Оптимизация, системный анализ и исследование операций

© 2019 г. Е.А. ВОРОНЦОВА, канд. физ.-мат. наук (vorontsovaea@gmail.com)
(Дальневосточный федеральный университет, Владивосток),
А.В. ГАСНИКОВ, д-р физ.-мат. наук (gasnikov@yandex.ru),
Э.А. ГОРБУНОВ (ed-gorbunov@yandex.ru)
(Московский физико-технический институт)

УСКОРЕННЫЙ СПУСК ПО СЛУЧАЙНОМУ НАПРАВЛЕНИЮ С НЕЕВКЛИДОВОЙ ПРОКС-СТРУКТУРОЙ[1]

Рассматриваются задачи гладкой выпуклой оптимизации, для численного решения которых полный градиент недоступен. В 2011 г. Ю.Е. Нестеровым были предложены ускоренные безградиентные методы решения таких задач. Поскольку рассматривались только задачи безусловной оптимизации, то использовалась евклидова прокс-структура. Однако если заранее знать, например, что решение задачи разреженно, а точнее, что расстояние от точки старта до решения в 1-норме и в 2-норме близки, то более выгодно выбирать не евклидову прокс-структуру, связанную с 2-нормой, а прокс-структуру, связанную с 1-нормой. Полное обоснование этого утверждения проводится в статье. Предлагается ускоренный метод спуска по случайному направлению с неевклидовой прокс-структурой для решения задачи безусловной оптимизации (в дальнейшем подход предполагается расширить на ускоренный безградиентный метод). Получены оценки скорости сходимости метода. Показаны сложности переноса описанного подхода на задачи условной оптимизации.

Ключевые слова: ускоренные методы первого порядка, выпуклая оптимизация, метод линейного каплинга, концентрация равномерной меры на единичной евклидовой сфере, неевклидова прокс-структура.

DOI: 10.1134/S000523101904007X

1. Введение

В [1] были предложены ускоренные оракульные[2] методы нулевого порядка (безградиентные методы) решения задач гладкой выпуклой безусловной оптимизации.

[1] Работа А.В. Гасникова по разделам 3 и 4 поддержана Российским фондом фундаментальных исследований (проект № 18-31-20005 мол_а_вед). Работа Э.А. Горбунова и Е.А. Воронцовой поддержана грантом Президента РФ МД-1320.2018.1. Работа А.В. Гасникова и Е.А. Воронцовой поддержана Российским фондом фундаментальных исследований (проект № 18-29-03071 мк).

[2] Здесь и далее под оракулом понимается подпрограмма расчета значений целевой функции и/или градиента (его части), а оптимальность метода на классе задач понимается в смысле Бахвалова–Немировского [2] как число обращений (по ходу работы метода) к оракулу для достижения заданной точности (по функции).

В рассуждениях [1] существенным образом использовалось то, что была выбрана евклидова прокс-структура (выпуклая гладкая функция, порождающая расстояние, и 1-сильно выпуклая относительно какой-то нормы (строгое определение см. в разделе 4)). Такой выбор прокс-структуры для задач безусловной оптимизации является вполне естественным (см., например, [3]). Однако в ряде задач имеется дополнительная информация, которая, например, позволяет рассчитывать на разреженность решения (в решении большая часть компонент нулевые). В таких случаях использование других прокс-структур бывает более выгодным. Для негладких задач стохастической условной оптимизации с оракулом нулевого порядка недавно было показано (см. [4, 5]), что в определенных ситуациях ускорение метода за счет перехода от евклидовой прокс-структуры, связанной с 2-нормой, к прокс-структурам, связанным с 1-нормой, может давать ускорение методу, по порядку равное размерности пространства, в котором происходит оптимизация. К сожалению, техника, использованная в [4, 5] существенным образом использовала неускоренную природу оптимальных методов для негладких задач. Другими словами, из [4, 5] непонятно, как получать аналогичные оценки для гладких задач. В настоящей статье на базе специального варианта быстрого (ускоренного) градиентного метода [6] строится ускоренный метод спуска по случайному направлению. Особенностью метода из [6] является представление быстрого градиентного метода как специальной выпуклой комбинации градиентного спуска и зеркального спуска. В [6], как и во всех известных авторам вариантах быстрого градиентного метода с двумя и более "проекциями", обе проекции осуществлялись в одной норме/прокс-структуре. Главной идеей настоящей статьи является использование разных норм/прокс-структур в этих проекциях, а именно: в градиентном шаге всегда используется обычная евклидова проекция, а вот в зеркальном шаге выбор прокс-структуры обусловлен априорной информацией о свойствах решения.

В настоящей статье на базе описанной конструкции для детерминированных задач безусловной гладкой выпуклой оптимизации строится ускоренный метод спуска по случайному направлению[3] (раздел 4).

В классе детерминированных спусков по направлению (к ним можно отнести и циклический координатный спуск) для получения лучших оценок необходимо вводить рандомизацию (доказательство см. в. [8]), поэтому рассматриваются сразу спуски по случайному направлению.

Показано, какие возникают сложности при попытке перенесения описанного подхода на задачи оптимизации на множествах простой структуры (раздел 4).

2. Постановка задачи

Рассматривается задача гладкой выпуклой оптимизации

$$f(x) \to \min_{x \in Q}, \tag{1}$$

[3] Подробно о разнице в подходах в случае детерминированной постановки, но с введением рандомизации и в случае задач стохастической оптимизации см. в [7].

где функция $f(x)$, заданная на выпуклом замкнутом множестве $Q \subseteq \mathbb{R}^n$, имеет липшицев градиент с константой L_2 (т.е. $f(x)$ — L_2-гладкая функция)

$$\|\nabla f(y) - \nabla f(x)\|_2 \leqslant L_2 \|y - x\|_2$$

и является μ-сильно выпуклой в p-норме ($1 \leqslant p \leqslant 2$) функцией (далее будем использовать и обозначение μ_p), при этом в точке минимума x_* выполнено равенство $\nabla f(x_*) = 0$, а итерационный процесс стартует с точки x_0.

В данной статье для решения задачи (1) вместо обычного градиента используется его стохастическая аппроксимация, построенная на базе производной по случайно выбранному направлению [9]

$$g(x, e) \stackrel{\text{def}}{=} n \langle \nabla f(x), e \rangle e,$$

где e — случайный вектор, равномерно распределенный на $S_2^n(1)$ — единичной сфере в 2-норме в пространстве $\mathbb{R}^n$ ($e \sim RS_2^n(1)$; под этой записью будем понимать, что случайный вектор e имеет равномерное распределение на n-мерной единичной евклидовой сфере), а угловые скобки $\langle \cdot, \cdot \rangle$ обозначают скалярное произведение[4].

Имеет место следующая лемма (доказательство см. в [10]), являющаяся следствием явления концентрации равномерной меры на сфере вокруг экватора (см. также [11]; северный полюс задается градиентом $\nabla f(x)$).

Лемма 1. Пусть $e \sim RS_2^n(1)$, $n \geqslant 8$, $s \in \mathbb{R}^n$, тогда

$$\mathbb{E}\left[\|e\|_q^2\right] \leqslant \min\{q - 1,\ 16 \ln n - 8\}\, n^{\frac{2}{q}-1}, \quad 2 \leqslant q \leqslant \infty, \tag{2}$$

$$\mathbb{E}\left[\langle s, e\rangle^2 \|e\|_q^2\right] \leqslant \sqrt{3}||s||_2^2 \min\{2q - 1,\ 32 \ln n - 8\}\, n^{\frac{2}{q}-2}, \quad 2 \leqslant q \leqslant \infty, \tag{3}$$

где под знаком $||\cdot||_q$ понимается векторная q-норма (норма Гельдера с показателем q).

Из (3) (см. также [4]) вытекает следующий факт.

Утверждение. Пусть $e \sim RS_2^n(1)$ и $g(x, e) = n \langle \nabla f(x), e \rangle e$, тогда

$$\mathbb{E}_e\left[\|g(x, e)\|_q^2\right] \leqslant \sqrt{3}\min\{2q - 1,\ 32 \ln n - 8\} n^{\frac{2}{q}} \|\nabla f(x)\|_2^2.$$

Используя данное утверждение, можно в сильно выпуклом случае при условии $\nabla f(x_*) = 0$ получить оценку необходимого числа обращений к оракулу за производной по направлению для достижения по функции точности ε в среднем [12]:

$$N(\varepsilon) = \mathrm{O}\left(n^{\frac{2}{q}} \ln n \frac{L_2}{\mu_p} \ln\left(\frac{\Delta f^0}{\varepsilon}\right)\right),$$

где $\Delta f^0 = f(x_0) - f(x_*)$.

[4] Отметим, что $\mathbb{E}_e[g(x,e)] = \nabla f(x)$, что следует из факта: $\mathbb{E}_e[|e_i|^2] = \frac{1}{n}$, где e_i — i-я компонента вектора e.

Имеется гипотеза (см., например, [3]), что, используя специальные ускоренные методы (типа Катюши (Katusha) из [13]), можно получить оценку

$$N(\varepsilon) = \mathrm{O}\left(n^{\frac{1}{q}+\frac{1}{2}} \ln n \sqrt{\frac{L_2}{\mu}} \ln\left(\frac{\Delta f^0}{\varepsilon}\right)\right).$$

Насколько известно авторам, эта гипотеза на данный момент не доказана и не опровергнута.

В разделе 4 данной статьи для случая $Q = \mathbb{R}^n$ доказывается оценка

$$\mathbb{E}[f(x_N)] - f(x_*) \leqslant C n^{\frac{2}{q}+1} \ln n \frac{L_2 R^2}{N^2}. \tag{4}$$

Эта оценка получается из приведенной выше оценки с помощью регуляризации $\mu_p = \varepsilon/R^2$ [7], где (с точностью до корня из логарифмического по n множителя) R — расстояние в p-норме от точки старта до решения.

3. Задача А.С. Немировского

Рассмотрим задачу (1) минимизации гладкого выпуклого функционала $f(x)$ с константой Липшица градиента L_2 в 2-норме на множестве $Q = B_1^n(R)$ (шар в пространстве $\mathbb{R}^n$ радиуса R в 1-норме). Тогда на рассматриваемом классе функции для любого итерационного метода, на каждой итерации которого только один раз можно обратиться к оракулу за градиентом функции, можно так подобрать функцию $f(x)$ из этого класса, что имеет место оценка скорости сходимости [14] в виде

$$f(x_N) - f(x_*) \geqslant \frac{\tilde{C}_1 L_2 R^2}{N^3}, \tag{5}$$

где $\tilde{C}_1$ — некоторая числовая константа[5], x^* — ближайшая к x_0 точка минимума функции $f(x)$. Поскольку $f(x_N) - f(x_*)$ должно быть меньше или равно ε, из (5) можно получить оценку на $N(\varepsilon)$.

С другой стороны, если использовать обычный быстрый градиентный метод с KL-прокс-структурой для этой же задачи, то [15]:

$$f(x_N) - f(x_*) \leqslant \frac{\tilde{C}_2 L_1 R^2 \ln n}{N^2},$$

где константа Липшица градиента L_1 в 1-норме удовлетворяет условию:

$$L_2/n \leqslant L_1 \leqslant {L_2}^6.$$

Отсюда нельзя сделать вывод, что нижняя оценка достигается. Достигается ли эта нижняя оценка и если достигается, то на каком методе? Насколько

[5] Как и $\tilde{C}_2$ далее. Здесь и далее все числовые константы не зависят от N и n.

[6] Здесь под L_1 понимается такое положительное число, что $||\nabla f(x) - \nabla f(y)||_\infty \leqslant \leqslant L_1 ||x - y||_1$.

авторам известно, это пока открытый вопрос, поставленный А.С. Немировским в 2015 г. (см. также [14]). Однако если оценивать не число итераций, а общее число арифметических операций и если ограничиться рассмотрением класса функций, для которых "стоимость" расчета производной по направлению примерно в n раз меньше "стоимости" расчета полного градиента[7], то при $N \leqslant n$ (точнее даже при $N \simeq n$) выписанная оценка (4) (в варианте для общего числа арифметических операций, необходимых для достижения заданной точности в среднем) с точностью до логарифмического множителя будет соответствовать нижней оценке (5), если последнюю понимать как

$$f\left(x_N\right) - f(x_*) \geqslant \frac{\tilde{C}_1 L_2 R^2}{nN^2},$$

т.е.

$$N\left(\varepsilon\right) = \mathrm{O}\left(\sqrt{\frac{L_2 R^2}{\varepsilon n}}\right).$$

Действительно ($q = \infty$), общее число арифметических операций равно

$$\mathrm{O}\left(n\right) \cdot \underbrace{\mathrm{O}\left(n^{\frac{1}{2}} \ln n \sqrt{\frac{L_2 R^2}{\varepsilon}}\right)}_{\text{число итераций}} \approx \mathrm{O}\left(n^2\right) \cdot \mathrm{O}\left(\sqrt{\frac{L_2 R^2}{\varepsilon n}}\right).$$

4. Обоснование формулы (4) в случае $Q = \mathbb{R}^n$

Введем дивергенцию Брэгмана [17] $V_z\left(y\right)$, связанную с p-нормой $(1 \leqslant p \leqslant 2)$

$$V_z(y) \stackrel{\text{def}}{=} d(y) - d(z) - \langle \nabla d(z),\, y - z \rangle,$$

где функция $d(x)$ является непрерывно дифференцируемой и сильно выпуклой с константой сильной выпуклости, равной единице. Например, для $p = 1$ функцию $d(x)$ можно выбрать так:

$$d(x) = \frac{1}{2(a-1)} ||x||_a^2,$$

где

$$a = \frac{2 \log n}{2 \log n - 1}.$$

[7] Из-за быстрого автоматического дифференцирования [16] это предположение довольно обременительное; но если функция задана моделью черного ящика, выдающего только значение функции, а градиент восстанавливается при $n+1$ таком обращении, то сделанное предположение кажется вполне естественным.

Функцию $d(x)$ будем называть *прокс-функцией* (или *прокс-структурой*), связанной с p-нормой. Кроме того, пусть q — такое число, что $\frac{1}{p}+\frac{1}{q}=1$. Далее будем следовать обозначениям из [12]. Пусть случайный вектор e равномерно распределен на поверхности евклидовой сферы единичного радиуса ($e \sim RS_2^n(1)$). Положим, что

$$\mathrm{Grad}_e(x) \stackrel{\mathrm{def}}{=} x - \frac{1}{L}\langle \nabla f(x), e\rangle e,$$

$$\mathrm{Mirr}_e(x,z,\alpha) \stackrel{\mathrm{def}}{=} \operatorname*{argmin}_{y\in\mathbb{R}^n}\{\alpha\langle n\langle \nabla f(x), e\rangle e, y-z\rangle + V_z(y)\},$$

где L — константа Липшица градиента функции $f(x)$ в 2-норме (индекс 2 не пишем, так как везде далее интересуемся константой Липшица в 2-норме).

Опишем ускоренный неевклидов спуск (английское название метода — Accelerated by Coupling Directional Search, ACDS), построенный на базе специальной комбинации спусков по направлению в форме градиентного спуска (Grad) и метода зеркального спуска (Mirr).

Алгоритм 1. Ускоренный неевклидов спуск (ACDS)

Вход: f — выпуклая дифференцируемая функция на $\mathbb{R}^n$ с липшицевым градиентом с константой L по отношению к 2-норме; x_0 — некоторая стартовая точка; N — количество итераций.

Выход: точка y_N, для которой выполняется $\mathbb{E}_{e_1,e_2,\ldots,e_N}[f(y_N)] - f(x^*) \leqslant \frac{4\Theta L C_{n,q}}{N^2}$.

$y_0 \leftarrow x_0$, $z_0 \leftarrow x_0$
for $k = 0, \ldots, N-1$ **do**
  $\alpha_{k+1} \leftarrow \frac{k+2}{2LC_{n,q}}$, $\tau_k \leftarrow \frac{1}{\alpha_{k+1}LC_{n,q}} = \frac{2}{k+2}$
  Генерируется $e_{k+1} \sim RS_2^n(1)$ независимо от предыдущих итераций
  $x_{k+1} \leftarrow \tau_k z_k + (1-\tau_k)y_k$
  $y_{k+1} \leftarrow \mathrm{Grad}_{e_{k+1}}(x_{k+1})$
  $z_{k+1} \leftarrow \mathrm{Mirr}_{e_{k+1}}(x_{k+1}, z_k, \alpha_{k+1})$
end for
return y_N

Теорема. Пусть $f(x)$ — выпуклая дифференцируемая функция на $Q=\mathbb{R}^n$ с константой Липшица для градиента, равной L в 2-норме, $d(x)$ — 1-сильно выпуклая в p-норме функция на Q, N – число итераций метода, x^* — точка минимума функции $f(x)$. Тогда ускоренный неевклидов спуск (ACDS) на выходе даст точку y_N, удовлетворяющую неравенству

$$\mathbb{E}_{e_1,e_2,\ldots,e_N}[f(y_N)] - f(x^*) \leqslant \frac{4\Theta L C_{n,q}}{N^2},$$

где

$$\Theta \stackrel{\mathrm{def}}{=} V_{x_0}(x^*), \quad C_{n,q} \stackrel{\mathrm{def}}{=} \sqrt{3}\min\{2q-1,\ 32\ln n - 8\}n^{\frac{2}{q}+1}, \quad \frac{1}{q}+\frac{1}{p}=1.$$

Сформулируем две ключевые леммы, которые понадобятся для доказательства теоремы (доказательства приведены в Приложении).

Лемма 2. Если $\tau_k = \frac{1}{\alpha_{k+1} L C_{n,q}}$, *то для всех* $u \in Q = \mathbb{R}^n$ *верны неравенства*

$$
\begin{gathered}
\alpha_{k+1}\langle \nabla f(x_{k+1}), z_k - u\rangle \leqslant \\
\leqslant \alpha_{k+1}^2 \cdot \frac{C_{n,q}}{2n} ||\nabla f(x_{k+1})||_2^2 + V_{z_k}(u) - \\
- \mathbb{E}_{e_{k+1}}[V_{z_{k+1}}(u) \mid e_1, e_2, \ldots, e_k] \leqslant \\
\leqslant \alpha_{k+1}^2 L C_{n,q} \cdot (f(x_{k+1}) - \mathbb{E}_{e_{k+1}}[f(y_{k+1}) \mid e_1, e_2, \ldots, e_k]) + \\
+ V_{z_k}(u) - \mathbb{E}_{e_{k+1}}[V_{z_{k+1}}(u) \mid e_1, e_2, \ldots, e_k].
\end{gathered} \tag{6}
$$

Лемма 3. Для всех $u \in Q = \mathbb{R}^n$ *выполнено неравенство*

$$
\begin{gathered}
\alpha_{k+1}^2 L C_n \mathbb{E}_{e_{k+1}}[f(y_{k+1}) \mid e_1, \ldots, e_k] - (\alpha_{k+1}^2 L C_n - \alpha_{k+1}) f(y_k) + \\
+ \mathbb{E}_{e_{k+1}}[V_{z_{k+1}}(u) \mid e_1, \ldots, e_k] - V_{z_k}(u) \leqslant \alpha_{k+1} f(u).
\end{gathered} \tag{7}
$$

Сложности возникают при попытке перенесения этого результата на случай $Q \neq \mathbb{R}^n$. Ограничимся рассмотрением случая $p = q = 2$, так как даже для него не удается обобщить рассуждения из [6]. Введем обозначение

$$
\mathrm{Prog}_s(x) \stackrel{\text{def}}{=} -\min_{y \in Q} \left\{ \langle s, y - x\rangle + \frac{L}{2} \|y - x\|_2^2 \right\}.
$$

Заметим, что вторая часть леммы 1 в евклидовом случае упрощается, а именно

$$
\mathbb{E}_e[\langle s, e\rangle^2] = \frac{||s||_2^2}{n},
$$

где $e \sim RS_2^n$ (см. лемму B.10 из [18]; формулировка данной леммы из [18] есть в Приложении — см. лемму (П.1)). Отсюда следует, что $C_{n,q} = n^2$. Если положить

$$
\mathrm{Grad}_e(x) \stackrel{\text{def}}{=} \operatorname*{argmin}_{y \in Q} \left\{ \langle \langle \nabla f(x), e\rangle e, x - y\rangle - \frac{L}{2} \|y - x\|_2^2 \right\}
$$

и

$$
\mathrm{Mirr}_e(x, z, \alpha) \stackrel{\text{def}}{=} \operatorname*{argmin}_{y \in Q} \{\alpha \langle n \langle \nabla f(x), e\rangle e, y - z\rangle + V_z(y)\},
$$

то чтобы обобщить приведенные рассуждения на условный случай, нужно оценить подходящим образом $\mathrm{Prog}_{n\langle \nabla f(x), e\rangle e}(x_{k+1})$ (точнее, его математическое ожидание по e_{k+1}), т.е., исходя из техники, используемой в [6], хотелось бы доказать оценку

$$
\mathbb{E}_{e_{k+1}}\left[\mathrm{Prog}_{n\langle \nabla f(x_{k+1}), e_{k+1}\rangle e_{k+1}}(x_{k+1})\right] \leqslant n^2 (f(x_{k+1}) - \mathbb{E}[f(y_{k+1})]), \tag{8}
$$

чтобы получить оценку скорости сходимости, как и в случае безусловной минимизации. К сожалению, существует пример (будет приведен далее) выпуклой L-гладкой функции и замкнутого выпуклого множества, для которых (8) не выполняется.

Сначала рассмотрим более детально $\mathrm{Prog}_\xi(x)$:

$$\mathrm{Prog}_\xi(x) = -\min_{y\in Q}\left\{\frac{L}{2}\Big|\Big|y-x\Big|\Big|_2^2 + \langle \xi,\, y-x\rangle\right\} =$$

$$= -\min_{y\in Q}\left\{\frac{L}{2}\Big|\Big|y-x\Big|\Big|_2^2 + \langle \xi,\, y-x\rangle + \frac{1}{2L}\Big|\Big|\xi\Big|\Big|_2^2\right\} + \frac{1}{2L}\Big|\Big|\xi\Big|\Big|_2^2 =$$

$$= -\min_{y\in Q}\left\{\Big|\Big|\frac{1}{\sqrt{2L}}\xi + \sqrt{\frac{L}{2}}\cdot y - \sqrt{\frac{L}{2}}\cdot x\Big|\Big|_2^2\right\} + \frac{1}{2L}\Big|\Big|\xi\Big|\Big|_2^2 =$$

$$= -\frac{L}{2}\min_{y\in Q}\left\{\Big|\Big|y - \left(x - \frac{1}{L}\xi\right)\Big|\Big|_2^2\right\} + \frac{1}{2L}\Big|\Big|\xi\Big|\Big|_2^2,$$

т.е. точка, в которой достигается этот минимум[8],

$$\hat{y} = \pi_Q\left(x - \frac{1}{L}\xi\right).$$

Тогда

$$y_{k+1} = \pi_Q\left(x - \frac{1}{L}s_{k+1}\right), \quad s_{k+1} \overset{\text{def}}{=} \langle \nabla f(x_{k+1}),\, e_{k+1}\rangle\, e_{k+1} = \frac{1}{n}g(x_{k+1}, e_{k+1}).$$

Кроме того, обозначим через $\widetilde{y}_{k+1}$ точку множества Q, в которой достигается минимум в формуле для $\mathrm{Prog}_{n\langle \nabla f(x_{k+1}),\, e_{k+1}\rangle\, e_{k+1}}(x_{k+1})$. Тогда

$$\widetilde{y}_{k+1} = \pi_Q\left(x - \frac{n}{L}s_{k+1}\right).$$

Также для удобства рассмотрим следующие представления для y_{k+1} и $\widetilde{y}_{k+1}$:

$$\begin{aligned} y_{k+1} &= x_{k+1} - \frac{1}{L}s_{k+1} + r_{k+1}, \\ \widetilde{y}_{k+1} &= x_{k+1} - \frac{n}{L}s_{k+1} + \widetilde{r}_{k+1}, \end{aligned} \tag{9}$$

где r_{k+1} и $\widetilde{r}_{k+1}$ будем называть векторами невязок.

Рассмотрим функцию

$$f(y) = f(x_{k+1}) + \langle \nabla f(x_{k+1}),\, y - x_{k+1}\rangle + \frac{L}{2}\Big|\Big|y - x_{k+1}\Big|\Big|_2^2 \tag{10}$$

и множество, изображенное на рис. 1 (в качестве $\nabla f(x_{k+1})$ можно выбрать любой ненулевой вектор, а в качестве Q — прямоугольный параллелепипед

[8] См. [7].

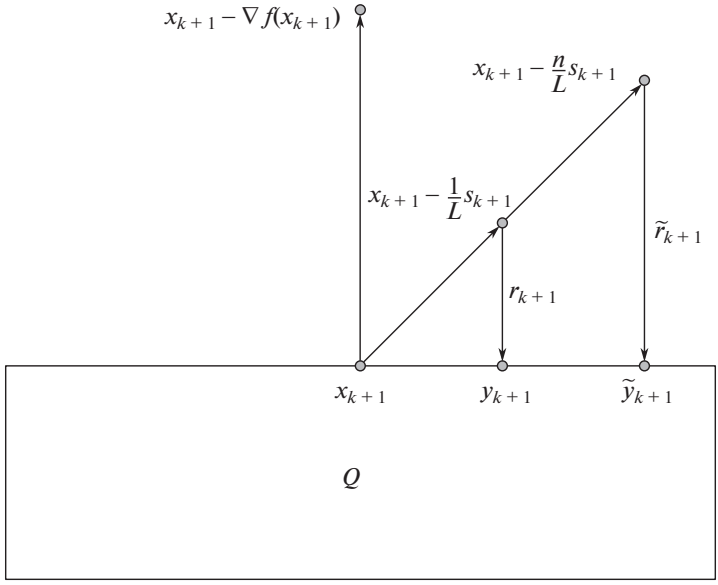

Рис. 1. Пример ситуации, когда ключевое неравенство не выполнено.

с достаточно длинными сторонами, в центре одной из гиперграней которого размещена точка x_{k+1}).

Подставим в (10) значение $y = y_{k+1}$ и воспользуемся представлением y_{k+1} из (9):

$$-\left\langle \nabla f(x_{k+1}),\, -\frac{1}{L} s_{k+1} + r_{k+1} \right\rangle - \frac{L}{2}\left|\left| r_{k+1} - \frac{1}{L} s_{k+1} \right|\right|_2^2 = f(x_{k+1}) - f(y_{k+1}).$$

Далее воспользуемся тем, что $s_{k+1} = \langle \nabla f(x_{k+1}),\, e_{k+1} \rangle\, e_{k+1}$:

$$\frac{1}{L}\langle \nabla f(x_{k+1}),\, e_{k+1}\rangle^2 - \langle \nabla f(x_{k+1}),\, r_{k+1}\rangle - \frac{L}{2}\left|\left| r_{k+1} \right|\right|_2^2 + \langle r_{k+1},\, s_{k+1}\rangle - \\ - \frac{1}{2L}\langle \nabla f(x_{k+1}),\, e_{k+1}\rangle^2 = f(x_{k+1}) - f(y_{k+1}),$$

или в более компактной форме

$$\frac{1}{2L}\langle \nabla f(x_{k+1}),\, e_{k+1}\rangle^2 - \frac{L}{2}\left|\left| r_{k+1} \right|\right|_2^2 + \langle r_{k+1},\, s_{k+1} - \nabla f(x_{k+1})\rangle = \\ = f(x_{k+1}) - f(y_{k+1}). \tag{11}$$

При таком выборе функции и множества получаем, что $n^2 \cdot ||r_{k+1}||_2^2 = ||\tilde{r}_{k+1}||_2^2$ для всех единичных e. Действительно, если

$$\mathrm{Prog}_{\langle \nabla f(x_{k+1}),\, e_{k+1}\rangle\, e_{k+1}}(x_{k+1}) = \frac{1}{2L}\langle \nabla f(x_{k+1}),\, e_{k+1}\rangle^2 - \frac{L}{2}\left|\left| r_{k+1} \right|\right|_2^2$$

и

$$\mathrm{Prog}_{n\langle \nabla f(x_{k+1}),\, e_{k+1}\rangle\, e_{k+1}}(x_{k+1}) = \frac{n^2}{2L}\langle \nabla f(x_{k+1}),\, e_{k+1}\rangle^2 - \frac{L}{2}\left|\left| \tilde{r}_{k+1} \right|\right|_2^2,$$

то

$$\mathrm{Prog}_{n\langle \nabla f(x_{k+1}),\, e_{k+1}\rangle\, e_{k+1}}(x_{k+1}) = n^2\, \mathrm{Prog}_{\langle \nabla f(x_{k+1}),\, e_{k+1}\rangle\, e_{k+1}}(x_{k+1}).$$

Отсюда и из (11) следует, что

$$\frac{1}{n^2}\mathrm{Prog}_{n\langle \nabla f(x_{k+1}),\, e_{k+1}\rangle\, e_{k+1}}(x_{k+1}) + \langle r_{k+1},\, s_{k+1} - \nabla f(x_{k+1})\rangle = \\ = f(x_{k+1}) - f(y_{k+1}).$$

Заметим, что вектор s_{k+1} всегда короче (точнее, не длиннее) вектора $\nabla f(x_{k+1})$ и направлен "вниз" (т.е. в то же полупространство, образованное гранью Q, на которой лежит точка x_{k+1}), как и $\nabla f(x_{k+1})$. Значит, разность $s_{k+1} - \nabla f(x_{k+1})$ будет направлена в противоположную часть пространства. А вектор r_{k+1} тоже направлен вниз. Следовательно, всегда выполняется $\langle r_{k+1},\, s_{k+1} - \nabla f(x_{k+1})\rangle \leqslant 0$, причем с ненулевой вероятностью выполнено строгое неравенство. Это означает, что

$$\mathbb{E}_{e_{k+1}}\left[\langle r_{k+1},\, s_{k+1} - \nabla f(x_{k+1})\rangle\right] < 0.$$

Поэтому

$$\mathbb{E}_{e_{k+1}}\left[\mathrm{Prog}_{n\langle \nabla f(x_{k+1}),\, e_{k+1}\rangle\, e_{k+1}}(x_{k+1})\right] = \\ = n^2(f(x_{k+1}) - \mathbb{E}_{e_{k+1}}[f(y_{k+1})]) - \mathbb{E}_{e_{k+1}}\left[\langle r_{k+1},\, s_{k+1} - \nabla f(x_{k+1})\rangle\right] > \\ > n^2\left(f(x_{k+1}) - \mathbb{E}_{e_{k+1}}\left[f(y_{k+1})\right]\right).$$

Представленный контр-пример показывает трудности в перенесении предлагаемого в статье метода на задачи условной оптимизации. Теорема утверждает, что ускоренный неевклидов спуск (алгоритм ACDS) через N итераций выдаст точку y_N, удовлетворяющую неравенству $\mathbb{E}[f(y_N)] - f(x^*) \leqslant \varepsilon$, $\varepsilon > 0$, если $N = O\left(\sqrt{\frac{\Theta L_2 C_{n,q}}{\varepsilon}}\right)$. По определению $C_{n,q} = \sqrt{3}\min\{2q-1,\, 32\ln n - 8\}n^{\frac{2}{q}+1}$, а в случае $p = q = 2$ можно взять $C_{n,q} = n^2$, что видно из леммы 1 и леммы B.10 из [18] (приведена в Приложении как лемма (П.1)), поэтому $N = O\left(\sqrt{\frac{\Theta L_2 n^2}{\varepsilon}}\right)$. Если же $p = 1$ и $q = \infty$, то $C_{n,q} = n\sqrt{3}\,(32\ln n - 8)$ и $N = O\left(\sqrt{\frac{\Theta L_2 n \ln n}{\varepsilon}}\right)$.

5. Численные эксперименты

Для практического применения предложенный ускоренный неевклидов спуск по случайному направлению ACDS был реализован на языке программирования Python. Код метода и демонстрация вычислительных свойств метода с построением графиков сходимости доступны как Jupyter Notebook и выложены в свободном доступе на Github [19].

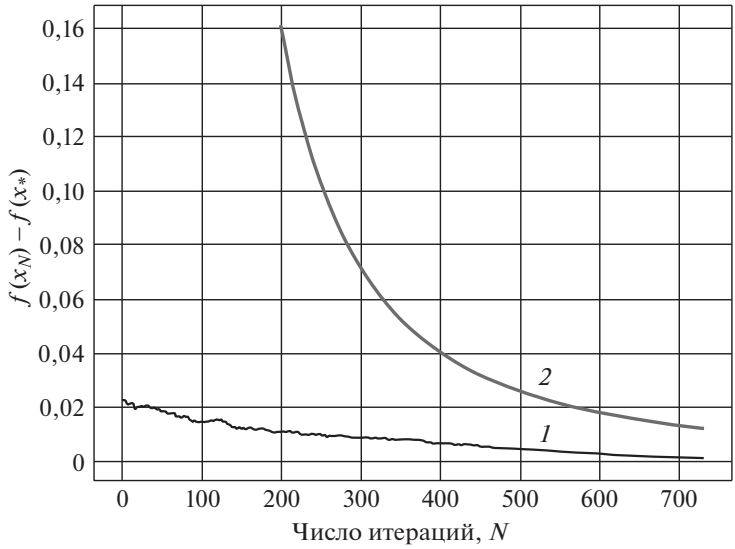

Рис. 2. Сходимость ускоренного неевклидового спуска ACDS для функции (12), размерность $n = 10$. Показана практическая зависимость точности нахождения минимума $f(x_N) - f(x_*)$ от числа итераций N алгоритма (график *1*) и теоретическая оценка $O\left(\frac{4\Theta LC_{n,q}}{N^2}\right)$ (график *2*).

Рассмотрим следующую *задачу*. *Пусть* A — *матрица размеров* $n \times n$ *со случайными независимыми элементами, равномерно распределенными на отрезке* $[0, 1]$, *а матрица* $B = \frac{A^\top A}{\lambda_{\max}(A^\top A)}$, *где* $\lambda_{\max}(A^\top A)$ — *максимальное собственное значение матрицы* $A^\top A$.

Необходимо минимизировать функцию

$$f = \frac{1}{2}\langle x - x_*,\, B(x - x_*)\rangle\,, \quad x \in \mathbb{R}^n, \tag{12}$$

где $x_* = (1, 0, 0, \ldots, 0)^\top$. Решение этой задачи известно и равно x_*, $f(x_*) = 0$. Начальная точка x_0 для всех экспериментов выбиралась как $(0, 0, 0, \ldots, 1)^\top$. Константа Липшица градиента целевой функции $L = 1$.

Следует отметить важность достаточно точного решения вспомогательной задачи минимизации для нахождения z_{k+1} на шаге 7 алгоритма 1 (зеркальный спуск). В рассматриваемом случае эту задачу с помощью метода множителей Лагранжа можно свести к задаче одномерной минимизации, подробности с формулами см. в [19]. В реализации метода одномерная минимизация выполняется с помощью обычной дихотомии с точностью, на один порядок превышающей заданную.

Для различных n и заданной точности ε были рассчитаны теоретически требуемые значения числа итераций по теореме и проведена проверка сходимости на практике. Для данной задачи во всех случаях практическая скорость сходимости по функции была выше. Так, например, для $n = 10$ и $\varepsilon = 10^{-3}$ заданная точность была достигнута за 729 итераций (см.

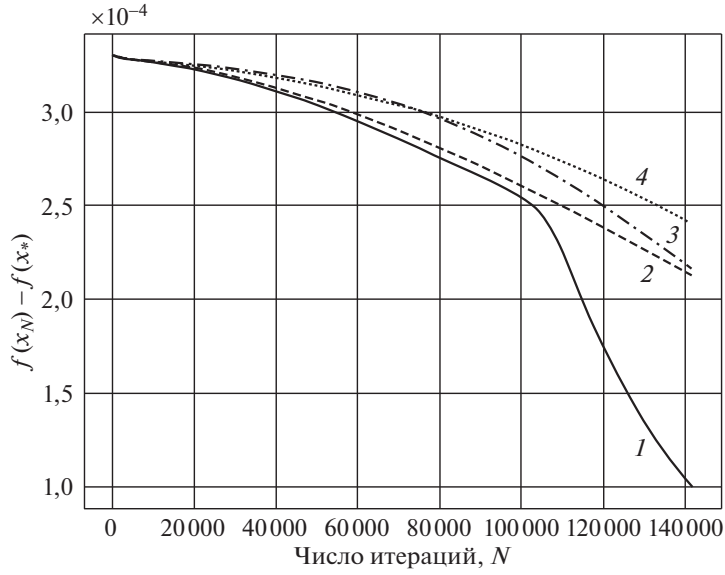

Рис. 3. Сходимость ускоренного неевклидового спуска ACDS для функции (12), размерность $n = 10^3$. Показана практическая зависимость точности нахождения минимума $f(x_N) - f(x_*)$ от числа итераций N алгоритма (сплошная линия — график *1*). Также для сравнения приведены результаты работы метода при других p (евклидова норма — график *2*; $p = 1{,}8$ — график *3*; $p = 1{,}9$ — график *4*) при тех же генерируемых векторах e и точке старта x_0.

рис. 2), а теоретическая оценка числа итераций дает 2537 итераций. Далее, для $n = 10^3$ и $\varepsilon = 10^{-4}$ теоретическая оценка числа итераций дает не более чем 255972 итерации. По факту алгоритм завершил работу за 141643 итерации (см. рис. 3). Медленный спуск в начале работы метода объясняется близостью начального значения целевой функции к оптимальному именно в данном примере ($f(x_0) = 0{,}00032983$).

При проведении численных экспериментов было обнаружено, что преимущество выбора прокс-структуры, связанной с 1-нормой, возникает только в пространствах от средней размерности (от $n = 1000$). Будет ли иметь преимущество предложенный метод, можно определить, сравнив теоретические оценки числа итераций предложенного метода для разных p. На рис. 3 показан именно случай, когда ускоренный спуск по направлению с неевклидовой прокс-структурой оказывается оптимальнее спуска по направлению с евклидовой прокс-структурой.

В целом, численные эксперименты с ускоренным спуском по случайному направлению подтверждают теоретические результаты.

6. Заключение

В статье предложен ускоренный неевклидов спуск по направлению для решения задачи выпуклой безусловной оптимизации. В отличие от известных вариантов методов спуска по направлению (см., например [1]) в данной статье рассматривается ускоренный спуск по направлению с неевклидовой

прокс-структурой. В случае когда 1-норма решения близка к 2-норме решения (это имеет место, например, если решение задачи разрежено — имеет много нулевых компонент), предлагаемый подход улучшает оценку на необходимое число итераций, полученную оптимальным методом из [1], приблизительно в $\sqrt{n}$ раз, где n – размерность пространства, в котором происходит оптимизация.

Данная статья открывает цикл работ, в которых планируется привести полные доказательства утверждений, полученных авторами в 2014–2016 гг. и приведенных (без доказательств) в [20]. В частности, далее планируется распространить приведенные в настоящей статье результаты на безградиентные методы, на задачи стохастической оптимизации и распространить все эти результаты на случай сильно выпуклой функции.

Открытой проблемой остается распространение полученных здесь результатов на случай задач оптимизации на множествах простой структуры. Напомним, что в статье существенным образом использовалось то, что оптимизация происходит на всем пространстве. Тем не менее в будущем планируется показать, что приведенные здесь результаты распространяются на задачи оптимизации на множествах простой структуры в случае, если градиент функционала в точке решения равен нулю (принцип Ферма).

Авторы выражают благодарность Павлу Двуреченскому и Александру Тюрину за помощь в работе.

ПРИЛОЖЕНИЕ

Доказательство леммы 2. Докажем сначала первую часть неравенства:

$$\begin{aligned}
&\alpha_{k+1}\langle n\langle \nabla f(x_{k+1}),\, e_{k+1}\rangle e_{k+1},\, z_k - u\rangle = \\
&= \langle \alpha_{k+1} n\langle \nabla f(x_{k+1}),\, e_{k+1}\rangle e_{k+1},\, z_k - z_{k+1}\rangle + \\
&+ \langle \alpha_{k+1} n\langle \nabla f(x_{k+1}),\, e_{k+1}\rangle e_{k+1},\, z_{k+1} - u\rangle \overset{①}{\leqslant} \\
&\overset{①}{\leqslant} \langle \alpha_{k+1} n\langle \nabla f(x_{k+1}), e_{k+1}\rangle e_{k+1}, z_k - z_{k+1}\rangle + \langle -\nabla V_{z_k}(z_{k+1}), z_{k+1} - u\rangle \overset{②}{=} \\
&\overset{②}{=} \langle \alpha_{k+1} n\langle \nabla f(x_{k+1}),\, e_{k+1}\rangle e_{k+1},\, z_k - z_{k+1}\rangle + V_{z_k}(u) - V_{z_{k+1}}(u) - V_{z_k}(z_{k+1}) \overset{③}{\leqslant} \\
&\overset{③}{\leqslant} \left(\langle \alpha_{k+1} n\langle \nabla f(x_{k+1}),\, e_{k+1}\rangle e_{k+1},\, z_k - z_{k+1}\rangle - \frac{1}{2}||z_k - z_{k+1}||_p^2 \right) + \\
&+ V_{z_k}(u) - V_{z_{k+1}}(u),
\end{aligned} \tag{П.1}$$

где ① выполнено в силу того, что

$$z_{k+1} = \operatorname*{argmin}_{z\in\mathbb{R}^n} \left\{ V_{z_k}(z) + \alpha_{k+1}\langle n\langle \nabla f(x_{k+1}),\, e_{k+1}\rangle e_{k+1},\, z\rangle \right\},$$

откуда следует, что $\langle \nabla V_{z_k}(z_{k+1}) + \alpha_{k+1} n\langle \nabla f(x_{k+1}),\, e_{k+1}\rangle e_{k+1},\, u - z_{k+1}\rangle \geqslant 0$ для всех $u \in Q = \mathbb{R}^n$, ② выполнено в силу равенства треугольника для дивер-

генции Брэгмана[9], ③ выполнено, так как $V_x(y) \geqslant \frac{1}{2}||x-y||_p^2$ в силу сильной выпуклости прокс-функции $d(x)$.

Теперь покажем, что

$$\langle \alpha_{k+1} n \langle \nabla f(x_{k+1}),\, e_{k+1}\rangle\, e_{k+1},\, z_k - z_{k+1}\rangle - \frac{1}{2}||z_k - z_{k+1}||_p^2 \leqslant$$
$$\leqslant \frac{\alpha_{k+1}^2 n^2}{2} |\langle \nabla f(x_{k+1}),\, e_{k+1}\rangle|^2 \cdot ||e_{k+1}||_q^2.$$

Действительно, в силу неравенства Гельдера

$$\langle \alpha_{k+1} n \langle \nabla f(x_{k+1}),\, e_{k+1}\rangle\, e_{k+1},\, z_k - z_{k+1}\rangle \leqslant$$
$$\leqslant \alpha_{k+1} n ||\langle \nabla f(x_{k+1}),\, e_{k+1}\rangle\, e_{k+1}||_q \cdot ||z_k - z_{k+1}||_p =$$
$$= \alpha_{k+1} n |\langle \nabla f(x_{k+1}),\, e_{k+1}\rangle| \cdot ||e_{k+1}||_q \cdot ||z_k - z_{k+1}||_p,$$

откуда

$$\text{(П.2)} \qquad \begin{gathered} \langle \alpha_{k+1} n \langle \nabla f(x_{k+1}),\, e_{k+1}\rangle\, e_{k+1},\, z_k - z_{k+1}\rangle - \frac{1}{2}||z_k - z_{k+1}||_p^2 \leqslant \\ \leqslant \alpha_{k+1} n |\langle \nabla f(x_{k+1}),\, e_{k+1}\rangle| \cdot ||e_{k+1}||_q \cdot ||z_k - z_{k+1}||_p - \frac{1}{2}||z_k - z_{k+1}||_p^2. \end{gathered}$$

Положим $t = ||z_k - z_{k+1}||_p$, $a = \frac{1}{2}$ и $b = \alpha_{k+1} n |\langle \nabla f(x_{k+1}),\, e_{k+1}\rangle| \cdot ||e_{k+1}||_q$, тогда правая часть в (П.2) имеет вид

$$bt - at^2.$$

Если рассматривать полученное выражение как функцию от $t \in \mathbb{R}$, то ее максимум при $t \in \mathbb{R}$ равен (а значит, при $t \in \mathbb{R}_+$ не превосходит) $\frac{b^2}{4a}$. Отсюда и из (П.2) следует неравенство

$$\text{(П.3)} \qquad \begin{gathered} \langle \alpha_{k+1} n \langle \nabla f(x_{k+1}),\, e_{k+1}\rangle\, e_{k+1},\, z_k - z_{k+1}\rangle - \frac{1}{2}||z_k - z_{k+1}||_p^2 \leqslant \\ \leqslant \frac{\alpha_{k+1}^2 n^2}{2} |\langle \nabla f(x_{k+1}),\, e_{k+1}\rangle|^2 \cdot ||e_{k+1}||_q^2. \end{gathered}$$

Итак, учитывая (П.1) и (П.3), получаем, что

$$\alpha_{k+1} \langle n \langle \nabla f(x_{k+1}),\, e_{k+1}\rangle\, e_{k+1},\, z_k - u\rangle \leqslant$$
$$\leqslant \frac{\alpha_{k+1}^2 n^2}{2} |\langle \nabla f(x_{k+1}),\, e_{k+1}\rangle|^2 \cdot ||e_{k+1}||_q^2 + V_{z_k}(u) - V_{z_{k+1}}(u).$$

[9] Действительно,

$$\forall x,\, y \in \mathbb{R}^n \quad \langle -\nabla V_x(y),\, y - u\rangle = \langle \nabla d(x) - \nabla d(y),\, y - u\rangle = (d(u) - d(x) - \langle \nabla d(x),\, u - x\rangle) -$$
$$- (d(u) - d(y) - \langle \nabla d(y),\, u - y\rangle) - (d(y) - d(x) - \langle \nabla d(x),\, y - x\rangle) = V_x(u) - V_y(u) - V_x(y).$$

Беря условное математическое ожидание $\mathbb{E}_{e_{k+1}}[\,\cdot \mid e_1, e_2, \ldots, e_k]$ от левой и правой частей последнего неравенства и пользуясь вторым неравенством из леммы 1, получаем, что

$$\alpha_{k+1}\langle \nabla f(x_{k+1}),\, z_k - u\rangle \leqslant$$
$$\leqslant \alpha_{k+1}^2 \cdot \frac{C_{n,q}}{2n}||\nabla f(x_{k+1})||_2^2 + V_{z_k}(u) - \mathbb{E}_{e_{k+1}}[V_{z_{k+1}}(u) \mid e_1, e_2, \ldots, e_k],$$

где $C_{n,q} = \sqrt{3}\min\{2q-1,\ 32\ln n - 8\}n^{\frac{2}{q}+1}$. Чтобы доказать вторую часть неравенства (6), покажем, что

$$||\nabla f(x_{k+1})||_2^2 \leqslant 2nL\left(f(x_{k+1}) - \mathbb{E}_{e_{k+1}}\left[f(y_{k+1}) \mid e_1, e_2, \ldots, e_k\right]\right). \tag{П.4}$$

Во-первых, для всех $x,\, y \in \mathbb{R}$

$$f(y) - f(x) = \int\limits_0^1 \langle \nabla f(x + \tau(y-x)),\, y - x\rangle d\tau =$$
$$= \langle \nabla f(x),\, y - x\rangle + \int\limits_0^1 \langle \nabla f(x + \tau(y-x)) - \nabla f(x),\, y - x\rangle d\tau \leqslant$$
$$\leqslant \langle \nabla f(x),\, y - x\rangle + \int\limits_0^1 ||\nabla f(x + \tau(y-x)) - \nabla f(x)||_2 \cdot ||y - x||_2 d\tau \leqslant$$
$$\leqslant \langle \nabla f(x),\, y - x\rangle + \int\limits_0^1 \tau L||y - x||_2 \cdot ||y - x||_2 d\tau =$$
$$= \langle \nabla f(x),\, y - x\rangle + \frac{L}{2}||y - x||_2^2,$$

т.е.

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

Беря в последнем неравенстве $x = x_{k+1}$, $y = \mathrm{Grad}_{e_{k+1}}(x_{k+1}) = x_{k+1} - -\frac{1}{L}\langle \nabla f(x_{k+1}),\, e_{k+1}\rangle\, e_{k+1}$, получим, что

$$f(x_{k+1}) - f(y_{k+1}) \geqslant \frac{1}{L}\langle \nabla f(x_{k+1}),\, e_{k+1}\rangle^2 - \frac{1}{2L}\langle \nabla f(x_{k+1}),\, e_{k+1}\rangle^2 \cdot ||e_{k+1}||_2^2 =$$
$$= \frac{1}{2L}\langle \nabla f(x_{k+1}),\, e_{k+1}\rangle^2.$$

Возьмем от этого неравенства условное математическое ожидание $\mathbb{E}_{e_{k+1}}[\,\cdot \mid e_1, e_2, \ldots, e_k]$, используя лемму B.10 из [18] (см. лемму (П.1)), и получим неравенство (П.4). Лемма 2 доказана.

Доказательство леммы 3. Доказательство состоит в выписывании цепочки неравенств:

$$\alpha_{k+1}(f(x_{k+1}) - f(u)) \leqslant$$
$$\leqslant \alpha_{k+1}\langle \nabla f(x_{k+1}),\, x_{k+1} - u\rangle =$$
$$= \alpha_{k+1}\langle \nabla f(x_{k+1}),\, x_{k+1} - z_k\rangle + \alpha_{k+1}\langle \nabla f(x_{k+1}),\, z_k - u\rangle \overset{①}{=}$$
$$\overset{①}{=} \frac{(1-\tau_k)\alpha_{k+1}}{\tau_k}\langle \nabla f(x_{k+1}),\, y_k - x_{k+1}\rangle + \alpha_{k+1}\langle \nabla f(x_{k+1}),\, z_k - u\rangle \overset{②}{\leqslant}$$
$$\overset{②}{\leqslant} \frac{(1-\tau_k)\alpha_{k+1}}{\tau_k}(f(y_k) - f(x_{k+1})) + \alpha_{k+1}\langle \nabla f(x_{k+1}),\, z_k - u\rangle \overset{③}{\leqslant}$$
$$\overset{③}{\leqslant} \frac{(1-\tau_k)\alpha_{k+1}}{\tau_k}(f(y_k) - f(x_{k+1})) +$$
$$+ \alpha_{k+1}^2 LC_{n,q}\cdot (f(x_{k+1}) - \mathbb{E}_{e_{k+1}}[f(y_{k+1}) \mid e_1,\, e_2,\, \ldots,\, e_k]) +$$
$$+ V_{z_k}(u) - \mathbb{E}_{e_{k+1}}[V_{z_{k+1}}(u) \mid e_1,\, e_2,\, \ldots,\, e_k] \overset{④}{=}$$
$$\overset{④}{=} (\alpha_{k+1}^2 LC_{n,q} - \alpha_{k+1})f(y_k) - \alpha_{k+1}^2 LC_{n,q}\mathbb{E}_{e_{k+1}}[f(y_{k+1}) \mid e_1,\, e_2,\, \ldots,\, e_k] +$$
$$+ \alpha_{k+1} f(x_{k+1}) + V_{z_k}(u) - \mathbb{E}_{e_{k+1}}[V_{z_{k+1}}(u) \mid e_1,\, e_2,\, \ldots,\, e_k].$$

Действительно, ① выполнено, так как $x_{k+1} \overset{\text{def}}{=} \tau_k z_k + (1-\tau_k)y_k \Leftrightarrow \tau_k(x_{k+1} - z_k) = (1-\tau_k)(y_k - x_{k+1})$, ② следует из выпуклости $f(\cdot)$ и неравенства $1 - \tau_k \geqslant 0$, ③ справедливо в силу леммы 2 и в ④ используется равенство $\tau_k = \frac{1}{\alpha_{k+1} LC_{n,q}}$. Лемма 3 доказана.

Доказательство теоремы. Заметим, что при $\alpha_{k+1} = \frac{k+2}{2LC_{n,q}}$ выполнено равенство

$$\alpha_k^2 LC_{n,q} = \alpha_{k+1}^2 LC_{n,q} - \alpha_{k+1} + \frac{1}{4LC_{n,q}}.$$

Возьмем для $k = 0, 1, \ldots, N-1$ от каждого неравенства (7) леммы 3 математическое ожидание по $e_1,\, e_2,\, \ldots,\, e_N$, просуммируем полученные неравенства и получим

$$\alpha_N^2 LC_{n,q}\mathbb{E}[f(y_N)] + \sum_{k=1}^{N-1}\frac{1}{4LC_{n,q}}\mathbb{E}[f(y_k)] + \mathbb{E}[V_{z_N}(u)] - V_{z_0}(u) \leqslant \sum_{k=1}^{N}\alpha_k f(u),$$

где $\mathbb{E}[\cdot] = \mathbb{E}_{e_1,\, e_2,\, \ldots,\, e_N}[\cdot]$. Положим $u = x^*$. Так как $\sum\limits_{k=1}^{N}\alpha_k = \frac{N(N+3)}{4LC_{n,q}}$, $\mathbb{E}[f(y_k)] \geqslant \geqslant f(x^*)$, $V_{z_N}(u) \geqslant 0$ и $V_{z_0}(x^*) = V_{x_0}(x^*) \leqslant \Theta$, то выполняется неравенство

$$\frac{(N+1)^2}{4LC_{n,q}}\mathbb{E}[f(y_N)] \leqslant \left(\frac{N(N+3)}{4LC_{n,q}} - \frac{N-1}{4LC_{n,q}}\right) f(x^*) + \Theta,$$

откуда следует, что $\mathbb{E}[f(y_N)] \leqslant f(x^*) + \frac{4\Theta LC_{n,q}}{(N+1)^2}$. Теорема доказана.

Приводим формулировку леммы B.10 из [18]. Отметим, что в доказательстве нигде не использовалось, что второй вектор в скалярном произведении (помимо e) есть градиент функции $f(x)$ (поэтому утверждение леммы (П.1) остается верным для произвольного вектора $s \in \mathbb{R}^n$ вместо $\nabla f(x)$).

Лемма П.1. Пусть $e \sim RS_2^n(1)$ и вектор $s \in \mathbb{R}^n$ — некоторый вектор. Тогда

$$\mathbb{E}_e[\langle s,\, e\rangle^2] = \frac{||s||_2^2}{n}.$$

СПИСОК ЛИТЕРАТУРЫ

1. *Nesterov Yu.* Random gradient-free minimization of convex functions // CORE Discussion Paper 2011/1. 2011.
2. *Немировский А.С., Юдин Д.Б.* Сложность задач и эффективность методов оптимизации. М.: Наука, 1979.
3. *Гасников А.В., Двуреченский П.Е., Нестеров Ю.Е.* Стохастические градиентные методы с неточным оракулом // Тр. МФТИ. 2016. Т. 8. № 1. С. 41–91. arXiv preprint arXiv:1411.4218.
4. *Гасников А.В., Лагуновская А.А., Усманова И.Н., Федоренко Ф.А.* Безградиентные прокс-методы с неточным оракулом для негладких задач выпуклой стохастической оптимизации на симплексе // АиТ. 2016. № 10. С. 57–77.

   Gasnikov A.V., Lagunovskaya A.A., Usmanova I.N., Fedorenko F.A. Gradient-Free Proximal Methods with Inexact Oracle for Convex Stochastic Nonsmooth Optimization Problems on the Simplex // Autom. Remote Control. 2016. V. 77. No. 11. P. 2018–2034.
5. *Гасников А.В., Крымова Е.А., Лагуновская А.А., Усманова И.Н., Федоренко Ф.А.* Стохастическая онлайн оптимизация. Одноточечные и двухточечные нелинейные многорукие бандиты. Выпуклый и сильно выпуклый случаи // АиТ. 2017. № 2. С. 36–49.

   Gasnikov A.V., Krymova E.A., Lagunovskaya A.A., Usmanova I.N., Fedorenko F.A. Stochastic Online Optimization. Single-Point and Multi-Point Non-Linear Multi-Armed Bandits. Convex and Strongly-Convex Case // Autom. Remote Control. 2017. V. 78. No. 2. P. 224–234.
6. *Allen-Zhu Z., Orecchia L.* Linear coupling: An ultimate unification of gradient and mirror descent // arXiv preprint arXiv:1407.1537.
7. *Гасников А.В.* Современные численные методы оптимизации. Универсальный градиентный спуск. Уч. пос. М.: МФТИ, 2018.

   URL: https://arxiv.org/ftp/arxiv/papers/1711/1711.00394.pdf
8. *Nesterov Yu.* Efficiency of Coordinate Descent Methods on Huge-Scale Optimization Problems // SIAM J. Optim. 2012. № 22 (2). P. 341–362.
9. *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) // arXiv preprint arXiv:1707.08486.
10. *Горбунов Э.А., Воронцова Е.А., Гасников А.В.* О верхней оценке математического ожидания нормы равномерно распределенного на сфере вектора и явлении концентрации равномерной меры на сфере // arXiv preprint arXiv:1804.03722.

11. *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 // АиТ. 2018. № 8. С. 38–49.

Bayandina A.S., Gasnikov A.V., Lagunovskaya A.A. Gradient-Free Two-Point Methods for Solving Stochastic Nonsmooth Convex Optimization Problems with Small Non-Random Noises // Autom. Remote Control. 2018. V. 79. No. 8. P. 1399–1408.

12. *Гасников А.В., Двуреченский П.Е., Усманова И.Н.* О нетривиальности быстрых (ускоренных) рандомизированных методов // Тр. МФТИ. 2016. Т. 8. № 2. С. 67–100. arXiv preprint arXiv:1508.02182.

13. *Allen-Zhu Z.* Katyusha: The First Direct Acceleration of Stochastic Gradient Methods // arXiv preprint arXiv:1603.05953.

14. *Guzman C., Nemirovski A.* On Lower Complexity Bounds for Large-Scale Smooth Convex Optimization // J. Complexity. February 2015. V. 31. Iss. 1. P. 1–14.

15. *Гасников А.В., Нестеров Ю.Е.* Универсальный метод для задач стохастической композитной оптимизации // ЖВМ и МФ. 2018. Т. 58. № 1. С. 52–69. arXiv preprint arXiv:1604.05275.

16. *Baydin A.G., Pearlmutter B.A., Radul A.A., Siskand J.M.* Automatic Differentiation in Machine Learning: a Survey // arXiv preprint arXiv:1502.05767.

17. *Брэгман Л.М.* Релаксационный метод нахождения общей точки выпуклых множеств и его применение для решения задач выпуклого программирования // ЖВМ и МФ. 1967. Т. 7. № 3. С. 200–217.

18. *Bogolubsky L., Dvurechensky P., Gasnikov A., Gusev G., Raigorodskii A., Tikhonov A., Zhukovskii M.* Learning Supervised PageRank with Gradient-Based and Gradient-Free Optimization Methods // 13th Annual Conf. on Neural Information Processing Systems (NIPS). 2016. arXiv preprint arXiv:1603.00717.

19. *ACDS method Python code.* URL: https://github.com/evorontsova/ACDS

20. *Гасников А.В.* Эффективные численные методы поиска равновесий в больших транспортных сетях. Дисс. на соискание уч. степ. д.ф.-м.н. по специальности 05.13.18 – Математическое моделирование, численные методы, комплексы программ. М.: МФТИ, 2016. arXiv preprint arXiv:1607.03142.

Статья представлена к публикации членом редколлегии Б.М. Миллером.

Поступила в редакцию 14.10.2017

После доработки 04.07.2018

Принята к публикации 08.11.2018