ISSN 0965-5425, Computational Mathematics and Mathematical Physics, 2017, Vol. 57, No. 8, pp. 1262–1276. © Pleiades Publishing, Ltd., 2017.
Original Russian Text © A.S. Anikin, A.V. Gasnikov, P.E. Dvurechensky, A.I. Tyurin, A.V. Chernov, 2017, published in Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki, 2017, Vol. 57, No. 8, pp. 1270–1284.

Dual Approaches to the Minimization of Strongly Convex Functionals with a Simple Structure under Affine Constraints

A. S. Anikin[a], A. V. Gasnikov[b, c]*, P. E. Dvurechensky[c, d], A. I. Tyurin[e], and A. V. Chernov[b]

[a] *Institute of System Dynamics and Control Theory, Siberian Branch, Russian Academy of Sciences, Irkutsk, 664033 Russia*

[b] *Moscow Institute of Physics and Technology, Dolgoprudnyi, Moscow oblast, 141700 Russia*

[c] *Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, 127051 Russia*

[d] *Weierstrass Institute of Applied Analysis and Stochastics, Berlin, 10117 Germany*

[e] *National Research University Higher School of Economics, Moscow, 101000 Russia*

**e-mail: gasnikov@yandex.ru*

Received February 3, 2016; in final form, May 12, 2016

Abstract—A strongly convex function of simple structure (for example, separable) is minimized under affine constraints. A dual problem is constructed and solved by applying a fast gradient method. The necessary properties of this method are established relying on which, under rather general conditions, the solution of the primal problem can be recovered with the same accuracy as the dual solution from the sequence generated by this method in the dual space of the problem. Although this approach seems natural, some previously unpublished rather subtle results necessary for its rigorous and complete theoretical substantiation in the required generality are presented.

Keywords: minimization of strongly convex functionals, primal-dual methods, fast gradient method, dual problem, regularization of dual problems, restart technique, strong convexity, PageRank problem.

DOI: 10.1134/S0965542517080048

1. INTRODUCTION

In this paper, we generalize the results of [1] (including an improvement of the convergence rate estimate for the method therein), where a technique was proposed for solving an entropy-linear programming (ELP) problem. Specifically, its solution was recovered (with the help of explicit formulas) from the solution of a specially regularized dual problem. This work develops the ideas of [2–13], where various primal-dual methods were proposed for a broad class of problems. The term "primal-dual" was coined in [5] for methods in which the solution of the primal (dual) problem is recovered (with the same accuracy but without considerable additional effort) from the solution of the corresponding dual (primal) problem. To make this paper self-contained, we tried to present all necessary derivations, although some of them are not original.

In Section 2, following [14], we describe a fast gradient method. In contrast to [1], we examine its primal-duality [2] and the following property: the sequence of points generated by the method lies in the ball centered at the solution of radius equal to the distance from the starting point of the method to the solution of the problem. Both these properties are used in Section 3 to substantiate a technique for recovering the solution of the primal problem (of minimizing a strongly convex function of simple structure under affine constraints) from the sequence generated by the method in the dual space. At the end of Section 3, we describe a direct generalization of the construction from [1] associated with the regularization of the dual problem. This generalization does not require that the method be primal-dual, but leads to a somewhat slower convergence rate. Specifically, as applied to various ELP problems, the method proposed in this paper finds solutions faster by several orders of magnitude than the method from [1].

2. PRIMAL-DUALITY OF THE FAST GRADIENT METHOD

Consider the convex optimization problem

$$f(x) \to \min_x. \tag{1}$$

By the solution of this problem, we mean $\overline{x}^N$ that such

$$f(\overline{x}^N) - f_* \le \varepsilon,$$

where $f_* = f(x_*)$ is an optimal value of the functional in problem (1) and x_* is the solution of problem (1). Define the set

$$B_R(x_*) = \{x: \|x - x_*\|_2 \le R\}.$$

Let

$$x^{k+1} = x^k - h\nabla f(x^k). \tag{2}$$

Additionally, for $x \in B_{\sqrt{2}R}(x_*)$, where

$$R = \|x^0 - x_*\|_2 = \|x_*\|_2,$$

assume that

$$\|\nabla f(x)\|_2 \le M.$$

Then, combining (2) with this inequality and (4) (see below) yields

$$\|x - x^{k+1}\|_2^2 = \|x - x^k + h\nabla f(x^k)\|_2^2 = \|x - x^k\|_2^2 + 2h\langle \nabla f(x^k), x - x^k\rangle + h^2\|\nabla f(x^k)\|_2^2$$
$$\le \|x - x^k\|_2^2 + 2h\langle \nabla f(x^k), x - x^k\rangle + h^2M^2.$$

From this (at $x = x_*$) it follows that

$$f\left(\frac{1}{N}\sum_{k=0}^{N-1} x^k\right) - f_* \le \frac{1}{N}\sum_{k=0}^{N-1} f(x^k) - f(x_*) \le \frac{1}{N}\sum_{k=0}^{N-1}\langle \nabla f(x^k), x^k - x_*\rangle$$
$$\le \frac{1}{2hN}\sum_{k=0}^{N-1}\left\{\|x_* - x^k\|_2^2 - \|x_* - x^{k+1}\|_2^2\right\} + \frac{hM^2}{2} = \frac{1}{2hN}\left(\|x_* - x^0\|_2^2 - \|x_* - x^N\|_2^2\right) + \frac{hM^2}{2}.$$

Choosing

$$h = \frac{R}{M\sqrt{N}}$$

and setting

$$\overline{x}^N = \frac{1}{N}\sum_{k=0}^{N-1} x^k,$$

we obtain

$$f(\overline{x}^N) - f_* \le \frac{MR}{\sqrt{N}}. \tag{3}$$

Note that

$$0 \le \frac{1}{2hk}\left(\|x_* - x^0\|_2^2 - \|x_* - x^k\|_2^2\right) + \frac{hM^2}{2}.$$

Therefore, for $k = 0, \ldots, N$, we have

$$\|x_* - x^k\|_2^2 \le \|x_* - x^0\|_2^2 + h^2M^2k \le 2\|x_* - x^0\|_2^2;$$

i.e.,

$$\left\|x^k - x_*\right\|_2 \le \sqrt{2}\left\|x^0 - x_*\right\|_2, \quad k = 0, \ldots, N. \tag{4}$$

For nonsmooth problems, estimate (3) is sharp up to a multiplicative constant [15] (here and below, talking about the sharpness of estimates, we assume that the space in which optimization is performed has a sufficiently high dimension, i.e., the number of iteration steps to be executed in the method does not exceed the space dimension). However, if the gradient of $f(\overline{x}^N)$ is additionally Lipschitz continuous, i.e.,

$$\left\|\nabla f(y) - \nabla f(x)\right\|_2 \le L\left\|y - x\right\|_2,$$

where $x, y \in B_R(x_*)$ (see (6)), then

$$\frac{1}{2L}\left\|\nabla f\left(x^k\right)\right\|_2^2 \le f\left(x^k\right) - f_*. \tag{5}$$

This inequality is a formal representation of the following simple geometric fact: if we draw the tangent to the function $f(x)$ at the point x^k, i.e.,

$$f\left(x^k\right) + \left\langle \nabla f\left(x^k\right), x - x^k\right\rangle,$$

and use this tangent to construct the parabola

$$f\left(x^k\right) + \left\langle \nabla f\left(x^k\right), x - x^k\right\rangle + \frac{L}{2}\left\|x - x^k\right\|_2^2,$$

then the latter majorizes $f(x)$, i.e.,

$$f(x) \le f\left(x^k\right) + \left\langle \nabla f\left(x^k\right), x - x^k\right\rangle + \frac{L}{2}\left\|x - x^k\right\|_2^2.$$

Specifically, this inequality holds at the minimum point of the parabola,

$$x^k - \frac{1}{L}\nabla f\left(x^k\right).$$

When the argument passes from the point x^k to the parabola's minimum point, the parabola increases by

$$\frac{1}{2L}\left\|\nabla f\left(x^k\right)\right\|_2^2.$$

Therefore, we obtain inequality (5), which allows us to refine the above argument. As before, we write

$$\left\|x^{k+1} - x_*\right\|_2^2 = \left\|x^k - x_*\right\|_2^2 - 2h\left\langle \nabla f\left(x^k\right), x^k - x_*\right\rangle + h^2\left\|\nabla f\left(x^k\right)\right\|_2^2$$

$$\le \left\|x^k - x_*\right\|_2^2 + 2Lh^2\left(f\left(x^k\right) - f_*\right) - 2h\left(f\left(x^k\right) - f_*\right) = \left\|x^k - x_*\right\|_2^2 + 2h(Lh - 1)\left(f\left(x^k\right) - f_*\right).$$

For $h \le 1/L$, it follows that

$$\left\|x^{k+1} - x_*\right\|_2^2 \le \left\|x^k - x_*\right\|_2^2, \quad k = 0, \ldots, N-1.$$

Therefore,

$$\left\|x^k - x_*\right\|_2 \le \left\|x^0 - x_*\right\|_2, \quad k = 0, \ldots, N. \tag{6}$$

Setting

$$h = \frac{1}{2L},$$

we obtain

$$f\left(\overline{x}^N\right) - f_* \le \frac{1}{N}\sum_{k=0}^{N-1} f\left(x^k\right) - f(x_*) \le \frac{2L}{N}\sum_{k=0}^{N-1}\left\{\left\|x^k - x_*\right\|_2^2 - \left\|x^{k+1} - x_*\right\|_2^2\right\} \le \frac{2LR^2}{N}.$$

Remark 1. The fact that $f\left(\bar{x}^N\right) - f_*$ depends only on $MR/\sqrt{N}$ and/or LR^2/N is a consequence of the Π-theorem from the theory of dimensions [16]. Introducing the accuracy

$$f\left(\bar{x}^N\right) - f_* \le \varepsilon,$$

we can show that there are only two (independent) dimensionless quantities in terms of the parameters introduced, namely,

$$\frac{M^2R^2}{\varepsilon^2} \quad \text{and} \quad \frac{LR^2}{\varepsilon}.$$

By the Π-theorem, any dimensionless variable has to be functionally expressed in terms of these two (basic) quantities. Specifically,

$$N = G\left(\frac{M^2R^2}{\varepsilon^2}, \frac{LR^2}{\varepsilon}\right).$$

When the Lipschitz continuity of the gradient cannot be guaranteed, the situation simplifies to

$$N = \tilde{G}\left(\frac{M^2R^2}{\varepsilon^2}\right).$$

Obtained by method (2), estimate (3) corresponds to this formula. Moreover, as was noted above, this estimate is sharp in the class of nonsmooth convex problems. Unfortunately, the convergence rate estimate obtained for method (2) with the step $h = 1/(2L)$ ceases to be optimal in the smooth case.

Incidentally, the Π-theorem also implies that the stepsize h in method (2) in the nonsmooth case has to be calculated using the formula

$$h = c\frac{\varepsilon}{M^2}, \quad c > 0,$$

which can be derived from by the above one

$$h = \frac{R}{M\sqrt{N}}$$

by expressing N in terms of ε with the help of (3) and setting

$$\varepsilon = \frac{MR}{\sqrt{N}}.$$

In the smooth case, h is determined by the relation

$$W\left(h\frac{M^2}{\varepsilon}, hL\right) = 1,$$

while, in the stochastic case [4] (the gradient is replaced by a stochastic gradient with variance σ^2), by the relation

$$\tilde{W}\left(h\frac{M^2}{\varepsilon}, hL, h\frac{\sigma}{R}\right) = 1.$$

If the gradient is Lipschitz continuous, the above convergence rate estimate can be improved [15] (e.g., by using the conjugate gradient method [15, 17]). The same follows from local convergence rate estimates of the heavy ball method [17]. Among the wide variety of "accelerated methods" that converge according to lower bounds, we distinguish the fast gradient method proposed by Yu.E. Nesterov in his candidate's dissertation in 1983. In addition to the fact that it was one of the first methods (making no use of auxiliary one- or two-dimensional optimization [15]) whose global convergence was rigorously proved on the basis of lower bounds (in the smooth case), the method was found to have good properties of type (6). However, the most important property to be used in this paper is its primal-duality. This method was further developed and applied in Nesterov's doctoral dissertation [5].

Let us construct a fast gradient method (FGM). In many respects, we follow the way of understanding the FGM proposed recently in [14]. Nevertheless, we need to obtain property (6) and primal-duality from

this argument. Whether or not the FGM has these properties was not examined in [14], so all the necessary arguments will be presented below to the required degree of detail.

Preliminarily, we define two numerical sequences of steps $\{\alpha_k, \tau_k\}$:

$$\alpha_1 = \frac{1}{L}, \quad \alpha_k^2 L = \alpha_{k+1}^2 L - \alpha_{k+1}, \quad \tau_k = \frac{1}{\alpha_{k+1} L}.$$

Explicit formulas can also be written. Below, we will use a simplified version of these sequences [14] defined as

$$\alpha_1 = \frac{1}{L}, \quad \alpha_k^2 L = \alpha_{k+1}^2 L - \alpha_{k+1} + \frac{1}{4L}, \quad \tau_k = \frac{1}{\alpha_{k+1} L}.$$

In this case,

$$\alpha_{k+1} = \frac{k+2}{2L}, \quad \tau_k = \frac{1}{\alpha_{k+1} L} = \frac{2}{k+2},$$

FGM ($x^0 = y^0 = z^0$),

1. $x^{k+1} = \tau_k z^k + (1 - \tau_k) y^k$;

2. $y^{k+1} = x^{k+1} - \frac{1}{L} \nabla f(x^{k+1})$;

3. $z^{k+1} = z^k - \alpha_{k+1} \nabla f(x^{k+1})$.

The last formula in the proof of Lemma 4.3 of [14] implies that (for all x)

$$\alpha_{k+1}^2 L f\left(y^{k+1}\right) - \left(\alpha_{k+1}^2 L - \alpha_{k+1}\right) f\left(y^k\right)$$
$$\leq \alpha_{k+1} \left\{ f\left(x^{k+1}\right) + \left\langle \nabla f\left(x^{k+1}\right), x - x^{k+1} \right\rangle \right\} + \frac{1}{2} \left\| z^k - x \right\|_2^2 - \frac{1}{2} \left\| z^{k+1} - x \right\|_2^2.$$

Summing up the results over $k = 0, \ldots, N-1$, we obtain

$$\begin{aligned} \alpha_N^2 L f\left(y^N\right) &\leq \min_x \left\{ \sum_{k=0}^{N-1} \alpha_{k+1} \left\{ f\left(x^{k+1}\right) + \left\langle \nabla f\left(x^{k+1}\right), x - x^{k+1} \right\rangle \right\} + \frac{1}{2} \left\| z^k - x \right\|_2^2 - \frac{1}{2} \left\| z^{k+1} - x \right\|_2^2 \right\} \\ &\leq \left(\sum_{k=0}^{N-1} \alpha_{k+1} \right) f_* + \frac{1}{2} \left\| z^0 - x_* \right\|_2^2 - \frac{1}{2} \left\| z^N - x_* \right\|_2^2. \end{aligned} \tag{7}$$

Note that

$$\max \left\{ \left\| x^k - x_* \right\|_2, \left\| y^k - x_* \right\|_2, \left\| z^k - x_* \right\|_2 \right\} \leq \left\| x^0 - x_* \right\|_2, \quad k = 0, \ldots, N. \tag{8}$$

Indeed, setting $N := k$ and $k := i$ in (7), taking into account that α_i are independent of k, and using $f(y_k) \geq f_*$ and $\alpha_k^2 L = \sum_{i=0}^{k-1} \alpha_{i+1}$, we obtain

$$\left\| z^k - x_* \right\|_2^2 \leq \left\| z^0 - x_* \right\|_2^2$$

(the same formula can be derived for the simplified stepsize selection scheme). Combining this relation with inequality (6) for y^{k+1} and taking into account that the squared Euclidean norm is convex, we have

$$\left\| y^{k+1} - x_* \right\|_2^2 \leq \left\| x^{k+1} - x_* \right\|_2^2 = \left\| \tau_k \left(z^k - x_* \right) + (1 - \tau_k) \left(y^k - x_* \right) \right\|_2^2$$
$$\leq \tau_k \left\| z^k - x_* \right\|_2^2 + (1 - \tau_k) \left\| y^k - x_* \right\|_2^2 \leq \tau_k \left\| z^0 - x_* \right\|_2^2 + (1 - \tau_k) \left\| y^k - x_* \right\|_2^2$$
$$= \tau_k \left\| x^0 - x_* \right\|_2^2 + (1 - \tau_k) \left\| y^k - x_* \right\|_2^2 = \tau_k \left\| y^0 - x_* \right\|_2^2 + (1 - \tau_k) \left\| y^k - x_* \right\|_2^2.$$

From this, (8) is obtained by applying an induction argument.

Returning to formula (7), we show that the FGM is primal-dual. For this purpose, (7) is rewritten for the simplified stepsize selection version:

$$\frac{(N+1)^2}{4L} f\left(y^N\right) + \sum_{k=1}^{N-1} \frac{1}{4L} f\left(y^k\right) \le \min_x \left\{ \sum_{k=0}^{N-1} \frac{k+2}{2L} \left\{ f\left(x^{k+1}\right) + \left\langle \nabla f\left(x^{k+1}\right), x - x^{k+1} \right\rangle \right\} + \frac{1}{2} \left\| z^0 - x \right\|_2^2 \right\},$$

i.e.,

$$f\left(\breve{y}^N\right) \le \frac{4L}{N(N+3)} \min_x \left\{ \sum_{k=0}^{N-1} \frac{k+2}{2L} \left\{ f\left(x^{k+1}\right) + \left\langle \nabla f\left(x^{k+1}\right), x - x^{k+1} \right\rangle \right\} + \frac{1}{2} \left\| z^0 - x \right\|_2^2 \right\}, \tag{9}$$

where

$$\breve{y}^N = \frac{1}{N(N+3)} \left(\sum_{k=0}^{N-1} y^k + (N+1)^2 y^N \right).$$

In fact, it is inequality (9) that allows us to recover the solution of the primal problem from that of the dual one by applying the FGM in the next section.

Below is the main result of this section.

Theorem 1. *Suppose that the functional of problem* (1) *has the property*

$$\left\| \nabla f(y) - \nabla f(x) \right\|_2 \le L \left\| y - x \right\|_2, \quad x, y \in B_R\left(x_*\right). \tag{10}$$

Then the FGM generates a sequence of points $\{x^k, y^k, z^k\}_{k=0}^N$ *that satisfies relations* (8) *and* (9). *Moreover, formula* (9) *can be rewritten as*

$$f\left(\breve{y}^N\right) - f_* \le \frac{2L \left\| z^0 - x_* \right\|_2^2}{N(N+3)} \le \frac{2L \left\| z^0 - x_* \right\|_2^2}{(N+1)^2} = \frac{2LR^2}{(N+1)^2}.$$

Remark 2. The novelty of this theorem as compared with all its analogues is that, although optimization problem (1) is solved on an unbounded set, the parameters involved in the convergence rate estimates are determined by the distance from the starting point to the solution. Thus, the entire iterative process belongs to the (Euclidean) ball centered at the solution of radius equal to the distance from the starting point to the solution. Since this distance is usually not known in advance, it may seem that this remark is of low value. However, we will show in the next section that this remark plays an important role in the substantiation of the proposed primal-dual approach (although there is another line of reasoning, which does not rely on a condition of type (8) [18]). Interestingly, if the solution of problem (1) is not unique (the solution set is denoted by X), then R^2 can be understood as $\min\limits_{x_* \in X} \left\| z^0 - x_* \right\|_2^2$.

Remark 3. Theorem 1 can be extended to the case where the set on which optimization is performed (optimization set) does not coincide with the entire space, instead being, for example, a nonnegative orthant or a simplex. In the general case, this requires that a prox-structure [5, 19] other than the Euclidean one (which was used above) be introduced into the problem. As a result, the presentation becomes somewhat more complicated. Specifically, the stepsizes in the FGM have to be other than those in the direct gradient method, namely, their proximal versions have to be used [4]. To the best of our knowledge, this has been performed (in the generality of Theorem 1) only for the Euclidean prox-structure, but with arbitrary optimization sets.

Remark 4. Actually, the construction described above can be extended (with the preservation of the main result—Theorem 1) to composite optimization problems [19, 20] and to problems in which the constant L is not known, but has to be chosen in the course of the solution process (see [5, 19]). Additionally, relying on the FGM and the remark made above, we can construct a corresponding version of the universal method from [21]. All the above generalizations can also be performed using the concept of an inexact oracle [4, 11, 22–25].

3. APPLICATION TO THE MINIMIZATION OF A STRONGLY CONVEX FUNCTIONAL OF SIMPLE STRUCTURE UNDER AFFINE CONSTRAINTS

Consider the problem

$$g(x) \to \min_{Ax=b,\ x\in Q}, \tag{11}$$

where $g(x)$ is a 1-strongly convex function in the p-norm $(1 \le p \le 2)$. Consider the dual problem

$$F(y) = \max_{x\in Q}\{\langle y, b - Ax\rangle - g(x)\} \to \min_{y}. \tag{12}$$

In many important applications, the basic contribution to the computational complexity of the inner maximization problem is made by the multiplication Ax $(A^{\mathrm{T}}y)$. For example, this occurs for separable functionals

$$g(x) = \sum_{k=1}^{n} g_k(x_k)$$

under box constraints Q. More specifically, this takes place for ELP problems [1] that involve an explicit formula for $x(y)$.

In the general case, the inner maximization problem is not solved exactly (with the use of explicit formulas). Nevertheless, since $g(x)$ is strongly convex (the same is true if $g(x)$ is separable rather than strongly convex), the complexity bound for the solution of this auxiliary problem depends logarithmically on the solution accuracy (at every iteration step of the outer method). Accordingly, the inaccuracy of the oracle producing the gradient for the outer minimization problem $F(y)$ can be neglected. An accurate account leads to only logarithmic correction terms in the resulting complexity bounds for the method (see, e.g., [8, 19, 24, 25]). Therefore, for illustrative purposes, we assume in what follows that there is an explicit formula for $x(y)$.

Applying the FGM from Section 2 ($z^0 = 0$) to problem (12) and using formula (9) (with substitutions $x \to y$, $\breve{y} \to \tilde{y}$), we obtain

$$F(\tilde{y}^N) \le \frac{4L}{N(N+3)} \min_{y}\left\{\sum_{k=0}^{N-1}\frac{k+2}{2L}\left\{F(y^{k+1}) + \left\langle \nabla F(y^{k+1}),\ y - y^{k+1}\right\rangle\right\} + \frac{1}{2}\left\|z^0 - y\right\|_2^2\right\}, \tag{13}$$

where (see, e.g., [5])

$$L = \max_{\|x\|_p \le 1} \|Ax\|_2^2.$$

Specifically, for an ELP problem, $p = 1$ (see [1, 13]) and

$$L = \max_{k=1,\ldots,n} \left\|A^{\langle k\rangle}\right\|_2^2,$$

where $A^{\langle k\rangle}$ is the kth column of the matrix $A^{\langle k\rangle}$. For the PageRank problem (see Remark 10 below), $p = 2$ and

$$L = \lambda_{\max}\left(A^{\mathrm{T}}A\right) = \sigma_{\max}(A).$$

Inequality (13) yields ($R^2 = \left\|z^0 - y_*\right\|_2^2 = \left\|y_*\right\|_2^2$, where y_* is the solution of problem (12))

$$F(\tilde{y}^N) - \min_{y\in B_{3R}(0)}\left\{\sum_{k=0}^{N-1}\frac{2(k+2)}{N(N+3)}\left\{F(y^{k+1}) + \left\langle \nabla F(y^{k+1}),\ y - y^{k+1}\right\rangle\right\}\right\} \le \frac{18LR^2}{(N+1)^2} \stackrel{\text{def}}{=} \gamma_N. \tag{14}$$

Define

$$\lambda_k = \frac{2(k+2)}{N(N+3)},$$

$$x^N = \sum_{k=0}^{N-1} \lambda_k x\left(y^{k+1}\right) = \frac{2}{N(N+3)} \sum_{k=0}^{N-1} (k+2)\, x\left(y^{k+1}\right) = \frac{N^2+N-2}{N(N+3)} x^{N-1} + 2\frac{N+1}{N(N+3)} x\left(y^N\right).$$

By the definition of $F(y)$ (see (12)) and $x(y)$, inequality (14) can be rewritten as

$$F\left(\tilde{y}^N\right) - \sum_{k=0}^{N-1} \lambda_k \left\langle y^{k+1}, b - Ax\left(y^{k+1}\right)\right\rangle + \sum_{k=0}^{N-1} \lambda_k g\left(x\left(y^{k+1}\right)\right)$$
$$- \min_{y \in B_{3R}(0)} \left\{ \sum_{k=0}^{N-1} \lambda_k \left\langle b - Ax\left(y^{k+1}\right), y - y^{k+1}\right\rangle \right\} \le \gamma_N,$$

which is similar to the argument used in [7, Section 3]. Taking into account

$$\sum_{k=0}^{N-1} \lambda_k = 1,$$

we obtain

$$F\left(\tilde{y}^N\right) + g\left(\sum_{k=0}^{N-1} \lambda_k x\left(y^{k+1}\right)\right) + \max_{y \in B_{3R}(0)} \left\{ \left\langle A\sum_{k=0}^{N-1} \lambda_k x\left(y^{k+1}\right) - b, y \right\rangle \right\} \le \gamma_N,$$

i.e.,

$$F\left(\tilde{y}^N\right) + g\left(x^N\right) + 3R\left\|Ax^N - b\right\|_2 \le \gamma_N.$$

From this (in many respects, the argument below repeats that used in [3, Section 6.11]), using

$$Ax_* = b$$

and the weak duality

$$-g(x_*) \le F(y_*),$$

we obtain

$$g\left(x^N\right) - g(x_*) \le g\left(x^N\right) + F(y_*) \le g\left(x^N\right) + F\left(\tilde{y}^N\right) \le g\left(x^N\right) + F\left(\tilde{y}^N\right) + 3R\left\|Ax^N - b\right\|_2 \le \gamma_N.$$

Applying the definition of $F(y)$ and property (8) yields

$$-g(x_*) = \left\langle y_*, b - Ax_*\right\rangle - g(x_*) = F(y_*) \ge \left\langle y_*, b - Ax^N\right\rangle - g\left(x^N\right)$$
$$\Rightarrow g(x_*) - g\left(x^N\right) \le R\left\|Ax^N - b\right\|_2,$$
$$R\left\|Ax^N - b\right\|_2 \le -g\left(x^N\right) + \left\langle \tilde{y}^N, b - Ax^N\right\rangle + g\left(x^N\right) + 3R\left\|Ax^N - b\right\|_2$$
$$\le F\left(\tilde{y}^N\right) + g\left(x^N\right) + 3R\left\|Ax^N - b\right\|_2 \le \gamma_N.$$

It follows that

$$\left|g\left(x^N\right) - g(x_*)\right| \le \gamma_N, \quad R\left\|Ax^N - b\right\|_2 \le \gamma_N.$$

Since

$$g\left(x^N\right) - g(x_*) \le F\left(\tilde{y}^N\right) + g\left(x^N\right) \le \gamma_N,$$

the following result is valid.

Theorem 2. *Suppose that problem* (11) *is solved by passing to problem* (12) *on the basis of the above-written formulas. Let the stopping criterion for the FGM be given by*

$$F(\tilde{y}^N) + g(x^N) \le \varepsilon, \quad \|Ax^N - b\|_2 \le \tilde{\varepsilon}.$$

Then the FGM is guaranteed to terminate after at most

$$\max\left\{\sqrt{\frac{18LR^2}{\varepsilon}}, \sqrt{\frac{18LR}{\tilde{\varepsilon}}}\right\}$$

iteration steps.

Remark 5. Note that the approach to the simultaneous solution of the primal and dual problems involves the unknown parameter R. However, this parameter is not involved in the algorithm or its stopping criterion. It appears only in the estimate for the number of iterations. Unfortunately, this result can rarely be achieved. It is nontrivial that we managed to achieve it in this context. Usually, in solving a dual problem, the optimization set (which is, as a rule, the entire space or the direct product of the space and the nonnegative orthant) is artificially compactified [2, 3, 11]. As a result, methods are used that require the projection onto a ball of beforehand unknown radius. This difficulty (the size of the dual solution is not known) is usually overcome using a restart procedure [1, 4, 24], which usually increases the number of iteration steps by at least one order of magnitude, or by applying Slater relaxation, which is frequently even more expensive in terms of the required number of iteration steps (see [1]).

Remark 6. Unfortunately, in some applications, $g(x)$ is only strictly (but not strongly) convex. In this case (although the dual problem is smooth) nothing can be said about the Lipschitz constant for the gradient, which is explicitly involved in every FGM step. As was noted above, this difficulty is resolved by adaptive selection of L and, in a more general case (when the smoothness of $g(x)$ cannot be guaranteed) by applying a universal method (see [11, 20, 24]). Nevertheless, how well the above-described constructions work as applied to a smooth dual problem (on an unbounded set) when the Lipschitz constant is not uniformly bounded (on this set) has earlier remained an open question. Indeed, assuming that the error of the method in iteration can be larger than the initial error and that this iteration error depends on the smoothness properties of the functional, we obtain a vicious circle. This indeed occurs in the case of inaccurate estimation. However, as was shown above for deterministic problem formulations, a suitable choice of primal-dual methods makes it possible to avoid this difficulty in a natural way, i.e., without using (conventional) artificial compactification, which leads to additional costs on restarts.

Remark 7. The class of problems to which the above-described approach applies can be significantly expanded, for example, if problem (11) admits inequality constraints of the form $Cx \le d$ or, in a more general case, of the form $Cx - d \in K$, where the cone K has a simple dual description [26]. Moreover, instead of problem (11), we can consider the minimization of a functional having a Legendre representation of form (12) (see [24]). In this case, minimization in y can be performed over an arbitrary convex set.

Remark 8. When the dual space is of low dimension, the FGM can be replaced by the ellipsoid method (which does not require the smoothness of the dual functional). This method is also primal-dual [3]. In the context of this construction, interesting examples appear when the dimension of the primal space is huge, but there is a linear minimization oracle that (despite the huge dimension of the primal space) efficiently calculates the gradient of the dual functional [9, 10, 27].

Remark 9. It is useful to note that, in many important cases, the approach described in this section (especially in the context of Remark 8) can be used to solve the problem of finding a gradient mapping, which arises (in projection onto a feasible set of rather complex structure) at every iteration step in most iterative methods [3, 5, 19]. In this case, the general divide-and-conquer idea as applied to the numerical solution of convex optimization problems has the form of an iterative process with a simpler problem (than the original one) solved at every step. The degree of difficulty of the problem to be solved at every step and the number of such steps can be varied. A good example is composite optimization [19, 20]. Some of the complexity of the problem formulation (in the form of a composite) is transferred completely (without linearization) to the problem to be solved at every iteration step. If the composite is fairly good, this operation does not have a large effect on the cost of an iteration step, but can substantially reduce the number of iteration steps (e.g., for a nonsmooth composite). Other examples concerning this subject can be found in [11, 12, 24, 25]). Here, the general line of reasoning can be roughly described as follows. As a rule, any complication of iteration leads to a reduced number of iteration steps. On the other hand, a single iteration step always involves the computation (updating) of the gradient or its (say, stochastic) analogue used in the method. For deterministic methods involving the gradient, the basic cost of an iteration step is formed by the computation of this gradient (as a rule, this means the multiplication of a matrix and a vector, i.e., $O(n^2)$ arithmetic operations). After the gradient has been computed, an iteration step requires at most $O(n\ln(n/\varepsilon))$ arithmetic operations. Thus, there is a rather large gap, which can cover additional computations transferred from the formulation of the problem to every step in hope of reducing the number of steps. Specifically, if we consider the problem (see [8, 25])

$$\frac{1}{2}\|Ax-b\|_2^2+\mu\sum_{k=1}^{n}x_k\ln x_k\to\min_{x\in S_n(1)}$$

with a sufficiently large $\mu>0$, then the strong convexity of the entropy composite has to be taken into account. At every iteration step, the usual FGM in the composite version as applied to this problem requires solving a nearly separable problem. The constructions described in this section turn out to be fairly useful for solving such a problem. Moreover, this problem is presented here, because, for a certain parameter value $\mu>0$, it becomes "equivalent" to the original problem (11) with an entropy-type functional. These different ways of understanding the same problem were discussed in [8] (in terms of its formulation and practical computations). Additionally, some constructions (similar to those described in this section) were presented that make it possible to determine the parameter $\mu>0$ at a low additional cost, so that the above-noted correspondence takes place (see also [25]).

Remark 10 (PageRank and lower bounds). At first glance, the estimates given in Theorem 2 seem to lead to contradictions. Let us explain this by using the following example [24].

The problem of finding $x^*\in\mathbb{R}^n$ such that

$$Ax^*=b$$

is reduced to the smooth convex optimization problem

$$f(x)=\|Ax-b\|_2^2\to\min_x.$$

The convergence rate of the solution to this problem [15] satisfies the lower bound

$$f\left(x^N\right)\geq\Omega\left(L_xR_x^2/N^2\right),\quad L_x=\sigma_{\max}(A),\quad R_x=\|x^*\|_2,$$

which implies that, only for

$$N\geq\Omega(\sqrt{L_x}R_x/\varepsilon),$$

we can guarantee the inequality

$$f\left(x^N\right)\leq\varepsilon^2,\quad\text{i.e.,}\quad\left\|Ax^N-b\right\|_2\leq\varepsilon.$$

However, for special matrices, this lower bound can be improved. Consider the problem of finding a PageRank vector [28] ($n\sim10^{10}$), which can be written as

$$Ax=\begin{pmatrix}\left(P^{\mathrm{T}}-I\right)\\ 1\ldots\ldots1\end{pmatrix}x=\begin{pmatrix}0\\1\end{pmatrix}=b,$$

where I is the identity matrix. By the Perron–Frobenius theorem [28], the solution of this system with an irreducible stochastic matrix P is unique and positive: $x>0$. This system of equations can be reduced to the degenerate convex optimization problem

$$\frac{1}{2}\|x\|_2^2\to\min_{Ax=b}.$$

Consider the dual problem for this one (since the system $Ax=b$ is consistent, Fredholm's theorem implies that there is no y such that $A^{\mathrm{T}}y=0$ and $\langle b,y\rangle>0$; therefore, the dual problem has a finite solution):

$$\min_{Ax=b}\frac{1}{2}\|x\|_2^2=\min_x\max_y\left\{\frac{1}{2}\|x\|_2^2+\langle b-Ax,y\rangle\right\}=\max_y\min_x\left\{\frac{1}{2}\|x\|_2^2+\langle b-Ax,y\rangle\right\}$$
$$=\max_y\left\{\langle b,y\rangle-\frac{1}{2}\left\|A^{\mathrm{T}}y\right\|_2^2\right\}.$$

Given the solution y^* of the dual problem (for example, with a minimum Euclidean norm)

$$\langle b,y\rangle-\frac{1}{2}\left\|A^{\mathrm{T}}y\right\|_2^2\to\max_y,$$

we can recover the solution of the primal problem

$$x(y) = A^{\mathrm{T}} y.$$

Additionally, if the dual problem is solved numerically by applying the FGM, then, by Theorem 2,

$$\left\| Ax^N - b \right\|_2 \le \frac{8 L_y R_y}{N^2},$$

where

$$L_y = \sigma_{\max}\left(A^{\mathrm{T}}\right) = \sigma_{\max}(A) = L_x, \quad R_y = \left\| y^* \right\|_2.$$

This seems to contradict the lower bound

$$\left\| Ax^N - b \right\|_2 \ge \Omega(\sqrt{L_x} R_x / N).$$

However, it is important to recall [15] that this lower bound was established for all $N \le n$ (n is the dimension of the vector x) and it can be improved by applying the above procedure only if the matrix A is additionally assumed to satisfy the condition

$$L_y R_y \ll n \sqrt{L_x} R_x,$$

which narrows down the class in which the lower bound

$$\Omega(\sqrt{L_x} R_x / N)$$

was obtained. In typical situations, it can be expected that $R_y \gg R_x$, which hinders the fulfillment of the required condition.

Remark 11. Certain nuances arise in an attempt to extend the results of this paper to the case where only a stochastic gradient is available instead of the gradient. We do not discuss this case in detail, but note that answers to many questions have been obtained for randomized componentwise methods [13]. Specifically, the PageRank problem from Remark 10 can be solved using a direct accelerated componentwise method or its dual. The estimates are as follows (see [13]):

$$E\left[\left\| Ax^N - b \right\|_2^2\right] = E\left[f\left(x^N\right)\right] - f_* = \mathrm{O}\left(n^2 \frac{\overline{L}_x R_x^2}{N^2}\right), \quad \overline{L}_x^{1/2} = \frac{1}{n}\sum_{k=1}^{n} \left\| A^{\langle k \rangle} \right\|_2 \le 2, \quad R_x^2 \le 2$$

(for the primal problem),

$$E\left[\left\| Ax^N - b \right\|_2\right] = \mathrm{O}\left(n^2 \frac{\overline{L}_y R_y}{N^2}\right), \quad \overline{L}_y^{1/2} = \frac{1}{n+1}\sum_{k=1}^{n+1} \left\| A_k \right\|_2$$

(for the dual problem).

Moreover, if the matrix P has sn nonzero elements ($s \ll n$), then one iteration step in both methods requires, on average, $\mathrm{O}(s)$ arithmetic operations. Without numerical experiments (relying only on the above estimates), it is difficult to determine which approach is preferable (basically, because R_y is unknown). This example is almost the only one where the computations can be organized so that the sparsity of the problem is used to a full extent (an iteration step requires $\mathrm{O}(s)$ arithmetic operations). Unfortunately, since x^N has to be computed (updated), at least $\mathrm{O}(n)$ arithmetic operations are usually executed per iteration step [13]. However, there is another (simpler) method for recovering the solution of the primal problem from the solution of the dual one, which is more suitable for taking into account sparsity. This method is described below.

Let us extend the approach described in this paper to the case where the dual functional in problem (11) is μ-strongly convex (concave) in the 2-norm. For this purpose, it is sufficient that the gradient of the functional in the primal problem have a uniformly bounded Lipschitz constant and the primal problem be solved in the entire space [8, 13] (i.e., there are no constraints other than $Ax = b$). Such an example was discussed in Remark 10. Let us use the restart technique (see, e.g., [5, 14, 25]), which, in this case, has the following form (see Theorem 2; note that the primal-duality of the FGM is not used in this estimate):

$$\frac{\mu}{2}\left\| \tilde{y}^{\bar{N}} - y_* \right\|_2^2 \le F\left(\tilde{y}^{\bar{N}}\right) - F(y_*) \le \frac{2L\left\| y^0 - y_* \right\|_2^2}{\bar{N}^2}.$$

Choosing

$$\bar{N} = \sqrt{\frac{8L}{\mu}},$$

we obtain

$$\left\|\tilde{y}^{\bar{N}} - y_*\right\|_2^2 \le \frac{1}{2}\left\|y^0 - y_*\right\|_2^2.$$

Using $\tilde{y}^{\bar{N}}$ as a starting point in the FGM, we again execute $\bar{N}$ iteration steps, etc. It is easy to see that, if the accuracy ε is desired for the function, then the number of such restarts in the FGM needs to be equal to

$$\left\lceil \log_2\left(\frac{\mu R^2}{\varepsilon}\right)\right\rceil$$

(here, $\lceil\cdot\rceil$ is the standard notation for the ceiling function; for example, $\lceil 0.2\rceil = 1$). Thus, the total number of iteration steps executed by the FGM can be estimated as

$$\sqrt{\frac{8L}{\mu}}\left\lceil \log_2\left(\frac{\mu R^2}{\varepsilon}\right)\right\rceil.$$

At first glance, it seems that we then do not control

$$\left\|Ax(y) - b\right\|_2 = \left\|\nabla F(y)\right\|_2.$$

Actually, for problems with a Lipschitz continuous gradient, $\left\|\nabla F(y)\right\|_2$ can always (rather than only for primal-dual methods) be controlled using inequality (5), namely,

$$\frac{1}{2L}\left\|\nabla F(y)\right\|_2^2 \le F(y) - F(y_*)$$

(if $F(y_*) \ne 0$, then, instead of the gradient $\nabla F(y)$, we can calculate a gradient mapping (see, e.g., [4, 5])). However, the problem is that this inequality is typically rather rough. Indeed, in the μ-strongly convex case,

$$\frac{1}{2L}\left\|\nabla F(y)\right\|_2^2 \le F(y) - F(y_*) \le \frac{1}{2\mu}\left\|\nabla F(y)\right\|_2^2.$$

If $L/\mu \gg 1$ (which is typical), then the use of inequality (5) may lead (and indeed leads [13]) to strongly overstated estimates. However, if along with (5), we take into account the geometric convergence rate of the FGM in view of strong convexity, then

$$\left\|Ax\left(y^N\right) - b\right\|_2 = \left\|\nabla F\left(y^N\right)\right\|_2 \le \sqrt{2L\left(F(y) - F(y_*)\right)} \le \sqrt{2L\mu R^2}\exp\left(-\frac{N}{2}\sqrt{\frac{\mu}{8L}}\right),$$

where $\{y^N\}$ denotes the sequence generated by the above-described FGM with restarts.

Let us describe a stopping criterion for the FGM with restarts. By the definition of $x(y)$, we have

$$g(x(y)) + \langle y, Ax(y) - b\rangle \le g(x_*),$$

whence

$$g(x(y)) - g(x_*) \le \left\|y\right\|_2\left\|Ax(y) - b\right\|_2.$$

Thus, the stopping criterion has the form

$$\left\|y^N\right\|_2\left\|Ax\left(y^N\right) - b\right\|_2 \le \varepsilon, \quad \left\|Ax\left(y^N\right) - b\right\|_2 \le \tilde{\varepsilon}. \tag{15}$$

Theorem 3. *Suppose that problem* (11) *is solved by passing to problem* (12) *with a* μ*-strongly convex functional in the* 2*-norm with the help of the formulas given above. Let the stopping criterion be given by* (15)*. Then the above-described FGM with restarts is guaranteed to terminate after at most*

$$\max\left\{\sqrt{\frac{8L}{\mu}}\left\lceil \log_2\left(\frac{2L\mu R^4}{\varepsilon^2}\right)\right\rceil, \sqrt{\frac{8L}{\mu}}\left\lceil \log_2\left(\frac{2L\mu R^2}{\tilde{\varepsilon}^2}\right)\right\rceil\right\}$$

iteration steps.

Remark 12. In fact, we have just described a rather general approach to solving a large number of problems via the transition to a smooth dual problem. It remains to be noted that the regularization $\mu \approx \varepsilon/R^2$ is artificially introduced into the dual problem if the latter is not strongly convex [25]. Accordingly, restarts with respect to μ arise, since R is a priori unknown [1, 4]. Actually, it is this approach that was proposed in [1] for solving an ELP problem. Theoretical estimates suggest that both methods have roughly the same running time (the present one is slightly better), but the experiments in [18] showed conclusively that the present method is faster than that from [1] by several orders of magnitude. The cause is that, in view of the restarts in μ, a prescribed number of iteration steps (no less) have to be executed (at every restart step), while the present method, first, does not require restarts (thus saving almost one order of magnitude) and, second, terminates according to a more flexible criterion (see Theorem 2), which admits fewer iteration steps than suggested by the estimate. Numerical experiments show that, due to this circumstance, several orders of magnitude are saved in the total running time of the new method.

Remark 13. Everything described above for the strongly convex case can be extended to arbitrary methods (not necessarily primal-dual), for example, to the conjugate gradient method or Newton's method and their modifications [17] (there is a general thesis due to A.S. Nemirovski that nearly any reasonable numerical method is either primal-dual or has a corresponding modification; the primal-duality of many important methods has been established in various works, but, to the best of our knowledge, this has not been done for the conjugate gradient and Newton methods). Specifically, stopping criterion (15) is a general technique for error control in the solution produced by an arbitrary method that simultaneously solves primal and dual problems. As was noted above, when the dual problem is only smooth (which is required for the validity of the presented arguments) but not strongly convex, available theoretical techniques for methods with stopping criterion (15) typically yield strongly overstated convergence rate estimates. This does not mean that these methods are poor, because, to the best of our knowledge, no accurate estimation methods are available at present. The problem of theoretical substantiation is solved by regularizing the dual problem (see Remark 12).

Remark 14. The above approach to obtaining FGM for strongly convex problems based on restarts in the distance from the current point to the solution has a serious disadvantage. At every restart step, the method has to execute a prescribed number of iterations, which is, as a rule, overestimated. An earlier termination is possible if there is a stopping criterion. However, $F(y_*)$ is usually not known. In practice, we can try to monitor the norm of the gradient (in the general case, the norm of the gradient mapping) and terminate the method when the square of this norm decreases by half. However, similar (sharp in order) estimates for the resulting convergence rate have not yet been proved for this approach. A way out of this situation is to replace the FGM with restarts by the FGM without restarts for strongly convex problems [5] or, even better, by the FGM without restarts for strongly convex problems with an adaptively selected Lipschitz constant for the gradient. Such a method (which was additionally continuous with respect to the strong convexity parameter) was described, for example, in [30]. If the strong convexity parameter is not known (see Remark 12), then, unfortunately, restarts cannot be avoided, but they are used only with respect to this parameter and the exit from the last restart (which is the most expensive and makes the basic contribution to the estimated total running time of the method) can be achieved (in contrast to the approach described above) by monitoring the smallness of the norm of the gradient (gradient mapping). Another method for improving the restart construction was proposed in [31].

ACKNOWLEDGMENTS

The authors are grateful to A.G. Biryukov, A.I. Golikov, Yu.E. Nesterov, A.S. Nemirovski, and P.I. Stetsyuk for their interest in this work.

Gasnikov and Dvurechensky's research presented in Section 2 was performed at the Institute for Information Transmission Problems of the Russian Academy of Sciences and was supported by the Russian Science Foundation, project no. 14-50-00150. Gasnikov's research presented in Section 3 was supported by the Russian Foundation for Basic Research (project no. 15-31-70001-mol_a_mos), while Dvurechensky's research in Section 3 was supported by a grant from the President of the Russian Federation (project no. MK-1806.2017.9).

REFERENCES

1. A. V. Gasnikov, E. V. Gasnikova, Yu. E. Nesterov, and A. V. Chernov, "Efficient numerical methods for entropy-linear programming problems," Comput. Math. Math. Phys. **56** (4), 514–524 (2016).
2. Y. Nesterov, "Primal-dual subgradient methods for convex problems," Math. Program. Ser. B **120** (1), 261–283 (2009).
3. A. Nemirovski, S. Onn, and U. G. Rothblum, "Accuracy certificates for computational problems with convex structure," Math. Operations Res. **35** (1), 52–78 (2010).
4. O. Devolder, PhD Thesis (CORE UCL, 2013).
5. Yu. E. Nesterov, Doctoral Dissertation in Mathematics and Physics (Moscow Inst. of Physics and Technology, Dolgoprudnyi, 2013). www.mathnet.ru/php/seminars.phtml?option_lang=rus&presentid=8313.
6. Yu. Nesterov, "New primal-dual subgradient methods for convex optimization problems with functional constraints," *International Workshop on Optimization and Statistical Learning, January 11–16, 2015* (France, Les Houches, 2015). http://lear.inrialpes.fr/workshop/osl2015/program.html.
7. Yu. Nesterov, "Complexity bounds for primal-dual methods minimizing the model of objective function," CORE Discussion Papers, 2015/03 (2015).
8. A. Anikin, P. Dvurechensky, A. Gasnikov, A. Golov, A. Gornov, Yu. Maximov, M. Mendel, and V. Spokoiny, "Modern efficient numerical approaches to regularized regression problems in application to traffic demands matrix calculation from link loads," *Proceedings of International Conference ITAS-2015, Sochi, Russia, September 2015* (2015). arXiv:1508.00858.
9. A. V. Gasnikov, E. V. Gasnikova, E. I. Ershov, P. E. Dvurechensky, and A. A. Lagunovskaya, "Search for stochastic equilibria in equilibrium traffic flow models," Tr. Mosk. Fiz.-Tekh. Inst. **7** (4), 114–128 (2015). arXiv:1505.07492.
10. A. V. Gasnikov, P. E. Dvurechensky, Yu. V. Dorn, and Yu. V. Maksimov, "Numerical methods of searching for equilibrium flow distributions in Beckmann's and stable dynamic models," Mat. Model. **28** (10), 40–64 (2016). arXiv:1506.00293.
11. A. V. Gasnikov, P. E. Dvurechensky, D. I. Kamzolov, Yu. E. Nesterov, V. G. Spokoinyi, P. I. Stetsyuk, A. L. Suvorikova, and A. V. Chernov, "Search for equilibria in multistage traffic flow models," Tr. Mosk. Fiz.-Tekh. Inst. **7** (4), 143–155 (2015). https://mipt.ru/upload/medialibrary/ffe/143-155.pdf.
12. A. V. Gasnikov, P. E. Dvurechensky, Yu. E. Nesterov, V. G. Spokoinyi, and A. L. Suvorikova, "Superposition of the balancing and universal gradient methods for searching for an entropy-smoothed Wasserstein barycenter and equilibria in multistage traffic flow models," Tr. Mosk. Fiz.-Tekh. Inst. **8** (3), 5–24 (2016). arXiv:1506.00292.
13. A. V. Gasnikov, P. E. Dvurechensky, and I. N. Usmanova, "On the nontriviality of fast (accelerated) randomized methods," Tr. Mosk. Fiz.-Tekh. Inst. **8** (2), 67–100 (2016). arXiv:1508.02182.
14. Z. Allen-Zhu and L. Orecchia, *Linear Coupling: An Ultimate Unification of Gradient and Mirror Descent*, e-print (2014). arXiv:1407.1537.
15. A. S. Nemirovski and D. B. Yudin, *Complexity of Problems and Efficiency of Optimization Methods* (Nauka, Moscow, 1979) [in Russian]. http://www2.isye.gatech.edu/~nemirovs/Lect_EMCO.pdf.
16. V. A. Zorich, *Mathematical Analysis of Problems in Natural Sciences* (MTsNMO, Moscow, 2008) [in Russian].
17. B. T. Polyak, *Introduction to Optimization* (Nauka, Moscow, 1983; Optimization Software, New York, 1987).
18. A. Chernov, P. Dvurechensky, and A. Gasnikov, "Fast primal-dual gradient method for strongly convex minimization problems with linear constraints," *International Conference on Discrete Optimization and Operations Research, Vladivostok, Russian Iceland, September 19–23, 2016* (Springer, Berlin, 2016), pp. 584–595. arXiv:1605.02970.
19. A. Nemirovski, *Lectures on Modern Convex Optimization Analysis, Algorithms, and Engineering Applications* (SIAM, Philadelphia, 2013). http://www2.isye.gatech.edu/~nemirovs/Lect_ModConvOpt.pdf.
20. Yu. Nesterov, "Gradient methods for minimizing composite functions," Math. Prog. **140** (1), 125–161 (2013).
21. Yu. Nesterov, "Universal gradient methods for convex optimization problems," CORE Discussion Paper, 2013/63 (2013).
22. A. V. Gasnikov and P. E. Dvurechensky, "Stochastic intermediate gradient method for convex optimization problems," Dokl. Math. **93** (2), 148–151 (2016).
23. P. Dvurechensky and A. Gasnikov, "Stochastic intermediate gradient method for convex problems with inexact stochastic oracle," J. Optim. Theory Appl., 1–25 (2016). arXiv:1411.2876.
24. A. V. Gasnikov, P. E. Dvurechensky, and Yu. E. Nesterov, "Stochastic gradient methods with an inexact oracle," Tr. Mosk. Fiz.-Tekh. Inst. **8** (1), 41–91 (2016). arxiv:1411.4218
25. A. V. Gasnikov, D. I. Kamzolov, and M. A. Mendel', "Basic constructions over convex optimization algorithms and their applications to deriving new estimates for strongly convex problems," Tr. Mosk. Fiz.-Tekh. Inst. **8** (3), 25–42 (2016). arXiv:1603.07701.

26. A. Pătraşcu, PhD Thesis (University Politehnica of Bucharest, Bucharest, 2015). http://acse.pub.ro/person/ion-necoara/.
27. B. Cox, A. Juditsky, and A. Nemirovski, *Decomposition Techniques for Bilinear Saddle Point Problems and Variational Inequalities with Affine Monotone Operators on Domains Given by Linear Minimization Oracles*, e-print (2015). arXiv:1506.02444.
28. A. V. Gasnikov and D. Yu. Dmitriev, "On efficient randomized algorithms for finding the PageRank vector," Comput. Math. Math. Phys. **55** (3), 349–365 (2015). arXiv:1410.3120.
29. H. Nikaido, *Convex Structures and Economic Theory* (Academic, New York, 1968; Mir, Moscow, 1972).
30. A. V. Gasnikov and Yu. E. Nesterov, "Universal method for stochastic composite optimization problems," arXiv:1604.05275.
31. B. O'Donoghue and E. Candes, "Adaptive restart for accelerated gradient schemes," Foundations Comput. Math. **15**, 715–732 (2015).

Translated by I. Ruzanova