ДОКЛАДЫ АКАДЕМИИ НАУК, 2019, том 484, № 6, с. 667–671

МАТЕМАТИКА

УДК 519.853.62

О ДОСТИЖИМОСТИ ОПТИМАЛЬНЫХ ОЦЕНОК СКОРОСТИ СХОДИМОСТИ ЧИСЛЕННЫХ МЕТОДОВ ВЫПУКЛОЙ ОПТИМИЗАЦИИ ВЫСОКИХ ПОРЯДКОВ

А. В. Гасников[1,2,*], Э. А. Горбунов[1], Д. А. Ковалев[1], А. А. М. Мохаммед[1], Е. О. Черноусова[1]

Представлено академиком РАН К.В. Рудаковым 05.09.2018 г.

Поступило 27.09.2018 г.

Рассматривается проксимальный быстрый градиентный метод Монтейро—Свайтера (2013 г.), в котором используется один шаг метода Ньютона для приближённого решения вспомогательной задачи на каждой итерации проксимального метода. Метод Монтейро—Свайтера является оптимальным (по числу вычислений градиента и гессиана оптимизируемой функции) для достаточно гладких задач выпуклой оптимизации в классе методов, использующих только градиент и гессиан оптимизируемой функции. За счёт замены шага метода Ньютона на шаг недавно предложенного тензорного метода Ю.Е. Нестерова (2018 г.), а также за счёт специального обобщения условия подбора шага в проксимальном внешнем быстром градиентном методе удалось предложить оптимальный тензорный метод, использующий старшие производные. В частности, такой тензорный метод, использующий производные до третьего порядка включительно, оказался достаточно практичным ввиду сложности итерации, сопоставимой со сложностью итерации метода Ньютона. Таким образом, получено конструктивное решение задачи, поставленной Ю.Е. Нестеровым в 2018 г., об устранении зазора в точных нижних и завышенных верхних оценках скорости сходимости для имеющихся на данный момент тензорных методов порядка $p \geq 3$.

Ключевые слова: проксимальный ускоренный метод, тензорный метод, метод Ньютона, нижние оценки.

DOI: https://doi.org/10.31857/S0869-56524846667-671

В работе рассматривается задача выпуклой безусловной оптимизации

$$f(x) \to \min_{x \in \mathbb{R}^n}, \tag{1}$$

где функция f имеет непрерывные частные производные порядка $p \geq 3$. Предположим, что тензоры частных производных $\nabla^r f(x) \overset{\text{def}}{=} \left\{ \dfrac{\partial \nabla^{r-1} f(x)}{\partial x} \right\}_{i=1}^{n}$ (в частности, при $r = 2$ получаем, что $\nabla^r f(x) = \nabla^2 f(x)$ — это матрица Гессе функции f в точке x) удовлетворяют условию Липшица с соответствующими константами M_r, т.е.

$$\|\nabla^r f(x) - \nabla^r f(y)\|_2 \leq M_r \|x - y\|_2 \quad \forall x, y \in \mathbb{R}^n, \tag{2}$$

где $\|\cdot\|_2$ — стандартная евклидовая норма в $\mathbb{R}^n$. Условие $x, y \in \mathbb{R}^n$ можно заменить условием $x, y \in \{z \in \mathbb{R}^n | \, f(x) \leq f(x^0)\}$, где x^0 — стартовая точка. Кроме того, отметим, что на $\nabla^r f(x)[\cdot]$ — это симметричная r-линейная форма и её норма определяется стандартным образом:

$$\|\nabla^r f(x)\|_2 \overset{\text{def}}{=} \sup_{h_1, \ldots, h_n} \{\nabla^r f(x)[h_1, \ldots, h_n] \,|\, \|h_i\|_2 \leq 1, \; i = 1, 2, \ldots, n\}.$$

В частности,

$$\|\nabla f(x)\|_2 = \sup_{\|y\|_2 \leq 1} \langle \nabla f(x), y \rangle,$$

$$\|\nabla^2 f(x)\|_2 = \sup_{\|y\|_2 \leq 1} \sup_{\|z\|_2 \leq 1} \langle \nabla^2 f(x) y, z \rangle.$$

Для класса методов, у которых на каждой итерации разрешается не более чем $O(1)$ раз обращаться к оракулу (подпрограмме) за значениями $\nabla^r f(x)$, $r \leq p$, $p \geq 2$, оценка числа итераций, необходимых для достижения ε-точности по функции, будет иметь вид

$$O\left(\min\left\{ n \ln\left(\frac{\Delta f}{\varepsilon}\right), \; \frac{M_0^2 R^2}{\varepsilon^2}, \; \left(\frac{M_1 R^2}{\varepsilon}\right)^{\frac{1}{2}}, \ldots, \; \left(\frac{M_p R^{p+1}}{\varepsilon}\right)^{\frac{2}{3p+1}} \right\} \right), \tag{3}$$

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

[2] *Институт проблем передачи информации им. А.А. Харкевича Российской Академии наук, Москва*

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

где $R = \|x_0 - x_*\|_2$ — расстояние от точки старта до её проекции на множество решений задачи (1).

Гипотеза 1 (см. [1–3]). Существует такой алгоритм, использующий только информацию о $\nabla^r f(x)$, $r \le p$, который сходится согласно оценке (3).

Для случая $p = 2$ такой алгоритм был построен в работе [2]. Ниже построены такие алгоритмы и для случая $p \ge 2$.

Заметим, что в общем случае оценка (3) не может быть улучшена, даже если дополнительно известно, что $M_{p+1} < \infty$ и $M_{p+2} < \infty$ (см. [1, 4]).

Следуя работе Ю.Е. Нестерова [2], введём оператор

$$T^F_{p,M}(x) \in \operatorname*{Argmin}_{y\in\mathbb{R}^n}\left\{\sum_{r=0}^{p}\frac{1}{r!}\nabla^r F(x)[\underbrace{y-x,\ \ldots,\ y-x}_{r}] + \frac{M}{(p+1)!}\|y-x\|_2^{p+1}\right\}. \quad (4)$$

Утверждение 1.

1) (см. Theorem 1 из [2]). *Задача* (4) *при* $M \ge pM_p$ *является задачей выпуклой оптимизации*.

2) (см. Lemma 1 из [2]). *Для всех* $x \in \mathbb{R}^n$ *имеет место неравенство*

$$\|\nabla F(T^F_{p,M}(x))\|_2 \le \frac{M + M_p}{p!}\|T^F_{p,M}(x) - x\|_2^p.$$

В частности, при $M = pM_p$

$$\|\nabla F(T^F_{p,pM_p}(x))\|_2 \le \frac{(1+p)M_p}{p!}\|T^F_{p,pM_p}(x) - x\|_2^p.$$

Сначала в общих чертах опишем идею предлагаемого подхода. Следуя работе Р. Монтейро и Б. Свайтера [5], введём семейство функций ($L \ge 0$ — параметр)

$$F_{L,\tilde{x}}(x) = f(x) + \frac{L}{2}\|x - \tilde{x}\|_2^2.$$

По функции $F_{L,\tilde{x}}(x)$ определим функцию

$$\tilde{f}_L(x) = \min_{y\in\mathbb{R}^n} F_{L,x}(y) = F_{L,x}(y_L(x)). \quad (5)$$

Для любого $L \ge 0$ имеет место неравенство

$$\tilde{f}_L(x) \le f(x),$$

причём функция $\tilde{f}_L(x)$ является выпуклой и имеет L-липшицев градиент. Кроме того,

$$x_* \in \operatorname*{Arg\,min}_{x\in\mathbb{R}^n}\tilde{f}_L(x) \Rightarrow x_* \in \operatorname*{Arg\,min}_{x\in\mathbb{R}^n} f(x),$$
$$\tilde{f}_L(x_*) = f(x_*).$$

Таким образом, вместо исходной задачи можно решать (сглаженную по Моро) задачу

$$\tilde{f}_L(x) \to \min_{x\in\mathbb{R}^n}. \quad (6)$$

Заметим, что

$$\nabla\tilde{f}_L(x) = -L(y_L(x) - x).$$

На задачу (6) можно смотреть как на обычную задачу гладкой выпуклой оптимизации. Согласно (3) для $p = 1$ сложность решения задачи (6) (число вычислений градиента $\nabla\tilde{f}_L(x)$, т.е. число решений вспомогательных задач вида (5)) быстрым градиентным методом [5–7] можно оценить следующим образом:

$$O\left(\sqrt{\frac{LR^2}{\varepsilon}}\right). \quad (7)$$

Чем меньше выбирается параметр L, тем оценка (7) будет лучше, но при этом тем сложнее на каждой итерации решать вспомогательную подзадачу (5), чтобы посчитать градиент $\nabla\tilde{f}_L(x)$ с нужной точностью. Идея подхода, восходящего к работе [5] при $p \ge 2$, состоит в следующем:

1) вместо задачи (6) с фиксированным L рассмотреть параметрическое семейство задач (6) со специальным образом убывающей (на внешних итерациях) последовательностью $\{L_k\}$. Все эти задачи имеют одинаковый минимум x_*, который необходимо найти. На k-й (внешней) итерации быстрого градиентного метода используется $\nabla\tilde{f}_{L_k}(x)$;

2) при этом считать точно $\nabla\tilde{f}_{L_k}(x)$ нет возможности, поэтому для решения задачи (5) используется всего одна итерация оператора (4) $T^{F_{L_k,x}}_{p,pM_p}(x)$.

Здесь и везде в дальнейшем

$$T^{F_{L,x}}_{p,pM_p}(x) = \operatorname*{arg\,min}_{y\in\mathbb{R}^n}\left\{\sum_{r=0}^{p}\frac{1}{r!}[\nabla^r_z F_{L,x}(z)]_{z=x} \times [\underbrace{y-x,\ \ldots,\ y-x}_{r}] + \frac{pM_p}{(p+1)!}\|y-x\|_2^{p+1}\right\}.$$

В работе [5] (см. Proposition 5.2) было показано, что если вместо точного решения $y_L(x)$ задачи (5), для которого $\|\nabla F_{L,x}(y_L(x))\|_2 = 0$, на каждой внешней итерации удаётся найти только такое решение $\tilde{y}_L(x)$, что

$$\|\nabla F_{L,x}(\tilde{y}_L(x))\|_2 \le \frac{L}{2}\|\tilde{y}_L(x) - x\|_2, \quad (8)$$

то быстрый градиентный метод для задачи (6) (с постоянным на итерациях L) будет также сходиться согласно оценке (7), несмотря на неточность решения вспомогательных задач на каждой итерации (5).

Оценка (8) соответствует концепции относительной точности решения вспомогательной задачи в популярном сейчас способе ускорения неускоренных методов Catalyst [8]. Также в работе [5] (см. Theorem 4.1) было показано при $p = 2$, что если дополнительно с выполнением условия (8)

$$\|\nabla F_{L_k,x^k}(\tilde{y}_{L_k}(x^k))\|_2 \le \frac{L_k}{2}\|\tilde{y}_{L_k}(x^k) - x^k\|_2, \qquad (9)$$

удаётся (за счёт специального подбора L_k на каждой итерации) ещё и обеспечить выполнение условия

$$\frac{2(p+1)M_p}{p!\,L_k}\|\tilde{y}_{L_k}(x^k) - x^k\|_2^{p-1} \ge \frac{1}{2}, \qquad (10)$$

то число внешних итераций такого метода (число решений вспомогательных подзадач (5)) будет определяться оценкой (3)

$$O\left(\left(\frac{M_p R^{p+1}}{\varepsilon}\right)^{\frac{2}{3p+1}}\right),$$

что существенно улучшает оценку (7) при $p \ge 2$. Константа $\frac{1}{2}$ в правой части неравенства (9) выбрана для определённости; важно только, чтобы это было число, строго меньшее 1. Проблема, однако, в том, как обеспечить одновременное выполнение условий (8) и (9). Оказывается, если выбирать

$$\tilde{y}_{L_k}(x^k) = T^{F_{L_k,x^k}}_{p,pM_p}(x^k),$$

то согласно утверждению 1

$$\left\|\nabla F\left(T^{F_{L_k,x^k}}_{p,pM_p}(x^k)\right)\right\|_2 \le \frac{(1+p)M_p}{p!\,L_k}\|\tilde{y}_{L_k}(x^k) - x^k\|_2^p.$$

Следовательно, если

$$\frac{2(1+p)M_p}{p!\,L_k}\|\tilde{y}_{L_k}(x^k) - x^k\|_2^{p-1} \le 1,$$

то условие (8) будет выполнено. Ввиду непрерывной зависимости $\tilde{y}_{L_k}(x^k)$ от L^k и достаточно очевидного факта, состоящего в том, что при $x^k \ne x^*$ найдётся такое, вообще говоря, достаточно маленькое значение $\hat{L}_k > 0$, что

$$\frac{2(1+p)M_p}{p!\,\hat{L}_k}\|\tilde{y}_{\hat{L}_k}(x^k) - x^k\|_2^{p-1} \ge 1,$$

и достаточно большое значение $\bar{L}_k > 0$, что

$$\frac{2(1+p)M_p}{p!\,\bar{L}_k}\|\tilde{y}_{\bar{L}_k}(x^k) - x^k\|_2^{p-1} \le \frac{1}{2},$$

имеем, что подобрать L_k можно с помощью процедуры одномерного поиска [5]. В типичных ситуациях можно ожидать, что число вызовов оператора (4) $T^{F_{L_k,x^k}}_{p,pM_p}(x^k)$ на одной итерации внешнего метода (быстрого градиентного метода) будет $O(1)$. При этом каждый вызов такого оператора порождает свою выпуклую задачу. Сложность решения такой задачи (т.е. вычисление (4)) с нужной точностью сопоставима при $p = 2, 3$ по объёму вычислений со сложностью итерации метода Ньютона, т.е. оценивается как $\tilde{O}(n^{2,37})$ [2, 9–11] ($\tilde{O}(\cdot) = O(\cdot)$ с точностью до логарифмических множителей).

Приведём теперь сам алгоритм (метод Монтейро—Свайтера—Нестерова порядка $p \ge 2$; см. алгоритм 1) и основную теорему данной работы о скорости сходимости предложенного алгоритма.

Теорема 1 (см. Theorem 6.4 из [5] для $p = 2$). *Методу Монтейро—Свайтера—Нестерова порядка $p \ge 2$ (алгоритм 1) для обеспечения условия*

$$f(y^N) - f(x_*) \le \varepsilon$$

достаточно сделать

$$O\left(\left(\frac{M_p R^{p+1}}{\varepsilon}\right)^{\frac{2}{3p+1}}\right)$$

итераций. На каждой итерации в среднем $O(1)$ раз необходимо решать задачу выпуклой оптимизации вида

$$\sum_{r=0}^{p}\frac{1}{r!}[\nabla_z^r F_{L,x}(z)]_{z=x}[\underbrace{y-x, \ldots, y-x}_{r}] + \frac{pM_p}{(p+1)!}\|y-x\|_2^{p+1} \to \min_{y\in\mathbb{R}},$$

где

$$F_{L,x}(z) = f(z) + \frac{L}{2}\|z-x\|_2^2.$$

Таким образом, сложность каждой итерации при $p = 2, 3$ составляет $\tilde{O}(n^{2,37})$.

Алгоритм 1. Метод Монтейро—Свайтера—Нестерова.

Вход: u_0, y_0 — стартовые точки; N — число итераций; $A_0 = 0$

Выход: y^N

1: **for** $k = 0, 1, 2, \ldots, N-1$.

2: Выбрать L_k так, чтобы выполнялось условие (условие Монтейро—Свайтера [5] при $p = 2$)

$$\frac{1}{2} \le \frac{2(p+1)M_p}{p!\,L_k}\|y^{k+1} - x^k\|_2^{p-1} \le 1$$

для

$$a_{k+1} = \frac{\frac{1}{L_k} + \sqrt{\frac{1}{L_k^2} + 4\frac{A_k}{L_k}}}{2},\ A_{k+1} = A_k + a_{k+1},$$

// отметим, что $L_k a_k^2 = A_k$

$$x^k = \frac{A_k}{A_{k+1}} y^k + \frac{a_{k+1}}{A_{k+1}} u^k,$$

$$y^{k+1} = \tilde{y}_{L_k}(x^k) = T_{p,pM_p}^{F_{L_k,x^k}}(x^k)$$

// тензорный шаг Ю.Е. Нестерова [2]

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

4: **end for**

5: **return** y^N

Основным вкладом данной работы является:

1) замена шага метода Ньютона на обобщённый метод Ньютона—Нестерова с регуляризацией;

2) обобщение условия Монтейро—Свайтера на случай $p \geq 2$.

Сочетание этих двух пунктов позволило предложить методы (для разных $p \geq 2$), закрывающие зазор (несовпадение нижних оценок скорости сходимости с верхними оценками для наилучших известных методов), который оставался в оценках скорости сходимости методов высоких порядков при $p \geq 3$. Более того, ввиду главы 5 из [2] в случае $p = 3$ можно ожидать, что предложенный выше алгоритм Монтейро—Свайтера—Нестерова, названный в честь учёных, на идеях которых он был построен, будет эффективным на практике для задач умеренной размерности $n \sim 10^3$.

Источники финансирования. Работа А.В. Гасникова поддержана грантом РФФИ 18–29–03071_мк, работа Э.А. Горбунова поддержана грантом Президента РФ МД-1320.2018.1.

СПИСОК ЛИТЕРАТУРЫ

1. *Arjevani Y.*, *Shamir O.*, *Shiff R.* Oracle Complexity of Second-Order Methods for Smooth Convex Optimization // Math. Programming. 2017. P. 1–34.
2. *Nesterov Yu.* Implementable Tensor Methods in Unconstrained Convex Optimization. Prepr. Univ. catholique de Louvain; Center for Operations Res. and Econometrics (CORE). 2018. № 2018005.
3. *Гасников А.В.*, *Ковалев Д.А.* Гипотеза об оптимальных оценках скорости сходимости численных методов выпуклой оптимизации высоких порядков // Компьют. исслед. и моделирование. 2018. Т. 10. № 3. С. 305–314.
4. *Немировский А.С.*, *Юдин Д.Б.* Сложность задач и эффективность методов оптимизации. М.: Наука, 1979.
5. *Monteiro R.*, *Svaiter B.* An Accelerated Hybrid Proximal Extragradient Method for Convex Optimization and its Implications to Second-Order Methods // SIAM J. Optim. 2013. V. 23. № 2. P. 1092–1125.
6. *Nesterov Yu.* Introductory Lectures on Convex Optimization: A Basic Course. B.: Springer Science & Business Media, 2013. 87 p.
7. *Нестеров Ю.Е.* Введение в выпуклую оптимизацию. М.: МЦНМО, 2010.
8. *Lin H.*, *Mairal J.*, *Harchaoui Z.* Catalyst Acceleration for First-Order Convex Optimization: from Theory to Practice // J. Machine Learning Res. 2018. V. 18. № 212. С. 1–54.
9. *Nesterov Yu.*, *Polyak B.* Cubic Regularization of Newton Method and Its Global Performance // Math. Program. 2006. V. 108. № 1. P. 177–205.
10. *Гасников А.В.* Современные численные методы оптимизации. Метод универсального градиентного спуска. М.: МФТИ, 2018. 166 с.
11. *Кормен Т.*, *Лейзерсон Ч.*, *Ривест Р.*, *Штайн К.* Алгоритмы. Построение и анализ. М.: Изд. дом “Вильямс”, 2009.

REACHABILITY OF OPTIMAL CONVERGENCE RATE ESTIMATES FOR HIGH-ORDER NUMERICAL CONVEX OPTIMIZATION METHODS

A. V. Gasnikov, E. A. Gorbunov, D. A. Kovalev, A. A. M. Mokhammed, E. A. Chernousova

Presented by Academician of the RAS K.V. Rudakov September 5, 2018

Received September 27, 2018

The Monteiro–Svaiter accelerated hybrid proximal extragradient method (2013) with one step of Newton's method used at every iteration for the approximate solution of an auxiliary problem is considered. The Monteiro–Svaiter method is optimal (with respect to the number of gradient and Hessian evaluations for the optimized function) for sufficiently smooth convex optimization problems in the class of methods using only the gradient and Hessian of the optimized function. An optimal tensor method involving higher derivatives is proposed by replacing Newton's step with a step of Yu.E. Nesterov's recently proposed tensor method (2018) and by using a special generalization of the step size selection condition in the outer accelerated proximal extragradient method. This tensor method with derivatives up to the third order inclusive is found fairly practical, since the complexity of its iteration is comparable with that of Newton's one. Thus, a constructive solution is obtained for Nesterov's problem (2018) of closing the gap between tight lower and overstated upper bounds for the convergence rate of existing tensor methods of order $p \geq 3$.

Keywords: accelerated proximal method, tensor method, Newton method, lower bounds.