ДОКЛАДЫ АКАДЕМИИ НАУК, 2019, том 485, № 1, с. 15–18

МАТЕМАТИКА

УДК 519.853.62

ПРЯМО-ДВОЙСТВЕННЫЙ УСКОРЕННЫЙ ГРАДИЕНТНЫЙ МЕТОД С ОДНОМЕРНЫМ ПОИСКОМ ДЛЯ ВЫПУКЛЫХ, НЕВЫПУКЛЫХ И НЕГЛАДКИХ ЗАДАЧ ОПТИМИЗАЦИИ

С. В. Гуминов[1,4,*], Ю. Е. Нестеров[2,3], П. Е. Двуреченский[4,5], А. В. Гасников[1,4]

Представлено академиком РАН К.В. Рудаковым 10.09.2018 г.

Поступило 27.09.2018 г.

Представлен новый ускоренный градиентный метод оптимизации. Данный метод не требует никакой априорной информации о целевой функции, использует процедуру одномерного поиска для ускорения сходимости на практике, сходится согласно известным нижним оценкам как для выпуклых, так и невыпуклых целевых функций и обладает свойством прямо-двойственности. Также представлена универсальная версия данного метода.

Ключевые слова: ускоренный градиентный спуск, одномерный поиск, прямо-двойственные методы, выпуклая оптимизация, невыпуклая оптимизация.

DOI: https://doi.org/10.31857/S0869-5652485115-18

В конце 80-х годов А.С. Немировский показал, что вспомогательная маломерная оптимизация не улучшает теоретическую оценку скорости сходимости оптимального метода первого порядка решения гладких выпуклых задач минимизации [1]. Однако на практике ускоренные методы с одномерным поиском (в частности, методы сопряжённых градиентов) обычно оказываются эффективнее своих аналогов с фиксированными длинами шагов в смысле числа итераций. Более того, подобные процедуры успешно применялись для решения невыпуклых задач оптимизации [2]. К сожалению, также хорошо известно, что прирост производительности за счёт использования процедур одномерного поиска значительно уменьшается из-за вычислительной сложности подобных процедур. В работе [3] было замечено, что для задач определённого вида, который часто встречается при решении двойственных задач, сложность выполнения шага одномерного поиска практически совпадает со сложностью выполнения обычного градиентного шага. Этот факт побуждает интерес к исследованию методов с одномерным поиском и их прямо-двойственности [4–8].

Рассмотрим задачу минимизации

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

Решение этой задачи обозначим за x_*. Будем предполагать, что целевая функция дифференцируема, а её градиент удовлетворяет условию Липшица с константой L: для всех $x,\ y \in \mathbb{R}^n$

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

Введём оценивающую последовательность $\{\psi_k(x)\}$ [1, 4, 9, 10] и последовательность коэффициентов $\{A_k\}$

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

$$\psi_{k+1}(x) = l_k(x) + \psi_0(x) =$$
$$= \psi_k(x) + a_{k+1}\{f(y^k) + \langle \nabla f(y^k),\ x - y^k\rangle\},$$
$$A_{k+1} = A_k + a_{k+1},\quad A_0 = 0.$$

Опишем ускоренный метод с одним одномерным поиском AGM.

Алгоритм 1 AGM

Input: $x^0 = v^0,\ L,\ N$

Output: x^N

1: $k = 0$

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

3: $\beta_k = \arg\min_{\beta \in [0,1]} f(v^k + \beta(x^k - v^k))$

---

[1] *Московский физико-технический институт (государственный университет), Долгопрудный Московской обл.*

[2] *Center for Operations Research and Econometrics (CORE), Catholic University of Louvain, Belgium*

[3] *Национальный исследовательский университет "Высшая школа экономики", Москва*

[4] *Институт проблем передачи информации Российской Академии наук, Москва*

[5] *Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany*

**E-mail*: *sergey.guminov@phystech.edu*

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: Выбрать a_{k+1} как решение $\frac{a_{k+1}^2}{A_{k+1}} = \frac{1}{L}$

7: $v^{k+1} = v^k - a_{k+1}\nabla f(y^k)$

$\triangleright$ $v^{k+1} = \arg\min_{x\in\mathbb{R}^n} \psi_{k+1}(x)$

8: $k = k + 1$

9: **end while**

Основным отличием данного алгоритма от известных похожих ускоренных градиентных методов [4, 10, 11] является выбор шага в строке 3. Предыдущие алгоритмы использовали фиксированную длину шага (например, $\beta_k = \frac{k}{k+2}$).

Вместо шага 5 могут быть использованы различные правила выбора шага, такие как правило Армихо [2] и его современные аналоги (как в универсальном быстром градиентном методе [12]). Версия метода, использующая для выбора шага точный одномерный поиск, будет называться ALSM.

Алгоритм 2 ALSM

Input: $x^0 = v^0$

Output: x^N

1: $k = 0$

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

3: $\beta_k = \arg\min_{\beta\in[0,1]} f(v^k + \beta(x^k - v^k))$

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

5: $h_{k+1} = \arg\min_{h\ge 0} f(y^k - h\nabla f(y^k))$

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

7: Выбрать a_{k+1} как решение

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

8: $v^{k+1} = v^k - a_{k+1}\nabla f(y^k)$

$\triangleright$ $v^{k+1} = \arg\min_{x\in\mathbb{R}^n} \psi_{k+1}(x)$

9: $k = k + 1$

10: **end while**

Сформулируем основные теоретические результаты для данных методов.

Теорема 1. *Для обоих методов* AGM *и* ALSM

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

Если $f(x)$ выпукла, то для обоих методов

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

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

где $R = \|x_* - x^0\|_2$.

Функция $f(x)$ называется γ-слабо-квазивыпуклой (где $\gamma \in (0, 1]$), если для всех $x \in \mathbb{R}^n$

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

γ-слабо-квазивыпуклые функции являются унимодальными, но в общем случае не являются выпуклыми. Если $f(x)$ γ-слабо-квазивыпукла, то можно рассмотреть метод AGM со следующей процедурой рестартов: как только

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

объявить $x_{i+1}^0 = x_i^N$ и перезапустить метод.

Теорема 2. *Если $f(x)$ γ-слабо-квазивыпукла, то для методов* AGM *и* ALSM *с вышеописанной процедурой рестартов верно*

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

где $R = \max_{x:\, f(x)\le f(x_0)}\|x\|_2$, а $\{\tilde{x}^i\}$ — *последовательность точек, которые генерирует метод по ходу всех запусков.*

Заметим, что метод SESOP [3] может быть применён к γ-слабо-квазивыпуклым задачам и имеет оценку скорости сходимости

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

с $R = \|x^0 - x_*\|_2$, но требует решения трёхмерной (возможно, невыпуклой) задачи на каждой итерации. Напротив, описанный в данной работе метод AGM требует лишь решения задачи минимизации на отрезке.

Теперь рассмотрим выпуклую задачу вида

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

В этом случае можно построить двойственную задачу минимизации

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

Согласно теореме Демьянова—Данскина $\nabla f(x) = Az(x)$. Предположим, что $\phi(x)$ μ-сильно выпукла. Тогда $\nabla f(x)$ удовлетворяет условию Липшица с константой $L = \frac{\lambda_{\max}(A)}{\mu}$. Применим к задаче (1) наши методы с $x^0 = v^0 = 0$. Определим

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

Теорема 3. *Для методов* AGM *и* ALSM

$$f(x^N) + \phi(\tilde{z}^N) \leq \frac{16LR^2}{N^2},$$
$$\|A\tilde{z}^N\|_2 \leq \frac{16LR}{N^2},$$

где $R = \|x_*\|_2$.

Рассмотрим класс задач, в котором целевая функция $f(x)$ не обязательно гладкая, и будем обозначать за $\nabla f(x)$ некоторый её субградиент. Предположим, что $\nabla f(x)$ удовлетворяет условию Гёльдера: для всех $x, y \in \mathbb{R}^n$ и некоторого $\nu \in [0, 1]$

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

Для решения данного класса задач можно предложить следующий метод ULSM.

Алгоритм 3 ULSM

Input: Начальная точка $x^0 = v^0$, точность ε

Output: x^N

1: $k = 0$

2: **while** $k \leq N - 1$ **do**

3: $\beta_k = \arg\min_{\beta\in[0,1]} f(v^k + \beta(x^k - v^k))$

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

5: $h_{k+1} = \arg\min_{h\geq 0} f(y^k - h\nabla f(y^k))$,
где $\langle \nabla f(y^k), v^k - y^k\rangle \geq 0$

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

7: Выбрать a_{k+1} как решение
$$f(x^{k+1}) = f(y^k) - \frac{a_{k+1}^2}{2A_{k+1}}\|\nabla f(y^k)\|_2^2 + \frac{\varepsilon a_{k+1}}{2A_{k+1}}$$

8: $v^{k+1} = v^k - a_{k+1}\nabla f(y^k)$
$\triangleright$ $v^{k+1} = \arg\min_{x\in\mathbb{R}^n} \psi_{k+1}(x)$

9: $k = k + 1$

10: **end while**

Заметим, что в отличие от других универсальных методов [12, 13] данный метод не требует оценивания требуемой длины шага во внутреннем цикле. Это приводит к несколько лучшей оценке скорости сходимости и в среднем меньшему числу обращений к оракулу за итерацию.

Теорема 4. *Если* $f(x)$ *выпукла и её субградиент удовлетворяет условию Гёльдера, то верно*

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

т.е. решение с точностью ε *генерируется методом за число итераций*

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

где $R = \|x_0 - x^*\|_2$.

Отметим, что если рассматриваемая задача сильно выпукла с известной константной μ, то рассмотрение оценивающей последовательности

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

приводит к аналогам вышеописанных методов, оптимальных (с точностью до мультипликативной константы) в классе сильно выпуклых задач.

Источник финансирования. Исследование выполнено за счёт гранта Российского научного фонда (проект 18–71–10108).

СПИСОК ЛИТЕРАТУРЫ

1. *Нестеров Ю.Е.* Эффективные методы в нелинейном программировании. М.: Радио и связь, 1989. 304 с.
2. *Waltz R.A.*, *Morales J.L.*, *Nocedal J.*, *Orban D.* An Interior Algorithm for Nonlinear Optimization that Combines Line Search and Trust Region Steps // Math. Progr. 2006. V. 107. № 3. P. 391–408.
3. *Narkiss G.*, *Zibulevsky M.* Sequential Subspace Optimization Method for Large-Scale Unconstrained Problems. Tech. Report CCIT № 559. Haifa: EE Dept. Technion, 2005.
4. *Nesterov Yu.* Smooth Minimization of Non-Smooth Functions // Math. Progr. 2005. V. 103. № 1. P. 127–152.
5. *Nesterov Yu.* Primal-Dual Subgradient Methods for Convex Problems // Math. Progr. 2009. V. 120. № 1. P. 221–259.
6. *Dvurechensky P.*, *Gasnikov A.*, *Kroshnin A.* Computational Optimal Transport: Complexity by Accelerated Gradient Descent Is Better Than by Sinkhorn's Algorithm. Proc. of the 35th International Conference on Machine Learning. Stockholm: PMLR, 2018. P. 1367–1376.
7. *Dvurechensky P.*, *Gasnikov A.*, *Matsievsky S.*, *Rodomanov A.*, *Usik I.* Primal-Dual Method for Searching

Equilibrium in Hierarchical Congestion Population Games CEUR-WS // Supplementary Proc. the 9th Intern. Conf. on Discrete Optimization and Operations Research and Scientific School (DOOR 2016). B.: Springer, 2016. P. 584–595. http://ceurws.org/Vol-1623/

8. *Chernov A.*, *Dvurechensky P.*, *Gasnikov A.* Fast Primal-Dual Gradient Method for Strongly Convex Minimization Problems with Linear Constraints // Proceeding of the 9th Intern. Conf. on Discrete Optimization and Operations Research (DOOR 2016). B.: Springer, 2016. P. 391–403.
9. *Нестеров Ю.Е.* Метод решения задач выпуклого программирования с трудоемкостью $O(1/k^2)$ // ДАН. 1983. Т. 269. № 3. С. 543–547.
10. *Нестеров Ю.Е.* Введение в выпуклую оптимизацию. М.: МЦНМО, 2010. 280 с.
11. *Allen-Zhu Z.*, *Orecchia L.* Linear Coupling: An Ultimate Unification of Gradient and Mirror Descent. Proc. of the 8th Innovations in Theoretical Computer Science. Saarbrücken: Schloss Dagstuhl, 2017.
12. *Nesterov Yu.* Universal Gradient Methods for Convex Optimization Problems // Math. Progr. 2015. V. 152. № 1/2. P. 381–404.
13. *Guminov S.*, *Gasnikov A.*, *Anikin A.*, *Gornov A.* A Universal Modification of the Linear Coupling Method // Optim. Met. and Soft. 2018. DOI: 10.1080/10556788.2018.1517158.

PRIMAL-DUAL ACCELERATED GRADIENT DESCENT WITH LINE SEARCH FOR CONVEX AND NONCONVEX OPTIMIZATION PROBLEMS

S. V. Guminov, Yu. E. Nesterov, P. E. Dvurechensky, A. V. Gasnikov

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

Received September 27, 2018

In this paper a new variant of accelerated gradient descent is proposed. The proposed method does not require any information about the objective function, uses exact line search for the practical accelerations of convergence, converges according to the well-known lower bounds for both convex and non-convex objective functions and possesses primal-dual properties. We also provide a universal version of said method, which converges according to the known lower bounds for both smooth and non-smooth problems.

Keywords: accelerated gradient descent, line-search, primal-dual methods, convex optimization, nonconvex optimization.