УДК 519.86

Обоснование гипотезы об оптимальных оценках скорости сходимости численных методов выпуклой оптимизации высоких порядков

The global rate of convergence for optimal tensor methods in smooth convex optimization

А.В. Гасников[а,б,в]*, д.ф.-м.н., доцент, в.н.с., Королёв (Мос. обл.), Россия*
Д.А. Ковалев[а]*, студент, Королёв (Мос. обл.), Россия*
А.А.М. Мохаммед[а]*, аспирант, Долгопрудный (Мос. обл.), Россия*
Е.О. Черноусова[а]*, к.ф.-м.н., доцент, Долгопрудный (Мос. обл.), Россия*

[а]Московский физико-технический институт,
141707 Россия, Долгопрудный (Московская область), Институтский пер., 9
[б] Институт проблем передачи информации РАН,
127051, Россия, Москва, Б. Каретный пер., 9
[в] Адыгейский государственный университет,
352700, Россия, республика Адыгея, Университетская ул., 208

A.V. Gasnikov, Doctor of Phys. And Math., associate professor, Korolev (Mos. Reg.), Russia
D.A. Kovalev, student, Korolev (Mos. Reg.), Russia
A. Mohammed, PhD student, Dolgoprudny (Mos. Reg), Russia
E.O. Chernousova, PhD, associate professor, Dolgoprudny (Mos. Reg), Russia

Работа в пп. 1–3 поддержана грантом РНФ 17-11-01027, а в пп. 4, 5 грантом РФФИ 18-29-03071 мк.

Абстракт

В данной работе рассматривается проксимальный быстрый градиентный метод Монтейро–Свайтера (2013 г.), в котором используются один шаг метода Ньютона для приближенного 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 счет замены шага метода Ньютона на шаг недавно предложенного тензорного метода Ю.Е. Нестерова (2018 г.), а также за счет 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 в виду сложности итерации сопоставимой со сложностью итерации метода Ньютона. Таким обра-

зом, 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: Метод Ньютона, матрица Гессе, нижние оценки, методы высокого порядка, тензорные методы, проксимальный быстрый градиентный метод.

Abstract

In this work we consider Monteiro–Svaiter accelerated hybrid proximal extragradient (A-HPE) framework and accelerated Newton proximal extragradient (A-NPE) framework. The last framework contains an optimal method for rather smooth convex optimization problems with second-order oracle. We generalize A-NPE framework for higher order derivative oracle (schemes). We replace Newton's type step in A-NPE that was used for auxiliary problem by Newton's regularized (tensor) type step (Yu. Nesterov, 2018). Moreover we generalize large step A-HPE/A-NPE framework by replacing Monteiro–Svaiter's large step condition so that this framework could work for high-order scheme. The main contribution of the paper is as follows: we propose optimal high-order methods for convex optimization problems. As far as we know for that moment there exist only zero, first and second order optimal methods that work according to the lower bounds. For higher order schemes there exists a gap between the lower bounds (Arjevani, Shamir, Shiff, 2017) and existing high-order (tensor) methods (Nesterov–Polyak, 2006; Yu. Nesterov, 2008; M. Baes, 2009; Yu. Nesterov, 2018). Asymptotically the ratio of the rates of convergences for the best existing methods and lower bounds is about 1.5. In this work we eliminate this gap and show that lower bounds are tight.

Key words: Newton method, Hesse matrix, lower bounds, tensor methods, proximal fast gradient descent.

1. Введение

В работе рассматривается задача выпуклой безусловной оптимизации

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

Предполагается, что

$$\left\| \nabla^r f(y) - \nabla^r f(x) \right\|_2 \leq M_r \left\| y - x \right\|_2, \; x, y \in \mathbb{R}^n, \; M_r \leq \infty, \; r = 0,1,2,\ldots$$

Условие $x, y \in \mathbb{R}^n$ можно заменить условием $x, y \in \left\{ z \in \mathbb{R}^n :\ f(z) \le f\left(x^0\right) \right\}$. Заметим, что $\nabla^r f(y)$ – тензор ранга r. В частности,

$$\nabla^2 f(x) = \left\{ \partial \nabla f(x) / \partial x_j \right\}_{j=1}^n = \left\| \partial^2 f(x) / \partial x_i \partial x_j \right\|_{i,j=1}^n$$

– матрица Гессе дважды гладкой функции $f(x)$. Аналогично можно определить

$$\nabla^{r+1} f(x) = \left\{ \partial \nabla^r f(x) / \partial x_j \right\}_{j=1}^n .$$

Поясним, что понимается под 2-нормой от тензора. Ограничимся случаем $r = 2$, тогда

$$\nabla^2 f(x) = \left\| \partial^2 f(x) / \partial x_i \partial x_j \right\|_{i,j=1}^n ,$$

$$\left\| \nabla^2 f(y) - \nabla^2 f(x) \right\|_2 = \sup_{\|h_1\|_2 \le 1} \sup_{\|h_2\|_2 \le 1} \left\langle \left(\nabla^2 f(y) - \nabla^2 f(x) \right)[h_1], h_2 \right\rangle =$$

$$= \sup_{\|x_1\|_2 \le 1} \sup_{\|x_2\|_2 \le 1} \left\langle \left(\nabla^2 f(y) - \nabla^2 f(x) \right) h_1, h_2 \right\rangle =$$

$$= \max \left\{ \lambda_{\max} \left(\nabla^2 f(y) - \nabla^2 f(x) \right), \left| \lambda_{\min} \left(\nabla^2 f(y) - \nabla^2 f(x) \right) \right| \right\}.$$

В общем случае см. [15]. Отметим также, что при $r = 0$: $\nabla^0 f(x) = f(x)$, а $\| \ \|_2 = | \ |$.

Целью данной работы является предложить оптимальные (тензорные) численные методы решения задачи (1), использующие старшие производные оптимизируемой функции. Оптимальность понимается в смысле Немировского–Юдина [11], см. также следующий пункт.

Оказалось, что построение оптимальных численных методов выпуклой оптимизации высокого порядка всецело завязано на идею ускорения методов первого порядка [12, 24, 28]. Другими словами потенциала, содержащегося в обычном быстром градиентном методе (оптимальном методе, работающим с информацией первого порядка: градиент и значение оптимизируемой функции) оказалось достаточно, чтобы на его основе строить оптимальные методы высокого порядка, использующие старшие производные оптимизируемой функции. На наш взгляд именно это наблюдение представляет наибольший интерес в данной работе.

2. Формулировка гипотезы

Для класса методов, у которых на каждой итерации разрешается не более чем $\mathrm{O}(1)$ раз обращаться к оракулу (подпрограмме) за $\nabla^r f(x)$, $r \le 1$, оценка числа итераций N, необходимых для достижения точности решения задачи ε (по функции)

$$f\left(x^N\right)-\min_{x\in\mathbb{R}^n} f\left(x\right)=f\left(x^N\right)-f\left(x_*\right)\le\varepsilon,$$

будет иметь вид

$$\mathrm{O}\left(\min\left\{n\ln\left(\frac{\Delta f}{\varepsilon}\right),\ \frac{M_0^2R^2}{\varepsilon^2},\left(\frac{M_1R^2}{\varepsilon}\right)^{1/2},\ \frac{M_0^2}{\mu\varepsilon},\left(\frac{M_1}{\mu}\right)^{1/2}\ln\left(\frac{\mu R^2}{\varepsilon}\right)\right\}\right), \tag{2}$$

где $R=\left\|x^0-x_*\right\|_2$ – расстояние от точки старта до решения. Если решение не единственно, то в определении R под x_* можно понимать (евклидову) проекцию точки x^0 на множество решений. Данная оценка в общем случае не может быть улучшена, даже если дополнительно известно, что, $M_2<\infty$, $M_3<\infty$, … [11]. При этом оценка (2) достигается [11, 12, 16, 25, 28].

В действительности, под M_0 можно понимать меньшую константу, которая только в худшем случае совпадает с введенной здесь [29]. Аналогичное замечание имеет место и по методам 2-го порядка [30] (и, вероятно, более высокого порядка). “Правильный” метод p-го порядка ($p\ge 1$) на первых итерациях “осуществляет” желаемую редукцию (уменьшение) констант гладкости $\left\{M_r\right\}_{r=0}^{p-1}$ за счет попадания в нужную область сходимости метода (причем, часто достаточно одной первой итерации [29, 30]).

Заметим, что если вместо $r=1$ имеет место $r=0$, то в приведенной оценке (2) все аргументы минимума следует домножить на размерность пространства n [1, 3, 4, 11, 13, 18, 31]. Отметим также, что у известных сейчас методов, отвечающих (с точностью до логарифмического множителя) первому аргументу минимума, достаточно дорогой является составляющая итерации, не связанная с вычислением градиента: $\gg n^2$ (см. также [11, 16, 22]).

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

$$\mathrm{O}\left(\min\left\{n\ln\left(\frac{\Delta f}{\varepsilon}\right),\ \frac{M_0^2R^2}{\varepsilon^2},\left(\frac{M_1R^2}{\varepsilon}\right)^{1/2},\left(\frac{M_2R^3}{\varepsilon}\right)^{2/7},\dots,\left(\frac{M_pR^{p+1}}{\varepsilon}\right)^{2/(3p+1)}\right\}\right). \tag{3}$$

Гипотеза (см. [5, 14, 27]): *Существует такой допустимый алгоритм (использующий только* $\nabla^r f\left(x\right)$, $r\le p$*), который сходится согласно оценке (3).*

Для случая $p=2$ такой алгоритм был построен в работе [24]. Ниже будут построены такие алгоритмы в общем случае $p\ge 2$.

Заметим, что в общем случае оценка (3) не может быть улучшена, даже если дополнительно известно, что $M_{p+1} < \infty$, $M_{p+2} < \infty$, … [11, 14]. Отметим, что здесь, также как и для градиентных методов [12, 16], можно строить «универсальные худшие в мире функции» в классе выпуклых полиномов степени $p+1$, которые обосновывают, что оценка (3) не может быть улучшена [14, 27]:

$$f_N(x) = \eta_{p+1}(A_{2N+1}x) - x_1, \ \eta_{p+1}(x) = \frac{1}{p+1}\sum_{i=1}^{n}|x_i|^{p+1},$$

$$A_m = \begin{pmatrix} U_m & 0 \\ 0 & I_{n-m} \end{pmatrix}, \ U_m = \underbrace{\begin{pmatrix} 1 -1 \ 0 \dots 0 \\ 0 \ 1 -1 \dots 0 \\ \dots\dots\dots\dots \\ 0 \ 0 \dots 1 -1 \\ 0 \ 0 \dots 0 \ 1 \end{pmatrix}}_{m}, \ I_{n-m} = \underbrace{\begin{pmatrix} 1\,0\,0 \dots 0 \\ 0\,1\,0 \dots 0 \\ \dots\dots\dots \\ 0\,0 \dots 0\,1 \end{pmatrix}}_{n-m}.$$

Точка старта выбирается следующим образом $x^0 = (0,\dots,0)^T$, а класс допустимых методов (алгоритмов) определяется условием

$$x^{k+1} \in x^0 + \sum_{l=0}^{k} S_f(x^l),$$

где

$$S_f(x) = \mathrm{Lin}\left\{ y_x(a_0,\dots,a_p;\gamma;q) : a_0,\dots,a_p \in \mathbb{R}; \gamma > 0; q = 1,2,3,\dots \right\},$$

$$y_x(a_0,\dots,a_p;\gamma;q) \in \mathrm{Arg}\min_{y\in\mathbb{R}^n}\left\{ \sum_{r=0}^{p} a_r \nabla^r f(x)\underbrace{[y-x,\dots,y-x]}_{r} + \gamma\|y-x\|_2^q \right\}.$$

В [27, Theorem 4] было показано, что для всех методов из описанного выше класса при $N \le (n-1)/2$ имеет место оценка

$$f_{2N+1}(x^N) - \min_{x\in\mathbb{R}^n} f_{2N+1}(x) \ge \frac{3p}{2^{p+1}(p+1)!}\frac{M_p R^{p+1}}{(2N+2)^{\frac{3p+1}{2}}}.$$

3. Схема оптимального метода

Следуя работе Ю.Е. Нестерова [27] введем оператор

$$T_{p,M}^{F}\left(x\right)\in \operatorname{Arg}\min_{y\in\mathbb{R}^n}\left\{\sum_{r=0}^{p}\frac{1}{r!}\nabla^r F\left(x\right)\underbrace{\left[y-x,...,y-x\right]}_{r}+\frac{M}{\left(p+1\right)!}\left\|y-x\right\|_2^{p+1}\right\}. \qquad (4)$$

Важным является следующее наблюдение работы [27]: для выпуклой функции $F\left(x\right)$ сумма, вообще говоря, невыпуклого по y многочлена Тейлора:

$$\sum_{r=0}^{p}\frac{1}{r!}\nabla^r F\left(x\right)\underbrace{\left[y-x,...,y-x\right]}_{r}$$

и

$$\frac{M}{\left(p+1\right)!}\left\|y-x\right\|_2^{p+1},\ M\geq pM_p$$

будет выпуклой по y функцией. Более того, при $p=2,3$ задача поиска $T_{p,M}^{F}\left(x\right)$ может быть эффективно решена (далее об этом будет написано немного подробнее). Заметим, что используемое здесь M в p раз больше M из [27].

Утверждение 1. ***1)*** **(см. [27, Theorem 1])** *Задач (4) при* $M\geq pM_p$ *является задачей выпуклой оптимизации.*

2) **(см. [27, Lemma 1])** *Для всех* $x\in\mathbb{R}^n$ *имеет место следующее неравенство*

$$\left\|\nabla F\left(T_{p,M}^{F}\left(x\right)\right)\right\|_2\leq\frac{M+M_p}{p!}\left\|T_{p,M}^{F}\left(x\right)-x\right\|^p.$$

При $M=pM_p$

$$\left\|\nabla F\left(T_{p,M}^{F}\left(x\right)\right)\right\|_2\leq\frac{\left(p+1\right)M_p}{p!}\left\|T_{p,M}^{F}\left(x\right)-x\right\|^p.$$

Сначала в общих чертах опишем идею предлагаемого подхода. Следуя работе Р. Монтейро и Б. Свайтера [24], введем семейство функций ($L\geq 0$ – параметр)

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

По функции $F_{L,\tilde{x}}\left(x\right)$ можно определить функцию

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

Для любого $L\geq 0$ имеет место неравенство

$$\tilde{f}_L\left(x\right)\leq f\left(x\right),$$

причем выпуклая функция $\tilde{f}_L\left(x\right)$ будет иметь L-Липшицев градиент. Кроме того,

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

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

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

Заметим, что

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

На задачу (6) можно смотреть как на обычную задачу гладкой выпуклой оптимизации. Согласно (2) сложность решения задачи (6) (число вычислений градиента $\nabla\tilde{f}_L(x)$, т.е. число решений вспомогательных задач вида (5)) быстрым градиентным методом [12, 24, 28] можно оценить следующим образом

$$\mathrm{O}\left(\left(\frac{LR^2}{\varepsilon}\right)^{1/2}\right). \tag{7}$$

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 [24] при $p = 2$, 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) $T_{p,pM_p}^{F_{L,x}}(x)$.

Здесь и везде в дальнейшем

$$T_{p,pM_p}^{F_{L,x}}(x) \in \operatorname{Arg}\min_{y\in\mathbb{R}^n}\left\{\sum_{r=0}^{p}\frac{1}{r!}\left[\nabla_z^r F_{L,x}(z)\right]_{z=x}\underbrace{\left[y-x,\dots,y-x\right]}_{r} + \frac{pM_p}{(p+1)!}\|y-x\|_2^{p+1}\right\}.$$

В работе [24, Proposition 5.2] было показано, что если вместо точного решения $y_L(x)$ задачи (5), для которого $\left\|\nabla F_{L,x}\left(y_L(x)\right)\right\|_2 = 0$, на каждой внешней итерации удается найти только такое решение $\tilde{y}_L(x)$, что

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

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 каталист [23]. Также в работе [24, Theorem 4.1] было показано при $p=2$, что если дополнительно с выполнением условия (8)

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

удается (за счет специального подбора L_k на каждой итерации) еще и обеспечить выполнение условия

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

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

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

что существенно улучшает оценку (7) при $p \ge 2$. Константа $1/2$ в правой части неравенства (10) выбрана для определенности; важно только, чтобы это было число строго меньшее 1. Проблема, однако, в том, как обеспечить одновременное выполнение условий (9) и (10)? Оказывается, если выбирать

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

то согласно утверждению 1

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

Следовательно, если

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

то условие (9) будет выполнено. В виду непрерывной зависимости $\tilde{y}_{L_k}\left(x^k\right)$ от L_k и достаточно очевидного наблюдения: при $x^k \neq x_*$ найдется такой, вообще говоря, достаточно маленький $\tilde{L}_k > 0$, что

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

имеем, что подобрать L_k можно с помощью процедуры вида: $L_k = 2L_{k-1}$, $L_k := L_k \big/ \sqrt{2}$. Конечно, есть риск “проскочить” нужный диапазон

$$\frac{1}{2} \leq \frac{2(p+1)}{p!}\frac{M_p}{L_k}\left\|\tilde{y}_{L_k}\left(x^k\right)-x^k\right\|_2^{p-1} \leq 1.$$

Несложно показать, что небольшой “проскок” не влияет на скорость сходимости (немного могут ухудшиться только числовые множители в оценках). Ну а если все-таки не хочется нарушать это условие, в таком случае можно предусмотреть процедуру “возврата” вида $L_k := \sqrt[4]{2}L_k$ и т.д. В типичных ситуация можно ожидать, что число вызовов оператора (4) $T_{p,pM_p}^{F_{L_k,x^k}}\left(x^k\right)$ на одной итерации внешнего метода (быстрого градиентного метода) будет $\mathrm{O}(1)$. 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 объему вычислений со сложностью итерации метода Ньютона, т.е. оценивается как $\tilde{\mathrm{O}}\left(n^{2.37}\right)$ [4, 10, 27, 30]. Также подбор параметра L_k можно осуществлять подобно процедуре Line search из работы [24, item 7].

Приведем теперь сам **Алгоритм** (метод Монтейро–Свайтера–Нестерова порядка $p \geq 2$) и **основную теорему** данной работы о скорости сходимости предложенного алгоритма.

Инициализация

Задаем x^0, y^0, $A_0 = 0$*;*

Основной цикл

Подбираем L_k *так, что*

$$a_{k+1} = \frac{1/L_k + \sqrt{1/L_k^2 + 4A_k/L_k}}{2},\quad A_{k+1} = A_k + a_{k+1},$$ // можно заметить, что $L_k a_{k+1}^2 = A_{k+1}$

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

$$\left| \begin{array}{l} y^{k+1} = \tilde{y}_{L_k}\left(x^k\right) = T_{p,pM_p}^{F_{L_k,x^k}}\left(x^k\right), \text{ // тензорный шаг Ю.Е. Нестерова [27]} \\ \dfrac{1}{2} \le \dfrac{2\left(p+1\right)}{p!}\dfrac{M_p}{L_k}\left\|\tilde{y}_{L_k}\left(x^k\right) - x^k\right\|_2^{p-1} \le 1; \text{ // условие Монтейро–Свайтера при } p=2 \text{ [24]} \end{array} \right.$$

$u^{k+1} = u^k - a_{k+1}\nabla f\left(\tilde{y}_{L_k}\left(x^k\right)\right).$

Теорема 1. (см. [24, Theorem 6.4] при $p=2$) *Описанному выше методу Монтейро–Свайтера–Нестерова порядка $p \ge 2$ для обеспечения условия*

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

достаточно сделать

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

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

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

где

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

Таким образом, сложность каждой итерации при $p = 2,3$ составляет $\tilde{\mathrm{O}}\left(n^{2.37}\right)$.

Основным вкладом данной работы является

1) Замена шага метода Ньютона на обобщенный метод Ньютона–Нестерова с регуляризацией;
2) Обобщение условия Монтейро–Свайтера на случай $p \ge 2$.

Сочетание этих двух пунктов позволило предложить методы (для разных $p \ge 2$), закрывающие зазор (не совпадение нижних оценок скорости сходимости с верхними оценками для наилучших известных методов), который оставался в оценках скорости сходимости методов высоких порядков при $p \ge 3$. Более того, в виду [27, item 5] в случае $p = 3$ можно ожидать, что предложенный выше алгоритм Монтейро–Свайтера–Нестерова, названный в честь ученых, на идеях которых он был построен, будет эффективным на практике для задач умеренной размерности $n = 10^3 - 10^4$.

4. Обсуждение полученных результатов

На момент написания статьи Р. Монтейро и Б. Свайтером [24] в 2011 году не было известно результата, полученного Ю.Е. Нестеровым в конце 2017 года [27], приведенного в утверждении 1. Поэтому исходный вариант метода Монтейро–Свайтера, отвечающий $p=2$, выглядит значительно более громоздко, чем приведенный выше метод. Собственно, приведенный выше метод отличается от метода Монтейро–Свайтера тем, что в последнем y^{k+1} пересчитывался исходя из шага метода Ньютона, а не с помощью (4). При $p=2$ итерация (4) отвечает так называемому методу Ньютона с кубической регуляризацией [30]. А вся процедура отчасти напоминает технику каталист [23], ускорения различных неускоренных методов, за счет общей ускоренной внешней оболочки. В данном случае мы ускоряли неускоренные методы вида (4), предложенные Ю.Е. Нестеровым в [27] и сходящиеся согласно оценкам

$$\mathrm{O}\left(\left(\frac{M_p R^{p+1}}{\varepsilon}\right)^{1/p}\right).$$

Отметим, что при этом процедура ускорения, предложенная в работах [15, 26, 27], не дает оптимальных оценок

$$\mathrm{O}\left(\left(\frac{M_p R^{p+1}}{\varepsilon}\right)^{1/(p+1)}\right).$$

И хотя, как отмечается в [26], разница

$$\frac{1}{p+1}-\frac{2}{3p+1}=\frac{p-1}{(p+1)(3p+1)}$$

является малой величиной, например, при $p=3$ эта разница равняется $1/20$, отношение степеней

$$\frac{1}{p+1}\bigg/\frac{2}{3p+1}=\frac{3p+1}{2(p+1)}\approx\frac{3}{2}$$

показывает, что порядки (степени) сходимости отличаются значительно – приблизительно в 1.5 раза. Таким образом, полученное в данной работе улучшение в оценке скорости сходимости может представлять не только теоретический интерес, но и практический. Впрочем, в работе [26] отсутствует необходимость в подборе L_k, что уменьшает стоимость итерации в несколько раз.

В работе Р. Монтейро и Б. Свайтера [24] предлагается способ обобщения

- на задачи композитной оптимизации (а значит и на задачи условной оптимизации на множествах простой структуры [4, § 2]);
- на случай, когда вспомогательная задача (4) решается неточно (в работе [24] вспомогательная задача отличается от (4)).

Все это можно (и несложно) перенести на предложенный в данной работе метод. Открытым является вопрос о возможности перенесения описанного в этой работе алгоритма на концепцию модели функции и более общие нормы/прокс-структуры [4, § 2, 3], [21].

Также как и для методов 1-го порядка для описанного выше семейства методов 2-го порядка и выше можно

- попробовать рассматривать их универсальные варианты [20];
- можно рассматривать работу методов в условиях наличия шума и различных концепций неточности, возникающей, при решении вспомогательных задач на каждой итерации [15, 19];
- также можно попробовать переносить и завязанные на наличие шумов конструкции, например, конструкцию mini-batching'а [4, 19].

По сути, конструкция Монтейро–Свайтера в исходном варианте [24] выполняла роль процедуры глобализации сходимости для метода Ньютона [7, 8, 17, 32]. Другой способ глобализации сходимости метода Ньютона обеспечивают методы внутренней точки Нестерова–Немировского [12, 16, 25, 28]. Оценки, полученные в последнем подходе при некоторых дополнительных предположениях могут доминировать оценки (3). Интересно было бы на практике сопоставить методы внутренней точки с предложенным в данной работе методом с $p = 3$.

В работах [5, 26, 28, 30] предлагается способ (рестарты) перенесения алгоритмов, работающих по оценкам (3), на случай, когда оптимизируемая функция $f(x)$ является также μ-сильно выпуклой функцией. Полученные в результате такого перенесения методы работают согласно следующим оценкам [5]

$$\mathrm{O}\left(\min\left\{ \left(\frac{M_1}{\mu}\right)^{1/2}, \left(\frac{M_2 R}{\mu}\right)^{2/7}, \ldots, \left(\frac{M_p R^{p-1}}{\mu}\right)^{2/(3p+1)} \right\} + \min_{r=2,\ldots,p} \log\left\lceil \log\left(\frac{\left(\mu^{r+1}/M_r^2\right)^{1/(r-1)}}{\varepsilon} \right) \right\rceil \right).$$

где $\lceil a \rceil = \max\{1, a\}$.

Для “правильных” методов второго порядка и выше в оценках скорости сходимости, по-видимому, можно писать вместо $M_r R^{r+1}$, $r \ge 2$ более простую величину $f\left(x^0\right) - f\left(x_*\right)$. Это отличает их от методов первого порядка, в которых вместо $M_1 R^2$ в общем случае можно писать $f\left(x^0\right) - f\left(x_*\right)$ только если дополнительно известно, что оптимизируемая функция сильно выпуклая [11, упражнение 6 § 3, гл. 4].

5. Схема доказательства основной теоремы 1

Сформулируем необходимые в дальнейшем результаты, полученные в [24] при $\sigma^2 = 1/2$, см. обозначения работы [24].

Алгоритм (метод Монтейро–Свайтера).

Инициализация

Задаем x^0, y^0, $A_0 = 0$;

Основной цикл

Подбираем L_k *и* y^{k+1} *так, что*

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

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

$$\left\| \nabla F_{L_k, x^k}\left(y^{k+1}\right) \right\|_2 \le \frac{L_k}{2} \left\| y^{k+1} - x^k \right\|_2 \quad \text{// условие Монтейро–Свайтера}$$

$$u^{k+1} = u^k - a_{k+1} \nabla f\left(y^{k+1}\right).$$

Теорема 2. (см. [24, Theorem 3.6]). *Для последовательности* $\left\{x^k, y^k, u^k\right\}$, *генерируемой методом Монтейро–Свайтера, справедливы следующие неравенства*

$$\frac{1}{2}\left\| u^N - x_* \right\|_2^2 + A_N \cdot \left(f\left(y^N\right) - f\left(x_*\right) \right) + \frac{1}{4} \sum_{k=1}^{N} A_k L_{k-1} \left\| y^k - x^{k-1} \right\|_2^2 \le \frac{1}{2} \left\| x^0 - x_* \right\|_2^2 = \frac{R^2}{2}, \tag{11}$$

$$f\left(y^N\right) - f\left(x_*\right) \le \frac{R^2}{2A_N}, \quad \left\| u^N - x_* \right\|_2 \le R, \tag{12}$$

$$\sum_{k=1}^{N} A_k L_{k-1} \left\| y^k - x^{k-1} \right\|_2^2 \le 2R^2. \tag{13}$$

◊ Наиболее важным результатом в теореме 2 является оценка (11). Эта оценка уточняет аналогичную оценку скорости сходимости для обычного (не проксимального) быстрого градиентного метода (см., например, [12, 28]) за счет наличия в левой части неравенства (11) суммы вида левой части неравенства (13). Именно этот «бонус» и дает возможность для «игры» на выборе L_k так, чтобы улучшить оценку (7), получаемую из (12) при стандартном выборе $L_k \equiv L$. Заметим также, что неравенства (12), (13) являются тривиальными следствиями основного неравенства (11). ◊

Утверждение 2. (см. [24, Lemma 3.7 a)]). *Для последовательности* $\{A_k, L_k\}$, *генерируемой методом Монтейро–Свайтера, справедливо следующее неравенство*

$$A_N \geq \frac{1}{4}\left(\sum_{k=1}^{N}\frac{1}{\sqrt{L_{k-1}}}\right)^2. \tag{14}$$

◊ Неравенство (14) следует из

$$a_{k+1} = \frac{1/L_k^2 + \sqrt{1/L_k^2 + 4A_k/L_k}}{2} \;\Rightarrow\; a_{k+1} \geq \frac{1}{2L_k} + \sqrt{\frac{A_k}{L_k}},$$

$$A_{k+1} = A_k + a_{k+1} \geq A_k + \sqrt{\frac{A_k}{L_k}} + \frac{1}{2L_k} \geq \left(\sqrt{A_k} + \frac{1}{2\sqrt{L_k}}\right)^2. \;◊$$

Предположим, что в алгоритме Монтейро–Свайтера удается дополнительно при подборе L_k и y^{k+1} обеспечивать выполнение условия типа (10)

$$\frac{1}{L_k}\left\|y^{k+1} - x^k\right\|_2^{p-1} \geq \theta. \tag{15}$$

Это условие было введено в [24, item 4] при $p = 2$, и названо условием «длинного шага». Формула (15) представляет собой естественное обобщение этого условия на случай $p \geq 2$.

Лемма 1 (см. [24, Lemma 4.2] при $p = 2$). *Для последовательности* $\{A_k, L_k\}$, *генерируемой методом Монтейро–Свайтера в предположении, что на каждой итерации удается так подобрать* L_k *и* y^{k+1}, *что дополнительно выполняется условие (10), справедливо следующее неравенство*

$$\sum_{k=1}^{N} A_k L_{k-1}^{\frac{p+1}{p-1}} \leq 2R^2\theta^{-\frac{2}{p-1}}. \tag{16}$$

◊ Действительно, из (13), (15) следует

$$\theta^{\frac{2}{p-1}}\sum_{k=1}^{N} A_k L_{k-1}^{\frac{p+1}{p-1}} \le \sum_{k=1}^{N} A_k L_{k-1}^{1+\frac{2}{p-1}} \left(\frac{1}{L_{k-1}}\left\|y^k - x^{k-1}\right\|_2^{p-1}\right)^{\frac{2}{p-1}} = \sum_{k=1}^{N} A_k L_{k-1}\left\|y^k - x^{k-1}\right\|_2^2 \le 2R^2 . \lozenge$$

Для того чтобы понять, как условие (16) может повлиять на оценку (14), определяющую скорость сходимости метода (см. (12)), сформулируем следующую лемму.

Лемма 1 (см. [24, Lemma A.1] при $p=2$). *Пусть справедливо неравенство (16) с* $A_k \ge 0$. *Тогда*

$$\sum_{k=1}^{N}\frac{1}{\sqrt{L_{k-1}}} \ge \frac{\theta^{\frac{1}{p+1}}}{\left(2R^2\right)^{\frac{p-1}{2(p+1)}}}\left(\sum_{k=1}^{N} A_k^{\frac{p-1}{3p+1}}\right)^{\frac{3p+1}{2(p+1)}} . \tag{17}$$

$\lozenge$ Введем новые переменные $z_k = 1\big/\sqrt{L_{k-1}}$ и рассмотрим следующую задачу выпуклой оптимизации

$$\sum_{k=1}^{N} z_k \to \min_{\sum_{k=1}^{N} A_k z_k^{-\gamma} \le C} , \tag{18}$$

где согласно (16)

$$\gamma = 2\frac{p+1}{p-1},\ C = 2R^2\theta^{-\frac{2}{p-1}} .$$

Задачу (18) можно явно решить с помощью принципа множителей Лагранжа в виду сепарабельной структуры функционала и ограничений

$$z_k = \left(\frac{1}{C}\sum_{j=1}^{N} A_j^{\frac{1}{\gamma+1}}\right)^{1/\gamma} A_k^{\frac{1}{\gamma+1}} .$$

Следовательно

$$\min_{\sum_{k=1}^{N} A_k z_k^{-\gamma} \le C} \sum_{k=1}^{N} z_k = \frac{1}{C^{1/\gamma}}\left(\sum_{k=1}^{N} A_k^{\frac{1}{\gamma+1}}\right)^{\frac{\gamma+1}{\gamma}} . \lozenge$$

Поскольку итоговый результат (теорема 1) сформулирован в категориях $\mathrm{O}(\ )$, то далее для наглядности и сокращения выкладок рассуждения будут проводиться в таких же категориях. Отметим, что все это можно сделать с явным выписыванием числовых множителей, см., например, [24, item 4].

Итак, из (14) и (17) имеем, что

$$A_N \geq \tilde{C} \cdot \left(\sum_{k=1}^{N} A_k^{\frac{p-1}{3p+1}} \right)^{\frac{3p+1}{(p+1)}}. \tag{19}$$

Подберем такой $\beta > 0$, чтобы для $A_N \sim N^{\beta}$ правая часть неравенства (19) асимптотически вела себя как $\sim N^{\beta}$. Для этого воспользуемся формулой Эйлера–Маклорена [6]

$$\sum_{k=1}^{N} k^{\beta \frac{p-1}{3p+1}} \sim N^{\beta \frac{p-1}{3p+1}+1}.$$

Таким образом, уравнение на β будет иметь вид

$$\beta \frac{p-1}{3p+1} + 1 = \beta \frac{p+1}{3p+1}.$$

Откуда можно получить, что

$$\beta = \frac{3p+1}{2}.$$

На базе этого наблюдения более аккуратные рассуждения (см., например, [24, item 4] при $p = 2$) позволяют строго доказать следующий факт.

Утверждение 3. *Существует такая числовая константа* $\tilde{c} > 0$*, что для любого* $N = 1, 2, 3, \ldots$ *выполняется неравенство*

$$A_N \geq \frac{\tilde{c}}{M_p R^{p-1}} N^{\frac{3p+1}{2}}. \tag{20}$$

Сочетание оценок (14), (20) вместе с анализом шага тензорного метода Ю.Е. Нестерова из работы [27]

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

с помощью утверждения 1 (соответствующие рассуждения были проведены в п. 3 выше), доказывает основную теорему 1.

Авторы выражают благодарность Ю.Е. Нестерову за ценное обсуждение результатов работы [27].

Список использованной литературы

1. *Баяндина А.С., Гасников А.В., Лагуновская А.А.* Безградиентные двухточечные методы решения задач стохастической негладкой выпуклой оптимизации при наличии малых шумов не случайной природы // Автоматика и Телемеханика. – 2018. – № 9. (в печати) – URL: https://arxiv.org/ftp/arxiv/papers/1701/1701.03821.pdf
2. *Васильев Ф.П.* Методы оптимизации. Т.2. – М.: МЦНМО, 2011. – 433 с.
3. *Воронцова Е.А., Гасников А.В., Горбунов Э.А.* Ускоренные спуски по случайному направлению и безградиентные методы с неевклидовой прокс-структурой // Автоматика и Телемеханика. – 2019. (в печати) – URL: https://arxiv.org/pdf/1710.00162.pdf
4. *Гасников А.В.* Современные численные методы оптимизации. Метод универсального градиентного спуска. – М.: МФТИ, 2018. – 166 с. – URL: https://arxiv.org/ftp/arxiv/papers/1711/1711.00394.pdf
5. *Гасников А.В., Ковалев Д.А.* Гипотеза об оптимальных оценках скорости сходимости численных методов выпуклой оптимизации высоких порядков // Компьютерные исследования и моделирование. – 2018. – Т. 10:3. – С. 305–314.
6. *Гельфонд А.О.* Исчисление конечных разностей. – М: ГИФМЛ, 1959. – 400 с.
7. *Денис Дж., Шнабель Р.* Численные методы безусловной оптимизации и решения нелинейных уравнений. – М.: Мир, 1988. – 440 с.
8. *Измаилов А.Ф., Солодов М.В.* Численные методы оптимизации. – М.: Физматлит, 2005. – 304 с.
9. *Карманов В.Г.* Математическое программирование. – М.: Наука, 1986. – 288 с.
10. *Кормен Т., Лейзерсон Ч., Ривест Р.* Алгоритмы: построение и анализ. – М.: МЦНМО, 2002. – 960 с.
11. *Немировский А.С., Юдин Д.Б.* Сложность задач и эффективность методов оптимизации. – М.: Наука, 1979. – 384 с.
12. *Нестеров Ю.Е.* Введение в выпуклую оптимизацию. – М.: МЦНМО, 2010. – 262 с.
13. *Протасов В.Ю.* К вопросу об алгоритмах приближенного вычисления минимума выпуклой функции по ее значениям // Мат. заметки. – 1996. – Т. 59 – № 1. С. 95–102.
14. *Arjevani Y., Shamir O., Shiff R.* Oracle complexity of second-order methods for smooth convex optimization // e-print, 2017. – URL: https://arxiv.org/pdf/1705.07260.pdf
15. *Baes M.* Estimate sequence methods: extensions and approximations // e-print, 2009. – URL: http://www.optimization-online.org/DB_FILE/2009/08/2372.pdf

16. *Bubeck S.* Convex optimization: algorithms and complexity // In Foundations and Trends in Machine Learning. – 2015. – V. 8. – no. 3-4. – P. 231–357.
17. *Conn A.B.*, *Gould N.I.M.*, *Toint P.L.* Trust region methods. – Philadelphia: SIAM, 2000.
18. *Dvurechensky P., Gasnikov A., Tiurin A.* Randomized Similar Triangles Method: A Unifying Framework for Accelerated Randomized Optimization Methods (Coordinate Descent, Directional Search, Derivative-Free Method) // e-print, 2017. – URL: https://arxiv.org/pdf/1707.08486.pdf
19. *Ghadimi S., Liu H., Zhang T.* Second-order methods with cubic regularization under inexact information // e-print, 2017. – URL: https://arxiv.org/pdf/1710.05782.pdf
20. *Grapiglia G.N., Nesterov Yu.* Regularized Newton methods for minimizing functions with Hölder continuous Hessian // SIAM J. Optim. – 2017. – V. 27(1). – P. 478–506.
21. *Hanzely F., Richtarik P., Xiao L.* Accelerated Bregman proximal gradient method for relatively smooth functions // arXiv.org e-Print archive. – 2018. – URL: https://arxiv.org/pdf/1808.03045.pdf
22. *Lee Y.-T., Sidford A., Wong S.C.-W.* A faster cutting plane method and its implications for combinatorial and convex optimization // e-print, 2015. – URL: https://arxiv.org/pdf/1508.04874.pdf
23. *Lin H.*, *Mairal J.*, *Harchaoui Z.* Catalyst acceleration for first-order convex optimization: from theory to practice // arXiv.org e-Print archive. 2017. – URL: https://arxiv.org/pdf/1712.05654.pdf
24. *Monteiro R., Svaiter B.* An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods // SIAM Journal on Optimization. – 2013. – V. 23(2). – P. 1092–1125.
25. *Nemirovski A.* Lectures on modern convex optimization analysis, algorithms, and engineering applications. – Philadelphia: SIAM, 2015. – URL: http://www2.isye.gatech.edu/~nemirovs/Lect_ModConvOpt.pdf
26. *Nesterov Yu*. Accelerating the cubic regularization of Newton's method on convex problems // Math. Prog., Ser. A. – 2008. – V. 112. – P. 159–181.
27. *Nesterov Yu.* Implementable tensor methods in unconstrained convex optimization // CORE Discussion Papers. 2018/5. – 2018. – 22 p. – URL: https://alfresco.uclouvain.be/alfresco/service/guest/streamDownload/workspace/SpacesStore/aabc2323-0bc1-40d4-9653-1c29971e7bd8/coredp2018_05web.pdf?guest=true
28. *Nesterov Yu.* Lectures on convex optimization. – Springer, 2018.

29. *Nesterov Yu.* Minimizing functions with bounded variation of subgradients // CORE Discussion Papers. 2005/79. – 2005. – 13 p. – URL: http://webdoc.sub.gwdg.de/ebook/serien/e/CORE/dp2005_79.pdf
30. *Nesterov Yu., Polyak P.* Cubic regularization of Newton method and its global performance // Math. Program. Ser. A. – 2006. – V. 108. – P. 177–205.
31. *Nesterov Yu., Spokoiny V.* Random gradient-free minimization of convex functions // Foundations of Computational Mathematics. – 2017. – V. 17(2). – P. 527–566.
32. *Nocedal J., Wright S.* Numerical optimization. – Springer, 2006.