Accelerated Gradient-Free Optimization Methods with a Non-Euclidean Proximal Operator

Evgeniya Vorontsova, Alexander Gasnikov, Eduard Gorbunov, Pavel Dvurechenskii

▶ **To cite this version:**

Evgeniya Vorontsova, Alexander Gasnikov, Eduard Gorbunov, Pavel Dvurechenskii. Accelerated Gradient-Free Optimization Methods with a Non-Euclidean Proximal Operator. Automation and Remote Control / Avtomatika i Telemekhanika, 2019, 80 (8), pp.1487-1501. ⟨10.1134/S0005117919080095⟩. ⟨hal-02321733⟩

HAL Id: hal-02321733

https://hal.science/hal-02321733v1

Submitted on 28 Oct 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

HAL Authorization

© 2019 г. Е.А. ВОРОНЦОВА, канд. физ.-мат. наук
(vorontsovaea@gmail.com)
(Дальневосточный федеральный университет, Владивосток;
Университет Гренобль Альпы, Гренобль, Франция),
А.В. ГАСНИКОВ, д-р физ.-мат. наук (gasnikov@yandex.ru)
(Московский физико-технический институт;
Национальный исследовательский университет "Высшая школа экономики", Москва;
Кавказский математический центр, Адыгейский государственный университет, Майкоп),
Э.А. ГОРБУНОВ (ed-gorbunov@yandex.ru)
(Московский физико-технический институт),
П.Е. ДВУРЕЧЕНСКИЙ, канд. физ.-мат. наук
(pavel.dvurechensky@gmail.com)
(Институт прикладного анализа и стохастики им. Вейерштрасса, Берлин)

УСКОРЕННЫЕ БЕЗГРАДИЕНТНЫЕ МЕТОДЫ ОПТИМИЗАЦИИ С НЕЕВКЛИДОВЫМ ПРОКСИМАЛЬНЫМ ОПЕРАТОРОМ [1]

Предлагается ускоренный безградиентный метод с неевклидовым проксимальным оператором, связанным с p-нормой ($1 \leqslant p \leqslant 2$). Получены оценки скорости сходимости метода в условиях малого шума, возникающего при вычислении значения функции. Представлены результаты вычислительных экспериментов.

Ключевые слова: ускоренные методы оптимизации, выпуклая оптимизация, безградиентные методы, неточный оракул, неевклидов проксимальный оператор, прокс-

[1]Исследование в разделе 3 выполнено за счет Российского научного фонда (проект № 17-11-01027). В остальных разделах работа А.В. Гасникова финансировалась в рамках Государственной поддержки ведущих университетов Российской Федерации "5-100" и была поддержана Российским фондом фундаментальных исследований (проект № 18-31-20005 мол-а-вед), работа Э.А. Горбунова была поддержана грантом Президента РФ МД-1320.2018.1, работа П.Е. Двуреченского и Е.А. Воронцовой была поддержана Российским фондом фундаментальных исследований (проект № 18-29-03071 мк).

структура.

1. Введение

В данной статье рассматриваются безградиентные методы выпуклой оптимизации, называемые также методами нулевого порядка [1–3]. Основной особенностью этих методов является предположение о доступности только значения целевой функции, но недоступности градиента или гессиана. Такая ситуация возникает, например, если значение функции вычисляется с помощью некоторой вспомогательной компьютерной программы и доработка этой программы для вычисления градиента стоит дороже, чем машинное время. Кроме того, в отличие от вычисления значения функции, вычисление градиента может быть вычислительно нестабильным, например, в глубинных нейронных сетях при использовании обратной прогонки (backpropagation) [4–7].

В [8,9] были предложены ускоренные методы нулевого порядка [2] (безградиентные методы) решения задач гладкой выпуклой безусловной оптимизации. В отличие от известных до [8, 9] методов ускоренные методы обладают более высокой скоростью сходимости. В рассуждениях [9] существенным образом использовалось то, что был выбран именно евклидов проксимальный оператор (далее будем называть его *прокс-структурой*, это выпуклая гладкая функция, порождающая расстояние, и 1-сильно выпуклая относительно какой-то нормы; строгое определение вводится в разделе 3). Такой выбор прокс-структуры для задач безусловной оптимизации является вполне естественным (см., например, [11]). Однако в ряде задач имеется дополнительная информация, которая, например, позволяет рассчитывать на разреженность решения (когда в решении бóльшая часть компонент нулевые). В частности, во многих задачах анализа данных существенным оказывается небольшое число признаков (потому и решения соответствующих задач обучения оказываются разреженными). Кроме

[2] Здесь и далее термин "ускоренный" для безградиентных методов означает точно то же самое, что и для методов первого порядка (градиентных), см., в частности, описание основополагающего и очень популярного в последние годы ускоренного (быстрого, моментного) градиентного метода Нестерова [10]. Ускоренный метод имеет скорость сходимости, пропорциональную $1/k^2$, где k – номер итерации. В отличие от него неускоренный метод имеет скорость сходимости по значению целевой функции, пропорциональную $1/k$.

того, во многих задачах обучения с подкреплением одновременно выполняются два условия: 1) нет возможности считать градиент, есть только возможность считать зашумленное значение целевой функции (см., например, [12]); 2) при выборе параметрической модели часто сложно сразу правильно угадать параметрическое представление и часть параметров оказываются малозначимыми (лишними), что и обеспечивает разреженность. Кроме того, тот же эффект наблюдается, если выбирать точку старта так, чтобы разность между точкой старта и решением была разреженным вектором. Наконец, даже если вдруг это не выполняется, полученные результаты свидетельствуют о том, что предложенный в статье метод с неевклидовой прокс-структурой будет работать не хуже, чем в евклидовом случае.

В [13] был предложен ускоренный спуск (опирающийся на метод из [14]) по случайному направлению и получена оценка скорости сходимости. В [13] метод использовал проекцию градиента на случайное направление. В данной статье используется не сама проекция градиента на случайное направление, а ее аппроксимация конечной разностью, т.е. приближение, которое не использует градиент. Кроме того, рассматривается случай, когда значения функции известны с некоторым ограниченным по абсолютной величине шумом.

2. Постановка задачи

Рассматривается разрешимая задача гладкой выпуклой оптимизации

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

где функция $f(x)$, заданная на $\mathbb{R}^n$, имеет липшицев градиент с константой L в 2-норме

$$\|\nabla f(y) - \nabla f(x)\|_2 \leqslant L \|y - x\|_2 \quad \forall\, x,\, y \in \mathbb{R}^n. \tag{2}$$

При этом в точке минимума x^* выполнено равенство $\nabla f(x^*) = 0$.

В [13] для решения задачи (1) вместо обычного градиента использовалась его стохастическая аппроксимация, построенная на базе производной по случайно выбранному направлению [15]

$$g(x,\, e) = n \langle \nabla f(x),\, e \rangle\, e,$$

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

В отличие от [13], будем использовать не $\langle \nabla f(x), e \rangle e$, а приближенный аналог

$$\frac{f(x+te)-f(x)}{t} e,\ t>0.$$

Более того, предположим, что оракул [3] может выдавать значение функции в любой точке, но с некоторым шумом $\delta(x)$, т.е. от оракула получаем значения $\tilde{f}(x) = f(x) + \delta(x)$, поэтому на практике будем вынуждены использовать

$$\frac{\tilde{f}(x+te)-\tilde{f}(x)}{t} e$$

вместо $\langle \nabla f(x), e \rangle e$.

Итак в силу сделанных предположений теперь работаем не с истинной производной по направлению, а с ее приближенным аналогом [4]

$$\langle \widetilde{\nabla}_{\delta,t} f(x), e \rangle e \stackrel{\text{def}}{=} (\langle \nabla f(x), e \rangle + \delta_\nabla(x, t, e)) e,$$

где $\delta_\nabla(x, t, e)$ — ошибка приближения значения $\langle \nabla f(x), e \rangle$, связанная с вычислительной ошибкой (значения функции известны с некоторым шумом) и с аппроксимационной ошибкой метода приближения, т.е.

$$\delta_\nabla(x, t, e) = \underbrace{\frac{\delta(x+te)-\delta(x)}{t}}_{\text{вычислительная ошибка}} + \underbrace{\frac{f(x+te)-f(x)}{t} - \langle \nabla f(x), e \rangle}_{\text{ошибка метода}}.$$

[3] 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 [16] — как число обращений (по ходу работы метода) к оракулу для достижения заданной точности (по функции).

[4] Строго говоря, вектор $\widetilde{\nabla}_{\delta,t} f(x)$ определяется из приведенного выражения не единственным образом, однако данное обозначение крайне удобно использовать для записи метода и доказательств. Кроме того, нигде в работе авторы не пользуются конкретным выбором $\widetilde{\nabla}_{\delta,t} f(x)$, везде существенную роль играет выражение $\langle \widetilde{\nabla}_{\delta,t} f(x), e \rangle e$, которое определяется однозначно.

Ограничимся рассмотрением следующей концепции абсолютной неточности оракула. Пусть про шум известно, что $\forall\, x \in \mathbb{R}^n \hookrightarrow |\delta(x)| \leqslant \delta$. Оценим $|\delta_\nabla(x,\, e,\, t)|$:

$$|\delta_\nabla(x,\, e,\, t)| \leqslant \tfrac{|\delta(x+te)|+|\delta(x)|}{t} + \left|\tfrac{f(x+te)-f(x)}{t} - \langle \nabla f(x),\, e\rangle\right| \leqslant \tfrac{2\delta}{t} + \tfrac{Lt}{2},$$

где последнее слагаемое в правой части возникает в силу неравенства

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

которое следует из (2). Выберем параметр $t = t^*$ так, чтобы минимизировать значение верхней оценки на $\delta_\nabla(x,\, e,\, t)$. Параметр t^* находится из условия

$$-\frac{2\delta}{(t^*)^2} + \frac{L}{2} = 0 \;\Rightarrow\; t^* = 2\sqrt{\frac{\delta}{L}}.$$

Тогда

$$|\delta_\nabla(x,\, e,\, t^*)| \leqslant 2\sqrt{L\delta} \stackrel{\text{def}}{=} \tilde{\delta}. \tag{3}$$

Везде далее будем использовать параметр $t = t^*$, поэтому можно писать $\delta_\nabla(x,\, e,\, t^*) = \delta_\nabla(x,\, e)$. Кроме того, для упрощения нотации будем использовать обозначение $\widetilde{\nabla} f(x)$ вместо $\widetilde{\nabla}_{\delta,\, t} f(x)$.

3. Ускоренный безградиентный метод

Введем дивергенцию Брэгмана $V_z\,(y)$, связанную с p-нормой:

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

где функция $d(x)$ является непрерывно дифференцируемой и сильно выпуклой с константой сильной выпуклости, равной единице (и называется *прокс-функцией*, или *прокс-структурой*, связанной с p-нормой). Например, для $p = 1$ функцию $d(x)$ можно взять такой:

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

где $a = \frac{2\log n}{2\log n - 1}$. Кроме того, пусть q — такое число, что $\frac{1}{p} + \frac{1}{q} = 1$.

Введем следующие объекты:

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

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

где

$$\langle \widetilde{\nabla} f(x),\, e\rangle\, e \stackrel{\text{def}}{=} \left(\langle \nabla f(x),\, e\rangle + \delta_\nabla(x,\, e)\right) e.$$

Также обозначим (см. теорему 1 из [17], формулировка приведена в Приложении)

$$C_{n,q} \stackrel{\text{def}}{=} \sqrt{3}\min\{2q-1, 32\ln n - 8\} n^{\frac{2}{q}+1}. \tag{4}$$

Опишем ускоренный безградиентный метод (Accelerated by Coupling Derivative-Free Method — ACDF) (см. алгоритм 1). Обращение к оракулу целевой функции происходит на шагах вычисления значений $\mathrm{Grad}_{e_{k+1}}(x_{k+1})$ и $\mathrm{Mirr}_{e_{k+1}}(x_{k+1},\, z_k,\, \alpha_{k+1})$.

Алгоритм 1. 1 ACDF

Вход: оракул, выдающий значение функции f (выпуклой дифференцируемой функции на $\mathbb{R}^n$ с липшицевым градиентом с константой L по отношению к 2-норме) в любой точке x с некоторым шумом $\delta(x)$; x_0 — некоторая стартовая точка; N — количество итераций.

Выход: точка y_N.

1: $y_0 \leftarrow x_0$, $z_0 \leftarrow x_0$
2: **for** $k = 0, \ldots, N-1$
3: $\quad \alpha_{k+1} \leftarrow \frac{k+2}{4LC_{n,q}}$, $\tau_k \leftarrow \frac{1}{2\alpha_{k+1}LC_{n,q}} = \frac{2}{k+2}$
4: $\quad$ Генерируется $e_{k+1} \in RS_2^n(1)$, независимо от предыдущих итераций
5: $\quad x_{k+1} \leftarrow \tau_k z_k + (1-\tau_k) y_k$
6: $\quad y_{k+1} \leftarrow \mathrm{Grad}_{e_{k+1}}(x_{k+1})$
7: $\quad z_{k+1} \leftarrow \mathrm{Mirr}_{e_{k+1}}(x_{k+1},\, z_k,\, \alpha_{k+1})$
8: **end for**
9: **return** y_N

Схема доказательства оценки скорости сходимости ACDF близка к схеме доказательства теоремы в [13] и опирается на две следующие леммы, в которых $\tilde{\delta}$ определена в (3).

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

$$\begin{aligned} &\alpha_{k+1}\langle \nabla f(x_{k+1}),\, z_k - u\rangle \leqslant \\ &\leqslant 2\alpha_{k+1}^2 LC_{n,q}(f(x_{k+1}) - \mathbb{E}_{e_{k+1}}[f(y_{k+1})]) + V_{z_k}(u) - \mathbb{E}_{e_{k+1}}[V_{z_{k+1}}(u)] + \\ &+\tfrac{7}{4}\alpha_{k+1}^2 C_{n,q}\tilde{\delta}^2 + \sqrt{n}\tilde{\delta}\alpha_{k+1}||u - z_k||_p,\ q \geqslant 2,\ n \geqslant 8. \end{aligned} \tag{5}$$

Лемма 2. Для всех $u \in \mathbb{R}^n$ *выполнено*

$$\begin{aligned} &2\alpha_{k+1}^2 LC_{n,q}\mathbb{E}_{e_{k+1}}[f(y_{k+1}) \mid e_1, \ldots, e_k] - \\ &-(2\alpha_{k+1}^2 LC_{n,q} - \alpha_{k+1})f(y_k) + \mathbb{E}_{e_{k+1}}[V_{z_{k+1}}(u)] - V_{z_k}(u) - \\ &-\tfrac{7}{4}\alpha_{k+1}^2 C_{n,q}\tilde{\delta}^2 - \sqrt{n}\tilde{\delta}\alpha_{k+1}||u - z_k||_p \leqslant \alpha_{k+1}f(u),\ q \geqslant 2,\ n \geqslant 8. \end{aligned} \tag{6}$$

Теорема. Пусть $f(x)$ *— выпуклая дифференцируемая функция на* $\mathbb{R}^n$ *с константой Липшица для градиента, равной* L, $d(x)$ *— 1-сильно выпуклая в p-норме функция на* $\mathbb{R}^n$, $N \in \mathbb{N}$. *Тогда ACDF на выходе даст точку* y_N, *удовлетворяющую неравенству*

$$\mathbb{E}[f(y_N)] - f(x^*) \leqslant \frac{16\Theta LC_{n,q}}{N^2} + \frac{35N\delta}{4} + \frac{16\sqrt{2\Theta nL\delta}}{N^2} + \frac{8nN^2\delta}{C_{n,q}}, \tag{7}$$

где $\Theta \overset{def}{=} V_{x_0}(x^*)$, $q \geqslant 2$, $n \geqslant 8$, *а* $C_{n,q}$ *определено в* (4).

Таким образом, теорема утверждает, что алгоритм ACDF через N итераций выдаст точку y_N, удовлетворяющую неравенству $\mathbb{E}[f(y_N)] - f(x^*) \leqslant \varepsilon$, $\varepsilon > 0$, если $N = O\left(\sqrt{\frac{\Theta LC_{n,q}}{\varepsilon}}\right)$ и шум δ такой, что [5]

$$\begin{aligned} \delta = O\left(\min\left\{\frac{\varepsilon^{\frac{3}{2}}}{\sqrt{\Theta LC_{n,q}}}, \frac{C_{n,q}^2\Theta L}{n}, \frac{\varepsilon^2}{n\Theta L}\right\}\right) = \\ = O\left(\min\left\{\frac{\varepsilon^{\frac{3}{2}}}{\sqrt{\Theta LC_{n,q}}}, \frac{\varepsilon^2}{n\Theta L}\right\}\right). \end{aligned} \tag{8}$$

По определению (см. (4)) $C_{n,q} = \sqrt{3}\min\{2q-1, 32\ln n - 8\}n^{\frac{2}{q}+1}$, а в случае $p = 2$, $q = 2$ можно взять $C_{n,q} = n^2$, что видно из теоремы 1 в [17] и леммы B.10 из [18] (формулировки приведены в Приложении), поэтому минимум достигается на втором

[5] Если $N \sim \sqrt{\frac{\Theta LC_{n,q}}{\varepsilon}}$ (оценивается порядок величины, числовые константы не учитываются), то

$$\begin{aligned} \tfrac{7(N+2)(2N+3)\delta}{6(N+1)} \sim N\delta \sim \sqrt{\tfrac{\Theta LC_{n,q}}{\varepsilon}}\delta \ &\Rightarrow\ \delta = O\left(\tfrac{\varepsilon^{\frac{3}{2}}}{\sqrt{\Theta LC_{n,q}}}\right), \\ \tfrac{16\sqrt{2\Theta nL\delta}}{(N+1)^2} \sim \tfrac{\varepsilon\sqrt{n\delta}}{\sqrt{\Theta LC_{n,q}}} \ &\Rightarrow\ \delta = O\left(\tfrac{\Theta LC_{n,q}^2}{n}\right), \\ \tfrac{8nN^4\delta}{C_{n,q}(N+1)^2} \sim \tfrac{nN^2\delta}{C_{n,q}} \sim \tfrac{n\Theta L\delta}{\varepsilon} \ &\Rightarrow\ \delta = O\left(\tfrac{\varepsilon^2}{n\Theta L}\right). \end{aligned}$$

аргументе, т.е. если не учитывать еще и константы Θ и L, то $\delta = O\left(\frac{\varepsilon^2}{n}\right)$ и $N = O\left(\sqrt{\frac{\Theta L n^2}{\varepsilon}}\right)$.

Если же $p = 1$, $q = \infty$, то $C_{n,q} = n\sqrt{3}\,(32 \ln n - 8)$ и

$$N = O\left(\sqrt{\frac{\Theta L n \ln n}{\varepsilon}}\right),\ \delta = O\left(\min\left\{\frac{\varepsilon^{\frac{3}{2}}}{\sqrt{\Theta L n \ln n}}, \frac{\varepsilon^2}{n\Theta L}\right\}\right).$$

4. Неускоренный безградиентный метод

Для сопоставления рассмотрим неускоренный метод

$$x_{k+1} = \underset{y \in \mathbb{R}^n}{\operatorname{argmin}}\{\alpha\langle n\langle \widetilde{\nabla} f(x_k), e_{k+1}\rangle e_{k+1}, y - x_k\rangle + V_{x_k}(y)\}.$$

Можно показать, что данный метод за N итераций генерирует последовательность точек x_0, x_1, …, x_N, такую что

$$\begin{aligned} &\mathbb{E}[f(\bar{x}^N)] - f(x^*) \leqslant \\ &\leqslant \tfrac{16\Theta L C_{n,q}}{nN} + \tfrac{8\sqrt{2n\Theta L\delta}}{N} + 3n\delta + \tfrac{8n^2 N\delta}{C_{n,q}}, \end{aligned} \tag{9}$$

где $\bar{x}^N = \frac{1}{N}\sum_{k=0}^{N-1} x_k$. Оказывается, что нужно брать $\delta = O\left(\frac{\varepsilon^2}{n}\right)$ (как и в ускоренном случае), чтобы получить ε-решение по функции.

5. Численные эксперименты

Для практического применения предложенный ускоренный безградиентный метод ACDF был реализован на языке программирования Python. Также была реализована концепция неточно заданного оракула. Был рассмотрен случай, когда неточность порождается только неточностью вычисления целевой функции. Зашумления производились на каждой итерации, независимо от предыдущих, случайным образом из отрезка $[-\delta, \delta]$, где максимально возможная для шума граница δ вычислялась по (8). Код метода и демонстрация вычислительных свойств метода с построением графиков сходимости доступны как Jupyter Notebook и выложены в свободном доступе на Github [19].

Рассмотрим следующую *задачу*. *Пусть* A — *матрица размеров* $n \times n$ *со случайными независимыми элементами, равномерно распределенными на* $[0, 1]$, *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

$$f(x) = \frac{1}{2}\langle x - x^*,\, B(x - x^*)\rangle\,, \quad x \in \mathbb{R}^n, \tag{10}$$

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 δ', где $\delta' \in [-\delta, \delta]$. Константа Липшица градиента целевой функции $L = 1$.

Для различных n, заданной точности ε и заданной границы максимального шума δ 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 $n = 10$, $\varepsilon = 10^{-4}$, $\delta = 2,1715 \cdot 10^{-10}$ заданная точность была достигнута за 1106 итераций (см. рис. 1), а теоретическая оценка числа итераций дает 17215 итераций. Далее, для $n = 10^3$, $\varepsilon = 10^{-4}$ теоретическая оценка числа итераций дает не более чем 527756 итераций. По факту алгоритм завершил работу за 141476 итераций (см. рис. 2).

Рис. 1

Рис. 2

При проведении численных экспериментов также были подтверждены результаты зкспериментов в [13]: и для ускоренного безградиентного метода преимущество выбора проксимального оператора, связанного с 1-нормой, возникает только для задач средней и большой размерности (от $n = 1000$). Будет ли иметь преимущество предложенный метод, можно определить, сравнив теоретические оценки числа итераций из теоремы для разных p. На рис. 2 показаны случаи, когда ускоренный безградиентный метод с неевклидовой прокс-структурой работает быстрее ускоренного безградиентного метода с евклидовой прокс-структурой. Сравнение этого эксперимента для ускоренного безградиентного метода (с неточным оракулом) с таким же, но для ускоренного спуска по случайному направлению (с точным оракулом) из [13] (см. рис. 2, *б*), показывает, что принципиальный результат тот же, но скорость сходимости оказывается хуже, что вполне естественно для метода нулевого порядка с неточным оракулом.

В целом, численные эксперименты с ускоренным безградиентным методом подтверждают теоретические результаты.

6. Заключение

В статье предложен ускоренный безградиентный метод и рассмотрен неускоренный безградиентный метод. Безградиентные методы рассматриваются в условиях наличия малого шума, возникающего при вычислении значения функции. Полученные оценки скорости сходимости ускоренного безградиентного метода в условиях малого шума подтверждены результатами вычислительных экспериментов.

В отличие от известных вариантов безградиентных методов (см., например, [9]) в данной статье рассматриваются безградиентные методы с неевклидовым проксимальным оператором. 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 [9], приблизительно в $\sqrt{n}$ раз, где n — размерность пространства, в котором происходит оптимизация.

Данная статья продолжает цикл работ, открытый публикацией [13] (см. также [20]). Далее планируется распространить приведенные в настоящей статье результаты на задачи стохастической оптимизации и распространить все эти результаты на случай сильно выпуклой функции.

ПРИЛОЖЕНИЕ

Доказательство леммы 1. Во-первых,

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

где ① выполнено в силу того, что $z_{k+1} = \operatorname{argmin}_{z\in\mathbb{R}^n}\left\{V_{z_k}(z) + \alpha_{k+1}\langle n\langle\widetilde{\nabla} f(x_{k+1}),\, e_{k+1}\rangle\, e_{k+1},\, z\rangle\right\}$, откуда следует, что $\langle \nabla V_{z_k}(z_{k+1}) + \alpha_{k+1} n\langle\widetilde{\nabla} f(x_{k+1}),\, e_{k+1}\rangle\, e_{k+1},\, u - z_{k+1}\rangle \geqslant 0$ для всех

$u \in \mathbb{R}^n$, ② выполнено в силу равенства треугольника для дивергенции Брэгмана [6], ③ выполнено, так как $V_x(y) \geqslant \frac{1}{2}||x-y||_p^2$ в силу сильной выпуклости прокс-функции $d(x)$.

Аналогично доказательству (П.3) из [13] можно показать, что

$$
\begin{aligned}
&\langle \alpha_{k+1} n \langle \widetilde{\nabla} f(x_{k+1}),\, e_{k+1}\rangle\, e_{k+1},\, z_k - z_{k+1}\rangle - \tfrac{1}{2}||z_k - z_{k+1}||_p^2 \leqslant \\
&\leqslant \tfrac{\alpha_{k+1}^2 n^2}{2} |\langle \widetilde{\nabla} f(x_{k+1}),\, e_{k+1}\rangle|^2 ||e||_q^2 = \\
&= \tfrac{\alpha_{k+1}^2 n^2}{2} (\langle \nabla f(x_{k+1}),\, e_{k+1}\rangle + \delta_\nabla(x_{k+1},\, e_{k+1}))^2 ||e||_q^2 \overset{①}{\leqslant} \\
&\overset{①}{\leqslant} \alpha_{k+1}^2 n^2 \langle \nabla f(x_{k+1}),\, e_{k+1}\rangle^2 ||e_{k+1}||_q^2 + \alpha_{k+1}^2 n^2 \delta_\nabla^2(x_{k+1}, e_{k+1}) ||e_{k+1}||_q^2 \overset{②}{\leqslant} \\
&\overset{②}{\leqslant} \alpha_{k+1}^2 n^2 \langle \nabla f(x_{k+1}),\, e_{k+1}\rangle^2 ||e_{k+1}||_q^2 + \alpha_{k+1}^2 n^2 \tilde{\delta}^2 ||e_{k+1}||_q^2,
\end{aligned}
\tag{П.2}
$$

где ① следует из неравенства $(a+b)^2 \leqslant 2a^2 + 2b^2$, $\tilde{\delta}$ определена в (3), ② выполнено в силу $|\delta_\nabla(x,\, e)| \leqslant \tilde{\delta}\ \forall x, e \in \mathbb{R}^n$. Используя неравенство (П.2), из неравенства (П.1) можно получить, что

$$
\begin{aligned}
&\alpha_{k+1} \langle n \langle \widetilde{\nabla} f(x_{k+1}),\, e_{k+1}\rangle\, e_{k+1},\, z_k - u\rangle \leqslant \\
&\leqslant \alpha_{k+1}^2 n^2 \langle \nabla f(x_{k+1}),\, e_{k+1}\rangle^2 ||e_{k+1}||_q^2 + \alpha_{k+1}^2 n^2 \tilde{\delta}^2 ||e_{k+1}||_q^2 + \\
&+ V_{z_k}(u) - V_{z_{k+1}}(u),
\end{aligned}
\tag{П.3}
$$

что можно записать в виде

$$
\begin{aligned}
&\alpha_{k+1} \langle n \langle \nabla f(x_{k+1}),\, e_{k+1}\rangle\, e_{k+1},\, z_k - u\rangle \leqslant \\
&\leqslant \alpha_{k+1}^2 n^2 \langle \nabla f(x_{k+1}),\, e_{k+1}\rangle^2 ||e_{k+1}||_q^2 + \alpha_{k+1}^2 n^2 \tilde{\delta}^2 ||e_{k+1}||_q^2 + \\
&+ V_{z_k}(u) - V_{z_{k+1}}(u) + \alpha_{k+1} n \langle \delta_\nabla(x_{k+1},\, e_{k+1}) e_{k+1},\, u - z_k\rangle \leqslant \\
&\leqslant \alpha_{k+1}^2 n^2 \langle \nabla f(x_{k+1}),\, e_{k+1}\rangle^2 ||e_{k+1}||_q^2 + V_{z_k}(u) - V_{z_{k+1}}(u) + \\
&+ \alpha_{k+1}^2 n^2 \tilde{\delta}^2 ||e_{k+1}||_q^2 + \alpha_{k+1} n |\langle \delta_\nabla(x_{k+1},\, e_{k+1}) e_{k+1},\, u - z_k\rangle|.
\end{aligned}
\tag{П.4}
$$

[6] Действительно,

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

В силу теоремы 1 из [17] и того что $C_{n,q}$ равно $\sqrt{3}\min\{2q-1, 32\ln n-8\}n^{\frac{2}{q}+1}$, получаем:

$$
\begin{aligned}
\mathbb{E}_{e_{k+1}}[\alpha_{k+1}^2 n^2\langle\nabla f(x_{k+1}),\, e_{k+1}\rangle^2||e_{k+1}||_q^2] &\leqslant \tfrac{\alpha_{k+1}^2 C_{n,q}}{n}||\nabla f(x_{k+1})||_2^2,\\
\mathbb{E}_{e_{k+1}}[\alpha_{k+1}^2 n^2\tilde{\delta}^2||e_{k+1}||_q^2] &\leqslant \tfrac{3}{4}\alpha_{k+1}^2 C_{n,q}\tilde{\delta}^2,\\
\mathbb{E}_{e_{k+1}}[\alpha_{k+1} n|\langle\delta_\nabla(x_{k+1},\, e_{k+1})e_{k+1},\, u-z_k\rangle|] &=\\
= \mathbb{E}_{e_{k+1}}[\alpha_{k+1} n|\delta_\nabla(x_{k+1},\, e_{k+1})|\cdot|\langle e_{k+1},\, u-z_k\rangle|] &\overset{①}{\leqslant}\\
\overset{①}{\leqslant} \alpha_{k+1} n\tilde{\delta}\mathbb{E}_{e_{k+1}}[|\langle e_{k+1},\, u-z_k\rangle|] \leqslant \alpha_{k+1} n\tilde{\delta}\sqrt{\mathbb{E}_{e_{k+1}}[|\langle e_{k+1},\, u-z_k\rangle|^2]} &\overset{②}{\leqslant}\\
\overset{②}{\leqslant} \alpha_{k+1} n\tilde{\delta}\sqrt{\tfrac{||u-z_k||_2^2}{n}} = \sqrt{n}\tilde{\delta}\alpha_{k+1}||u-z_k||_2 &\overset{③}{\leqslant} \sqrt{n}\tilde{\delta}\alpha_{k+1}||u-z_k||_p,
\end{aligned}
\tag{П.5}
$$

где ① следует из $|\delta_\nabla(x_{k+1},\, e_{k+1})| \leqslant \tilde{\delta}$, ② выполнено в силу леммы B.10 из [18], ③ получено при помощи следующего факта: $\forall\, x \in \mathbb{R}^n\ \forall\, 1 \leqslant p \leqslant q \leqslant \infty \hookrightarrow ||x||_p \geqslant ||x||_q$ (данное неравенство доказывается, например, при помощи рассмотрения p-нормы для фиксированного x как функции, зависящей от p, а точнее логарифма p-нормы [7]). Беря от (П.4) математическое ожидание по e_{k+1} и пользуясь неравенствами (П.5), получим, что

$$
\begin{aligned}
\alpha_{k+1}\langle\nabla f(x_{k+1}),\, z_k-u\rangle &\leqslant\\
\leqslant \tfrac{\alpha_{k+1}^2 C_{n,q}}{n}||\nabla f(x_{k+1})||_2^2 + V_{z_k}(u) - \mathbb{E}_{e_{k+1}}[V_{z_{k+1}}(u)] &+\\
+\tfrac{3}{4}\alpha_{k+1}^2 C_{n,q}\tilde{\delta}^2 + \sqrt{n}\tilde{\delta}\alpha_{k+1}||u-z_k||_p.&
\end{aligned}
\tag{П.6}
$$

Покажем теперь, что

$$
||\nabla f(x_{k+1})||_2^2 \leqslant 2nL(f(x_{k+1}) - \mathbb{E}_{e_{k+1}}[f(y_{k+1})]) + n\tilde{\delta}^2.
\tag{П.7}
$$

[7] Пусть $g_x(p) \overset{\text{def}}{=} \ln||x||_p = \ln\left(\sum_{k=1}^n |x_k|^p\right)^{\frac{1}{p}} = \frac{1}{p}\ln\left(\sum_{k=1}^n |x_k|^p\right)$. Тогда

$$
\frac{dg_x(p)}{dp} = -\frac{1}{p^2}\ln\left(\sum_{k=1}^n |x_k|^p\right) + \frac{1}{p}\cdot\frac{\sum\limits_{k=1}^n \ln(|x_k|)\cdot|x_k|^p}{\sum\limits_{k=1}^n |x_k|^p}.
$$

Так как $\ln y$ — вогнутая по y функция, то по неравенству Йенсена получаем, что

$$
\frac{dg_x(p)}{dp} \leqslant \frac{1}{p}\ln\left(\sum_{k=1}^n |x_k|^p\right)^{-\frac{1}{p}} + \frac{1}{p}\ln\left(\sum_{k=1}^n |x_k|\cdot\frac{|x_k|^p}{\sum\limits_{k=1}^n |x_k|^p}\right) = \frac{1}{p}\ln\left(\frac{\sum\limits_{k=1}^n |x_k|^{p+1}}{\left(\sum\limits_{k=1}^n |x_k|^p\right)^{\frac{p+1}{p}}}\right) \leqslant \frac{1}{p}\ln\left(\frac{\sum\limits_{k=1}^n |x_k|^{p+1}}{\sum\limits_{k=1}^n \left(|x_k|^p\right)^{\frac{p+1}{p}}}\right) = 0,
$$

т.е. функция $g_x(p)$ — невозрастающая функция на $[1, +\infty)$.

Во-первых, для всех $x,\, y \in \mathbb{R}$ в силу (2)

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

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

$$\begin{aligned} f(x_{k+1}) - f(y_{k+1}) \geqslant \tfrac{1}{L}\langle \widetilde{\nabla} f(x_{k+1}),\, e_{k+1}\rangle \cdot \langle \nabla f(x_{k+1}),\, e_{k+1}\rangle - \\ - \tfrac{1}{2L}\langle \widetilde{\nabla} f(x_{k+1}),\, e_{k+1}\rangle^2 \cdot ||e_{k+1}||_2^2 = \\ = \tfrac{1}{L}\langle \nabla f(x_{k+1}),\, e_{k+1}\rangle^2 + \tfrac{1}{L}\langle \nabla f(x_{k+1}),\, e_{k+1}\rangle \delta_\nabla(x_{k+1}, e_{k+1}) - \\ - \tfrac{1}{2L}\langle \nabla f(x_{k+1}),\, e_{k+1}\rangle^2 - \tfrac{1}{L}\langle \nabla f(x_{k+1}),\, e_{k+1}\rangle \delta_\nabla(x_{k+1}, e_{k+1}) - \\ - \tfrac{1}{2L}\delta_\nabla^2(x_{k+1}, e_{k+1}) = \tfrac{1}{2L}\langle \nabla f(x_{k+1}),\, e_{k+1}\rangle^2 - \tfrac{1}{2L}\delta_\nabla^2(x_{k+1}, e_{k+1}), \end{aligned}$$

откуда получаем неравенство

$$\begin{aligned} \langle \nabla f(x_{k+1}),\, e_{k+1}\rangle^2 \leqslant 2L(f(x_{k+1}) - f(y_{k+1})) + \delta_\nabla^2(x_{k+1}, e_{k+1}) \leqslant \\ \leqslant 2L(f(x_{k+1}) - f(y_{k+1})) + \tilde{\delta}^2, \end{aligned}$$

так как $\forall x, e \in \mathbb{R}^n \hookrightarrow |\delta_\nabla(x, e)| \leqslant \tilde{\delta}$. Возьмем от этого неравенства условное математическое ожидание $\mathbb{E}_{e_{k+1}}[\,\cdot\,]$, используя лемму B.10 из [18], и получим (П.7).

Наконец, из (П.6) и (П.7) получим, что

$$\begin{aligned} \alpha_{k+1}\langle \nabla f(x_{k+1}),\, z_k - u\rangle \leqslant \\ \leqslant 2\alpha_{k+1}^2 L C_{n,q}(f(x_{k+1}) - \mathbb{E}_{e_{k+1}}[f(y_{k+1})]) + V_{z_k}(u) - \mathbb{E}_{e_{k+1}}[V_{z_{k+1}}(u)] + \\ + \tfrac{7}{4}\alpha_{k+1}^2 C_{n,q}\tilde{\delta}^2 + \sqrt{n}\tilde{\delta}\alpha_{k+1}||u - z_k||_p. \end{aligned}$$

Лемма 1 доказана.

Доказательство леммы 2. Запишем цепочку неравенств:

$$\begin{aligned}
\alpha_{k+1}(f(x_{k+1}) - f(u)) \leqslant \\
\leqslant \alpha_{k+1}\langle \nabla f(x_{k+1}),\, x_{k+1} - u\rangle = \\
= \alpha_{k+1}\langle \nabla f(x_{k+1}),\, x_{k+1} - z_k\rangle + \alpha_{k+1}\langle \nabla f(x_{k+1}),\, z_k - u\rangle \overset{①}{=} \\
\overset{①}{=} \tfrac{(1-\tau_k)\alpha_{k+1}}{\tau_k}\langle \nabla f(x_{k+1}),\, y_k - x_{k+1}\rangle + \alpha_{k+1}\langle \nabla f(x_{k+1}),\, z_k - u\rangle \overset{②}{\leqslant} \\
\overset{②}{\leqslant} \tfrac{(1-\tau_k)\alpha_{k+1}}{\tau_k}(f(y_k) - f(x_{k+1})) + \alpha_{k+1}\langle \nabla f(x_{k+1}),\, z_k - u\rangle \overset{③}{\leqslant} \\
\overset{③}{\leqslant} \tfrac{(1-\tau_k)\alpha_{k+1}}{\tau_k}(f(y_k) - f(x_{k+1})) + \\
+2\alpha_{k+1}^2 LC_{n,q}\cdot(f(x_{k+1}) - \mathbb{E}_{e_{k+1}}[f(y_{k+1}) \mid e_1,\, e_2,\, \ldots,\, e_k]) + \\
+V_{z_k}(u) - \mathbb{E}_{e_{k+1}}[V_{z_{k+1}}(u) \mid e_1,\, e_2,\, \ldots,\, e_k] + \\
+\tfrac{7}{4}\alpha_{k+1}^2 C_{n,q}\tilde{\delta}^2 + \sqrt{n}\tilde{\delta}\alpha_{k+1}||u - z_k||_p \overset{④}{=} \\
\overset{④}{=} (2\alpha_{k+1}^2 LC_{n,q} - \alpha_{k+1})f(y_k) - 2\alpha_{k+1}^2 LC_{n,q}\mathbb{E}_{e_{k+1}}[f(y_{k+1}) \mid e_1,\, e_2,\, \ldots,\, e_k] + \\
+\alpha_{k+1}f(x_{k+1}) + V_{z_k}(u) - \mathbb{E}_{e_{k+1}}[V_{z_{k+1}}(u) \mid e_1,\, e_2,\, \ldots,\, e_k] + \\
+\tfrac{7}{4}\alpha_{k+1}^2 C_{n,q}\tilde{\delta}^2 + \sqrt{n}\tilde{\delta}\alpha_{k+1}||u - z_k||_p.
\end{aligned}$$

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 1 и в ④ используется равенство $\tau_k = \frac{1}{2\alpha_{k+1}LC_{n,q}}$. Лемма 2 доказана.

Доказательство теоремы. Заметим, что в силу выбора $\alpha_k = \frac{k+1}{4LC_{n,p}}$ выполняется равенство

$$2\alpha_k^2 LC_{n,q} = 2\alpha_{k+1}^2 LC_{n,q} - \alpha_{k+1} + \tfrac{1}{8LC_{n,q}}. \tag{П.8}$$

Действительно,

$$\begin{aligned}
2\alpha_k^2 LC_{n,q} = \tfrac{(k+1)^2}{8LC_{n,q}}, \\
2\alpha_{k+1}^2 LC_{n,q} - \alpha_{k+1} + \tfrac{1}{8LC_{n,q}} = \tfrac{(k+2)^2}{8LC_{n,q}} - \tfrac{k+2}{4LC_{n,q}} + \tfrac{1}{8LC_{n,q}} = \\
= \tfrac{(k+1)^2 + 2(k+1) + 1 - 2(k+2) + 1}{8LC_{n,q}} = \tfrac{(k+1)^2}{8LC_{n,q}} = 2\alpha_k^2 LC_{n,q}.
\end{aligned}$$

Возьмем для $k = 0, 1, \ldots, N-1$ математическое ожидание по $e_1, e_2, \ldots, e_N$ от неравенств (6) и сложим получившиеся неравенства, учитывая (П.8):

$$\begin{aligned}
2\alpha_N^2 LC_{n,q}\mathbb{E}[f(y_N)] + \sum_{k=1}^{N-1}\tfrac{\mathbb{E}[f(y_k)]}{8LC_{n,q}} + \mathbb{E}[V_{z_N}(u)] - V_{z_0}(u) - \\
-\tfrac{7}{4}C_{n,q}\tilde{\delta}^2\sum_{k=0}^{N-1}\alpha_{k+1}^2 - \sqrt{n}\tilde{\delta}\sum_{k=0}^{N-1}\alpha_{k+1}\mathbb{E}[||u - z_k||_p] \leqslant \sum_{k=0}^{N-1}\alpha_{k+1}f(u)
\end{aligned} \tag{П.9}$$

для любого $u \in \mathbb{R}^n$, здесь также $\tilde{\delta}$ определена в (3). Положим $u = x^*$. Так как $\sum_{k=1}^{N} \alpha_k = \frac{N(N+3)}{8LC_{n,q}}$, $\mathbb{E}[f(y_k)] \geqslant f(x^*)$ и $V_{z_0}(x^*) = V_{x_0}(x^*) = \Theta$, то из (П.9) следует, что

$$\begin{aligned} &\tfrac{(N+1)^2}{8LC_{n,q}}\mathbb{E}[f(y_N)] - \left(\tfrac{N(N+3)}{8LC_{n,q}} - \tfrac{N-1}{8LC_{n,q}}\right) f(x^*) \leqslant \\ &\leqslant \Theta - \mathbb{E}[V_{z_N}(x^*)] + \tfrac{7}{4}C_{n,q}\tilde{\delta}^2 \sum_{k=0}^{N-1} \alpha_{k+1}^2 + \sqrt{n}\tilde{\delta} \sum_{k=0}^{N-1} \alpha_{k+1}\mathbb{E}[R_k], \end{aligned} \tag{П.10}$$

где $R_k \overset{\text{def}}{=} ||x^* - z_k||_p$. После простых преобразований неравенство (П.10) запишется в виде

$$\begin{aligned} &0 \leqslant \tfrac{(N+1)^2}{8LC_{n,q}}(\mathbb{E}[f(y_N)] - f(x^*)) \leqslant \\ &\leqslant \Theta - \mathbb{E}[V_{z_N}(x^*)] + \tfrac{7}{4}C_{n,q}\tilde{\delta}^2 \sum_{k=0}^{N-1} \alpha_{k+1}^2 + \sqrt{n}\tilde{\delta} \sum_{k=0}^{N-1} \alpha_{k+1}\mathbb{E}[R_k], \end{aligned} \tag{П.11}$$

откуда следует еще одно полезное неравенство:

$$\mathbb{E}[V_{z_N}(x^*)] \leqslant \Theta + \frac{7}{4}C_{n,q}\tilde{\delta}^2 \sum_{k=0}^{N-1} \alpha_{k+1}^2 + \sqrt{n}\tilde{\delta} \sum_{k=0}^{N-1} \alpha_{k+1}\mathbb{E}[R_k]. \tag{П.12}$$

Докажем индукцией по N неравенство

$$\begin{aligned} &\Theta + \tfrac{7}{4}C_{n,q}\tilde{\delta}^2 \sum_{k=0}^{N-1} \alpha_{k+1}^2 + \sqrt{n}\tilde{\delta} \sum_{k=0}^{N-1} \alpha_{k+1}\mathbb{E}[R_k] \leqslant \\ &\leqslant \left(\sqrt{\Theta + \tfrac{7}{4}C_{n,q}\tilde{\delta}^2 \sum_{k=0}^{N-1} \alpha_{k+1}^2 + \sqrt{2\Theta n}\tfrac{\tilde{\delta}}{2LC_{n,q}}} + \sqrt{2n}\tfrac{N^2\tilde{\delta}}{4LC_{n,q}}\right)^2. \end{aligned} \tag{П.13}$$

Для $N = 1$ неравенство (П.13) выполнено, так как

$$\begin{aligned} &\Theta + \tfrac{7}{4}C_{n,q}\tilde{\delta}^2\alpha_1^2 + \sqrt{n}\tilde{\delta}\alpha_1 \underbrace{\mathbb{E}[R_0]}_{R_0} \overset{①}{\leqslant} \\ &\overset{①}{\leqslant} \Theta + \tfrac{7}{4}C_{n,q}\tilde{\delta}^2\alpha_1^2 + \sqrt{2\Theta n}\tfrac{\tilde{\delta}}{2LC_{n,q}} \leqslant \\ &\leqslant \left(\sqrt{\Theta + \tfrac{7}{4}C_{n,q}\tilde{\delta}^2\alpha_1^2 + \sqrt{2\Theta n}\tfrac{\tilde{\delta}}{2LC_{n,q}}} + \sqrt{2n}\tfrac{\tilde{\delta}}{4LC_{n,q}}\right)^2, \end{aligned}$$

где ① следует из неравенства $R_0 \leqslant \sqrt{2V_{z_0}(x^*)} = \sqrt{2\Theta}$ ($V_x(y) \geqslant \frac{1}{2}||x - y||_p^2$ в силу сильной выпуклости прокс-функции $d(x)$) и равенства $\alpha_1 \overset{\text{def}}{=} \frac{2}{4LC_{n,q}} = \frac{1}{2LC_{n,q}}$. Таким образом, база индукции доказана. Докажем теперь шаг индукции: предположим, что (П.13) выполнено для некоторого натурального N и докажем, что тогда оно выполнено и для $N + 1$. Во-первых, из предположения индукции и (П.12) следует, что

$$\begin{aligned} &\tfrac{1}{2}(\mathbb{E}[R_N])^2 \leqslant \tfrac{1}{2}\mathbb{E}[R_N^2] \leqslant \mathbb{E}[V_{z_N}(x^*)] \overset{\text{(П.12)}}{\leqslant} \\ &\overset{\text{(П.12)}}{\leqslant} \Theta + \tfrac{7}{4}C_{n,q}\tilde{\delta}^2 \sum_{k=0}^{N-1} \alpha_{k+1}^2 + \sqrt{n}\tilde{\delta} \sum_{k=0}^{N-1} \alpha_{k+1}\mathbb{E}[R_k] \overset{\text{(П.13)}}{\leqslant} \\ &\overset{\text{(П.13)}}{\leqslant} \left(\sqrt{\Theta + \tfrac{7}{4}C_{n,q}\tilde{\delta}^2 \sum_{k=0}^{N-1} \alpha_{k+1}^2 + \sqrt{2\Theta n}\tfrac{\tilde{\delta}}{2LC_{n,q}}} + \sqrt{2n}\tfrac{N^2\tilde{\delta}}{4LC_{n,q}}\right)^2, \end{aligned}$$

откуда следует, что

$$\mathbb{E}[R_N] \leqslant \sqrt{2}\left(\sqrt{\Theta + \frac{7}{4}C_{n,q}\tilde{\delta}^2\sum_{k=0}^{N-1}\alpha_{k+1}^2 + \frac{\sqrt{2\Theta n}\tilde{\delta}}{2LC_{n,q}}} + \sqrt{2n}\frac{N^2\tilde{\delta}}{4LC_{n,q}}\right). \tag{П.14}$$

Тогда получаем оценку

$$\begin{aligned}
&\Theta + \tfrac{7}{4}C_{n,q}\tilde{\delta}^2\sum_{k=0}^{N}\alpha_{k+1}^2 + \sqrt{n}\tilde{\delta}\sum_{k=0}^{N}\alpha_{k+1}\mathbb{E}[R_k] \leqslant \\
&\leqslant \left(\sqrt{\Theta + \tfrac{7}{4}C_{n,q}\tilde{\delta}^2\sum_{k=0}^{N-1}\alpha_{k+1}^2 + \sqrt{2\Theta n}\tfrac{\tilde{\delta}}{2LC_{n,q}}} + \sqrt{2n}\tfrac{N^2\tilde{\delta}}{4LC_{n,q}}\right)^2 + \\
&+\tfrac{7}{4}C_{n,q}\tilde{\delta}\alpha_{N+1}^2 + \sqrt{n}\tilde{\delta}\alpha_{N+1}\mathbb{E}[R_N] \overset{\text{(П.14)}}{\leqslant} \\
&\overset{\text{(П.14)}}{\leqslant} \Theta + \tfrac{7}{4}C_{n,q}\tilde{\delta}^2\sum_{k=0}^{N}\alpha_{k+1}^2 + \sqrt{2\Theta n}\tfrac{\tilde{\delta}}{2LC_{n,q}} + \\
&+2\sqrt{2n}\tfrac{N^2\tilde{\delta}}{4LC_{n,q}}\sqrt{\Theta + \tfrac{7}{4}C_{n,q}\tilde{\delta}^2\sum_{k=0}^{N-1}\alpha_{k+1}^2 + \sqrt{2\Theta n}\tfrac{\tilde{\delta}}{2LC_{n,q}}} + \left(\sqrt{2n}\tfrac{N^2\tilde{\delta}}{4LC_{n,q}}\right)^2 + \\
&+\sqrt{2n}\tilde{\delta}\alpha_{N+1}\left(\sqrt{\Theta + \tfrac{7}{4}C_{n,q}\tilde{\delta}^2\sum_{k=0}^{N-1}\alpha_{k+1}^2 + \tfrac{\sqrt{2\Theta n}\tilde{\delta}}{2LC_{n,q}}} + \sqrt{2n}\tfrac{N^2\tilde{\delta}}{4LC_{n,q}}\right) \overset{①}{\leqslant} \\
&\overset{①}{\leqslant} \Theta + \tfrac{7}{4}C_{n,q}\tilde{\delta}^2\sum_{k=0}^{N}\alpha_{k+1}^2 + \sqrt{2\Theta n}\tfrac{\tilde{\delta}}{2LC_{n,q}} + \\
&+2\sqrt{2n}\tfrac{(N+1)^2\tilde{\delta}}{4LC_{n,q}}\sqrt{\Theta + \tfrac{7}{4}C_{n,q}\tilde{\delta}^2\sum_{k=0}^{N}\alpha_{k+1}^2 + \sqrt{2\Theta n}\tfrac{\tilde{\delta}}{2LC_{n,q}}} + \left(\sqrt{2n}\tfrac{(N+1)^2\tilde{\delta}}{4LC_{n,q}}\right)^2 = \\
&= \left(\sqrt{\Theta + \tfrac{7}{4}C_{n,q}\tilde{\delta}^2\sum_{k=0}^{N}\alpha_{k+1}^2 + \sqrt{2\Theta n}\tfrac{\tilde{\delta}}{2LC_{n,q}}} + \sqrt{2n}\tfrac{(N+1)^2\tilde{\delta}}{4LC_{n,q}}\right)^2,
\end{aligned}$$

где ① следует из неравенств

$$\begin{aligned}
&\tfrac{2N^2}{4LC_{n,q}} + \alpha_{N+1} = \tfrac{2N^2}{4LC_{n,q}} + \tfrac{N+2}{4LC_{n,q}} \leqslant \tfrac{2(N+1)^2}{4LC_{n,q}}, \\
&\left(\sqrt{2n}\tfrac{N^2\tilde{\delta}}{4LC_{n,q}}\right)^2 + \sqrt{2n}\tilde{\delta}\alpha_{N+1}\cdot\sqrt{2n}\tfrac{N^2\tilde{\delta}}{4LC_{n,q}} = \\
&= \left(\sqrt{2n}\tfrac{\tilde{\delta}}{4LC_{n,q}}\right)^2(N^4 + (N+2)N^2) \leqslant \left(\sqrt{2n}\tfrac{(N+1)^2\tilde{\delta}}{4LC_{n,q}}\right)^2, \\
&\sqrt{\Theta + \tfrac{7}{4}C_{n,q}\tilde{\delta}^2\sum_{k=0}^{N-1}\alpha_{k+1}^2 + \sqrt{2\Theta n}\tfrac{\tilde{\delta}}{2LC_{n,q}}} \leqslant \\
&\leqslant \sqrt{\Theta + \tfrac{7}{4}C_{n,q}\tilde{\delta}^2\sum_{k=0}^{N}\alpha_{k+1}^2 + \sqrt{2\Theta n}\tfrac{\tilde{\delta}}{2LC_{n,q}}}.
\end{aligned}$$

Итак, неравенство (П.13) доказано.

Из (П.11), (П.13) и $V_{z_N}(x^*) \geqslant 0$ получаем неравенство

$$
\begin{aligned}
0 \leqslant \tfrac{(N+1)^2}{8LC_{n,q}}(\mathbb{E}[f(y_N)] - f(x^*)) \leqslant \\
\leqslant \left(\sqrt{\Theta + \tfrac{7}{4}C_{n,q}\tilde{\delta}^2\sum_{k=0}^{N-1}\alpha_{k+1}^2 + \sqrt{2\Theta n}\tfrac{\tilde{\delta}}{2LC_{n,q}}} + \sqrt{2n}\tfrac{N^2\tilde{\delta}}{4LC_{n,q}}\right)^2 \overset{①}{\leqslant} \\
\overset{①}{\leqslant} 2\Theta + \tfrac{7}{2}C_{n,q}\tilde{\delta}^2\sum_{k=0}^{N-1}\alpha_{k+1}^2 + \tfrac{\sqrt{2\Theta n}\tilde{\delta}}{LC_{n,q}} + \tfrac{nN^4\tilde{\delta}^2}{4L^2C_{n,q}^2} \overset{②}{\leqslant} \\
\overset{②}{\leqslant} 2\Theta + \tfrac{7\tilde{\delta}^2(N+1)(N+2)(2N+3)}{192L^2C_{n,q}} + \tfrac{\sqrt{2\Theta n}\tilde{\delta}}{LC_{n,q}} + \tfrac{nN^4\tilde{\delta}^2}{4L^2C_{n,q}^2} \overset{③}{=} \\
\overset{③}{=} 2\Theta + \tfrac{7(N+1)(N+2)(2N+3)\delta}{48LC_{n,q}} + \tfrac{2\sqrt{2\Theta n\delta}}{\sqrt{L}C_{n,q}} + \tfrac{nN^4\delta}{LC_{n,q}^2},
\end{aligned}
\tag{П.15}
$$

где ① следует из неравенства $(a+b)^2 \leqslant 2a^2 + 2b^2 \quad \forall\, a, b \in \mathbb{R}$, ② получается из неравенства

$$
\begin{aligned}
\sum_{k=0}^{N-1}\alpha_{k+1}^2 = \sum_{k=0}^{N-1}\tfrac{(k+2)^2}{16L^2C_{n,q}^2} = \tfrac{1}{16L^2C_{n,q}^2}\sum_{k=2}^{N+1}k^2 = \\
= \tfrac{1}{16L^2C_{n,q}^2}\cdot\left(\tfrac{(N+1)(N+2)(2N+3)}{6} - 1\right) \leqslant \tfrac{(N+1)(N+2)(2N+3)}{96L^2C_{n,q}^2},
\end{aligned}
$$

а ③ верно в силу (3). Поделим неравенство (П.15) на $\frac{(N+1)^2}{8LC_{n,q}}$ и получим окончательно, что

$$
\mathbb{E}[f(y_N)] - f(x^*) \leqslant \tfrac{16\Theta LC_{n,q}}{(N+1)^2} + \tfrac{7(N+2)(2N+3)\delta}{6(N+1)} + \tfrac{16\sqrt{2\Theta nL\delta}}{(N+1)^2} + \tfrac{8nN^4\delta}{C_{n,q}(N+1)^2}.
$$

Теорема доказана.

Приводим формулировку теоремы 1 из [17].

Т е о р е м а П.1. Пусть $e \in RS_2^n(1)$, $n \geqslant 8$, $s \in \mathbb{R}^n$, *тогда*

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

$$
\mathbb{E}[\langle s,\, e\rangle^2||e||_q^2] \leqslant \sqrt{3}||s||_2^2\min\{2q-1, 32\ln n - 8\}n^{\frac{2}{q}-2}, \quad 2 \leqslant q \leqslant \infty,
\tag{П.17}
$$

где под знаком $||\cdot||_q$ *понимается векторная q-норма.*

Кроме того, 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 остается верным для произвольного вектора $s \in \mathbb{R}^n$ вместо $\nabla f(x)$).

Л е м м а П.1. Пусть $e \in RS_2^n(1)$ *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. *Rosenbrock H.H.* An Automatic Method for Finding the Greatest or Least Value of a Function // Comput. J. 1960. V. 3. Iss. 3. P. 175–184. doi: 10.1093/comjnl/3.3.175.

2. *Brent R.P.* Algorithms for Minimization Without Derivatives // Dover Books on Mathematics. Dover Publications, 1973. ISBN 9780486419985. URL: https://books.google.de/books?id=6Ay2biHG-GEC.

3. *Spall J.C.* Introduction to Stochastic Search and Optimization. 1 edition. N.Y.: John Wiley & Sons, Inc., 2003.

4. *Rumelhart D.E., Hinton G.E., Williams R.J.* Learning Representations by Back-Propagating errors // Nature. 1986. № 323. P. 533–536.

5. *Schmidhuber J.* Deep Learning in Neural Networks: An Overview // Neural Networks. 2015. V. 61. P. 85–117. arXiv:1404.7828

6. *Goodfellow I., Bengio Y., Courville A.* Deep Learning. Cambridge: MIT Press, 2016.

7. *Николенко С., Кадурин А., Архангельская Е.* Глубокое обучение. Погружение в мир нейронных сетей. СПб.: Питер, 2018.

8. *Nesterov Yu.* Random Gradient-Free Minimization of Convex Functions // Université catholique de Louvain, Center for Operations Research and Econometrics (CORE). № 2011001, 2011.

9. *Nesterov Yu., Spokoiny V.* Random Gradient-Free Minimization of Convex Functions // Found. Comput. Math. 2017. V. 17. Iss. 2. P. 527–566.

10. *Nesterov Yu.E.* A Method of Solving a Convex Programming Problem with Convergence Rate $O(1/k^2)$ // Soviet Math. Dokl. 1983. V. 27. № 2. P. 372–376.

11. *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.

12. *Chopra P.* Reinforcement Learning without Gradients: Evolving Agents using Genetic Algorithms. URL: https://towardsdatascience.com/reinforcement-learning-without-gradients-evolving-agents-using-genetic-algorithms-

13. *Воронцова Е.А., Гасников А.В., Горбунов Э.А.* Ускоренные спуски по случайному направлению с неевклидовой прокс-структурой // АиТ. 2019. № 4. С. 126–143. arXiv:1710.00162.

14. *Allen-Zhu Z., Orecchia L.* Linear Coupling: An Ultimate Unification of Gradient and Mirror Descent // arXiv preprint arXiv:1407.1537.

15. *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.

16. *Немировский А.С., Юдин Д.Б.* Сложность задач и эффективность методов оптимизации. М.: Наука, 1979.

17. *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.

18. *Bogolubsky L., Dvurechensky P., Gasnikov A., Gusev G., Nesterov Y., Raigorodskii A., Tikhonov A., Zhukovskii M.* Learning Supervised PageRank with Gradient-Based and Gradient-Free Optimization Methods // D.D. Lee, M. Sugiyama, U.V. Luxburg, I. Guyon, and R. Garnett, ed. Proc. NIPS'16 Proc. 30th Int. Conf. on Neural Information Processing Systems. P. 4914–4922. Curran Associates, Inc., 2016. arXiv:1603.00717.

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

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

Список подрисуночных подписей

Рис. 1. Сходимость метода ACDF для функции (10), размерность $n = 10$. Показана практическая зависимость точности нахождения минимума $f(y_N) - f(x^*)$ от числа итераций N алгоритма — темный график и теоретическая оценка $O\left(\frac{16\Theta LC_{n,q}}{N^2}\right)$ — светлый график.

Рис. 2. Сходимость метода ACDF для функции (10), размерность $n = 10^3$, точность $\varepsilon = 10^{-4}$. Показана практическая зависимость точности нахождения минимума $f(y_N) - f(x^*)$ от числа итераций N алгоритма (a), график *1*). Также для сравнения на рис. 2, *а* приведены результаты работы метода при других p (евклидова норма — график *2*; $p = 1,8$ — график *3*; $p = 1,9$ — график *4*) при одних и тех же генерируемых e_k и точке старта x_0. На рис. 2, *б* приведены результаты сравнения на той же задаче работы метода ACDF (пунктирная линия) и ускоренного неевклидового спуска ACDS [13] (сплошная линия, график *1*), предназначенного для работы с точным оракулом первого порядка. Также на рис. 2, *б* приведены результаты работы указанных методов при $p = 2$ (график *2*, штрихпунктирная линия —ACDF, сплошная линия —ACDS).

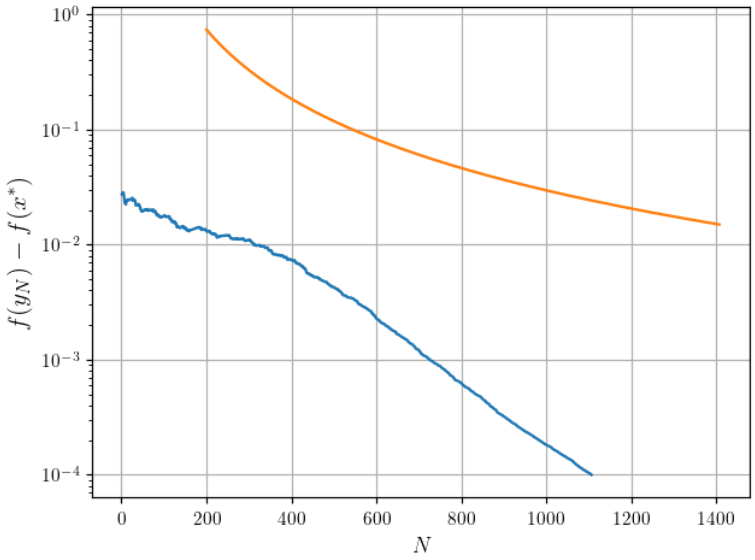
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
$f(y_N) - f(x^*)$
0
200
400
600
800
1000
1200
1400
N

Рис. 1.

а *б*

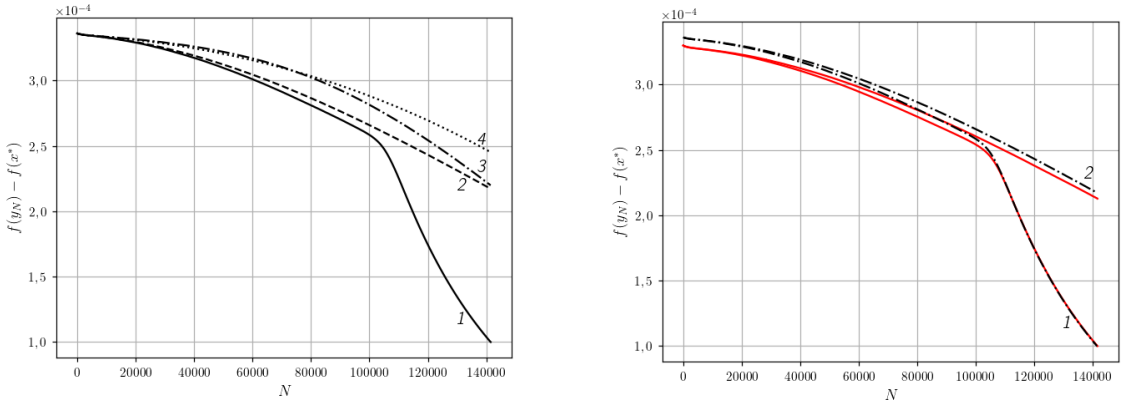
$\times 10^{-4}$
3,0
2,5
2,0
1,5
1,0
$f(y_N) - f(x^*)$
0
20000
40000
60000
80000
100000
120000
140000
N
1
2
3
4

Рис. 2.