative-free methods (established usually on an empirical level) is far below the efficiency of the usual optimization schemes.

On the other hand, as it was established in the beginning of 1980s, any function, represented by an explicit sequence of differentiable operations, can be automatically equipped with a program for computing the whole vector of its partial derivatives. Moreover, the complexity of this program is at most four times bigger than the complexity of computation of the initial function (this technique is called *Fast Differentiation*). It seems that this observation destroyed the last arguments for supporting the idea of derivative-free optimization. During several decades, these methods were almost out of computational practice.

However, in the last years, we can see a restoration of the interest to this topic. The current state of the art in this field was recently updated by a comprehensive monograph [2]. It appears that, despite to very serious theoretical objections, the derivative-free methods can probably find their place on the software market. For that, there exist at least several reasons.

- In many applied fields, there exist some models, which are represented by an old black-box software for computing only the values of the functional characteristics of the problem. Modification of this software is either too costly, or impossible.
- There exist some restrictions for applying the Fast Differentiation technique. In particular, it is necessary to store the results of *all* intermediate computations. Clearly, for some applications, this is impossible by memory limitations.
- In any case, creation of a program for computing partial derivatives requires some (substantial) efforts of a qualified programmer. Very often his/her working time is much more expensive than the computational time. Therefore, in some situations it is reasonable to buy a cheaper software preparing for waiting more for the computational results.
- Finally, the extension of the notion of the gradient onto nonsmooth case is a nontrivial operation. The generalized gradient *cannot* be formed by partial derivatives. The most popular framework for defining the *set* of local differential characteristics (*Clarke subdifferential* [1]) suffers from an incomplete Chain Rule. The only known technique for automatic computations of such characteristics (*lexicographic differentiation* [9]) requires an increase of complexity of function evaluation in $O(n)$ times, where n is the number of variables.

Thus, it is interesting to develop the derivative-free optimization methods and obtain the theoretical bounds for their performance. It is interesting that such bounds are almost

absent in this field (see, for example, [2]). One of a few exception is a derivative-free version of cutting plane method presented in Section 9.2 of [7] and improved by [12].

In this paper, we present several *random* derivative-free methods, and provide them with some complexity bounds for different classes of *convex* optimization problems. As we will see, the complexity analysis is crucial for finding the reasonable values of their parameters.

Our approach can be seen as a combination of several popular ideas. First of all, we mention the *random optimization approach* [4], as applied to the problem

$$\min_{x \in R^n} f(x), \tag{1}$$

where f is a differentiable function. It was suggested to sample a point y randomly around the current position x (in accordance to Gaussian distribution), and move to y if $f(y) < f(x)$. The performance of this technique for nonconvex functions was estimated in [3], and criticized by [13] from the numerical point of view.

Different improvements of the random search idea were discussed in Section 3.4 [11]. In particular, it was mentioned that the scheme

$$x_{k+1} \quad = \quad x_k - h_k \frac{f(x_k + \mu_k u) - f(x_k)}{\mu_k} u, \tag{2}$$

where u is a random vector distributed uniformly over the unit sphere, converges under assumption $\mu_k \to 0$. However, no explicit rules for choosing the parameters were given, and no particular rate of convergence was established.

The main goal of this paper is the complexity analysis of different variants of method (2) and its accelerated versions. We study these methods both for smooth and nonsmooth optimization problems. It appears that the most powerful version of the scheme (2) corresponds to $\mu_k \to 0$. Then we get the following process:

$$x_{k+1} \quad = \quad x_k - h_k f'(x_k, u) u, \tag{3}$$

where $f'(x, u)$ is a directional derivative of function $f(x)$ along $u \in R^n$. As compared with the gradient, directional derivative is a much simpler object. Its exact value can be easily computed even for nonconvex nonsmooth functions by a forward differentiation. Or, it can be approximated very well by finite differences. Note that in the gradient schemes the target accuracy ϵ for problem (1) is not very high. Hence, as we will see, the accuracy of the finite differences can be kept on a reasonable level.

For our technique, it is convenient to work with a normally distributed Gaussian vector $u \in R^n$. Then we can define

$$g_0(x) \quad \stackrel{\text{def}}{=} \quad E_u(f'(x, u)u).$$

It appears that for convex f, vector $g_0(x)$ is always a subgradient of f at x.

Thus, we can treat the process (3) as a method with *random oracle*. Usually, these methods are analyzed in the framework of Stochastic Approximation (see [6] for the state of art of the field). However, our random oracle is very special. The standard assumption in Stochastic Approximation is the boundedness of the second moment of the random estimate $\nabla_x F(x, u)$ of the gradient for the objective function $f(x) = E_u(F(x, u))$:

$$E_u(\|\nabla_x F(x, u)\|_2^2) \quad \leq \quad M^2, \quad x \in R^n. \tag{4}$$

(see, for example, condition (2.5) in [6]). However, in our case, if f is differentiable at x, then

$$E_u(\|g_0(x)\|_2^2) \quad \leq \quad (n+4)\|\nabla f(x)\|_2^2.$$

This relation makes the analysis of our methods much simpler and leads to the faster schemes. In particular, for the method (3) as applied to Lipschitz continuous functions, we can prove that the expected rate of convergence of the objective function is of the order $O(\sqrt{\frac{n}{k}})$. If functions has Lipschitz-continuous gradient, then the rate is increased up to $O(\frac{n}{k})$. If in addition, our function is strongly convex, then we have a global linear rate of convergence. Note that in the smooth case, using the technique of estimate sequences (e.g. Section 2.2 in [8]), we can accelerate method (3) up to convergence rate $O(\frac{n^2}{k^2})$.

For justifying the versions of random search methods with $\mu_k > 0$, we use a smoothed version of the objective function

$$f_\mu(x) \quad = \quad E_u(f(x+\mu u)). \tag{5}$$

This object is classical in Optimization Theory. For the complexity analysis of the random search methods it was used, for example, in Section 9.3 [7][1] However, in their analysis the authors used the first part of the representation

$$\nabla f_\mu(x) \quad = \quad \tfrac{1}{\mu} E_u(f(x+\mu u)u) \stackrel{(!)}{=} \tfrac{1}{\mu} E_u([f(x+\mu u) - f(x)]u).$$

In our analysis, we use the second part, which is bounded in μ. Hence, our conclusions are more optimistic.

Contents. In Section 2, we introduce the *Gaussian smoothing* (5) and study its properties. In particular, for different functional classes, we estimate the error of approximation of the objective function and the gradient with respect to the smoothing parameter μ. In Section 3, we introduce the *random gradient-free oracles*, which are based either on finite differences, or on directional derivatives. The main results of this section are the upper bounds for the expected values of squared norms of these oracles. In Section 4, we apply the simple random search method to a nonsmooth convex optimization problem with simple convex constraints. We show that the scheme (3) works at most in $O(n)$ times slower than the usual subgradient method. For the finite-difference version (2), this factor is increased up to $O(n^2)$. Both methods can be naturally modified to be used for Stochastic Programming Problems.

In Section 5, we estimate the performance of method (2) on smooth optimization problems. We show that, under proper choice of parameters, it works at most n times slower than the usual Gradient Method. In Section 6, we consider an accelerated version of this scheme with the convergence rate $O(\frac{n^2}{k^2})$. For all methods we derive the upper bounds for the value of the smoothing parameter μ. It appears that in all situations their dependence in ϵ and n is quite moderate. For example, for the fast random search presented in Section 6, the average size of the trial step μu is of the order $O(n^{-1/2}\epsilon^{3/4})$, where ϵ is the target accuracy for solving (1). For the simple random, this average size is even better: $O(n^{-1/2}\epsilon^{1/2})$.

[1] In [7], u was uniformly distributed over a unit ball. In our comparison, we use a direct translation of the constructions in [7] into the language of the normal Gaussian distribution.

In Section 7 we estimate a rate of convergence the random search methods to a stationary point of nonconvex function (in terms of the norm of the gradient). We consider both smooth and nonsmopth cases. Finally, in Section 8, we present the preliminary computational results. In the tested methods, we were checking the validity of our theoretical conclusions on stability and the rate of convergence of the scheme, as compared with the prototype gradient methods.

Notation. For a finite-dimensional space E, we denote by E^* its dual space. The value of a linear function $s \in E^*$ at point $x \in E$ is denoted by $\langle s, x \rangle$. We endow the spaces E and E^* with Euclidean norms

$$\|x\| \;=\; \langle Bx, x \rangle^{1/2},\; x \in E, \quad \|s\|_* \;=\; \langle s, B^{-1}s \rangle^{1/2},\; s \in E^*,$$

where $B = B^* \succ 0$ is a linear operator from E to E^*. For any $u \in E$ we denote by uu^* a linear operator from E^* to E, which acts as follows:

$$uu^*(s) \;=\; u \cdot \langle s, u \rangle, \quad s \in E^*.$$

In this paper, we consider functions with different level of smoothness. It is indicated by the following notation.

- $f \in C^{0,0}(E)$ if $|f(x) - f(y)| \le L_0(f)\|x - y\|$, $x, y \in E$.
- $f \in C^{1,1}(E)$ if $\|\nabla f(x) - \nabla f(y)\|_* \le L_1(f)\|x - y\|$, $x, y \in E$. This condition is equivalent to the following inequality:

$$|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| \;\le\; \tfrac{1}{2} L_1(f)\|x - y\|^2, \quad x, y \in E. \tag{6}$$

- $f \in C^{2,2}(E)$ if $\|\nabla^2 f(x) - \nabla^2 f(y)\| \le L_2(f)\|x - y\|$, $x, y \in E$. This condition is equivalent to the inequality

$$\begin{aligned} &|f(y) - f(x) - \langle \nabla f(x), y - x \rangle - \tfrac{1}{2}\langle \nabla^2 f(x)(y - x), y - x \rangle| \\ &\qquad \le\; \tfrac{1}{6} L_2(f)\|x - y\|^3, \quad x, y \in E. \end{aligned} \tag{7}$$

We say that $f \in C^1(E)$ is strongly convex, if for any x and $y \in E$ we have

$$f(y) \;\ge\; f(x) + \langle \nabla f(x), y - x \rangle + \tfrac{\tau(f)}{2}\|y - x\|^2, \tag{8}$$

where $\tau(f) \ge 0$ is the *convexity parameter*.

Let $\epsilon \ge 0$. For convex function f, we denote by $\partial f_\epsilon(x)$ its ϵ-subdifferential at $x \in E$:

$$f(y) \;\ge\; f(x) - \epsilon + \langle g, y - x \rangle, \quad g \in \partial f_\epsilon(x),\; y \in E.$$

If $\epsilon = 0$, we simplify this notation up to $\partial f(x)$.

2 Gaussian smoothing

Consider a function $f : E \to R$. We assume that at each point $x \in E$ it is differentiable along any direction. Let us form its *Gaussian approximation*

$$f_\mu(x) \;=\; \tfrac{1}{\kappa} \int\limits_E f(x + \mu u) e^{-\frac{1}{2}\|u\|^2} du, \tag{9}$$

where

$$\kappa \stackrel{\text{def}}{=} \int\limits_E e^{-\frac{1}{2}\|u\|^2} du = \frac{(2\pi)^{n/2}}{[\det B]^{1/2}}. \tag{10}$$

In this definition, $\mu \geq 0$ plays a role of smoothing parameter. Clearly, $\frac{1}{\kappa} \int\limits_E u e^{-\frac{1}{2}\|u\|^2} du = 0$. Therefore, if f is convex and $g \in \partial f(x)$, then

$$f_\mu(x) \geq \frac{1}{\kappa} \int\limits_E [f(x) + \mu \langle g, u \rangle] e^{-\frac{1}{2}\|u\|^2} du = f(x). \tag{11}$$

Note that in general, f_μ has better properties than f. At least, all initial characteristics of f are preserved by any f_μ with $\mu \geq 0$.

- If f is convex, then f_μ is also convex.
- If $f \in C^{0,0}$, then $f_\mu \in C^{0,0}$ and $L_0(f_\mu) \leq L_0(f)$. Indeed, for all $x, y \in E$ we have

$$|f_\mu(x) - f_\mu(y)| \leq \frac{1}{\kappa} \int\limits_E |f(x + \mu u) - f(y + \mu u)| e^{-\frac{1}{2}\|u\|^2} du \leq L_0(f) \|x - y\|.$$

- If $f \in C^{1,10}$, then $f_\mu \in C^{1,1}$ and $L_1(f_\mu) \leq L_1(f)$:

$$\begin{aligned} \|\nabla f_\mu(x) - \nabla f_\mu(y)\|_* &\leq \frac{1}{\kappa} \int\limits_E \|\nabla f(x + \mu u) - \nabla f(y + \mu u)\|_* e^{-\frac{1}{2}\|u\|^2} du \\ &\leq L_1(f) \|x - y\|, \quad x, y \in E. \end{aligned} \tag{12}$$

From definition (10), we get also the identity

$$\ln \int\limits_E e^{-\frac{1}{2}\langle Bu, u \rangle} du \equiv \frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln \det B.$$

Differentiating this identity in B, we get the following representation:

$$\frac{1}{\kappa} \int\limits_E u u^* e^{-\frac{1}{2}\|u\|^2} du = B^{-1}. \tag{13}$$

Taking a scalar product of this equality with B, we obtain

$$\frac{1}{\kappa} \int\limits_E \|u\|^2 e^{-\frac{1}{2}\|u\|^2} du = n. \tag{14}$$

In what follows, we often need upper bounds for the moments $M_p \stackrel{\text{def}}{=} \frac{1}{\kappa} \int\limits_E \|u\|^p e^{-\frac{1}{2}\|u\|^2} du$. We have exact simple values for two cases:

$$M_0 \stackrel{(10)}{=} 1, \quad M_2 \stackrel{(14)}{=} n. \tag{15}$$

For other cases, we will use the following simple bounds.

Lemma 1 *For $p \in [0, 2]$, we have*

$$M_p \leq n^{p/2}. \tag{16}$$

If $p \geq 2$, then we have two-side bounds

$$n^{p/2} \leq M_p \leq (p + n)^{p/2}. \tag{17}$$

Proof:
Denote $\psi(p) = \ln M_p$. This function is convex in p. Let us represent $p = (1-\alpha)\cdot 0 + \alpha \cdot 2$ (thus, $\alpha = \frac{p}{2}$). For $p \in [0,2]$, we have $\alpha \in [0,1]$. Therefore,

$$\psi(p) \;\leq\; (1-\alpha)\psi(0) + \alpha\psi(2) \;\stackrel{(14)}{=}\; \tfrac{p}{2}\ln n.$$

This is the upper bound (16). If $p \geq 2$, then $\alpha \geq 1$, and $\alpha\psi(2)$ becomes a lower bound for $\psi(p)$. It remains to prove the upper bound in (17).

Let us fix some $\tau \in (0,1)$. Note that for any $t \geq 0$ we have

$$t^p e^{-\frac{\tau}{2}t^2} \;\leq\; (\tfrac{p}{\tau e})^{p/2}. \tag{18}$$

Therefore,

$$M_p \;=\; \tfrac{1}{\kappa}\int\limits_E \|u\|^p e^{-\frac{1}{2}\|u\|^2}du \;=\; \tfrac{1}{\kappa}\int\limits_E \|u\|^p e^{-\frac{\tau}{2}\|u\|^2} e^{-\frac{1-\tau}{2}\|u\|^2}du$$

$$\stackrel{(18)}{\leq} \tfrac{1}{\kappa}(\tfrac{p}{\tau e})^{p/2}\int\limits_E e^{-\frac{1-\tau}{2}\|u\|^2}du \;=\; (\tfrac{p}{\tau e})^{p/2}\tfrac{1}{(1-\tau)^{n/2}}.$$

The minimum of the right-hand side in $\tau \in (0,1)$ is attained at $\tau = \frac{p}{p+n}$. Thus,

$$M_p \;\leq\; (\tfrac{p}{e})^{p/2}\left(1+\tfrac{n}{p}\right)^{p/2}(1+\tfrac{p}{n})^{n/2} \;\leq\; (p+n)^{p/2}.$$

□

Now we can prove the following useful result.

Theorem 1 *Let $f \in C^{0,0}(E)$, then,*

$$|f_\mu(x) - f(x)| \;\leq\; \mu L_0(f) n^{1/2}, \quad x \in E. \tag{19}$$

If $f \in C^{1,1}(E)$, then

$$|f_\mu(x) - f(x)| \;\leq\; \tfrac{\mu^2}{2} L_1(f) n, \quad x \in E. \tag{20}$$

Finally, if $f \in C^{2,2}(E)$, then

$$|f_\mu(x) - f(x) - \tfrac{\mu^2}{2}\langle \nabla^2 f(x), B^{-1}\rangle| \;\leq\; \tfrac{\mu^3}{3} L_2(f)(n+3)^{3/2}, \quad x \in E. \tag{21}$$

Proof:
Indeed, for any $x \in E$ we have $f_\mu(x) - f(x) = \frac{1}{\kappa}\int\limits_E [f(x+\mu u) - f(x)] e^{-\frac{1}{2}\|u\|^2}du$. Therefore,

$$|f_\mu(x) - f(x)| \;\leq\; \tfrac{1}{\kappa}\int\limits_E |f(x+\mu u) - f(x)| e^{-\frac{1}{2}\|u\|^2}du$$

$$\leq\; \tfrac{\mu L_0(f)}{\kappa}\int\limits_E \|u\| e^{-\frac{1}{2}\|u\|^2}du \;\stackrel{(16)}{\leq}\; \mu L_0(f) n^{1/2}.$$

Further, if f is differentiable at x, then

$$f_\mu(x) - f(x) \;=\; \tfrac{1}{\kappa} \int\limits_E [f(x+\mu u) - f(x) - \mu \langle \nabla f(x), u \rangle] e^{-\frac{1}{2}\|u\|^2} du.$$

Therefore,

$$|f_\mu(x) - f(x)| \;\stackrel{(6)}{\leq}\; \tfrac{\mu^2 L_1(f)}{2\kappa} \int\limits_E \|u\|^2 e^{-\frac{1}{2}\|u\|^2} du \;\stackrel{(14)}{=}\; \tfrac{\mu^2 L_1(f)}{2} n.$$

Finally, if f is twice differentiable at x, then

$$\tfrac{1}{\kappa} \int\limits_E [f(x+\mu u) - f(x) - \mu \langle \nabla f(x), u \rangle - \tfrac{\mu^2}{2} \langle \nabla^2 f(x) u, u \rangle] e^{-\frac{1}{2}\|u\|^2} du$$

$$\stackrel{(13)}{=}\; f_\mu(x) - f(x) - \tfrac{\mu^2}{2} \langle \nabla^2 f(x), B^{-1} \rangle.$$

Therefore,

$$|f_\mu(x) - f(x) - \tfrac{\mu^2}{2} \langle \nabla^2 f(x), B^{-1} \rangle| \;\stackrel{(7)}{\leq}\; \tfrac{\mu^3 L_2(f)}{6\kappa} \int\limits_E \|u\|^3 e^{-\frac{1}{2}\|u\|^2} du \;\stackrel{(17)}{=}\; \tfrac{\mu^3 L_1(f)}{6} (n+3)^{3/2}.$$

□

Inequality (21) shows that that the increase of the level of smoothness of function f, as compared with $C^{1,1}(E)$, cannot improve the quality of approximation of f by f_μ. If, for example, f is quadratic and $\nabla^2 f(x) \equiv G$, then

$$f_\mu(x) \;\stackrel{(21)}{=}\; f(x) + \tfrac{\mu^2}{2} \langle G, B^{-1} \rangle.$$

The constant term in this identity can reach the right-hand side of inequality (20).

For any positive μ, function f_μ is differentiable. Let us obtain a convenient expression for its gradient. For that, we rewrite definition (9) in another form by introducing a new integration variable $y = x + \mu u$:

$$f_\mu(x) \;=\; \tfrac{1}{\mu^n \kappa} \int\limits_E f(y) e^{-\frac{1}{2\mu^2}\|y-x\|^2} dy.$$

Then,

$$\begin{aligned} \nabla f_\mu(x) \;&=\; \tfrac{1}{\mu^{n+2}\kappa} \int\limits_E f(y) e^{-\frac{1}{2\mu^2}\|y-x\|^2} B(y-x)\, dy \\ &=\; \tfrac{1}{\mu\kappa} \int\limits_E f(x+\mu u) e^{-\frac{1}{2}\|u\|^2} Bu\, du \qquad (22) \\ &=\; \tfrac{1}{\kappa} \int\limits_E \tfrac{f(x+\mu u) - f(x)}{\mu} e^{-\frac{1}{2}\|u\|^2} Bu\, du. \end{aligned}$$

It appears that this gradient is Lipschitz-continuous.

Lemma 2 *Let $f \in C^{0,0}(E)$ and $\mu > 0$. Then $f_\mu \in C^{1,1}(E)$ with*

$$L_1(f_\mu) \;=\; \tfrac{2n^{1/2}}{\mu} L_0(f). \qquad (23)$$

Proof:
Indeed, for all x and y in E, we have

$$\begin{aligned} \|\nabla f_\mu(x) - \nabla f_\mu(y)\|_* &\overset{(22)}{\leq} \frac{1}{\kappa\mu} \int\limits_E |f(x+\mu u) - f(x) + f(y) - f(y+\mu u)| \|u\| e^{-\frac{1}{2}\|u\|^2} du \\ &\leq \frac{2}{\kappa\mu} L_0(f) \int\limits_E \|u\| e^{-\frac{1}{2}\|u\|^2} du. \end{aligned}$$

It remains to apply (16). □

Denote by $f'(x,u)$ the directional derivative of f at point x along direction u:

$$f'(x,u) = \lim_{\alpha \downarrow 0} \frac{1}{\alpha}[f(x+\alpha u) - f(x)]. \tag{24}$$

Then we can define the limiting vector of the gradients (22):

$$\nabla f_0(x) = \frac{1}{\kappa} \int\limits_E f'(x,u) e^{-\frac{1}{2}\|u\|^2} Bu \; du. \tag{25}$$

Note that at each $x \in E$ the vector (25) is uniquely defined. If f is differentiable at x, then

$$\nabla f_0(x) = \frac{1}{\kappa} \int\limits_E \langle \nabla f(x), u \rangle e^{-\frac{1}{2}\|u\|^2} Bu \; du \overset{(13)}{=} \nabla f(x). \tag{26}$$

Let us prove that in convex case $\nabla f_\mu(x)$ always belongs to some ϵ-subdifferential of function f.

Theorem 2 *Let f be convex and Lipschitz continuous. Then, for any $x \in E$ and $\mu \geq 0$ we have*

$$\nabla f_\mu(x) \in \partial_\epsilon f(x), \quad \epsilon = \mu L_0(f) n^{1/2}.$$

Proof:
Let $\mu > 0$. Since f_μ is convex, for all x and $y \in E$ we have

$$f(y) + \mu L_0(f) n^{1/2} \overset{(19)}{\geq} f_\mu(y) \geq f_\mu(x) + \langle \nabla f_\mu(x), y - x \rangle \overset{(11)}{\geq} f(x) + \langle \nabla f_\mu(x), y - x \rangle.$$

Taking now the limit as $\mu \to 0$, we prove the statement for $\mu = 0$. □

Note that expression (22) can be rewritten in the following form:

$$\begin{aligned} \nabla f_\mu(x) &= \frac{1}{\kappa} \int\limits_E \frac{f(x) - f(x-\mu u)}{\mu} e^{-\frac{1}{2}\|u\|^2} Bu \; du \\ &\overset{(22)}{=} \frac{1}{\kappa} \int\limits_E \frac{f(x+\mu u) - f(x-\mu u)}{2\mu} e^{-\frac{1}{2}\|u\|^2} Bu \; du. \end{aligned} \tag{27}$$

Lemma 3 *If $f \in C^{1,1}(E)$ with constant $L_1(f)$, then*

$$\|\nabla f_\mu(x) - \nabla f(x)\|_* \;\leq\; \tfrac{\mu}{2} L_1(f)(n+3)^{3/2}. \tag{28}$$

For $f \in C^{2,2}(E)$ with constant $L_2(f)$, we can guarantee that

$$\|\nabla f_\mu(x) - \nabla f(x)\|_* \;\leq\; \tfrac{\mu^2}{6} L_2(f)(n+4)^2. \tag{29}$$

Proof:
Indeed, for function $f \in C^{1,1}(E)$, we have

$$\begin{aligned}\|\nabla f_\mu(x) - \nabla f(x)\|_* &\overset{(26)}{\leq} \tfrac{1}{\kappa\mu} \int\limits_E |f(x+\mu u) - f(x) - \mu\langle \nabla f(x), u\rangle| \cdot \|u\| e^{-\frac{1}{2}\|u\|^2} du \\ &\overset{(6)}{\leq} \tfrac{\mu L_1(f)}{2\kappa} \int\limits_E \|u\|^3 e^{-\frac{1}{2}\|u\|^2} du \overset{(17)}{\leq} \tfrac{\mu}{2} L_1(f)(n+3)^{3/2}.\end{aligned}$$

Let $f \in C^{2,2}(E)$. Denote $a_u(\tau) = f(x+\tau u) - f(x) - \tau\langle \nabla f(x), u\rangle - \frac{\tau^2}{2}\langle \nabla^2 f(x)u, u\rangle$. Then, $|a_u(\pm\mu)| \overset{(7)}{\leq} \frac{\mu^3}{6} L_2(f)\|u\|^3$. Since

$$\nabla f_\mu(x) - \nabla f(x) \overset{(13)}{=} \tfrac{1}{2\kappa\mu} \int\limits_E [f(x+\mu u) - f(x-\mu u) - 2\mu\langle \nabla f(x), u\rangle] \cdot Bu e^{-\frac{1}{2}\|u\|^2} du,$$

we have

$$\begin{aligned}\|\nabla f_\mu(x) - \nabla f(x)\|_* &\leq \tfrac{1}{2\kappa\mu} \int\limits_E |f(x+\mu u) - f(x-\mu u) - 2\mu\langle \nabla f(x), u\rangle| \cdot \|u\| e^{-\frac{1}{2}\|u\|^2} du \\ &= \tfrac{1}{2\kappa\mu} \int\limits_E |a_u(\mu) - a_u(-\mu)| \cdot \|u\| e^{-\frac{1}{2}\|u\|^2} du \\ &\leq \tfrac{\mu^2 L_2(f)}{6\kappa} \int\limits_E \|u\|^4 e^{-\frac{1}{2}\|u\|^2} du \overset{(17)}{\leq} \tfrac{\mu^2}{6} L_2(f)(n+4)^2.\end{aligned}$$

□

Finally, we prove one more relation between the gradients of f and f_μ.

Lemma 4 *Let $f \in C^{1,1}(E)$, Then, for any $x \in E$ we have*

$$\|\nabla f(x)\|_*^2 \;\leq\; 2\|\nabla f_\mu(x)\|_*^2 + \tfrac{\mu^2}{2} L_1^2(f)(n+4)^2. \tag{30}$$

Proof:

Indeed,

$$\|\nabla f(x)\|_*^2 \overset{(13)}{=} \|\tfrac{1}{\kappa}\int\limits_E \langle \nabla f(x), u\rangle B u e^{-\frac{1}{2}\|u\|^2} du\|_*^2$$

$$= \|\tfrac{1}{\kappa\mu}\int\limits_E \left([f(x+\mu u) - f(x)] - [f(x+\mu u) - f(x) - \mu\langle \nabla f(x), u\rangle]\right) B u e^{-\frac{1}{2}\|u\|^2} du\|_*^2$$

$$\overset{(27)}{\leq} 2\|\nabla f_\mu(x)\|_*^2 + \tfrac{2}{\mu^2}\|\tfrac{1}{\kappa}\int\limits_E [f(x+\mu u) - f(x) - \mu\langle \nabla f(x), u\rangle] B u e^{-\frac{1}{2}\|u\|^2} du\|_*^2$$

$$\leq 2\|\nabla f_\mu(x)\|_*^2 + \tfrac{2}{\mu^2\kappa}\int\limits_E [f(x+\mu u) - f(x) - \mu\langle \nabla f(x), u\rangle]^2 \|u\|^2 e^{-\frac{1}{2}\|u\|^2} du$$

$$\overset{(6)}{\leq} 2\|\nabla f_\mu(x)\|_*^2 + \tfrac{\mu^2}{2} L_1^2(f) M_4.$$

It remains to use inequality (17). □

3 Random gradient-free oracles

Let random vector $u \in E$ have Gaussian distribution with correlation operator B^{-1}. Denote by $E_u(\psi(u))$ the expectation of corresponding random variable. For $\mu \geq 0$, using expressions (22), (27), and (25), we can define the following *random gradient-free oracles*:

1. Generate random $u \in E$ and return $g_\mu(x) = \frac{f(x+\mu u) - f(x)}{\mu} \cdot Bu$.

2. Generate random $u \in E$ and return $\hat{g}_\mu(x) = \frac{f(x+\mu u) - f(x-\mu u)}{2\mu} \cdot Bu$. (31)

3. Generate random $u \in E$ and return $g_0(x) = f'(x,u) \cdot Bu$.

As we will see later, oracles g_μ and $\hat{g}_\mu$ are more suitable for minimizing smooth functions. Oracle g_0 is more universal. It can be also used for minimizing nonsmooth convex functions. Recall that in view of (25) and Theorem 2, we have

$$E_u(g_0(x)) \;=\; \nabla f_0(x) \;\in\; \partial f(x). \tag{32}$$

We can establish now the following upper bounds.

Theorem 3 *1. If f is differentiable at x, then*

$$E_u(\|g_0(x)\|_*^2) \;\leq\; (n+4)\|\nabla f_0(x)\|_*^2. \tag{33}$$

2. Let f be convex. Denote $D(x) = \operatorname{diam} \partial f(x)$. Then, for any $x \in E$ we have

$$E_u(\|g_0(x)\|_*^2) \;\leq\; (n+4)\left(\|\nabla f_0(x)\|_*^2 + nD^2(x)\right). \tag{34}$$

Proof:
Indeed, let us fix $\tau \in (0,1)$. Then,

$$\begin{aligned} E_u(\|g_0(x)\|_*^2) &\overset{(31)}{=} \tfrac{1}{\kappa} \int\limits_E \|u\|^2 e^{-\frac{1}{2}\|u\|^2} f'(x,u)^2 du \\ &= \tfrac{1}{\kappa} \int\limits_E \|u\|^2 e^{-\frac{\tau}{2}\|u\|^2} f'(x,u)^2 e^{-\frac{1-\tau}{2}\|u\|^2} du \\ &\overset{(18)}{\leq} \tfrac{2}{\kappa\tau e} \int\limits_E f'(x,u)^2 e^{-\frac{1-\tau}{2}\|u\|^2} du \\ &= \tfrac{2}{\kappa\tau(1-\tau)^{1+n/2} e} \int\limits_E f'(x,u)^2 e^{-\frac{1}{2}\|u\|^2} du. \end{aligned}$$

The minimum of the right-hand side in τ is attained for $\tau_* = \frac{2}{n+4}$. In this case,

$$\tau_*(1-\tau_*)^{\frac{n+2}{2}} = \tfrac{2}{n+4} \left(\tfrac{n+2}{n+4}\right)^{\frac{n+2}{2}} > \tfrac{2}{(n+4)e}.$$

Therefore,

$$E_u(\|g_0(x)\|_*^2) \leq \tfrac{n+4}{\kappa} \int\limits_E f'(x,u)^2 e^{-\frac{1}{2}\|u\|^2} du.$$

If f is differentiable at x, then $f'(x,u) = \langle \nabla f(x), u \rangle$, and we get (33) from (13).

Suppose that f is convex and not differentiable at x. Denote

$$g(u) \in \operatorname{Arg}\max_g \{\langle g, u \rangle : \; g \in \partial f(x)\}.$$

Then $f'(x,u)^2 = (\langle \nabla f_0(x), u \rangle + \langle g(u) - \nabla f_0(x), u \rangle)^2$. Note that

$$\begin{aligned} E_u(\langle \nabla f_0(x), u \rangle \cdot \langle g(u) - \nabla f_0(x), u \rangle) &\overset{(13)}{=} E_u(\langle \nabla f_0(x), u \rangle \cdot f'(x,u)) - \|\nabla f_0(x)\|_*^2 \\ &= \langle \nabla f_0(x), E_u(u \cdot f'(x,u)) \rangle - \|\nabla f_0(x)\|_*^2 \\ &\overset{(25)}{=} 0. \end{aligned}$$

Therefore,

$$\begin{aligned} E_u(\|g_0(x)\|_*^2) &\leq \tfrac{n+4}{\kappa} \int\limits_E (\langle \nabla f_0(x), u \rangle^2 + D^2(x)\|u\|^2) e^{-\frac{1}{2}\|u\|^2} du \\ &\overset{(13)}{=} (n+4) \left(\|\nabla f_0(x)\|_*^2 + \tfrac{D^2(x)}{\kappa} \int\limits_E \|u\|^2 e^{-\frac{1}{2}\|u\|^2} du \right) \\ &\overset{(14)}{=} (n+4) \left(\|\nabla f_0(x)\|_*^2 + n D^2(x) \right). \end{aligned}$$

□

Let us prove now the similar bounds for oracles g_μ and $\hat{g}_\mu$.

Theorem 4 *1. If $f \in C^{0,0}(E)$, then*

$$E_u(\|g_\mu(x)\|_*^2) \;\leq\; L_0^2(f)(n+4)^2. \tag{35}$$

2. If $f \in C^{1,1}(E)$, then

$$\begin{aligned} E_u(\|g_\mu(x)\|_*^2) &\leq \tfrac{\mu^2}{2} L_1^2(f)(n+6)^3 + 2(n+4)\|\nabla f(x)\|_*^2, \\ E_u(\|\hat{g}_\mu(x)\|_*^2) &\leq \tfrac{\mu^2}{8} L_1^2(f)(n+6)^3 + 2(n+4)\|\nabla f(x)\|_*^2. \end{aligned} \tag{36}$$

3. If $f \in C^{2,2}(E)$, then

$$E_u(\|\hat{g}_\mu(x)\|_*^2) \;\leq\; \tfrac{\mu^4}{18} L_2^2(f)(n+8)^4 + 2(n+4)\|\nabla f(x)\|_*^2. \tag{37}$$

Proof:
Note that $E_u(\|g_\mu(x)\|_*^2) = \frac{1}{\mu^2} E_u\left([f(x+\mu u) - f(x)]^2 \|u\|^2\right)$. If $f \in C^{0,0}(E)$, then we obtain (35) directly from the definition of the functional class and (17).

Let $f \in C^{1,1}(E)$. Since

$$\begin{aligned} [f(x+\mu u) - f(x)]^2 &= [f(x+\mu u) - f(x) - \mu\langle \nabla f(x), u\rangle + \mu\langle \nabla f(x), u\rangle]^2 \\ &\overset{(6)}{\leq} 2\left(\tfrac{\mu^2}{2} L_1(f)\|u\|^2\right)^2 + 2\mu^2\langle \nabla f(x), u\rangle^2, \end{aligned}$$

we get

$$\begin{aligned} E_u(\|g_\mu(x)\|_*^2) &\leq \tfrac{\mu^2}{2} L_1^2(f) E_u(\|u\|^6) + 2E_u(\|g_0(x)\|_*^2) \\ &\overset{(17),(33)}{\leq} \tfrac{\mu^2}{2} L_1^2(f)(n+6)^3 + 2(n+4)\|\nabla f(x)\|_*^2. \end{aligned}$$

For the symmetric oracle $\hat{g}_\mu$, we have

$$\begin{aligned} f(x+\mu u) - f(x-\mu u) &= [f(x+\mu u) - f(x)] + [f(x) - f(x-\mu u)] \\ &\overset{(6)}{\leq} [\mu\langle \nabla f(x), u\rangle + \tfrac{\mu^2}{2} L_1(f)\|u\|^2] + \mu\langle \nabla f(x), u\rangle. \end{aligned}$$

Similarly, we have $f(x+\mu u) - f(x-\mu u) \geq 2\mu\langle \nabla f(x), u\rangle - \frac{\mu^2}{2} L_1(f)\|u\|^2$. Therefore,

$$\begin{aligned} E_u(\|\hat{g}_\mu(x)\|_*^2) &= \tfrac{1}{4\mu^2} E_u\left([f(x+\mu u) - f(x-\mu u)]^2 \|u\|^2\right) \\ &\leq \tfrac{1}{2\mu^2}\left[E_u\left(\tfrac{\mu^4}{4} L_1^2(f)\|u\|^6\right) + E_u\left(4\mu^2\langle \nabla f(u), u\rangle^2 \|u\|^2\right)\right] \\ &\overset{(17),(33)}{\leq} \tfrac{\mu^2}{8} L_1^2(f)(n+6)^3 + 2(n+4)\|\nabla f(x)\|_*^2. \end{aligned}$$

Let $f \in C^{2,2}(E)$. We will use notation of Lemma 3. Since

$$\begin{aligned} [f(x+\mu u) - f(x-\mu u)]^2 &= [f(x+\mu u) - f(x-\mu u) - 2\mu\langle \nabla f(x), u\rangle + 2\mu\langle \nabla f(x), u\rangle]^2 \\ &\leq 2[a_u(\mu) - a_u(-\mu)]^2 + 8\mu^2\langle \nabla f(x), u\rangle^2 \\ &\overset{(7)}{\leq} \tfrac{2\mu^6}{9} L_2^2(f)\|u\|^6 + 8\mu^2\langle \nabla f(x), u\rangle^2, \end{aligned}$$

we get

$$E_u(\|\hat{g}_\mu(x)\|_*^2) \leq \frac{\mu^4}{18} L_2^2(f) E_u(\|u\|^8) + 2E_u(\|g_0(x)\|_*^2)$$

$$\overset{(17),(33)}{\leq} \frac{\mu^4}{18} L_2^2(f)(n+8)^4 + 2(n+4)\|\nabla f(x)\|_*^2.$$

□

Corollary 1 *Let $f \in C^{1,1}(E)$. The, for any $x \in E$ we have*

$$E_u(\|g_\mu(x)\|_*^2) \leq 4(n+4)\|\nabla f_\mu(x)\|_*^2 + \frac{3\mu^2}{2} L_1^2(f)(n+5)^3. \tag{38}$$

Proof:
Indeed,

$$E_u(\|g_\mu(x)\|_*^2) \overset{(36)}{\leq} \frac{\mu^2}{2} L_1^2(f)(n+6)^3 + 2(n+4)\|\nabla f(x)\|_*^2$$

$$\overset{(30)}{\leq} \frac{\mu^2}{2} L_1^2(f)(n+6)^3 + 2(n+4)\left(2\|\nabla f_\mu(x)\|_*^2 + \frac{\mu^2}{2} L_1^2(f)(n+4)^2\right).$$

It remains to note that $(n+6)^3 + 2(n+4)^3 \leq 3(n+5)^3$. □

Example $f(x) = \|x\|$, $x = 0$, shows that the pessimistic bound (35) cannot be significantly improved.

4 Random search for nonsmooth and stochastic optimization

From now on, we assume that f is convex. Let us show how to use the oracles (31) for solving the following nonsmooth optimization problem:

$$f^* \overset{\text{def}}{=} \min_{x \in Q} f(x), \tag{39}$$

where $Q \subseteq E$ is a closed convex set, and f is a nonsmooth convex function on E. Denote by $x^* \in Q$ one of its optimal solutions.

Let us choose a sequence of positive steps $\{h_k\}_{k \geq 0}$. Consider the following method.

Method $\mathcal{RS}_\mu$: Choose $x_0 \in Q$. If $\mu = 0$, we need $D(x_0) = 0$.
Iteration $k \geq 0$. a). Generate u_k and corresponding $g_\mu(x_k)$. b). Compute $x_{k+1} = \pi_Q\left(x_k - h_k B^{-1} g_\mu(x_k)\right)$.

(40)

Note that this method generates a random sequence $\{x_k\}_{k\geq 0}$. Denote by

$$\mathcal{U}_k \;=\; (u_0, \ldots, u_k)$$

a random vector composed by i.i.d. variables $\{u_k\}_{k\geq 0}$ attached to each iteration of the scheme. Denote $\phi_0 = f(x_0)$, and $\phi_k \stackrel{\rm def}{=} E_{\mathcal{U}_{k-1}}(f(x_k))$, $k \geq 1$.

Theorem 5 *Let sequence $\{x_k\}_{k\geq 0}$ be generated by $\mathcal{RS}_0$. Then, for any $N \geq 0$ we have*

$$\sum\limits_{k=0}^{N} h_k(\phi_k - f^*) \;\leq\; \tfrac{1}{2}\|x_0 - x^*\|^2 + \tfrac{n+4}{2} L_0^2(f) \sum\limits_{k=0}^{N} h_k^2. \tag{41}$$

Proof:
Let point x_k with $k \geq 1$ be generated after k iterations of the scheme (40). Denote $r_k = \|x_k - x^*\|$. Then

$$r_{k+1}^2 \;\leq\; \|x_k - h_k g_0(x_k) - x^*\|^2 \;=\; r_k^2 - 2h_k\langle g_0(x_k), x_k - x^*\rangle + h_k^2\|g_0(x_k)\|_*^2.$$

Note that function f is differentiable at x_k with probability one. Therefore, using representation (26) and the estimate (33), we get

$$\begin{aligned} E_{u_k}\left(r_{k+1}^2\right) \;&\leq\; r_k^2 - 2h_k\langle \nabla f(x_k), x_k - x^*\rangle + h_k^2(n+4)L_0^2(f) \\ &\leq\; r_k^2 - 2h_k(f(x_k) - f^*) + h_k^2(n+4)L_0^2(f). \end{aligned}$$

Taking now the expectation in $\mathcal{U}_{k-1}$, we obtain

$$E_{\mathcal{U}_k}\left(r_{k+1}^2\right) \;\leq\; E_{\mathcal{U}_{k-1}}\left(r_k^2\right) - 2h_k(\phi_k - f^*) + h_k^2(n+4)L_0^2(f).$$

Using the same reasoning, we get

$$E_{\mathcal{U}_0}\left(r_1^2\right) \;\leq\; r_0^2 - 2h_0(f(x_0) - f^*) + h_0^2(n+4)L_0^2(f).$$

Summing up these inequalities, we come to (41). □

Denote $S_N = \sum\limits_{k=0}^{N} h_k$, and define $\hat{x}_N = \arg\min\limits_x [f(x) :\; x \in \{x_0, \ldots, x_N\}]$. Then

$$\begin{aligned} E_{\mathcal{U}_{N-1}}\left(f(\hat{x}_N)\right) - f^* \;&\leq\; E_{\mathcal{U}_{N-1}}\left(\tfrac{1}{S_N}\sum\limits_{k=0}^{n} h_k(f(x_k) - f^*)\right) \\ &\stackrel{(41)}{\leq}\; \tfrac{1}{S_N}\left[\tfrac{1}{2}\|x_0 - x^*\|^2 + \tfrac{n+4}{2}L_0^2(f)\sum\limits_{k=0}^{N} h_k^2\right]. \end{aligned}$$

In particular, if the number of steps N is fixed, and $\|x_0 - x^*\| \leq R$, we can choose

$$h_k \;=\; \tfrac{R}{(n+4)^{1/2}(N+1)^{1/2}L_0(f)}, \quad k = 0, \ldots, N. \tag{42}$$

Then we obtain the following bound:

$$E_{\mathcal{U}_{N-1}}\left(f(\hat{x}_N)\right) - f^* \;\leq\; L_0(f)R\left[\tfrac{n+4}{N+1}\right]^{1/2}. \tag{43}$$

Hence, inequality $E_{\mathcal{U}_{N-1}}(f(\hat{x}_N)) - f^* \leq \epsilon$ can be ensured by $\mathcal{RS}_0$ in

$$\frac{n+4}{\epsilon^2} L_0^2(f) R^2 \tag{44}$$

iterations.

Same as in the standard nonsmooth minimization, instead of fixing the number of steps apriori, we can define

$$h_k \;=\; \frac{R}{(n+4)^{1/2}(k+1)^{1/2} L_0(f)}, \quad k \geq 0. \tag{45}$$

This modification results in a multiplication of the right-hand side of the estimate (43) by a factor $O(\ln N)$ (e.g. Section 3.2 in [8]).

Let us consider now the random search method (40) with $\mu > 0$.

Theorem 6 *Let sequence $\{x_k\}_{k \geq 0}$ be generated by $\mathcal{RS}_\mu$ with $\mu > 0$. Then, for any $N \geq 0$ we have*

$$\frac{1}{S_N} \sum_{k=0}^{N} h_k(\phi_k - f^*) \;\leq\; \mu L_0(f) n^{1/2} + \frac{1}{S_N} \left[\frac{1}{2}\|x_0 - x^*\|^2 + \frac{(n+4)^2}{2} L_0^2(f) \sum_{k=0}^{N} h_k^2 \right]. \tag{46}$$

Proof:
Let point x_k with $k \geq 1$ be generated after k iterations of the scheme (40). Denote $r_k = \|x_k - x^*\|$. Then

$$r_{k+1}^2 \;\leq\; \|x_k - h_k g_\mu(x_k) - x^*\|^2 \;=\; r_k^2 - 2h_k \langle g_\mu(x_k), x_k - x^* \rangle + h_k^2 \|g_\mu(x_k)\|_*^2.$$

Using representation (22) and the estimate (35), we get

$$\begin{aligned} E_{u_k}\left(r_{k+1}^2\right) \;&\leq\; r_k^2 - 2h_k \langle \nabla f_\mu(x_k), x_k - x^* \rangle + h_k^2 (n+4)^2 L_0^2(f) \\ &\overset{(11)}{\leq}\; r_k^2 - 2h_k (f(x_k) - f_\mu(x^*)) + h_k^2 (n+4)^2 L_0^2(f). \end{aligned}$$

Taking now the expectation in $\mathcal{U}_{k-1}$, we obtain

$$E_{\mathcal{U}_k}\left(r_{k+1}^2\right) \;\leq\; E_{\mathcal{U}_{k-1}}(r_k^2) - 2h_k(\phi_k - f_\mu(x^*)) + h_k^2 (n+4)^2 L_0^2(f).$$

It remains to note that $f_\mu(x^*) \overset{(19)}{\leq} f^* + \mu L_0(f) n^{1/2}$. □

Thus, in order to guarantee inequality $E_{\mathcal{U}_{N-1}}(f(\hat{x}_N)) - f^* \leq \epsilon$ by method $\mathcal{RS}_\mu$, we can choose

$$\begin{aligned} \mu \;&=\; \frac{\epsilon}{2L_0(f) n^{1/2}}, \quad h_k \;=\; \frac{R}{(n+4)(N+1)^{1/2} L_0(f)},\; k = 0, \ldots, N, \\ N \;&=\; \frac{4(n+4)^2}{\epsilon^2} L_0^2(f) R^2. \end{aligned} \tag{47}$$

Note that this complexity bound is in $O(n)$ times worse than the complexity bound (44) of the method $\mathcal{RS}_0$. This can be explained by the different upper bounds provided by inequalities (33) and (35). It is interesting that the smoothing parameter μ is not used

in the definition (47) of the step sizes and in the total length of the process generated by method $\mathcal{RS}_\mu$.

Finally, let us compare our results with the following *Random Coordinate Method*:

$$\begin{array}{l} 1. \ \text{Generate a uniformly distributed number } i_k \in \{1, \dots, n\}. \\ 2. \ \text{Update } x_{k+1} = \pi_Q\left(x_k - h e_{i_k} \langle g(x_k), e_{i_k} \rangle\right), \end{array} \tag{48}$$

where e_i is a coordinate vector in R^n, and $g(x_k) \in \partial f(x_k)$. By the same reasoning as in Theorem 5, we can show that (compare with [10])

$$\frac{1}{N+1} \sum_{k=0}^{N} (\phi_k - f^*) \quad \leq \quad \frac{nR^2}{2(N+1)h} + \frac{h}{2} L_0^2(f).$$

Thus, under an appropriate choice of h, method (48) has the same complexity bound (44) as $\mathcal{RS}_0$. However, note that (48) requires computation of the coordinates of the subgradient $g(x_k)$. This computation *cannot* be arranged with directional derivatives, or with function values. Therefore, method 48) cannot be transformed in a gradient-free form.

A natural modification of method (40) can be applied to the problems of Stochastic Optimization. Indeed, assume that the objective function in (39) has the following form:

$$f(x) \quad = \quad E_\xi \left[F(x, \xi)\right] \stackrel{\text{def}}{=} \int\limits_{\Xi} F(x, \xi) dP(\xi), \quad x \in Q, \tag{49}$$

where ξ is a random vector with probability distribution $P(\xi)$, $\xi \in \Xi$. We assume that $f \in C^{0,0}(E)$ is convex (this is a relaxation of the standard assumption that $F(x, \xi)$ is convex in x for any $\xi \in \Xi$). Similarly to (31), we can define *random stochastic gradient-free oracles*:

$$\begin{array}{ll} \textbf{1.} & \text{Generate random } u \in E,\ \xi \in \Xi. \text{ Return } s_\mu(x) = \frac{F(x+\mu u, \xi) - F(x, \xi)}{\mu} \cdot Bu. \\ \textbf{2.} & \text{Generate random } u \in E,\ \xi \in \Xi. \text{ Return } \hat{s}_\mu(x) = \frac{F(x+\mu u, \xi) - F(x-\mu u, \xi)}{2\mu} \cdot Bu. \\ \textbf{3.} & \text{Generate random } u \in E,\ \xi \in \Xi. \text{ Return } s_0(x) = F'_x(x, \xi) \cdot Bu. \end{array} \tag{50}$$

Consider the following method with smoothing parameter $\mu > 0$.

$$\begin{array}{|l|} \hline \textbf{Method } \mathcal{SS}_\mu\textbf{:} \text{ Choose } x_0 \in Q. \\ \hline \textbf{Iteration } k \geq 0\textbf{.} \\ \text{a). For } x_k \in Q, \text{ generate random vectors } \xi_k \in \Xi \text{ and } u_k. \\ \text{b). Compute } s_\mu(x_k), \text{ and } x_{k+1} = \pi_Q\left(x_k - h_k B^{-1} s_\mu(x_k)\right). \\ \hline \end{array} \tag{51}$$

Its justification is very similar to the proof of Theorem 6.

Theorem 7 *Let $L_0(F(\cdot,\xi)) \le L$ for all $\xi \in \Xi$. Assume the sequence $\{x_k\}_{k\ge 0}$ be generated by $\mathcal{SS}_\mu$ with $\mu > 0$. Then, for any $N \ge 0$ we have*

$$\frac{1}{S_N}\sum_{k=0}^{N} h_k(\phi_k - f^*) \;\le\; \mu L n^{1/2} + \frac{1}{S_N}\left[\frac{1}{2}\|x_0 - x^*\|^2 + \frac{(n+4)^2}{2}L^2\sum_{k=0}^{N} h_k^2\right], \tag{52}$$

where $\phi_k = E_{\mathcal{U}_{k-1},\mathcal{P}_{k-1}}(f(x_k))$, and $\mathcal{P}_k = \{\xi_0,\dots,\xi_k\}$.

Proof:
In the notation of Theorem 6, we have

$$r_{k+1}^2 \;\le\; r_k^2 - 2h\langle s_\mu(x_k), x_k - x^*\rangle + h^2\|s_\mu(x_k)\|^2.$$

In view of our assumptions, $\|s_\mu(x_k)\| \le L\|u_k\|^2$. Since $E_\xi\,(s_\mu(x)) = g_\mu(x)$, we have

$$\begin{aligned} E_{u_k,\xi_k}(r_{k+1}^2) \;&\le\; r_k^2 + E_{u_k}\left(-2h\langle g_\mu(x_k), x_k - x^*\rangle + h^2L^2\|u_k\|^4\right) \\ &\overset{(22),(17)}{\le}\; r_k^2 - 2h_k\langle \nabla f_\mu(x_k), x_k - x^*\rangle + h^2(n+4)L^2 \\ &\le\; r_k^2 - 2h_k(f_\mu(x_k) - f_\mu(x^*)) + h^2(n+4)L^2. \end{aligned}$$

Taking now the expectation in $\mathcal{U}_{k-1}$ and $\mathcal{P}_{k-1}$, we get

$$E_{\mathcal{U}_k,\mathcal{P}_k}(r_{k+1}^2) \overset{(11)}{\le} E_{\mathcal{U}_{k-1},\mathcal{P}_{k-1}}(r_k^2) - 2h_k(\phi_k - f_\mu(x^*)) + h^2(n+4)L^2.$$

It remains to note that $f_\mu(x^*) \overset{(19)}{\le} f^* + \mu L n^{1/2}$. □

Thus, choosing the parameters of method $\mathcal{SS}_\mu$ in accordance to (47), we can solve the Stochastic Programming Problem (39), (49) in $O(\frac{n^2}{\epsilon^2})$ iterations. To the best of our knowledge, method (51) is the first zero-order method in Stochastic Optimization. A similar analysis can be applied to the method $\mathcal{SS}_0$.

5 Simple random search for smooth optimization

Consider the following smooth unconstrained optimization problem:

$$f^* \overset{\text{def}}{=} \min_{x\in E} f(x), \tag{53}$$

where f is a smooth convex function on E. Denote by x^* one of its optimal solutions. For the sake of notation, we assume that $\dim E \ge 2$.

Consider the following method.

Method $\mathcal{RG}_\mu$: Choose $x_0 \in E$.
Iteration $k \geq 0$. a). Generate u_k and corresponding $g_\mu(x_k)$. b). Compute $x_{k+1} = x_k - hB^{-1}g_\mu(x_k)$.

(54)

This is a random version of the standard primal gradient method. A version of method (54) with oracle $\hat{g}_\mu$ will be called $\widehat{\mathcal{RG}}_\mu$.

Since the bounds (36) and (37) are continuous in μ, we can justify all variants of method $\mathcal{RG}_\mu$, $\mu \geq 0$, by a single statement.

Theorem 8 *Let $f \in C^{1,1}(E)$, and sequence $\{x_k\}_{k\geq 0}$ be generated by $\mathcal{RG}_\mu$ with*

$$h \;=\; \tfrac{1}{4(n+4)L_1(f)}. \tag{55}$$

Then, for any $N \geq 0$, we have

$$\tfrac{1}{N+1}\sum_{k=0}^{N}(\phi_k - f^*) \;\leq\; \tfrac{4(n+4)L_1(f)\|x_0-x^*\|^2}{N+1} + \tfrac{9\mu^2(n+4)^2L_1(f)}{25}. \tag{56}$$

Let function f be strongly convex. Denote $\delta_\mu = \frac{18\mu^2(n+4)^2}{25\tau(f)}L_1(f)$. Then

$$\phi_N - f^* \;\leq\; \tfrac{1}{2}L_1(f)\left[\delta_\mu + \left(1 - \tfrac{\tau(f)}{8(n+4)L_1(f)}\right)^N(\|x_0 - x^*\|^2 - \delta_\mu)\right]. \tag{57}$$

Proof:
Let point x_k with $k \geq 0$ be generated after k iterations of the scheme (54). Denote $r_k = \|x_k - x^*\|$. Then

$$r_{k+1}^2 \;=\; r_k^2 - 2h\langle g_\mu(x_k), x_k - x^*\rangle + h^2\|g_\mu(x_k)\|_*^2.$$

Using representation (27) and the estimate (36), we get

$$\begin{aligned}
E_{u_k}\left(r_{k+1}^2\right) \;&\leq\; r_k^2 - 2h\langle \nabla f_\mu(x_k), x_k - x^*\rangle + h^2[\tfrac{\mu^2(n+6)^3}{2}L_1^2(f) + 2(n+4)\|\nabla f(x)\|_*^2] \\
&\overset{(11)}{\leq}\; r_k^2 - 2h(f(x_k) - f_\mu(x^*)) + h^2[\tfrac{\mu^2(n+6)^3}{2}L_1^2(f) + 4(n+4)L_1(f)(f(x_k) - f^*)] \\
&\overset{(20)}{\leq}\; r_k^2 - 2h(1 - 2h(n+4)L_1(f))(f(x_k) - f^*) + \mu^2 nhL_1(f) + \tfrac{\mu^2(n+6)^3}{2}h^2L_1^2(f) \\
&\overset{(55)}{=}\; r_k^2 - \tfrac{f(x_k)-f^*}{4(n+4)L_1(f)} + \tfrac{\mu^2}{4}\left[\tfrac{n}{n+4} + \tfrac{(n+6)^3}{8(n+4)^2}\right] \;\leq\; r_k^2 - \tfrac{f(x_k)-f^*}{4(n+4)L_1(f)} + \tfrac{9\mu^2(n+4)}{100}.
\end{aligned}$$

Taking now the expectation in $\mathcal{U}_{k-1}$, we obtain

$$\rho_{k+1} \stackrel{\text{def}}{=} E_{\mathcal{U}_k}\left(r_{k+1}^2\right) \;\leq\; \rho_k - \tfrac{\phi_k - f^*}{4(n+4)L_1(f)} + \tfrac{9\mu^2(n+4)}{100}.$$

Summing up these inequalities for $k = 0, \ldots, N$, and dividing the result by $N+1$, we get (56).

Assume now that f is strongly convex. As we have seen,

$$E_{u_k}\left(r_{k+1}^2\right) \;\leq\; r_k^2 - \tfrac{f(x_k) - f^*}{4(n+4)L_1(f)} + \tfrac{9\mu^2(n+4)}{100} \stackrel{(8)}{\leq} \left(1 - \tfrac{\tau(f)}{8(n+4)L_1(f)}\right) r_k^2 + \tfrac{9\mu^2(n+4)}{100}.$$

Taking the expectation in $\mathcal{U}_{k-1}$, we get

$$\rho_{k+1} \;\leq\; \left(1 - \tfrac{\tau(f)}{8(n+4)L_1(f)}\right) \rho_k + \tfrac{9\mu^2(n+4)}{100}.$$

This inequality is equivalent to the following one:

$$\rho_{k+1} - \delta_\mu \;\leq\; \left(1 - \tfrac{\tau(f)}{8(n+4)L_1(f)}\right)(\rho_k - \delta_\mu) \;\leq\; \left(1 - \tfrac{\tau(f)}{8(n+4)L_1(f)}\right)^{k+1}(\rho_0 - \delta_\mu).$$

It remains to note that $\phi_k - f^* \stackrel{(6)}{\leq} \frac{1}{2}L_1(s)\rho_k$. □

Let us discuss the choice of parameter μ in method $\mathcal{RG}_\mu$. Consider first the minimization of functions from $C^{1,1}(E)$. Clearly, the estimate (56) is valid also for $\hat{\phi}_N \stackrel{\text{def}}{=} E_{\mathcal{U}_{k-1}}(f(\hat{x}_N))$, where $\hat{x}_N = \arg\min_x [f(x) : \; x \in \{x_0, \ldots, x_N\}]$. In order to get the final accuracy ϵ for the objective function, we need to choose μ sufficiently small:

$$\mu \;\leq\; \tfrac{5}{3(n+4)}\sqrt{\tfrac{\epsilon}{2L_1(f)}}. \tag{58}$$

Taking into account that $E_u(\|u\|) = O(n^{1/2})$, we can see that the average length of the finite-difference step in computation of the oracle g_μ is of the order $O\left(\sqrt{\frac{\epsilon}{nL_1(f)}}\right)$. It is interesting that this bound is much more relaxed with respect to ϵ than the bound (47) for nonsmooth version of the random search. However, it depends now on the dimension of the space of variables. At the same time, inequality $\hat{\phi}_N - f^* \leq \epsilon$ is satisfied at most in $O(\frac{n}{\epsilon}L_1(f)R^2)$ iterations.

Consider now the strongly convex case. Then, we need to choose μ satisfying the equation $\frac{1}{2}L_1(f)\delta_\mu \leq \frac{\epsilon}{2}$. This is

$$\mu \;\leq\; \tfrac{5}{3(n+4)}\sqrt{\tfrac{\epsilon}{2L_1(f)} \cdot \tfrac{\tau(f)}{L_1(f)}}. \tag{59}$$

The number iterations of this method is of the order $O\left(\frac{nL_1(f)}{\tau(f)} \ln \frac{L_1(f)R^2}{\epsilon}\right)$. It is natural that a faster scheme needs a higher accuracy of the finite-difference oracle (or, a smaller value of μ).

The complexity analysis of the method $\widehat{\mathcal{RG}}_\mu$ can be done in a similar way. In accordance to the estimate (36), the corresponding results will have slightly better dependence in μ. Note that our complexity results are also valid for the limiting version $\mathcal{RG}_0 \equiv \widehat{\mathcal{RG}}_0$.

6 Accelerated random search

Let us apply to problem (53) a random variant of the fast gradient method. We assume that function $f \in C^{1,1}(E)$ is strongly convex with convexity parameter $\tau(f) \geq 0$. Denote by $\kappa(f) \stackrel{\text{def}}{=} \frac{\tau(f)}{L_1(f)}$ its *condition number*. And let $\theta_n = \frac{1}{16(n+1)^2 L_1(f)}$, $h_n = \frac{1}{4(n+4)L_1(f)}$.

Method $\mathcal{FG}_\mu$: Choose $x_0 \in E$, $v_0 = x_0$, and a positive $\gamma_0 \geq \tau(f)$.
Iteration $k \geq 0$: a) Compute $\alpha_k > 0$ satisfying $\theta_n^{-1}\alpha_k^2 = (1-\alpha_k)\gamma_k + \alpha_k\tau(f) \equiv \gamma_{k+1}$. b) Set $\lambda_k = \frac{\alpha_k}{\gamma_{k+1}}\tau(f)$, $\beta_k = \frac{\alpha_k\gamma_k}{\gamma_k+\alpha_k\tau(f)}$, and $y_k = (1-\beta_k)x_k + \beta_k v_k$. c). Generate random u_k and compute corresponding $g_\mu(y_k)$. d). Set $x_{k+1} = y_k - h_n B^{-1} g_\mu(y_k)$, $v_{k+1} = (1-\lambda_k)v_k + \lambda_k y_k - \frac{\theta_n}{\alpha_k} B^{-1} g_\mu(y_k)$.

(60)

Note that the parameters of this method satisfy the following relations:

$$1-\lambda_k \;=\; (1-\alpha_l)\tfrac{\gamma_k}{\gamma_{k+1}}, \quad 1-\beta_k \;=\; \tfrac{\gamma_{k+1}}{\gamma_k+\alpha_k\tau(f)}, \quad (1-\lambda_k)\tfrac{1-\beta_k}{\beta_k} \;=\; \tfrac{1-\alpha_k}{\alpha_k}. \tag{61}$$

Theorem 9 *For all $k \geq 0$, we have*

$$\phi_k - f^* \;\leq\; \psi_k \cdot [f(x_0) - f(x^*) + \tfrac{\gamma_0}{2}\|x_0 - x^*\|^2] + \mu^2 L_1(f)\left(n + \tfrac{3(n+8)}{32}C_k\right), \tag{62}$$

where $\psi_k \leq \min\left\{\left(1 - \frac{\kappa^{1/2}(f)}{4(n+4)}\right)^k, \left(1 + \frac{k}{8(n+4)}\sqrt{\frac{\gamma_0}{L_1(f)}}\right)^{-2}\right\}$, *and* $C_k \leq \min\left\{k, \frac{4(n+4)}{\kappa^{1/2}(f)}\right\}$.

Proof:
Assume that after k iterations, we have generated points x_k and v_k. Then we can compute y_k and generate $g_\mu(y_k)$. Taking a random step from this point, we get

$$f_\mu(x_{k+1}) \;\stackrel{(12)}{\leq}\; f_\mu(y_k) - h_n\langle \nabla f_\mu(y_k), B^{-1}g_\mu(x_k)\rangle + \tfrac{h_n^2}{2}L_1(f)\|g_\mu(y_k)\|_*^2.$$

Therefore,

$$E_{u_k}\left(f_\mu(x_{k+1})\right) \;\stackrel{(27)}{\leq}\; f_\mu(y_k) - h_n\|\nabla f_\mu(y_k)\|_*^2 + \tfrac{h_n^2}{2}L_1(f)E_{u_k}\left(\|g_\mu(y_k)\|_*^2\right)$$

$$\stackrel{(38)}{\leq}\; f_\mu(y_k) - \tfrac{h_n}{4(n+4)}\left(E_{u_k}\left(\|g_\mu(y_k)\|_*^2\right) - \tfrac{3\mu^2}{2}L_1^2(f)(n+5)^3\right) + \tfrac{h_n^2}{2}L_1(f)E_{u_k}\left(\|g_\mu(y_k)\|_*^2\right)$$

$$=\; f_\mu(y_k) - \tfrac{1}{2}\theta_n E_{u_k}\left(\|g_\mu(y_k)\|_*^2\right) + \xi_\mu,$$

where $\xi_\mu \stackrel{\text{def}}{=} \frac{3(n+5)^3\mu^2}{32(n+4)^2} L_1(f)$. Note that $\frac{(n+5)^3}{(n+4)^2} \leq n+8$ for $n \geq 2$.

Let us fix an arbitrary $x \in E$. Note that

$$\begin{aligned}
\delta_{k+1}(x) &\stackrel{\text{def}}{=} \tfrac{\gamma_{k+1}}{2}\|v_{k+1}-x\|^2 + f_\mu(x_{k+1}) - f_\mu(x) \\
&= \tfrac{\gamma_{k+1}}{2}\|(1-\lambda_k)v_k + \lambda_k y_k - x\|^2 - \tfrac{\theta_n \gamma_{k+1}}{\alpha_k}\langle g_\mu(y_k), (1-\lambda_k)v_k + \lambda_k y_k - x\rangle \\
&\quad + \tfrac{\theta_n^2 \gamma_{k+1}}{2\alpha_k^2}\|g_\mu(y_k)\|_*^2 + f_\mu(x_{k+1}) - f_\mu(x).
\end{aligned}$$

Taking the expectation in u_k, and using the equation of Step a) in (60), we get

$$\begin{aligned}
E_{u_k}(\delta_{k+1}(x)) &\stackrel{(22)}{\leq} \tfrac{\gamma_{k+1}}{2}\|(1-\lambda_k)v_k + \lambda_k y_k - x\|^2 - \alpha_k\langle \nabla f_\mu(y_k), (1-\lambda_k)v_k + \lambda_k y_k - x\rangle \\
&\quad + \tfrac{1}{2}\theta_n E_{u_k}\left(\|g_\mu(y_k)\|_*^2\right) + E_{u_k}\left(f_\mu(x_{k+1})\right) - f_\mu(x) \\
&\leq \tfrac{\gamma_{k+1}}{2}\|(1-\lambda_k)v_k + \lambda_k y_k - x\|^2 + \alpha_k\langle \nabla f_\mu(y_k), x - (1-\lambda_k)v_k - \lambda_k y_k\rangle \\
&\quad + f_\mu(y_k) - f_\mu(x) + \xi_\mu.
\end{aligned}$$

Note that $v_k = y_k + \frac{1-\beta_k}{\beta_k}(y_k - x_k)$. Therefore,

$$(1-\lambda_k)v_k + \lambda_k y_k = y_k + (1-\lambda_k)\tfrac{1-\beta_k}{\beta_k}(y_k - x_k) \stackrel{(61)}{=} y_k + \tfrac{1-\alpha_k}{\alpha_k}(y_k - x_k).$$

Hence,

$$\begin{aligned}
&f_\mu(y_k) + \alpha_k\langle \nabla f_\mu(y_k), x - (1-\lambda_k)v_k - \lambda_k y_k\rangle - f_\mu(x) \\
&= f_\mu(y_k) + \langle \nabla f_\mu(y_k), \alpha_k x + (1-\alpha_k)x_k - y_k\rangle - f_\mu(x) \\
&\stackrel{(8)}{\leq} (1-\alpha_k)(f(x_k) - f_\mu(x)) - \tfrac{1}{2}\alpha_k\tau(f)\|x - y_k\|^2,
\end{aligned}$$

and we can continue:

$$\begin{aligned}
E_{u_k}(\delta_{k+1}(x)) &\leq \tfrac{\gamma_{k+1}}{2}\|(1-\lambda_k)v_k + \lambda_k y_k - x\|^2 + \xi_\mu \\
&\quad + (1-\alpha_k)(f(x_k) - f_\mu(x)) - \tfrac{1}{2}\alpha_k\tau(f)\|x - y_k\|^2 \\
&\leq \tfrac{\gamma_{k+1}}{2}(1-\lambda_k)\|v_k - x\|^2 + \tfrac{\gamma_{k+1}}{2}\lambda_k\|y_k - x\|^2 + \xi_\mu \\
&\quad + (1-\alpha_k)(f(x_k) - f_\mu(x)) - \tfrac{1}{2}\alpha_k\tau(f)\|x - y_k\|^2 \\
&\stackrel{(61)}{=} (1-\alpha_k)\delta_k(x) + \xi_\mu.
\end{aligned}$$

Denote $\phi_k(\mu) = E_{\mathcal{U}_{k-1}}(f_\mu(x_k))$, $\rho_k = \frac{\gamma_k}{2}E_{\mathcal{U}_{k-1}}(\|v_k - x^*\|^2)$. Then, taking the expectation of the latter inequality in $\mathcal{U}_{k-1}$, we get

$$\begin{aligned}
\phi_{k+1}(\mu) - f_\mu(x) + \rho_{k+1} &\leq (1-\alpha_k)(\phi_k(\mu) - f_\mu(x^*) + \rho_k) + \xi_\mu \\
&\leq \ldots \leq \psi_{k+1}\cdot\left(f_\mu(x_0) - f_\mu(x) + \tfrac{\gamma_0}{2}\|x_0 - x\|^2\right) + \xi_\mu \cdot C_{k+1},
\end{aligned}$$

where $\psi_k = \prod_{i=0}^{k-1}(1-\alpha_i)$, and $C_k = 1 + \sum_{i=1}^{k-1}\prod_{j=k-i}^{k-1}(1-\alpha_j)$, $k \geq 1$. Defining $\psi_0 = 1$ and $C_0 = 0$, we get $C_k \leq k$, $k \geq 0$. On the other hand, by induction it is easy to see that $\gamma_k \geq \tau(f)$ for all $k \geq 0$. Therefore,

$$\alpha_k \;\geq\; [\tau(f)\theta_n]^{1/2} \;=\; \tfrac{\kappa^{1/2}(f)}{4(n+4)} \;\stackrel{\text{def}}{=}\; \omega_n, \quad k \geq 0.$$

Then, $C_k \leq 1 + \sum_{i=1}^{k-1}\prod_{j=k-i}^{k-1}(1-\omega_n)^i = 1 + (1-\omega_n)\frac{(1-(1-\omega_n)^k)}{\omega_n} \leq \omega_n^{-1}$. Thus,

$$C_k \;\leq \min\left\{k, \tfrac{4(n+4)}{\kappa^{1/2}(f)}\right\}, \quad \psi_k \;\leq\; \left(1 - \tfrac{\kappa^{1/2}(f)}{4(n+4)}\right)^k, \quad k \geq 0.$$

Further,[2] let us prove that $\gamma_k \geq \gamma_0\psi_k$. For $k = 0$ this is true. Assume it is true for some $k \geq 0$. Then

$$\gamma_{k+1} \;\geq\; (1-\alpha_k)\gamma_k \;\geq\; \gamma_0\psi_{k+1}.$$

Denote $a_k = \frac{1}{\psi_k^{1/2}}$. Then, in view of the established inequality we have:

$$\begin{aligned} a_{k+1} - a_k \;&=\; \tfrac{\psi_k^{1/2} - \psi_{k+1}^{1/2}}{\psi_k^{1/2}\psi_{k+1}^{1/2}} \;=\; \tfrac{\psi_k - \psi_{k+1}}{\psi_k^{1/2}\psi_{k+1}^{1/2}(\psi_k^{1/2} + \psi_{k+1}^{1/2})} \;\geq\; \tfrac{\psi_k - \psi_{k+1}}{2\psi_k\psi_{k+1}^{1/2}} \\ &=\; \tfrac{\psi_k - (1-\alpha_k)\psi_k}{2\psi_k\psi_{k+1}^{1/2}} \;=\; \tfrac{\alpha_k}{2\psi_{k+1}^{1/2}} \;=\; \tfrac{\gamma_{k+1}^{1/2}\theta_n^{1/2}}{2\psi_{k+1}^{1/2}} \;\geq\; \tfrac{1}{8(n+4)}\sqrt{\tfrac{\gamma_0}{L_1(f)}}. \end{aligned}$$

Hence, $\frac{1}{\psi_k 1/2} \geq 1 + \frac{k}{8(n+4)}\sqrt{\frac{\gamma_0}{L_1(f)}}$ for all $k \geq 0$. It remains to note that

$$\begin{aligned} E_{\mathcal{U}_{k-1}}(f(x_k)) - f(x^*) \;&\stackrel{(11)}{\leq}\; \phi_k(\mu) - f(x^*) \;\stackrel{(20)}{\leq}\; \phi_k(\mu) - f_\mu(x^*) + \tfrac{\mu^2}{2}L_1(f)n \\ &\leq\; \psi_k \cdot (f_\mu(x_0) - f_\mu(x^*) + \tfrac{\gamma_0}{2}\|x_0 - x^*\|^2) + \xi_\mu \cdot C_k + \tfrac{\mu^2}{2}L_1(f)n \\ &\stackrel{(20)}{\leq}\; \psi_k \cdot (f(x_0) - f(x^*) + \tfrac{\gamma_0}{2}\|x_0 - x^*\|^2) + \xi_\mu \cdot C_k + \mu^2 L_1(f)n. \end{aligned}$$

It remains to apply the upper bounds for ψ_k. □

Let us discuss the complexity estimates of the method (60) for $\tau(f) = 0$. In order to get accuracy ϵ for the objective function, both terms in the right-hand side of inequality (62) must be smaller than $\frac{\epsilon}{2}$. Thus, we need

$$N(\epsilon) \;=\; O\left(\frac{nL_1^{1/2}(f)R}{\epsilon^{1/2}}\right) \tag{63}$$

iterations. Similarly to the simple random search method (40), this estimate is n times larger than the estimate of the corresponding scheme with full computation of the gradient.

[2]The rest of the proof is very similar to the proof of Lemma 2.2.4 in [8]. We present it here just for the reader convenience.

The parameter of the oracle μ must be chosen as

$$\begin{aligned} \mu &\leq O\left(\frac{\epsilon^{1/2}}{L_1^{1/2}(f)(n\cdot N(\epsilon))^{1/2}}\right) = O\left(\frac{\epsilon^{3/4}}{nL_1^{3/4}(f)R^{1/2}}\right) \\ &= O\left(\frac{1}{n}\left[\frac{\epsilon}{L_1(f)}\cdot\left[\frac{\epsilon}{L_1(f)R^2}\right]^{1/2}\right]^{1/2}\right). \end{aligned} \tag{64}$$

As compared with (58), the average size of the trial step μu is a tighter function of ϵ. This is natural, since the method (54) is much faster. On the other hand, this size is still quite moderate (this is good for numerical stability of the scheme).

Remark 1 *1. Method (60) can be seen as a variant of the Constant Step Scheme (2.2.8) in [8]. Therefore, the sequence $\{v_k\}$ can be expressed in terms of $\{x_k\}$ and $\{y_k\}$ (see Section 2.2.1 in [8] for details).*

2. Linear convergence of method (60) for strongly convex functions allows an efficient generation of random approximations to the solution of problem (53) with arbitrary high confidence level. This can be achieved by an appropriate regularization of the initial problem, as suggested in Section 3 of [10].

7 Nonconvex problems

Consider now the problem

$$\min_{x\in E} f(x), \tag{65}$$

where the objective function f is nonconvex. Let us apply to it method (40). Now it has the following form:

Method $\widehat{\mathcal{RS}}_\mu$: Choose $x_0 \in E$.
Iteration $k \geq 0$. a). Generate u_k and corresponding $g_\mu(x_k)$. b). Compute $x_{k+1} = x_k - h_k B^{-1} g_\mu(x_k)$.

(66)

Let us estimate the evolution of the value of function f_μ after one step of this scheme. Since f_μ has Lipschitz-continuous gradient, we have

$$f_\mu(x_{k+1}) \overset{(6)}{\leq} f_\mu(x_k) - h\langle \nabla f_\mu(x_k), B^{-1} g_\mu(x_k)\rangle + \tfrac{1}{2}h^2 L_1(f_\mu)\|g_\mu\|_*^2.$$

Taking now the expectation in u_k, we obtain

$$E_{u_k}(f_\mu(x_{k+1})) \overset{(22)}{\leq} f_\mu(x_k) - h_k\|\nabla f_\mu(x_k)\|_*^2 + \tfrac{1}{2}h_k^2 L_1(f_\mu) E_{u_k}\left(\|g_\mu\|_*^2\right). \tag{67}$$

Consider now two cases.

1. $f \in C^{1,1}(E)$. Then

$$\begin{aligned} E_{u_k}(f_\mu(x_{k+1})) &\overset{(38)}{\leq} f_\mu(x_k) - h_k\|\nabla f_\mu(x_k)\|_*^2 \\ &\quad + \tfrac{1}{2}h_k^2 L_1(f)\left(4(n+4)\|\nabla f_\mu(x_k)\|_*^2 + \tfrac{3\mu^2}{2}L_1^2(f)(n+5)^3\right) \end{aligned}$$

Choosing now $h_k = \hat{h} \overset{\text{def}}{=} \frac{1}{4(n+4)L_1(f)}$, we obtain

$$E_{u_k}(f_\mu(x_{k+1})) \overset{(38)}{\leq} f_\mu(x_k) - \tfrac{1}{2}\hat{h}\|\nabla f_\mu(x_k)\|_*^2 + \tfrac{3\mu^2}{16}L_1(f)\tfrac{(n+5)^3}{(n+4)^2}$$

Since $(n+5)^3 \leq (n+8)(n+4)^2$, taking the expectation of this inequality in $\mathcal{U}_k$, we get

$$\phi_{k+1} \leq \phi_k - \tfrac{1}{2}\hat{h}\eta_k^2 + \tfrac{3\mu^2(n+8)}{16}L_1(f),$$

where $\eta_k^2 \overset{\text{def}}{=} E_{\mathcal{U}_k}\left(\|\nabla f_\mu(x_k)\|_*^2\right)$. Assuming now that $f(x) \geq f^*$ for all $x \in E$, we get

$$\tfrac{1}{N+1}\sum_{k=0}^{N}\eta_k^2 \leq 8(n+4)L_1(f)\left[\tfrac{f(x_0)-f^*}{N+1} + \tfrac{3\mu^2(n+8)}{16}L_1(f)\right]. \tag{68}$$

Since $\theta_k^2 \overset{\text{def}}{=} E_{\mathcal{U}_k}\left(\|\nabla f(x_k)\|_*^2\right) \overset{(30)}{\leq} 2\eta_k^2 + \frac{\mu^2(n+4)^2}{2}L_1^2(f)$, the expected rate of decrease of θ_k is of the same order as (68). In order to get $\frac{1}{N+1}\sum\limits_{k=0}^{N}\theta_k^2 \leq \epsilon^2$, we need to choose

$$\mu \leq O\left(\tfrac{\epsilon}{nL_1(f)}\right).$$

Then, the upper bound for the expected number of steps is $O(\frac{n}{\epsilon^2})$.

2. $f \in C^{0,0}(E)$. Then,

$$\begin{aligned} E_{u_k}(f_\mu(x_{k+1})) &\overset{(35)}{\leq} f_\mu(x_k) - h_k\|\nabla f_\mu(x_k)\|_*^2 + \tfrac{1}{2}h_k^2 L_1(f_\mu)\cdot L_0^2(f)(n+4)^2 \\ &\overset{(23)}{=} f_\mu(x_k) - h_k\|\nabla f_\mu(x_k)\|_*^2 + \tfrac{1}{\mu}h_k^2 n^{1/2}(n+4)^2\cdot L_0^3(f). \end{aligned}$$

Assume $f(x) \geq f^*$, $x \in E$, and denote $S_N \overset{\text{def}}{=} \sum\limits_{k=0}^{n} h_k$. Taking the expectation of the latter inequality in $\mathcal{U}_k$, and summing them up, we get

$$\begin{aligned} \tfrac{1}{S_N}\sum_{k=0}^{N}h_k\eta_k^2 &\leq \tfrac{1}{S_N}\left[(f_\mu(x_0) - f^*) + C(\mu)\sum_{k=0}^{N}h_k^2\right], \\ C(\mu) &\overset{\text{def}}{=} \tfrac{1}{\mu}n^{1/2}(n+4)^2\cdot L_0^3(f). \end{aligned} \tag{69}$$

Thus, we can guarantee a convergence of the process (66) to a stationary point of the function f_μ, which is a smooth approximation of f. In order to bound the gap in this

approximation by ϵ, we need to choose $\mu \leq \bar{\mu} \stackrel{(19)}{=} \frac{\epsilon}{n^{1/2} L_0(f)}$. Let us assume for simplicity that we are using a constant-step scheme: $h_k \equiv h$, $k \geq 0$. Then the right-hand side of inequality (69) becomes

$$\frac{f_{\bar{\mu}}(x_0) - f^*}{(N+1)h} + \frac{h}{\epsilon} n(n+4)^2 L_0^4(f) \;\leq\; \frac{L_0(f) R}{(N+1)h} + \frac{h}{\epsilon} n(n+4)^2 L_0^4(f) \stackrel{\text{def}}{=} \rho(h).$$

Minimizing this upper bound in h, we get is optimal value:

$$h^* \;=\; \left[\frac{\epsilon R}{n(n+4)^2 L_0^3(f)(N+1)}\right]^{1/2}, \quad \rho(h^*) \;=\; 2\left[\frac{n(n+4)^2 L_0^5(f) R}{\epsilon (N+1)}\right]^{1/2}.$$

Thus, in order to guarantee the expected squared norm of the gradient of function $f_{\bar{\mu}}$ of the order δ, we need

$$O\left(\frac{n(n+4)^2 L_0^5(f) R}{\epsilon \delta^2}\right)$$

iteration of the scheme (66). To the best of our knowledge, this is the first complexity bound for the methods for minimizing nonsmooth nonconvex functions. Note that allowing in the method (66) $h_k \to 0$ and $\mu \to 0$, we can ensure a convergence of the scheme to a stationary point of the initial function f. But this proof is quite long and technical. Therefore, we omit it.

8 Preliminary computational experiments

The main goal of our experiment was the investigation of the impact of the random oracle on the actual convergence of the minimization methods. We compared the performance of the randomized gradient-free methods with the classical gradient schemes. As suggested by our efficiency estimates, it is normal if the former methods need n times more iterations as compared with the classical ones. Let us describe our results.

8.1 Smooth minimization

We checked the performance of the methods (54) and (60) on the following test function:

$$f_n(x) \;=\; \tfrac{1}{2}(x^{(1)})^2 + \tfrac{1}{2}\sum_{i=1}^{n-1}(x^{(i+1)} - x^{(i)})^2 + \tfrac{1}{2}(x^{(n)})^2 - x^{(1)}, \quad x_0 = 0. \qquad (70)$$

This function was used in Section 2.1 in [8] for proving the lower complexity bound for the gradient methods as applied to functions from $C^{1,1}(R^n)$. It has the following parameters:

$$L_1(f_n) \;\leq\; 4, \quad R^2 = \|x_0 - x^*\|^2 \;\leq\; \tfrac{n+1}{3}.$$

These values were used for defining the trial step size μ by (58) and (64). We also tested the versions of corresponding methods with $\mu = 0$. Finally, we compared these results with the usual gradient and fast gradient method.

Our results for the simple gradient schemes are presented in the following table. The first column of the table indicates the current level of relative accuracy with respect to the scale $S \stackrel{\text{def}}{=} \frac{1}{2} L_1(f_n) R^2$. The kth row of the table, $k = 2 \ldots 9$, shows the number of

iterations spent for achieving the absolute accuracy $2^{-(k+7)}S$. This table aggregates the results of 20 attempts of the method $\mathcal{RG}_0$ and $\mathcal{RG}_\mu$ to minimize the function (70). The columns 2-4 of the table represent the minimal, maximal and average number of blocks by n iterations, executed by $\mathcal{RG}_0$ in order to reach corresponding level of accuracy. The next three columns represent this information for $\mathcal{RG}_\mu$ with μ computed by (58) with $\epsilon = 2^{-16}$. The last column contains the results for the standard gradient method with constant step $h = \frac{1}{L_1(f_n)}$.

	$\mu = 0$			$\mu = 8.9 \cdot 10^{-6}$			
Accuracy	min	max	Mean	min	max	Mean	GM
$2.0 \cdot 10^{-3}$	3	4	4.0	3	4	3.9	1
$9.8 \cdot 10^{-4}$	20	22	21.3	21	22	21.3	5
$4.9 \cdot 10^{-4}$	85	89	86.8	85	89	86.8	22
$2.4 \cdot 10^{-4}$	329	343	335.5	327	342	335.4	83
$1.2 \cdot 10^{-4}$	1210	1254	1232.8	1204	1246	1231.8	304
$6.1 \cdot 10^{-5}$	4129	4242	4190.3	4155	4235	4190.4	1034
$3.1 \cdot 10^{-5}$	12440	12611	12536.7	12463	12645	12538.1	3092
$1.5 \cdot 10^{-5}$	30883	31178	31054.6	30939	31269	31058.1	7654

Table 1. Simple Random Search $\mathcal{RG}_\mu$.

We can see a very small variance of the results presented in each column. Moreover, the finite-difference version with an appropriate value of μ demonstrates practically the same performance as the version based on the directional derivative. Moreover, the number of blocks by n iterations of the random schemes is practically equal to the number of iterations of the standard gradient method multiplied by four. A plausible explanation of this phenomena is related to the choice of the step size $h = \frac{1}{\mathbf{4}\cdot(n+4)L_1(f)}$. However, we prefer to use this value since there is no theoretical justification for a larger step.

Let us present the results of 20 runs of the accelerated schemes. The structure of Table 2 is similar to that of Table 1. Since these methods are faster, we give the results

for a more accurate solution, up to $\epsilon = 2^{-30}$.

	$\mu = 0$			$\mu = 3.5 \cdot 10^{-10}$			
Accuracy	min	max	Mean	min	max	Mean	FGM
$2.0 \cdot 10^{-3}$	7	7	7.0	7	7	7.0	1
$9.8 \cdot 10^{-4}$	21	22	21.1	21	22	21.1	4
$4.9 \cdot 10^{-4}$	45	47	45.8	46	47	46.2	10
$2.4 \cdot 10^{-4}$	93	96	94.1	93	96	94.5	22
$1.2 \cdot 10^{-4}$	182	187	184.7	180	188	185.4	44
$6.1 \cdot 10^{-5}$	338	350	345.4	342	349	346.6	84
$3.1 \cdot 10^{-5}$	597	611	603.2	599	609	604.3	147
$1.5 \cdot 10^{-5}$	944	967	953.1	948	964	954.9	233
$7.6 \cdot 10^{-6}$	1328	1355	1339.6	1332	1351	1341.5	328
$3.8 \cdot 10^{-6}$	1671	1695	1679.4	1671	1688	1680.3	411
$1.9 \cdot 10^{-6}$	1915	1934	1922.6	1916	1928	1923.1	471
$9.5 \cdot 10^{-7}$	2070	2083	2075.3	2070	2080	2075.7	508
$4.8 \cdot 10^{-7}$	2177	2189	2182.1	2177	2187	2182.6	535
$2.4 \cdot 10^{-7}$	2270	2281	2274.4	2268	2279	2274.4	557
$1.2 \cdot 10^{-7}$	2360	2375	2366.8	2355	2375	2366.3	580
$6.0 \cdot 10^{-8}$	4294	4308	4299.9	4291	4308	4300.9	1056
$3.0 \cdot 10^{-8}$	4396	4410	4402.4	4392	4411	4403.6	1081
$1.5 \cdot 10^{-8}$	4496	4521	4506.9	4495	4518	4508.0	1107
$7.5 \cdot 10^{-9}$	6519	6537	6529.0	6517	6540	6529.1	1604
$3.7 \cdot 10^{-9}$	6624	6669	6646.2	6623	6672	6644.4	1633
$1.9 \cdot 10^{-9}$	8680	8718	8700.3	8682	8712	8699.1	2139
$9.3 \cdot 10^{-10}$	10770	10805	10789.9	10779	10808	10791.2	2653

Table 2. Fast Random Search $\mathcal{FG}_\mu$.

As we can see, the accelerated schemes are indeed faster than the simple random search. On the other hand, same as in Table 1, the variance of the results in each line is very small. Method with $\mu = 0$ demonstrates almost the same efficiency as the method with μ defined by (64). And again, the number of the blocks by n iterations of the random methods is proportional to the number of iterations of the standard gradient methods multiplied by four.

8.2 Nonsmooth minimization

For nonsmooth problems, we present first the computational results of two variants of method (40) on the following test function:

$$F_n(x) \;=\; |x^{(1)} - 1| + \sum_{i=1}^{n-1} |1 + x^{(i+1)} - 2x^{(i)}|, \quad x_0 = 0. \tag{71}$$

It has the following parameters:

$$L_0(F_n) \;\le\; 3n^{1/2}, \quad R^2 = \|x_0 - x^*\|^2 \;\le\; n.$$

We compared the version $\mathcal{RS}_0$ and version $\mathcal{RS}_\mu$ with μ defined by (47) with the standard subgradient method (e.g. Section 3.2.3 in [8]). The results are presented in Tables 3-5. The first columns of these tables show the required accuracy as compared with the scale $L_0(F_n)R$. In this case, the theoretical upper bound for achieving this level of accuracy is $\frac{\kappa}{\epsilon^2}$, where κ is an absolute constant. The columns of the tables correspond to the test problems of dimension $n = 2^p$, $p = 2 \dots 8$. Each cell shows the number of blocks of n iterations, which were necessary to reach this level of accuracy. If this was impossible after 10^5 iterations, we put in the cell the best value found by the scheme. In Table 5, representing the results of the standard subgradient scheme, we show the usual number of iterations. We show the results only for a single run since the variability in the performance of the random scheme is very small.

Table 3. Method $\mathcal{RS}_0$, Limit = 10^5

$\epsilon \backslash n$	8	16	32	64	128	256
2.5E-1	1	4	2	2	5	4
1.3E-1	10	7	7	7	13	11
6.3E-2	16	11	18	25	27	21
3.1E-2	22	27	49	59	61	74
1.6E-2	50	104	156	187	218	263
7.8E-3	65	328	480	685	885	1045
3.9E-3	477	1086	1812	2749	3397	3848
2.0E-3	533	4080	6834	10828	12872	14773
9.8E-4	5784	10809	27341	41896	51072	54615
4.9E-4	60089	39157	84009	6.0E-4	6.8E-4	7.5E-4
2.4E-4	3.6E-4	3.0E-4	4.8E-4			

As compared with the theoretical upper bounds, all methods perform much better. Note that we observe an unexpectedly good performance of the method $\mathcal{RS}_\mu$. It is always better than its variant with exact directional derivative. Moreover, it is very often better than the usual subgradient method. Let us compare these schemes on a more sophisticated test problem.

Table 4. Method $\mathcal{RS}_\mu$, μ = 3.2E-7

$\epsilon \backslash n$	8	16	32	64	128	256
2.5E-1	2	1	4	9	17	33
1.3E-1	12	18	32	58	113	221
6.3E-2	25	38	58	105	199	381
3.1E-2	30	60	78	137	258	482
1.6E-2	38	88	94	161	296	546
7.8E-3	41	108	107	180	323	590
3.9E-3	65	114	126	199	347	624
2.0E-3	130	273	210	221	364	656
9.8E-4	1293	884	966	698	451	698
4.9E-4	9489	3714	3044	2213	1772	981
2.4E-4	3.5E-4	11156	9589	9506	6591	3759
1.2E-4		26608	47570	37870	25565	14691

Table 5. Subgradient Method

$\epsilon \backslash n$	8	16	32	64	128	256
2.5E-1	1	1	1	1	1	1
1.3E-1	9	4	3	3	3	3
6.3E-2	13	12	6	4	4	4
3.1E-2	73	30	14	10	6	4
1.6E-2	261	40	24	24	14	14
7.8E-3	1274	94	90	48	44	36
3.9E-3	3858	248	592	118	128	94
2.0E-3	8609	3866	2042	368	342	202
4.9E-4	28607	17698	3442	904	648	392
9.8E-4	93886	46218	9280	3570	5.8E-4	566
2.4E-4	3.9E-4	85778	13684	18354		904
1.2E-4		2.2E-4	1.8E-4	1.8E-4		1.7E-4

Let $\Delta_m \subset R^m$ be a standard simplex. Consider the following matrix game:

$$\min_{x \in \Delta_m} \max_{y \in \Delta_m} \langle Ax, y \rangle \quad = \max_{y \in \Delta_m} \min_{x \in \Delta_m} \langle Ax, y \rangle, \tag{72}$$

where A is an $m \times m$-matrix. Define the following function:

$$f(x,y) \quad = \quad \max \left\{ \max_{1 \le i,j \le m} \left[\langle A^T e_i, x \rangle - \langle A e_j, y \rangle \right], \ |\langle \bar{e}, x \rangle - 1|, \ |\langle \bar{e}, y \rangle - 1| \right\},$$

where $e_i \in R^m$ are coordinate vectors, and $\bar{e} \in R^m$ is the vector of all ones. Clearly, the problem (72) is equivalent to the following minimization problem:

$$\min_{x,y \ge 0} \ f(x,y). \tag{73}$$

The optimal value of this problem is zero. We choose the starting points $x_0 = \frac{\bar{e}}{m}$, $y_0 = \frac{\bar{e}}{m}$, and generate A with random entries uniformly distributed in the interval $[-1, 1]$. Then the parameters of problem (39) are as follows:

$$n \ = \ 2m, \quad Q \ = \ R^n_+, \quad L_0(f) \ \le \ n^{1/2}, \quad R \ \le \ 2.$$

In Table 6, we present the computational results for two variants of method $\mathcal{RS}_\mu$ and the subgradient scheme. For problems (73) of dimension $n = 2^p$, $p = 3 \ldots 16$, we report the best accuracy achieved by the schemes after 10^5 iterations (as usual, for random methods, we count the blocks of n iterations). The parameter μ of method $\mathcal{RS}_\mu$ was computed by (47) with target accuracy $\epsilon =$ 9.5E-7.

Table 6. Saddle point problem

Dim	$\mathcal{RS}_0$	$\mathcal{RS}_\mu$	SubGrad
8	1.3E-5	5.3E-6	1.4E-4
16	3.3E-5	8.3E-6	1.3E-4
32	4.80E-5	7.0E-6	1.3E-4
64	2.3E-4	2.2E-4	2.4E-4
128	9.3E-5	3.1E-5	1.6E-4
256	9.3E-5	2.1E-5	1.7E-4

Clearly, in this competition method $\mathcal{RS}_\mu$ is a winner. Two other methods demonstrate equal performance.

8.3 Conclusion

Our experiments confirm the following conclusion. If the computation of the gradient is feasible, then the cost of the iteration of random methods, and the cost of iteration of the gradients methods are the same. In this situation, the total time spent by the random methods is typically in $O(n)$ times bigger than the time of the gradient schemes. Hence, the random gradient-free methods should be used only if creation of the code for computing the gradient is too costly or just impossible.

In the latter case, for smooth functions, the accelerated scheme (60) demonstrates better performance. This practical observation is confirmed by the theoretical results. For nonsmooth problems, the situation is more delicate. In our experiments, the finite-difference version $\mathcal{RS}_\mu$ was always better than the method $\mathcal{RS}_0$, based on the exact directional derivative. Up to now, we did not manage to find a reasonable explanation for this phenomena. It remains an interesting topic for the future research.

References

[1] F.Clarke. Optimization and nonsmooth analysis. Wliley, New York (1983).

[2] A.Conn, K.Scheinberg, and L.Vicente. Introduction to derivative-free optimization. MPS-SIAM series on optimization, 8, SIAM, Philadelphia (2009).

[3] C.Dorea. "Expected number of steps of a random optimization method." *JOTA*, **39**(2), 165-171 (1083).

[4] J.Matyas. "Random optimization." *Automation and Remote Control*, **26**, 246-253 (1965).

[5] Nelder, John A.; R. Mead. "A simplex method for function minimization". *Computer Journal*, **7**, 308313 (1965).

[6] A.Nemirovski, A.Juditsky, G.Lan, and A.Shapiro. "Robust Stochastic Approximation approach to Stochastic Programming." *SIAM Journal on Optimization* **19**(4), 1574-1609 (2009).

[7] A.Nemirovsky, and D.Yudin. Problem complexity and method efficiency in optimization. John Wiley and Sons, New York (1983).

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

[9] Yu. Nesterov. "Lexicographic differentiation of nonsmooth functions". *Mathematical Programming* (B), **104** (2-3), 669-700 (2005).

[10] Yu.Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. CORE Discussion Paper #2010/02, CORE (2010).

[11] B.Polyak. Introduction to Optimization. Optimization Software - Inc., Publications Division, New York (1987).

[12] V.Protasov. "Algorithms for approximate calculation of the minimum of a convex function from its values." *Mathematical Notes*, **59**(1), 69-74 (1996).

[13] M.Sarma. "On theconvergence of the Baba and Dorea random optimization methods." *JOTA*, **66**(2), 337-343 (1990).