

Primal–Dual Mirror Descent Method for Constraint Stochastic Optimization Problems

A. S. Bayandina^{a,*}, A. V. Gasnikov^{b,c,f,**}, E. V. Gasnikova^{d,***}, and S. V. Matsievskii^{e,****}

^a Department of Control and Applied Mathematics, Moscow Institute of Physics and Technology, Dolgoprudnyi, Moscow oblast, 141700 Russia

^b Chair of Mathematical Foundations of Control, Moscow Institute of Physics and Technology, Dolgoprudnyi, Moscow oblast, 141700 Russia

^c Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, 127051 Russia

^d Laboratory of Structural Analysis Methods in Predictive Simulation, Moscow Institute of Physics and Technology, Dolgoprudnyi, Moscow oblast, 141700 Russia

^e Kant Baltic Federal University, Kaliningrad, 236016 Russia

^f Adygeya State University, Maykop, 352700 Russia

*e-mail: anast.bayandina@gmail.com

**e-mail: gasnikov@yandex.ru

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

****e-mail: matsievsky@newmail.ru

Received December 9, 2016; in final form, September 29, 2017

Abstract—Extension of the mirror descent method developed for convex stochastic optimization problems to constrained convex stochastic optimization problems (subject to functional inequality constraints) is studied. A method that performs an ordinary mirror descent step if the constraints are insignificantly violated and performs a mirror descent step with respect to the violated constraint if this constraint is significantly violated is proposed. If the method parameters are chosen appropriately, a bound on the convergence rate (that is optimal for the given class of problems) is obtained and sharp bounds on the probability of large deviations are proved. For the deterministic case, the primal–dual property of the proposed method is proved. In other words, it is proved that, given the sequence of points (vectors) generated by the method, the solution of the dual method can be reconstructed up to the same accuracy with which the primal problem is solved. The efficiency of the method as applied for problems subject to a huge number of constraints is discussed. Note that the bound on the duality gap obtained in this paper does not include the unknown size of the solution to the dual problem.

Keywords: Mirror descent method, convex stochastic optimization, constrained optimization, probability of large deviations, randomization.

DOI: 10.1134/S0965542518110039

1. INTRODUCTION

In [1], a theory of lower (oracle) bounds on the complexity of solving constrained and stochastic convex optimization problems on sets of a simple structure was developed. In [1], methods that converge according to these lower bounds (up to logarithmic factors) were also proposed. In particular, a class of nonsmooth (stochastic) convex optimization was considered. For this class, a special method called the mirror descent method was proposed, which is optimal for this class of problems. It was known that the mirror descent method can be extended to constrained optimization problems (with the loss of a logarithmic factor compared with lower bounds). The subsequent development of numerical methods for convex optimization showed that the mirror descent method is in many respects convenient for various problems, including huge-scale optimization problems (problems in huge-dimensional spaces or with a huge number of constraints). In particular, such problems arise in truss topology design (see Example 1 in Section 4). There are various simplifications and generalizations of this method. In this paper, we follow [2] to extend the modern version of the mirror descent method to constrained stochastic optimization problems. In distinction from [1, 2], we obtain sharp bounds on the probability of large deviations and prove the primal–dual property of the method in the deterministic case. The bounds on the convergence rate of the

method correspond to the lower bounds obtained in [1]; i.e., the logarithmic gap mentioned above is eliminated. An important feature of the proposed method is its simplicity, which helps use the sparseness of the problem.

In Section 2, we describe the method and prove a theorem establishing a bound on the convergence rate of the method. Note that we consider the oracle that produces not only stochastic (sub)gradients of the objective functional and functional constraint but also the value (but not realization) of the functional constraint at the point of interest. Using the theory of fast automatic differentiation (see [3]), we may conclude that such an oracle can also produce the gradient of the functional constraint (rather than the stochastic gradient). However, the fast automatic differentiation technique is developed for smooth problems (while lexicographic differentiation is used for nonsmooth problems [4]) and, secondly, certain applications require the use of the sparse stochastic (sub)gradient rather than the nonsparse full (sub)gradient. Some examples of such problems are discussed in Section 4.

In Section 3, we prove (in the deterministic case) the primal–dual property of the method. This property is useful, for instance, in the application of the proposed method to truss topology design, in which the primal and the dual problems must be solved simultaneously.

In the concluding Section 4, we discuss possible generalizations and applications. A more detailed comparative analysis with other available results is also performed.

2. MIRROR DESCENT METHOD FOR CONVEX CONSTRAINED OPTIMIZATION PROBLEMS

Consider the convex constrained optimization problem

$$f(x) \rightarrow \min_{g(x) \leq 0, x \in Q}. \quad (1)$$

By an $(\varepsilon_f, \varepsilon_g, \sigma)$ -solution of this problem, we mean an $\bar{x}^N \in Q \subseteq \mathbb{R}^n$ such that the inequality

$$f(\bar{x}^N) - f_* \leq \varepsilon_f = \frac{M_f}{M_g} \varepsilon_g, \quad g(\bar{x}^N) \leq \varepsilon_g, \quad (2)$$

holds with a probability $\geq 1 - \sigma$, where $f_* = f(x_*)$ is the optimal value of the objective functional in problem (1) and x_* is the solution of problem (1).

Define a norm $\| \cdot \|$ in the primal space (the adjoint norm will be denoted by $\| \cdot \|_*$) and a proximity function $d(x)$ that is strongly convex with respect to this norm with a strong convexity constant ≥ 1 . Choose an initial point

$$x^1 = \arg \min_{x \in Q} d(x),$$

where we assume that

$$d(x^1) = 0, \quad \nabla d(x^1) = 0.$$

Define the Bregman “distance”

$$V_x(y) = d(y) - d(x) - \langle \nabla d(x), y - x \rangle.$$

The “size” of the solution is defined by

$$d(x_*) = V_{x^1}(x_*) = R^2,$$

and the size of the set Q (for clarity, we assume that this set is bounded, but in the general case the reasoning below can be conducted accurately under certain additional assumptions, see [5]) is defined by

$$\max_{x, y \in Q} V_x(y) = \bar{R}^2.$$

We assume that there is a sequence of independent random variables $\{\xi^k\}$ and sequences $\{\nabla_x f(x, \xi^k)\}$, $\{\nabla_x g(x, \xi^k)\}$ ($k = 1, \dots, N$) such that the following relations hold:

$$E_{\xi^k} [\nabla_x f(x, \xi^k)] = \nabla f(x), \quad E_{\xi^k} [\nabla_x g(x, \xi^k)] = \nabla g(x); \quad (3)$$

$$\|\nabla_x f(x, \xi^k)\|_*^2 \leq M_f^2, \quad \|\nabla_x g(x, \xi^k)\|_*^2 \leq M_g^2 \quad (4)$$

or

$$E_{\xi^k} \left[\|\nabla_x f(x, \xi^k)\|_*^2 \right] \leq M_f^2, \quad E_{\xi^k} \left[\|\nabla_x g(x, \xi^k)\|_*^2 \right] \leq M_g^2. \quad (4')$$

At each iteration step $k = 1, \dots, N$, we have a stochastic (sub)gradient $\nabla_x f(x, \xi^k)$ or $\nabla_x g(x, \xi^k)$ at a point x^k chosen by the method.

Let us describe the stochastic version of the mirror descent method for problems with functional constraints (this method dates to [1]).

Define the “projection” operator associated with the Bregman distance by

$$\text{Mirr}_{x^k}(v) = \arg \min_{y \in Q} \left\{ \langle v, y - x^k \rangle + V_{x^k}(y) \right\}.$$

The mirror descent method for problem (1) has the form (e.g., see [2])

$$\begin{aligned} x^{k+1} &= \text{Mirr}_{x^k} \left(h_f \nabla_x f(x^k, \xi^k) \right) & \text{if } g(x^k) \leq \varepsilon_g, \\ x^{k+1} &= \text{Mirr}_{x^k} \left(h_g \nabla_x g(x^k, \xi^k) \right) & \text{if } g(x^k) > \varepsilon_g, \end{aligned} \quad (5)$$

where $h_g = \varepsilon_g / M_g^2$, $h_f = \varepsilon_g / (M_f M_g)$, $k = 1, \dots, N$. Denote by I the set of indexes k for which $g(x^k) \leq \varepsilon_g$. We also introduce the notation

$$[N] = \{1, \dots, N\}, \quad J = [N] \setminus I, \quad N_I = |I|, \quad N_J = |J|, \quad \bar{x}^N = \frac{1}{N_I} \sum_{k \in I} x^k.$$

In the theorems formulated below, it is assumed that the sequence $\{x^k\}_{k=1}^{N+1}$ is generated by method (5).

Theorem 1. *Let conditions (3) and (4') hold. Then, for*

$$N > \frac{2M_g^2 R^2}{\varepsilon_g^2} \stackrel{\text{def}}{=} N(\varepsilon_g)$$

it holds that $N_I \geq 1$ (with a probability $\geq 1/2$) and

$$E \left[f(\bar{x}^N) \right] - f_* \leq \varepsilon_f, \quad g(\bar{x}^N) \leq \varepsilon_g.$$

Let conditions (3) and (4) hold. Then, for

$$N \geq \frac{81M_g^2 \bar{R}^2}{\varepsilon_g^2} \ln \left(\frac{1}{\sigma} \right) \quad (6)$$

it holds that $N_I \geq 1$ and

$$f(\bar{x}^N) - f_* \leq \varepsilon_f, \quad g(\bar{x}^N) \leq \varepsilon_g$$

with a probability $\geq 1 - \sigma$; i.e., inequalities (2) are satisfied.

Proof. The first part of this theorem was proved in [2]. Here we prove the second part. According to [6], for every $x \in Q$ such that $g(x) \leq 0$, it holds that

$$\begin{aligned} h_f N_I (f(\bar{x}^N) - f(x)) &\leq h_f \sum_{k \in I} \left\langle E_{\xi^k} \left[\nabla_x f(x^k, \xi^k) \right], x^k - x \right\rangle \leq \frac{h_f^2}{2} \sum_{k \in I} \|\nabla_x f(x^k, \xi^k)\|_*^2 \\ &+ h_f \sum_{k \in I} \left\langle E_{\xi^k} \left[\nabla_x f(x^k, \xi^k) \right] - \nabla_x f(x^k, \xi^k), x^k - x \right\rangle + h_g \sum_{k \in J} \underbrace{\left\langle E_{\xi^k} \left[\nabla_x g(x^k, \xi^k) \right], x^k - x \right\rangle}_{\geq g(x^k) - g(x) > \varepsilon_g} \\ &+ \frac{h_g^2}{2} \sum_{k \in J} \|\nabla_x g(x^k, \xi^k)\|_*^2 + h_g \sum_{k \in J} \left\langle E_{\xi^k} \left[\nabla_x g(x^k, \xi^k) \right] - \nabla_x g(x^k, \xi^k), x^k - x \right\rangle + \sum_{k \in [N]} (V_{x^k}(x) - V_{x^{k+1}}(x)). \end{aligned}$$

Put $x = x_*$ and define

$$\delta_N = h_f \sum_{k \in I} \langle \nabla f(x^k) - \nabla_x f(x^k, \xi^k), x^k - x_* \rangle + h_g \sum_{k \in J} \langle \nabla g(x^k) - \nabla_x g(x^k, \xi^k), x^k - x_* \rangle.$$

Then

$$\begin{aligned} h_f N_I (f(\bar{x}^N) - f(x_*)) &\leq \frac{1}{2} h_f^2 M_f^2 N_I - \frac{1}{2M_g^2} \varepsilon_g^2 N_J + V_{x^1}(x_*) - V_{x^{N+1}}(x_*) + \delta_N \\ &= \frac{1}{2} \left(h_f^2 M_f^2 + \frac{\varepsilon_g^2}{M_g^2} \right) N_I - \frac{1}{2M_g^2} \varepsilon_g^2 N + R^2 - V_{x^{N+1}}(x_*) + \delta_N \\ &= \varepsilon_f h_f N_I - \frac{1}{2M_g^2} \varepsilon_g^2 N + R^2 - V_{x^{N+1}}(x_*) + \delta_N \leq \varepsilon_f h_f N_I + \left(R^2 + \delta_N - \frac{1}{2M_g^2} \varepsilon_g^2 N \right). \end{aligned} \tag{7}$$

The Azuma–Hoeffding inequality [7] implies that

$$P\left(\delta_N \geq 2\sqrt{2}\bar{R}\Lambda\sqrt{h_f^2 M_f^2 N_I + h_g^2 M_g^2 N_J}\right) \leq \exp(-\Lambda^2/2);$$

i.e.,

$$P\left(\delta_N \geq \frac{4\bar{R}\varepsilon_g}{M_g} \sqrt{N \ln\left(\frac{1}{\sigma}\right)}\right) \leq \sigma$$

with a probability $\geq 1 - \sigma$. We assume that (the constant 81 can be decreased to $(4 + \sqrt{18})^2$):

$$N \geq \frac{81M_g^2 \bar{R}^2}{\varepsilon_g^2} \ln\left(\frac{1}{\sigma}\right).$$

Then, with a probability $\geq 1 - \sigma$, the expression in parentheses in (7) is strictly less than zero; therefore, we have the inequalities

$$f(\bar{x}^N) - f_* \leq \varepsilon_f, \quad g(\bar{x}^N) \leq \varepsilon_g.$$

The last inequality follows from the fact that $g(x^k) \leq \varepsilon_g$ for $k \in I$ and from the convexity of the function $g(x)$.

3. THE PROMAL–DUAL PROPERTY OF THE METHOD

Let $g(x) = \max_{l=1, \dots, m} g_l(x)$. Consider the dual problem

$$\varphi(\lambda) = \min_{x \in Q} \left\{ f(x) + \sum_{l=1}^m \lambda_l g_l(x) \right\} \rightarrow \max_{\lambda \geq 0}. \tag{8}$$

We have the inequality (weak duality)

$$0 \leq f(x) - \varphi(\lambda) \stackrel{\text{def}}{=} \Delta(x, \lambda), \quad x \in Q, \quad g(x) \leq 0, \quad \lambda \geq 0.$$

Denote the solution to problem (8) by λ_* . Assume that Slater’s conditions hold, i.e., there exists an $\tilde{x} \in Q$ such that $g(\tilde{x}) < 0$. Then

$$f_* = f(x_*) = \varphi(\lambda_*) \stackrel{\text{def}}{=} \varphi_*.$$

In this case, the “quality” of the pair (x^N, λ^N) is naturally assessed by the size of the duality gap $\Delta(x^N, \lambda^N)$. The less the gap, the higher is the quality.

Let (we restrict ourselves to the deterministic case)

$$g(x^k) = g_{l(k)}(x^k), \quad \nabla g(x^k) = \nabla g_{l(k)}(x^k), \quad k \in J.$$

Set

$$\lambda_l^N = \frac{1}{h_f N_I} \sum_{k \in J} h_g I[l(k) = l],$$

$$I[\text{predicate}] = \begin{cases} 1, & \text{predicate} = \text{true}, \\ 0, & \text{predicate} = \text{false}. \end{cases}$$

Theorem 2. *Let*

$$\|\nabla f(x)\|_*^2 \leq M_f^2, \quad \|\nabla g(x)\|_*^2 \leq M_g^2.$$

Then, for

$$N \geq \frac{2M_g^2 \bar{R}^2}{\varepsilon_g^2} + 1,$$

it holds that $N_I \geq 1$ *and*

$$\Delta(\bar{x}^N, \lambda^N) \leq \varepsilon_f, \quad g(\bar{x}^N) \leq \varepsilon_g.$$

Proof. According to [6], we have

$$\begin{aligned} h_f N_I f(\bar{x}^N) &\leq \min_{x \in Q} \left\{ h_f N_I f(x) + h_f \sum_{k \in I} \langle \nabla f(x^k), x^k - x \rangle \right\} \leq \min_{x \in Q} \left\{ h_f N_I f(x) + \frac{h_f^2}{2} \sum_{k \in I} \|\nabla f(x^k)\|_*^2 \right. \\ &- h_g \sum_{k \in J} \underbrace{\langle \nabla g(x^k), x^k - x \rangle}_{\geq g_{l(k)}(x^k) - g_{l(k)}(x)} + \frac{h_g^2}{2} \sum_{k \in J} \|\nabla g(x^k)\|_*^2 + \sum_{k \in [N]} (V_{x^k}(x) - V_{x^{k+1}}(x)) \left. \right\} \leq \frac{1}{2} h_f^2 M_f^2 N_I - \frac{1}{2M_g^2} \varepsilon_g^2 N_J + \bar{R}^2 \\ &+ h_f N_I \min_{x \in Q} \left\{ f(x) + \sum_{l=1}^m \lambda_l^N g_l(x) \right\} = \varepsilon_f h_f N_I + \left(\bar{R}^2 - \frac{1}{2M_g^2} \varepsilon_g^2 N \right) + h_f N_I \varphi(\lambda^N). \end{aligned}$$

The subsequent reasoning repeats the reasoning in the proof of Theorem 1 (see formula (7) and the text following it).

4. CONCLUDING REMARKS

In Remarks 1 and 2, the results obtained in Sections 2 and 3 are compared with other known results.

Remark 1. The results of Theorems 1 and 2 can be found in [8, 9] for the deterministic case; however, they were proved for other methods that are close to (5) but still different from it. As in [8, 9] the main advantage of method (5) is that the bounds on its convergence rate do not include the size of the dual solution, which is involved in the bounds of other primal–dual methods and approaches (e.g., see [10]).

Remark 2. In [11, 12], another method for deriving results close to those obtained in this paper in the deterministic case is proposed. The approach used in those papers is based on the ellipsoid method instead of the mirror descent method (5). Note that in Section 5 of [11], it is shown how the violation of the constraint $g(\bar{x}^N) \leq \varepsilon_g$ can be avoided. In the approach described in that paper, which follows the series of works [2, 8, 9], the constraint was perturbed to ensure proper bounds on the rate of the duality gap decrease. In [11, 12], the ellipsoid method was used that guaranteed the desired bounds on the convergence rate of the accuracy certificate, which majorizes the duality gap, without constraint relaxation.

The results of Sections 2 and 3 can be further elaborated. This will be briefly described in Remarks 3–7. A more detailed presentation will be made in a future paper.

Remark 3. Using the constructs described in [13] (also see [14]), the results obtained above can be extended for the case of small nonrandom noise.

Remark 4. The method described in Section 3 can be extended for the case of arbitrary ε_g and ε_f not satisfying the relation $\varepsilon_f = M_f \varepsilon_g / M_g$. In Sections 2 and 3, this relation was assumed only to simplify the calculations.

Remark 5. Similarly to [5, 13, 15], we can propose an adaptive version of the method described in Section 3 that does not require the bounds M_f and M_g to be known a priori. Furthermore, in the case of a bounded set Q , an adaptive version of the method described in Section 2 for stochastic optimization problems can be proposed (see [15]), as well as the corresponding generalization of the method AdaGrad [15]. Moreover, using the results obtained in [16] and a special choice of steps in the adaptive method, one can obtain (in the deterministic case) sharper bounds on the convergence rate that admit, e.g., the unboundedness of the Lipschitz constant of the functional on an unbounded set Q .

Remark 6. The method described in Section 3 can be extended for composite optimization problems [17] in the case when the function and the functional have a common composite.

Remark 7. Using restarts as in [18], the method described in Section 3 can be extended to strongly convex problem statements (when both the functional and the constraints are strongly convex). The key observation is as follows. If $f(x)$ and $g(x)$ are μ -strongly convex functions with respect to the norm $\| \cdot \|$ on the convex set Q and $x_* = \arg \min_{x \in Q, g(x) \leq 0} f(x)$ for $x \in Q$, then $f(x) - f(x_*) \leq \epsilon_f$ and $g(x) \leq \epsilon_g$ imply that

$$\frac{\mu}{2} \|x - x_*\|^2 \leq \max\{\epsilon_f, \epsilon_g\}.$$

Examples 1 and 2 discussed below demonstrate the possible fields of application of the proposed version of the mirror descent method. Pay attention to how the number of constraints m appears in these bounds. In the sparse case, formulas (9) and (10) seem to be very optimistic.

Example 1. The main field of application of the proposed approach is convex problems of the form (see [2, 8])

$$f(c^T x) \rightarrow \min_{\substack{\max_{k=1, \dots, m} \sigma_k(A_k^T x) \leq 0, \\ x \geq 0}},$$

$m \gg 1$, where $f(\cdot)$ and $\sigma_k(\cdot)$ are convex functions (of a scalar argument) with the Lipschitz constant uniformly bounded by a known number M and the (sub)gradient of each such function can be computed in an amount of time $O(1)$. As applied to truss topology design, these functions may be assumed to be linear [8]. Define the matrix

$$A = [A_1, \dots, A_m]^T,$$

and assume that each column of the matrix A contains not more than $s_m \ll m$ nonzero elements and each row contains not more than $s_n \ll n$ such elements (the vector c has no more than s_n nonzero elements as well). The results obtained in Sections 2 and 3 imply that the proposed version of the mirror descent method (with the choice of $\| \cdot \| = \| \cdot \|_2$, $d(x) = \|x\|_2^2/2$) requires (Theorem 2)

$$O\left(\frac{M^2 \max\left\{\max_{k=1, \dots, m} \|A_k\|_2^2, \|c\|_2^2\right\} R_2^2}{\epsilon^2}\right)$$

iteration steps, where R_2^2 is the Euclidean distance from the start point to the solution squared and each iteration step (except for the first one) requires (see [2, 19] and the reasoning in Example 2 below)

$$O(s_n s_m \log_2 m)$$

operations. This requires $O(m + n)$ operations for additional preprocessing (to prepare the memory in a proper way). Thus, the total number of arithmetic operations is

$$O\left(s_n s_m \log_2 m \frac{M^2 \max\left\{\max_{k=1, \dots, m} \|A_k\|_2^2, \|c\|_2^2\right\} R_2^2}{\epsilon^2}\right). \tag{9}$$

Example 2. Assume that the matrix A and the vector c in Example 1 are not sparse. We try to introduce randomization into the approach described in Example 1. To this end, we perform some additional preprocessing to construct a probability distribution vector from nonsparse vectors A_k . Represent these vectors by

$$A_k = A_k^+ - A_k^-,$$

where each vector A_k^+ and A_k^- has nonnegative components. According to this representation, we prepare the memory in such a way that the time needed to generate random variables based on the probability distributions $A_k^+ / \|A_k^+\|_1$ and $A_k^- / \|A_k^-\|_1$ takes a time $O(\log_2 n)$. This can always be done as was proved in [19]. However, this requires a fairly large number of the corresponding “trees” to be stored in fast memory. The time taken by the preprocessing procedure and the amount of required memory are proportional to the number of nonzero elements in the matrix A , which is too large in the case of huge-scale problems. Nevertheless, we below assume that such a preprocessing can be performed and (which is the main thing) such an amount of memory is available. In practice the preprocessing is often needed only for a small number of constraints and the problem functional, so that the required amount of memory is available. Define the stochastic (sub)gradient (the same can be made for the functional)

$$\nabla_x g(x, \xi^k) = \left(\|A_{k(x)}^+\|_1 e_{i(\xi^k)} - \|A_{k(x)}^-\|_1 e_{j(\xi^k)} \right) \sigma_k'(A_{k(x)}^T x),$$

where

$$k(x) \in \text{Arg max}_{k=1, \dots, m} \sigma_k(A_k^T x);$$

moreover, it is of no importance which representative of Arg max is chosen;

$$e_i = \underbrace{(0, \dots, 0, 1, 0, \dots, 0)}_i;$$

$$i(\xi^k) = i \quad \text{with the probability} \quad A_{k(x)i}^+ / \|A_{k(x)}^+\|_1, \quad i = 1, \dots, n;$$

$$j(\xi^k) = j \quad \text{with the probability} \quad A_{k(x)j}^- / \|A_{k(x)}^-\|_1, \quad j = 1, \dots, n.$$

The results of Sections 2 and 3 imply that the proposed version of the mirror descent method (with the choice of $\| \cdot \| = \| \cdot \|_2$, $d(x) = \|x\|_2^2/2$) requires (here, for clearness, we restrict ourselves to the convergence in expectation in Theorem 1, i.e. without bounds on the probabilities of large deviations)

$$O \left(\frac{M^2 \max \left\{ \max_{k=1, \dots, m} \|A_k\|_1^2, \|c\|_1^2 \right\} R_2^2}{\epsilon^2} \right)$$

iteration steps. The main computational complexity is in computing $k(x)$. However, except for the first iteration step, the repeated solution of this problem can be organized efficiently. Indeed, assume that $k(x^l)$ has already been calculated and we want to calculate $k(x^{l+1})$. Since x^{l+1} can differ from x^l only in two components (see [2]), $\max_{k=1, \dots, m} \sigma_k(A_k^T x^{l+1})$ can be recalculated in time $O(s_m \log_2 m)$ based on the known $\max_{k=1, \dots, m} \sigma_k(A_k^T x^l)$ (e.g., see [19]). Thus, the expected total number of arithmetic operations in the randomized mirror descent method is

$$O \left(s_m \log_2 m \frac{M^2 \max \left\{ \max_{k=1, \dots, m} \|A_k\|_1^2, \|c\|_1^2 \right\} R_2^2}{\epsilon^2} \right). \tag{10}$$

For the matrices A and the vector c all nonzero elements of which have the same order of magnitude, e.g., $O(1)$, we have

$$\max_{k=1,\dots,m} \|A_k\|_2^2 \approx s_n, \quad \max_{k=1,\dots,m} \|A_k\|_1^2 \approx s_n^2, \quad \|c\|_2^2 \approx s_n, \quad \|c\|_1^2 \approx s_n^2.$$

In this case, no advantages can be expected because formulas (9) and (10) will be similar. However, if this condition (that the nonzero elements of A and c have the same order of magnitude) holds not very accurately, certain advantages can be expected.

Numerical experiments confirm the estimates presented in these examples and in [2]. A more detailed description of the numerical experiments can be found in [20].

Already after this paper has been prepared for publication, we got to know the works [21–23], in which similar results were obtained. In connection with this, we note a different (simpler) technique used in this paper to obtain the main results and the treatment of the primal–dual property of the proposed method in the deterministic case.

ACKNOWLEDGMENTS

We are grateful to Yu.E. Nesterov and A.S. Nemirovski for discussions of parts of this paper. We are also grateful to the reviewer for valuable remarks.

The work by A.V. Gasnikov was performed in the Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences and supported by the Russian Science Foundation (project no. 14-50-00150). The work by E.V. Gasnikova was supported by the Russian Foundation for Basic Research, project no. 15-31-20571-mol_a_ved.

REFERENCES

1. A. S. Nemirovski and D. B. Yudin, *Problem Complexity and Method Efficiency in Optimization, Interscience Series in Discrete Mathematics* (Nauka, Moscow, 1979; Wiley, 1983), Vol. XV.
2. A. S. Anikin, A. V. Gasnikov, and A. Yu. Gornov, “Randomization and sparseness in huge-scale optimization problems using the mirror descent method as an example,” *Trudy Mosk. Fiz.-Tekhn. Inst.* **8** (1), 11–24 (2016). arXiv:1602.00594
3. K. Kim, Yu. Nesterov, V. Skokov, and B. Cherkasskii, “Efficient differentiation algorithms and extreme problems,” *Ekon. Mat. Metody* **20**, 309–318 (1984).
4. Yu. Nesterov, “Lexicographic differentiation of nonsmooth functions,” *Math. Program.* **104**, 669–700 (2005).
5. A. V. Gasnikov, P. E. Dvurechensky, Yu. V. Dorn, and Yu. V. Maksimov, “Numerical methods for the problem of traffic flow equilibrium in the Backmann and the stable dynamics models,” *Mat. Model.* **28** (10), 40–64 (2016). arXiv:1506.00293
6. A. Juditsky, G. Lan, A. Nemirovski, and A. Shapiro, “Stochastic approximation approach to stochastic programming,” *SIAM J. Optim.* **19**, 1574–1609 (2009).
7. S. Boucheron, G. Lugoshi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence* (Oxford Univ. Press, 2013).
8. Yu. Nesterov and S. Shpirko, “Primal-dual subgradient method for huge-scale linear conic problem,” *SIAM J. Optim.* **24**, 1444–1457 (2014). http://www.optimization-online.org/DB_FILE/2012/08/3590.pdf
9. Yu. Nesterov, “New primal-dual subgradient methods for convex optimization problems with functional constraints,” *Int. Workshop “Optimization and Statistical Learning”*, Les Houches, France, 2015. <http://lear.inria-alpes.fr/workshop/osl2015/program.html>
10. A. S. Anikin, A. V. Gasnikov, P. E. Dvurechensky, 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**, 1262–1275 (2017). arXiv:1602.01686
11. A. Nemirovski, S. Onn, and U. G. Rothblum, “Accuracy certificates for computational problems with convex structure,” *Math. Oper. Res.* **35** (1), 52–78 (2010).
12. 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,” 2015. arXiv:1506.02444

13. A. Juditsky and A. Nemirovski, “First order methods for nonsmooth convex large-scale optimization, I, II,” in *Optimization for Machine Learning*, ed. by S. Sra, S. Nowozin, and S. Wright (MIT Press, 2012).
14. A. V. Gasnikov, E. A. Krymova, A. A. Lagunovskaya, I. N. Usmanova, and F. A. Fedorenko, “Stochastic online optimization. Single-point and multi-point non-linear multi-armed bandits. Convex and strongly-convex case,” *Autom. Remote Control* **78**, 224–234 (2017). arXiv:1509.01679
15. J. C. Duchi, *Introductory Lectures on Stochastic Optimization, IAS/Park City Mathematics Series*.(2016), pp. 1–84. <http://stanford.edu/~jduchi/PCMICConvex/Duchi16.pdf>
16. Yu. Nesterov, “Subgradient methods for convex function with nonstandart growth properties,” 2016. http://www.mathnet.ru:8080/PresentFiles/16179/growthbm_nesterov.pdf
17. J. C. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari, “Composite objective mirror descent,” *Proc. of COLT*, 2010, pp. 14–26.
18. A. Juditsky and Yu. Nesterov, “Deterministic and stochastic primal-dual subgradient algorithms for uniformly convex minimization,” *Stoch. Syst.* **4** (1), 44–80 (2014).
19. A. S. Anikin, A. V. Gasnikov, A. Yu. Gornov, D. I. Kamzolov, Yu. V. Maksimov, and Yu. E. Nesterov, “Efficient Numerical Solution of the PageRank problem for doubly sparse matrices,” *Trudy Mosk. Fiz.-Tekhn. Inst.*, **7** (4), 74–94 (2015). arXiv:1508.07607
20. <https://github.com/anastasiabayandina/Mirror>
21. A. Beck, A. Ben-Tal, N. Guttman-Beck, and L. Tretuashvili, “The CoMirror algorithm for solving nonsmooth constrained convex problems,” *Oper. Res. Letts.* **38**, 493–498 (2011).
22. A. Juditsky, A. Nemirovski, and C. Tauvel, “Solving variational inequalities with stochastic mirror-prox algorithm,” *Stochastic Syst.* **1** (1), 17–58 (2011).
23. G. Lan and Z. Zhou, “Algorithms for stochastic optimization with expectation constraints,” 2016. <http://pwp.gatech.edu/guanghui-lan/wp-content/uploads/sites/330/2016/08/SPCS8-19-16.pdf>

Translated by A. Klimontovich