ISSN 0965-5425, Computational Mathematics and Mathematical Physics, 2019, Vol. 59, No. 7, pp. 1085–1097. © Pleiades Publishing, Ltd., 2019.
Russian Text © The Author(s), 2019, published in Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki, 2019, Vol. 59, No. 7, pp. 1137–1150.

Fast Gradient Descent for Convex Minimization Problems with an Oracle Producing a (δ, L)-Model of Function at the Requested Point

A. V. Gasnikov[a,b,c] and A. I. Tyurin[a,*]

[a] *State University—Higher School of Economics, Moscow, 125319 Russia*
[b] *Moscow Institute of Physics and Technology, Dolgoprudnyi, Moscow oblast, 141700 Russia*
[c] *Kharkevich Institute for Information Transmission Problems, Moscow, 127051 Russia*
**e-mail: atyurin@hse.ru*

Received November 8, 2017; revised November 8, 2017; accepted March 11, 2019

Abstract—A new concept of (δ, L)-model of a function that is a generalization of the Devolder–Glineur–Nesterov (δ, L)-oracle is proposed. Within this concept, the gradient descent and fast gradient descent methods are constructed and it is shown that constructs of many known methods (composite methods, level methods, conditional gradient and proximal methods) are particular cases of the methods proposed in this paper.

Keywords: gradient descent, fast gradient descent, model of function, universal method, conditional gradient method, composite optimization

DOI: 10.1134/S0965542519070078

1. INTRODUCTION

1.1. Motivation

Gradient descent and fast gradient descent are perhaps the most popular numerical optimization methods. Both methods are based on the idea of approximating the function at the initial point (the current point of the method) by a majorizing paraboloid of revolution and on choosing the minimizer of this paraboloid as the next point of the method. Thus, the divide and conquer principle is implemented, i.e., the initially hard problem is decomposed into a set of simpler problems. The second-order, quasi-Newtonian, cubic regularization, Chebychev's, composite, and level methods suggest that it is not necessary to use paraboloids of revolution in the above approach. One can use more complex functions that provide a more accurate local model of the function at the point under examination, which ultimately gives faster convergence of the method. There are two approaches. One of them introduces higher order derivatives into the function model. The other approach is based on including a part of the problem formulation into the model; e.g., if the function to be optimized is a sum of two functions, then one of them can be replaced by a paraboloid of revolution in the model and the other function remains as it is. The second approach is new and currently does not form a special direction of research. As far as we know, until the present time no attempts have been made at integrating the available uncoordinated results. In this paper, we make such an attempt. Note that, in the case of paraboloid of revolution, the solution of the auxiliary problem is typically easy; often, it can be solved using explicit formulas, i.e., exactly. In the second approach, the situation is quite different; more precisely, the auxiliary problem can usually be solved only approximately. For this reason, we consider in this paper the case when the auxiliary problem is solved approximately.

The generic optimization problem, in which only information about the Lipschitzness of the gradient or of the function is available, is well studied and lower and upper bounds for this problem are known (see [1, 2]). Recently, structural optimization has gained popularity, in which a priori information about the structure of the problem is available. Additional information about the problem makes it possible to find new methods and improve upper bounds. In particular, the composite statement of the problem, in which the given function is the sum of a smooth and nonsmooth functions [3] can be solved with the rate of the fast gradient method using additional information about the nonsmooth term in the sum. Moreover, the minimum of the nonsmooth function can often be found with the rate of the fast gradient

method even though this is not possible for nonsmooth problems in the general case. An example is the minimax problem, which is not smooth but for which a fast gradient method was proposed [1]. The main purpose of this paper is to make an attempt at unifying different approaches and propose a method based on such a unified approach that should include all earlier proposed concepts. In Section 4, we give a large number of examples of optimization problems that can be solved using the proposed unified method. For this purpose, we define the concept of (δ, L)-model, which is in essence a generalization of the definition of Lipschitzness of the gradient or a generalization of the concept of the Devolder–Glineur–Nesterov (δ, L)-oracle.

The paper is organized as follows. In Section 2, we introduce a novel concept of (δ, L)-oracle [4] and discuss the concept of the (δ, L)-model of function and allow the auxiliary problem to be solved inexactly [5]. In this section, we also generalize the gradient descent method for the case of dealing with the new model. In Section 3, the results obtained in Section 2 for the gradient descent method are extended to the fast gradient descent method. Section 4 describes the application of the methods proposed in Sections 2 and 3 within the concept of (δ, L)-model to various problem statements. In this section, we also show that the proposed concept and methods make it possible to unify the results available in this direction of research. The Appendix contains the justification of the fact that the concept of accuracy of solving the auxiliary problem, which is adopted following Nemirovski, is quite effective, and for bounded smooth statements of the problem is reduced to the conventional concept of function convergence (note that there are other concepts [6–8]).

2. THE GRADIENT DESCENT METHOD WITH AN ORACLE USING THE (δ, L)-MODEL

First, we give the general formulation of the convex optimization problem [1]. Let a function $F(x) : Q \to \mathbb{R}$ and an arbitrary norm $\| \ \|$ in $\mathbb{R}^n$ be given. The adjoint norm is defined by

$$\|\lambda\|_* = \max_{\|v\| \le 1; v \in \mathbb{R}^n} \langle \lambda, v \rangle \quad \forall \lambda \in \mathbb{R}^n. \tag{1}$$

We assume that

1. $Q \subseteq \mathbb{R}^n$ is a convex closed set.

2. $F(x)$ is a continuous convex function on Q.

3. $F(x)$ is bounded from below on Q and attains its minimum at a certain point (not necessarily unique) $x_* \in Q$.

Consider the optimization problem

$$F(x) \to \min_{x \in Q}. \tag{2}$$

Let us introduce the concepts of prox-function and Bregman divergence [9].

Definition 1. $d(x) : Q \to \mathbb{R}$ is called a *prox-function* if $d(x)$ is continuously differentiable on $\operatorname{int} Q$ and $d(x)$ is 1-strongly convex with respect to the norm $\| \ \|$ on $\operatorname{int} Q$.

Definition 2. The *Bregman divergence* is

$$V(x, y) \stackrel{\text{def}}{=} d(x) - d(y) - \langle \nabla d(y), x - y \rangle, \tag{3}$$

where $d(x)$ is an arbitrary prox-function. It is easy to verify (see [5]) that

$$V(x, y) \ge \frac{1}{2}\|x - y\|^2.$$

Next, we define the (δ, L)-model of function [10], which is a direct generalization of the (δ, L)-oracle [4, 11, 12].

Definition 3. The pair $(F_\delta(y), \psi_\delta(x, y))$ is called a (δ, L)-model of the function $F(x)$ at the point y if the inequality

$$0 \le F(x) - F_\delta(y) - \psi_\delta(x, y) \le \frac{L}{2}\|x - y\|^2 + \delta \tag{4}$$

holds for every $x \in Q$,

$$\psi_\delta(x, x) = 0 \quad \forall x \in Q, \tag{5}$$

and $\psi_\delta(x, y)$ is a convex function in x $\forall y \in Q$.

We assume that, for $F(x)$, there are δ and L such that at every point $x \in Q$ there exists a (δ, L)-model. Examples are given in Section 4.

Corollary 1. Take $x = y$ in (4) and use (5); then

$$F_\delta(y) \le F(y) \le F_\delta(y) + \delta \quad \forall y \in Q. \tag{6}$$

Consider the concept of inexact solution of problem described in [5].

Definition 4. Consider the problem

$$\psi(x) \to \min_{x \in Q},$$

where $\psi(x)$ is convex. Then $\operatorname{Arg\,min}^{\tilde\delta}_{x\in Q} \psi(x)$ is the set of all $\tilde x$ such that

$$\exists h \in \partial\psi(\tilde x), \quad \langle h, x - \tilde x\rangle \ge -\tilde\delta \quad \forall x \in Q.$$

An arbitrary element in $\operatorname{Arg\,min}^{\tilde\delta}_{x\in Q} \psi(x)$ will be denoted by $\arg\min^{\tilde\delta}_{x\in Q} \psi(x)$.

Consider a simple consequence. Let $\tilde x \in \operatorname{Arg\,min}^{\tilde\delta}_{x\in Q} \psi(x)$. Then, the convexity implies that $\psi(x) \ge \psi(\tilde x) + \langle h, x - \tilde x\rangle \ge \psi(\tilde x) - \tilde\delta$, where $h \in \partial\psi(\tilde x)$. Take $x = x_*$; then $\psi(\tilde x) - \psi(x_*) \le \tilde\delta$. That is, $\tilde x \in \operatorname{Arg\,min}^{\tilde\delta}_{x\in Q} \psi(x)$ implies that $\tilde x$ is a $\tilde\delta$-optimal solution. The converse is generally not true, and throughout this paper we will essentially use the condition on the $\tilde\delta$-solution from Definition 4, which is more rigorous.

Consider a generalization of the gradient descent algorithm for problem (2) [10]. In this algorithm, we assume that an initial point x_0 is given, N is the number of steps of the method, and L_0 is a constant that has the sense of tentative "local" Lipschitz constant of the gradient at the point x_0. At the input of the algorithm are also the sequences $\{\delta_k\}_{k=0}^{N-1}$ and $\{\tilde\delta_k\}_{k=0}^{N-1}$, wher $\{\delta_k\}_{k=0}^{N-1}$ is a sequence such that, for every k, there exists a (δ_k, L)-model for $F(x)$ at any point $x \in Q$, and $\{\tilde\delta_k\}_{k=0}^{N-1}$ are the errors of the solution from Definition 4, which may be zero, constant, or vary from iteration to iteration in different problems.

Note that we never explicitly use the constant L in the proposed method. We assume that $L_0 \le L$; otherwise, we will set $L := \max(L_0, L)$ in all the bounds below.

Let us describe the gradient descent algorithm with an oracle that uses the (δ, L)-model.

Algorithm

Input data: x_0 is the initial point, N is the number of steps, $\{\delta_k\}_{k=0}^{N-1}$ and $\{\tilde\delta_k\}_{k=0}^{N-1}$ are sequences, and $L_0 > 0$.

Step 0:

$$L_1 := \frac{L_0}{2}.$$

Step k + 1:

$$\alpha_{k+1} := \frac{1}{L_{k+1}},$$

$$\phi_{k+1}(x) = V(x, x_k) + \alpha_{k+1}\psi_{\delta_k}(x, x_k), \tag{7}$$

$$x_{k+1} := \arg\min^{\tilde\delta_k}_{x\in Q} \phi_{k+1}(x).$$

If it holds that

$$F_{\delta_k}(x_{k+1}) \le F_{\delta_k}(x_k) + \psi_{\delta_k}(x_{k+1}, x_k) + \frac{L_{k+1}}{2}\|x_{k+1} - x_k\|^2 + \delta_k, \tag{8}$$

then set

$$L_{k+2} := \frac{L_{k+1}}{2}$$

and go to the next step; otherwise, set

$$L_{k+1} := 2L_{k+1}$$

and repeat the current step.

Remark 1. For all $k \geq 0$, it holds that

$$L_k \leq 2L.$$

For $k = 0$, this inequality holds because $L_0 \leq L$. For $k \geq 1$, this follows from the fact that we will exit the inner loop, in which L_k is fitted, earlier than L_k becomes greater than $2L$. The exit from the loop is guaranteed by the condition that a (δ_k, L)-model for $F(x)$ exists at every point $x \in Q$.

We now prove an important lemma.

Lemma 1. *Let $\psi(x)$ be a convex function and*

$$y = \arg\min_{x \in Q}^{\tilde{\delta}} \{\psi(x) + V(x, z)\}.$$

Then

$$\psi(x) + V(x, z) \geq \psi(y) + V(y, z) + V(x, y) - \tilde{\delta} \quad \forall x \in Q.$$

Proof. By Definition 4, we have

$$\exists g \in \partial\psi(y), \quad \langle g + \nabla_y V(y, z), x - y\rangle \geq -\tilde{\delta} \quad \forall x \in Q.$$

Now the inequality

$$\psi(x) - \psi(y) \geq \langle g, x - y\rangle \geq \langle \nabla_y V(y, z),\ y - x\rangle - \tilde{\delta}$$

and the equality

$$\langle \nabla_y V(y, z), y - x\rangle = \langle \nabla d(y) - \nabla d(z), y - x\rangle = d(y) - d(z) - \langle \nabla d(z), y - z\rangle$$
$$+ d(x) - d(y) - \langle \nabla d(y), x - y\rangle - d(x) + d(z) + \langle \nabla d(z), x - z\rangle = V(y, z) + V(x, y) - V(x, z)$$

complete the proof.

Lemma 2. *For every $x \in Q$, it holds that*

$$\alpha_{k+1}F(x_{k+1}) - \alpha_{k+1}F(x) \leq V(x, x_k) - V(x, x_{k+1}) + \tilde{\delta}_k + 2\delta_k\alpha_{k+1}.$$

Proof. Consider the chain of inequalities

$$F(x_{k+1}) \overset{(8),(6)}{\leq} F_{\delta_k}(x_k) + \psi_{\delta_k}(x_{k+1}, x_k) + \frac{L_{k+1}}{2}\|x_{k+1} - x_k\|^2 + 2\delta_k$$

$$\leq F_{\delta_k}(x_k) + \psi_{\delta_k}(x_{k+1}, x_k) + \frac{1}{\alpha_{k+1}}V(x_{k+1}, x_k) + 2\delta_k$$

$$\leq_{①} F_{\delta_k}(x_k) + \psi_{\delta_k}(x, x_k) + \frac{1}{\alpha_{k+1}}V(x, x_k) - \frac{1}{\alpha_{k+1}}V(x, x_{k+1}) + \frac{\tilde{\delta}_k}{\alpha_{k+1}} + 2\delta_k$$

$$\overset{(4)}{\leq} F(x) + \frac{1}{\alpha_{k+1}}V(x, x_k) - \frac{1}{\alpha_{k+1}}V(x, x_{k+1}) + \frac{\tilde{\delta}_k}{\alpha_{k+1}} + 2\delta_k.$$

Inequality ① follows from Lemma 1 with $\psi(x) = \alpha_{k+1}\psi_{\delta_k}(x, x_k)$ and the left-hand side of (4).

Theorem 1. *Let $V(x_*, x_0) \le R^2$, where x_0 is the initial point, x_* be the closest minimizer to the point x_0 in the sense of the Bregman divergence, and*

$$\overline{x}_N = \frac{1}{A_N}\sum_{k=0}^{N-1}\alpha_{k+1}x_{k+1}.$$

For the proposed algorithm, it holds that

$$F(\overline{x}_N) - F(x_*) \le \frac{2LR^2}{N} + \frac{2L}{N}\sum_{k=0}^{N-1}\tilde{\delta}_k + \frac{2}{A_N}\sum_{k=0}^{N-1}\alpha_{k+1}\delta_k.$$

Proof. Let us sum the inequality in Lemma 2 over $k = 0,\ldots, N-1$:

$$\sum_{k=0}^{N-1}\alpha_{k+1}F(x_{k+1}) - A_N F(x) \le V(x, x_0) - V(x, x_N) + \sum_{k=0}^{N-1}\tilde{\delta}_k + 2\sum_{k=0}^{N-1}\alpha_{k+1}\delta_k.$$

Take $x = x_*$ to obtain

$$\sum_{k=0}^{N-1}\alpha_{k+1}F(x_{k+1}) - A_N F(x_*) \le R^2 - V(x_*, x_N) + \sum_{k=0}^{N-1}\tilde{\delta}_k + 2\sum_{k=0}^{N-1}\alpha_{k+1}\delta_k.$$

Since $V(x_*, x_N) \ge 0$, we have

$$\sum_{k=0}^{N-1}\alpha_{k+1}F(x_{k+1}) - A_N F(x_*) \le R^2 + \sum_{k=0}^{N-1}\tilde{\delta}_k + 2\sum_{k=0}^{N-1}\alpha_{k+1}\delta_k.$$

Divide both sides by A_N:

$$\frac{1}{A_N}\sum_{k=0}^{N-1}\alpha_{k+1}F(x_{k+1}) - F(x_*) \le \frac{R^2}{A_N} + \frac{1}{A_N}\sum_{k=0}^{N-1}\tilde{\delta}_k + \frac{2}{A_N}\sum_{k=0}^{N-1}\alpha_{k+1}\delta_k.$$

Using the convexity of $F(x)$, we finally obtain

$$\begin{aligned} F(\overline{x}_N) - F(x_*) &\le \frac{R^2}{A_N} + \frac{1}{A_N}\sum_{k=0}^{N-1}\tilde{\delta}_k + \frac{2}{A_N}\sum_{k=0}^{N-1}\alpha_{k+1}\delta_k \\ &\le_{①} \frac{2LR^2}{N} + \frac{2L}{N}\sum_{k=0}^{N-1}\tilde{\delta}_k + \frac{2}{A_N}\sum_{k=0}^{N-1}\alpha_{k+1}\delta_k. \end{aligned}$$

Inequality ① follows from Remark 1 and the fact that $\alpha_{k+1} = 1/L_{k+1}$.

3. THE FAST GRADIENT METHOD WITH AN ORACLE USING THE (δ, L)-MODEL

Consider the fast version of the algorithm described in Section 2.

Algorithm

Input data: x_0 is the initial point, N is the number of steps, $\{\delta_k\}_{k=0}^{N-1}$ and $\{\tilde{\delta}_k\}_{k=0}^{N-1}$ are sequences, and $L_0 > 0$.

Step 0: $y_0 := x_0, \quad u_0 := x_0, \quad L_1 := \frac{L_0}{2}, \quad \alpha_0 := 0, \quad A_0 := \alpha_0.$

Step k + 1:

Find the greatest root:

$$\begin{gathered} \alpha_{k+1} : A_k + \alpha_{k+1} = L_{k+1}\alpha_{k+1}^2, \\ A_{k+1} := A_k + \alpha_{k+1}, \\ y_{k+1} := \frac{\alpha_{k+1}u_k + A_k x_k}{A_{k+1}}, \end{gathered} \tag{9}$$

$$\phi_{k+1}(x) = V(x,u_k) + \alpha_{k+1}\psi_{\delta_k}(x,y_{k+1}),$$
$$u_{k+1} := \arg\min_{x\in Q}{}^{\tilde{\delta}_k}\ \phi_{k+1}(x), \tag{10}$$

$$x_{k+1} := \frac{\alpha_{k+1}u_{k+1} + A_k x_k}{A_{k+1}}. \tag{11}$$

If it holds that

$$F_{\delta_k}(x_{k+1}) \le F_{\delta_k}(y_{k+1}) + \psi_{\delta_k}(x_{k+1},y_{k+1}) + \frac{L_{k+1}}{2}\|x_{k+1} - y_{k+1}\|^2 + \delta_k, \tag{12}$$

then set

$$L_{k+2} := \frac{L_{k+1}}{2}$$

and to the next step; otherwise, set

$$L_{k+1} := 2L_{k+1}$$

and repeat the current step.

Lemma 3. *Let the sequence α_k satisfy the conditions*

$$\alpha_0 = 0, \quad A_k = \sum_{i=0}^{k}\alpha_i, \quad A_k = L_k\alpha_k^2,$$

where $L_k \le 2L$ for every $k \ge 0$ (Remark 1). Then, the following inequality holds for every $k \ge 1$:

$$A_k \ge \frac{(k+1)^2}{8L}. \tag{13}$$

Proof. Let $k = 1$, i.e.,

$$\alpha_1 = L_1\alpha_1^2$$

and

$$A_1 = \alpha_1 = \frac{1}{L_1} \ge \frac{1}{2L}.$$

Let $k \ge 2$; then

$$L_{k+1}\alpha_{k+1}^2 = A_{k+1} \Leftrightarrow L_{k+1}\alpha_{k+1}^2 = A_k + \alpha_{k+1} \Leftrightarrow L_{k+1}\alpha_{k+1}^2 - \alpha_{k+1} - A_k = 0.$$

We solve this quadratic equation and take its greatest root. Then

$$\alpha_{k+1} = \frac{1 + \sqrt{1 + 4L_{k+1}A_k}}{2L_{k+1}}.$$

By induction, let inequality (13) holds for k; then,

$$\alpha_{k+1} = \frac{1}{2L_{k+1}} + \sqrt{\frac{1}{4L_{k+1}^2} + \frac{A_k}{L_{k+1}}} \ge \frac{1}{2L_{k+1}} + \sqrt{\frac{A_k}{L_{k+1}}} \ge \frac{1}{4L} + \frac{1}{\sqrt{2L}}\frac{k+1}{2\sqrt{2L}} = \frac{k+2}{4L}.$$

The last inequality follows from the induction hypothesis. Ultimately, we obtain

$$\alpha_{k+1} \ge \frac{k+2}{4L}$$

and

$$A_{k+1} = A_k + \alpha_{k+1} = \frac{(k+1)^2}{8L} + \frac{k+2}{4L} \ge \frac{(k+2)^2}{8L}.$$

We now formulate and prove the main lemma.

Lemma 4. *For every $x \in Q$, it holds that*

$$A_{k+1}F(x_{k+1}) - A_kF(x_k) + V(x,u_{k+1}) - V(x,u_k) \le \alpha_{k+1}F(x) + 2\delta_k A_{k+1} + \tilde{\delta}_k.$$

Proof. Consider the chain of inequalities

$$F(x_{k+1}) \le F_{\delta_k}(y_{k+1}) + \psi_{\delta_k}(x_{k+1}, y_{k+1}) + \frac{L_{k+1}}{2}\|x_{k+1} - y_{k+1}\|^2 + 2\delta_k \overset{(11)}{=} F_{\delta_k}(y_{k+1}) + \psi_{\delta_k}\left(\frac{\alpha_{k+1}u_{k+1} + A_k x_k}{A_{k+1}}, y_{k+1}\right)$$

$$+ \frac{L_{k+1}}{2}\left\|\frac{\alpha_{k+1}u_{k+1} + A_k x_k}{A_{k+1}} - y_{k+1}\right\|^2 + 2\delta_k \overset{\text{Def. 3, (9)}}{\le} F_{\delta_k}(y_{k+1}) + \frac{\alpha_{k+1}}{A_{k+1}}\psi_{\delta_k}(u_{k+1}, y_{k+1})$$

$$+ \frac{A_k}{A_{k+1}}\psi_{\delta_k}(x_k, y_{k+1}) + \frac{L_{k+1}\alpha_{k+1}^2}{2A_{k+1}^2}\|u_{k+1} - u_k\|^2 + 2\delta_k = \frac{A_k}{A_{k+1}}(F_{\delta_k}(y_{k+1}) + \psi_{\delta_k}(x_k, y_{k+1}))$$

$$+ \frac{\alpha_{k+1}}{A_{k+1}}(F_{\delta_k}(y_{k+1}) + \psi_{\delta_k}(u_{k+1}, y_{k+1})) + \frac{L_{k+1}\alpha_{k+1}^2}{2A_{k+1}^2}\|u_{k+1} - u_k\|^2 + 2\delta_k$$

$$=_{①} \frac{A_k}{A_{k+1}}(F_{\delta_k}(y_{k+1}) + \psi_{\delta_k}(x_k, y_{k+1})) + \frac{\alpha_{k+1}}{A_{k+1}}(F_{\delta_k}(y_{k+1}) + \psi_{\delta_k}(u_{k+1}, y_{k+1}) + \frac{1}{2\alpha_{k+1}}\|u_{k+1} - u_k\|^2) + 2\delta_k$$

$$\le \frac{A_k}{A_{k+1}}(F_{\delta_k}(y_{k+1}) + \psi_{\delta_k}(x_k, y_{k+1})) + \frac{\alpha_{k+1}}{A_{k+1}}(F_{\delta_k}(y_{k+1}) + \psi_{\delta_k}(u_{k+1}, y_{k+1}) + \frac{1}{\alpha_{k+1}}V(u_{k+1}, u_k)) + 2\delta_k$$

$$\le_{②} \frac{A_k}{A_{k+1}}F(x_k) + \frac{\alpha_{k+1}}{A_{k+1}}\left(F_{\delta_k}(y_{k+1}) + \psi_{\delta_k}(x, y_{k+1}) + \frac{1}{\alpha_{k+1}}V(x, u_k) - \frac{1}{\alpha_{k+1}}V(x, u_{k+1}) + \frac{\tilde{\delta}_k}{\alpha_{k+1}}\right) + 2\delta_k$$

$$\overset{(4)}{\le} \frac{A_k}{A_{k+1}}F(x_k) + \frac{\alpha_{k+1}}{A_{k+1}}F(x) + \frac{1}{A_{k+1}}V(x, u_k) - \frac{1}{A_{k+1}}V(x, u_{k+1}) + 2\delta_k + \frac{\tilde{\delta}_k}{\alpha_{k+1}}.$$

Inequality ① follows from the equality $A_k = L_k\alpha_k^2$. Inequality ② follows from Lemma 1 with $\psi(x) = \alpha_{k+1}\psi_{\delta_k}(x, y_{k+1})$ and the left-hand side of (4).

Theorem 2. *Let $V(x_*, x_0) \le R^2$, where x_0 is the initial point, and let x_* be the closest minimizer to x_0 in the sense of the Bregman divergence. For the proposed algorithm, the following inequality holds*:

$$F(x_N) - F(x_*) \le \frac{8LR^2}{(N+1)^2} + \frac{2\sum_{k=0}^{N-1}\delta_k A_{k+1}}{A_N} + \frac{8L\sum_{k=0}^{N-1}\tilde{\delta}_k}{(N+1)^2}.$$

Proof. Sum the inequality in Lemma 4 over $k = 0, \ldots, N-1$ to obtain

$$A_N F(x_N) - A_0 F(x_0) + V(x, u_N) - V(x, u_0) \le (A_N - A_0)F(x) + 2\sum_{k=0}^{N-1}\delta_k A_{k+1} + \sum_{k=0}^{N-1}\tilde{\delta}_k$$

$$\Leftrightarrow A_N F(x_N) + V(x, u_N) - V(x, u_0) \le A_N F(x) + 2\sum_{k=0}^{N-1}\delta_k A_{k+1} + \sum_{k=0}^{N-1}\tilde{\delta}_k.$$

Set $x = x_*$. Then, we have

$$A_N(F(x_N) - F_*) \le R^2 + 2\sum_{k=0}^{N-1}\delta_k A_{k+1} + \sum_{k=0}^{N-1}\tilde{\delta}_k.$$

Divide both parts of this inequality by A_N to obtain

$$F(x_N) - F_* \le \frac{R^2}{A_N} + \frac{2\sum_{k=0}^{N-1}\delta_k A_{k+1}}{A_N} + \frac{\sum_{k=0}^{N-1}\tilde{\delta}_k}{A_N} \le_{①} \frac{8LR^2}{(N+1)^2} + \frac{2\sum_{k=0}^{N-1}\delta_k A_{k+1}}{A_N} + \frac{8L\sum_{k=0}^{N-1}\tilde{\delta}_k}{(N+1)^2}.$$

Inequality ① follows from Lemma 3.

4. CONSEQUENCES

4.1. The Fast Gradient Method

Assume that $F(x)$ is a smooth convex function with an L-Lipschitzian gradient in the norm $\|\ \|$. Then (see [11]),

$$0 \le F(x) - F(y) - \langle \nabla F(y), x - y\rangle \le \frac{L}{2}\|x - y\|^2 \quad \forall x, y \in Q. \tag{14}$$

Thus, we obtain that $\psi_{\delta_k}(x, y) = \langle \nabla F(y), x - y\rangle$, $F_{\delta_k}(y) = F(y)$, and $\delta_k = 0\ \forall k$. In addition, we assume that the auxiliary problem can be solved exactly, i.e., $\tilde{\delta}_k = 0\ \forall k$. Hence, we obtain the following convergence rate for the fast version of the method (Section 3 and Theorem 2):

$$F(x_N) - F(x_*) \le \frac{8LR^2}{(N+1)^2}.$$

This convergence rate is optimal up to a numerical factor (the constant cannot be less than two (see [6]), and we have the constant 8).

4.2. Comparison of the Gradient and Fast Gradient Methods

Assume that there is a (δ, L)-oracle for problem (2) (see [11]), and assume that the auxiliary problem in the sense of Definition 4 can be solved at each step with an error not exceeding $\tilde{\delta}$. Then, due to Theorems 1 and 2, it holds that

$$F(\overline{x}_N) - F(x_*) \le \frac{2LR^2}{N} + 2L\tilde{\delta} + 2\delta,$$
$$F(x_N) - F(x_*) \le \frac{8LR^2}{(N+1)^2} + \frac{8L\tilde{\delta}}{N+1} + 2N\delta,$$

where $\overline{x}_N$ is the point mentioned in Theorem 1 and x_N is the point mentioned in Theorem 2. We conclude that the fast version is more stable to the errors $\tilde{\delta}$ of solving the auxiliary problems; however, this version accumulates the errors δ occurring when the (δ, L)-oracle is called. Note that there is an intermediate gradient method [13] (for the stochastic case, [14]) for which the following bound can be obtained:

$$\mathbb{O}(1)\frac{LR^2}{N^p} + \mathbb{O}(1)N^{1-p}\tilde{\delta} + \mathbb{O}(1)N^{p-1}\delta;$$

here $p \in [1, 2]$ can be chosen arbitrarily, and the proper choice can reduce the noise, which, however, deteriorates the bound on the convergence rate.

4.3. The Universal Method

Consider the fast version of the gradient method (Section 3 and Theorem 2).

The universal method described in [15] makes it possible to apply the concept of (δ, L)-oracle (see [11, 15]) for solving nonsmooth problems. We assume that the Hölder condition holds, i.e., there exists a $\nu \in [0, 1]$ such that

$$\|\nabla F(x) - \nabla F(y)\|_* \le L_\nu \|x - y\|^\nu \quad \forall x, y \in Q.$$

Then (see [15]),

$$0 \le F(x) - F(y) - \langle \nabla F(y), x - y\rangle \le \frac{L(\delta)}{2}\|x - y\|^2 + \delta \quad \forall x, y \in Q, \tag{15}$$

where

$$L(\delta) = L_\nu\left[\frac{L_\nu}{2\delta}\frac{1-\nu}{1+\nu}\right]^{\frac{1-\nu}{1+\nu}}$$

and $\delta > 0$ is a free parameter. We obtain $\psi_{\delta_k}(x, y) = \langle \nabla F(y), x - y \rangle$ and $F_{\delta_k}(y) = F(y)$. Assume that the auxiliary problem can be solved exactly, i.e., $\tilde{\delta}_k = 0$ for every k. Set

$$\delta_k = \epsilon \frac{\alpha_{k+1}}{4A_{k+1}} \quad \forall k, \tag{16}$$

where ϵ is the desired accuracy of solution (with respect to the function).

Theorem 2 and the assumptions made above imply the convergence rate

$$f(x_N) - f(x_*) \le \frac{R^2}{A_N} + \frac{\epsilon}{2}. \tag{17}$$

As in [15], we can prove the inequality

$$A_N \ge \frac{N^{\frac{1+3\nu}{1+\nu}} \epsilon^{\frac{1-\nu}{1+\nu}}}{2^{\frac{2+4\nu}{1+\nu}} L_\nu^{\frac{2}{1+\nu}}}.$$

Hence, we conclude that

$$N \le \inf_{\nu \in [0,1]} \left[2^{\frac{3+5\nu}{1+3\nu}} \left(\frac{L_\nu R^{1+\nu}}{\epsilon} \right)^{\frac{2}{1+3\nu}} \right].$$

This bound is optimal up to a numerical factor [16].

4.4. The Conditional Gradient Method

In practice, the auxiliary problem (10) cannot be solved in reasonable time [5, 17]. It was shown in [18] that the Frank–Wolfe conditional gradient method [5, 18, 19] can be very effective for a certain class of problems. For this reason, instead of $\phi_{k+1}(x) = V(x, u_k) + \alpha_{k+1}\psi_{\delta_k}(x, y_{k+1})$, $\tilde{\phi}_{k+1}(x) = \alpha_{k+1}\psi_{\delta_k}(x, y_{k+1})$ is used in (10). Consider this replacement from the viewpoint of the error $\tilde{\delta}_k$. Below, we assume that $F(x)$ is a smooth function with L-Lipschitzian gradient in the norm $\| \ \|$ and $V(x, y) \le R_Q^2$ for all $x, y \in Q$. Furthermore, let $u_{k+1} = \left(\arg\min_{x \in Q}^{\tilde{\delta}_k} \phi_{k+1}(x) \stackrel{\text{def}}{=} \arg\min_{x \in Q} \tilde{\phi}_{k+1}(x) \right)$. Then

$$\exists h \in \partial\phi_{k+1}(u_{k+1}), \quad \exists g \in \partial\tilde{\phi}_{k+1}(u_{k+1}), \quad \langle h, x - u_{k+1} \rangle = \langle g, x - u_{k+1} \rangle + \langle \nabla_{u_{k+1}} V(u_{k+1}, u_k), x - u_{k+1} \rangle$$
$$\ge \langle \nabla_{u_{k+1}} V(u_{k+1}, u_k), x - u_{k+1} \rangle = -V(u_{k+1}, u_k) - V(x, u_{k+1}) + V(x, u_k) \ge -2R_Q^2.$$

In the algorithm, we assume that $\tilde{\delta}_k = 2R_Q^2$ for each k. The further reasoning is similar to the smooth case with an L-Lipschitzian gradient in the norm $\| \ \|$. Then, we obtain the following convergence rate for the fast version of the method (Section 3 and Theorem 2):

$$F(x_N) - F(x_*) \le \frac{8LR^2}{(N+1)^2} + \frac{16LR_Q^2}{N+1}.$$

This bound is optimal up to a numerical factor—it cannot be improved for the method under examination (see [20]).

4.5. Composite Optimization

Consider the composite optimization problem [3]

$$F(x) \stackrel{\text{def}}{=} f(x) + h(x) \to \min_{x \in Q}, \tag{18}$$

where $f(x)$ is a smooth function with the L-Lipschitzian gradient in the norm $\| \ \|$ and $h(x)$ is a convex function (not necessarily smooth). For this problem we have the inequality

$$0 \le F(x) - F(y) - \langle \nabla f(y), x - y \rangle - h(x) + h(y) \le \frac{L}{2}\|x - y\|^2 \quad \forall x, y \in Q. \tag{19}$$

Therefore, we may use $\psi_{\delta_k}(x,y) = \langle \nabla f(y), x - y\rangle + h(x) - h(y)$, $F_{\delta_k}(y) = F(y)$ and $\delta_k = 0$ for each k. It turns out that the ordinary and fast versions of the method in this problem work without any changes. Note that we thus move a part of the problem complexity to (7) or (10). While the auxiliary problem in the smooth case includes the function $V(x,u_k) + \alpha_{k+1}\langle \nabla f(y_{k+1}), x - y_{k+1}\rangle$, the term $h(x)$ is added in problem (18), and we therefore must solve at each step the more difficult problem $V(x,u_k) + \alpha_{k+1}\left(\langle \nabla f(y_{k+1}), x - y_{k+1}\rangle + h(x) - h(y_{k+1})\right)$.

4.6. The Prox-Method

Consider the problem

$$F(x) \to \min_{x\in Q}, \tag{20}$$

where $F(x)$ is generally a nonsmooth convex function. In the approach described above, we can set $\psi_{\delta_k}(x,y) = F(x) - F(y)$, $F_{\delta_k}(y) = F(y)$ and $\delta_k = 0$ for each k. Condition (4) holds for any $L \geq 0$. The methods described in Sections 2 and 3 are, generally speaking, adaptive in the sense that the "local" Lipschitz constant L_k of the gradient is fitted during the method operation. Let us fix an arbitrary constant $L \geq 0$ and set all L_k equal to L rather than fitting them in the inner loop. In this case, it is easy to verify that the method and all the bounds do not change. Then, the intermediate step of the method in Section 2 is written as

$$x_{k+1} := \arg\min_{x\in Q}{}^{\delta_k}\left[LV(x,x_k) + F(x)\right]. \tag{21}$$

This is called the proximal method (see [20, 21]). It can be effective in certain problems (see [22]). For nonsmooth functions, the algorithm described in Section 3 converges due to bounds proved in Theorem 2, which contradicts the lower bounds for nonsmooth functions (see [1, 2]). However, problem (21) can generally be solved only approximately (see [23, 24]). A more detailed analysis in [10] shows that the total number of oracle calls for obtaining the subgradient of the function $f(x)$ does not contradict the lower bounds in [1, 2] for nonsmooth problems and even agrees with them.

4.7. Superposition of Functions

Consider the problem (see [25–27])

$$F(x) \stackrel{\text{def}}{=} f(f_1(x),\ldots,f_m(x)) \to \min_{x\in Q}, \tag{22}$$

where $f_k(x)$ is a smooth function with the L_k-Lipschitzian gradient in the norm $\|\ \|$ for every k and $f(x)$ is an M-Lipschitzian convex function with respect to the L_1-norm that is nondecreasing in each of its arguments. Consequently (see [26, 28]), $F(x)$ is a convex function as well, and it holds that (see [26])

$$0 \leq F(x) - f(f_1(y) + \langle \nabla f_1(y), x - y\rangle, \ldots, f_m(y) + \langle \nabla f_m(y), x - y\rangle) \leq M\frac{\sum_{i=1}^m L_i}{2}\|x - y\|^2 \quad \forall x, y \in Q.$$

Moreover, we have

$$0 \leq F(x) - F(y) - f(f_1(y) + \langle \nabla f_1(y), x - y\rangle, \ldots, f_m(y) + \langle \nabla f_m(y), x - y\rangle) + F(y)$$
$$\leq M\frac{\sum_{i=1}^m L_i}{2}\|x - y\|^2 \quad \forall x, y \in Q.$$

We may set $\psi_{\delta_k}(x,y) = f(f_1(y) + \langle \nabla f_1(y), x - y\rangle,\ldots, f_m(y) + \langle \nabla f_m(y), x - y\rangle) - F(y)$, $F_{\delta_k}(y) = F(y)$, and $\delta_k = 0$ for each k. As in problem (18), the estimates of the convergence rate remain the same, but the auxiliary problems (7) and (10) can become significantly more complicated. This problem can include a large number of specific cases (see [25, 26]), such as smooth optimization, nonsmooth optimization, minimax problem [1], composite optimization, and the problem with regularization.

4.8. Additional Examples

Consider without going into details additional examples of problem statements in which the concept of the model of function introduced in Section 2 can be useful.

1. Consider the following minimin problem [29]

$$f(x) \stackrel{\text{def}}{=} \min_{y\in Q} F(y,x) \to \min_{x\in\mathbb{R}^n}. \tag{23}$$

Let $F(y,x)$ be a smooth function and

$$\|\nabla F(y',x') - \nabla F(y,x)\|_2 \le L\|(y',x') - (y,x)\|_2 \quad \forall y,y'\in Q, \quad \forall x,x'\in\mathbb{R}^n.$$

Then (see [30]), if there exists a $\tilde{y}_\delta(x)\in Q$ such that

$$\langle \nabla_y F(\tilde{y}_\delta(x),x), y - \tilde{y}_\delta(x)\rangle \ge -\delta \quad \forall y\in Q,$$

then

$$F(\tilde{y}_\delta(x),x) - f(x) \le \delta, \quad \|\nabla f(x') - \nabla f(x)\|_2 \le L\|x' - x\|_2,$$

and

$$\left(F_\delta(x) = F(\tilde{y}_\delta(x),x) - 2\delta, \psi_\delta(z,x) = \langle \nabla_y F(\tilde{y}_\delta(x),x), z - x\rangle\right)$$

is a $(6\delta, 2L)$-model of the function $f(x)$ at the point x.

Thus, we obtain a $(6\delta, 2L)$-model that can be used for solving problem (23).

2. Consider the problem of finding the saddle point [29]

$$f(x) \stackrel{\text{def}}{=} \max_{y\in Q}[\langle x, b - Ay\rangle - \phi(y)] \to \min_{x\in\mathbb{R}^n}, \tag{24}$$

where $\phi(y)$ is μ-strongly convex with respect to the p-norm, $1 \le p \le 2$. Then, as was shown in [11], $f(x)$ is a smooth function with the Lipschitz constant of the gradient in the 2-norm

$$L = \frac{1}{\mu}\max_{\|y\|_p\le 1}\|Ay\|_2^2.$$

If $y_\delta(x)$ is the solution of the auxiliary maximization problem accurate to δ with respect to the function, then the pair

$$\left(F_\delta(x) = \langle x, b - Ay_\delta(x)\rangle - \phi(y_\delta(x)), \psi_\delta(z,x) = \langle b - Ay_\delta(x), z - x\rangle\right)$$

is a $(\delta, 2L)$-model of $f(x)$ at the point x.

3. Consider the function (cf. (21))

$$f(x) \stackrel{\text{def}}{=} \min_{y\in Q} \underbrace{\left\{\phi(y) + \frac{L}{2}\|y - x\|_2^2\right\}}_{\Lambda(x,y)}. \tag{25}$$

Let $\phi(y)$ be a convex function and

$$\max_{y\in Q}\left\{\Lambda(x,y(x)) - \Lambda(x,y) + \frac{L}{2}\|y - y(x)\|_2^2\right\} \le \delta.$$

Then (see [11]), it holds that

$$\left(F_\delta(x) = \phi(y(x)) + \frac{L}{2}\|y(x) - x\|_2^2 - \delta, \psi_\delta(z,x) = \langle L(x - y(x)), z - x\rangle\right)$$

is a (δ, L)-model of the function $f(x)$ at the point x.

5. CONCLUSIONS

We presented the gradient and fast gradient methods for the (δ, L)-model. Algorithms developed for these methods and estimates of the convergence rates were obtained. In Section 4, it was shown that these methods provide a powerful tool for solving a large class of problems. Note that the problems listed in Sec-

tion 4 do not exhaust the potential capabilities of the proposed concept. We believe that this approach can be used in many other problems, including stochastic, component-wise, and gradient free optimization [31]. Furthermore, it can be shown that the algorithms proposed in this paper are primal-dual [32]. Details will be presented in further studies.

APPENDIX

In this paper, we essentially used the fact that we are solving the auxiliary problem with an error not exceeding $\tilde{\delta}$ using the concept of Definition 4. It was shown that the $\tilde{\delta}$-solution in the sense of Definition 4 implies the $\tilde{\delta}$-optimal solution. The converse is generally not true; however, we try to give fairly general examples in which the converse result holds. The trivial case is $\tilde{\delta} = 0$. In this case, the first-order optimality criterion implies that these two definitions of $\tilde{\delta}$-solution are equivalent.

Assume that the following problem is being solved:

$$\alpha_k \psi(x) + V(x, x_k) \to \min_{x \in Q}, \tag{26}$$

where $\psi(x)$ is a convex function and $V(x, x_k)$ is a strongly convex function with the strong convexity constant equal to one. The auxiliary problem in the iterations of optimization methods often has this form. Certainly, there are cases in which this problem can be solved analytically, e.g., when the main problem is the smooth optimization without constraints and with the Euclidean prox-structure $V(x, y) = \frac{1}{2}\|x - y\|_2^2$. If problem (26) can be solved only numerically, then various approaches depending on the problem can be used.

Consider the case when

$$\psi(x) + V(x, x_k) = \sum_{i=1}^{n} [\psi_i(x_i) + V_i(x_i)].$$

Under this condition, problem (26) is separable. Therefore, it is sufficient to solve n one-dimensional problems each of which can be solved using the bisection method [33] in time $\mathbb{O}\left(\ln\left(\frac{1}{\varepsilon}\right)\right)$, where ϵ is the error with respect to the function.

If we additionally assume that $\psi(x)$ has an L-Lipschitzian gradient in the norm $\|\ \|$, then two approaches can be used. If $V(x, x_k)$ has an L-Lipschitzian gradient in the norm $\|\ \|$, then the problem can be solved in linear time $\mathbb{O}\left(\ln\left(\frac{1}{\varepsilon}\right)\right)$ [1]. If $V(x, x_k)$ does not have an L-Lipschitzian gradient in the norm $\|\ \|$, then $V(x, x_k)$ in problem (26) can be considered as a composite one. In this case, in order to obtain a linear convergence rate, the restart technique [14, 34] can be used.

ACKNOWLEDGMENTS

We are grateful to P. Dvurechenskii for a number of sources of literature.

FUNDING

The work by Tyurin was supported by the program of support of leading Russian universities, project no. 5-100. The work by Gasnikov was supported by the Russian Foundation for Basic Research, project no. 18-31-20005 mol_a_ved (the main part of the paper) and by the Russian Science Foundation, project 17-11-01027 (the Appendix).

REFERENCES

1. Yu. E. Nesterov, *Introduction to Convex Optimization* (Mosk. Tsentr Nepreryvnogo Matematicheskogo Obrazovaniya, Moscow, 2010) [in Russian].
2. A. S. Nemirovski and D. B. Yudin, *Complexity of Problems and Efficiency of Optimization Methods* (Nauka, Moscow, 1979) [in Russian].
3. Yu. Nesterov, "Gradient methods for minimizing composite functions," Math. Program. **140** (2), 125–161 (2013).
4. O. Devolder, F. Glineur, and Yu. Nesterov, "First-order methods with inexact oracle: the strongly convex case," CORE Discussion Papers, 2013/16 (2013).
https://www.uclouvain.be/cps/ucl/doc/core/documents/ coredp2013_16web.pdf

5. A. Ben-Tal and A. Nemirovski, *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications,* http://www2.isye.gatech.edu/~nemirovs/Lect_ModConvOpt.pdf
6. A. Taylor, J. Hendrickx, and F. Glineur, "Exact worst-case performance of first-order methods for composite convex optimization," arXiv:1512.07516
7. P. Ochs, J. Fadili, and T. Brox, "Non-smooth non-convex Bregman minimization: Unification and new algorithms," arXiv:1707.02278
8. J. Miral, "Optimization with first-order surrogate functions," in *Int. Conf. on Machine Learning (ICML-2013),* 2013, Vol. 28, pp. 783–791.
9. M. D. Gupta and T. Huang, "Bregman distance to l1 regularized logistic regression," in *19th International Conference on Pattern Recognition,* 2008, pp. 1–4.
10. A. V. Gasnikov, "Modern numerical optimization methods. The universal gradient descent method," arXiv:1711.00394
11. O. Devolder, F. Glineur, and Yu. Nesterov, "First-order methods of smooth convex optimization with inexact oracle," Math. Program **146**, 37–75 (2014).
12. O. Devolder, "Exactness, inexactness and stochasticity in first-order methods for large-scale convex optimization," PhD Thesis, ICTEAM and CORE, Universite Catholique de Louvain. 2013.
13. O. Devolder, F. Glineur, and Yu. Nesterov, "Intermediate gradient methods for smooth convex problems with inexact oracle," *Techn. Report of Universite catholique de Louvain, Center for Operations Research and Econometrics (CORE),* 2013.
14. P. Dvurechensky and A. Gasnikov, "Stochastic intermediate gradient method for convex problems with stochastic inexact oracle," J. Optim. Theory Appl. **171** (1), 121–145 (2016).
15. Yu. Nesterov, "Universal gradient methods for convex optimization problems," Math. Program. **152**, 381–404 (2015).
16. C. Guzman and A. Menirovski, "On lower complexity bounds for large–scale smooth convex optimization," J. Complexity **31** (1), 1–14.
17. Yu. Nesterov, "Complexity bounds for primal–dual methods minimizing the model of objective function," Math. Program. **171**, 311–330 (2018).
18. M. Jaggi, "Revisiting Frank–Wolfe: Projection-free sparse convex optimization," in *Int. Conf. on Machine Learning (ICML-2013),* 2013, pp. 427–435.
19. Z. Harchaoui, A. Juditsky, and A. Nemirovski, "Conditional gradient algorithms for norm-regularized smooth convex optimization," Math. Program. **152** (1–2), 75–112 (2015).
20. B. T. Polyak, Introduction to Optimization, (Nauka, Moscow, 1983; Optimization Software, New York, 1987).
21. N. Parikh and S. Boyd, "Proximal algorithms," Foundations Trends Optim. **1** (3), 127–239 (2014).
22. H. Lin, J. Mairal, and Z. Harchaoui, "A universal catalyst for first-order optimization," in *Advances in Neural Information Processing Systems,* 2015, pp. 3384–3392.
23. A. Rakhlin, O. Shamir, and K. Sridharan, "Making gradient descent optimal for strongly convex stochastic optimization," in *Proc. of the 29th Int. Conf. on Machine Learning (ICML-12),* 2012, pp. 449–456.
24. A. Juditsky and A. Nemirovski, "First order methods for nonsmooth convex large-scale optimization, I: General purpose methods," *Optimization for Machine Learning* (MIT Press, Cambridge, 2011), pp. 121–148.
25. A. Nemirovski, *Information-Based Complexity of Convex Programming,* Technion, Fall Semester 1994/95. http://www2.isye.gatech.edu/~nemirovs/Lec_EMCO.pdf
26. G. Lan, "Bundle-level type methods uniformly optimal for smooth and nonsmooth convex optimization," Math. Program. **149** (1–2), 1–45 (2015).
27. A. S. Nemirovskii and Yu. E. Nesterov, "Optimal methods of smooth convex minimization," Comput. Math. Math. Phys. **25** (2), 21–30 (1985).
28. S. Boyd and L. Vandenberghe, *Convex Optimization* (Cambridge Univ. Press, Cambridge, 2004).
29. A. V. Gasnikov, Efficient numerical methods for finding equilibriums in large transportation networks, Doctoral Dissertation in Mathematics and Physics, (Moscow, 2016).
30. A. V. Gasnikov, P. E. Dvurechensky, and Yu. E. Nesterov, "Stochastic gradient methods with inexact oracle," Trudy Mosc. Fiz.-Tekhn. Inst. **8** (1), 41–91 (2016).
31. A. Tyurin, "Mirror version of similar triangles method for constrained optimization problems," arXiv: 1705.09809
32. A. S. Anikin, A. V. Gasnikov, P. E. Dvurechenskii, A. I. Tyurin, and A. V. Chernov, "Dual Approaches to the Minimization of Strongly Convex Functionals with a Simple Structure under Affine Constraints," Comput. Math. Math. Phys. **57**, 1262–1276 (2017).
33. F. P. Vasil'ev, *Optimization Methods* (Faktorial, Moscow, 2011), Vol. 1 [in Russian].
34. A. Juditsky and Yu. Nesterov, "Deterministic and stochastic primal–dual subgradient algorithms for uniformly convex minimization," Stochastic Syst. **4** (1), 44–80 (2014).

Translated by A. Klimontovich