**КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2018 Т. 10 № 6 С. 737–753**
DOI: 10.20537/2076-7633-2018-10-6-737-753
МАТЕМАТИЧЕСКИЕ ОСНОВЫ И ЧИСЛЕННЫЕ МЕТОДЫ МОДЕЛИРОВАНИЯ

УДК: 519.86

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

А. В. Гасников[1,2,3,a], Э. А. Горбунов[1,b], Д. А. Ковалев[1,c], А. А. М. Мохаммед[1,d], Е. О. Черноусова[1,e]

[1]Московский физико-технический институт,
Россия, 141701, г. Долгопрудный, Институтский пер., д. 9
[2]Институт проблем передачи информации РАН,
Россия, 127051, г. Москва, Б. Каретный пер., д. 9
[3]Кавказский математический центр, Адыгейский государственный университет,
Россия, 352700, Республика Адыгея, ул. Университетская, д. 208

E-mail: [a] gasnikov.av@mipt.ru, [b] ed-gorbunov@yandex.ru, [c] dakovalev1@mail.ru,
[d] a.nafea@science.sohag.edu.eg, [e] lena-ezhova@rambler.ru

Получено 03.09.2018, после доработки — 19.11.2018.
Принято к публикации 03.12.2018.

В данной работе рассматривается проксимальный быстрый градиентный метод Монтейро – Свайтера (2013 г.), в котором используется 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 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 $p \geqslant 3$.

Ключевые слова: метод Ньютона, матрица Гессе, нижние оценки, методы высокого порядка, тензорные методы, проксимальный быстрый градиентный метод

Работа в пп. 1, 2 поддержана грантом РНФ 17-11-01027, а в пп. 3, 4 — грантом РФФИ 18-29-03071 мк и грантом РФФИ 18-31-20005 мол_а_вед.

COMPUTER RESEARCH AND MODELING
2018 VOL. 10 NO. 6 P. 737–753
DOI: 10.20537/2076-7633-2018-10-6-737-753

Ки&М

MATHEMATICAL MODELING AND NUMERICAL SIMULATION

UDC: 519.86

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

A. V. Gasnikov[1,2,3,a], E. A. Gorbunov[1,b], D. A. Kovalev[1,c], A. A. M. Mohammed[1,d], E. O. Chernousova[1,e]

[1]Moscow Institute of Physics and Technology,
9 Institute lane, Dolgoprudny, 141701, Russia
[2]Institute for Information Transmission Problems RAS,
9 B. Karetny lane, Moscow, 127051, Russia
[3]Caucasian Mathematical Center, Adygea State University,
208 University st., Republic of Adygea, 352700, Russia

E-mail: [a] gasnikov.av@mipt.ru, [b] ed-gorbunov@yandex.ru, [c] dakovalev1@mail.ru,
[d] a.nafea@science.sohag.edu.eg, [e] lena-ezhova@rambler.ru

Received 03.09.2018, after completion — 19.11.2018.
Accepted for publication 03.12.2018.

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 schemes. 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. We also consider rather smooth strongly convex optimization problems and show how to generalize the proposed methods to this case. The basic idea is to use restart technique until iteration sequence reach the region of quadratic convergence of Newton method and then use Newton method. One can show that the considered method converges with optimal rates up to a logarithmic factor. Note, that proposed in this work technique can be generalized in the case when we can't solve auxiliary problem exactly, moreover we can't even calculate the derivatives of the functional exactly. Moreover, the proposed technique can be generalized to the composite optimization problems and in particular to the constraint convex optimization problems. We also formulate a list of open questions that arise around the main result of this paper (optimal universal method of high order e.t.c.).

Keywords: Newton method, Hesse matrix, lower bounds, tensor methods, proximal fast gradient descent

Citation: *Computer Research and Modeling*, 2018, vol. 10, no. 6, pp. 737–753 (Russian).

This work was supported by RNF 17-11-01027 in sections 1, 2, by RFFI 18-29-03071 mk in sections 3, 4 and by RFFI 18-31-20005 mol_a_ved.

Введение

В работе рассматривается задача выпуклой безусловной оптимизации

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

Предполагается, что

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

Условие $x, y \in \mathbb{R}^n$ можно заменить условием $x, y \in \left\{z \in \mathbb{R}^n : f(z) \leqslant f\left(x^0\right)\right\}$ (на самом деле, и это условие можно ослабить аналогично § 2 [Гасников, 2017]), где x^0 — точка старта. Заметим, что $\nabla^r f(y)$ — тензор ранга r. В частности,

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

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

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

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

$$\begin{aligned}
\nabla^2 f(x) &= \left\|\partial^2 f(x)\big/\partial x_i \partial x_j\right\|_{i,j=1}^n, \\
\left\|\nabla^2 f(y) - \nabla^2 f(x)\right\|_2 &= \sup_{\|h_1\|_2\leqslant 1} \sup_{\|h_2\|_2\leqslant 1} \left\langle \left(\nabla^2 f(y) - \nabla^2 f(x)\right)[h_1], h_2 \right\rangle = \\
&= \sup_{\|h_1\|_2\leqslant 1} \sup_{\|h_2\|_2\leqslant 1} \left\langle \left(\nabla^2 f(y) - \nabla^2 f(x)\right) h_1, h_2 \right\rangle = \\
&= \max\left\{\lambda_{\max}\left(\nabla^2 f(y) - \nabla^2 f(x)\right), \left|\lambda_{\min}\left(\nabla^2 f(y) - \nabla^2 f(x)\right)\right|\right\},
\end{aligned}$$

где $\lambda_{\max}()$ и $\lambda_{\min}()$ — максимальное и минимальное собственное значение соответствующей матрицы. В общем случае см. [Baes, 2009]. Отметим также, что при $r = 0$ $\nabla^0 f(x) = f(x)$, а $\|\cdot\|_2 = |\cdot|$.

Цель данной работы — предложить оптимальные (тензорные) численные методы решения задачи (1), использующие старшие производные оптимизируемой функции. Оптимальность понимается в смысле Немировского – Юдина [Немировский, Юдин, 1979] (см. также следующий пункт).

Оказалось, что построение оптимальных численных методов выпуклой оптимизации высокого порядка всецело завязано на идею ускорения методов первого порядка [Нестеров, 2010; Monteiro, Svaiter, 2013; Nesterov, 2018b]. 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.

Формулировка гипотезы

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

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

будет иметь вид

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

где $R = \left\|x^0 - x_*\right\|_2$ — расстояние от точки старта x^0 до решения x_*, $\Delta f = f\left(x^0\right) - f(x_*)$. Если решение не единственно, то в определении R под x_* можно понимать расстояние до (евклидовой) проекции точки x^0 на множество решений. Данная оценка в общем случае не может быть улучшена, даже если дополнительно известно, что $M_2 < \infty$, $M_3 < \infty$, ... [Немировский, Юдин, 1979]. При этом оценка (2) достигается [Немировский, Юдин, 1979; Нестеров, 2010; Bubeck, 2015; Nemirovski, 2015; Nesterov, 2018b].

В действительности под M_0 можно понимать меньшую константу, которая только в худшем случае совпадает с введенной в [Nesterov, 2005]. Аналогичное замечание имеет место и по методам 2-го порядка [Nesterov, Polyak, 2006] (и, вероятно, более высокого порядка). «Правильный» метод p-го порядка ($p \geqslant 1$) на первых итерациях «осуществляет» желаемую редукцию (уменьшение) констант гладкости $\{M_r\}_{r=0}^{p-1}$ за счет попадания в нужную область сходимости метода (причем часто достаточно одной первой итерации [Nesterov, 2005; Nesterov, Polyak, 2006]).

Заметим, что если вместо $r = 1$ имеет место $r = 0$, то в приведенной оценке (2) все аргументы минимума следует домножить на размерность пространства n [Баяндина и др., 2018; Воронцова и др., 2019; Гасников, 2017; Немировский, Юдин, 1979; Протасов, 1996; Dvurechensky, 2017; Nesterov, Spokoiny, 2017]. 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 [Немировский, Юдин, 1979; Bubeck, 2015; Lee et al., 2015]).

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

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

Гипотеза 1 (см. [Гасников, Ковалев, 2018; Arjevani et al., 2017; Nesterov, 2018a]). *Существует такой допустимый алгоритм (использующий только $\nabla^r f(x)$, для всех $r \leqslant p$), который сходится согласно оценке* (3).

Для случая $p = 2$ такой алгоритм был построен в работе [Monteiro, Svaiter, 2013]. Ниже будут построены такие алгоритмы в общем случае $p \geqslant 2$.

Заметим, что в общем случае оценка (3) не может быть улучшена, даже если дополнительно известно, что $M_{p+1} < \infty$, $M_{p+2} < \infty$, ... [Немировский, Юдин, 1979; Arjevani et al., 2017]. Отметим, что здесь, так же как и для градиентных методов [Нестеров, 2010; Bubeck, 2015], можно строить «универсальные худшие в мире функции» в классе выпуклых полиномов степени $p + 1$, которые обосновывают, что оценка (3) не может быть улучшена [Arjevani et al., 2017; Nesterov, 2018a]:

$$f_m(x) = \eta_{p+1}(A_m x) - x_1, \quad \eta_{p+1}(x) = \frac{1}{p+1}\sum_{i=1}^{n}|x_i|^{p+1},$$

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

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

$$x^{k+1} \in x^0 + \sum_{l=0}^{k} S_f\left(x^l\right) \subseteq \{x \in \mathbb{R}^n : x_i = 0, \quad i = k+2, \ldots, n\},$$

где

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

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

В Theorem 4 [Nesterov, 2018a] было показано, что для всех методов из описанного выше класса при $N \leqslant (n-1)/2$ имеет место оценка

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

Схема оптимального метода

Следуя работе Ю. Е. Нестерова [Nesterov, 2018a], введем оператор

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

Важным является следующее наблюдение работы [Nesterov, 2018a]: для выпуклой функции $F(x)$ сумма, вообще говоря, невыпуклого по y многочлена Тейлора:

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

и

$$\frac{M}{(p+1)!} \|y - x\|_2^{p+1}, \quad M \geqslant p M_p,$$

будет выпуклой по y функцией. Более того, при $p = 2, 3$ задача поиска $T^F_{p,M}(x)$ может быть эффективно решена (далее об этом будет написано немного подробнее). Заметим, что используемое здесь M в p раз больше M из [28].

Предложение 1.

1. (*См. Theorem 1 [Nesterov, 2018a].*) *Задача* (4) *при* $M \geqslant pM_p$ *является задачей выпуклой оптимизации.*

2) (*См. Lemma 1 [Nesterov, 2018a].*) *Для всех* $x \in \mathbb{R}^n$ *имеет место следующее неравенство:*

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

При $M = pM_p$

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

Сначала в общих чертах опишем идею предлагаемого подхода. Следуя работе Р. Монтейро и Б. Свайтера [Monteiro, Svaiter, 2013], введем семейство функций ($L \geqslant 0$ — параметр)

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

По функции $F_{L,\overline{x}}(x)$ можно определить функцию

$$\widetilde{f_L}(x) = \min_{y\in\mathbb{R}^n} F_{L,x}(y) = F_{L,x}(y_L(x)). \tag{5}$$

Для любого $L \geqslant 0$ имеет место неравенство

$$\widetilde{f_L}(x) \leqslant f(x),$$

причем выпуклая функция $\widetilde{f_L}(x)$ будет иметь L-липшицев градиент. Кроме того,

$$x_* \in \operatorname{Arg}\min_{x\in\mathbb{R}^n} f(x) \Rightarrow x_* \in \operatorname{Arg}\min_{x\in\mathbb{R}^n} \widetilde{f}(x), \quad \widetilde{f_L}(x_*) = f(x_*).$$

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

$$\widetilde{f_L}(x) \to \min_{x\in\mathbb{R}^n}. \tag{6}$$

Заметим, что

$$\nabla\widetilde{f_L}(x) = L\cdot(x - y_L(x)).$$

На задачу (6) можно смотреть как на обычную задачу гладкой выпуклой оптимизации. Согласно (2) сложность решения задачи (6) (число вычислений градиента $\nabla\widetilde{f_L}(x)$, т. е. число решений вспомогательных задач вида (5)) быстрым градиентным методом [Нестеров, 2010; Monteiro, Svaiter, 2013; Nesterov, 2018b] 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 $\nabla\widetilde{f_L}(x)$ с нужной точностью. Идея подхода, восходящего к работе [Monteiro, Svaiter, 2013] 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 $\nabla\widetilde{f_{L_k}}(x)$.

2. При этом считать точно $\nabla\widetilde{f_{L_k}}(x)$ 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

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

В работе [Monteiro, Svaiter, 2013] (Proposition 5.2) было показано, что если вместо точного решения $y_L(x)$ задачи (5), для которого $\left\|\nabla F_{L,x}(y_L(x))\right\|_2 = 0$, на каждой внешней итерации удается найти только такое решение $\widetilde{y}_L(x)$, что

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

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 [Lin et al., 2018]: из оценки (8) следует, что

$$\|\nabla \widetilde{f}_L(x) - L(x - \widetilde{y}_L(x))\|_2 \leqslant \|\nabla \widetilde{f}_L(x)\|_2.$$

Также в работе [Monteiro, Svaiter, 2013] (Theorem 4.1) было показано при $p = 2$, что если дополнительно с выполнением условия (8)

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

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

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

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 $p \geqslant 2$. 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

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

то согласно предположению 1

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

Следовательно, если выполнено условие

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

то условие (9) также будет выполнено. Далее заметим, что при $x^k \neq x_*$ найдется такое, вообще говоря, достаточно маленькое значение $\breve{L}_k > 0$, что

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

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

$$\frac{2(p+1)}{p!}\frac{M_p}{\bar{L}_k}\left\|\widetilde{y}_{\bar{L}_k}\left(x^k\right) - x^k\right\|_2^{p-1} \leqslant \frac{1}{2}.$$

Отсюда, ввиду непрерывной зависимости $\widetilde{y}_{L_k}\left(x^k\right)$ от L_k, имеем, что подобрать L_k, для которого будут одновременно выполняться условия (9) и (10), можно с помощью процедуры вида $L_k = 2L_{k-1}$, $L_k := L_k \big/ \sqrt{2}$. Конечно, есть риск «проскочить» нужный диапазон:

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

Несложно показать, что небольшой «проскок» не влияет на скорость сходимости (немного могут ухудшиться только числовые множители в оценках). 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 (в [Nesterov, 2018a] предлагается осуществлять вычисление $\left[\nabla_z^3 F_{L,x}(z)\right]_{z=x} \underbrace{[y-x, y-x, y-x]}_{r}$ с помощью автоматического дифференцирования), т. е. оценивается как $\widetilde{\mathrm{O}}\left(n^{2.37}\right)$ [Гасников, 2017; Кормен и др., 2009; Nesterov, 2018a; Nesterov, Polyak, 2006], где $\widetilde{\mathrm{O}} = \mathrm{O}$ с точностью до логарифмических множителей. Также подбор параметра L_k можно осуществлять подобно процедуре Line search из работы [Monteiro, Svaiter, 2013] (item 7).

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

Алгоритм 1. Метод Монтейро – Свайтера – Нестерова

Input: u_0, y_0 — стартовые точки; N — число итераций; $A_0 = 0$
Output: y^N

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

2: Выбрать L_k так, чтобы выполнялось условие (условие Монтейро – Свайтера [Monteiro, Svaiter, 2013] при $p = 2$)

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

для

$$a_{k+1} = \frac{\frac{1}{L_k} + \sqrt{\frac{1}{L_k^2} + 4\frac{A_k}{L_k}}}{2}, A_{k+1} = A_k + a_{k+1}, \ // \text{ отметим, что } L_k a_k^2 = A_k$$

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

$$y^{k+1} = \widetilde{y}_{L_k}(x^k) = T_{p,pM_p}^{F_{L_k,x^k}}(x^k) \ // \text{ тензорный шаг Ю. Е. Нестерова [Nesterov, 2018a]}$$

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

4: **end for**

5: **return** y^N

Теорема 1 (см. Theorem 6.4 [Monteiro, Svaiter, 2013] при $p = 2$). *Описанному выше методу Монтейро – Свайтера – Нестерова порядка $p \geqslant 2$ для обеспечения условия*

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

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*

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

где

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

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

Основным вкладом данной работы является:

1) замена шага метода Ньютона на обобщенный метод Ньютона – Нестерова с регуляризацией;
2) обобщение условия Монтейро – Свайтера на случай $p \geqslant 2$.

Сочетание этих двух пунктов позволило предложить методы (для разных $p \geqslant 2$), закрывающие зазор (несовпадение нижних оценок скорости сходимости с верхними оценками для наилучших известных методов), который оставался в оценках скорости сходимости методов высоких порядков при $p \geqslant 3$. Более того, в виду [Nesterov, 2018a] (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 $n \sim 10^3$–10^4.

Обсуждение полученных результатов

На момент написания статьи Р. Монтейро и Б. Свайтером [Monteiro, Svaiter, 2013] в 2011 году не было известно результата, полученного Ю. Е. Нестеровым в конце 2017 года [Nesterov, 2018a], приведенного в предложении 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 [Nesterov, Polyak, 2006]. А вся процедура отчасти напоминает технику каталист [Lin et al., 2018], ускорения различных неускоренных методов, за счет общей ускоренной внешней оболочки. В данном случае мы ускоряли неускоренные методы вида (4), предложенные Ю. Е. Нестеровым в [Nesterov, 2018a] 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 [Baes, 2009; Nesterov, 2008; Nesterov, 2018a], не дает оптимальных оценок, а дает оценки вида

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

И хотя, как отмечается в [Nesterov, 2008], 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 [Nesterov, 2008] отсутствует необходимость в подборе L_k, что уменьшает стоимость итерации в несколько раз.

В работе Р. Монтейро и Б. Свайтера [Monteiro, Svaiter, 2013] предлагается способ обобщения

- на задачи композитной оптимизации (а значит, и на задачи условной оптимизации на множествах простой структуры § 2 [Гасников, 2017]);
- на случай, когда $\nabla f(y)$ задан неточно.

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 § 2, 3 [Гасников, 2017], [Hanzely et al., 2018].

Так же как и для методов 1-го порядка, для описанного выше семейства методов 2-го порядка и выше можно

- попробовать рассматривать их универсальные варианты [Grapiglia, Nesterov, 2017];
- рассматривать работу методов в условиях наличия шума и различных концепций неточности, возникающей при решении вспомогательных задач на каждой итерации [Baes, 2009; Ghadimi et al., 2017];
- попробовать переносить и завязанные на наличие шумов конструкции, например конструкцию mini-batching'а [Гасников, 2017; Ghadimi et al., 2017].

По сути, конструкция Монтейро – Свайтера в исходном варианте [Monteiro, Svaiter, 2013] выполняла роль процедуры глобализации сходимости для метода Ньютона [Денис, Шнабель, 1988; Измаилов, Солодов, 2005; Conn et al., 2000; Nocedal, Wright, 2006]. Другой способ глобализации сходимости метода Ньютона обеспечивают методы внутренней точки Нестерова – Немировского [Нестеров, 2010; Bubeck, 2015; Nemirovski, 2015; Nesterov, 2018b]. 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 [Гасников, Ковалев, 2018; Nesterov, 2008; Nesterov, 2018b; Nesterov, Polyak, 2006] 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 следующей оценке [Гасников, Ковалев, 2018]:

$$\widetilde{\mathrm{O}}\left(\min\left\{\left(\frac{M_1}{\mu}\right)^{\frac{1}{2}}, \left(\frac{M_2R}{\mu}\right)^{\frac{2}{7}}, \ldots, \left(\frac{M_pR^{p-1}}{\mu}\right)^{\frac{2}{3p+1}}\right\} + \min_{r=2,\ldots,p}\log\left\lceil\log\left(\frac{\left(\frac{\mu^{r+1}}{M_r^2}\right)^{\frac{1}{r-1}}}{\varepsilon}\right)\right\rceil\right),$$

где $\lceil a\rceil = \max\{1, a\}$. Выписанная оценка не может быть улучшена [Гасников, Ковалев, 2018; Arjevani et al., 2017]. Действительно, из оценки (3) имеем

$$f\left(x^N\right) - f\left(x_*\right) \leqslant \frac{CM_p\left\|x^0 - x_*\right\|_2^{p+1}}{N^{(3p+1)/2}}.$$

Из μ-сильно выпуклости функции $f(x)$ отсюда имеем

$$\frac{\mu}{2}\left\|x^N - x_*\right\|_2^2 \leqslant f\left(x^N\right) - f\left(x_*\right) \leqslant \frac{CM_p\left\|x^0 - x_*\right\|_2^{p+1}}{N^{(3p+1)/2}}.$$

Выберем N таким образом, чтобы

$$\left\|x^N - x_*\right\|_2^2 \leqslant \frac{1}{2}\left\|x^0 - x_*\right\|_2^2.$$

Для этого достаточно сделать

$$\left(\frac{4M_pR^{p-1}}{\mu}\right)^{2/(3p+1)}$$

итераций. Процедуру можно повторить (рестартовать). Рестарты заканчиваются в момент попадания в окрестность квадратичной скорости сходимости для метода Ньютона [Arjevani et al., 2017]. Оценим эту окрестность. Напомним, что метод Ньютона имеет вид

$$\begin{aligned} x^{k+1} &= \arg\min_{x\in\mathbb{R}^n}\left\{f\left(x^k\right) + \left\langle\nabla f\left(x^k\right), x - x^k\right\rangle + \frac{1}{2}\left\langle\nabla^2 f\left(x^k\right)\left(x - x^k\right), x - x^k\right\rangle\right\} = \\ &= x^k - \left[\nabla^2 f\left(x^k\right)\right]^{-1}\nabla f\left(x^k\right), \end{aligned}$$

следовательно,

$$\begin{aligned} \left\|\nabla f\left(x^{k+1}\right)\right\|_2 &= \left\|\nabla f\left(x^{k+1}\right) - \nabla f\left(x^k\right) - \nabla^2 f\left(x^k\right)\left(x^{k+1} - x^k\right)\right\|_2 \leqslant M_2\left\|x^{k+1} - x^k\right\|_2^2 = \\ &= M_2\left\|\left[\nabla^2 f\left(x^k\right)\right]^{-1}\nabla f\left(x^k\right)\right\|_2^2 \leqslant \frac{M_2}{\mu^2}\left\|\nabla f\left(x^k\right)\right\|_2^2, \end{aligned}$$

$$\frac{M_2}{\mu^2}\left\|\nabla f\left(x^k\right)\right\|_2 < 1 \quad \Rightarrow \quad \frac{\mu}{2}\left\|x^k - x_*\right\|_2^2 \leqslant f\left(x^k\right) - f\left(x_*\right) \leqslant \frac{1}{2\mu}\left\|\nabla f\left(x^k\right)\right\|_2^2 < \frac{\mu^3}{2M_2^2}.$$

Значит, окрестность квадратичной скорости сходимости содержит открытый шар:

$$\|x - x_*\|_2^2 < \frac{\mu^2}{M_2^2}.$$

Таким образом, число рестартов можно оценить сверху следующим образом:

$$\left\lceil \log_2\left(\frac{M_2^2R^2}{\mu^2}\right)\right\rceil.$$

Вместо метода Ньютона здесь можно было бы использовать методы более высокого порядка сходимости [Гасников, 2017; Евтушенко, 2013; Nesterov, Polyak, 2006]. При попадании в окрестность сверхлинейной (квадратичной, кубической и т. д.) скорости сходимости таких методов желаемая относительная точность по функции $\widetilde{\varepsilon}$ будет достигнута после $\sim \log\log\left(\widetilde{\varepsilon}^{-1}\right)$ [Евтушенко, 2013; Немировский, Юдин, 1979; Arjevani et al., 2017] дополнительных итераций (основания логарифмов зависят от порядка выбранного метода).

Следуя [Nesterov, 2018b] (item 5.2.3), можно провести проделанные выше выкладки немного по-другому. А именно, из

$$f\left(x^N\right) - f\left(x_*\right) \leqslant \frac{CM_p\left\|x^0 - x_*\right\|_2^{p+1}}{N^{(3p+1)/2}}$$

и μ-сильно выпуклости функции $f(x)$ имеем

$$f\left(x^N\right) - f(x_*) \leqslant \frac{CM_p \left\|x^0 - x_*\right\|_2^{p+1}}{N^{(3p+1)/2}} \leqslant \frac{2^{(p+1)/2} CM_p}{\mu^{(p+1)/2} N^{(3p+1)/2}} \left(f\left(x^0\right) - f(x_*)\right)^{(p+1)/2}.$$

Выберем N таким образом, чтобы

$$\frac{2^{(p+1)/2} CM_p}{\mu^{(p+1)/2} N^{(3p+1)/2}} \left(f\left(x^0\right) - f(x_*)\right)^{(p+1)/2} \leqslant \frac{1}{2}(\underbrace{f\left(x^0\right) - f(x_*)}_{\Delta f}).$$

Для этого достаточно сделать

$$\left(\frac{2^{(p+3)/2} M_p \Delta f^{(p-1)/2}}{\mu^{(p+1)/2}}\right)^{2/(3p+1)}$$

итераций. Процедуру можно повторить (рестартовать). Таким образом,

$$\mathrm{O}\left(\left(\frac{M_p R^{p-1}}{\mu}\right)^{2/(3p+1)}\right) \to \mathrm{O}\left(\left(\frac{M_p \Delta f^{(p-1)/2}}{\mu^{(p+1)/2}}\right)^{2/(3p+1)}\right).$$

Однако такой переход можно было бы сделать и проще, заметив, что ввиду μ-сильно выпуклости функции $f(x)$

$$\frac{\mu}{2} R^2 = \frac{\mu}{2} \left\|x^0 - x_*\right\|_2^2 \leqslant f\left(x^0\right) - f(x_*) = \Delta f,$$

т. е.

$$R^{p-1} \leqslant (2\Delta f/\mu)^{(p-1)/2}.$$

Как и ожидалось, оценка из [Nesterov, 2018b] (item 5.2.3) соответствует приведенной выше оценке [Гасников, Ковалев, 2018; Arjevani et al., 2017].

Схема доказательства основной теоремы 1

Сформулируем необходимые в дальнейшем результаты, полученные в [Monteiro, Svaiter, 2013] при $\sigma^2 = 1/4$ (см. обозначения работы [Monteiro, Svaiter, 2013]).

Алгоритм 2. Метод Монтейро – Свайтера

Input: u_0, y_0 — стартовые точки; N — число итераций; $A_0 = 0$
Output: y^N

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

2: Выбрать L_k и y^{k+1} так, что $\left\|\nabla F_{L_k, x^k}\left(y^{k+1}\right)\right\|_2 \leqslant \frac{L_k}{2}\left\|y^{k+1} - x^k\right\|_2$ (условие Монтейро – Свайтера) для

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

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

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

4: **end for**

5: **return** y^N

Теорема 2 (см. Theorem 3.6 [Monteiro, Svaiter, 2013]). *Для последовательности* $\left\{x^k, y^k, u^k\right\}$, *генерируемой методом Монтейро – Свайтера, справедливы следующие неравенства:*

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

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

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

Доказательство. Наиболее важным результатом в теореме 2 является оценка (11). Она уточняет аналогичную оценку скорости сходимости для обычного (не проксимального) быстрого градиентного метода (см., например, [Нестеров, 2010; Nesterov, 2018b]) 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). □

Предложение 2 (см. Lemma 3.7a [Monteiro, Svaiter, 2013]). *Для последовательности* $\{A_k, L_k\}$, *генерируемой методом Монтейро – Свайтера, справедливо следующее неравенство:*

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

Доказательство. Неравенство (14) следует из

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

$$A_{k+1} = A_k + a_{k+1} \geqslant A_k + \sqrt{\tfrac{A_k}{L_k}} + \tfrac{1}{2L_k} \geqslant \left(\sqrt{A_k} + \tfrac{1}{2\sqrt{L_k}}\right)^2. \qquad \square$$

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

$$\frac{1}{L_k}\left\|y^{k+1} - x^k\right\|_2^{p-1} \geqslant \theta, \quad \theta > 0. \tag{15}$$

Это условие было введено в [Monteiro, Svaiter, 2013] (item 4) при $p = 2$ и названо условием «длинного шага». Формула (15) представляет собой естественное обобщение этого условия на случай $p \geqslant 2$.

Лемма 1 (см. Lemma 4.2 [Monteiro, Svaiter, 2013] при $p = 2$). *Для последовательности* $\{A_k, L_k\}$, *генерируемой методом Монтейро – Свайтера в предположении, что на каждой итерации удается так подобрать* $L_k y^{k+1}$, *что дополнительно выполняется условие* (10), *справедливо следующее неравенство:*

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

Доказательство. Действительно, из (13), (15) следует $\theta^{\frac{2}{p-1}} \sum\limits_{k=1}^{N} A_k L_{k-1}^{\frac{p+1}{p-1}} \leqslant$ $\leqslant \sum\limits_{k=1}^{N} A_k L_{k-1}^{1+\frac{2}{p-1}} \left(\frac{1}{L_{k-1}} \left\|y^k - x^{k-1}\right\|_2^{p-1}\right)^{\frac{2}{p-1}} = \sum\limits_{k=1}^{N} A_k L_{k-1} \left\|y^k - x^{k-1}\right\|_2^2 \leqslant 2R^2.$ □

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

Лемма 2 (см. Lemma A.1 [Monteiro, Svaiter, 2013] при $p = 2$). *Пусть справедливо неравенство* (16) *с* $A_k \geqslant 0$.

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

Доказательство. Введем новые переменные $z_k = 1\Big/\sqrt{L_{k-1}}$ и рассмотрим следующую задачу выпуклой оптимизации:

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

где согласно (16)

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

Задачу (18) можно явно решить с помощью принципа множителей Лагранжа ввиду 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,

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

□

Поскольку итоговый результат (теорема 1) сформулирован в категориях 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, [Monteiro, Svaiter, 2013] (item 4)).

Итак, из (14) и (17) имеем, что

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

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 [Гельфонд, 1959]:

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

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, [Monteiro, Svaiter, 2013] (item 4) при $p = 2$) позволяют строго доказать следующий факт.

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

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

Сочетание оценок (14), (20), вместе с анализом шага тензорного метода Ю. Е. Нестерова из работы [Nesterov, 2018a]:

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

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

Авторы выражают благодарность Ю. Е. Нестерову за ценное обсуждение результатов работы [Nesterov, 2018a].

Список литературы (References)

Баяндина А. С., Гасников А. В., Гулиев Ф. Ш., 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 (дата обращения: 03.09.2018).

Bayandina A. S., Gasnikov A. V., Guliev F. S., Lagunovskaya A. A. Bezgradientnye dvukhtochechnye metody resheniya zadach stokhasticheskoi negladkoi vypukloi optimizatsii pri nalichii malykh shumov ne sluchainoi prirody [Gradient-free two-points optimal method for non smooth stochastic convex optimization problem with additional small noise] // Avtomatika i telemekhanika [Automation and Remote Control]. — 2018. — № 9 (in print). — URL: https://arxiv.org/ftp/arxiv/papers/1701/1701.03821.pdf (accessed: 03.09.2018) (in Russian).

Васильев Ф. П. Методы оптимизации. Т. 2. — М.: МЦНМО, 2011. — 433 с.

Vasiliev F. P. Metody optimizatsii [Methods of Optimization]. Vol. 2. — Moscow: MTSNMO, 2011. — 433 p. (in Russian).

Воронцова Е. А., Гасников А. В., Горбунов Э. А. Ускоренные спуски по случайному направлению с неевклидовой прокс-структурой // Автоматика и Телемеханика. — 2019 (в печати). — URL: https://arxiv.org/pdf/1710.00162.pdf (дата обращения: 03.09.2018).

Vorontsova E. A., Gasnikov A. V., Gorbunov E. A. Uskorennye spuski po sluchainomu napravleniyu s neevklidovoi proks-strukturoi [Accelerated Directional Search with non-Euclidean prox-structure] // Avtomatika i telemekhanika [Automation and Remote Control]. — 2019 (in print). — URL: https://arxiv.org/pdf/1710.00162.pdf (accessed: 03.09.2018) (in Russian).

Гасников А. В. Современные численные методы оптимизации. Метод универсального градиентного спуска. — М.: МФТИ, 2018. — 166 с.

Gasnikov A. V. Sovremennye chislennye metody optimizatsii. Metod universalnogo gradientnogo spuska [Universal gradient descent]. — Moscow: MIPT, 2018. — 166 p. (in Russian).

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.
Gasnikov A. V., Kovalev D. A. Gipoteza ob optimalnykh otsenkakh skorosti skhodimosti chislennykh metodov vypukloi optimizatsii vysokikh poryadkov [Hypothesis of optimal estimates of the rate of convergence of numerical methods of convex optimization of high orders] // Kompyuternye issledovaniya i modelirovanie [Computer Research and Modeling]. — 2018. — Vol. 10, No. 3. — P. 305–314 (in Russian).

Гельфонд А. О. Исчисление конечных разностей. — М: ГИФМЛ, 1959. — 400 с.
Gelfond A. O. Ischislenie konechnykh raznostei [Calculus of finite differences]. — Moscow: GIFML, 1959. — 400 p. (in Russian).

Евтушенко Ю. Г. Оптимизация и быстрое автоматическое дифференцирование. — М.: ВЦ РАН, 2013. — 144 с. –– URL: http://www.ccas.ru/personal/evtush/p/198.pdf (дата обращения: 03.09.2018).
Evtushenko Yu. G. Optimizatsiya i bystroe avtomaticheskoe differentsirovanie [Optimization and fast automatic differentiation]. — Moscow: VC RAN, 2013. — 144 p. — URL: http://www.ccas.ru/personal/evtush/p/198.pdf (accessed: 03.09.2018) (in Russian).

Денис Дж., Шнабель Р. Численные методы безусловной оптимизации и решения нелинейных уравнений. — М.: Мир, 1988. — 440 с.
Dennis J., Schnabel R. Chislennye metody bezuslovnoi optimizatsii i resheniya nelineinykh uravnenii [Numerical methods for unconditional optimization and nonlinear equations]. — Moscow: Mir, 1988. — 440 p. (in Russian).

Измаилов А. Ф., Солодов М. В. Численные методы оптимизации. — М.: Физматлит, 2005. — 304 с.
Izmailov A. P., Solodov M. V. Chislennye metody optimizatsii [Numerical Optimization Methods]. — Moscow: Fizmatlit, 2005. — 304 p. (in Russian).

Карманов В. Г. Математическое программирование. — М.: Наука, 1986. — 288 с.
Karmanov V. G. Matematicheskoe programmirovanie [Mathematical programming]. — Moscow: Nauka, 1986. — 288 p. (in Russian).

Кормен Т., Лейзерсон Ч., Ривест Р., Штайн К. Алгоритмы. Построение и анализ: [пер. с англ.]. — М.: Издательский дом Вильямс, 2009.
Cormen T., Leiserson C., Rivest R., Stein C. Algoritmy. Postroenie i analiz [Introduction to algorithms]. — Moscow: Izdatelskii dom Vilyams, 2009 (in Russian).

Немировский А. С., Юдин Д. Б. Сложность задач и эффективность методов оптимизации. — М.: Наука, 1979.
Nemirovski A. S., Yudin D. B. Slozhnost zadach i effektivnost metodov optimizatsii [Problem Complexity and method Efficiency in Optimization]. — Moscow: Nauka, 1979 (in Russian).

Нестеров Ю. Е. Введение в выпуклую оптимизацию. — М.: МЦНМО, 2010. — 262 с.
Nesterov Yu. E. Vvedenie v vypukluyu optimizatsiyu [Introduction to convex optimization]. — Moscow: MTSNMO, 2010. — 262 p. (in Russian).

Протасов В. Ю. К вопросу об алгоритмах приближенного вычисления минимума выпуклой функции по ее значениям // Мат. заметки. — 1996. — Т. 59, № 1. — С. 95–102.
Protasov V. Yu. K voprosu ob algoritmakh priblizhennogo vychisleniya minimuma vypukloi funktsii po ee znacheniyam [On the question of algorithms for the approximate calculation of a minimum of a convex function from its values] // Mat. zametki [Math. notes]. — 1996. — Vol. 59, No. 1. — P. 95–102 (in Russian).

Arjevani Y., Shamir O., Shiff R. Oracle complexity of second-order methods for smooth convex optimization // Mathematical Programming. — 2017. — P. 1–34.

Baes M. Estimate sequence methods: extensions and approximations // e-print, 2009. — URL: http://www.optimization-online.org/DB_FILE/2009/08/2372.pdf (accessed: 03.09.2018).

Bubeck S. Convex optimization: algorithms and complexity // In Foundations and Trends in Machine Learning. — 2015. — Vol. 8, No. 3–4. — P. 231—357.

Conn A. B., Gould N. I. M., Toint P. L. Trust region methods. — Philadelphia: SIAM, 2000.

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 (accessed: 03.09.2018).

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 (accessed: 03.09.2018).

Grapiglia G. N., Nesterov Yu. Regularized Newton methods for minimizing functions with Hölder continuous Hessian // SIAM J. Optim. — 2017. — Vol. 27 (1). — P. 478–506.

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 (accessed: 03.09.2018).

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 (accessed: 03.09.2018).

Lin H., Mairal J., Harchaoui Z. Catalyst Acceleration for First-order Convex Optimization: from Theory to Practice // Journal of Machine Learning Research. — 2018. — Vol. 18:212. — P. 1–54.

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. — Vol. 23:2. — P. 1092–1125.

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 (accessed: 03.09.2018).

Nesterov Yu. Accelerating the cubic regularization of Newton's method on convex problems // Math. Prog., Ser. A. — 2008. — Vol. 112. — P. 159–181.

Nesterov Yu. Implementable tensor methods in unconstrained convex optimization // Université catholique de Louvain, Center for Operations Research and Econometrics (CORE). — 2018a. — No. 2018005.

Nesterov Yu. Lectures on convex optimization. — Springer, 2018b.

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 (accessed: 03.09.2018).

Nesterov Yu., Polyak B. Cubic regularization of Newton method and its global performance // Mathematical Programming. — 2006. — Vol. 108:1. — P. 177–205.

Nesterov Yu., Spokoiny V. Random gradient-free minimization of convex functions // Foundations of Computational Mathematics. — 2017. — Vol. 17 (2). — P. 527–566.

Nocedal J., Wright S. Numerical optimization. — Springer, 2006.