ISSN 1064-5624, Doklady Mathematics, 2019, Vol. 99, No. 2, pp. 125–128. © Pleiades Publishing, Ltd., 2019.
Russian Text © The Author(s), 2019, published in Doklady Akademii Nauk, 2019, Vol. 485, No. 1.

MATHEMATICS

Accelerated Primal-Dual Gradient Descent with Linesearch for Convex, Nonconvex, and Nonsmooth Optimization Problems

S. V. Guminov[a,*], Yu. E. Nesterov[b,c], P. E. Dvurechensky[d], and A. V. Gasnikov[a,d]

Presented by Academician of the RAS K.V. Rudakov September 10, 2018

Received September 27, 2018

Abstract—A new version of accelerated gradient descent is proposed. The method does not require any a priori information on the objective function, uses a linesearch procedure for convergence acceleration in practice, converge according to well-known lower bounds for both convex and nonconvex objective functions, and has primal-dual properties. A universal version of this method is also described.

DOI: 10.1134/S1064562419020042

In the late 1980s, A.S. Nemirovski showed that auxiliary low-dimensional optimization does not improve the theoretical worst-case rate of convergence of a first-order optimal gradient-type method for smooth convex minimization problems [1]. However, in practice, accelerated methods with linesearch (in particular, conjugate gradient methods) are usually more efficient than their fixed-stepsize counterparts in terms of the number of iterations. Moreover, such procedures have been successfully applied to nonconvex optimization problems [2]. Unfortunately, it is also well known that the gain in performance due to the use of linesearch is significantly reduced by the computational complexity of such procedures. It was noted in [3] that, for problems of a certain type frequently occurring in solving dual problems, the complexity of executing a linesearch step nearly coincides with the complexity of a usual gradient step. This fact motivates the study of methods with linesearch and their primal-dual properties [4–8].

Consider the minimization problem

$$f(x) \to \min_{x \in \mathbb{R}^n}.$$

Its solution is denoted by x_*. Assume that the objective function is differentiable and its gradient satisfies the Lipschitz condition with a constant L: for all x, $y \in \mathbb{R}^n$,

$$\|\nabla f(y) - \nabla f(x)\|_2 \le L\|x - y\|_2.$$

We introduce an estimating sequence $\{\psi_k(x)\}$ [1, 4, 9, 10] and a sequence of coefficients $\{A_k\}$:

$$l_k(x) = \sum_{i=0}^{k} a_{i+1}\{f(y^i) + \langle \nabla f(y^i), x - y^i \rangle\},$$

$$\begin{aligned}\psi_{k+1}(x) &= l_k(x) + \psi_0(x)\\ &= \psi_k(x) + a_{k+1}\{f(y^k) + \langle \nabla f(y^k), x - y^k \rangle\},\end{aligned}$$

$$A_{k+1} = A_k + a_{k+1}, \quad A_0 = 0.$$

Let us describe an accelerated gradient method (AGM) with single linesearch.

[a] Moscow Institute of Physics and Technology, Dolgoprudnyi, Moscow oblast, 141700 Russia
[b] Center for Operations Research and Econometrics (CORE), Catholic University of Louvain, Louvain-la-Neuve, Belgium
[c] National Research University Higher School of Economics, Moscow, 101000 Russia
[d] Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, 127051 Russia
*e-mail: sergey.guminov@phystech.edu

Algorithm 1: AGM

Input: $x^0 = v^0$, L, N
Output: x^N
1: $k = 0$
2: **while** $k \le N - 1$ **do**
3: $\beta_k = \operatorname*{argmin}_{\beta \in [0,1]} f(v^k + \beta(x^k - v^k))$
4: $y^k = v^k + \beta_k(x^k - v^k)$
5: $x^{k+1} = y^k - \frac{1}{L}\nabla f(y^k)$
6: Choose a_{k+1} by solving $\frac{a_{k+1}^2}{A_{k+1}} = \frac{1}{L}$
7: $v^{k+1} = v^k - a_{k+1}\nabla f(y^k)$ $\quad \triangleright v^{k+1} = \operatorname*{argmin}_{x \in \mathbb{R}^n} \psi_{k+1}(x)$
8: $k = k + 1$
9: **end while**

The main difference of this algorithm from well-known similar accelerated gradient methods [4, 10, 11] is the stepsize selection in line 3. The previous algorithms used a fixed stepsize (e.g., $\beta_k = \frac{k}{k+2}$).

Instead of Step 5, one can use different stepsize selection procedures, such as the Armijo rule [2] and its modern analogues (as in the universal fast gradient method [12]). The version of the method using exact linesearch for stepsize selection will be referred to as ALSM.

Algorithm 2: ALSM

Input: $x^0 = v^0$
Output: x^N
1: $k = 0$
2: **while** $k \le N - 1$ **do**
3: $\beta_k = \operatorname*{argmin}_{\beta \in [0,1]} f(v^k + \beta(x^k - v^k))$
4: $y^k = v^k + \beta_k(x^k - v^k)$
5: $h_{k+1} = \operatorname*{argmin}_{h \ge 0} f(y^k - h\nabla f(y^k))$
6: $x^{k+1} = y^k - h_{k+1}\nabla f(y^k)$
7: Choose a_{k+1} by solving $f(y^k) - \frac{a_{k+1}^2}{2A_{k+1}}\|\nabla f(y^k)\|_2^2 = f(x^{k+1})$
8: $v^{k+1} = v^k - a_{k+1}\nabla f(y^k)$ $\quad \triangleright v^{k+1} = \operatorname*{argmin}_{x \in \mathbb{R}^n} \psi_{k+1}(x)$
9: $k = k + 1$
10: **end while**

Let us formulate the main theoretical results for these methods.

Theorem 1. *For both* AGM *and* ALSM,

$$\min_{k=0,\dots,N} \|\nabla f(y^k)\|_2^2 \le \frac{2L(f(x^0) - f(x_*))}{N}.$$

If $f(x)$ is convex, then, for both methods

$$\min_{k=[N/2],\ldots,N} \|\nabla f(y^k)\|_2^2 \le \frac{32L^2R^2}{N^3},$$

$$f(x^N) - f(x_*) \le \frac{2LR^2}{N^2},$$

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

The function $f(x)$ is called γ-weakly quasiconvex (where $\gamma \in (0,1]$) if, for all $x \in \mathbb{R}^n$,

$$\gamma(f(x) - f(x_*)) \le \langle \nabla f(x), x - x_* \rangle.$$

Note that γ-weakly quasiconvex functions are unimodal, but, in the general case, are not convex. If $f(x)$ is γ-weakly quasiconvex, the AGM method can be considered with the following restarting procedure: as soon as

$$f(x_i^N) - f(x_*) \le \left(1 - \frac{\gamma}{2}\right)(f(x_i^0) - f(x_*)),$$

set $x_{i+1}^0 = x_i^N$ and restart the method.

Theorem 2. *If $f(x)$ is γ-weakly quasiconvex, then, for the AGM and ALSM methods with the above-described restarting procedure,*

$$f(\tilde{x}^N) - f(x_*) = O\left(\frac{LR^2}{\gamma^3 N^2}\right),$$

where $R = \max\limits_{x: f(x) \le f(x_0)} \|x\|_2$ *and* $\{\tilde{x}^i\}$ *is the sequence of points generated by the method in the course of all starts.*

It can be shown that the SESOP method [3] can be applied to γ-weakly quasiconvex problems and has the convergence rate estimate

$$f(\tilde{x}^N) - f(x_*) = O\left(\frac{LR^2}{\gamma^2 N^2}\right)$$

with $R = \|x^0 - x_*\|_2$, but it requires solving a three-dimensional (possibly nonconvex) problem at every iteration step. On the contrary, the AGM method requires only solving a minimization problem on an interval.

Now we consider a convex optimization problem of the form

$$\phi(z) \to \min_{Az=0}. \qquad (1)$$

In this case, a dual minimization problem can be constructed, namely,

$$f(x) = \max_z \{\langle x, Az \rangle - \phi(z)\}$$
$$= \langle x, Az(x) \rangle - \phi(z(x)) \to \min_{x \in \mathbb{R}}.$$

According to the Demyanov–Danskin theorem, $\nabla f(x) = Az(x)$. Assume that $\phi(z)$ is μ-strongly convex. Then $\nabla f(x)$ satisfies the Lipschitz condition with the constant $L = \dfrac{\lambda_{\max}(A)}{\mu}$. Let us apply our methods to problem (1) with $x^0 = v^0 = 0$. Define

$$\tilde{z}^N = \frac{1}{A_N} \sum_{k=0}^{N-1} a_{k+1} z(y^k).$$

Theorem 3. *For the* AGM *and* ALSM *methods*,

$$f(x^N) + \phi(\tilde{z}^N) \le \frac{16LR^2}{N^2},$$

$$\|A\tilde{z}^N\|_2 \le \frac{16LR}{N^2},$$

where $R = \|x_*\|_2$.

Consider a class of problems in which the objective function $f(x)$ is not necessarily smooth. Let $\nabla f(x)$ denote some subgradient of $f(x)$. Assume that $\nabla f(x)$ satisfies the Hölder condition: for all $x, y \in \mathbb{R}^n$ and some $u \in [0,1]$,

$$\|\nabla f(y) - \nabla f(x)\|_2 \le M_\nu \|x - y\|_2^\nu.$$

The following ULSM method can be proposed for solving problems of this class.

Algorithm 3: ULSM

Input: Initial point $x^0 = v^0$, accuracy ε

Output: x^N

$k = 0$

while $k \le N - 1$ **do**

$\beta_k = \operatorname*{argmin}\limits_{\beta \in [0,1]} f(v^k + \beta(x^k - v^k))$

$y^k = v^k + \beta_k(x^k - v^k)$

$h_{k+1} = \operatorname*{argmin}\limits_{h \ge 0} f(y^k - h\nabla f(y^k))$, where $\langle \nabla f(y^k), v^k - y^k \rangle \ge 0$

$x^{k+1} = y^k - h_{k+1}\nabla f(y^k)$

Choose a_{k+1} by solving $f(x^{k+1}) = f(y^k) - \frac{a_{k+1}^2}{2A_{k+1}}\|\nabla f(y^k)\|_2^2 + \frac{\varepsilon a_{k+1}}{2A_{k+1}}$

$v^{k+1} = v^k - a_{k+1}\nabla f(y^k)$ $\qquad \triangleright\ v^{k+1} = \underset{x\in\mathbb{R}^n}{\operatorname{argmin}}\, \psi_{k+1}(x)$

$k = k + 1$

end while

Note that, in contrast to other universal methods [12, 13], this one does not require estimating the necessary stepsize in an inner loop. This leads to a somewhat better estimate for the rate of convergence and, on average, to a smaller number of oracle calls per iteration step.

Theorem 4. *If $f(x)$ is convex and its subgradient satisfies the Hölder condition, then*

$$f(x^N) - f(x_*)\frac{1}{2A_N}\|x_* - x^0\|_2^2 + \frac{\varepsilon}{2},$$

i.e., the method generates an ε-accurate solution after N iterations, where

$$N \le \inf_{\nu\in[0,1]} 2\left[\frac{1-\nu}{1+\nu}\right]^{\frac{1-\nu}{1+3\nu}}\left[\frac{M_\nu}{\varepsilon}\right]^{\frac{2}{1+3\nu}} R^{\frac{2+2\nu}{1+3\nu}}$$

with $R = \|x_0 - x^*\|_2$.

If the problem under consideration is strongly convex with a given constant μ, then use of the estimating sequence

$$\psi_{k+1}(x) = l_k(x) + \psi_0(x)$$
$$= \psi_k(x) + a_{k+1}\left\{f(y^k) + \langle\nabla f(y^k), x - y^k\rangle + \frac{\mu}{2}\|x - y^k\|_2^2\right\}$$

leads to optimal (up to a multiplicative constant) analogues of the above-described methods in the class of strongly convex problems.

ACKNOWLEDGMENTS

This work was supported by the Russian Science Foundation, project no. 18-71-10108.

REFERENCES

1. Yu. E. Nesterov, *Efficient Methods in Nonlinear Programming* (Radio i Svyaz', Moscow, 1989) [in Russian].
2. R. A. Waltz, J. L. Morales, J. Nocedal, and D. Orban, Math. Program. **107** (3), 391–408 (2006).
3. G. Narkiss and M. Zibulevsky, "Sequential subspace optimization method for large-scale unconstrained problems," Tech. Rep. CCIT No. 559 (Dep. Electr. Eng. Tech., Haifa, 2005).
4. Yu. Nesterov, Math. Program. **103** (1), 127–152 (2005).
5. Yu. Nesterov, Math. Program. **120** (1), 221–259 (2009).
6. P. Dvurechensky, A. Gasnikov, and A. Kroshnin, "Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorn's algorithm," *Proceedings of the 35th International Conference on Machine Learning* (PMLR, Stockholm, 2018), pp. 1367–1376.
7. P. Dvurechensky, A. Gasnikov, E. Gasnikova, S. Matsievsky, A. Rodomanov, and I. Usik, "Primal-dual method for searching equilibrium in hierarchical congestion population games," *CEUR-WS Supplementary Proceedings of the 9th International Conference on Discrete Optimization and Operations Research and Scientific School* (*DOOR 2016*) (Springer, Berlin, 2016), pp. 584–595. http://ceur-ws.org/Vol-1623.
8. A. Chernov, P. Dvurechensky, and A. Gasnikov, "Fast primal-dual gradient method for strongly convex minimization problems with linear constraints," *Proceedings of the 9th International Conference on Discrete Optimization and Operations Research* (*DOOR 2016*) (Springer, Berlin, 2016), pp. 391–403.
9. Yu. E. Nesterov, Dokl. Akad. Nauk SSSR **269** (3), 543–547 (1983).
10. Yu. E. Nesterov, *Introduction to Convex Optimization* (MTsNMO, Moscow, 2010) [in Russian].
11. Z. Allen-Zhu and L. Orecchia, "Linear coupling: An ultimate unification of gradient and mirror descent," *Proceedings of the 8th Innovations in Theoretical Computer Science* (Schloss Dagstuhl, Saarbrücken, 2017).
12. Yu. Nesterov, Math. Program. **152** (1–2), 381–404 (2015).
13. S. Guminov, A. Gasnikov, A. Anikin, and A. Gornov, "A universal modification of the linear coupling method," Optim. Methods Software (2019). doi 10.1080/10556788.2018.1517158

Translated by I. Ruzanova