ГОУ ВПО «Московский физико-технический институт (государственный университет)»

Факультет управления и прикладной математики

Кафедра проблем передачи информации и анализа данных, ИППИ РАН

Работа допущена к защите

зав. кафедрой

_____ Соболевский А. Н.

«____» _____ 2018 г.

# Выпускная квалификационная работа на соискание степени

# БАКАЛАВРА

Тема: **Ускоренные спуски по случайному направлению и безградиентные методы с неевклидовой прокс-структурой**

Направление: 03.03.01 – Прикладные математика и физика

Выполнил студент гр. 477б _____ Горбунов Эдуард Александрович

Научный руководитель,

д. ф.-м. н. _____ Гасников А. В.

Москва – 2018

# Оглавление

## 1. Введение

Безградиентная оптимизация имеет долгую историю, начиная с 1960 года [1] (см. также [2–4]), а методы первого порядка были впервые рассмотрены ещё в 19 веке [5]. В данной работе рассматривается в некотором смысле промежуточный класс задач оптимизации, который как мы увидим далее изучен не достаточно хорошо, несмотря на долгую историю отмеченных классов. Будем предполагать, что на каждой итерации мы имеем доступ к производной целевой функции по любому направлению. Такого рода методы называют обычно *спусками по случайному направлению*. Отметим, что области безградиентной оптимизации и стохастической оптимизации первого порядка развиты достаточно хорошо, в то время как спуски по случайному направлению для задач стохастической оптимизации представлены в литературе в недостаточном объёме и в этой области остаются открытые вопросы. Однако рассматривать этот класс методов могут мотивировать следующие три ситуации.

Первый сюжет связан с автоматическим дифференцированием [6]. Предположим, что целевая функция задаётся некоторым деревом вычислений, включающим в себя элементарные арифметические операции и элементарные оценки функций. Автоматическое дифференцирование позволяет в таком случае вычислять градиент целевой функции, причём дополнительно требует произвести не более, чем в 5 раз больше элементарных операций, чем при вычислении значения целевой функции. Однако у такого подхода есть недостаток: он требует хранить в памяти результаты всех промежуточных вычислений, которые получались при вычислении значения функции. Напротив, вычисление производной по направлению — это более простая операция, чем вычисление полного градиента, и требует тот же по порядку объём памяти, что и при вычислении значения целевой функции [7]. Так как некоторый случайный вектор может быть частью входа программы, вычисляющей функцию, или некоторая случайность может быть использована внутри самого алгоритма, то естественно рассматривать задачи стохастической оптимизации.

Второй сюжет связан с квази-вариационными неравенствами, которые используются при моделировании различных явлений, таких как формирование и рост песчанников [8], определение озёр и речных сетей [9], и сверхпроводимость [10]. Оказывается, что в таких ситуациях производные по направлению могут вычисляться [11] как решения вспомогательных задач. Эти подзадачи могут быть решены точно не

всегда, откуда естественным образом берётся некоторый шум в вычисленных производных по направлению. Если рассматриваемое физическое явление происходит в некоторых случайных средах, стохастическая оптимизация кажется вполне естественным подходом к решению.

Третий сюжет связан с задачами безградиентной стохастической оптимизации. В этой ситуации используется аппроксимация градиента, построенная на основе конечных разностей стохастических аппроксимаций значений целевой функции в двух близких точках, которая может быть рассмотрена как зашумлённая производная по направлению, которое задаётся разностью между этими точками [12]. В таком случае безградиентная стохастическая оптимизация может быть рассмотрена как частный случай класса методов, использующих производные по направлению для задач стохастической оптимизации.

Мотивируясь тем, что в используемой производной по направлению может присутствовать шум нестохастической природы (см. второй пример), мы будем предполагать, что шум состоит из двух частей. Как и в задачах стохастической оптимизации первая часть имеет стохастическую природу. Напротив, вторая часть — это шум неизвестной природы, но ограниченный по своему абсолютному значению. Более формально: рассмотрим следующую задачу оптимизации

$$\min_{x \in \mathbb{R}^n} \left\{ f(x) := \mathbb{E}_\xi[F(x,\xi)] = \int_{\mathcal{X}} F(x,\xi) dP(x) \right\}, \tag{1}$$

где $\xi$ — случайная величина с функцией распределения $P(\xi)$, $\xi \in \mathcal{X}$, и для $P$-почти всех $\xi \in \mathcal{X}$, функция $F(x,\xi)$ замкнута, а функция $f(x)$ — выпукла (отметим, что достаточно потребовать выпуклость только мат. ожидания, а не почти всех реализаций). Более того, предположим, что для $P$-почти всех $\xi$, у функции $F(x,\xi)$ существует градиент $g(x,\xi)$, который является липшицевым с константой $L(\xi)$ в евклидовой норме, и $L_2 := \sqrt{\mathbb{E}_\xi L(\xi)^2} < +\infty$. При сделанных предположениях $\mathbb{E}_\xi g(x,\xi) = \nabla f(x)$ и $f$ имеет липшицев градиент с константой $L_2$ в евклидовой норме. Кроме того, мы предполагаем, что

$$\mathbb{E}_\xi[\|g(x,\xi) - \nabla f(x)\|_2^2] \leqslant \sigma^2, \tag{2}$$

где $\|\cdot\|_2$ обозначает евклидову норму.

Наконец, будем предполагать, что оракул, который используется в нашем методе, при заданных точке $x \in \mathbb{R}^n$, направлении $e \in S_2(1)$ и $\xi$, которые сэмплиру-

ются независимо от всех предыдущих итераций из распределения $P$, возвращает зашумлённую стохастическую аппроксимацию $\widetilde{f}'(x,\xi,e)$ производной по направлению $\langle g(x,\xi),e\rangle$:

$$\begin{aligned}\widetilde{f}'(x,\xi,e) &= \langle g(x,\xi),e\rangle + \zeta(x,\xi,e) + \eta(x,\xi,e),\\ \mathbb{E}_\xi(\zeta(x,\xi,e))^2 &\leqslant \Delta_\zeta,\ \forall x\in\mathbb{R}^n, \forall e\in S_2(1),\\ |\eta(x,\xi,e)| &\leqslant \Delta_\eta,\ \forall x\in\mathbb{R}^n, \forall e\in S_2(1),\ \text{для п.в. } \xi,\end{aligned} \tag{3}$$

где $S_2(1)$ — единичная евклидова сфера в $\mathbb{R}^n$ с центром в начале координат, а значения $\Delta_\zeta$ и $\Delta_\eta$ контролируются нами и могут быть сделаны настолько малыми, насколько нам нужно. Отметим, что мы используем гладкость $F(\cdot,\xi)$ чтобы записать производную по направлению в виде $\langle g(x,\xi),e\rangle$, однако мы *не предполагаем*, что имеем доступ ко всему стохастическому градиенту $g(x,\xi)$.

Хорошо известно [13–15], что если доступен стохастический градиент $g(x,\xi)$, то ускоренный градиентный метод требует $O\left(\max\left\{\sqrt{L_2/\varepsilon}, \sigma^2/\varepsilon^2\right\}\right)$ итераций, чтобы достичь $\varepsilon$-решения по функции в среднем. В работе даётся ответ на следующий вопрос.

*Можно ли решать задачи гладкой стохастической оптимизации с той же самой зависимостью числа итераций от $\varepsilon$, используя только зашумлённые производные по направлению?*

### 1.1. Близкие работы

Рассмотрим сначала связанные с нашим методом работы по спускам по случайному направлению и близкий к ним класс безградиентных двухточечных методов (последнее означает, что рассматриваемые методы на каждой итерации используют два значения (возможно, зашумлённые) функции. Поскольку все рассматриваемые методы рандомизированные, мы рассматриваем оценки оракульной сложности по математическому ожиданию, то есть число обращений к оракулу за производной по направлению, которое гарантирует, что для заданной точности $\varepsilon$ выполняется $\mathbb{E}f(\hat{x}) - f^* \leqslant \varepsilon$, где $\hat{x}$ — это точка, которую возвращает алгоритм и $f^*$ — оптимальное значение функции $f$.

### Спуски по направлению

**Детерминированные гладкие задачи оптимизации.** В работе [16] авторы рассмотрели евклидов случай и предложили неускоренный и ускоренный спуски по направлениям для гладких выпуклых задач и доказали оценки на сложность $O(nL_2/\varepsilon)$ и $O(n\sqrt{L_2/\varepsilon})$ соответственно. Кроме того, они предложили неускоренный и ускоренный методы для минимизации сильно выпуклых функций, имеющие сложность $O(nL_2/\mu\log_2(1/\varepsilon))$ и $O(n\sqrt{L_2/\mu}\log_2(1/\varepsilon))$ соответственно. Для более общего случая, когда производные по направлению известны с аддитивным ограниченным шумом, но по-прежнему для евклидового случая, был предложен ускоренный спуск по случайному направлению [12] и доказана оценка на сложность $O(n\sqrt{L_2/\varepsilon})$.

Также стоит отметить координатные спуски. В плодотворной работе [17] был предложен рандомизированый координатный спуск для задач гладкой выпуклой и $\mu$-сильно выпуклой оптимизации и были доказаны оценки $O(L/\varepsilon)$ и $O(L/\mu\log_2(1/\varepsilon))$ на сложность, где $L$ — эффективная константа Липшица градиента, которая меняется от $n$ до некоторого среднего покоординатных констант Липшица. В той же статье был предложен ускоренный рандомизированный координатный спуск для выпуклых задач и доказана оценка $O(n\sqrt{L/\varepsilon})$ на сложность. Работы [18–21] обобщают ускоренный рандомизированный координатный спуск на различных задач, включая $\mu$-сильно выпуклые задачи, и [22–24] дают оценки $O(\sqrt{L/\varepsilon})$ и $O(\sqrt{L/\mu}\log_2(1/\varepsilon))$ на сложность, где $L$ — эффективная константа Липшица градиента, которая меняется от $n$ до некоторого среднего покоординатных констант Липшица и которая в лучшем случае не зависит от размерности. Ускоренный рандоизированный координатный спуск, использующий неточные частные производные был предложен в [12], была доказана оценка $O(n\sqrt{L/\varepsilon})$ на сложность и, кроме того, было представлен универсальный взгляд на спуски по направлениям, координатные спуски и безградиентные методы.

**Задачи стохастической оптимизации.** Спуск по направлениям для задач неглакой стохастической выпуклой оптимизации был предложен в [16] с оценкой $O(n^2/\varepsilon^2)$ на сложность. Рандомизированный координатный спуск для задач негладкой стохастической выпуклой и $\mu$-сильно выпуклой оптимизации был представлен в [25] и доказаны оценки $O(n/\varepsilon^2)$ и $O(n/\mu\varepsilon)$ на сложность соответственно.

**Безградиентные методы**

**Детерминированные гладкие задачи оптимизации.** Для данных типов задач в работе [16] были предложены неускоренный и ускоренный безградиентные методы для евклидового случая с оценками $O(nL_2/\varepsilon)$ и $O(n\sqrt{L_2/\varepsilon})$ на сложность соответственно. В той же статье были представлены неускоренный и ускоренный методы для $\mu$-сильно выпуклых задач с оценками на сложность $O(nL_2/\mu \log_2(1/\varepsilon))$ и $O(n\sqrt{L_2/\mu}\log_2(1/\varepsilon))$ соответственно. Неускоренный безградиентный метод для детерминированных задач, в которых значения функции доступны лишь с некоторым аддитивным шумом, был предложен в [26] вместе с оценкой на сложность $O(nL_2/\varepsilon)$ и применении к параметрической PageRank модели. Кроме того, детерминированные задачи, в которых значение функции известно также с аддитивным ограниченным по абсолютной величине шумом, также были рассмотрены в работе [12], в которой было предложено несколько ускоренных безградиентных методов, включая безградиентный блочно-координатный спуск, и доказана оценка на сложность $O(n\sqrt{L/\varepsilon})$, где $L$ зависит от метода и, в некотором смысле, характирует среднее по блочно-координатным константам Липшица производной из блока.

**Задачи стохастической оптимизации.** Многие авторы в этой области рассматривали так называемую задачу bandit convex optimization и получали оценки, так называемого, регрета. Хорошо известно [27], что оценка на регрет может дать оценку средней ошибки. Задачи негладкой стохастической оптимизации были рассмотрены в [16], где была доказана оценка $O(n^2/\varepsilon^2)$ на сложность для безградиентного метода. Эта оценка была улучшена в работах [28–33] до[1] $\widetilde{O}(n^{2/q}R_p^2/\varepsilon^2)$, где $p \in \{1, 2\}$, $\frac{1}{p} + \frac{1}{q} = 1$ и $R_p$ — радиус множества, на котором работает метод, в $p$-норме $\|\cdot\|_p$. Для негладких $\mu_p$-сильно выпуклых относительно $p$-нормы задач, авторы [30, 32] доказали оценку на сложность $\widetilde{O}(n^{2/q}/(\mu_p\varepsilon))$.

Промежуточный случай, когда использовалось предположение о частичной гладкости задачи с ограничительным предположением ограниченности $\mathbb{E}\,\|g(x,\xi)\|^2$, был рассмотрен в [28], где было доказано, что правильная модификация зеракального спуска с безградиентной аппроксимацией градиента даёт оценку на сложность $O(n^{2/q}R_p^2/\varepsilon^2)$ для выпуклых задач, улучшенную по сравнению с $\widetilde{O}(n^2/\varepsilon^2)$ из [34]. Для сильно выпуклых в 2-норме задач, авторы [34] получили оценку на сложность

[1] $\widetilde{O}$ содержит полилогарифмический множитель $(\ln n)^c$, $c > 0$.

$\widetilde{O}(n^2/\varepsilon)$, которая была расширена на класс $\mu_p$-сильно выпуклых задач и улучшена до $\widetilde{O}(n^{2/q}/(\mu_p\varepsilon))$ в [30].

В полностью глаком случае без предположения $\mathbb{E}\|g(x,\xi)\|^2 < +\infty$, в статьях [35, 36] предложен безградиентный метод для евклидового случая с оценкой на сложность

$$\widetilde{O}\left(\max\left\{\frac{nL_2R_2^2}{\varepsilon}, \frac{n\sigma^2R_2^2}{\varepsilon^2}\right\}\right).$$

### 1.2. Обзор полученных результатов

Как видно выше, по нашим сведениям только два результата по спускам по направлениям для негладких стохастических выпуклых задач оптимизации представлены в литературе, и, насколько нам известно, ничего не было предложено до этого момента по теме спуском по направлению для задач гладкой *стохастической* выпуклой оптимизации, даже в хорошо развитой области рандомизированных координатных спуском. Основной вклад этой работы состоит в закрытии данного пробела в теории спуском по направлению для стохастической оптимизации и рассмотрении даже более общего случая с аддитивным шумом неизвестной природы в производной по направлению.

Наш метод основан на двух прокс-структурах [37], характеризующихся[2] числом $p \in \{1, 2\}$ и его сопряжённым $q \in \{2, \infty\}$, которое задаётся соотношением $\frac{1}{p} + \frac{1}{q} = 1$. Случай $p = 1$ соответствует выбору 1-нормы в $\mathbb{R}^n$ и соответствующей прокс-функции, которая сильно выпукла в данной норме (детали приведены ниже). Случай $p = 2$ соответствует выбору евклидовой 2-нормы в $\mathbb{R}^n$ и квадрату евклидовой нормы в качестве прокс-функции. Основной результат этой работы состоит в том, что мы предложили ускоренный спуск по случайному направлению ARDD (Accelerated Randomized Directional Derivative algorithm) для задач гладкой стохастической оптимизации, использующий зашумлённые значения производных по направлению целевой функции. Наш метод имеет следующую оценку на сложность:

$$\widetilde{O}\left(\max\left\{n^{\frac{1}{2}+\frac{1}{q}}\sqrt{\frac{L_2R_p^2}{\varepsilon}}, \frac{n^{\frac{2}{q}}\sigma^2R_p^2}{\varepsilon^2}\right\}\right), \tag{4}$$

где $R_p$ характеризует расстояние в $p$-норме между стартовой точкой алгоритма и точкой, в которой достигается решение задачи (1). Наш алгоритм в случае $p = 1$, $q = \infty$

[2] Вообще говоря, наш анализ проведён для всех промежуточных случаев $p \in [1, 2]$, но нас не интересуют прокс-структуры, задаваемые числом $p \notin \{1, 2\}$

основан на новой идее комбинирования шага градиентного метода, соответствующего выбору евклидовой прокс-структуры, и шага зеркального спуска, соответствующего выбору 1-нормы для прокс-структуры. Отметим, что использование разных норм и прокс-структур позволяет нам избавиться от множителя $\sqrt{n}$ в оценке сложности в случае $p = 1$ по отношению со случаем, когда выбирается евклидова норма и прокс-структура.

Во-вторых, был предложен неускоренный спуск по случайному направлению RDD (non-accelerated Randomized Directional Derivative algorithm) и доказана оценка на сложность

$$\widetilde{O}\left(\max\left\{\frac{n^{\frac{2}{q}}L_2R_p^2}{\varepsilon}, \frac{n^{\frac{2}{q}}\sigma^2R_p^2}{\varepsilon^2}\right\}\right). \tag{5}$$

Интересен тот факт, что в случае $p = 1$, $q = \infty$ оценка не зависит от размерности пространства, причём алгоритм использует только зашумлённые производные по направлению.

Отметим, что в случае, когда задача (1) имеет разреженное решение, наши оценки для случая $p = 1$ позволяют сократить множитель $\sqrt{n}$ в оценке сложности ускоренного метода и множитель $n$ в оценке сложности неускоренного метода по сравнению с евклидовым случаем $p = 2$. Действительно, разреженность решения $x^*$ означает, что $\|x^*\|_1 = O(1) \cdot \|x^*\|_2$ и, если выбрать стартовой точкой начало координат, мы получим $R_1^2 = \|x^*\|_1^2 = O(1) \cdot \|x^*\|_2^2 = O(1)R_2^2$. Отсюда следует, что оценки для случаев $p = 1$ и $p = 2$ можно сравнивать и оценки для случая $p = 1$, $q = \infty$ получаются лучше отсносительно порядка вхождения размерности пространства $n$.

Отметим, что существенным в нашем алгоримте является тот факт, что случайные направления генерируются из равномерного распределения на единичной евклидовой сфере. Наши результаты нельзя получить, если использовать в случае $p = 1$ рандомизированный координатный спуск.

Кроме того, мы развили результаты, полученные выше, на случаи, когда целевая функция является $\mu$-сильно выпуклой относительно $p$-нормы[3]. Для этого случая были предложены ускоренный и неускоренный алгоритмы и доказаны соответству-

[3] Результаты для сильно выпуклых задач были получены при определяющем вкладе П. Е. Двуреченского.

ющие оценки на сложность:

$$\widetilde{O}\left(\max\left\{n^{\frac{1}{2}+\frac{1}{q}}\sqrt{\frac{L_2}{\mu_p}}\log_2\frac{\mu_p R_p^2}{\varepsilon}, \frac{n^{\frac{2}{q}}\sigma^2}{\mu_p\varepsilon}\right\}\right), \quad \widetilde{O}\left(\max\left\{\frac{n^{\frac{2}{q}}L_2}{\mu_p}\log_2\frac{\mu_p R_p^2}{\varepsilon}, \frac{n^{\frac{2}{q}}\sigma^2}{\mu_p\varepsilon}\right\}\right). \tag{6}$$

Наконец, мы рассмотрели задачи безградиентной гладкой стохастической оптимизации, в которых используются лишь зашумлённые стохастические аппроксимации значений целевой функции, как частный случай задач оптимизации, в которых используется зашумлённые производные по направлению. Были получены в этом случае и новые оценки на сложность для сильно выпуклых задач, которых до этого в литературе не встречалось.

Обычно задачи стохастической оптимизации решаются на выпуклых замкнутых множествах, существенно опираясь на то, что у множества есть некоторый диаметр $R$ [14, 15]. Мы же рассматриваем задачи стохастической оптимизации на всём пространстве (диаметр неограничен). Более того, мы решаем эти задачи используя оракул нулевого порядка с аддитивным шумом стохастичсекой природы. Все эти вещи делают рассмотрение гораздо более трудоёмким. Кроме того, мы разработали технику (см. Леммы 12, 13), которая может быть полезна также и в более простых ситуациях и позволяет получить новые результаты даже в таких ситуациях (например, для задач с нулевым шумом, использующих градиент целиком).

## 2. Алгоритмы и основные результаты

Здесь предсталены неускоренный и ускоренный методы для выпуклых и сильно выпуклых задач, а также приведены теоремы об оценках сложности алгоритмов. Доказательства можно найти[4] в [40]. Для удобства все доказательства также приведены в разделах 4-6.

### 2.1. Обозначения и вспомогательные факты

**Прокс-структура.** Пусть $p \in [1, 2]$ и $\|x\|_p$ — это $p$-норма вектора $x \in \mathbb{R}^n$, которая задаётся равенством

$$\|x\|_p^p = \sum_{i=1}^n |x_i|^p, \quad x \in \mathbb{R}^n,$$

[4] Данная работа — это результат цикла статей: [38–41]

$\|\cdot\|_q$ — это сопряжённая к ней норма, определяемая равенством $\|g\|_q = \max\limits_x \left\{\langle g, x\rangle, \|x\|_p \leq 1\right\}$, причём $q \in [2, \infty]$ — это сопряжённое к $p$ число, то есть $\frac{1}{p} + \frac{1}{q} = 1$. Для $q = \infty$ по определению будем считать $\|x\|_\infty = \max\limits_{i=1,\ldots,n} |x_i|$.

Рассмотрим *прокс-функцию* $d(x)$, то есть такую функцию, которая непрерывна и 1-сильно выпукла на $\mathbb{R}^n$ относительно нормы $\|\cdot\|_p$, то есть для любых $x, y \in \mathbb{R}^n$ $d(y) - d(x) - \langle \nabla d(x), y - x\rangle \geq \frac{1}{2}\|y - x\|_p^2$. Не умаляя общности, будем считать, что $\min\limits_{x \in \mathbb{R}^n} d(x) = 0$. Определим соответствующую *дивергенцию Брегмана* $V[z](x) = d(x) - d(z) - \langle \nabla d(z), x - z\rangle$, $x, z \in \mathbb{R}^n$. Отметим, что из сильной выпуклости $d$ следует

$$V[z](x) \geq \frac{1}{2}\|x - z\|_p^2, \quad x, z \in \mathbb{R}^n. \tag{7}$$

Для случая $p = 1$ мы рассмотрим следующую прокс-функцию [37, 42]

$$d(x) = \frac{en^{(\kappa-1)(2-\kappa)/\kappa} \ln n}{2}\|x\|_\kappa^2, \quad \kappa = 1 + \frac{1}{\ln n} \tag{8}$$

и для случая $p = 2$ в качестве прокс-функции мы выберем функцию, пропорциональную евклидовой норме с коэффициентом $\frac{1}{2}$:

$$d(x) = \frac{1}{2}\|x\|_2^2. \tag{9}$$

**Основная лемма.** В доказательствах наших теорем мы существенно опираемся на следующую лемму, доказанную в [39, 41].

**Лемма 1.** *Пусть $e \in RS_2(1)$ — случайный вектор из равномерного распределения на единичной евклидовой сфере в $\mathbb{R}^n$, $p \in [1, 2]$ и $q$ задаётся соотношением $\frac{1}{p} + \frac{1}{q} = 1$. Тогда для всех $n \geqslant 8$ и $\rho_n = \min\{q - 1, 16 \ln n - 8\} n^{\frac{2}{q}-1}$ выполняются следующие неравенства*

$$\mathbb{E}_e \|e\|_q^2 \leq \rho_n, \tag{10}$$

$$\mathbb{E}_e \left(\langle s, e\rangle^2 \|e\|_q^2\right) \leq \frac{6\rho_n}{n}\|s\|_2^2, \quad \forall s \in \mathbb{R}^n. \tag{11}$$

Данный факт и сам по себе заслуживает отдельного внимания, так как связан с таким интересным явлением как концентрация равномерной меры на евклидовой сфере. Рассмотрим следующую интерпретацию, проливающей свет на связь с указанным явлением.

Пусть задан некоторый (неслучайный) вектор $s$ с единичной евклидовой сферы. Не умаляя общности, мы будем считать, что вектор $s$ направлен вдоль первой

координатной оси (если это не так, то мы можем перейти к нужному базису). Тогда с вероятностью хотя бы $1 - \frac{2}{c}e^{-\frac{c^2}{2}}$ будет выполнено неравенство $|\langle s,\, e\rangle| \leqslant \frac{c}{\sqrt{n-1}}$ (см. теорему 2.7 и рисунок 2.2 из [43] и [44]). То есть, если взять $c = 10$, то получим, что с большой вероятностью выполнено неравенство $\langle s,\, e\rangle^2 \leqslant \frac{100}{n}$ (множество, на котором $\langle s,\, e\rangle^2 \leqslant \frac{100}{n}$, обозначим через $A_s$; как мы видим, при достаточно больших $n$ вероятностная мера множества $A_s$ велика). Кроме того, можно показать, что $\mathbb{E}[\langle s,\, e\rangle^2] = \frac{1}{n}$ (см., например, лемму B.10 из [26]).

Рассмотрим $\infty$-норму, которая для произвольного вектора $x \in \mathbb{R}^n$ задаётся формулой $\|x\|_\infty = \max\limits_{1 \leqslant i \leqslant n} |x_i|$, где $x = (x_1,\, x_2,\, \ldots,\, x_n)^\top$. Заметим, что функция $f(e) = \|e\|_\infty$ является липшицевой с константой 1 в евклидовой норме. Рассмотрим константу $M_f$ такую, что $\mathbb{P}_e\{f(e) \geqslant M_f\} \geqslant \frac{1}{2}$ и $\mathbb{P}_e\{f(e) \leqslant M_f\} \geqslant \frac{1}{2}$. Тогда верно неравенство (см. [45], [46])

$$\mathbb{P}_e\{|f(e) - M_f| > t\} \leqslant 4e^{-\frac{t^2}{4}}, \quad t > 0.$$

Это означает, что случайная величина $\|e\|_\infty$ принимает очень близкие к $\mathbb{E}[\|e\|_\infty]$ ($M_f$ и $\mathbb{E}[f(e)]$ асимптотически близки, см. [47]) значения на множестве достаточно большой меры. Кроме того, можно показать, что максимальная по модулю компонента вектора $e$ с вероятностью не меньше $1 - \frac{1}{n\sqrt{n}}$ принимает значения по модулю меньшие $\frac{2\sqrt{\ln n}}{\sqrt{n-1}}$ (множество, на котором $\|e\|_\infty \leqslant \frac{2\sqrt{\ln n}}{\sqrt{n-1}}$, обозначим через $B_\infty$). Тогда $\mathbb{E}[\langle s,\, e\rangle^2 \|e\|_\infty^2]$ близко к среднему значению случайной величины $\langle s,\, e\rangle^2 \|e\|_\infty^2$ на множестве $A_e \cap B_\infty$ (чья вероятностная мера по-прежнему велика), на котором она не превосходит $400 \ln n / n^2$. Константа в этой оценке сильно завышена. Однако такого рода рассуждения, вытекающие из явления концентрации равномерной меры на сфере, поясняют причины возникновения такой оценки, а также её целесообразность в терминах вхождения размерности пространства $n$.

**Стохастическая аппроксимация градиента.** Используя зашумлённые стохастические (3) производные по направлениям, мы сформируем следующую стохастическую аппроксимацию $\nabla f(x)$

$$\widetilde{\nabla}^m f(x) = \frac{1}{m}\sum_{i=1}^{m} \widetilde{f}'(x, \xi_i, e)e, \tag{12}$$

где $e \in RS_2(1)$, $\xi_i$, $i = 1, ..., m$ — независимые реализации $\xi$, $m$ — *размер батча*.

### 2.2. Алгоритмы и основные результаты для выпуклых задач

Наш ускоренный спуск по случайному направлению (ARDD) обозначен в тексте как Algorithm 1. Отметим, что $y_{k+1}$ обозначает градиентный шаг из точки $x_{k+1}$ и $z_{k+1}$ обозначает шаг зеркального спуска из $z_k$. Отсюда видно, что наш алгоритм для $p = 1$, $q = \infty$ основан на новой идее комбинирования градиентного шага в евклидовой прокс-структуре и шага зеркального спуска в прокс-структуре, связанной с 1-нормой[5]. Такое комбинирование позволяет нам избавиться от мультипликативной константы $\sqrt{n}$ в оценке сложности для случая $p = 1$ по сравнению со стандартным выбором $p = 2$.

**Algorithm 1** Accelerated Randomized Directional Derivative (ARDD) method

**Input:** $x_0$ —стартовая точка; $N \geqslant 1$ — число итераций; $m \geqslant 1$ — размер батча.

**Output:** точка $y_N$.

1: $y_0 \leftarrow x_0$, $z_0 \leftarrow x_0$.

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

3: $\alpha_{k+1} \leftarrow \frac{k+2}{96n^2\rho_n L_2}$, $\tau_k \leftarrow \frac{1}{48\alpha_{k+1}n^2\rho_n L_2} = \frac{2}{k+2}$.

4: Сгенерировать $e_{k+1} \in RS_2(1)$ независимо от предыдущих итераций и $\xi_i$, $i = 1, ..., m$ – независимые реализации $\xi$.

5: Вычислить

$$\widetilde{\nabla}^m f(x_{k+1}) = \frac{1}{m}\sum_{i=1}^{m} \widetilde{f}'(x_{k+1}, \xi_i, e_{k+1}) e_{k+1}.$$

6: $x_{k+1} \leftarrow \tau_k z_k + (1-\tau_k) y_k$.

7: $y_{k+1} \leftarrow x_{k+1} - \frac{1}{2L_2}\widetilde{\nabla}^m f(x_{k+1})$.

8: $z_{k+1} \leftarrow \underset{z\in\mathbb{R}^n}{\operatorname{argmin}} \left\{ \alpha_{k+1} n \left\langle \widetilde{\nabla}^m f(x_{k+1}),\, z - z_k \right\rangle + V[z_k](z) \right\}$.

9: **end for**

10: **return** $y_N$

**Теорема 1.** *Пусть метод ARDD применяется для решения задачи (1). Тогда*

$$\begin{aligned} \mathbb{E}[f(y_N)] - f(x^*) \leqslant & \frac{384\Theta_p n^2 \rho_n L_2}{N^2} + \frac{4N}{nL_2}\cdot\frac{\sigma^2}{m} + \frac{61N}{24L_2}\Delta_\zeta + \frac{122N}{3L_2}\Delta_\eta^2 \\ & + \frac{12\sqrt{2n\Theta_p}}{N^2}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right) + \frac{N^2}{12n\rho_n L_2}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right)^2, \end{aligned} \tag{13}$$

[5] Идея комбинирования градиентного шага и шага зеркального спуска для детерминированной оптимизации первого порядка была предложена в [48]. Однако в данной работе использовалась одна и та же прокс-структура для обоих шагов.

*где* $\Theta_p = V[z_0](x^*)$ *определяется выбранной прокс-структурой и* $\mathbb{E}[\cdot] = \mathbb{E}_{e_1,\ldots,e_N,\xi_{1,1},\ldots,\xi_{N,m}}[\cdot]$.

Перед тем, как мы рассмотрим неускоренный метод, обсудим выбор параметров $N$, $m$ и уровнях шума в производной по направлению $\Delta_\zeta$, $\Delta_\eta$. Эти параметры выбираются так, чтобы правая часть неравенства (13) была не больше $\varepsilon$. Для простоты мы опустили числовые константы и представили полученные значения в Таблице 1, расположенной ниже. Последняя строчка представляет собой общее число обращений к оракулу $Nm$, то есть число обращений за производной по направлению, которое было заявлено в (4).

Наш неускоренный спуск по случайному направлению (RDD) обозначен в тексте как Algorithm 2.

**Algorithm 2** Randomized Directional Derivative (RDD) method

**Input:** $x_0$ —стартовая точка; $N \geqslant 1$ — число итераций; $m \geqslant 1$ — размер батча.

**Output:** точка $\bar{x}_N$.

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

2: $\alpha \leftarrow \frac{1}{48n\rho_n L_2}$.

3: Сгенерировать $e_{k+1} \in RS_2(1)$ независимо от предыдущих итераций и $\xi_i$, $i = 1, ..., m$ – независимые реализации случайной величины $\xi$.

4: Вычислить

$$\widetilde{\nabla}^m f(x_k) = \frac{1}{m}\sum_{i=1}^{m} \widetilde{f}'(x_k, \xi_i, e_{k+1}) e_{k+1}.$$

5: $x_{k+1} \leftarrow \underset{x\in\mathbb{R}^n}{\operatorname{argmin}} \left\{ \alpha n \left\langle \widetilde{\nabla}^m f(x_k),\, x - x_k \right\rangle + V[x_k](x) \right\}$.

6: **end for**

7: **return** $\bar{x}_N \leftarrow \frac{1}{N}\sum\limits_{k=0}^{N-1} x_k$

**Теорема 2.** *Пусть метод RDD применяется для решения задачи* (1). *Тогда*

$$\begin{aligned}\mathbb{E}[f(\bar{x}_N)] - f(x_*) \;&\leqslant \tfrac{384n\rho_n L_2 \Theta_p}{N} + \tfrac{2}{L_2}\tfrac{\sigma^2}{m} + \tfrac{n}{12L_2}\Delta_\zeta + \tfrac{4n}{3L_2}\Delta_\eta^2 \\ &+ \tfrac{8\sqrt{2n\Theta_p}}{N}\left(\tfrac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right) + \tfrac{N}{3L_2\rho_n}\left(\tfrac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right)^2,\end{aligned} \tag{14}$$

*где* $\Theta_p = V[z_0](x^*)$ *определяется выбранной прокс-структурой и* $\mathbb{E}[\cdot] = \mathbb{E}_{e_1,\ldots,e_N,\xi_{1,1},\ldots,\xi_{N,m}}[\cdot]$.

Перед тем как продолжить, обсудим выбор параметров $N$, $m$ и уровнях шума в производной по направлению $\Delta_\zeta$, $\Delta_\eta$. Эти параметры выбираются так, чтобы правая часть

неравенства (14) была не больше $\varepsilon$. Для простоты мы опустили числовые константы и представили полученные значения в Таблице 2, расположенной ниже. Последняя строчка представляет собой общее число обращений к оракулу $Nm$, то есть число обращений за производной по направлению, которое было заявлено в (5).

### 2.3. Расширения для сильно выпуклых задач

В этом разделе мы будем дополнительно предполагать, что $f$ является $\mu_p$-сильно выпуклой относительно $p$-нормы. Наши алгоритмы и доказательства опираются на следующий факт. Пусть $x_*$ — некоторая фиксированная точка и $x$ — такая случайная точка, что $\mathbb{E}_x\left[\|x - x_*\|_p^2\right] \leqslant R_p^2$, тогда

$$\mathbb{E}_x d\left(\frac{x - x_*}{R_p}\right) \leqslant \frac{\Omega_p}{2}, \tag{15}$$

где $\mathbb{E}_x$ обозначает математическое ожидание по вектору $x$, а $\Omega_p$ определяется следующим образом. Для $p = 1$ и нашего выбора прокс-структуры (8), $\Omega_p = en^{(\kappa-1)(2-\kappa)/\kappa}\ln n = O(\ln n)$ для нашего выбора $\kappa = 1 + \frac{1}{\ln n}$, см. [42, 49]. Для $p = 2$ и нашего выбора прокс-структуры (9), $\Omega_p = 1$. Наш ускоренный спуск по случайному направлению для сильно выпуклых функций (ARDDsc) обозначен в тексте как Algorithm 3.

**Теорема 3.** *Пусть $f$ в задаче (1) является $\mu_p$-сильно выпуклой, и метод ARDDsc применяется для решения этой задачи. Тогда*

$$\mathbb{E}f(u_K) - f^* \leqslant \tfrac{\mu_p R_p^2}{2} \cdot 2^{-K} + 2\Delta. \tag{18}$$

*where* $\Delta = \frac{61N_0}{24L_2}\Delta_\zeta + \frac{122N_0}{3L_2}\Delta_\eta^2 + \frac{12\sqrt{2nR_p^2\Omega_p}}{N_0^2}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right) + \frac{N_0^2}{12n\rho_n L_2}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right)^2$. *Более того, если параметры $\Delta_\zeta$ и $\Delta_\eta$ выбираются так, что $2\Delta \leqslant \varepsilon/2$, число обращений к оракулу для достижения $\varepsilon$-решения по функции в среднем:*

$$\widetilde{O}\left(\max\left\{n^{\frac{1}{2}+\frac{1}{q}}\sqrt{\frac{L_2\Omega_p}{\mu_p}}\log_2\frac{\mu_p R_p^2}{\varepsilon}, \frac{n^{\frac{2}{q}}\sigma^2\Omega_p}{\mu_p\varepsilon}\right\}\right).$$

Перед тем, как мы рассмотрим неускоренный метод для сильно выпуклых задач, обсудим выбор параметров $N$, $m$ и уровнях шума в производной по направлению $\Delta_\zeta$, $\Delta_\eta$. Эти параметры выбираются так, чтобы правая часть неравенства (18) была не больше $\varepsilon$. Для простоты мы опустили числовые константы и представили полученные значения в Таблице 3, расположенной ниже. Последняя строчка представляет собой

**Algorithm 3** Accelerated Randomized Directional Derivative method for strongly convex functions (ARDDsc)

**Input:** $x_0$ —стартовая точка, такая что $\|x_0 - x_*\|_p^2 \leq R_p^2$; $K \geqslant 1$ — число итераций; $\mu_p$ – константа сильной выпуклости в $p$-норме.

**Output:** точка $u_K$.

1: Задать

$$N_0 = \left\lceil \sqrt{\frac{8aL_2\Omega_p}{\mu_p}} \right\rceil, \tag{16}$$

где $a = 384n^2\rho_n$.

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

3: Задать

$$m_k := \max\left\{1, \left\lceil \frac{8b\sigma^2 N_0 2^k}{L_2\mu_p R_p^2} \right\rceil \right\}, \quad R_k^2 := R_p^2 2^{-k} + \frac{4\Delta}{\mu_p}\left(1 - 2^{-k}\right), \tag{17}$$

где $b = \frac{4}{n}$.

4: Задать $d_k(x) = R_k^2 d\left(\frac{x-u_k}{R_k}\right)$.

5: Запустить ARDD из стартовой точки $u_k$ с прокс-функцией $d_k(x)$ на $N_0$ итераций с размером батча $m_k$.

6: Задать $u_{k+1} = y_{N_0}$, $k = k+1$.

7: **end for**

8: **return** $u_K$

общее число обращений к оракулу $Nm$, то есть число обращений за производной по направлению, которое было заявлено в (6).

Наш неускоренный спуск по случайному направлению (RDDsc) обозначен в тексте как Algorithm 4.

---

**Algorithm 4** Randomized Directional Derivative method for strongly convex functions (RDDsc)

---

**Input:** $x_0$ —стартовая точка, такая что $\|x_0 - x_*\|_p^2 \leq R_p^2$; $K \geqslant 1$ — число итераций; $\mu_p$ – константа сильной выпуклости в $p$-норме.

**Output:** точка $u_K$.

1: Задать

$$N_0 = \left\lceil \frac{8aL_2\Omega_p}{\mu_p} \right\rceil, \tag{19}$$

где $a = 384n\rho_n$.

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

3: Задать

$$m_k := \max\left\{1, \left\lceil \frac{8b\sigma^2 2^k}{L_2\mu_p R_p^2} \right\rceil \right\}, \quad R_k^2 := R_p^2 2^{-k} + \frac{4\Delta}{\mu_p}\left(1 - 2^{-k}\right), \tag{20}$$

где $b = 2$

4: Задать $d_k(x) = R_k^2 d\left(\frac{x-u_k}{R_k}\right)$.

5: Запустить RDD из стартовой точки $u_k$ с прокс-функцией $d_k(x)$ на $N_0$ итераций с размером батча $m_k$.

6: Задать $u_{k+1} = y_{N_0}$, $k = k+1$.

7: **end for**

8: **return** $u_K$

---

**Теорема 4.** *Пусть $f$ в задаче (1) является $\mu_p$-сильно выпуклой, и метод RDDsc применяется для решения этой задачи. Тогда*

$$\mathbb{E}f(u_K) - f^* \leqslant \tfrac{\mu_p R_p^2}{2} \cdot 2^{-K} + 2\Delta. \tag{21}$$

*где $\Delta = \frac{n}{12L_2}\Delta_\zeta + \frac{4n}{3L_2}\Delta_\eta^2 + \frac{8\sqrt{2nR_p^2\Omega_p}}{N_0}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right) + \frac{N_0}{3L_2\rho_n}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right)^2$. Более того, если параметры $\Delta_\zeta$ и $\Delta_\eta$ выбираются так, что $2\Delta \leqslant \varepsilon/2$, число обращений к оракулу*

*для достижения $\varepsilon$-решения по функции в среднем:*

$$\widetilde{O}\left(\max\left\{\frac{n^{\frac{2}{q}}L_2\Omega_p}{\mu_p}\log_2\frac{\mu_p R_p^2}{\varepsilon}, \frac{n^{\frac{2}{q}}\sigma^2\Omega_p}{\mu_p\varepsilon}\right\}\right).$$

Обсудим выбор параметров $N$, $m$ и уровнях шума в производной по направлению $\Delta_\zeta$, $\Delta_\eta$. Эти параметры выбираются так, чтобы правая часть неравенства (21) была не больше $\varepsilon$. Для простоты мы опустили числовые константы и представили полученные значения в Таблице 4, расположенной ниже. Последняя строчка представляет собой общее число обращений к оракулу $Nm$, то есть число обращений за производной по направлению, которое было заявлено в (6).

### 2.4. Следствия для задач безградиентной оптимизации

В этом разделе, следуя [39], рассматриваются задачи безградиентной гладкой стохастической выпуклой оптимизации с двухточечным оракулом. Предположим, что метод может для любой пары точек $(x, y) \in \mathbb{R}^{2n}$ получать пару значений зашумлённой стохастической релизации
$(\widetilde{f}(x,\xi), \widetilde{f}(y,\xi))$ значений функции $f$, где

$$\widetilde{f}(x,\xi) = F(x,\xi) + \Xi(x,\xi), \quad |\Xi(x,\xi)| \leqslant \Delta, \ \forall x \in \mathbb{R}^n, \text{ для п.в. } \xi, \tag{22}$$

причём $\xi$ сэмплируется из распределения $P$ независимо от предыдущих сэмплирований.

Используя эти пары зашумлённых значений функции $f$, мы формируем следующую стохастическую аппроксимацию $\nabla f(x)$

$$\begin{aligned}\widetilde{\nabla}^m f^t(x) &= \frac{1}{m}\sum_{i=1}^m \frac{\widetilde{f}(x+te,\xi_i)-\widetilde{f}(x,\xi_i)}{t}e \\ &= \left(\langle g^m(x,\xi_{\mathbf{m}}), e\rangle + \frac{1}{m}\sum_{i=1}^m(\zeta(x,\xi_i,e)+\eta(x,\xi_i,e))\right)e,\end{aligned} \tag{23}$$

где $e \in RS_2(1)$, $\xi_i$, $i = 1, ..., m$ — независимые реализации случайной величины $\xi$, $m$ — *размер батча*, $t$ — некоторое небольшое положительное число, которое мы называем *параметром сглаживания*, $g^m(x, \xi_{\mathbf{m}}) := \frac{1}{m}\sum_{i=1}^m g(x,\xi_i)$ и

$$\begin{aligned}\zeta(x,\xi_i,e) &= \tfrac{F(x+te,\xi_i)-F(x,\xi_i)}{t} - \langle g(x,\xi_i),\, e\rangle, \\ \eta(x,\xi_i,e) &= \tfrac{\Xi(x+te,\xi_i)-\Xi(x,\xi_i)}{t}, \quad i = 1, ..., m.\end{aligned}$$

Из липшицевости градиента $F(\cdot,\xi)$, мы имеем $|\zeta(x,\xi,e)| \leqslant \frac{L(\xi)t}{2}$ для всех $x \in \mathbb{R}^n$ и $e \in S_2(1)$. Поэтому $\mathbb{E}_\xi(\zeta(x,\xi,e))^2 \leqslant \frac{L_2^2 t^2}{4}$ для всех $x \in \mathbb{R}^n$ и $e \in S_2(1)$. В то же время

из (22) мы имеем $|\eta(x,\xi,e)| \leqslant \frac{2\Delta}{t}$ для всех $x \in \mathbb{R}^n$, $e \in S_2(1)$ и почти всех $\xi$. Применяя Теорему 1 и Теорему 2 для $\Delta_\zeta = \frac{L_2^2 t^2}{4}$ и $\Delta_\eta = \frac{2\Delta}{t}$, мы восстанавливаем соответственно результаты Теоремы 2 и Теоремы 3 из [39].

Применяя Теорему 3 и Теорему 4 для $\Delta_\zeta = \frac{L_2^2 t^2}{4}$ и $\Delta_\eta = \frac{2\Delta}{t}$, мы получаем оценку на сложность (6) для безградиентной гладкой стохастической сильно выпуклой оптимизации, что до этого не было сделано в литературе.

## 3. Доказательство основного результата для ARDD метода

Доказательство Теоремы 1 состоит из двух больших шагов. Сначала, чтобы упростить выкладки, мы доказываем эту теорему, предполагая, что выполнены два неравенства, которые связывают зашумлённую стохастическую аппроксимацию градиента с настоящим градиентом и значениями функции. Этот результат сформулирован в виде Леммы 2. Затем в Лемме 3 мы показываем, что наша аппроксимация градиента (12) действительно удовлетворяет этим двум неравенствам.

**Лемма 2.** *Пусть точки $\{x_k, y_k, z_k\}$, $k \geqslant 0$ генерируется методом ARDD. Предположим, что существуют такие числа $\delta_1 > 0$,$\delta_2 > 0$, что для всех $k \geqslant 0$*

$$\mathbb{E}\left[\left\langle \widetilde{\nabla}^m f(x_{k+1}),\, z_k - x_* \right\rangle\right] \geqslant \frac{1}{n}\mathbb{E}\left[\langle \nabla f(x_{k+1}),\, z_k - x_* \rangle\right] - \delta_1 \mathbb{E}\left[\|z_k - x_*\|\right] \tag{24}$$

*и*

$$\mathbb{E}\left[\|\widetilde{\nabla}^m f(x_{k+1})\|_q^2\right] \leqslant 96\rho_n L_2 \left(\mathbb{E}[f(x_{k+1})] - \mathbb{E}[f(y_{k+1})]\right) + \delta_2, \tag{25}$$

*где математическое ожидание берётся относительно всей случайности и $x^*$ — решение (1). Тогда*

$$\mathbb{E}[f(y_N)] - f(x^*) \leqslant \frac{384\Theta_p n^2 \rho_n L_2}{N^2} + \frac{12n\sqrt{2\Theta_p}}{N^2}\delta_1 + \frac{N}{24\rho_n L_2}\delta_2 + \frac{N^2}{12\rho_n L_2}\delta_1^2, \tag{26}$$

*где $\Theta_p = V[z_0](x^*)$ определяется выбором прокс-структуры и математическое ожидание берётся относительно всей случайности.*

Этот результат доказан в разделе 3.1.

**Лемма 3.** *Пусть точки $\{x_k, y_k, z_k\}$, $k \geqslant 0$ генерируются методом ARDD. Тогда неравенства (24) и (25) выполняются с параметрами*

$$\delta_1 = \frac{\sqrt{\Delta_\zeta}}{2\sqrt{n}} + \frac{2\Delta_\eta}{\sqrt{n}} \tag{27}$$

$u$

$$\delta_2 = \frac{96\rho_n}{n} \cdot \frac{\sigma^2}{m} + 61\rho_n\Delta_\zeta + 976\rho_n\Delta_\eta^2. \tag{28}$$

Этот результат доказан в разделе 3.2.

**Доказательство Теоремы 1.** Соединяя результаты Леммы 2 и Леммы 3, мы получаем (13). Теорема доказана.

### 3.1. Доказательство Леммы 2

Следующая лемма оценивает прогресс шага 8 метода ARDD (и шага 5 метода RDD), который является шагом зеркального спуска.

**Лемма 4.** *Пусть* $z_+ = \operatorname{argmin}_{v\in\mathbb{R}^n}\left\{\alpha n\left\langle \widetilde{\nabla}^m f(x),\, v-z\right\rangle + V[z]\,(v)\right\}$. *Тогда для любой фиксированной точки* $u \in \mathbb{R}^n$,

$$\alpha n\mathbb{E}\left[\langle \widetilde{\nabla}^m f(x),\, z-u\rangle\right] \leqslant \tfrac{\alpha^2n^2}{2}\mathbb{E}\left[\|\widetilde{\nabla}^m f(x)\|_q^2\right] + \mathbb{E}\left[V[z](u)\right] - \mathbb{E}\left[V[z_+](u)\right], \tag{29}$$

*где математическое ожидание берётся относительно всей случайности.*

**Доказательство леммы.** Для всех $u \in \mathbb{R}^n$ имеем

$$\begin{aligned}
\alpha n\langle \widetilde{\nabla}^m f(x),\, z-u\rangle &= \alpha n\langle \widetilde{\nabla}^m f(x),\, z-z_+\rangle + \alpha n\langle \widetilde{\nabla}^m f(x),\, z_+-u\rangle \\
&\overset{①}{\leqslant} \alpha n\langle \widetilde{\nabla}^m f(x),\, z-z_+\rangle + \langle -\nabla V[z](z_+),\, z_+-u\rangle \\
&\overset{②}{=} \alpha n\langle \widetilde{\nabla}^m f(x),\, z-z_+\rangle \\
&\quad + V[z](u) - V[z_+](u) - V[z](z_+) \\
&\overset{③}{\leqslant} \left(\alpha n\langle \widetilde{\nabla}^m f(x),\, z-z_+\rangle - \tfrac{1}{2}\|z-z_+\|_p^2\right) \\
&\quad + V[z](u) - V[z_+](u) \\
&\overset{④}{\leqslant} \tfrac{\alpha^2n^2}{2}\|\widetilde{\nabla}^m f(x)\|_q^2 + V[z](u) - V[z_+](u),
\end{aligned} \tag{30}$$

где ① следует из определения $z_+$, откуда $\langle \nabla V[z](z_+) + \alpha n\widetilde{\nabla}^m f^t(x),\, u-z_+\rangle \geqslant 0$ для всех $u \in \mathbb{R}^n$; ② следует из «магического равенства» (см. Fact 5.3.3 («magic identity») в [37]) для дивергенции Брегмана; ③ следует из (7) и ④ вытекает из неравенства Фенхеля $\zeta\langle s, z\rangle - \frac{1}{2}\|z\|_p^2 \leq \frac{\zeta^2}{2}\|s\|_q^2$. Беря полное математическое ожидание, мы получаем (29). Лемма доказана.

Теперь докажем лемму, которая оценивает прогресс одной итерации всего алгоритма.

**Лемма 5.** *Пусть точки* $\{x_k, y_k, z_k, \alpha_k, \tau_k\}$, $k \geqslant 0$ *генерируются методом ARDD. Тогда в предположении Леммы 2 выполняется неравенство:*

$$\begin{aligned} &48n^2\rho_n L_2\alpha_{k+1}^2\mathbb{E}[f(y_{k+1})] - (48n^2\rho_n L_2\alpha_{k+1}^2 - \alpha_{k+1})\mathbb{E}\left[f(y_k)\right] \\ &-\mathbb{E}\left[V[z_k](x_*)\right] + \mathbb{E}[V[z_{k+1}](x_*)] - \alpha_{k+1}\delta_1 n\mathbb{E}\left[\|z_k - x_*\|_p\right] \\ &-\tfrac{\alpha_{k+1}^2 n^2}{2}\delta_2 \leqslant \alpha_{k+1}f(x_*), \end{aligned} \tag{31}$$

*где математическое ожидание берётся относительно всей случайности,* $x^*$ *— решение задачи* (1).

**Доказательство леммы.** Соединяя (24), (25) и (29), мы получим

$$\begin{aligned} \alpha_{k+1}\mathbb{E}\left[\langle \nabla f(x_{k+1}), z_k - x_*\rangle\right] &\leqslant 48\alpha^2 n^2\rho_n L_2\left(\mathbb{E}\left[f(x_{k+1})\right] - \mathbb{E}\left[f(y_{k+1})\right]\right) \\ &+\mathbb{E}\left[V_{z_k}(x_*)\right] - \mathbb{E}[V[z_{k+1}](x_*)] \\ &+\alpha_{k+1}\delta_1 n\mathbb{E}\left[\|z_k - x_*\|_p\right] + \tfrac{\alpha_{k+1}^2 n^2}{2}\delta_2. \end{aligned} \tag{32}$$

Далее

$$\begin{aligned} \alpha_{k+1}\left(\mathbb{E}\left[f(x_{k+1})\right] - f(x_*)\right) &\leqslant \alpha_{k+1}\mathbb{E}\left[\langle \nabla f(x_{k+1}),\, x_{k+1} - x_*\rangle\right] \\ &= \alpha_{k+1}\mathbb{E}\left[\langle \nabla f(x_{k+1}),\, x_{k+1} - z_k\rangle\right] \\ &+\alpha_{k+1}\mathbb{E}\left[\langle \nabla f(x_{k+1}),\, z_k - x_*\rangle\right] \\ &\overset{①}{=} \tfrac{(1-\tau_k)\alpha_{k+1}}{\tau_k}\mathbb{E}\left[\langle \nabla f(x_{k+1}),\, y_k - x_{k+1}\rangle\right] \\ &+\alpha_{k+1}\mathbb{E}\left[\langle \nabla f(x_{k+1}),\, z_k - x_*\rangle\right] \\ &\overset{②}{\leqslant} \tfrac{(1-\tau_k)\alpha_{k+1}}{\tau_k}\left(\mathbb{E}\left[f(y_k)\right] - \mathbb{E}\left[f(x_{k+1})\right]\right) \\ &+\alpha_{k+1}\mathbb{E}\left[\langle \nabla f(x_{k+1}),\, z_k - x_*\rangle\right] \\ &\overset{(32)}{\leqslant} \tfrac{(1-\tau_k)\alpha_{k+1}}{\tau_k}\left(\mathbb{E}\left[f(y_k)\right] - \mathbb{E}\left[f(x_{k+1})\right]\right) \\ &+48\alpha^2 n^2\rho_n L_2\left(\mathbb{E}\left[f(x_{k+1})\right] - \mathbb{E}\left[f(y_{k+1})\right]\right) \\ &+\mathbb{E}\left[V_{z_k}(x_*)\right] - \mathbb{E}[V[z_{k+1}](x_*)] \\ &+\alpha_{k+1}\delta_1 n\mathbb{E}\left[\|z_k - x_*\|_p\right] + \tfrac{\alpha_{k+1}^2 n^2}{2}\delta_2 \\ &\overset{③}{=} (48\alpha_{k+1}^2 n^2\rho_n L_2 - \alpha_{k+1})\mathbb{E}\left[f(y_k)\right] \\ &-48\alpha_{k+1}^2 n^2\rho_n L_2\mathbb{E}[f(y_{k+1})] \\ &+\alpha_{k+1}\mathbb{E}[f(x_{k+1})] + \mathbb{E}\left[V_{z_k}(x_*)\right] - \mathbb{E}[V[z_{k+1}](x_*)] \\ &+\alpha_{k+1}\delta_1 n\mathbb{E}\left[\|z_k - x_*\|_p\right] + \tfrac{\alpha_{k+1}^2 n^2}{2}\delta_2. \end{aligned}$$

Здесь ① вытекает из $x_{k+1} := \tau_k z_k + (1-\tau_k)y_k \;\Leftrightarrow\; \tau_k(x_{k+1} - z_k) = (1-\tau_k)(y_k - x_{k+1})$, ② следует из выпуклости функции $f$ и неравенства $1 - \tau_k \geqslant 0$, ③ верно в силу $\tau_k = \frac{1}{48\alpha_{k+1}n^2\rho_n L_2}$. Сокращая подобные слагаемые, получаем (31). Лемма доказана.

Теперь всё готово к тому, чтобы завершить доказательство Леммы 2.

**Доказательство Леммы 2.** Заметим, что $48n^2\rho_n L_2\alpha_{k+1}^2 - \alpha_{k+1} + \frac{1}{192n^2\rho_n L_2} = 48n^2\rho_n L_2\alpha_k^2$. Действительно,

$$\begin{aligned} 48n^2\rho_n L_2\alpha_{k+1}^2 - \alpha_{k+1} + \tfrac{1}{192n^2\rho_n L_2} &= \tfrac{(k+2)^2}{192n^2\rho_n L_2} - \tfrac{k+2}{96n^2\rho_n L_2} + \tfrac{1}{192n^2\rho_n L_2} \\ &= \tfrac{k^2+4k+4-2k-4+1}{192n^2\rho_n L_2} \\ &= \tfrac{(k+1)^2}{192n^2\rho_n L_2} \\ &= 48n^2\rho_n L_2\alpha_k^2. \end{aligned}$$

Сложим неравенства (31) (получится телескопическая сумма) для $k = 0, 1, 2, \ldots, l-1$, где $l \leqslant N$ мы получим[6]

$$\begin{aligned} 48n^2\rho_n L_2\alpha_l^2\mathbb{E}[f(y_l)] + \sum_{k=1}^{l-1}\tfrac{1}{192n^2\rho_n L_2}\mathbb{E}[f(y_k)] - V[z_0](x_*) + \mathbb{E}[V[z_l](x_*)] \\ -\zeta_1\sum_{k=0}^{l-1}\alpha_{k+1}\mathbb{E}[\|u - z_k\|_p] - \zeta_2\sum_{k=0}^{l-1}\alpha_{k+1}^2 \leqslant \sum_{k=0}^{l-1}\alpha_{k+1}f(u), \end{aligned} \tag{33}$$

где мы ввели обозначения:

$$\zeta_1 := \delta_1 n, \quad \zeta_2 := \frac{n^2}{2}\delta_2. \tag{34}$$

Кроме того, обозначим $\Theta := V[z_0](x^*)$, $R_k := \mathbb{E}[\|x^* - z_k\|_p]$. Также отметим, что из неравенства (7) мы имеем: $\zeta_1\alpha_1 R_0 \leq \frac{\sqrt{2\Theta}\zeta_1}{48n^2\rho_n L_2}$. Для упрощения выкладок будем использовать обозначение $B_l := \zeta_2\sum_{k=0}^{l-1}\alpha_{k+1}^2 + \Theta + \frac{\sqrt{2\Theta}\zeta_1}{48n^2\rho_n L_2}$. В силу $\sum_{k=0}^{l-1}\alpha_{k+1} = \frac{l(l+3)}{192n^2\rho_n L_2}$ и того, что для всех $i = 1, \ldots, N$, $f(y_i) \leqslant f(x^*)$, получаем из (33) следующее неравенство:

$$\begin{aligned} \tfrac{(l+1)^2}{192n^2\rho_n L_2}\mathbb{E}[f(y_l)] &\leqslant f(x^*)\left(\tfrac{(l+3)l}{192n^2\rho_n L_2} - \tfrac{l-1}{192n^2\rho_n L_2}\right) + B_l \\ &\quad -\mathbb{E}[V[z_l](x^*)] + \zeta_1\sum_{k=1}^{l-1}\alpha_{k+1}R_k, \\ 0 \leqslant \tfrac{(l+1)^2}{192n^2\rho_n L_2}\left(\mathbb{E}[f(y_l)] - f(x^*)\right) &\leqslant B_l - \mathbb{E}[V[z_l](x^*)] + \zeta_1\sum_{k=1}^{l-1}\alpha_{k+1}R_k, \end{aligned} \tag{35}$$

откуда следует, что

$$\mathbb{E}[V[z_l](x^*)] \leqslant B_l + \zeta_1\sum_{k=1}^{l-1}\alpha_{k+1}R_k. \tag{36}$$

Более того,

$$\begin{aligned} \tfrac{1}{2}\left(\mathbb{E}[\|z_l - x^*\|_p]\right)^2 &\leqslant \tfrac{1}{2}\mathbb{E}[\|z_l - x^*\|_p^2] \leqslant \mathbb{E}[V[z_l](x^*)] \\ &\overset{(36)}{\leqslant} B_l + \zeta_1\sum_{k=1}^{l-1}\alpha_{k+1}R_k, \end{aligned} \tag{37}$$

откуда

$$R_l \leqslant \sqrt{2}\cdot\sqrt{B_l + \zeta_1\sum_{k=1}^{l-1}\alpha_{k+1}R_k}. \tag{38}$$

[6] Отметим, что $\alpha_1 = \frac{2}{96n^2\rho_n L_2} = \frac{1}{48n^2\rho_n L_2}$ и поэтому $48n^2\rho_n L_2\alpha_1^2 - \alpha_1 = 0$.

Применяя Лемму 12 (см. раздел 6) для $a_0 = \zeta_2\alpha_1^2 + \Theta + \frac{\sqrt{2\Theta}\zeta_1}{48n^2\rho_n L_2}, a_k = \zeta_2\alpha_{k+1}^2, b = \zeta_1$ для $k = 1, \ldots, N-1$, мы получаем

$$B_l + \zeta_1 \sum_{k=1}^{l-1} \alpha_{k+1} R_k \leqslant \left(\sqrt{B_l} + \sqrt{2}\zeta_1 \cdot \tfrac{l^2}{96n^2\rho_n L_2}\right)^2, \; l = 1, \ldots, N \tag{39}$$

Так как $V[z](x^*) \geqslant 0$, из неравенства (35) для $l = N$ и определения $B_l$, мы имеем

$$\begin{aligned} \tfrac{(N+1)^2}{192n^2\rho_n L_2}\left(\mathbb{E}[f(y_N)] - f(x^*)\right) &\leqslant \left(\sqrt{B_N} + \sqrt{2}\zeta_1 \cdot \tfrac{N^2}{96n^2\rho_n L_2}\right)^2 \\ &\overset{①}{\leqslant} 2B_N + 4\zeta_1^2 \cdot \tfrac{N^4}{(96n^2\rho_n L_2)^2} \\ &= 2\zeta_2 \sum_{k=0}^{l-1} \alpha_{k+1}^2 + 2\Theta + \tfrac{\sqrt{2\Theta}\zeta_1}{24n^2\rho_n L_2} + 4\zeta_1^2 \cdot \tfrac{N^4}{(96n^2\rho_n L_2)^2} \\ &\overset{②}{\leqslant} 2\Theta + \tfrac{\sqrt{2\Theta}\zeta_1}{24n^2\rho_n L_2} + \tfrac{2\zeta_2(N+1)^3}{(96n^2\rho_n L_2)^2} + 4\zeta_1^2 \cdot \tfrac{N^4}{(96n^2\rho_n L_2)^2} \end{aligned} \tag{40}$$

где ① выполнено в силу того, что $\forall a, b \in \mathbb{R} \quad (a+b)^2 \leqslant 2a^2 + 2b^2$ и ② следует из $\sum\limits_{k=0}^{N-1} \alpha_{k+1}^2 = \frac{1}{(96n^2\rho_n L_2)^2} \sum\limits_{k=2}^{N+1} k^2 \leqslant \frac{1}{(96n^2\rho_n L_2)^2} \cdot \frac{(N+1)(N+2)(2N+3)}{6} \leqslant \frac{1}{(96n^2\rho_n L_2)^2} \cdot \frac{(N+1)2(N+1)3(N+1)}{6} = \frac{(N+1)^3}{(96n^2\rho_n L_2)^2}$. Деля неравенство (40) на $\frac{(N+1)^2}{192n^2\rho_n L_2}$ и подставляя $\zeta_1, \zeta_2$ из (34), мы получаем

$$\begin{aligned} \mathbb{E}[f(y_N)] - f(x^*) &\leqslant \tfrac{384\Theta n^2\rho_n L_2}{(N+1)^2} + \tfrac{12\sqrt{2\Theta}}{(N+1)^2}\zeta_1 + \tfrac{(N+1)\zeta_2}{24n^2\rho_n L_2} + \tfrac{N^4\zeta_1^2}{12n^2\rho_n L_2(N+1)^2} \\ &\leqslant \tfrac{384\Theta n^2\rho_n L_2}{N^2} + \tfrac{12n\sqrt{2\Theta}}{N^2}\delta_1 + \tfrac{N}{24\rho_n L_2}\delta_2 + \tfrac{N^2}{12\rho_n L_2}\delta_1^2. \end{aligned}$$

Лемма доказана.

### 3.2. Доказательство Леммы 3

Начнём мы со следующего технического результата, который связывает зашёмлённую аппроксимацию стохастического градиента (12) с самим стохастическим градиентом и $\nabla f$.

**Лемма 6.** *Для всех $x, s \in \mathbb{R}^n$ имеем*

$$\mathbb{E}_e\|\widetilde{\nabla}^m f(x)\|_q^2 \leqslant \tfrac{12\rho_n}{n}\|g^m(x, \xi_{\mathbf{m}})\|_2^2 + \tfrac{\rho_n}{m}\sum_{i=1}^m \zeta(x, \xi_i)^2 + 16\rho_n\Delta_\eta^2, \tag{41}$$

$$\mathbb{E}_e\|\widetilde{\nabla}^m f(x)\|_2^2 \geqslant \frac{1}{2n}\|g^m(x, \xi_{\mathbf{m}})\|_2^2 - \frac{1}{2m}\sum_{i=1}^m \zeta(x, \xi_i)^2 - 8\Delta_\eta^2, \tag{42}$$

$$\mathbb{E}_e\langle\widetilde{\nabla}^m f(x), s\rangle \geqslant \tfrac{1}{n}\langle g^m(x, \xi_{\mathbf{m}}), s\rangle - \tfrac{\|s\|_p}{2m\sqrt{n}}\sum_{i=1}^m |\zeta(x, \xi_i)| - \tfrac{2\Delta_\eta\|s\|_p}{\sqrt{n}}, \tag{43}$$

$$\begin{aligned} \mathbb{E}_e\|\langle\nabla f(x), e\rangle e - \widetilde{\nabla}^m f(x)\|_2^2 &\leqslant \tfrac{2}{n}\|\nabla f(x) - g^m(x, \xi_{\mathbf{m}})\|_2^2 \\ &\quad + \tfrac{1}{m}\sum_{i=1}^m \zeta(x, \xi_i)^2 + 16\Delta_\eta^2, \end{aligned} \tag{44}$$

*где $g^m(x, \xi_{\mathbf{m}}) := \frac{1}{m}\sum\limits_{i=1}^m g(x, \xi_i)$, а $\zeta(x, \xi_i)$ и $\Delta_\eta$ определены в (3).*

**Доказательство леммы.** Для начала перепишем $\widetilde{\nabla}^m f(x)$ в следующем виде

$$\widetilde{\nabla}^m f(x) = \left( \langle g^m(x, \xi_{\mathbf{m}}), e \rangle + \frac{1}{m} \sum_{i=1}^{m} \theta(x, \xi_i, e) \right) e,$$

где

$$\theta(x, \xi_i, e) = \zeta(x, \xi_i) + \eta(x, \xi_i, e), \quad i = 1, ..., m.$$

Из (3) мы имеем

$$|\theta(x, \xi_i, e)| \leq |\zeta(x, \xi_i)| + \Delta_\eta. \tag{45}$$

**Доказательство** (41).

$$\begin{aligned} \mathbb{E}_e \|\widetilde{\nabla}^m f(x)\|_q^2 &= \mathbb{E}_e \left\| \left( \langle g^m(x, \xi_{\mathbf{m}}), e \rangle + \tfrac{1}{m} \sum_{i=1}^{m} \theta(x, \xi_i, e) \right) e \right\|_q^2 \\ &\overset{①}{\leqslant} 2\mathbb{E}_e \|\langle g^m(x, \xi_{\mathbf{m}}), e \rangle e\|_q^2 + 2\mathbb{E}_e \left\| \tfrac{1}{m} \sum_{i=1}^{m} \theta(x, \xi_i, e) e \right\|_q^2 \\ &\overset{②}{\leqslant} \tfrac{12\rho_n}{n} \|g^m(x, \xi_{\mathbf{m}})\|_2^2 + \tfrac{2\rho_n}{m} \sum_{i=1}^{m} \left( |\zeta(x, \xi_i)| + \Delta_\eta \right)^2 \\ &\leqslant \tfrac{12\rho_n}{n} \|g^m(x, \xi_{\mathbf{m}})\|_2^2 + \tfrac{\rho_n}{m} \sum_{i=1}^{m} \zeta(x, \xi_i)^2 + 16\rho_n \Delta_\eta^2, \end{aligned} \tag{46}$$

где ① выполнено в силу $\|x+y\|_q^2 \leqslant 2\|x\|_q^2 + 2\|y\|_q^2, \forall x, y \in \mathbb{R}^n$; ② следует из неравенств (10),(11), (45) и простого факта, что для любых чисел $a_1, a_2, \ldots, a_m > 0$ выполняется неравенство $\left( \sum_{i=1}^{m} a_i \right)^2 \leqslant m \sum_{i=1}^{m} a_i^2$.

**Доказательство** (42).

$$\begin{aligned} \mathbb{E}_e \|\widetilde{\nabla}^m f(x)\|_2^2 &= \mathbb{E}_e \left\| \left( \langle g^m(x, \xi_{\mathbf{m}}), e \rangle + \tfrac{1}{m} \sum_{i=1}^{m} \theta(x, \xi_i, e) \right) e \right\|_2^2 \\ &\overset{①}{\geqslant} \tfrac{1}{2} \mathbb{E}_e \|\langle g^m(x, \xi_{\mathbf{m}}), e \rangle e\|_2^2 - \tfrac{1}{m} \sum_{i=1}^{m} \left( |\zeta(x, \xi_i)| + \Delta_\eta \right)^2 \\ &\overset{②}{\geqslant} \tfrac{1}{2n} \|g^m(x, \xi_{\mathbf{m}})\|_2^2 - \tfrac{1}{2m} \sum_{i=1}^{m} \zeta(x, \xi_i)^2 - 8\Delta_\eta^2, \end{aligned} \tag{47}$$

где ① вытекает из (45) и неравенства $\|x+y\|_2^2 \geqslant \frac{1}{2}\|x\|_2^2 - \|y\|_2^2, \forall x, y \in \mathbb{R}^n$; ② следует из $e \in S_2(1)$ и Леммы B.10 из [26], которая утверждает, что для любого $s \in \mathbb{R}^n$, $\mathbb{E}\langle s, e \rangle^2 = \frac{1}{n}\|s\|_2^2$.

**Доказательство** (43).

$$\begin{aligned} \mathbb{E}_e \langle \widetilde{\nabla}^m f(x), s \rangle &= \mathbb{E}_e \langle \langle g^m(x, \xi_{\mathbf{m}}), e \rangle e, s \rangle + \mathbb{E}_e \tfrac{1}{m} \sum_{i=1}^{m} \theta(x, \xi_i, e) \langle e, s \rangle \\ &\overset{①}{\geqslant} \tfrac{1}{n} \langle g^m(x, \xi_{\mathbf{m}}), s \rangle - \tfrac{1}{m} \sum_{i=1}^{m} \left( |\zeta(x, \xi_i)| + \Delta_\eta \right) \mathbb{E}_e |\langle e, s \rangle| \\ &\overset{②}{\geqslant} \tfrac{1}{n} \langle g^m(x, \xi_{\mathbf{m}}), s \rangle - \tfrac{\|s\|_p}{2m\sqrt{n}} \sum_{i=1}^{m} |\zeta(x, \xi_i)| - \tfrac{2\Delta_\eta \|s\|_p}{\sqrt{n}} \end{aligned} \tag{48}$$

ult ① выполнено в силу $\mathbb{E}_e[n\langle g,e\rangle e]=g$, $\forall g\in\mathbb{R}^n$ и (45); ② следует из Леммы B.10 из [26], неравенства $\mathbb{E}|\langle s,e\rangle|\leq\sqrt{\mathbb{E}\langle s,e\rangle^2}$ и простого факта о том, что $\|x\|_2\leqslant\|x\|_p$ для $p\leqslant 2$.

**Доказательство** (44).

$$\begin{aligned}
&\mathbb{E}_e\|\langle\nabla f(x),e\rangle e-\widetilde{\nabla}^m f(x)\|_2^2\\
&=\mathbb{E}_e\left\|\langle\nabla f(x),e\rangle e-\langle g^m(x,\xi_{\mathbf{m}}),e\rangle e-\tfrac{1}{m}\sum_{i=1}^m\theta(x,\xi_i,e)e\right\|_2^2\\
&\overset{①}{\leqslant}2\mathbb{E}_e\|\langle\nabla f(x)-g^m(x,\xi_{\mathbf{m}}),e\rangle e\|_2^2+2\mathbb{E}_e\left\|\tfrac{1}{m}\sum_{i=1}^m\theta(x,\xi_i,e)e\right\|_2^2\\
&\overset{②}{\leqslant}\tfrac{2}{n}\|\nabla f(x)-g^m(x,\xi_{\mathbf{m}})\|_2^2+\tfrac{1}{m}\sum_{i=1}^m\zeta(x,\xi_i)^2+16\Delta_\eta^2,
\end{aligned}\tag{49}$$

где ① выполнено в силу $\|x+y\|_2^2\leqslant 2\|x\|_2^2+2\|y\|_2^2,\forall x,y\in\mathbb{R}^n$; ② следует из $e\in S_2(1)$, Леммы B.10 из [26] и (45). Лемма доказана.

Теперь докажем следующую лемму, которая оценивает прогресс шага 7 метода ARDD, который является шагом градиентного спуска.

**Лемма 7.** *Пусть* $y=x-\frac{1}{2L_2}\widetilde{\nabla}^m f(x)$. *Тогда,*

$$\begin{aligned}
\|g^m(x,\xi_{\mathbf{m}})\|_2^2 &\leqslant 8nL_2(f(x)-\mathbb{E}_e f(y))+8\|\nabla f(x)-g^m(x,\xi_{\mathbf{m}})\|_2^2\\
&\quad+\tfrac{5n}{m}\sum_{i=1}^m\zeta(x,\xi_i)^2+80n\Delta_\eta^2,
\end{aligned}\tag{50}$$

*где выражение* $g^m(x,\xi_{\mathbf{m}})$ *определено в Лемме 6, а* $\zeta(x,\xi_i)$ *и* $\Delta_\eta$ *определены в* (3).

**Доказательство леммы.** Так как вектор $\widetilde{\nabla}^m f(x)$ коллинеарен вектору $e$, то существует число $\gamma\in\mathbb{R}$, такое что $y-x=\gamma e$. Тогда с учётом $\|e\|_2=1$ получаем

$$\langle\nabla f(x),y-x\rangle=\langle\nabla f(x),e\rangle\gamma=\langle\nabla f(x),e\rangle\langle e,y-x\rangle=\langle\langle\nabla f(x),e\rangle e,\,y-x\rangle.$$

Отсюда и из $L_2$-гладкости функции $f$ мы получаем

$$\begin{aligned}
f(y) &\leqslant f(x)+\langle\langle\nabla f(x),e\rangle e,\,y-x\rangle+\tfrac{L_2}{2}||y-x||_2^2\\
&\leqslant f(x)+\langle\widetilde{\nabla}^m f(x),\,y-x\rangle+L_2||y-x||_2^2\\
&+\langle\langle\nabla f(x),e\rangle e-\widetilde{\nabla}^m f(x),\,y-x\rangle-\tfrac{L_2}{2}||y-x||_2^2\\
&\overset{①}{\leqslant} f(x)+\langle\widetilde{\nabla}^m f(x),\,y-x\rangle+L_2||y-x||_2^2+\tfrac{1}{2L_2}\|\langle\nabla f(x),e\rangle e-\widetilde{\nabla}^m f(x)\|_2^2,
\end{aligned}$$

где ① следует из неравенства Фенхеля: $\langle s,z\rangle-\frac{\zeta}{2}\|z\|_2^2\leq\frac{1}{2\zeta}\|s\|_2^2$. Подставляя $y=x-\frac{1}{2L_2}\widetilde{\nabla}^m f(x)$, мы получим

$$\tfrac{1}{4L_2}\|\widetilde{\nabla}^m f(x)\|_2^2\leqslant f(x)-f(y)+\tfrac{1}{2L_2}\|\langle\nabla f(x),e\rangle e-\widetilde{\nabla}^m f(x)\|_2^2$$

Беря математическое ожидание по $e$ и применяя (42), (44), получим

$$\begin{aligned}\tfrac{1}{4L_2}\left(\tfrac{1}{2n}\|g^m(x,\xi_{\mathbf{m}})\|_2^2-\tfrac{1}{2m}\sum_{i=1}^m\zeta(x,\xi_i)^2-8\Delta_\eta^2\right)&\leqslant\tfrac{1}{4L_2}\mathbb{E}_e\|\widetilde{\nabla}^m f(x)\|_2^2\\&\leqslant f(x)-\mathbb{E}_e f(y)+\tfrac{1}{2L_2}\mathbb{E}_e\|\langle\nabla f(x),e\rangle e-\widetilde{\nabla}^m f(x)\|_2^2\\&\leqslant f(x)-\mathbb{E}_e f(y)+\tfrac{1}{2L_2}\left(\tfrac{2}{n}\|\nabla f(x)-g^m(x,\xi_{\mathbf{m}})\|_2^2+\tfrac{t^2}{m}\sum_{i=1}^m\zeta(x,\xi_i)^2+16\Delta_\eta^2\right),\end{aligned}$$

Приводя подобные слагаемые, получим утверждение леммы. Лемма доказана.

Теперь всё готово для доказательства Леммы 3.

**Доказательство Леммы 3.** Беря математическое ожидание относительно всей случайности[7] от неравенства (43) и используя неравентсво

$$\mathbb{E}[|\zeta(x,\xi_i)|]\leqslant\sqrt{\mathbb{E}[|\zeta(x,\xi_i)|^2]}\overset{(3)}{\leqslant}\sqrt{\Delta_\zeta},$$

мы получим неравенство (24) с $\delta_1=\frac{\sqrt{\Delta_\zeta}}{2\sqrt{n}}+\frac{2\Delta_\eta}{\sqrt{n}}$. Соединяя вместе неравенства (41) и (50), беря полное математическое ожидание и используя неравенство $\mathbb{E}[\|\nabla f(x)-g^m(x,\xi)\|_2^2]\leqslant\frac{\sigma^2}{m}$, которое следует из (2), мы получаем (25) с $\delta_2=\frac{96\rho_n}{n}\cdot\frac{\sigma^2}{m}+61\rho_n\Delta_\zeta+976\rho_n\Delta_\eta^2$. Лемма доказана.

## 4. Доказательство основного результата для RDD метода

Как и в предыдущем разделе мы разделили доказательство Теоремы 2 на два больших шага. Сначала, чтобы упростить выкладки, мы доказываем эту теорему, предполагая истинность двух неравентсв, которые связывают нашу зашумлённую стохастическую аппроксимацию градиента (12) с самим градиентом и значениями функции. Затем мы показываем, что наша аппроксимация градиента (12) действительно удовлетворяет этим двум неравенствам.

**Лемма 8.** *Пусть точки $\{x_k,y_k,z_k\}$, $k\geqslant 0$ генерируются RDD методом. Предположим, что существуют числа $\delta_1>0$,$\delta_2>0$, такие что для всех $k\geqslant 0$ выполняются неравенства*

$$\mathbb{E}\left[\left\langle\widetilde{\nabla}^m f(x_k),x_k-x_*\right\rangle\right]\geqslant\frac{1}{n}\mathbb{E}\left[\langle\nabla f(x_k),\,x_k-x_*\rangle\right]-\delta_1\mathbb{E}\left[\|x_k-x_*\|_p\right]\qquad(51)$$

[7] Отметим, что мы используем $s=z_k-x_*$, который не зависит от $\xi_1,\xi_2,\ldots,\xi_m$ с $(k+1)$-й итерации и не зависит от $e_{k+1}$. Поэтому мы можем использовать «башенное свойство» («tower property») математического ожидания и брать сначала условное математическое ожидание по $\xi_1,\ldots,\xi_m$ и после этого брать полное математическое ожидание.

$$\mathbb{E}\left[\|\widetilde{\nabla}^m f(x_k)\|_q^2\right] \leqslant \frac{48\rho_n L_2}{n}\left(\mathbb{E}\left[f(x_k)\right] - f(x_*)\right) + \delta_2, \tag{52}$$

*где математическое ожидание берётся относительно всей случайности и* $x_*$ *— это решение задачи* (1)*. Тогда*

$$\mathbb{E}[f(\bar{x}_N)] - f(x_*) \leqslant \frac{384n\rho_n L_2\Theta_p}{N} + \frac{n}{12\rho_n L_2}\delta_2 + \frac{8n\sqrt{2\Theta_p}}{N}\delta_1 + \frac{nN}{3L_2\rho_n}\delta_1^2, \tag{53}$$

*где* $\Theta_p = V[z_0](x^*)$ *задаётся выбором прокс-структуры и математическое ожидание берётся относительно всей случайности.*

Этот результат доказывается в разделе 4.1.

**Лемма 9.** *Пусть точки* $\{x_k, y_k, z_k\}$, $k \geqslant 0$ *генерируются RDD методом. Тогда неравенства* (51) *и* (52) *выполнены с*

$$\delta_1 = \frac{\sqrt{\Delta_\zeta}}{2\sqrt{n}} + \frac{2\Delta_\eta}{\sqrt{n}} \tag{54}$$

*и*

$$\delta_2 = \frac{24\rho_n}{n}\cdot\frac{\sigma^2}{m} + \rho_n\Delta_\zeta + 16\rho_n\Delta_\eta^2. \tag{55}$$

Этот результат доказывается в разделе 4.2.

**Доказательство Теоремы 2.** Соединяя результаты Леммы 8 и Леммы 9, мы получаем (14). Теорема доказана.

### 4.1. Доказательство Леммы 8

Используя (29), (51) и (52), мы получаем

$$\begin{aligned}\alpha\mathbb{E}\left[\langle\nabla f(x_k),\, x_k - x_*\rangle\right] &\leqslant 24\alpha^2 n\rho_n L_2\left(\mathbb{E}\left[f(x_k)\right] - f(x_*)\right)\\ &+\alpha\delta_1 n\mathbb{E}\left[\|x_k - x_*\|_p\right] + \tfrac{\alpha^2n^2}{2}\delta_2\\ &+\mathbb{E}\left[V[x_k](x_*)\right] - \mathbb{E}\left[V[x_{k+1}](x_*)\right],\end{aligned}$$

откуда и из выпуклости $f$ мы имеем

$$\begin{aligned}\underbrace{(\alpha - 24\alpha^2 n\rho_n L_2)}_{\frac{\alpha}{4}}\left(\mathbb{E}[f(x_k)] - f(x_*)\right) &\leqslant \alpha\delta_1 n\mathbb{E}\left[\|x_k - x_*\|_p\right] + \tfrac{\alpha^2n^2}{2}\delta_2\\ &+\mathbb{E}[V[x_k](x_*)] - \mathbb{E}[V[x_{k+1}](x_*)],\end{aligned} \tag{56}$$

так как $\alpha = \frac{1}{48n\rho_n L_2}$. Суммируя неравенства (56) для $k = 0, \ldots, l-1$, где $l \leqslant N$, получаем

$$\begin{aligned} 0 \leqslant \tfrac{N\alpha}{4}\left(\mathbb{E}[f(\bar{x}_l)] - f(x_*)\right) &\leqslant \tfrac{\alpha^2 n^2 l}{2}\delta_2 + \alpha\delta_1 n \sum_{k=0}^{l-1}\mathbb{E}[\|x_k - x_*\|_p] \\ &\quad + \underbrace{V[x_0](x_*)}_{\Theta_p} - \mathbb{E}[V[x_l](x^*)], \end{aligned} \tag{57}$$

где $\bar{x}_l \overset{\text{def}}{=} \frac{1}{l}\sum\limits_{k=0}^{l-1} x_k$. Из предыдущего неравенства имеем

$$\begin{aligned} \tfrac{1}{2}\left(\mathbb{E}[\|x_l - x_*\|_p]\right)^2 &\leqslant \tfrac{1}{2}\mathbb{E}[\|x_l - x_*\|_p^2] \leqslant \mathbb{E}[V[x_l](x_*)] \\ &\leqslant \Theta_p + l \cdot \tfrac{\alpha^2 n^2}{2}\delta_2 + \alpha\delta_1 n \sum_{k=0}^{l-1}\mathbb{E}[\|x_k - x_*\|_p], \end{aligned} \tag{58}$$

откуда $\forall l \leqslant N$ мы получаем

$$\mathbb{E}[\|x_k - x_*\|_p] \leqslant \sqrt{2}\sqrt{\Theta_p + l \cdot \frac{\alpha^2 n^2}{2}\delta_2 + \alpha\delta_1 n \sum_{k=0}^{l-1}\mathbb{E}[\|x_k - x_*\|_p]}. \tag{59}$$

Обозначим $R_k = \mathbb{E}[\|x^* - x_k\|_p]$ для $k = 0, \ldots, N$. Применяя Лемму 13 (см. раздел 6) для $a_0 = \Theta_p + \alpha\delta_1 n\mathbb{E}[\|x_0 - x_*\|_p] \leqslant \Theta_p + \alpha n\sqrt{2\Theta_p}\delta_1, a_k = \frac{\alpha^2 n^2}{2}\delta_2, b = n\delta_1$ для $k = 1, \ldots, N-1$, мы получаем для $l = N$

$$\begin{aligned} \tfrac{N\alpha}{4}\left(\mathbb{E}[f(\bar{x}_N)] - f(x_*)\right) &\leqslant \left(\sqrt{\Theta_p + N \cdot \tfrac{\alpha^2 n^2}{2}\delta_2 + \alpha n\sqrt{2\Theta_p}\delta_1} + \sqrt{2}n\delta_1\alpha N\right)^2 \\ &\overset{①}{\leqslant} 2\Theta_p + N\alpha^2 n^2\delta_2 + 2\alpha n\sqrt{2\Theta_p}\delta_1 + 4n^2\delta_1^2\alpha^2 N^2, \end{aligned}$$

откуда

$$\mathbb{E}[f(\bar{x}_N)] - f(x_*) \leqslant \tfrac{384 n\rho_n L_2 \Theta_p}{N} + \tfrac{n}{12\rho_n L_2}\delta_2 + \tfrac{8n\sqrt{2\Theta_p}}{N}\delta_1 + \tfrac{nN}{3L_2\rho_n}\delta_1^2,$$

так как $\alpha = \frac{1}{48n\rho_n L_2}$. Лемма доказана.

### 4.2. Доказательство Леммы 9

Беря математическое ожидание относительно всей случайности от неравенства (43), мы получаем[8] неравенство (51) с $\delta_1 = \frac{\sqrt{\Delta_\zeta}}{2\sqrt{n}} + \frac{2\Delta_\eta}{\sqrt{n}}$, так как $\mathbb{E}[|\zeta(x, \xi_i)|] \leqslant \sqrt{\mathbb{E}[|\zeta(x, \xi_i)|^2]} \overset{(3)}{\leqslant}$

[8] Отметим, что мы используем вектор $s = x_k - x_*$, который не зависит от $\xi_1, \xi_2, \ldots, \xi_m$ с $(k+1)$-й итерации и не зависит от $e_{k+1}$. Поэтому мы можем использовать «башенное свойство» (tower property) математического ожидания и взять сначала условное математическое ожидание по $\xi_1, \ldots, \xi_m$ и после этого взять полное математическое ожидание.

$\sqrt{\Delta_\zeta}$. Используя неравенства (41) и

$$\begin{aligned} \|g^m(x,\xi_m)\|_2^2 &\leqslant 2\|\nabla f(x)\|_2^2 + 2\|\nabla f(x) - g^m(x,\xi_m)\|_2^2 \\ &\leqslant 4L_2\left(\mathbb{E}[f(x)] - f(x_*)\right) + 2\|\nabla f(x) - g^m(x,\xi_m)\|_2^2, \\ \mathbb{E}[\|\nabla f(x) - g^m(x,\xi_m)\|_2^2] &\leqslant \frac{\sigma^2}{m} \end{aligned}$$

и беря полное математическое ожидание, мы получаем (52) с $\delta_2 = \frac{24\rho_n}{n}\cdot\frac{\sigma^2}{m} + \rho_n\Delta_\zeta + 16\rho_n\Delta_\eta^2$. Лемма доказана.

## 5. Доказательства для сильно выпуклых задач

### 5.1. Ускоренный метод

**Лемма 10.** *Пусть мы запускаем метод ARDD (Algorithm 1) из случайной точки $x_0$, такой что $\mathbb{E}_{x_0}\|x^* - x_0\|_p^2 \leqslant R_p^2$, и используем функцию $R_p^2 d\left(\frac{x-x_0}{R_p}\right)$ в качестве прокс-функции, причём запускаем метод ARDD на $N_0$ итераций. Тогда*

$$\mathbb{E}[f(y_{N_0})] - f^* \leqslant \frac{aL_2R_p^2\Omega_p}{N_0^2} + \frac{b\sigma^2 N_0}{mL_2} + \Delta,$$

*где $a = 384n^2\rho_n$, $b = \frac{4}{n}$, $\Delta = \frac{61N_0}{24L_2}\Delta_\zeta + \frac{122N_0}{3L_2}\Delta_\eta^2 + \frac{12\sqrt{2nR_p^2\Omega_p}}{N_0^2}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right) + \frac{N_0^2}{12n\rho_n L_2}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right)^2$ и математическое ожидание берётся относительно всей случайности.*

**Доказательство леммы.** Отметим, что $R_p^2 d\left(\frac{x-x_0}{R_p}\right)$ является сильно выпуклой функцией с константой 1 относительно нормы $\|\cdot\|_p$. Так как $0 = \arg\min d(x)$, мы получаем, что для прокс-функции $\bar{d}(x) = R_p^2 d\left(\frac{x-x_0}{R_p}\right)$ и соответствующей дивергенции Брегмана $\bar{V}[x_0](x)$,

$$\Theta_p = \bar{V}[x_0](x_*) = \bar{d}(x_*) - \bar{d}(x_0) - \langle\nabla\bar{d}(x_0), x_* - x_0\rangle = \bar{d}(x_*) \leq \frac{R_p^2\Omega_p}{2}.$$

Применяя Теорему 1 и беря дополнительно математическое ожидание по $x_0$, мы завершаем доказательство леммы. Лемма Доказана.

**Доказательство Теоремы 3** Докажем по индукции, что

$$\mathbb{E}\|u_k - x^*\|_p^2 \leq R_k^2 = R_p^2 2^{-k} + \frac{4\Delta}{\mu_p}\left(1 - 2^{-k}\right). \tag{60}$$

Для $k = 0$ это неравенство очевидно выполнено. Предположим, что оно выполнено для некоторого $k \geq 0$, и докажем индукционный переход. Применяя Лемму 10 на шаге $k$ Алгоритма 3, получаем

$$\mathbb{E}f(u_{k+1}) - f^* = \mathbb{E}f(y_{N_0}) - f^* \leqslant \frac{aL_2R_k^2\Omega_p}{N_0^2} + \frac{b\sigma^2 N_0}{m_k L_2} + \Delta.$$

По определению $N_0$ мы имеем

$$\frac{aL_2R_k^2\Omega_p}{N_0^2} \leqslant \frac{aL_2R_k^2\Omega_p}{\frac{8aL_2\Omega_p}{\mu_p}} = \frac{\mu_p R_k^2}{8}.$$

Из определения $m_k$ мы получаем

$$m_k \geqslant \frac{8b\sigma^2 N_0}{L_2\mu_p R_p^2 2^{-k}} \geqslant \frac{8b\sigma^2 N_0}{L_2\mu_p\left(R_p^2 2^{-k} + \frac{4\Delta}{\mu_p}\left(1-2^{-k}\right)\right)} = \frac{8b\sigma^2 N_0}{L_2\mu_p R_k^2}$$

и

$$\frac{b\sigma^2 N_0}{m_k L_2} \leqslant \frac{b\sigma^2 N_0}{L_2\frac{8b\sigma^2 N_0}{L_2\mu_p R_k^2}} = \frac{\mu_p R_k^2}{8}.$$

Поэтому

$$\begin{aligned}\mathbb{E}f(u_{k+1}) - f^* &\leqslant \tfrac{\mu_p R_k^2}{4} + \Delta = \tfrac{\mu_p}{4}\left(R_p^2 2^{-k} + \tfrac{4\Delta}{\mu_p}\left(1-2^{-k}\right)\right) + \Delta \\ &= \tfrac{\mu_p}{2}\left(R_p^2 2^{-(k+1)} + \tfrac{4\Delta}{\mu_p}\left(1-2^{-(k+1)}\right)\right) \\ &= \tfrac{\mu_p R_{k+1}^2}{2}.\end{aligned}$$

Так как $f$ сильно выпукла, мы имеем

$$\mathbb{E}\|u_{k+1} - x^*\|_p^2 \leqslant \frac{2}{\mu_p}(\mathbb{E}f(u_{k+1}) - f^*) \leqslant R_{k+1}^2.$$

Это завершает доказательство индукционного шага. Попутно мы получили неравенство (18).

Остаётся оценить сложность. Чтобы сделать правую часть неравенства (18) меньше $\varepsilon$, достаточно выбрать $K = \left\lceil \log_2 \frac{\mu_p R_p^2}{\varepsilon} \right\rceil$. Чтобы оценить полное число обращений к оракулу, запишем

$$\begin{aligned}\text{Число вызовов} &= \textstyle\sum_{k=0}^{K-1} N_0 m_k \leqslant \sum_{k=0}^{K-1} N_0\left(1 + \frac{8b\sigma^2 N_0 2^k}{L_2\mu_p R_p^2}\right) \\ &\leqslant KN_0 + \frac{8b\sigma^2 N_0^2 2^K}{L_2\mu_p R_p^2} \\ &\leqslant \sqrt{\frac{8aL_2\Omega_p}{\mu_p}}\log_2\frac{\mu_p R_p^2}{\varepsilon} + \frac{8b\sigma^2}{L_2\mu_p R_p^2}\cdot\frac{8aL_2\Omega_p}{\mu_p}\cdot\frac{\mu_p R_p^2}{\varepsilon} \\ &\leqslant \sqrt{\frac{8aL_2\Omega_p}{\mu_p}}\log_2\frac{\mu_p R_p^2}{\varepsilon} + \frac{64ab\sigma^2\Omega_p}{\mu_p\varepsilon} \\ &= \widetilde{O}\left(\max\left\{n^{\frac{1}{2}+\frac{1}{q}}\sqrt{\frac{L_2\Omega_p}{\mu_p}}\log_2\frac{\mu_p R_p^2}{\varepsilon}, \frac{n^{\frac{2}{q}}\sigma^2\Omega_p}{\mu_p\varepsilon}\right\}\right),\end{aligned}$$

где мы использовали, что $a = 384n^2\rho_n$, $b = \frac{4}{n}$ и $\rho_n$ задаётся в Лемме 1. Теорема доказана.

### 5.2. Неускоренный метод

**Лемма 11.** *Пусть мы запускаем метод RDD (Algorithm 2) из случайной точки $x_0$, такой что $\mathbb{E}_{x_0}\|x^* - x_0\|_p^2 \leqslant R_p^2$, и используем функцию $R_p^2 d\left(\frac{x-x_0}{R_p}\right)$ в качестве прокс-функции, причём запускаем метод RDD на $N_0$ итераций. Тогда*

$$\mathbb{E}[f(y_{N_0})] - f^* \leqslant \frac{aL_2R_p^2\Omega_p}{N_0} + \frac{b\sigma^2}{mL_2} + \Delta,$$

*где* $a = 192n\rho_n$, $b = 2$, $\Delta = \frac{n}{12L_2}\Delta_\zeta + \frac{4n}{3L_2}\Delta_\eta^2 + \frac{8\sqrt{2nR_p^2\Omega_p}}{N_0}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right) + \frac{N_0}{3L_2\rho_n}\left(\frac{\sqrt{\Delta_\zeta}}{2} + 2\Delta_\eta\right)^2$ *и математическое ожидание берётся относительно всей случайности.*

**Доказательство леммы.** Заметим, что $R_p^2 d\left(\frac{x-x_0}{R_p}\right)$ является сильно выпуклой функцией с константой 1 относительно нормы $\|\cdot\|_p$. Так как $0 = \arg\min d(x)$, мы получаем, что для прокс-функции $\bar{d}(x) = R_p^2 d\left(\frac{x-x_0}{R_p}\right)$ и соответствующей дивергенции Брегмана $\bar{V}[x_0](x)$,

$$\Theta_p = \bar{V}[x_0](x_*) = \bar{d}(x_*) - \bar{d}(x_0) - \langle \nabla\bar{d}(x_0), x_* - x_0\rangle = \bar{d}(x_*) \leq \frac{R_p^2\Omega_p}{2}.$$

Применяя Теорему 2 и беря дополнительное математическое ожидание по $x_0$, мы завершаем доказательство леммы. Лемма доказана.

**Доказательство Теоремы 4.** Докажем по индукции, что

$$\mathbb{E}\|u_k - x^*\|_p^2 \leq R_k^2 = R_p^2 2^{-k} + \frac{4\Delta}{\mu_p}\left(1 - 2^{-k}\right). \tag{61}$$

Для $k = 0$ это неравенство очевидно выполнено. Предположим, что оно выполнено для некоторого $k \geq 0$ и докажем индукционный переход. Применяя Лемму 11 для шага $k$ Алгоритма 4, мы получаем

$$\mathbb{E}f(u_{k+1}) - f^* = \mathbb{E}f(y_{N_0}) - f^* \leqslant \frac{aL_2R_k^2\Omega_p}{N_0} + \frac{b\sigma^2}{m_kL_2} + \Delta.$$

По определению $N_0$ имеем

$$\frac{aL_2R_k^2\Omega_p}{N_0} \leqslant \frac{aL_2R_k^2\Omega_p}{\frac{8aL_2\Omega_p}{\mu_p}} = \frac{\mu_pR_k^2}{8}.$$

Из определения $m_k$ получаем

$$m_k \geqslant \frac{8b\sigma^2}{L_2\mu_pR_p^2 2^{-k}} \geqslant \frac{8b\sigma^2}{L_2\mu_p\left(R_p^2 2^{-k} + \frac{4\Delta}{\mu_p}\left(1 - 2^{-k}\right)\right)} = \frac{8b\sigma^2}{L_2\mu_pR_k^2}$$

и

$$\frac{b\sigma^2}{m_k L_2} \leqslant \frac{b\sigma^2}{L_2 \frac{8b\sigma^2}{L_2\mu_p R_k^2}} = \frac{\mu_p R_k^2}{8}.$$

Поэтому

$$\begin{aligned}\mathbb{E}f(u_{k+1}) - f^* &\leqslant \tfrac{\mu_p R_k^2}{4} + \Delta = \tfrac{\mu_p}{4}\left(R_p^2 2^{-k} + \tfrac{4\Delta}{\mu_p}\left(1 - 2^{-k}\right)\right) + \Delta \\ &= \tfrac{\mu_p}{2}\left(R_p^2 2^{-(k+1)} + \tfrac{4\Delta}{\mu_p}\left(1 - 2^{-(k+1)}\right)\right) \\ &= \tfrac{\mu_p R_{k+1}^2}{2}.\end{aligned}$$

Так как $f$ сильно выпукла, мы имеем

$$\mathbb{E}\|u_{k+1} - x^*\|_p^2 \leqslant \frac{2}{\mu_p}(\mathbb{E}f(u_{k+1}) - f^*) \leqslant R_{k+1}^2.$$

Это завершает доказательство индукционного шага. Попутно мы получили неравенство (21).

Остаётся оценить сложность. Чтобы сделать правую часть (21) меньше $\varepsilon$, достаточно выбрать $K = \left\lceil \log_2 \frac{\mu_p R_p^2}{\varepsilon} \right\rceil$. Чтобы оценить общее число обращений к оракулу, запишем

$$\begin{aligned}\text{Число обращений} &= \textstyle\sum_{k=0}^{K-1} N_0 m_k \leqslant \sum_{k=0}^{K-1} N_0 \left(1 + \frac{8b\sigma^2 2^k}{L_2\mu_p R_p^2}\right) \\ &\leqslant KN_0 + \tfrac{8b\sigma^2 N_0 2^K}{L_2\mu_p R_p^2} \\ &\leqslant \tfrac{8aL_2\Omega_p}{\mu_p}\log_2\tfrac{\mu_p R_p^2}{\varepsilon} + \tfrac{8b\sigma^2}{L_2\mu_p R_p^2}\cdot\tfrac{8aL_2\Omega_p}{\mu_p}\cdot\tfrac{\mu_p R_p^2}{\varepsilon} \\ &\leqslant \tfrac{8aL_2\Omega_p}{\mu_p}\log_2\tfrac{\mu_p R_p^2}{\varepsilon} + \tfrac{64ab\sigma^2\Omega_p}{\mu_p\varepsilon} \\ &= \widetilde{O}\left(\max\left\{\tfrac{n^{\frac{2}{q}}L_2\Omega_p}{\mu_p}\log_2\tfrac{\mu_p R_p^2}{\varepsilon}, \tfrac{n^{\frac{2}{q}}\sigma^2\Omega_p}{\mu_p\varepsilon}\right\}\right),\end{aligned}$$

где мы использовали, что $a = 192n\rho_n$, $b = 2$ и $\rho_n$ задано в Лемме 1. Теорема доказана.

## 6. Технические результаты

### 6.1. Доказательство основной технической леммы (Леммы 1)

Докажем вспомогательное неравенство:

$$\mathbb{E}[||e||_q^2] \leqslant (q-1)n^{\frac{2}{q}-1}, \quad 2 \leqslant q < \infty. \tag{62}$$

Во-первых,

$$\mathbb{E}[||e||_q^2] = \mathbb{E}\left[\left(\sum_{k=1}^n |e_k|^q\right)^{\frac{2}{q}}\right] \overset{①}{\leqslant} \left(\mathbb{E}\left[\sum_{k=1}^n |e_k|^q\right]\right)^{\frac{2}{q}} \overset{②}{=} \left(n\mathbb{E}[|e_2|^q]\right)^{\frac{2}{q}}, \tag{63}$$

где ① выполнено в силу вероятностного неравенства Йенсена (функция $\varphi(x) = x^{\frac{2}{q}}$ является вогнутой, так как $q \geqslant 2$), а переход ② корректен в силу линейности математического ожидания и одинаковой распределённости компонент вектора $e$.

Во-вторых, воспользуемся тем фактом (лемма Пуанкаре), что

$$e \stackrel{d}{=} \frac{\xi}{\sqrt{\xi_1^2 + \dots + \xi_n^2}}, \tag{64}$$

где $\xi = (\xi_1, \xi_2, \dots, \xi_n)^\top$ — $n$-мерный гауссовский случайный вектор с нулевым математическим ожиданием и единичной ковариационной матрицей, а $\stackrel{d}{=}$ обозначает равенство по распределению. Тогда

$$\begin{aligned}
\mathbb{E}[|e_2|^q] &= \mathbb{E}\left[\frac{|\xi_2|^q}{\left(\xi_1^2+\dots+\xi_n^2\right)^{\frac{q}{2}}}\right] \\
&= \int\limits_{\mathcal{R}^n} \cdots \int |x_2|^q \left(\sum_{k=1}^n x_k^2\right)^{-\frac{q}{2}} \cdot \frac{1}{(2\pi)^{\frac{n}{2}}} \cdot \exp\left(-\frac{1}{2}\sum_{k=1}^n x_k^2\right) dx_1 \dots dx_n.
\end{aligned}$$

Перейдём к сферическим координатам:

$$\begin{aligned}
x_1 &= r\cos\varphi \sin\theta_1 \dots \sin\theta_{n-2}, \\
x_2 &= r\sin\varphi \sin\theta_1 \dots \sin\theta_{n-2}, \\
x_3 &= r\cos\theta_1 \sin\theta_2 \dots \sin\theta_{n-2}, \\
x_4 &= r\cos\theta_2 \sin\theta_3 \dots \sin\theta_{n-2}, \\
&\dots \\
x_n &= r\cos\theta_{n-2}, \\
&r > 0,\ \varphi \in [0, 2\pi),\ \theta_i \in [0, \pi],\ i = \overline{1, n-2},
\end{aligned}$$

якобиан преобразования координат равен

$$\det\left(\frac{\partial(x_1, \dots, x_n)}{\partial(r, \varphi, \theta_1, \theta_2, \dots, \theta_{n-2})}\right) = r^{n-1}\sin\theta_1(\sin\theta_2)^2 \dots (\sin\theta_{n-2})^{n-2}.$$

Тогда математическое ожидание $\mathbb{E}[|e_2|^q]$ можно записать в виде:

$$\begin{aligned}
\mathbb{E}[|e_2|^q] &= \int\limits_{\substack{r>0,\,\varphi\in[0,2\pi),\\ \theta_i\in[0,\pi],\,i=\overline{1,n-2}}} \cdots \int r^{n-1}|\sin\varphi|^q |\sin\theta_1|^{q+1}|\sin\theta_2|^{q+2} \dots |\sin\theta_{n-2}|^{q+n-2} \\
&\cdot \frac{e^{-\frac{r^2}{2}}}{(2\pi)^{\frac{n}{2}}} dr \dots d\theta_{n-2} \\
&= \frac{1}{(2\pi)^{\frac{n}{2}}} I_r \cdot I_\varphi \cdot I_{\theta_1} \cdot I_{\theta_2} \cdot \dots \cdot I_{\theta_{n-2}},
\end{aligned}$$

где

$$\begin{aligned}
I_r &= \int\limits_0^{+\infty} r^{n-1} e^{-\frac{r^2}{2}} dr, \\
I_\varphi &= \int\limits_0^{2\pi} |\sin\varphi|^q d\varphi = 2\int\limits_0^{\pi} |\sin\varphi|^q d\varphi, \\
I_{\theta_i} &= \int\limits_0^{\pi} |\sin\theta_i|^{q+i} d\theta_i,\ i = \overline{1,\, n-2}.
\end{aligned}$$

Вычислим эти интегралы. Начнём с $I_r$:

$$I_r = \int\limits_0^{+\infty} r^{n-1} e^{-\frac{r^2}{2}} dr = /\text{замена } r = \sqrt{2t}/ = \int\limits_0^{+\infty} (2t)^{\frac{n}{2}-1} e^{-t} dt = 2^{\frac{n}{2}-1}\Gamma(\tfrac{n}{2}).$$

Чтобы вычислить остальные интегралы, будет полезно рассмотреть следующий интеграл ($\alpha > 0$):

$$\begin{aligned}\int\limits_0^{\pi} |\sin\varphi|^{\alpha} d\varphi &= 2\int\limits_0^{\frac{\pi}{2}} |\sin\varphi|^{\alpha} d\varphi = 2\int\limits_0^{\frac{\pi}{2}} (\sin^2\varphi)^{\frac{\alpha}{2}} d\varphi = /\text{замена } t = \sin^2\varphi/ \\ &= \int\limits_0^1 t^{\frac{\alpha-1}{2}} (1-t)^{-\frac{1}{2}} dt = B(\tfrac{\alpha+1}{2},\, \tfrac{1}{2}) = \tfrac{\Gamma(\frac{\alpha+1}{2})\Gamma(\frac{1}{2})}{\Gamma(\frac{\alpha+2}{2})} = \sqrt{\pi}\tfrac{\Gamma(\frac{\alpha+1}{2})}{\Gamma(\frac{\alpha+2}{2})}.\end{aligned}$$

Отсюда получаем, что

$$\begin{aligned}\mathbb{E}[|e_2|^q] &= \tfrac{1}{(2\pi)^{\frac{n}{2}}} I_r \cdot I_{\varphi} \cdot I_{\theta_1} \cdot I_{\theta_2} \cdot \ldots \cdot I_{\theta_{n-2}} \\ &= \tfrac{2^{\frac{n}{2}-1}}{(2\pi)^{\frac{n}{2}}} \cdot \Gamma(\tfrac{n}{2}) \cdot 2\tfrac{\sqrt{\pi}\Gamma(\frac{q+1}{2})}{\Gamma(\frac{q+2}{2})} \cdot \tfrac{\sqrt{\pi}\Gamma(\frac{q+2}{2})}{\Gamma(\frac{q+3}{2})} \cdot \tfrac{\sqrt{\pi}\Gamma(\frac{q+3}{2})}{\Gamma(\frac{q+4}{2})} \cdot \ldots \cdot \tfrac{\sqrt{\pi}\Gamma(\frac{q+n-1}{2})}{\Gamma(\frac{q+n}{2})} \\ &= \tfrac{1}{\sqrt{\pi}} \cdot \tfrac{\Gamma(\frac{n}{2})\Gamma(\frac{q+1}{2})}{\Gamma(\frac{q+n}{2})}.\end{aligned} \tag{65}$$

Покажем, что $\forall\, q \geqslant 2$

$$\frac{1}{\sqrt{\pi}} \cdot \frac{\Gamma(\frac{n}{2})\Gamma(\frac{q+1}{2})}{\Gamma(\frac{q+n}{2})} \leqslant \left(\frac{q-1}{n}\right)^{\frac{q}{2}}. \tag{66}$$

Сначала убедимся, что неравенство (66) выполнено для $q = 2$ (и произвольного $n$):

$$\frac{1}{\sqrt{\pi}} \cdot \frac{\Gamma(\frac{n}{2})\Gamma(\frac{2+1}{2})}{\Gamma(\frac{2+n}{2})} - \frac{1}{n} = \frac{1}{\sqrt{\pi}} \cdot \frac{\Gamma(\frac{n}{2}) \cdot \frac{1}{2}\Gamma(\frac{1}{2})}{\frac{n}{2}\Gamma(\frac{n}{2})} - \frac{1}{n} = \frac{1}{n} - \frac{1}{n} = 0 \leqslant 0.$$

Рассмотрим функцию (вообще говоря, двух аргументов)

$$f_n(q) = \frac{1}{\sqrt{\pi}} \cdot \frac{\Gamma(\frac{n}{2})\Gamma(\frac{q+1}{2})}{\Gamma(\frac{q+n}{2})} - \left(\frac{q-1}{n}\right)^{\frac{q}{2}}$$

при $q \geqslant 2$. Также введём в рассмотрение функцию $\psi(x) = \frac{d(\ln(\Gamma(x)))}{dx}$ при $x > 0$ (*дигамма-функция*). Для гамма-функции выполняется тождество

$$\Gamma(x+1) = x\Gamma(x),\ x > 0.$$

Возьмём от этого тождества логарифм и продифференцируем по $x$:

$$\begin{gathered}\ln\Gamma(x+1) = \ln\Gamma(x) + \ln x, \\ \tfrac{d(\ln(\Gamma(x+1)))}{dx} = \tfrac{d(\ln(\Gamma(x)))}{dx} + \tfrac{1}{x},\end{gathered}$$

что можно записать через дигамма-функцию:

$$\psi(x+1) = \psi(x) + \frac{1}{x}. \tag{67}$$

Покажем, что дигамма-функция возрастает при $x > 0$. Для этого докажем неравенство:

$$\left(\Gamma'(x)\right)^2 < \Gamma(x)\Gamma''(x). \tag{68}$$

Действительно,

$$\begin{aligned}\left(\Gamma'(x)\right)^2 &= \left(\int\limits_0^{+\infty} e^{-t}\ln t \cdot t^{x-1}dt\right)^2 \\ &\overset{①}{<} \int\limits_0^{+\infty}\left(e^{-\frac{t}{2}}t^{\frac{x-1}{2}}\right)^2 dt \cdot \int\limits_0^{+\infty}\left(e^{-\frac{t}{2}}t^{\frac{x-1}{2}}\ln t\right)^2 dt \\ &= \underbrace{\int\limits_0^{+\infty} e^{-t}t^{x-1}dt}_{\Gamma(x)} \cdot \underbrace{\int\limits_0^{+\infty} e^{t}t^{x-1}\ln^2 t dt}_{\Gamma''(x)},\end{aligned}$$

где ① следует из неравенства Коши-Буняковского (причём неравенство строгое, ибо функции $e^{-\frac{t}{2}}t^{\frac{x-1}{2}}$ и $e^{-\frac{t}{2}}t^{\frac{x-1}{2}}\ln t$ линейно независимы). Из неравенства (68) следует, что

$$\frac{d^2(\ln\Gamma(x))}{dx^2} = \left(\frac{\Gamma'(x)}{\Gamma(x)}\right)' = \frac{\Gamma''(x)}{\Gamma(x)} - \frac{\left(\Gamma'(x)\right)^2}{\left(\Gamma(x)\right)^2} \overset{(68)}{>} 0,$$

то есть дигамма-функция возрастает.

Теперь покажем, что $f_n(q)$ убывает на отрезке $[2, +\infty)$. Для этого достаточно рассмотреть $\ln(f(q))$:

$$\begin{aligned}\ln(f_n(q)) &= \ln\left(\tfrac{\Gamma(\frac{n}{2})}{\sqrt{\pi}}\right) + \ln\left(\Gamma\left(\tfrac{q+1}{2}\right)\right) - \ln\left(\Gamma\left(\tfrac{q+n}{2}\right)\right) - \tfrac{q}{2}\left(\ln(q-1) - \ln n\right), \\ \tfrac{d(\ln(f_n(q)))}{dq} &= \tfrac{1}{2}\psi\left(\tfrac{q+1}{2}\right) - \tfrac{1}{2}\psi\left(\tfrac{q+n}{2}\right) - \tfrac{1}{2}\ln(q-1) - \tfrac{q}{2(q-1)} + \tfrac{1}{2}\ln n.\end{aligned}$$

Покажем, что $\frac{d(\ln(f_n(q)))}{dq} < 0$ при $q \geqslant 2$. Пусть $k = \lfloor \frac{n}{2} \rfloor$ (ближайшее целое число, не превосходящее $\frac{n}{2}$). Тогда $\psi\left(\frac{q+n}{2}\right) > \psi\left(k - 1 + \frac{q+1}{2}\right)$ и $\ln n \leqslant \ln(2k+1)$, откуда следует, что

$$\begin{aligned}\tfrac{d(\ln(f_n(q)))}{dq} &< \tfrac{1}{2}\left(\psi\left(\tfrac{q+1}{2}\right) - \psi\left(k - 1 + \tfrac{q+1}{2}\right)\right) - \tfrac{1}{2}\ln(q-1) - \tfrac{q}{2(q-1)} + \tfrac{1}{2}\ln(2k+1) \\ &\overset{(67)}{=} \tfrac{1}{2}\left(\psi\left(\tfrac{q+1}{2}\right) - \sum_{i=1}^{k-1}\tfrac{1}{\frac{q+1}{2}+k-i-1} - \psi\left(\tfrac{q+1}{2}\right)\right) - \tfrac{q}{2(q-1)} + \tfrac{1}{2}\ln\left(\tfrac{2k+1}{q-1}\right) \\ &\overset{①}{\leqslant} -\tfrac{1}{2}\sum_{i=1}^{k-1}\tfrac{2}{q-1+2k-2i} - \tfrac{1}{q-1} + \tfrac{1}{2}\ln\left(\tfrac{2k+1}{q-1}\right) \\ &= -\tfrac{1}{2}\left(\tfrac{2}{q-1} + \tfrac{2}{q+1} + \tfrac{2}{q+3} + \ldots + \tfrac{2}{q+2k-3}\right) + \tfrac{1}{2}\ln\left(\tfrac{2k+1}{q-1}\right) \\ &\overset{②}{<} -\tfrac{1}{2}\ln\left(\tfrac{q+2k-1}{q-1}\right) + \tfrac{1}{2}\ln\left(\tfrac{2k+1}{q-1}\right) \overset{③}{\leqslant} -\tfrac{1}{2}\ln\left(\tfrac{2k+1}{q-1}\right) + \tfrac{1}{2}\ln\left(\tfrac{2k+1}{q-1}\right) = 0,\end{aligned}$$

где ① и ③ выполнены в силу неравенства $q \geqslant 2$, ② следует из оценки сверху интеграла от функции $\frac{1}{x}$ интегралом от её верхней ступенчатой мажоранты $g(x) = \frac{1}{q-1+2i}$, $x \in$

$[q-1+2i, q-1+2i+2]$, $i = \overline{0, 2k-1}$:

$$\frac{2}{q-1} + \frac{2}{q+1} + \frac{2}{q+3} + \ldots + \frac{2}{q+2k-3} > \int\limits_{q-1}^{q+2k-1} \frac{1}{x} dx = \ln\left(\frac{q+2k-1}{q-1}\right).$$

Итак, мы показали, что $\frac{d(\ln(f_n(q)))}{dq} < 0$ для $q \geqslant 2$ и произвольного натурального $n$. Следовательно, для любого фиксированного $n$ функция $f_n(q)$ убывает по $q$, а значит, $f_n(q) \leqslant f_n(2) = 0$, то есть справедливо неравенство (66). Отсюда и из (63), (65) получаем, что для любого $q \geqslant 2$

$$\mathbb{E}[||e||_q^2] \overset{(63)}{\leqslant} (n\mathbb{E}[|e_2|^q])^{\frac{2}{q}} \overset{(65),(66)}{\leqslant} (q-1)n^{\frac{2}{q}-1}. \tag{69}$$

Неравенство (69) нет смысла использовать при больших $q$ (относительно $n$). Рассмотрим правую часть неравенства (69) как функцию $q$ и найдём её минимум при $q \geqslant 2$. Рассмотрим $h_n(q) = \ln(q-1) + \left(\frac{2}{q} - 1\right)\ln n$ (логарифм правой части (69)). Производная $h(q)$:

$$\begin{aligned} \tfrac{dh(q)}{dq} &= \tfrac{1}{q-1} - \tfrac{2\ln n}{q^2}, \\ \tfrac{1}{q-1} - \tfrac{2\ln n}{q^2} &= 0, \\ q^2 - 2q\ln n + 2\ln n &= 0. \end{aligned}$$

Если $n \geqslant 8$, то точка минимума на множестве $[2, +\infty)$ есть

$$q_0 = \ln n\left(1 + \sqrt{1 - \frac{2}{\ln n}}\right)$$

(в случае $n \leqslant 7$ оказывается, что $q_0 = 2$; везде далее считаем, что $n \geqslant 8$). Поэтому для всех $q > q_0$ более точная оценка будет следующей:

$$\begin{aligned} \mathbb{E}[||e||_q^2] &\overset{①}{<} \mathbb{E}[||e||_{q_0}^2] \overset{(69)}{\leqslant} (q_0-1)n^{\frac{2}{q_0}-1} \overset{②}{\leqslant} (2\ln n - 1)n^{\frac{2}{\ln n}-1} \\ &= (2\ln n - 1)e^2\tfrac{1}{n} \leqslant (16\ln n - 8)\tfrac{1}{n} \leqslant (16\ln n - 8)n^{\frac{2}{q}-1}, \end{aligned} \tag{70}$$

где ① верно в силу Леммы 1, ② следует из неравенств $q_0 \leqslant 2\ln n$, $q_0 \geqslant \ln n$. Объединяя оценки (69) и (70), получаем неравенство (**??**).

Теперь перейдём к доказательству неравенства (**??**). Во-первых, получим оценку для $\sqrt{\mathbb{E}[||e||_q^4]}$. В силу вероятностного неравенства Йенсена ($q \geqslant 2$)

$$\begin{aligned} \mathbb{E}[||e||_q^4] &= \mathbb{E}\left[\left(\left(\sum_{k=1}^n |e_k|^q\right)^2\right)^{\frac{2}{q}}\right] \leqslant \left(\mathbb{E}\left[\left(\sum_{k=1}^n |e_k|^q\right)^2\right]\right)^{\frac{2}{q}} \\ &\overset{①}{\leqslant} \left(\mathbb{E}\left[\left(n\sum_{k=1}^n |e_k|^{2q}\right)\right]\right)^{\frac{2}{q}} \overset{②}{=} (n^2\mathbb{E}[|e_2|^{2q}])^{\frac{2}{q}} \\ &\overset{(65),(66)}{\leqslant} n^{\frac{4}{q}}\left(\left(\tfrac{2q-1}{n}\right)^{\frac{2q}{2}}\right)^{\frac{2}{q}} = (2q-1)^2 n^{\frac{4}{q}-2}, \end{aligned}$$

где ① следует из неравенства $\left(\sum\limits_{k=1}^{n} x_k\right)^2 \leqslant n\sum\limits_{k=1}^{n} x_k^2$ для $x_1, x_2, \ldots, x_n \in \mathcal{R}$, а ② есть следствие линейности математического ожидания и одинаковой распределённости компонент вектора $e$. Отсюда получаем оценку

$$\sqrt{\mathbb{E}[||e||_q^4]} \leqslant (2q-1)n^{\frac{2}{q}-1}. \tag{71}$$

Рассмотрим правую часть неравенства (71) как функцию $q$ и найдём её минимум при $q \geqslant 2$. Рассмотрим $h_n(q) = \ln(2q-1) + \left(\frac{2}{q} - 1\right)\ln n$ (логарифм правой части (71)). Производная $h(q)$:

$$\begin{aligned} \tfrac{dh(q)}{dq} &= \tfrac{2}{2q-1} - \tfrac{2\ln n}{q^2}, \\ \tfrac{2}{2q-1} - \tfrac{2\ln n}{q^2} &= 0, \\ q^2 - 2q\ln n + \ln n &= 0. \end{aligned}$$

Если $n \geqslant 3$, то точка минимума на множестве $[2, +\infty)$ есть

$$q_0 = \ln n\left(1 + \sqrt{1 - \frac{1}{\ln n}}\right)$$

(в случае $n \leqslant 2$ оказывается, что $q_0 = 2$; везде далее считаем, что $n \geqslant 3$). Поэтому для всех $q > q_0$ более точная оценка будет следующей:

$$\begin{aligned} \sqrt{\mathbb{E}[||e||_q^4]} &\overset{①}{<} \sqrt{\mathbb{E}[||e||_{q_0}^4]} \overset{(71)}{\leqslant} (2q_0-1)n^{\frac{2}{q_0}-1} \overset{②}{\leqslant} (4\ln n - 1)n^{\frac{2}{\ln n}-1} \\ &= (4\ln n - 1)e^2\tfrac{1}{n} \leqslant (32\ln n - 8)\tfrac{1}{n} \leqslant (32\ln n - 8)n^{\frac{2}{q}-1}, \end{aligned} \tag{72}$$

где ① верно в силу неравенства $||e||_q < ||e||_{q_0}$ для $q > q_0$, ② следует из неравенств $q_0 \leqslant 2\ln n$, $q_0 \geqslant \ln n$. Объединяя оценки (71) и (72), получаем неравенство

$$\sqrt{\mathbb{E}[||e||_q^4]} \leqslant \min\{2q-1, 32\ln n - 8\}n^{\frac{2}{q}-1}. \tag{73}$$

Теперь найдём $\mathbb{E}[\langle s, e\rangle^4]$, где $s \in \mathcal{R}^n$ — некоторый вектор. Пусть $S_n(r)$ — площадь поверхности $n$-мерной евклидовой сферы радиуса $n$, $d\sigma(e)$ — ненормированная равномерная мера на $n$-мерной евклидовой сфере. В данных обозначениях $S_n(r) = S_n(1)r^{n-1}$, $\frac{S_{n-1}(1)}{S_n(1)} = \frac{n-1}{n\sqrt{\pi}}\frac{\Gamma(\frac{n+2}{2})}{\Gamma(\frac{n+1}{2})}$. Кроме того, пусть $\varphi$ — угол между $s$ и $e$. Тогда

$$\begin{aligned} \mathbb{E}[\langle s, e\rangle^4] &= \tfrac{1}{S_n(1)}\int\limits_S \langle s, e\rangle^4 d\sigma(\varphi) = \tfrac{1}{S_n(1)}\int\limits_0^\pi ||s||_2^4 \cos^3\varphi S_{n-1}(\sin\varphi)d\varphi \\ &= ||s||_2^4 \tfrac{S_{n-1}(1)}{S_n(1)}\int\limits_0^\pi \cos^4\varphi \sin^{n-2}\varphi d\varphi \\ &= ||s||_2^4 \cdot \tfrac{n-1}{n\sqrt{\pi}}\tfrac{\Gamma(\frac{n+2}{2})}{\Gamma(\frac{n+1}{2})}\int\limits_0^\pi \cos^4\varphi \sin^{n-2}\varphi d\varphi. \end{aligned} \tag{74}$$

Отдельно вычислим интеграл:

$$\begin{aligned}\int\limits_0^{\pi}\cos^4\varphi\sin^{n-2}\varphi d\varphi &= 2\int\limits_0^{\frac{\pi}{2}}\cos^4\varphi\sin^{n-2}\varphi d\varphi = /\text{замена } t=\sin^2\varphi/\\ &= \int\limits_0^{\frac{\pi}{2}} t^{\frac{n-3}{2}}(1-t)^{\frac{3}{2}}dt = B(\tfrac{n-1}{2},\tfrac{5}{2}) = \tfrac{\Gamma(\frac{5}{2})\Gamma(\frac{n-1}{2})}{\Gamma(\frac{n+4}{2})}\\ &= \tfrac{\frac{3}{2}\cdot\frac{1}{2}\Gamma(\frac{1}{2})\Gamma(\frac{n-1}{2})}{\frac{n+2}{2}\cdot\Gamma(\frac{n+2}{2})} = \tfrac{3}{n+2}\cdot\tfrac{\sqrt{\pi}\Gamma(\frac{n-1}{2})}{2\Gamma(\frac{n+2}{2})}.\end{aligned}$$

Отсюда и из (74) получаем, что

$$\begin{aligned}\mathbb{E}[\langle s,\, e\rangle^4] &= ||s||_2^4\cdot\tfrac{n-1}{n\sqrt{\pi}}\tfrac{\Gamma(\frac{n+2}{2})}{\Gamma(\frac{n+1}{2})}\cdot\tfrac{3}{n+2}\cdot\tfrac{\sqrt{\pi}\Gamma(\frac{n-1}{2})}{2\Gamma(\frac{n+2}{2})}\\ &= ||s||_2^4\cdot\tfrac{3(n-1)}{2n(n+2)}\cdot\tfrac{\Gamma(\frac{n-1}{2})}{\frac{n-1}{2}\Gamma(\frac{n-1}{2})} = \tfrac{3||s||_2^4}{n(n+2)} \overset{①}{\leqslant} \tfrac{3||s||_2^4}{n^2}.\end{aligned} \tag{75}$$

Чтобы доказать неравенство (**??**), осталось воспользоваться (73), (75) и неравенством Коши-Буняковского ($(\mathbb{E}[XY])^2 \leqslant \mathbb{E}[X^2]\cdot\mathbb{E}[Y^2]$):

$$\mathbb{E}[\langle s,\, e\rangle^2||e||_q^2] \overset{①}{\leqslant} \sqrt{\mathbb{E}[\langle s,\, e\rangle^4]\cdot\mathbb{E}[||e||_q^4]} \leqslant \sqrt{3}||s||_2^2\min\{2q-1, 32\ln n - 8\}n^{\frac{2}{q}-2}.$$

Лемма доказана.

### 6.2. Остальные технические результаты

**Лемма 12.** *Пусть $a_0,\ldots,a_{N-1},b,R_1,\ldots,R_{N-1}$ являются такими неотрицательными числами, что*

$$R_l \leqslant \sqrt{2}\cdot\sqrt{\left(\sum_{k=0}^{l-1}a_k + b\sum_{k=1}^{l-1}\alpha_{k+1}R_k\right)} \quad l=1,\ldots,N, \tag{76}$$

*где $\alpha_{k+1} = \frac{k+2}{96n^2\rho_n L_2}$ для всех $k\in\mathbb{N}$. Тогда для $l=1,\ldots,N$ выполнено неравенство*

$$\sum_{k=0}^{l-1}a_k + b\sum_{k=1}^{l-1}\alpha_{k+1}R_k \leqslant \left(\sqrt{\sum_{k=0}^{l-1}a_k} + \sqrt{2}b\cdot\frac{l^2}{96n^2\rho_n L_2}\right)^2. \tag{77}$$

**Доказательство леммы.** Для $l=1$ это неравенство тривиально. Предположим, что неравенство (77) выполнено для некотрого $l<N$, и докажем его для $l+1$. По предположению индукции и (76) мы получаем

$$R_l \leqslant \sqrt{2}\left(\sqrt{\sum_{k=0}^{l-1}a_k} + \sqrt{2}b\cdot\tfrac{l^2}{96n^2\rho_n L_2}\right), \tag{78}$$

откуда

$$\begin{aligned}
\sum_{k=0}^{l} a_k + b\sum_{k=1}^{l}\alpha_{k+1}R_k &= \sum_{k=0}^{l-1} a_k + b\sum_{k=1}^{l-1}\alpha_{k+1}R_k + a_l + b\alpha_{l+1}R_l \\
&\overset{①}{\leqslant} \left(\sqrt{\sum_{k=0}^{l-1} a_k} + \sqrt{2}b\cdot\tfrac{l^2}{96n^2\rho_n L_2}\right)^2 + a_l \\
&+\sqrt{2}b\alpha_{l+1}\left(\sqrt{\sum_{k=0}^{l-1} a_k} + \sqrt{2}b\cdot\tfrac{l^2}{96n^2\rho_n L_2}\right) \\
&= \sum_{k=0}^{l} a_k + 2\sqrt{\sum_{k=0}^{l-1} a_k}\cdot\sqrt{2}b\tfrac{l^2}{96n^2\rho_n L_2} + 2b^2\tfrac{l^4}{(96n^2\rho_n L_2)^2} \\
&+\sqrt{2}b\alpha_{l+1}\left(\sqrt{\sum_{k=0}^{l-1} a_k} + \sqrt{2}b\cdot\tfrac{l^2}{96n^2\rho_n L_2}\right) \\
&= \sum_{k=0}^{l} a_k + 2\sqrt{\sum_{k=0}^{l-1} a_k}\cdot\sqrt{2}b\left(\tfrac{l^2}{96n^2\rho_n L_2} + \tfrac{\alpha_{l+1}}{2}\right) \\
&+2b^2\left(\tfrac{l^4}{(96n^2\rho_n L_2)^2} + \alpha_{l+1}\cdot\tfrac{l^2}{96n^2\rho_n L_2}\right) \\
&\overset{②}{\leqslant} \sum_{k=0}^{l} a_k + 2\sqrt{\sum_{k=0}^{l} a_k}\cdot\sqrt{2}b\tfrac{(l+1)^2}{96n^2\rho_n L_2} + 2b^2\tfrac{(l+1)^4}{(96n^2\rho_n L_2)^2} \\
&= \left(\sqrt{\sum_{k=0}^{l} a_k} + \sqrt{2}b\cdot\tfrac{(l+1)^2}{96n^2\rho_n L_2}\right)^2,
\end{aligned}$$

где ① следует из предположения индукции и (78), ② выполнено в силу $\sum\limits_{k=0}^{l-1} a_k \leqslant \sum\limits_{k=0}^{l} a_k$ и

$$\begin{aligned}
\tfrac{l^2}{96n^2\rho_n L_2} + \tfrac{\alpha_{l+1}}{2} &= \tfrac{2l^2+l+2}{192n^2\rho_n L_2} \leqslant \tfrac{(l+1)^2}{96n^2\rho_n L_2}, \\
\tfrac{l^4}{(96n^2\rho_n L_2)^2} + \alpha_{l+1}\cdot\tfrac{l^2}{96n^2\rho_n L_2} &\leqslant \tfrac{l^4+(l+2)l^2}{(96n^2\rho_n L_2)^2} \leqslant \tfrac{(l+1)^4}{(96n^2\rho_n L_2)^2}.
\end{aligned}$$

Лемма доказана.

**Лемма 13.** *Пусть $a_0, \ldots, a_{N-1}, b, R_1, \ldots, R_{N-1}$ являются такими неотрицательными числами, что*

$$R_l \leqslant \sqrt{2}\cdot\sqrt{\left(\sum_{k=0}^{l-1} a_k + b\alpha\sum_{k=1}^{l-1} R_k\right)} \quad l = 1, \ldots, N. \tag{79}$$

*Тогда для всех $l = 1, \ldots, N$ выполнено неравенство*

$$\sum_{k=0}^{l-1} a_k + b\alpha\sum_{k=1}^{l-1} R_k \leqslant \left(\sqrt{\sum_{k=0}^{l-1} a_k} + \sqrt{2}b\alpha l\right)^2. \tag{80}$$

**Доказательство леммы.** Для $l = 1$ это ривиальное неравенство. Предположим, что (80) выполнено для некоторого $l < N$, и докажем его для $l + 1$. Из предположения

индукции и (79) получаем

$$R_l \leqslant \sqrt{2}\left(\sqrt{\sum_{k=0}^{l-1} a_k} + \sqrt{2}b\alpha l\right), \tag{81}$$

откуда

$$\begin{aligned}
\sum_{k=0}^{l} a_k + b\alpha \sum_{k=1}^{l} R_k &= \sum_{k=0}^{l-1} a_k + b\alpha \sum_{k=1}^{l-1} R_k + a_l + b\alpha R_l \\
&\overset{①}{\leqslant} \left(\sqrt{\sum_{k=0}^{l-1} a_k} + \sqrt{2}b\alpha l\right)^2 + a_l \\
&+\sqrt{2}b\alpha\left(\sqrt{\sum_{k=0}^{l-1} a_k} + \sqrt{2}b\alpha l\right) \\
&= \sum_{k=0}^{l} a_k + 2\sqrt{\sum_{k=0}^{l-1} a_k}\cdot\sqrt{2}b\alpha l + 2b^2\alpha^2 l^2 \\
&+\sqrt{2}b\alpha\left(\sqrt{\sum_{k=0}^{l-1} a_k} + \sqrt{2}b\alpha l\right) \\
&= \sum_{k=0}^{l} a_k + 2\sqrt{\sum_{k=0}^{l-1} a_k}\cdot\sqrt{2}b\alpha\left(l+\tfrac{1}{2}\right) + 2b^2\alpha^2\left(l^2+l\right) \\
&\overset{②}{\leqslant} \sum_{k=0}^{l} a_k + 2\sqrt{\sum_{k=0}^{l} a_k}\cdot\sqrt{2}b\alpha(l+1) + 2b^2\alpha^2(l+1)^2 \\
&= \left(\sqrt{\sum_{k=0}^{l} a_k} + \sqrt{2}b\alpha(l+1)\right)^2,
\end{aligned}$$

где ① следует из индукционного предположения и (81), ② выполнено в силу $\sum\limits_{k=0}^{l-1} a_k \leqslant \sum\limits_{k=0}^{l} a_k$. Лемма доказана.

## 7. Обсуждение результатов

В данной работе предлагаются следующие новые рандомизированные методы для решения задач гладкой стохастической выпуклой оптимизации: 1) Accelerated Randomized Directional Derivative (ARDD) method, 2) Randomized Directional Derivative (RDD) method, 3) Accelerated Randomized Directional Derivative method for strongly convex functions (ARDDsc), 4) Randomized Directional Derivative method for strongly convex functions (RDDsc). Для каждого из методов получены оценки на скорость сходимости, а также допустимый уровень шума для производных по направлению, генерируемых оракулом. Показано, как полученные результаты могут быть применимы для задач безградиентной гладкой стохастической выпуклой оптимизации. Кроме того, доказаны неравенства, связанные с математическим ожиданием векторной

q-нормы (для q не меньших 2) равномерно распределённого на n-мерной единичной сфере вектора, и показана связь этих неравенств с явлением концентрации равномерной меры на евклидовой сфере. Полученные новые методы позволили, среди прочего, решать задачи выпуклой гладкой оптимизации, для которых 1- и 2-нормы решения близки, почти в корень из размерности пространства раз быстрее (в смысле числа обращений к оракулу), чем предписывают известные нижние оценки, полученные без сделанного выше предположения о структуре решения.

Однако есть некоторые трудности при перенесении описанного подхода на задачи оптимизации на выпуклых замкнутых множествах. Рассмотрим частный случай задачи (1), когда оракул может по заданной точке $x$ и направлению $e$ вернуть точное значение производной по направлению $\langle \nabla f(x), e\rangle e$. Кроме того, ограничимся выбором евклидовой прокс-структуры, так как даже в таком простом случае возникают проблемы с обобщением результатов статьи [48] на случай спуска по случайному направлению на множестве. Заметим, что в такой постановке первое неравенство в Лемме 1 становится тривиальным, а второе неравенство можно усилить, используя Лемму B.10 из [26]:

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

Следуя [48], введём следующее обозначение

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

Чтобы обобщить приведённые выше 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 $e_{k+1}$), то есть, исходя из техники, используемой в [48], хотелось бы доказать оценку

$$\mathbb{E}_{e_{k+1}}\left[\mathrm{Prog}_{n\langle \nabla f(x_{k+1}),\, e_{k+1}\rangle\, e_{k+1}}(x_{k+1})\right] \leqslant n^2\left(f(x_{k+1}) - \mathbb{E}\left[f\left(y_{k+1}\right)\right]\right), \tag{82}$$

чтобы получить оценку скорости сходимости как и в случае безусловной минимизации. К сожалению, существует пример (приведён далее) выпуклой $L$-гладкой функции и замкнутого выпуклого множества, для которых (82) не выполняется.

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

$$\begin{aligned}\text{Prog}_\xi(x) &= -\min_{y\in Q}\left\{\tfrac{L}{2}\left|\left|y-x\right|\right|_2^2 + \langle \xi,\, y-x\rangle\right\}\\ &= -\min_{y\in Q}\left\{\tfrac{L}{2}\left|\left|y-x\right|\right|_2^2 + \langle \xi,\, y-x\rangle + \tfrac{1}{2L}\left|\left|\xi\right|\right|_2^2\right\} + \tfrac{1}{2L}\left|\left|\xi\right|\right|_2^2\\ &= -\min_{y\in Q}\left\{\left|\left|\tfrac{1}{\sqrt{2L}}\xi + \sqrt{\tfrac{L}{2}}\cdot y - \sqrt{\tfrac{L}{2}}\cdot x\right|\right|_2^2\right\} + \tfrac{1}{2L}\left|\left|\xi\right|\right|_2^2\\ &= -\tfrac{L}{2}\min_{y\in Q}\left\{\left|\left|y-\left(x-\tfrac{1}{L}\xi\right)\right|\right|_2^2\right\} + \tfrac{1}{2L}\left|\left|\xi\right|\right|_2^2,\end{aligned}$$

то есть точка, в которой достигается этот минимум,

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

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 $\text{Prog}_{n\langle \nabla f(x_{k+1}),\, e_{k+1}\rangle\, e_{k+1}}(x_{k+1})$. 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 $\widetilde{y}_{k+1}$:

$$\begin{aligned} y_{k+1} &= x_{k+1} - \frac{1}{L}s_{k+1} + r_{k+1}, \\ \widetilde{y}_{k+1} &= x_{k+1} - \frac{n}{L}s_{k+1} + \widetilde{r}_{k+1},\end{aligned} \tag{83}$$

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

Рассмотрим функцию

$$f(y) = f(x_{k+1}) + \langle \nabla f(x_{k+1}),\, y - x_{k+1}\rangle + \frac{L}{2}\left|\left|y-x_{k+1}\right|\right|_2^2 \tag{84}$$

и множество, изображённое на рисунке 1 (в качестве $\nabla f(x_{k+1})$ можно выбрать любой ненулевой вектор, а в качестве $Q$ — прямоугольный параллелепипед с достаточно длинными сторонами, в центре одной из гиперграней которого размещена точка $x_{k+1}$). Подставим в (84) значение $y = y_{k+1}$ и воспользуемся представлением $y_{k+1}$ из (83):

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

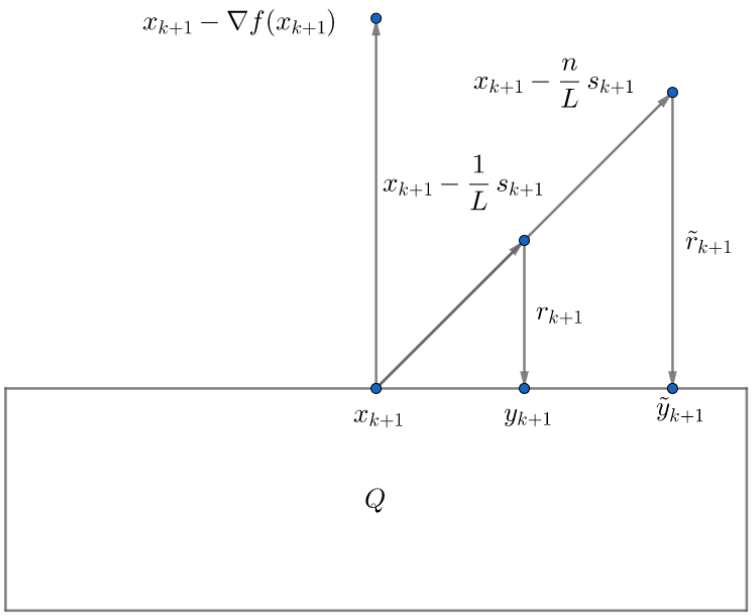


Рис. 1. Пример ситуации, когда ключевое неравенство не выполнено

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

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

или в более компактной форме

$$\begin{gathered} \tfrac{1}{2L}\langle \nabla f(x_{k+1}),\, e_{k+1}\rangle^2 - \tfrac{L}{2}\left|\left|r_{k+1}\right|\right|_2^2 + \langle r_{k+1},\, s_{k+1} - \nabla f(x_{k+1})\rangle \\ = f(x_{k+1}) - f(y_{k+1}). \end{gathered} \tag{85}$$

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,

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

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 (85) следует, что

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

Заметим, что вектор $s_{k+1}$ всегда короче (точнее, не длиннее) вектора $\nabla f(x_{k+1})$ и направлен «вниз» (то есть 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 $\langle r_{k+1},\, s_{k+1} - \nabla f(x_{k+1})\rangle \leqslant 0$, причём 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

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

Представленный контр-пример показывает трудности в перенесении предлагаемого ускоренного метода на задачи условной оптимизации.

Таким образом, одной из тем для дальнейшей работы является вопрос о перенесении результатов на задачи оптимизации на множествах простой структуры.

| | $p = 1$ |
|---|---|
| $N$ | $O\left(\sqrt{\frac{n \ln n L_2 \Theta_1}{\varepsilon}}\right)$ |
| $m$ | $O\left(\max\left\{1, \sqrt{\frac{\ln n}{n}} \cdot \frac{\sigma^2}{\varepsilon^{3/2}} \cdot \sqrt{\frac{\Theta_1}{L_2}}\right\}\right)$ |
| $\Delta_\zeta$ | $O\left(\min\left\{n(\ln n)^2 L_2^2 \Theta_1, \frac{\varepsilon^2}{n\Theta_1}, \frac{\varepsilon^{\frac{3}{2}}}{\sqrt{n \ln n}} \cdot \sqrt{\frac{L_2}{\Theta_1}}\right\}\right)$ |
| $\Delta_\eta$ | $O\left(\min\left\{\sqrt{n} \ln n L_2 \sqrt{\Theta_1}, \frac{\varepsilon}{\sqrt{n\Theta_1}}, \frac{\varepsilon^{\frac{3}{4}}}{\sqrt[4]{n \ln n}} \cdot \sqrt[4]{\frac{L_2}{\Theta_1}}\right\}\right)$ |
| O-le calls | $O\left(\max\left\{\sqrt{\frac{n \ln n L_2 \Theta_1}{\varepsilon}}, \frac{\sigma^2 \Theta_1 \ln n}{\varepsilon^2}\right\}\right)$ |
| | $p = 2$ |
| $N$ | $O\left(\sqrt{\frac{n^2 L_2 \Theta_2}{\varepsilon}}\right)$ |
| $m$ | $O\left(\max\left\{1, \frac{\sigma^2}{\varepsilon^{3/2}} \cdot \sqrt{\frac{\Theta_2}{L_2}}\right\}\right)$ |
| $\Delta_\zeta$ | $O\left(\min\left\{n^3 L_2^2 \Theta_2, \frac{\varepsilon}{n\Theta_2}, \frac{\varepsilon^{\frac{3}{2}}}{n} \cdot \sqrt{\frac{L_2}{\Theta_2}}\right\}\right)$ |
| $\Delta_\eta$ | $O\left(\min\left\{n^{\frac{3}{2}} L_2 \sqrt{\Theta_2}, \frac{\varepsilon}{\sqrt{n\Theta_2}}, \frac{\varepsilon^{\frac{3}{4}}}{\sqrt{n}} \cdot \sqrt[4]{\frac{L_2}{\Theta_2}}\right\}\right)$ |
| O-le calls | $O\left(\max\left\{\sqrt{\frac{n^2 L_2 \Theta_2}{\varepsilon}}, \frac{\sigma^2 \Theta_2 n}{\varepsilon^2}\right\}\right)$ |

Таблица 1. Параметры Алгоритма 1 для случаев $p = 1$ и $p = 2$.

| | $p = 1$ |
|---|---|
| $N$ | $O\left(\frac{L_2\Theta_1 \ln n}{\varepsilon}\right)$ |
| $m$ | $O\left(\max\left\{1, \frac{\sigma^2}{\varepsilon L_2}\right\}\right)$ |
| $\Delta_\zeta$ | $O\left(\min\left\{\frac{(\ln n)^2}{n} L_2^2\Theta_1,\ \frac{\varepsilon^2}{n\Theta_1},\ \frac{\varepsilon L_2}{n}\right\}\right)$ |
| $\Delta_\eta$ | $O\left(\min\left\{\frac{\ln n}{\sqrt{n}} L_2\sqrt{\Theta_1},\ \frac{\varepsilon}{\sqrt{n\Theta_1}},\ \sqrt{\frac{\varepsilon L_2}{n}}\right\}\right)$ |
| O-le calls | $O\left(\max\left\{\frac{L_2\Theta_1 \ln n}{\varepsilon}, \frac{\sigma^2\Theta_1 \ln n}{\varepsilon^2}\right\}\right)$ |
| | $p = 2$ |
| $N$ | $O\left(\frac{nL_2\Theta_2}{\varepsilon}\right)$ |
| $m$ | $O\left(\max\left\{1, \frac{\sigma^2}{\varepsilon L_2}\right\}\right)$ |
| $\Delta_\zeta$ | $O\left(\min\left\{nL_2^2\Theta_2,\ \frac{\varepsilon^2}{n\Theta_2},\ \frac{\varepsilon L_2}{n}\right\}\right)$ |
| $\Delta_\eta$ | $O\left(\min\left\{\sqrt{n} L_2\sqrt{\Theta_2},\ \frac{\varepsilon}{\sqrt{n\Theta_2}},\ \sqrt{\frac{\varepsilon L_2}{n}}\right\}\right)$ |
| O-le calls | $O\left(\max\left\{\frac{nL_2\Theta_2}{\varepsilon}, \frac{n\sigma^2\Theta_2}{\varepsilon^2}\right\}\right)$ |

Таблица 2. Параметры Алгоритма 2 для случаев $p = 1$ и $p = 2$.

| | |
|---|---|
| | $p = 1$ |
| $\Delta_\zeta$ | $O\left(\min\left\{\varepsilon\sqrt{\frac{L_2\mu_1}{n\ln n\Omega_1}},\ \varepsilon^2\frac{n(\ln n)^2L_2^2\Omega_1}{R_1^2\mu_1^2},\ \varepsilon\cdot\frac{\mu_1}{n\Omega_1}\right\}\right)$ |
| $\Delta_\eta$ | $O\left(\min\left\{\sqrt{\varepsilon}\sqrt[4]{\frac{L_2\mu_1}{n\ln n\Omega_1}},\ \varepsilon\frac{\sqrt{n}\ln nL_2\sqrt{\Omega_1}}{R_1\mu_1},\ \sqrt{\varepsilon}\cdot\sqrt{\frac{\mu_1}{n\Omega_1}}\right\}\right)$ |
| O-le calls | $\widetilde{O}\left(\max\left\{\sqrt{\frac{n\ln nL_2\Omega_1}{\mu_1}}\log_2\frac{\mu_1R_1^2}{\varepsilon},\ \frac{\sigma^2\Omega_1\ln n}{\mu_1\varepsilon}\right\}\right)$ |
| | $p = 2$ |
| $\Delta_\zeta$ | $O\left(\min\left\{\varepsilon\sqrt{\frac{L_2\mu_2}{n^2\Omega_2}},\ \varepsilon^2\frac{n^3L_2^2\Omega_2}{R_2^2\mu_2^2},\ \varepsilon\cdot\frac{\mu_2}{n\Omega_2}\right\}\right)$ |
| $\Delta_\eta$ | $O\left(\min\left\{\sqrt{\varepsilon}\sqrt[4]{\frac{L_2\mu_2}{n^2\Omega_2}},\ \varepsilon\frac{\sqrt{n^3}L_2\sqrt{\Omega_2}}{R_2\mu_2},\ \sqrt{\varepsilon}\cdot\sqrt{\frac{\mu_2}{n\Omega_2}}\right\}\right)$ |
| O-le calls | $\widetilde{O}\left(\max\left\{n\sqrt{\frac{L_2\Omega_2}{\mu_2}}\log_2\frac{\mu_2R_2^2}{\varepsilon},\ \frac{n\sigma^2\Omega_2}{\mu_2\varepsilon}\right\}\right)$ |

Таблица 3. Параметры Алгоритма 3 для случаев $p = 1$ и $p = 2$.

| | $p = 1$ |
|---|---|
| $\Delta_\zeta$ | $O\left(\min\left\{\frac{\varepsilon L_2}{n},\ \varepsilon^2\frac{(\ln n)^2 L_2^2}{nR_1^2\mu_1^2},\ \varepsilon\frac{\mu_1}{n\Omega_1}\right\}\right)$ |
| $\Delta_\eta$ | $O\left(\min\left\{\sqrt{\frac{\varepsilon L_2}{n}},\ \varepsilon\frac{\ln n L_2}{\sqrt{n}R_1\mu_1},\ \sqrt{\varepsilon\frac{\mu_1}{n\Omega_1}}\right\}\right)$ |
| O-le calls | $\widetilde{O}\left(\max\left\{\frac{L_2\Omega_1\ln n}{\mu_1}\log_2\frac{\mu_1 R_1^2}{\varepsilon},\ \frac{\sigma^2\Omega_1}{\mu_1\varepsilon}\right\}\right)$ |
| | $p = 2$ |
| $\Delta_\zeta$ | $O\left(\min\left\{\frac{\varepsilon L_2}{n},\ \varepsilon^2\frac{nL_2^2}{R_2^2\mu_2^2},\ \varepsilon\frac{\mu_2}{n\Omega_2}\right\}\right)$ |
| $\Delta_\eta$ | $O\left(\min\left\{\sqrt{\frac{\varepsilon L_2}{n}},\ \varepsilon\frac{\sqrt{n}L_2}{R_2\mu_2},\ \sqrt{\varepsilon\frac{\mu_2}{n\Omega_2}}\right\}\right)$ |
| O-le calls | $\widetilde{O}\left(\max\left\{\frac{nL_2\Omega_2}{\mu_2}\log_2\frac{\mu_2 R_2^2}{\varepsilon},\ \frac{n\sigma^2\Omega_2}{\mu_2\varepsilon}\right\}\right)$ |

Таблица 4. Параметры Алгоритма 4 для случаев $p = 1$ и $p = 2$.

# Список литературы


1. Rosenbrock H. H. An Automatic Method for Finding the Greatest or Least Value of a Function // The Computer Journal. 1960. Vol. 3, no. 3. P. 175–184. URL: http://dx.doi.org/10.1093/comjnl/3.3.175.
2. Brent R. 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. New York, NY, USA: John Wiley & Sons, Inc., 2003. ISBN: 0471330523.
4. Conn A., Scheinberg K., Vicente L. Introduction to Derivative-Free Optimization. Society for Industrial and Applied Mathematics, 2009. URL: http://epubs.siam.org/doi/abs/10.1137/1.9780898718768.
5. Cauchy A. Méthode générale pour la résolution des systémes d'équations simultanées // Comptes rendus hebdomadaires des séances de l'Académie des sciences. 1847. Vol. 55. P. 536–538.
6. Wengert R. E. A Simple Automatic Derivative Evaluation Program // Commun. ACM. 1964. — . Vol. 7, no. 8. P. 463–464. URL: http://doi.acm.org/10.1145/355586.364791.
7. Kim K., Nesterov Y., Skokov V., Cherkasskii B. Effektivnii algoritm vychisleniya proisvodnyh i ekstremalnye zadachi (Efficient algorithm for calculation of derivatives and extreme problems) // Ekonomika i matematicheskie metody. 1984. Vol. 20, no. 2. P. 309–318.
8. Prigozhin L. Variational model of sandpile growth // European Journal of Applied Mathematics. 1996. Vol. 7, no. 3. P. 225–235.
9. Barrett J. W., Prigozhin L. Lakes and rivers in the landscape: A quasi-variational inequality approach // Interfaces and Free Boundaries. 2014. Vol. 16, no. 2. P. 269–296.
10. Barrett J. W., Prigozhin L. A QUASI-VARIATIONAL INEQUALITY PROBLEM IN SUPERCONDUCTIVITY // Mathematical Models and Methods in Applied Sciences. 2010. Vol. 20, no. 5. P. 679–706.
11. Mordukhovich B. S., Outrata J. V. Coderivative Analysis of Quasi-variational Inequalities with Applications to Stability and Optimization // SIAM Journal on Optimiza-

tion. 2007. Vol. 18, no. 2. P. 389–412. URL: https://doi.org/10.1137/060665609.

12. 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:1707.08486. 2017.
13. Lan G. An optimal method for stochastic composite optimization // Mathematical Programming. 2012. — Jun. Vol. 133, no. 1. P. 365–397. Firs appeared in June 2008. URL: https://doi.org/10.1007/s10107-010-0434-y.
14. Devolder O. Stochastic first order methods in smooth convex optimization // CORE Discussion Paper 2011/70. 2011.
15. Dvurechensky P., Gasnikov A. Stochastic Intermediate Gradient Method for Convex Problems with Stochastic Inexact Oracle // Journal of Optimization Theory and Applications. 2016. Vol. 171, no. 1. P. 121–145. URL: http://dx.doi.org/10.1007/s10957-016-0999-6.
16. Nesterov Y., Spokoiny V. Random Gradient-Free Minimization of Convex Functions // Found. Comput. Math. 2017. — . Vol. 17, no. 2. P. 527–566. First appeared in 2011 as CORE discussion paper 2011/16. URL: https://doi.org/10.1007/s10208-015-9296-2.
17. Nesterov Y. Efficiency of Coordinate Descent Methods on Huge-Scale Optimization Problems // SIAM Journal on Optimization. 2012. Vol. 22, no. 2. P. 341–362. First appeared in 2010 as CORE discussion paper 2010/2. URL: https://doi.org/10.1137/100802001.
18. Lee Y. T., Sidford A. Efficient Accelerated Coordinate Descent Methods and Faster Algorithms for Solving Linear Systems // Proceedings of the 2013 IEEE 54th Annual Symposium on Foundations of Computer Science. FOCS '13. Washington, DC, USA: IEEE Computer Society, 2013. P. 147–156. First appeared in arXiv:1305.1922. URL: http://dx.doi.org/10.1109/FOCS.2013.24.
19. Fercoq O., Richtárik P. Accelerated, parallel, and proximal coordinate descent // SIAM Journal on Optimization. 2015. Vol. 25, no. 4. P. 1997–2023. First appeared in arXiv:1312.5799.
20. Lin Q., Lu Z., Xiao L. An Accelerated Proximal Coordinate Gradient Method // Advances in Neural Information Processing Systems 27 / Ed. by Z. Ghahramani, M. Welling, C. Cortes et al. Curran Associates, Inc., 2014. P. 3059–3067.

First appeared in arXiv:1407.1296. URL: http://papers.nips.cc/paper/5356-an-accelerated-proximal-coordinate-gradient-method.pdf.

21. Shalev-Shwartz S., Zhang T. Accelerated Proximal Stochastic Dual Coordinate Ascent for Regularized Loss Minimization // Proceedings of the 31st International Conference on Machine Learning / Ed. by E. P. Xing, T. Jebara. Vol. 32 of Proceedings of Machine Learning Research. Bejing, China: PMLR, 2014. — 22–24 Jun. P. 64–72. First appeared in arXiv:1309.2375. URL: http://proceedings.mlr.press/v32/shalev-shwartz14.html.
22. Nesterov Y., Stich S. U. Efficiency of the Accelerated Coordinate Descent Method on Structured Optimization Problems // SIAM Journal on Optimization. 2017. Vol. 27, no. 1. P. 110–123. First presented in May 2015 http://www.mathnet.ru:8080/PresentFiles/11909/7_nesterov.pdf. URL: https://doi.org/10.1137/16M1060182.
23. Allen-Zhu Z., Qu Z., Richtarik P., Yuan Y. Even Faster Accelerated Coordinate Descent Using Non-Uniform Sampling // Proceedings of The 33rd International Conference on Machine Learning / Ed. by M. F. Balcan, K. Q. Weinberger. Vol. 48 of Proceedings of Machine Learning Research. New York, New York, USA: PMLR, 2016. — 20–22 Jun. P. 1110–1119. First appeared in arXiv:1512.09103. URL: http://proceedings.mlr.press/v48/allen-zhuc16.html.
24. Gasnikov A., Dvurechensky P., Usmanova I. On accelerated randomized methods // Proceedings of Moscow Institute of Physics and Technology. 2016. Vol. 8, no. 2. P. 67–100. In Russian, first appeared in arXiv:1508.02182.
25. Dang C. D., Lan G. Stochastic Block Mirror Descent Methods for Nonsmooth and Stochastic Optimization // SIAM J. on Optimization. 2015. — . Vol. 25, no. 2. P. 856–881. URL: https://doi.org/10.1137/130936361.
26. Bogolubsky L., Dvurechensky P., Gasnikov A. et al. Learning Supervised PageRank with Gradient-Based and Gradient-Free Optimization Methods // Advances in Neural Information Processing Systems 29 / Ed. by D. D. Lee, M. Sugiyama, U. V. Luxburg et al. Curran Associates, Inc., 2016. P. 4914–4922. arXiv:1603.00717. URL: http://papers.nips.cc/paper/6565-learning-supervised-pagerank-with-gradient-based-and-gradient-free-optimiz pdf.

27. Cesa-bianchi N., Conconi A., Gentile C. On the Generalization Ability of On-Line Learning Algorithms // Advances in Neural Information Processing Systems 14 / Ed. by T. G. Dietterich, S. Becker, Z. Ghahramani. MIT Press, 2002. P. 359–366. URL: http://papers.nips.cc/paper/2113-on-the-generalization-ability-of-on-line-learning-algorithms.pdf.
28. Duchi J. C., Jordan M. I., Wainwright M. J., Wibisono A. Optimal Rates for Zero-Order Convex Optimization: The Power of Two Function Evaluations // IEEE Trans. Information Theory. 2015. Vol. 61, no. 5. P. 2788–2806. arXiv:1312.2139. URL: https://doi.org/10.1109/TIT.2015.2409256.
29. 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 // Automation and Remote Control. 2016. — Nov. Vol. 77, no. 11. P. 2018–2034. arXiv:1412.3890. URL: http://dx.doi.org/10.1134/S0005117916110114.
30. Gasnikov A. V., Krymova E. A., Lagunovskaya A. A. et al. Stochastic online optimization. Single-point and multi-point non-linear multi-armed bandits. Convex and strongly-convex case // Automation and Remote Control. 2017. — Feb. Vol. 78, no. 2. P. 224–234. arXiv:1509.01679. URL: http://dx.doi.org/10.1134/S0005117917020035.
31. Shamir O. An Optimal Algorithm for Bandit and Zero-Order Convex Optimization with Two-Point Feedback // Journal of Machine Learning Research. 2017. Vol. 18. P. 52:1–52:11. First appeared in arXiv:1507.08752. URL: http://jmlr.org/papers/v18/papers/v18/16-632.html.
32. Bayandina A., Gasnikov A., Lagunovskaya A. Gradient-free two-points optimal method for non smooth stochastic convex optimization problem with additional small noise // Automation and remote control. 2018. Vol. 79, no. 7. arXiv:1701.03821.
33. Hu X., L.A. P., György A., Szepesvari C. (Bandit) Convex Optimization with Biased Noisy Gradient Oracles // Proceedings of the 19th International Conference on Artificial Intelligence and Statistics / Ed. by A. Gretton, C. C. Robert. Vol. 51 of Proceedings of Machine Learning Research. Cadiz, Spain: PMLR, 2016. — 09–11 May. P. 819–828. URL: http://proceedings.mlr.press/v51/hu16b.html.
34. Agarwal A., Dekel O., Xiao L. Optimal Algorithms for Online Convex Optimization

with Multi-Point Bandit Feedback // COLT 2010 - The 23rd Conference on Learning Theory. 2010.

35. Ghadimi S., Lan G., Zhang H. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization // Mathematical Programming. 2016. Vol. 155, no. 1. P. 267–305. arXiv:1308.6594. URL: http://dx.doi.org/10.1007/s10107-014-0846-1.
36. Ghadimi S., Lan G. Stochastic First- and Zeroth-Order Methods for Nonconvex Stochastic Programming // SIAM Journal on Optimization. 2013. Vol. 23, no. 4. P. 2341–2368. arXiv:1309.5549. URL: https://doi.org/10.1137/120880811.
37. Ben-Tal A., Nemirovski A. Lectures on Modern Convex Optimization (Lecture Notes). Personal web-page of A. Nemirovski, 2015. URL: http://www2.isye.gatech.edu/~nemirovs/Lect_ModConvOpt.pdf.
38. Vorontsova E., Gasnikov A., Gorbunov E. Accelerated Directional Search with non-Euclidean prox-structure // arXiv:1710.00162. 2017.
39. Dvurechensky P., Gasnikov A., Gorbunov E. An Accelerated Method for Derivative-Free Smooth Stochastic Convex Optimization // arXiv:1802.09022. 2018.
40. Dvurechensky P., Gasnikov A., Gorbunov E. An Accelerated Directional Derivative Method for Smooth Stochastic Convex Optimization // arXiv:1804.02394. 2018.
41. Gorbunov E., Vorontsova E., Gasnikov A. On the upper bound for the mathematical expectation of the norm of a vector uniformly distributed on the sphere and the phenomenon of concentration of uniform measure on the sphere // arXiv:1804.03722. 2018.
42. Nemirovsky A., Yudin D. Problem Complexity and Method Efficiency in Optimization. J. Wiley & Sons, New York, 1983.
43. Blum A., Hopcroft J., Kannan R. Foundations of Data Science. Vorabversion eines Lehrbuchs, 2016.
44. Ball K. An elementary introduction to modern convex geometry. Cambridge University Press, 1997. Vol. 31.
45. Boucheron S., Lugosi G., Massart P. Concentration inequalities: A nonasymptotic theory of independence. Oxford university press, 2013.
46. Milman V., Schechtman G. Asymptotic Theory of Finite Dimensional Normed Spaces. (With an Appendix by M. Gromov). Springer-Verlag, 1986.

47. Zorich V. A. Mathematical analysis in problems of natural science. MCCME, 2008.
48. Allen-Zhu Z., Orecchia L. Linear coupling: An ultimate unification of gradient and mirror descent // arXiv:1407.1537. 2014.
49. Juditsky A., Nesterov Y. Deterministic and Stochastic Primal-Dual Subgradient Algorithms for Uniformly Convex Minimization // Stochastic Systems. 2014. Vol. 4, no. 1. P. 44–80. URL: https://doi.org/10.1287/10-SSY010.