КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2018 Т. 10 № 3 С. 305–314
DOI: 10.20537/2076-7633-2018-10-3-305-314

Ки&М

СПЕЦИАЛЬНЫЙ ВЫПУСК

УДК: 519.86

Гипотеза об оптимальных оценках скорости сходимости численных методов выпуклой оптимизации высоких порядков

А. В. Гасников[1,2,a], Д. А. Ковалёв[1,b]

[1]Московский физико-технический институт,
Россия, 141707, Московская область, г. Долгопрудный, Институтский пер., д. 9
[2]Институт передачи информации РАН,
127051, Россия, г. Москва, Большой Каретный пер., д. 19

E-mail: [a] gasnikov@yandex.ru, [b] dakovalev1@mail.ru

Получено 28.02.2018, после доработки — 12.05.2018.
Принято к публикации 24.05.2018.

В данной работе приводятся нижние оценки скорости сходимости для класса численных методов выпуклой оптимизации первого порядка и выше, т. е. использующих градиент и старшие производные. Обсуждаются вопросы достижимости данных оценок. Приведенные в статье оценки замыкают известные на данный момент результаты в этой области. Отметим, что замыкание осуществляется без должного обоснования, поэтому в той общности, в которой данные оценки приведены в статье, их стоит понимать как гипотезу. Опишем более точно основной результат работы. Пожалуй, наиболее известным методом второго порядка является метод Ньютона, использующий информацию о градиенте и матрице Гессе оптимизируемой функции. Однако даже для сильно выпуклых функций метод Ньютона сходится лишь локально. Глобальная сходимость метода Ньютона обеспечивается с помощью кубической регуляризации оптимизируемой на каждом шаге квадратичной модели функции [Nesterov, Polyak, 2006]. Сложность решения такой вспомогательной задачи сопоставима со сложностью итерации обычного метода Ньютона, т. е. эквивалентна по порядку сложности обращения матрицы Гессе оптимизируемой функции. В 2008 году Ю. Е. Нестеровым был предложен ускоренный вариант метода Ньютона с кубической регуляризацией [Nesterov, 2008]. В 2013 г. Monteiro – Svaiter сумели улучшить оценку глобальной сходимости ускоренного метода с кубической регуляризацией [Monteiro, Svaiter, 2013]. В 2017 году Arjevani – Shamir – Shiff показали, что оценка Monteiro – Svaiter оптимальна (не может быть улучшена более чем на логарифмический множитель на классе методов 2-го порядка) [Arjevani et al., 2017]. Также удалось получить вид нижних оценок для методов порядка $p \geq 2$ для задач выпуклой оптимизации. Отметим, что при этом для сильно выпуклых функций нижние оценки были получены только для методов первого и второго порядка. В 2018 году Ю. Е. Нестеров для выпуклых задач оптимизации предложил методы 3-го порядка, которые имеют сложность итерации сопоставимую со сложностью итерации метода Ньютона и сходятся почти по установленным нижним оценкам [Nesterov, 2018]. Таким образом, было показано, что методы высокого порядка вполне могут быть практичными. В данной работе приводятся нижние оценки для методов высокого порядка $p \geq 3$ для сильно выпуклых задач безусловной оптимизации. Работа также может рассматриваться как небольшой обзор современного состояния развития численных методов выпуклой оптимизации высокого порядка.

Ключевые слова: метод Ньютона, матрица Гессе, нижние оценки, чебышёвские методы, сверхлинейная сходимость

Работа поддержана грантом РНФ № 17-11-01027.

© 2018 Александр Владимирович Гасников, Дмитрий Александрович Ковалёв

COMPUTER RESEARCH AND MODELING
2018 VOL. 10 NO. 3 P. 305–314
DOI: 10.20537/2076-7633-2018-10-3-305-314

Ки&М

SPECIAL ISSUE

UDC: 519.86

A hypothesis about the rate of global convergence for optimal methods (Newton's type) in smooth convex optimization

A. V. Gasnikov[1,2,a], D. A. Kovalev[1,b]

[1]Moscow Institute of Physics and Technology,
9 Institutskij per., Dolgoprudny, Moscow Region, 141707, Russia
[2]Institute for Information Transmission Problems of the Russian Academy of Sciences,
19 Bolshoy Karetny per., Moscow, 127051, Russia

E-mail: [a] gasnikov@yandex.ru, [b] dakovalev1@mail.ru

Received 28.02.2018, after completion — 12.05.2018.
Accepted for publication 24.05.2018.

In this paper we discuss lower bounds for convergence of convex optimization methods of high order and attainability of this bounds. We formulate a hypothesis that covers all the cases. It is noticeable that we provide this statement without a proof. Newton method is the most famous method that uses gradient and Hessian of optimized function. However, it converges locally even for strongly convex functions. Global convergence can be achieved with cubic regularization of Newton method [Nesterov, Polyak, 2006], whose iteration cost is comparable with iteration cost of Newton method and is equivalent to inversion of Hessian of optimized function. Yu. Nesterov proposed accelerated variant of Newton method with cubic regularization in 2008 [Nesterov, 2008]. R. Monteiro and B. Svaiter managed to improve global convergence of cubic regularized method in 2013 [Monteiro, Svaiter, 2013]. Y. Arjevani, O. Shamir and R. Shiff showed that convergence bound of Monteiro and Svaiter is optimal (cannot be improved by more than logarithmic factor with any second order method) in 2017 [Arjevani et al., 2017]. They also managed to find bounds for convex optimization methods of p-th order for $p \geq 2$. However, they got bounds only for first and second order methods for strongly convex functions. In 2018 Yu. Nesterov proposed third order convex optimization methods with rate of convergence that is close to this lower bounds and with similar to Newton method cost of iteration [Nesterov, 2018]. Consequently, it was showed that high order methods can be practical. In this paper we formulate lower bounds for p-th order methods for $p \geq 3$ for strongly convex unconstrained optimization problems. This paper can be viewed as a little survey of state of the art of high order optimization methods.

Keywords: Newton method, Hesse matrix, lower bounds, Chebyshev's type methods, superliner rate of convergence

Citation: *Computer Research and Modeling*, 2018, vol. 10, no. 3, pp. 305–314 (Russian).

This work was supported by RSCF grant No. 17-11-01027.

© 2018 Alexander V. Gasnikov, Dmitry A. Kovalev

Введение

В работе рассматривается задача выпуклой безусловной оптимизации

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

Предполагается, что

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

и $f(x)$ является μ-сильно выпуклой в 2-норме функцией ($\mu \ge 0$), т. е. для любых $x, y \in \mathrm{R}^n$

$$f(y) \ge f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|_2^2.$$

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

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

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

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

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

В общем случае см. [Baes, 2009]. Отметим также, что при $r = 0$ $\nabla^0 f(x) = f(x)$, а $\|\cdot\|_2 = |\cdot|$.

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, необходимых для достижения точности решения задачи ε (по функции), будет иметь вид

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

где $R = \left\|x^0 - x_*\right\|_2$ — расстояние от точки старта до решения. 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 данная оценка достигается [Немировский, Юдин, 1979; Нестеров, 2010; Bubeck, 2015].

В действительности под M_0 можно понимать меньшую константу, которая только в худшем случае совпадает с введенной здесь [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 ($p \geq 1$) 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 все аргументы минимума следует домножить на размерность пространства n [Баяндина и др., 2018; Воронцова и др., 2018; Гасников, 2016; Немировский, Юдин, 1979; Протасов, 1996; Dvurechensky et al., 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 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 точности ε (по функции), будет иметь вид

$$\mathrm{O}\left(\min\left\{n\ln\left(\frac{\Delta f}{\varepsilon}\right),\ \frac{M_0^2R^2}{\varepsilon^2},\ \left(\frac{M_1R^2}{\varepsilon}\right)^{1/2},\ \left(\frac{M_2R^3}{\varepsilon}\right)^{2/7},\ \ldots,\ \left(\frac{M_pR^{p+1}}{\varepsilon}\right)^{2/(3p+1)},\right.\right.$$
$$\left.\left.\min\left\{\left(\frac{M_1}{\mu}\right)^{1/2},\ \left(\frac{M_2R}{\mu}\right)^{2/7},\ \ldots,\left(\frac{M_pR^{p-1}}{\mu}\right)^{2/(3p+1)}\right\}+\min_{r=2,\ldots,p}\log\log\left(\frac{\left(\mu^{r+1}\middle/M_r^2\right)^{1/(r-1)}}{\varepsilon}\right)\right\}\right).$$

Гипотеза 1. *Данная оценка в общем случае не может быть улучшена, даже если дополнительно известно, что* $M_{p+1} < \infty, M_{p+2} < \infty, \ldots$.

В случае $\mu = 0$ или при $r \leq 2$ данная гипотеза верна (см. [Немировский, Юдин, 1979; Arjevani et al., 2017]). При этом недавно было также установлено [Monteiro, Svaiter, 2013; Arjevani et al., 2017], что выписанная оценка при $p = 2$ достигается в выпуклом случае и с точностью до множителя $\log\left(M_1M_2^2R^2\middle/\mu^3\right)$ в сильно выпуклом случае.

Несильно выпуклая часть оценки (кроме первых двух аргументов минимума) получается из сильно выпуклой с помощью регуляризации $\mu \simeq \varepsilon\middle/R^2$ [Васильев, 2011, гл. 9].

В [Arjevani et al., 2017; Nesterov, 2018] были получены нижние оценки для методов порядка p для выпуклых функций на число итераций, необходимых для достижения точности ε по функции:

$$\Omega\left(\left(\frac{M_pR^{p+1}}{\varepsilon}\right)^{2/(3p+1)}\right).$$

Покажем, как из этого факта получить оценки снизу для сильно выпуклых функций. Предположим, что есть метод порядка p для сильно выпуклых функций, который для достижения точности ε по функции требует меньше итераций, чем

$$O\left(\left(\frac{M_pR^{p-1}}{\mu}\right)^{2/(3p+1)}+\log\max\left\{1,\log\left(\frac{\left(\mu^{p+1}/M_p^2\right)^{1/(p-1)}}{\varepsilon}\right)\right\}\right).$$

Покажем противоречие с нижними оценками для не сильно выпуклого случая. Рассмотрим новую сильно выпуклую задачу с коэффициентом сильной выпуклости $\mu = \dfrac{\varepsilon}{R^2}$:

$$g(x) = f(x) + \frac{\varepsilon}{2R^2}\left\|x^0 - x\right\|_2^2 \to \min_{x\in\mathrm{R}^n}.$$

По $\varepsilon/2$ решению этой задачи можно восстановить решение исходной задачи. Действительно, если $g(x) - g_* \le \varepsilon/2$, то

$$f(x) - f_* \le g(x) - f_* \le g_* + \varepsilon/2 - f_* \le g(x_*) + \varepsilon/2 - f_* = \varepsilon/2 + \frac{\varepsilon}{2R^2}\left\|x^0 - x_*\right\|_2^2 = \varepsilon.$$

Применим данный метод для новой задачи. По предположению, для получения точности по функции $\varepsilon/2$ понадобится меньше итераций, чем

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

Пренебрегая вторым членом оценки, можно получить точность ε по функции для исходной задачи за меньшее число итераций, чем

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

Таким образом, получено противоречие с нижними оценками для не сильно выпуклого случая.

Поясним, как можно получить оценку сверху на число итераций в сильно выпуклом случае, предполагая, что верна оценка сверху в несильно выпуклом случае. Для этого рассмотрим метод Ньютона:

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

Считая, что $M_2 < \infty$, $\mu > 0$, получим

$$\left\|\nabla f\left(x^{k+1}\right)\right\|_2 = \left\|\nabla f\left(x^{k+1}\right) - \nabla f\left(x^k\right) - \nabla^2 f\left(x^k\right)\left(x^{k+1} - x^k\right)\right\|_2 \le M_2\left\|x^{k+1} - x^k\right\|_2^2 =$$
$$= M_2\left\|\left[\nabla^2 f\left(x^k\right)\right]^{-1}\nabla f\left(x^k\right)\right\|_2^2 \le \frac{M_2}{\mu^2}\left\|\nabla f\left(x^k\right)\right\|_2^2.$$

Заметим, что если последовательность положительных чисел $\left\{c^k\right\}_{k=0,1,2,\ldots}$ удовлетворяет условию

$$c^{k+1} \le \mathrm{const}\cdot\left(c^k\right)^\gamma, \quad \gamma > 1,$$

и c^0 достаточно мало, то после $N = \mathrm{O}\left(\log\log\left(c^0/\varepsilon\right)\right)$ итераций $c^N \le \varepsilon$. Для метода Ньютона $c_k = \left\|\nabla f\left(x^k\right)\right\|_2$, $\gamma = 2$; для класса чебышёвских методов высокого порядка $c_k = \left\|x^k - x_*\right\|_2$, $\gamma = 3, 4, 5, \ldots$ [Евтушенко, 2013, п. 2.9], [Карманов, 1986, п. 9.5.10]; для методов Ньютона с кубической регуляризацией $c_k = f\left(x^k\right) - f\left(x_*\right)$, $\gamma = 4/3$, причем в последнем случае сильную выпуклость можно заменить градиентным доминированием [Nesterov, Polyak, 2006].

С помощью неравенства (сильной выпуклости)

$$f(x) - f(x_*) \le \frac{1}{2\mu}\left\|\nabla f(x)\right\|_2^2$$

для метода Ньютона можно оценить окрестность *квадратичной скорости сходимости* метода:

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

Оказавшись в этой окрестности, можно достичь желаемой точности за $\log\log\left(\left(\mu^3/M_2^2\right)/\varepsilon\right)$ итераций.

Чтобы оказаться в этой окрестности, можно использовать технику рестартов (см. [Arjevani et al., 2017; Nesterov, 2008]), примененную к методам, обеспечивающим несильно выпуклую составляющую рассматриваемой оценки. Для данных методов имеем оценку числа итераций

$$k \le \text{const} \cdot \left(\frac{M_r \left\| x^0 - x_* \right\|^{r+1}}{f(x^k) - f(x_*)} \right)^{2/(3r+1)}, \qquad r = 1 \dots p,$$

откуда с учетом неравенства (сильной выпуклости)

$$\frac{\mu \left\| x^0 - x_* \right\|^2}{2} \le f(x_0) - f(x_*)$$

следует

$$f(x^k) - f(x_*) \le \frac{cM_r \left\| x^0 - x_* \right\|^{r+1}}{k^{(3r+1)/2}} \le \frac{2cM_r \left\| x^0 - x_* \right\|^{r-1}}{\mu k^{(3r+1)/2}} \left(f(x^0) - f(x_*) \right), \qquad r = 1 \dots p.$$

Произведя $N = \min\limits_{r=1\dots p} \left(\frac{4cM_r \left\| x^0 - x_* \right\|^{r-1}}{\mu} \right)^{\frac{2}{3r+1}}$ итераций метода, получим уменьшение невязки по функции

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

после чего с учетом $\left\| x^N - x_* \right\| \le \left\| x^0 - x_* \right\|$ можно рестартовать метод и получить следующую оценку невязки по функции:

$$f(x^n) - f(x_*) \le \text{const} \cdot \frac{f(x^0) - f(x_*)}{2^{n/N}},$$

откуда получаем оценку числа итераций, необходимых для достижения точности ε по функции:

$$O\left(N \log_2 \left(\frac{f(x^0) - f(x_*)}{\varepsilon} \right) \right).$$

Подставляя точность $\varepsilon = \frac{\mu^3}{2M_2^2}$, необходимую для попадания в окрестность квадратичной сходимости метода Ньютона, а также с учетом неравенства (липшицевости градиента)

$$f(x^0) - f(x_*) \le \frac{M_1 \left\| x^0 - x_* \right\|^2}{2}$$

получим окончательную оценку числа итераций, необходимых для попадания в окрестность квадратичной сходимости:

$$O\left(\min_{r=1\dots p} \left(\frac{M_r R^{r-1}}{\mu} \right)^{\frac{2}{3r+1}} \log_2 \left(\frac{M_1 M_2^2 R^2}{\mu^3} \right) \right).$$

Таким образом, получается вторая (сильно выпуклая) часть рассматриваемой оценки (с точностью до логарифмического множителя).

Стоит, однако, отметить, что если вместо сильной выпуклости и достаточной гладкости предполагать *самосогласованность* оптимизируемой функции, то, используя специальную локальную норму Дикина [Дикин, 2010] $\|u\|_x = \left\langle u, \nabla^2 f(x)\, u \right\rangle^{1/2}$, в классе методов 2-го порядка можно улучшить рассматриваемую оценку [Нестеров, 2010; Nemirovski, 2015]:

$$\mathrm{O}\left(f\left(x^0\right) - f(x_*) + \log\log(1/\varepsilon)\right).$$

Выше использовалось понятие самосогласованной функции. Поясним его.

Введем $g(t) = f(w + t\mathrm{v})$. Самосогласованность $f(x)$ означает, что для любых w и v справедливо неравенство $|g'''(t)| \le 2\,(g''(t))^{3/2}$ для всех t, причем множитель 2 здесь выбран для определенности. Отметим, что в описанном подходе константы имеют «физическую» размерность, поэтому в отличие от остальных формул не стоит пытаться проверять корректность формул в случае самосогласованной функции из соображений размерности [Зорич, 2017, гл. 1]. Если дополнительно предполагать, что оптимизируемая функция является ν-самосогласованным барьером (в интересных случаях удается конструктивно показать, что $\nu \le n$), то в классе методов 2-го порядка также можно достичь следующей оценки числа итераций (Нестеров – Немировский, 1988): $\mathrm{O}\left(\sqrt{\nu}\ln(\nu/\varepsilon)\right)$ [Нестеров, 2010; Nemirovski, 2015].

Обсуждение

Так же как и для методов 1-го порядка, для методов 2-го порядка и выше можно рассматривать их универсальные варианты [Grapiglia, Nesterov, 2017], можно рассматривать работу методов в условиях наличия шума и неточностей, возникающих при решении вспомогательных задач на каждой итерации [Baes, 2009; Ghadimi et al., 2017], также можно переносить и завязанные на наличие шумов конструкции, например конструкцию mini-batching'a [Ghadimi et al., 2017].

Однако при использовании методов 2-го порядка и выше появляется много новых вопросов относительно сильного проигрыша методам первого порядка (градиентного типа) по стоимости итерации и требуемой памяти. Так, для честного осуществления шага метода Ньютона необходимо обратить матрицу Гессе оптимизируемой функции в текущей точке. Эта задача по сложности эквивалентна задаче умножения двух матриц такого же по порядку размера [Кормен и др., 2002, гл. 31], что типично в n раз дороже, чем осуществление шага метода типа градиентного спуска (умножение матрицы на вектор).

Замечание. На самом деле это не так. Умножение двух матриц $n \times n$ современными алгоритмами может быть осуществлено за время $\mathrm{O}\left(n^{2.37}\right)$; см. [Разборов, 2016] и цитированную там литературу. Однако такого рода результаты проявляются только при очень больших значениях n.

В последнее время было предложено несколько подходов, имеющих своей целью хотя бы частичное устранение такого большого зазора в стоимости итерации между методами 1-го и 2-го порядка. Одна из идей активно используется в машинном обучении, когда функционал имеет вид суммы (среднего арифметического) большого числа однотипных слагаемых. Идея заключается в том, чтобы формировать матрицу Гессе оптимизируемой функции исходя из матриц Гессе относительно небольшого числа случайно выбранных слагаемых [Ghadimi et al., 2017]. Другая идея заключается в отказе от обращения матрицы Гессе на итерации, вместо этого предлагается использовать информацию о собственном векторе, отвечающем наименьшему собственному значению [Agarwal et al., 2017; Carmon et al., 2017]. Для приближенного вычисления такого вектора вполне достаточно уметь умножать матрицу Гессе на произвольный вектор:

$$\nabla^2 f(x)\,\mathrm{v} \approx \frac{\nabla f(x + \tau\mathrm{v}) - \nabla f(x)}{\tau},$$

что может быть сделано с помощью автоматического дифференцирования за то же по порядку время, что и вычисление градиента [Baydin et al., 2015; Nocedal, Wright, 2006]. Эта идея сейчас активно развивается в связи с поиском наиболее эффективных методов обучения глубоких нейронных сетей [Гудфеллоу и др., 2017].

Перспективной также представляется довольно старая идея спуска в область квадратичной сходимости с помощью методов типа градиентного спуска (с дешевыми итерациями) и последующая квадратичная сходимость с использованием, например, метода Ньютона. Проблема в таком подходе — детектирование момента попадания в нужную окрестность. В качестве возможного решения проблемы можно, например, действовать таким образом: через каждые $\sim \sqrt{n}$ итераций метода типа градиентного спуска проверять условие $\left\|\nabla f\left(x^k\right)\right\|_2 \ll 1$; если оно выполняется, то делать «пристрелочный» шаг метода Ньютона. Если в результате такого шага выполняется еще и условие $\left\|\nabla f\left(x^{k+1}\right)\right\|_2 \ll \left\|\nabla f\left(x^k\right)\right\|_2^{3/2}$, то нужно продолжать делать шаги метода Ньютона, каждый раз проверяя это условие. Если хотя бы одно из этих условий не выполняется, то следует вернуться к методу типа градиентного спуска. Можно показать, что при естественных условиях такой способ приводит к наилучшей по порядку оценке общего времени работы метода.

Список литературы (References)

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. — URL: https://arxiv.org/ftp/arxiv/papers/1701/1701.03821.pdf

Bayandina A. S., Gasnikov A. V., Lagunovskaya A. A. Bezgradiyentnyye dvukhtochechnyye metody resheniya zadach stokhasticheskoy negladkoy vypukloy optimizatsii pri nalichii malykh shumov ne sluchaynoy prirody [Gradient-less two-point methods for solving stochastic nonsmooth convex optimization problems in the presence of small non-random noises] // Automatics and telemechanics. — 2018. — URL: https://arxiv.org/ftp/arxiv/papers/1701/1701.03821.pdf (in Russian).

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

Vasiliev F. P. Metody optimizatsii [Optimization methods]. — Vol. 2. — Moscow: MCCME, 2011. — P. 433. (in Russian).

Воронцова Е. А., Гасников А. В., Горбунов Э. А. Ускоренные спуски по случайному направлению и безградиентные методы с неевклидовой прокс-структурой // Автоматика и телемеханика. — 2018. — URL: https://arxiv.org/pdf/1710.00162.pdf

Vorontsova E. A., Gasnikov A. V., Gorbunov E. A. Uskorennyye spuski po sluchaynomu napravleniyu i bezgradiyentnyye metody s neyevklidovoy proks-strukturoy [Accelerated descents in a random direction and gradientless methods with non-euclidean prox-structure] // Automatics and telemechanics. — 2018. — URL: https://arxiv.org/pdf/1710.00162.pdf (in Russian).

Гасников А. В. Эффективные численные методы поиска равновесий в больших транспортных сетях: диссертация на соискание ученой степени д. ф.-м. н. по специальности 05.13.18 — Математическое моделирование, численные методы, комплексы программ. — М.: МФТИ, 2016. — 487 с.

Gasnikov A. V. Effektivnyye chislennyye metody poiska ravnovesiya v bol'shikh transportnykh setyakh: dissertatsiya na soiskaniye uchenoy stepeni d. f.-m. n. po spetsial'nosti 05.13.18 [Effective numerical methods for finding equilibrium in large transport networks: thesis for PhD on the specialty 05.13.18] — Matematicheskoye modelirovaniye, chislennyye metody, kompleksy programm [Mathematical modeling, numerical methods, program complexes]. — Moscow: MFTI, 2016. — 487 p. (in Russian).

Гудфеллоу Я., Бенджио И., Курвиль А. Глубокое обучение. — ДМК Пресс, 2017. — 652 с.

Goodfellow Ya., Bengio I., Courville A. Glubokoye obucheniye [Deep Learning]. — DMK Press, 2017. — 652 p. (in Russian).

Дикин И. И. Метод внутренних точек в линейном и нелинейном программировании. — М.: КРАСАНД, 2010. — 120 с.

Dikin I. Metod vnutrennikh tochek v lineynom i nelineynom programmirovanii [Interior point methods in linear and nonlinear programming]. — Moscow: KRASAND, 2010. — 120 p. (in Russian).

Зорич В. А. Математический анализ задач естествознания. — М.: МЦНМО, 2017. — 160 с.
Zorich V. A. Matematicheskiy analiz zadach yestestvoznaniya [Mathematical analysis of problems in the natural sciences]. — Moscow: MCCME, 2017. — 160 p. (in Russian).

Евтушенко Ю. Г. Оптимизация и быстрое автоматическое дифференцирование. — М.: ВЦ РАН, 2013. — 144 с.
Evtushenko Yu. G. Optimizatsiya i bystroye avtomaticheskoye differentsirovaniye [Evtushenko. Optimization and fast automatic differentiation]. — Moscow: CC RAS, 2013. — 144 p. (in Russian).

Карманов В. Г. Математическое программирование. — М.: Наука, 1986. — 288 с.
Karmanov V. G. Matematicheskoye programmirovaniye [Mathematical programming]. — M.: Science, 1986. — 288 p. (in Russian).

Кормен Т., Лейзерсон Ч., Ривест Р. Алгоритмы: построение и анализ. — М.: МЦНМО, 2002. — 960 с.
Cormen T., Leiserson C., Rivest R. Algoritmy: postroyeniye i analiz [Introduction to Algorithms]. — Moscow: MCCME, 2002. — 960 p. (in Russian).

Немировский А. С., Юдин Д. Б. Сложность задач и эффективность методов оптимизации. — М.: Наука, 1979. — 384 с.
Nemirovsky A. S., Yudin D. B. Slozhnost' zadach i effektivnost' metodov optimizatsii [Problem complexity and method efficiency in optimization]. — Moscow: Science, 1979. — 384 p. (in Russian).

Нестеров Ю. Е. Введение в выпуклую оптимизацию. — М.: МЦНМО, 2010. — 262 с.
Nesterov Yu. E. Vvedeniye v vypukluyu optimizatsiyu [Introductory lectures on convex optimization]. — Moscow: MCCME, 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 vypukloy funktsii po yeye znacheniyam [On the question of algorithms for the approximate calculation of the minimum of a convex function from its values] // Mat. zametki [Math. notes]. — 1996. — Vol. 59, No. 1. — P. 95–102. (in Russian).

Разборов А. А. Алгебраическая сложность. — М.: МЦНМО, 2016. — 32 с.
Razborov A. A. Algebraicheskaya slozhnost' [Algebraic complexity]. — Moscow: MCCME, 2016. — 32 p. (in Russian).

Agarwal N., Allen-Zhu Z., Bullins B., Hazan E., Ma T. Finding approximate local minima faster than gradient descent // In Proceedings of the Forty-Ninth Annual ACM Symposium on the Theory of Computing, 2017.

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

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

Baydin A. G., Pearlmutter B. A., Radul A. A., Siskand J. M. Automatic differentiation in machine learning: a survey // e-print, 2015. — URL: https://arxiv.org/pdf/1502.05767.pdf

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

Carmon Y., Duchi J. C., Hinder O., Sidford A. Accelerated methods for non-convex optimization // e-print, 2017. — URL: https://arxiv.org/pdf/1611.00756.pdf

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) // SIAM J. Optim. — 2017 (Submitted). — URL: https://arxiv.org/pdf/1707.08486.pdf

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

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

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

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

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 optmization // CORE Discussion Papers 2018005. — 2018. — URL: https://ideas.repec.org/p/cor/louvco/2018005.html

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

Nesterov Yu., Polyak P. Cubic regularization of Newton method and its global performance // Math. Program. Ser. A. — 2006. — Vol. 108. — 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.