ISSN 1064-5624, Doklady Mathematics, 2016, Vol. 93, No. 2, pp. 148–151. © Pleiades Publishing, Ltd., 2016.
Original Russian Text © A.V. Gasnikov, P.E. Dvurechensky, 2016, published in Doklady Akademii Nauk, 2016, Vol. 467, No. 2, pp. 131–134.

MATHEMATICS

Stochastic Intermediate Gradient Method for Convex Optimization Problems

A. V. Gasnikov and P. E. Dvurechensky

Presented by Academician of the RAS A.P. Kuleshov October 16, 2015

Received October 16, 2015

Abstract—New first-order methods are introduced for solving convex optimization problems from a fairly broad class. For composite optimization problems with an inexact stochastic oracle, a stochastic intermediate gradient method is proposed that allows using an arbitrary norm in the space of variables and a prox-function. The mean rate of convergence of this method and the probability of large deviations from this rate are estimated. For problems with a strongly convex objective function, a modification of this method is proposed and its rate of convergence is estimated. The resulting estimates coincide, up to a multiplicative constant, with lower complexity bounds for the class of composite optimization problems with an inexact stochastic oracle and for all usually considered subclasses of this class.

DOI: 10.1134/S1064562416020071

First-order methods were among the first to be developed in optimization theory; their descriptions can be found in classical books, such as [1–3]. After publishing [4], special attention was given to convergence rate estimates for developed methods and to lower complexity bounds for various classes of problems (see also [5]). Later, more efficient methods, such as the ellipsoid method and interior point methods, were developed for convex optimization problems. These methods have a high rate of convergence, but the number of arithmetic operations required at every iteration step is on the order of $n^3 - n^4$ [5], which makes them inefficient for large-scale problems (with $n > 10^5$). Over the last decade, large-scale optimization problems have attracted much interest motivated by numerous applications, such as transportation modeling, web page ranking, and the design of mechanical structures. In such problems, the solution is usually not required to be highly accurate. As a result, they can be effectively solved by applying first-order methods, for which the estimated number of iterations required for finding a solution with prescribed accuracy is usually nearly independent of the dimension of the problem and the number of arithmetic operations required at every iteration step is on the order of n^2 or lower. Accordingly, an important issue is to develop new efficient first-order methods.

Institute for Information Transmission Problems, Russian Academy of Sciences, Bol'shoi Karetnyi per. 19/1, Moscow, 127994 Russia
e-mail: gasnikov@yandex.ru, pavel.dvurechensky@gmail.com

Now, we describe the formulation of the problem. Let E be a finite-dimensional vector space and E^* be its adjoint. The value of a linear functional $g \in E^*$ at a point $x \in E$ is denoted by $\langle g, x\rangle$. Let E be equipped with some norm $\|\cdot\|$. In this paper, we consider composite optimization problems of the form

$$\min_{x \in Q} \{\varphi(x) := f(x) + h(x)\}, \qquad (1)$$

where $Q \subseteq E$ is a convex closed set, $h(x)$ is a simple convex function (for example, $\|x\|_1$, which is used in LASSO problems), and $f(x)$ is a convex function with an inexact stochastic oracle. This means that, for any $x \in Q$, there are $f_{\delta, L}(x) \in \mathbb{R}$ and $g_{\delta, L}(x) \in E^*$ such that

$$0 \le f(y) - f_{\delta, L}(x) - \langle g_{\delta, L}(x), y - x\rangle \le \frac{L}{2}\|x - y\|^2 + \delta,$$

$\forall y \in Q$ and that instead of the pair $(f_{\delta, L}(x), g_{\delta, L}(x))$ (which is referred to as a (δ, L)-oracle) one can use only their stochastic approximations $(F_{\delta, L}(x, \xi), G_{\delta, L}(x, \xi))$. The last means that every point $x \in Q$ is associated with a random variable ξ such that $\mathbb{E}_\xi F_{\delta, L}(x, \xi) = f_{\delta, L}(x)$, $\mathbb{E}_\xi G_{\delta, L}(x, \xi) = g_{\delta, L}(x)$, and $\mathbb{E}_\xi(\|G_{\delta, L}(x, \xi) - g_{\delta, L}(x)\|_*)^2 \le \sigma^2$. Here, $\|\cdot\|_*$ is the dual norm defined in the standard manner as $\|g\|_* = \sup_{y \in E}\{\langle g, y\rangle : \|y\|_E \le 1\}$.

First-order methods for various special cases of the above class of problems were proposed in [6–10]. Below, general methods are suggested for solving problems of this class. Specifically, for the composite optimization problem (1), we propose a stochastic intermediate gradient method (SIGM) (Algorithm 1) that allows using an arbitrary norm on E and a prox-

function $d(x)$ (see the rigorous definition below) and has an $O\left(\frac{LR^2}{k^p} + \frac{\sigma R}{\sqrt{k}} + k^{p-1}\delta\right)$ convergence rate (Theorem 1), where k is the iteration number, R is an estimate of the distance between the starting point of the algorithm and the solution, and $p \in [1, 2]$ is a prescribed number. Theorem 1 also gives an estimate for the probability of large deviations from the convergence rate. Additionally, a modification of this method (Algorithm 2) is proposed for problems with a strongly convex function φ and its rate of convergence is estimated (see Theorem 2). The results of [4, 10] imply that our estimates coincide, up to a multiplicative constant, with lower bounds not only for the considered class of composite optimization problems with an inexact stochastic oracle, but also for all usually considered subclasses of this class. Additionally, we describe Algorithm 3 for controlling large deviations from the resulting convergence rate in the strongly convex case (Theorem 3).

The algorithms proposed have the following advantages.

(1) They are applicable to a wide range of problems: stochastic optimization, problems with an error in gradient evaluation, smooth and nonsmooth problems, and problems with a Hölder subgradient (see [6]), and strongly convex problems.

(2) For fixed values of δ, L, and the number of iterations, $p \in [1, 2]$ can be chosen so as to minimize the error of the resulting solution approximation. For $p = 1$, we obtain the gradient method and, for $p = 2$, a fast gradient method.

(3) Artificial randomization can be used in an initially deterministic problem if the computation of a stochastic oracle requires fewer arithmetic operations than that of the original deterministic oracle.

(4) The computational costs required for solving an auxiliary problem at Step 3 in Algorithm 1 can be reduced due to the use of an arbitrary norm and a prox-functions $d(x)$. For example, if Q is a unit simplex in an n-dimensional space, $d(x) = -\ln n + \sum_{i=1}^{n} x_i \ln x_i$, and $h(x) = 0$, then minimization in such auxiliary problems can be done by explicit formulas [4].

Let us describe the methods. Assume that E is equipped with some norm $\|\cdot\|$ and $d(x)$ is a differentiable function that is strongly convex with parameter 1 on Q with respect to the chosen norm (referred to as a prox-function). The Bregman distance is defined as

$$V(x, z) = d(x) - d(z) - \langle \nabla d(z), x - z\rangle.$$

We choose numbers $p \in [1, 2]$, $a = 2^{\frac{2p-1}{2}}$, $b = 2^{\frac{5-2p}{4}} p^{\frac{1-2p}{2}}$, and R: $\sqrt{2d(x^*)} \le R$, where x^* is a solution of problem (1), and sequences $\alpha_i = \frac{1}{a}\left(\frac{i+p}{p}\right)^{p-1}$, $\beta_i = L + \frac{b\sigma}{R}(i+p+1)^{\frac{2p-1}{2}}$, $B_i = a\alpha_i^2$, $\tau_i = \frac{\alpha_{i+1}}{B_{i+1}}$ $\forall i \ge 0$, $A_k = \sum_{i=0}^{k} \alpha_i$, $N \in \mathbb{N}$. Below, we describe Algorithm 1, which outputs a point y_k.

ALGORITHM 1: SIGM 1

Step 1. $x_0 = \operatorname{argmin}_{x \in Q}\{d(x)\}$.

Step 2. Let ξ_0 be a realization of the random variable ξ. Compute $G_{\delta, L}(x_0, \xi_0)$.

Step 3. Compute $y_0 = \operatorname{argmin}_{x \in Q}\{\beta_0 d(x) + \alpha_0\langle G_{\delta, L}(x_0, \xi_0), x - x_0\rangle + h(x)\}$. Set $k = 0$.

repeat

Step 4. Compute $z_k = \operatorname{argmin}_{x \in Q}\{\beta_k d(x) + \sum_{i=0}^{k} \alpha_i \langle G_{\delta, L}(x_i, \xi_i), x - x_i\rangle + A_k h(x)\}$.

Step 5. Set $x_{k+1} = \tau_k z_k + (1 - \tau_k) y_k$.

Step 6. Let ξ_{k+1} be a realization of the random variable ξ. Compute $G_{\delta, L}(x_{k+1}, \xi_{k+1})$.

Step 7. Compute $\xi_{k+1} = \operatorname{argmin}_{x \in Q}\{\beta_k V(x, z_k) + \alpha_{k+1}\langle G_{\delta, L}(x_{k+1}, \xi_{k+1}), x - z_k\rangle + \alpha_{k+1} h(x)\}$.

Step 8. Set $w_{k+1} = \tau_k \hat{x}_{k+1} + (1 - \tau_k) y_k$, $y_{k+1} = \frac{A_{k+1} - B_{k+1}}{A_{k+1}} y_k + \frac{B_{k+1}}{A_{k+1}} w_{k+1}$.

until $k > N$.

Theorem 1. 1. *Let the function f be equipped with an inexact stochastic oracle. Then the sequence y_k generated by Algorithm* 1 *as applied to problem* (1) *satisfies the relation*

$$\mathbb{E}_{\xi_0, \ldots, \xi_k} \varphi(y_k) - \varphi^* \le 48\left(\frac{LR^2}{k^p} + \frac{\sigma R}{\sqrt{k}} + k^{p-1}\delta\right).$$

2. *Let, additionally, $\xi_0, \ldots, \xi_k$ be independent identically distributed random variables,*

$$\mathbb{E}_{\xi}\left[\exp\left(\frac{\|G_{\delta, L}(x, \xi) - g_{\delta, L}(x)\|_*^2}{\sigma^2}\right)\right] \le \exp(1),$$

also set Q limited and

$$D = \max_{x, y \in Q} \|x - y\|.$$

Then the sequence y_k generated by Algorithm 1 *as applied to problem* (1) *satisfies the relation*

$$\mathbb{P}\Big(\varphi(y_k) - \varphi^* > 48\Big(\frac{LR^2}{k^p} + \frac{(1+\Omega)\sigma R}{\sqrt{k}} + k^{p-1}\delta + \frac{D\sigma\sqrt{\Omega}}{\sqrt{k}}\Big)\Big) \le 3\exp(-\Omega).$$

We introduce the following additional assumptions:

(i) E is a Euclidean space with inner product $\langle\cdot,\cdot\rangle$ and the norm $\|x\| = \sqrt{\langle x, Hx\rangle}$, where H is a symmetric positive definite matrix.

(ii) Without loss of generality, the minimum value of $d(x)$ is 0 and is reached at the point $0 \in E$.

(iii) The function $d(x)$ has quadratic growth with a constant V^2 with respect to the chosen norm: $d(x) \le \frac{V^2}{2}\|x\|^2$, $\forall x \in E$.

(iv) The function $\varphi(x)$ is strongly convex, i.e.,

$$\frac{\mu}{2}\|x-y\|^2 \le \varphi(y) - \varphi(x) - \langle g(x), y-x\rangle, \quad \forall x, y \in Q, \quad g(x) \in \partial\varphi(x). \tag{2}$$

Here, $\partial\varphi(x)$ is the subdifferential of $\varphi(x)$ at the point x.

Given an initial point u_0 and numbers R_0: $\|u_0 - x^*\| \le R_0$, $p \in [1, 2]$, and $N \in \mathbb{N}$, Algorithm 2 described below outputs a point u_{k+1}.

ALGORITHM 2: SIGM 2

Step 1. Set $k = 0$. Define $N_0 = \left\lceil \left(\frac{16e\sqrt{2}LV^2}{\mu}\right)^{\frac{1}{p}} \right\rceil$.

repeat

Step 2. Set

$$m_k = \max\left\{1, \left\lceil \frac{8192 e^{k+2}\sigma^2 V^2}{\mu^2 R_0^2 N_0} \right\rceil\right\},$$

$$R_k^2 = R_0^2 e^{-k} + \frac{48e2^p\delta}{\mu(e-1)}\left(\frac{16e\sqrt{2}LV^2}{\mu}\right)^{\frac{p-1}{p}}(1-e^{-k}).$$

Step 3. Take N_0 steps of Algorithm 1 beginning at the point $x_0 = u_k$ with the prox-function $d\left(\frac{x-u_k}{R_k}\right)$ and with the oracle $G_{\delta,L}(x_i, \xi_i)$ replaced by $\tilde{G}^k_{\delta,L}(x_i) = \frac{1}{m}\sum_{j=1}^{m_k} G_{\delta,L}(x_i, \xi_j)$ at Steps 2–4, 6, and 7. Set $u_{k+1} = y_{N_k}$, $k = k+1$.

until $k > N$.

Theorem 2. *Let assumptions* (i)–(iv) *and the assumptions from the first part of Theorem* 1 *hold. Then, after* $k \ge 1$ *outer iterations of Algorithm* 2, *we obtain*

$$\mathbb{E}\varphi(u_k) - \varphi^* \le \frac{\mu R_0^2}{2}e^{-k} + \frac{48e2^{p-1}}{e-1}\left(\frac{16e\sqrt{2}LV^2}{\mu}\right)^{\frac{p-1}{p}}\delta.$$

Moreover, if the oracle error δ *is chosen so that* $\delta \le \frac{\varepsilon(e-1)}{48e2^p}\left(\frac{16e\sqrt{2}LV^2}{\mu}\right)^{\frac{1-p}{p}}$, *then a point* u_N *satisfying* $\mathbb{E}\varphi(u_N) - \varphi^* \le \varepsilon$ *can be found after* $N = \max\left\{1, \left\lceil \ln\left(\frac{\mu R_0^2}{\varepsilon}\right)\right\rceil\right\}$ *outer iterations and* $O\left(\left(\frac{LV^2}{\mu}\right)^{\frac{1}{p}}\ln\left(\frac{\mu R_0^2}{\varepsilon}\right) + \frac{\sigma^2V^2}{\mu\varepsilon}\right)$ *oracle calls.*

Now let error tolerance ε and confidence level Λ be also given. An output of Algorithm 3, stated below, is the point u_N.

ALGORITHM 3: SIGM 3

Step 1. Set $N = \max\left\{1, \left\lceil \ln\left(\frac{\mu R_0^2}{\varepsilon}\right)\right\rceil\right\}$, $N_0 = \left\lceil \left(\frac{24e\sqrt{2}LV^2}{\mu}\right)^{\frac{1}{p}} \right\rceil$, $k = 0$.

repeat

Step 2. Set

$$m_k = \max\left\{1, \left\lceil \frac{18432\, e^{k+2}\sigma^2V^2\left(1+\ln\left(\frac{3N}{\Lambda}\right)\right)^2}{\mu^2R_0^2N_0} \right\rceil, \left\lceil \frac{6912\, e^{k+2}\sigma^2\ln\left(\frac{3N}{\Lambda}\right)}{\mu^2R_0^2N_0} \right\rceil\right\},$$

$$R_k^2 = R_0^2e^{-k} + \frac{48e2^p\delta}{\mu(e-1)}\left(\frac{24e\sqrt{2}LV^2}{\mu}\right)^{\frac{p-1}{p}}(1-e^{-k}),$$

$$Q_k = \{x \in Q: \|x - u_k\|^2 \le R_k^2\}.$$

Step 3. Take N_0 steps of Algorithm 1 for solving the problem $\min\limits_{x\in Q_k}\varphi(x)$ beginning at the point $x_0 = u_k$ with the prox-function $d\left(\frac{x-u_k}{R_k}\right)$ and with the oracle $G_{\delta,L}(x_i, \xi_i)$ replaced by $\tilde{G}^k_{\delta,L}(x_i) = \frac{1}{m_k}\sum_{j=1}^{m_k} G_{\delta,L}(x_i, \xi_j)$ at Steps 2–4, 6, and 7. Set $u_{k+1} = y_{N_k}$ and $k = k+1$.

until $k = N - 1$.

Theorem 3. *Let assumptions* (i)–(iv) *and the assumptions of Theorem* 1 *hold. Suppose that the oracle error* δ *satisfies the relation* $\delta \le \frac{\varepsilon(e-1)}{48e2^p}\left(\frac{24e\sqrt{2}LV^2}{\mu}\right)^{\frac{1-p}{p}}$. *Then Algorithm* 3 *finds an* (ε, Λ)*-solution* u_N *satisfying* $\mathbb{P}\{\varphi(u_N) - \varphi^* > \varepsilon\} \le \Lambda$ *after* $O\left(\left(\frac{LV^2}{\mu}\right)^{\frac{1}{p}}\ln\frac{\mu R_0^2}{\varepsilon} + \frac{\sigma^2V^2}{\mu\varepsilon}\left(\ln\left(\frac{3}{\Lambda}\ln\frac{\mu R_0^2}{\varepsilon}\right)\right)^2 + \frac{\sigma^2}{\mu\varepsilon}\ln\left(\frac{3}{\Lambda}\ln\frac{\mu R_0^2}{\varepsilon}\right)\right)$ *oracle calls.*

Note that, by using the strong convexity of φ, estimates of the same form as in Theorems 2 and 3 can be obtained for the squared error with respect to the argument.

ACKNOWLEDGMENTS

The authors are grateful to Professors Yu.E. Nesterov and A.S. Nemirovsky for a number of helpful discussions.

This study 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.

REFERENCES

1. Yu. G. Evtushenko, *Methods for Solving Optimization Problems and Applications to Optimization Systems* (Nauka, Moscow, 1982) [in Russian].
2. B. T. Polyak, *Introduction to Optimization* (Nauka, Moscow, 1983; Optimization Software, New York, 1987).
3. F. P. Vasil'ev, *Optimization Methods* (Faktorial, Moscow, 2002) [in Russian].
4. A. S. Nemirovsky and D. B. Yudin, *Problem Complexity and Method Efficiency in Optimization* (Nauka, Moscow, 1979; Wiley-Interscience, New York, 1983).
5. Yu. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course* (Kluwer Academic, Dordrecht, 2004).
6. O. Devolder, F. Glineur, and Yu. Nesterov, Math. Progr. A **146** (1), 37–75 (2014).
7. S. Ghadimi and G. Lan, SIAM J. Optim. **22** (4), 1469–1492 (2012).
8. S. Ghadimi and G. Lan, SIAM J. Optim. **23** (4), 2061–2089 (2013).
9. O. Devolder, F. Glineur, and Yu. Nesterov, "Intermediate gradient methods for smooth convex problems with inexact oracle," CORE Discussion Paper, 2013/17.
10. O. Devolder, "Exactness, Inexactness and stochasticity in first-order methods for large-scale convex optimization," PhD Thesis, 2013.

Translated by I. Ruzanova