

© 2025 г. А. АХАВАН, д-р (arya.akhavan@stats.ox.ac.uk)
(Оксфордский университет, Великобритания),
А.Б. ЦЫБАКОВ, д-р физ.-мат. наук (alexandre.tsybakov@ensae.fr)
(Центр исследований по экономике и статистике (CREST);
Парижская национальная школа статистики
и экономического управления (ENSAE);
Политехнический институт Парижа, Франция)

БЕЗГРАДИЕНТНАЯ СТОХАСТИЧЕСКАЯ ОПТИМИЗАЦИЯ ДЛЯ АДДИТИВНЫХ МОДЕЛЕЙ¹

Рассматривается задача оптимизации нулевого порядка по зашумленным наблюдениям для целевой функции, удовлетворяющей условию Поляка–Лоясевича или условию сильной выпуклости. Кроме того, предполагается, что целевая функция имеет аддитивную структуру и удовлетворяет свойству гладкости высокого порядка, характеризующему гильбертовым семейством функций. Аддитивная модель для гильбертовых классов функций хорошо изучена в литературе по непараметрическому оцениванию функций; в частности, показано, что точность оценивания для такой модели существенно лучше, чем для гильбертовой модели без аддитивной структуры. В данной статье аддитивная модель изучается в задаче безградиентной оптимизации. Предлагается рандомизированная оценка градиента, позволяющая при подключении к алгоритму градиентного спуска достичь минимаксно-оптимальной ошибки оптимизации порядка $dT^{-(\beta-1)/\beta}$, где d – размерность задачи, T – количество пробных точек, а $\beta \geq 2$ – гильбертова степень гладкости. Устанавливается, что, в отличие от непараметрических задач оценивания, использование аддитивных моделей в безградиентной оптимизации не приводит к существенному выигрышу в точности.

Ключевые слова: аддитивная модель, безградиентная оптимизация, минимаксная оптимальность, условие Поляка–Лоясевича.

DOI: 10.31857/S0005231025090025, **EDN:** VMQENP

1. Введение

Аддитивное моделирование является распространенным подходом к снижению размерности в задачах непараметрического оценивания [1–3]. Оно заключается в следующем. Предполагается, что неизвестная функция $f : \mathbb{R}^d \rightarrow \mathbb{R}$, которую необходимо оценить по доступным данным, имеет вид $f(\mathbf{x}) = \sum_{j=1}^d f_j(x_j)$, где x_j – координаты вектора $\mathbf{x} \in \mathbb{R}^d$, а f_j – неизвестные функции одной переменной. Основное свойство, доказанное в литературе по аддитивным моделям непараметрической регрессии, можно сформулировать так.

¹ Исследование Арьи Ахаван поддержано Фондом исследований и инноваций Великобритании (UKRI) в рамках государственной гарантии финансирования “Горизонт Европа” (Horizon Europe) правительства Великобритании (грант № EP/Y028333/1).

Если каждая из функций f_j является β -гельдеровой (см. определение 1 ниже), то минимаксная ошибка оценивания f , поточечно или по норме L_2 , имеет порядок $n^{-\beta/(2\beta+1)}$, где n – число наблюдений [1]. Данный результат сильно отличается от того, что можно получить в задаче оценивания β -гельдеровых функций на $\mathbb{R}^d$ без аддитивной структуры: для таких функций минимаксная ошибка имеет порядок $n^{-\beta/(2\beta+d)}$ [4–7]. Таким образом, при переходе от общих моделей непараметрической регрессии к аддитивным существенно уменьшается ошибка оценивания.

В настоящей работе показывается, что подобное свойство снижения размерности не имеет места в задаче безградиентной оптимизации. Рассматривается задача минимизации неизвестной функции $f : \mathbb{R}^d \rightarrow \mathbb{R}$, удовлетворяющей аддитивной модели, когда доступны только последовательные зашумленные оценки значений f . Предполагается, что функция f либо сильно выпукла, либо удовлетворяет условию Поляка–Лоясевича (условию PL) [8, 9] и допускает аддитивное представление (см. выше) с β -гельдеровыми компонентами f_j . Функции, удовлетворяющие условию Поляка–Лоясевича, будем для краткости называть PL-функциями.

Рассматриваемая постановка относится к семейству задач безградиентной стохастической оптимизации (оптимизации нулевого порядка), по которым имеется богатая литература, см. публикации [10–23] и ссылки в них. В этих работах аддитивная структура f не предполагалась. В [13] доказано, что для β -гельдеровой ($\beta \geq 2$) функции f , удовлетворяющей условию квадратичного роста, минимаксно-оптимальная ошибка оптимизации, как функция количества T последовательно выбранных пробных точек, имеет порядок $T^{-(\beta-1)/\beta}$ с точностью до неопределенного множителя, зависящего от размерности d . Дальнейшие исследования были посвящены зависимости минимаксной ошибки от d в предположении, что функция f является β -гельдеровой с $\beta \geq 2$ и удовлетворяет либо условию сильной выпуклости [14, 15, 18, 21, 23, 24], либо условию PL [21, 25]. В случае PL-функций изучалась минимизация без ограничений, в то время как сильно выпуклый случай анализировался как при наличии ограничений, так и без них. Достигнут значительный прогресс, хотя полностью (для всех $\beta \geq 2$) задача пока не решена. Получена минимаксная нижняя граница для класса β -гельдеровых и сильно выпуклых функций, которая имеет порядок $dT^{-(\beta-1)/\beta}$. Данный результат доказан в [15] для $\beta = 2$ и гауссовского шума и в [19] для всех $\beta \geq 2$ и более общего шума; кроме того, более общая нижняя граница представлена в [21]. С другой стороны, для $\beta = 2$ существует алгоритм, достигающий ошибки такого же порядка при общих условиях на шум (без предположения о независимости или нулевом среднем), см. [19]. Таким образом, для $\beta = 2$ известно, что минимаксная ошибка имеет порядок $d/\sqrt{T}$. Для $\beta > 2$ в литературе приведены различные зависимости верхних границ от d , определяемые геометрией условия β -гельдеровости. К примеру, для гильбертовых классов, определяемых через приближение многочленом Тейлора, наилучшая известная верхняя граница

имеет порядок $d^{2-1/\beta}T^{-(\beta-1)/\beta}$ в случае сильно выпуклых функций [23, 24]. С другой стороны, для гильдеровых классов, заданных условиями тензорного типа, можно достичь ошибки порядка $d^{2-2/\beta}T^{-(\beta-1)/\beta}$ как для сильно выпуклых функций, так и для PL-функций [21]. Наконец, также в предположении сильной выпуклости, в недавней работе [26] рассмотрен класс функций, у которых гессиян удовлетворяет условию Липшица (разновидность условия Гильдера для $\beta = 3$) и получена верхняя граница порядка $dT^{-2/3}$. Отметим, что нижняя граница из [19] с ошибкой $dT^{-(\beta-1)/\beta}$ справедлива для всех вышеперечисленных гильдеровых классов, поскольку она получена для аддитивных функций, принадлежащих всем этим классам. Таким образом, ошибка $dT^{-(\beta-1)/\beta}$ оказывается минимаксно-оптимальной не только для $\beta = 2$, но и для $\beta = 3$ при подходящем определении 3-гильдерового класса сильно выпуклых функций.

Основным результатом настоящей работы является верхняя граница $dT^{-(\beta-1)/\beta}$, установленная для ошибки оптимизации в безградиентной постановке для класса β -гильдеровых функций при наличии шума, описываемых аддитивной моделью и удовлетворяющих условию PL или условию сильной выпуклости. Вместе с нижней границей, доказанной в [19], это означает, что $dT^{-(\beta-1)/\beta}$ – минимаксно-оптимальная ошибка оптимизации в данной постановке для всех $\beta \geq 2$. Это заключение является довольно неожиданным, так как оно противоречит интуитивным представлениям на основе упомянутых выше классических результатов по непараметрическому оцениванию. Действительно, по крайней мере для $\beta \in \{2, 3\}$, при переходе от общей β -гильдеровой модели к аддитивной β -гильдеровой модели ошибка не улучшается, ни по T , ни по d . Улучшение может проявляться только в величине множителя, не зависящего от T и d . Подобное свойство можно объяснить сравнительной простотой оптимизационной постановки по сравнению с непараметрическим оцениванием функций в том смысле, что ее целью является оценка функционала от неизвестной функции f (ее минимума), а не самой функции f . Приведем другой, схожий факт из области безградиентной стохастической оптимизации. А именно, нет существенной разницы между сложностью минимизации сильно выпуклой функции с липшицевым градиентом (что соответствует рассмотренному выше случаю $\beta = 2$ с ошибкой порядка $d/\sqrt{T}$) и сложностью минимизации выпуклой функции без дополнительных свойств, для которой можно построить алгоритм, сходящийся с ошибкой в пределах от $d^{1.5}/\sqrt{T}$ до $d^{1.75}/\sqrt{T}$ (с точностью до логарифмического множителя) [27].

2. Постановка задачи

Пусть Θ – замкнутое выпуклое подмножество пространства $\mathbb{R}^d$. Рассмотрим задачу минимизации неизвестной функции $f : \mathbb{R}^d \rightarrow \mathbb{R}$ на множестве Θ по наблюдениям значений f на фоне шума в пробных точках, которые могут быть выбраны последовательно в зависимости от предыдущих наблюдений. А именно, предполагаем, что на шаге $t \in \{1, \dots, T\}$ наблюдаются два зашум-

ленных значения f в точках $\mathbf{z}_t, \mathbf{z}'_t \in \mathbb{R}^d$:

$$y_t = f(\mathbf{z}_t) + \xi_t, \quad y'_t = f(\mathbf{z}'_t) + \xi'_t,$$

где ξ_t, ξ'_t – скалярные шумы, а пробные точки $\mathbf{z}_t, \mathbf{z}'_t$ могут быть выбраны в зависимости от $\{\mathbf{z}_i, \mathbf{z}'_i, y_i, y'_i\}_{i=1}^{t-1}$ и от возможной рандомизации.

Далее везде предполагается, что f описывается аддитивной моделью

$$f(\mathbf{x}) = \sum_{j=1}^d f_j(x_j),$$

где x_j – координаты вектора $\mathbf{x} \in \mathbb{R}^d$, а f_j – неизвестные функции одной переменной.

Предполагается, что каждая из функций $f_j : \mathbb{R} \rightarrow \mathbb{R}$, $j = 1, \dots, d$, принадлежит классу β -гельдеровых функций с $\beta \geq 2$, который определяется следующим образом.

Определение 1. Для $\beta > 0$ и $L > 0$ обозначим через $\mathcal{F}_\beta(L)$ множество всех функций $f : \mathbb{R} \rightarrow \mathbb{R}$, которые $\ell = \lfloor \beta \rfloor$ раз дифференцируемы и удовлетворяют для всех $x, z \in \mathbb{R}$ условию

$$(1) \quad \left| f(z) - \sum_{m=0}^{\ell} \frac{1}{m!} f^{(m)}(x)(z-x)^m \right| \leq L|z-x|^\beta,$$

где $f^{(m)}$ – производная функции f порядка m , а $\lfloor \beta \rfloor$ – максимальное целое число, меньшее β . Элементы класса $\mathcal{F}_\beta(L)$ будем называть β -гельдеровыми функциями.

Если $\beta > 2$, то из $f_j \in \mathcal{F}_\beta(L)$ не следует, что $f_j \in \mathcal{F}_2(L)$, однако понадобится и последнее условие. Удобно использовать его в несколько иной форме, задаваемой следующим определением.

Определение 2. Функция $f : \mathbb{R} \rightarrow \mathbb{R}$ называется $\bar{L}$ -гладкой, если она дифференцируема на $\mathbb{R}$ и существует число $\bar{L} > 0$ такое, что для любых $x, x' \in \mathbb{R}$ справедливо

$$|f'(x) - f'(x')| \leq \bar{L}|x - x'|.$$

Обозначим через $\mathcal{F}'_2(\bar{L})$ класс всех $\bar{L}$ -гладких функций.

Также предположим, что f является либо α -сильно выпуклой функцией, либо α -PL функцией (т.е. удовлетворяет условию PL с заданным α); см. два определения ниже.

Определение 3. Пусть $\alpha > 0$. Функция $f : \mathbb{R}^d \rightarrow \mathbb{R}$ называется α -PL функцией, если она дифференцируема на $\mathbb{R}^d$ и

$$2\alpha \left(f(\mathbf{x}) - \min_{\mathbf{z} \in \mathbb{R}^d} f(\mathbf{z}) \right) \leq \|\nabla f(\mathbf{x})\|^2 \quad \text{для всех } \mathbf{x} \in \mathbb{R}^d,$$

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

Функции, удовлетворяющие условию PL, не обязательно являются выпуклыми. Условие PL является полезным инструментом в задачах оптимизации, поскольку, как показано Б.Т. Поляком [8], оно обеспечивает линейную сходимость алгоритма градиентного спуска без свойства выпуклости. Более подробное обсуждение условия PL можно найти в [28].

Определение 4. Пусть $\alpha > 0$. Функция $f : \mathbb{R}^d \rightarrow \mathbb{R}$ называется α -сильно выпуклой, если она дифференцируема на $\mathbb{R}^d$ и

$$f(\mathbf{x}) - f(\mathbf{x}') \leq \langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{x}' \rangle - \frac{\alpha}{2} \|\mathbf{x} - \mathbf{x}'\|^2 \quad \text{для всех } \mathbf{x}, \mathbf{x}' \in \mathbb{R}^d.$$

Для минимизации f применим вариант метода проекции градиента (см. алгоритм 1). Пусть $\{\eta_t\}_{t=1}^T$ – последовательность положительных чисел, $\{\mathbf{g}_t\}_{t=1}^T$ – последовательность случайных векторов и $\mathbf{x}_1 \in \mathbb{R}^d$ – произвольный фиксированный вектор. Определим векторы $\mathbf{x}_t$, $t = 2, \dots, T$ с помощью рекуррентного соотношения

$$(2) \quad \mathbf{x}_{t+1} = \text{Proj}_\Theta(\mathbf{x}_t - \eta_t \mathbf{g}_t),$$

где $\text{Proj}_\Theta(\cdot)$ обозначает евклидову проекцию на Θ .

Алгоритм 1

Вход: Множество Θ , функция $K : [-1, 1] \rightarrow \mathbb{R}$, размер шага $\eta_t > 0$ и параметр возмущения $h_t > 0$ для $t = 1, \dots, T$.

Инициализация: Сгенерировать векторы $\mathbf{r}_t = (r_{t,1}, \dots, r_{t,d}) \in \mathbb{R}^d$, $t = 1, \dots, T$, с независимыми равномерно распределенными на $[-1, 1]$ компонентами и выбрать $\mathbf{x}_1 \in \Theta$.

Для $t = 1, \dots, T$ выполнить:

- Получить наблюдения $y_t = f(\mathbf{x}_t + h_t \mathbf{r}_t) + \xi_t$ и $y'_t = f(\mathbf{x}_t - h_t \mathbf{r}_t) + \xi'_t$
- **Для $j = 1, \dots, d$ выполнить:**
 - $g_{t,j} := \frac{d}{2h_t}(y_t - y'_t)K(r_{t,j})$
- $\mathbf{g}_t := (g_{t,1}, \dots, g_{t,d})$ // оценка градиента
- $\mathbf{x}_{t+1} = \text{Proj}_\Theta(\mathbf{x}_t - \eta_t \mathbf{g}_t)$ // обновление

В данной работе оценка градиента $\mathbf{g}_t = (g_{t,1}, \dots, g_{t,d}) \in \mathbb{R}^d$ на шаге $t \in \{1, \dots, T\}$ алгоритма определяется следующим образом. Для заданного $\beta \geq 2$ и $\ell = \lfloor \beta \rfloor$, пусть $K : [-1, 1] \rightarrow \mathbb{R}$ – функция такая, что

$$(3) \quad \int u K(u) du = 1, \quad \int u^j K(u) du = 0, \quad j = 0, 2, 3, \dots, \ell,$$

$$\kappa_\beta \equiv \int |u|^\beta |K(u)| du < \infty.$$

Предполагается, что величина $\kappa := \int K^2(r)dr$ конечна. Несложно построить функции K , удовлетворяющие этим условиям. В частности, можно использовать полиномы Лежандра, например см. [13, 18, 29].

На каждом шаге t алгоритма генерируем случайный вектор $\mathbf{r}_t = (r_{t,1}, \dots, r_{t,d}) \in \mathbb{R}^d$, где компоненты $r_{t,j}$ независимы и распределены равномерно на $[-1, 1]$. Взяв $h_t > 0$, произведем два наблюдения

$$y_t = f(\mathbf{x}_t + h_t \mathbf{r}_t) + \xi_t, \quad y'_t = f(\mathbf{x}_t - h_t \mathbf{r}_t) + \xi'_t,$$

и определим величины

$$(4) \quad g_{t,j} = \frac{1}{2h_t}(y_t - y'_t)K(r_{t,j}),$$

где $j \in \{1, \dots, d\}$. Будем рассматривать оценку градиента вида $\mathbf{g}_t = (g_{t,1}, \dots, g_{t,d})$.

Заметим, что применяя другие оценки градиента можно получить результаты, аналогичные представленным ниже, в частности, используя оценки, основанные на конечно-разностной аппроксимации с учетом гладкости более высокого порядка. В отличие от (4), такие конечно-разностные схемы более высокого порядка имеют сложный вид и для них требуется много пробных точек на каждом шаге алгоритма.

Допустим, что переменные шума ξ_t, ξ'_t и рандомизирующие переменные $r_{t,j}$ удовлетворяют следующим условиям.

Предположение 1. Существует величина $\sigma^2 > 0$ такая, что для всех $t \in \{1, \dots, T\}$ справедливо:

- (i) случайные величины $r_{t,j} \sim U[-1, 1]$, $j = 1, \dots, d$, независимы от $\mathbf{x}_t$ и условно независимы от ξ_t, ξ'_t при заданном векторе $\mathbf{x}_t$;
- (ii) $\mathbb{E}[\xi_t^2] \leq \sigma^2$, $\mathbb{E}[(\xi'_t)^2] \leq \sigma^2$.

Пункт (i) предположения 1 можно рассматривать не как ограничение, а как часть определения алгоритма, связанную с выбором рандомизирующих переменных $r_{t,j}$. Эти переменные естественным образом выбираются независимо от всех других источников случайности. Для доказательств оказывается достаточным даже более слабое свойство, которое и приведено здесь в виде предположения, чтобы можно было сослаться на него в дальнейшем. Отметим также, что в предположении 1 не требуется, чтобы шумы ξ_t, ξ'_t имели нулевое среднее. Более того, они могут быть неслучайными, и не предполагается независимость этих шумов на разных шагах алгоритма. Тот факт, что сходимость безградиентных алгоритмов может быть достигнута при общих условиях подобного типа на шумы, не требующих независимости и нулевых средних, восходит к работам [30, 31].

3. Основные результаты

В этом разделе представлены верхние границы для ошибки оптимизации алгоритма, определенного в разделе 2. Предположим сначала, что функция f

описывается аддитивной моделью и является α -PL функцией. Заметим, что при наложении этого условия на сумму f все компоненты f_j являются PL-функциями. В классе PL-функций рассмотрим задачу минимизации без ограничений:

$$f^* = \min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}).$$

Обозначим через $[n]$ множество натуральных чисел, не превышающих заданного натурального числа n .

Теорема 1. Пусть $\alpha > 0$, $\beta \geq 2$, $\bar{L}, L > 0$, и функции $f_j : \mathbb{R} \rightarrow \mathbb{R}$, $j \in [d]$, таковы, что $f_j \in \mathcal{F}'_2(\bar{L}) \cap \mathcal{F}_\beta(L)$. Допустим, что функция $f : \mathbb{R}^d \rightarrow \mathbb{R}$ представима в виде $f(\mathbf{x}) = \sum_{j=1}^d f_j(x_j)$ и является α -PL функцией. Пусть выполнено предположение 1 и пусть $\{\mathbf{x}_t\}_{t=1}^T$ – реализация алгоритма 1 с $\Theta = \mathbb{R}^d$ и

$$\eta_t = \min \left(\frac{4}{\alpha t}, \frac{1}{18\bar{L}d\kappa} \right),$$

$$h_t = \left(\frac{3\bar{L}}{\alpha} \frac{\kappa\sigma^2}{2L^2\kappa_\beta^2} \right)^{\frac{1}{2\beta}} \begin{cases} t^{-\frac{1}{2\beta}} & \text{при } \eta_t = \frac{4}{\alpha t}, \\ T^{-\frac{1}{2\beta}} & \text{при } \eta_t = \frac{1}{18\bar{L}d\kappa}. \end{cases}$$

Тогда

$$\mathbf{E}[f(\mathbf{x}_T) - f^*] \leq \frac{A_0}{T}(f(\mathbf{x}_1) - f^*) + A \frac{d}{\alpha} \left(\left(\frac{1}{\alpha T} \right)^{\frac{\beta-1}{\beta}} + \frac{d}{T} \left(\frac{1}{\alpha T} \right)^{\frac{1}{\beta}} \right),$$

где $A_0 = \max(1, 144\bar{L}d\kappa/\alpha)$ и $A > 0$ зависит только от L , $\bar{L}$, β и σ^2 .

Следствие 1. В условиях теоремы 1, если $T \geq \alpha^{\beta/2-1} d^{\beta/2}$, то

$$\mathbf{E}[f(\mathbf{x}_T) - f^*] \leq \frac{A_0}{T}(f(\mathbf{x}_1) - f^*) + A \frac{d}{\alpha} \left(\frac{1}{\alpha T} \right)^{\frac{\beta-1}{\beta}},$$

где $A > 0$ зависит только от L , $\bar{L}$, β и σ^2 .

Заметим, что при $2 \leq \beta \leq 3$ условие $T \geq Cd^{\beta/2}$ с некоторой константой $C > 0$ (см. следствие 1) не является ограничительным, поскольку оно слабее, чем задание диапазона для T , при котором граница следствия 1 представляет интерес. Действительно, если $T \leq Cd^{\beta/2}$ и $2 \leq \beta \leq 3$, то данная граница больше некоторой константы, не зависящей от T и d .

Исследуем теперь ошибку оптимизации для алгоритма, определенного в разделе 2, в предположении α -сильной выпуклости целевой функции. В отличие от случая PL-функций, рассмотрим задачу минимизации с ограничениями:

$$\min_{\mathbf{x} \in \Theta} f(\mathbf{x}),$$

где Θ – компактное выпуклое подмножество $\mathbb{R}^d$.

Теорема 2. Пусть $\alpha > 0$, $\beta \geq 2$, $\bar{L}, L > 0$, пусть Θ – компактное выпуклое подмножество $\mathbb{R}^d$ и функции $f_j : \mathbb{R} \rightarrow \mathbb{R}$, $j \in [d]$, таковы, что $f_j \in \mathcal{F}'_2(\bar{L}) \cap \mathcal{F}_\beta(L)$. Допустим, что функция $f : \mathbb{R}^d \rightarrow \mathbb{R}$ представима в виде $f(\mathbf{x}) = \sum_{j=1}^d f_j(x_j)$, является α -сильно выпуклой функцией и удовлетворяет условию $\max_{\mathbf{x} \in \Theta} \|\nabla f(\mathbf{x})\| \leq G$. Пусть выполнено предположение 1 и пусть $\{\mathbf{x}_t\}_{t=1}^T$ – реализация алгоритма 1 с

$$\eta_t = \frac{4}{\alpha(t+1)}, \quad h_t = \left(\frac{3}{2t} \frac{\kappa \sigma^2}{\kappa_\beta^2 L^2} \right)^{\frac{1}{2\beta}}.$$

Рассмотрим взвешенную оценку

$$\bar{\mathbf{x}}_T = \frac{2}{T(T+1)} \sum_{t=1}^T t \mathbf{x}_t.$$

Тогда для любого $\mathbf{x} \in \Theta$ имеем

$$\mathbf{E}[f(\bar{\mathbf{x}}_T) - f(\mathbf{x})] \leq \frac{36G^2 d \kappa}{\alpha T} + A \frac{d}{\alpha} \left(1 + dT^{-\frac{2}{\beta}}\right) T^{-\frac{\beta-1}{\beta}},$$

где $A > 0$ зависит только от L , $\bar{L}$, β и σ^2 .

Следствие 2. В условиях теоремы 2, если $T \geq d^{\beta/2}$, то

$$\mathbf{E}\left[f(\bar{\mathbf{x}}_T) - \min_{\mathbf{x} \in \Theta} f(\mathbf{x})\right] \leq A \frac{d}{\alpha} T^{-\frac{\beta-1}{\beta}},$$

где $A > 0$ не зависит от T, d и α .

Так же, как и в связи со следствием 1, можно отметить следующий факт: при $2 \leq \beta \leq 3$ условие $T \geq Cd^{\beta/2}$ с константой $C > 0$ в следствии 2 не является ограничительным, а указывает на значимый диапазон величины T , поскольку при $T \leq Cd^{\beta/2}$ граница следствия 2 больше константы.

Замечание 1. Ввиду сильной выпуклости теорема 2 и следствие 2 немедленно приводят к соответствующим границам для ошибки оценивания $\mathbf{E}[\|\bar{\mathbf{x}}_T - \mathbf{x}^*\|^2]$, где $\mathbf{x}^*$ – точка минимума f на Θ , если в ней достигается глобальный минимум. Таким образом, при условиях следствия 2, если $\nabla f(\mathbf{x}^*) = 0$, то

$$\mathbf{E}[\|\bar{\mathbf{x}}_T - \mathbf{x}^*\|^2] \leq 2A \frac{d}{\alpha^2} T^{-\frac{\beta-1}{\beta}},$$

где $A > 0$ – константа из следствия 2.

Согласно следствию 2 и нижним границам, доказанным в [19, 21], при неограничительных условиях на параметры задачи, ошибка $\frac{d}{\alpha} T^{-\frac{\beta-1}{\beta}}$ является минимаксно-оптимальной в классе аддитивных функций f , удовлетворяющих условиям теоремы 2. Искомая нижняя граница не указана в явном виде

в [19, 21], но непосредственно следует из представленных там доказательств, так как нижние границы в [19, 21] получены на аддитивных функциях. Для полноты изложения приведем здесь утверждение о нижней границе для аддитивных функций, основанное на результатах [21].

Рассмотрим все стратегии выбора пробных точек в виде

$$\mathbf{z}_t = \Phi_t((\mathbf{z}_i, y_i)_{i=1}^{t-1}, (\mathbf{z}'_i, y'_i)_{i=1}^{t-1}, \tau_t) \text{ и } \mathbf{z}'_t = \Phi'_t((\mathbf{z}_i, y_i)_{i=1}^{t-1}, (\mathbf{z}'_i, y'_i)_{i=1}^{t-1}, \tau_t) \text{ при } t \geq 2,$$

где Φ_t и Φ'_t – измеримые функции, $\mathbf{z}_1, \mathbf{z}'_1 \in \mathbb{R}^d$ – любые случайные величины, а $\{\tau_t\}$ – последовательность случайных величин со значениями в измеримом пространстве $(\mathcal{Z}, \mathcal{U})$ такая, что τ_t не зависит от $((\mathbf{z}_i, y_i)_{i=1}^{t-1}, (\mathbf{z}'_i, y'_i)_{i=1}^{t-1})$. Обозначим через Π_T множество всех таких стратегий выбора пробных точек вплоть до шага $t = T$. Класс Π_T содержит последовательную стратегию алгоритма из раздела 2 с оценкой градиента (4). В этом случае $\tau_t = \mathbf{r}_t$, $\mathbf{z}_t = \mathbf{x}_t + h_t \mathbf{r}_t$ и $\mathbf{z}'_t = \mathbf{x}_t - h_t \mathbf{r}_t$.

Используемая здесь нижняя граница из [21] доказана при следующем предположении на шумы (ξ, ξ') . Пусть $H^2(\cdot, \cdot)$ – квадрат расстояния Хеллингера, который определяется для двух вероятностных мер $\mathbf{P}, \mathbf{P}'$ на измеримом пространстве $(\Omega, \mathcal{A})$ по формуле

$$H^2(\mathbf{P}, \mathbf{P}') = \int (\sqrt{d\mathbf{P}} - \sqrt{d\mathbf{P}'})^2.$$

Предположение 2. Для каждого шага $t \geq 1$ справедливо следующее:

- функция распределения $F_t : \mathbb{R}^2 \rightarrow \mathbb{R}$ случайной величины (ξ_t, ξ'_t) такова, что

$$(5) \quad H^2(P_{F_t(\cdot, \cdot)}, P_{F_t(\cdot+v, \cdot+w)}) \leq I_0 \max(v^2, w^2), \quad |v|, |w| \leq v_0,$$

для некоторых $0 < I_0 < \infty$, $0 < v_0 \leq \infty$. Здесь $P_{F(\cdot, \cdot)}$ обозначает вероятностную меру, соответствующую функции распределения $F(\cdot, \cdot)$;

- случайная величина (ξ_t, ξ'_t) не зависит от $((\mathbf{z}_i, y_i)_{i=1}^{t-1}, (\mathbf{z}'_i, y'_i)_{i=1}^{t-1}, \tau_t)$.

Пусть $\Theta = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| \leq 1\}$. При фиксированных $\alpha, L, \bar{L} > 0$, $G > \alpha$, $\beta \geq 2$ обозначим через $\mathcal{F}$ множество всех функций f , удовлетворяющих условиям теоремы 2 и достигающих своего минимума на $\mathbb{R}^d$ в множестве Θ .

Теорема 3. Пусть $\Theta = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| \leq 1\}$ и верно предположение 2. Допустим, что $\alpha > T^{-1/2+1/\beta}$ и $T \geq d^\beta$. Тогда для любой оценки $\tilde{\mathbf{x}}_T$ по наблюдениям $((\mathbf{z}_t, y_t), (\mathbf{z}'_t, y'_t), t = 1, \dots, T)$, где $((\mathbf{z}_t, \mathbf{z}'_t), t = 1, \dots, T)$ получены с помощью произвольной стратегии из класса Π_T , имеем

$$(6) \quad \sup_{f \in \mathcal{F}} \mathbf{E}[f(\tilde{\mathbf{x}}_T) - \min_{\mathbf{x} \in \Theta} f(\mathbf{x})] \geq C \frac{d}{\alpha} T^{-\frac{\beta-1}{\beta}}$$

с константой $C > 0$, не зависящей от T, d и α .

Теорема 3 непосредственно следует из доказательства в [21, теорема 22], так как семейство используемых там функций принадлежит классу $\mathcal{F}$. Нижняя граница из [21, теорема 22] имеет вид

$$C \min \left(\max(\alpha, T^{-1/2+1/\beta}), \frac{d}{\sqrt{T}}, \frac{d}{\alpha} T^{-\frac{\beta-1}{\beta}} \right)$$

и сводится к $C \frac{d}{\alpha} T^{-\frac{\beta-1}{\beta}}$ при предположениях на T, d и α , использованных в теореме 3.

Замечание 2. Поскольку сильная выпуклость и свойство RL справедливы для каждой аддитивной компоненты функции f , возможной альтернативой была бы покоординатная процедура (минимизация каждой компоненты f_j отдельно). Однако подобный подход приводит к худшему результату. Действительно, в этом случае на каждом шаге необходимо параллельно выбирать $2d$ пробных точек (по две на каждую компоненту) и, таким образом, можно сделать только $\sim T/d$ шагов при общем количестве пробных точек, равном T . В итоге, применяя теорему 1 или 2 в одномерном случае, для каждой компоненты получим ошибку порядка $(T/d)^{-(\beta-1)/\beta}$. Эта ошибка не может быть улучшена, что следует из одномерного случая теоремы 3. Суммируя по d компонентам, получаем общую ошибку порядка $d(T/d)^{-(\beta-1)/\beta} = d^{2-1/\beta} T^{-(\beta-1)/\beta}$, т.е. ошибка будет зависеть от d неоптимальным образом.

4. Доказательства

Начнем с доказательства вспомогательных лемм.

Лемма 1. Пусть $f : \mathbb{R}^d \rightarrow \mathbb{R}$ – дифференцируемая функция такая, что $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}')\| \leq \bar{L} \|\mathbf{x} - \mathbf{x}'\|$ для всех $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$, где $\bar{L} > 0$. Пусть выполнено предположение 1 и пусть $\{\mathbf{x}_t\}_{t=1}^T$ – реализация алгоритма 1 с $\Theta = \mathbb{R}^d$. Тогда для всех $t \in [T]$ справедливо неравенство

$$(7) \quad \mathbf{E}[f(\mathbf{x}_{t+1})|\mathbf{x}_t] \leq f(\mathbf{x}_t) - \frac{\eta_t}{2} \|\nabla f(\mathbf{x}_t)\|^2 + \\ + \frac{\eta_t}{2} \|\mathbf{E}[\mathbf{g}_t|\mathbf{x}_t] - \nabla f(\mathbf{x}_t)\|^2 + \frac{\bar{L}\eta_t^2}{2} \mathbf{E}[\|\mathbf{g}_t\|^2|\mathbf{x}_t].$$

Доказательство. Из условия на f следует, что

$$f(\mathbf{x}_{t+1}) = f(\mathbf{x}_t - \eta_t \mathbf{g}_t) \leq f(\mathbf{x}_t) - \eta_t \langle \nabla f(\mathbf{x}_t), \mathbf{g}_t \rangle + \frac{\bar{L}\eta_t^2}{2} \|\mathbf{g}_t\|^2.$$

Прибавляя и вычитая $\eta_t \|\nabla f(\mathbf{x}_t)\|^2$, получаем

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \eta_t \|\nabla f(\mathbf{x}_t)\|^2 - \eta_t \langle \nabla f(\mathbf{x}_t), \mathbf{g}_t - \nabla f(\mathbf{x}_t) \rangle + \frac{\bar{L}\eta_t^2}{2} \|\mathbf{g}_t\|^2.$$

Взяв условное математическое ожидание, будем иметь

$$\begin{aligned} \mathbf{E}[f(\mathbf{x}_{t+1})|\mathbf{x}_t] &\leq f(\mathbf{x}_t) - \eta_t \|\nabla f(\mathbf{x}_t)\|^2 - \eta_t \langle \nabla f(\mathbf{x}_t), \mathbf{E}[\mathbf{g}_t|\mathbf{x}_t] - \nabla f(\mathbf{x}_t) \rangle + \\ &\quad + \frac{\bar{L}\eta_t^2}{2} \mathbf{E}[\|\mathbf{g}_t\|^2|\mathbf{x}_t] \leq \\ &\leq f(\mathbf{x}_t) - \eta_t \|\nabla f(\mathbf{x}_t)\|^2 + \eta_t \|\nabla f(\mathbf{x}_t)\| \mathbf{E}[\|\mathbf{g}_t|\mathbf{x}_t] - \nabla f(\mathbf{x}_t)\| + \\ &\quad + \frac{\bar{L}\eta_t^2}{2} \mathbf{E}[\|\mathbf{g}_t\|^2|\mathbf{x}_t]. \end{aligned}$$

Используя неравенство $2ab \leq a^2 + b^2$, $\forall a, b \in \mathbb{R}$, приходим к требуемому результату. Лемма доказана.

Лемма 2. Пусть функции $f_j : \mathbb{R} \rightarrow \mathbb{R}$, $j \in [d]$, таковы, что $f_j \in \mathcal{F}_\beta(L)$, где $\beta \geq 2$ и $L > 0$. Допустим, что функция $f : \mathbb{R}^d \rightarrow \mathbb{R}$ задается аддитивной моделью $f(\mathbf{x}) = \sum_{j=1}^d f_j(x_j)$ и верно предположение 1(i). Тогда для всех $t \in [T]$ справедливо неравенство

$$\|\mathbf{E}[\mathbf{g}_t|\mathbf{x}_t] - \nabla f(\mathbf{x}_t)\| \leq \kappa_\beta L \sqrt{d} h_t^{\beta-1}.$$

Доказательство. Используя предположение 1(i) для любых $j \in [d]$ и $t \in [T]$ имеем $\mathbf{E}(K(r_{t,j})) = 0$, $\mathbf{E}[\xi_t K(r_{t,j})|\mathbf{x}_t] = 0$ и $\mathbf{E}[\xi_t' K(r_{t,j})|\mathbf{x}_t] = 0$. Таким образом,

$$\begin{aligned} \mathbf{E}[\mathbf{g}_{t,j}|\mathbf{x}_t] &= \frac{1}{2h_t} \mathbf{E}[(f_j(x_{t,j} + h_t r_{t,j}) - f_j(x_{t,j} - h_t r_{t,j}))K(r_{t,j})|\mathbf{x}_t] + \\ &\quad + \frac{1}{2h_t} \sum_{m \neq j} \mathbf{E}[(f_m(x_{t,m} + h_t r_{t,m}) - f_m(x_{t,m} - h_t r_{t,m}))K(r_{t,j})|\mathbf{x}_t] = \\ &= \frac{1}{2h_t} \mathbf{E}[(f_j(x_{t,j} + h_t r_{t,j}) - f_j(x_{t,j} - h_t r_{t,j}))K(r_{t,j})|\mathbf{x}_t], \end{aligned}$$

где использовано равенство $\mathbf{E}(K(r_{t,j})) = 0$ и независимость $r_{t,j}$ от $r_{t,m}$ при $m \neq j$, а также от $\mathbf{x}_t$. В силу разложения Тейлора,

$$\begin{aligned} \frac{1}{2h_t} (f_j(x_{t,j} + h_t r_{t,j}) - f_j(x_{t,j} - h_t r_{t,j})) &= f_j'(x_{t,j}) r_{t,j} + \\ &\quad + \frac{1}{h_t} \sum_{1 \leq m \leq \ell, m \text{ нечетно}} \frac{h_t^m}{m!} f_j^{(m)}(x_{t,j}) r_{t,j}^m + \frac{R(h_t r_{t,j}) - R(-h_t r_{t,j})}{2h_t}, \end{aligned}$$

где $|R(-h_t r_{t,j})|, |R(h_t r_{t,j})| \leq L |r_{t,j}|^\beta h_t^\beta$. Умножив обе части этого неравенства на $K(r_{t,j})$ и взяв условное математическое ожидание, получаем

$$|\mathbf{E}[\mathbf{g}_{t,j}|\mathbf{x}_t] - f_j'(x_{t,j})| \leq L \mathbf{E}[|r_{t,j}|^\beta K(r_{t,j})] h_t^{\beta-1} = \kappa_\beta L h_t^{\beta-1}.$$

Лемма следует из этого неравенства в сочетании с оценкой

$$\|\mathbf{E}[\mathbf{g}_t|\mathbf{x}_t] - \nabla f(\mathbf{x}_t)\| \leq \sqrt{d} \max_{j=1, \dots, d} |\mathbf{E}[\mathbf{g}_{t,j}|\mathbf{x}_t] - f_j'(x_{t,j})|.$$

Лемма 3. Пусть функции $f_j : \mathbb{R} \rightarrow \mathbb{R}$, $j \in [d]$, таковы, что $f_j \in \mathcal{F}'_2(\bar{L})$, где $\bar{L} > 0$. Допустим, что функция $f : \mathbb{R}^d \rightarrow \mathbb{R}$ задается аддитивной моделью $f(\mathbf{x}) = \sum_{j=1}^d f_j(x_j)$ и верно предположение 1. Тогда для всех $t \in [T]$ справедливо неравенство

$$\mathbf{E} [\|\mathbf{g}_t\|^2 | \mathbf{x}_t] \leq \frac{3}{2} \kappa d \left(\frac{3}{4} (d\bar{L}^2 h_t^2 + 8 \|\nabla f(\mathbf{x}_t)\|^2) + \frac{\sigma^2}{h_t^2} \right).$$

Доказательство. Для $i \in [d]$ определим

$$G_i = f_i(x_{t,i} + h_t r_{t,i}) - f_i(x_{t,i} - h_t r_{t,i}).$$

Тогда

$$\mathbf{E} [g_{t,j}^2 | \mathbf{x}_t] = \frac{1}{4h_t^2} \mathbf{E} \left[\left(\sum_{i=1}^d G_i + \xi_t - \xi'_t \right)^2 K^2(r_{t,j}) | \mathbf{x}_t \right].$$

Заметим, что $r_{t,i}$ и $-r_{t,i}$ распределены одинаково. Следовательно, $\mathbf{E}[G_i | \mathbf{x}_t] = 0$ и можно записать

$$\begin{aligned} \mathbf{E} [g_{t,j}^2 | \mathbf{x}_t] &\leq \frac{3}{4h_t^2} \mathbf{E} \left[\left(\left(\sum_{i=1}^d G_i \right)^2 + \xi_t^2 + (\xi'_t)^2 \right) K^2(r_{t,j}) | \mathbf{x}_t \right] \leq \\ &\leq \frac{3}{4h_t^2} \mathbf{E} \left[\sum_{i=1}^d G_i^2 K^2(r_{t,j}) + \sum_{i,k=1, i \neq k}^d G_i G_k K^2(r_{t,j}) | \mathbf{x}_t \right] + \frac{3\sigma^2 \kappa}{2h_t^2} = \\ &= \frac{3}{4h_t^2} \mathbf{E} \left[\sum_{i=1}^d G_i^2 K^2(r_{t,j}) | \mathbf{x}_t \right] + \frac{3\sigma^2 \kappa}{2h_t^2}. \end{aligned}$$

Так как $f_i \in \mathcal{F}'_2(\bar{L})$, для всех $i \in [d]$ имеем

$$\begin{aligned} G_i^2 &= \left((f_i(x_{t,i} + h_t r_{t,i}) - f(x_{t,i}) - f'_i(x_{t,i}) h_t r_{t,i}) - \right. \\ &\quad \left. - (f_i(x_{t,i} - h_t r_{t,i}) - f(x_{t,i}) + f'_i(x_{t,i}) h_t r_{t,i}) + 2f'_i(x_{t,i}) h_t r_{t,i} \right)^2 \leq \\ &\leq 3 \left((f_i(x_{t,i} + h_t r_{t,i}) - f(x_{t,i}) - f'_i(x_{t,i}) h_t r_{t,i})^2 + \right. \\ &\quad \left. + (f_i(x_{t,i} - h_t r_{t,i}) - f(x_{t,i}) + f'_i(x_{t,i}) h_t r_{t,i})^2 + 4(f'_i(x_{t,i}) h_t r_{t,i})^2 \right) \leq \\ &\leq 3 \left(\frac{\bar{L}^2}{2} h_t^4 + 4(f'_i(x_{t,i}))^2 h_t^2 \right). \end{aligned}$$

Таким образом,

$$\mathbf{E} [g_{t,j}^2 | \mathbf{x}_t] \leq \frac{9}{4} \kappa \left(\frac{d\bar{L}^2}{2} h_t^2 + 4 \sum_{i=1}^d (f'_i(x_{t,i}))^2 \right) + \frac{3\sigma^2 \kappa}{2h_t^2},$$

что завершает доказательство леммы.

Лемма 4. Пусть функции $f_j : \mathbb{R} \rightarrow \mathbb{R}$, $j \in [d]$, таковы, что $f_j \in \mathcal{F}'_2(\bar{L}) \cap \mathcal{F}_\beta(L)$, где $\beta \geq 2$ и $\bar{L}, L > 0$. Допустим, что функция $f : \mathbb{R}^d \rightarrow \mathbb{R}$ задается аддитивной моделью $f(\mathbf{x}) = \sum_{j=1}^d f_j(x_j)$. Пусть выполнено предположение 1 и пусть $\{\mathbf{x}_t\}_{t=1}^T$ – реализация алгоритма 1 с $\Theta = \mathbb{R}^d$ и оценками градиента (4). Тогда для всех $t \in [T]$ справедливо неравенство

$$\begin{aligned} \mathbf{E}[f(\mathbf{x}_{t+1})|\mathbf{x}_t] &\leq f(\mathbf{x}_t) - \frac{\eta_t}{2} (1 - 9\bar{L}d\kappa\eta_t) \|\nabla f(\mathbf{x}_t)\|^2 + \\ &+ \frac{\eta_t}{2} d \left(L\kappa_\beta h_t^{\beta-1} \right)^2 + \frac{3\bar{L}\eta_t^2}{4} \kappa d \left(\frac{\sigma^2}{h_t^2} + \frac{3\bar{L}^2 d}{4} h_t^2 \right). \end{aligned}$$

Доказательство. Требуемый результат следует из лемм 1, 2 и 3.

Лемма 5. Пусть функция f α -сильно выпукла с $\alpha > 0$. Предположим, что $\{\mathbf{x}_t\}_{t=1}^T$ – реализация алгоритма 1. Тогда для всех $t \in [T]$ и $\mathbf{x} \in \Theta$ справедливо неравенство

$$\begin{aligned} f(\mathbf{x}_t) - f(\mathbf{x}) &\leq (2\eta_t)^{-1} (\|\mathbf{x}_t - \mathbf{x}\|^2 - \mathbf{E}[\|\mathbf{x}_{t+1} - \mathbf{x}\|^2|\mathbf{x}_t]) + \\ &+ \frac{1}{\alpha} \|\nabla f(\mathbf{x}_t) - \mathbf{E}[\mathbf{g}_t|\mathbf{x}_t]\|^2 + \frac{\eta_t}{2} \mathbf{E}[\|\mathbf{g}_t\|^2|\mathbf{x}_t] - \frac{\alpha}{4} \|\mathbf{x}_t - \mathbf{x}\|^2. \end{aligned}$$

Доказательство. В силу определения евклидовой проекции, для фиксированного вектора $\mathbf{x} \in \Theta$ имеем $\|\mathbf{x}_{t+1} - \mathbf{x}\|^2 \leq \|\mathbf{x}_t - \eta_t \mathbf{g}_t - \mathbf{x}\|^2$. Это неравенство можно записать в эквивалентном виде

$$\langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{x} \rangle \leq (2\eta_t)^{-1} (\|\mathbf{x}_t - \mathbf{x}\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}\|^2) + \frac{\eta_t}{2} \|\mathbf{g}_t\|^2.$$

Пусть $a_t = \|\mathbf{x}_t - \mathbf{x}\|^2$. Так как функция f α -сильно выпукла,

$$\begin{aligned} f(\mathbf{x}_t) - f(\mathbf{x}) &\leq \langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x} \rangle - \frac{\alpha}{2} a_t = \\ &= \langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{x} \rangle + \langle \nabla f(\mathbf{x}_t) - \mathbf{g}_t, \mathbf{x}_t - \mathbf{x} \rangle - \frac{\alpha}{2} a_t \leq \\ &\leq (2\eta_t)^{-1} (a_t - a_{t+1}) + \langle \nabla f(\mathbf{x}_t) - \mathbf{g}_t, \mathbf{x}_t - \mathbf{x} \rangle + \frac{\eta_t}{2} \|\mathbf{g}_t\|^2 - \frac{\alpha}{2} a_t. \end{aligned}$$

Взяв условное математическое ожидание при заданном $\mathbf{x}_t$ и воспользовавшись неравенством $ab \leq a^2/\lambda + \lambda b^2/4$, справедливым для всех $a, b \in \mathbb{R}$ и $\lambda > 0$, получим

$$\begin{aligned} f(\mathbf{x}_t) - f(\mathbf{x}) &\leq (2\eta_t)^{-1} (a_t - \mathbf{E}[a_{t+1}|\mathbf{x}_t]) + \langle \nabla f(\mathbf{x}_t) - \mathbf{E}[\mathbf{g}_t|\mathbf{x}_t], \mathbf{x}_t - \mathbf{x} \rangle + \\ &+ \frac{\eta_t}{2} \mathbf{E}[\|\mathbf{g}_t\|^2|\mathbf{x}_t] - \frac{\alpha}{2} a_t \leq \\ &\leq (2\eta_t)^{-1} (a_t - \mathbf{E}[a_{t+1}|\mathbf{x}_t]) + \|\nabla f(\mathbf{x}_t) - \mathbf{E}[\mathbf{g}_t|\mathbf{x}_t]\| \|\mathbf{x}_t - \mathbf{x}\| + \\ &+ \frac{\eta_t}{2} \mathbf{E}[\|\mathbf{g}_t\|^2|\mathbf{x}_t] - \frac{\alpha}{2} a_t \leq \\ &\leq (2\eta_t)^{-1} (a_t - \mathbf{E}[a_{t+1}|\mathbf{x}_t]) + \frac{1}{\alpha} \|\nabla f(\mathbf{x}_t) - \mathbf{E}[\mathbf{g}_t|\mathbf{x}_t]\|^2 + \\ &+ \frac{\eta_t}{2} \mathbf{E}[\|\mathbf{g}_t\|^2|\mathbf{x}_t] - \frac{\alpha}{4} a_t. \end{aligned}$$

Доказательство теоремы 1. Поскольку $\eta_t \leq 1/(18\bar{L}d\kappa)$, из леммы 4 следует, что

$$\begin{aligned} \mathbf{E}[f(\mathbf{x}_{t+1})|\mathbf{x}_t] &\leq f(\mathbf{x}_t) - \frac{\eta_t}{4} \|\nabla f(\mathbf{x}_t)\|^2 + \\ &+ \frac{\eta_t}{2} d \left(L\kappa_\beta h_t^{\beta-1} \right)^2 + \frac{3\bar{L}\eta_t^2}{4} \kappa d \left(\frac{\sigma^2}{h_t^2} + \frac{3\bar{L}^2 d}{4} h_t^2 \right). \end{aligned}$$

Взяв математическое ожидание обеих частей этого неравенства и воспользовавшись тем, что $f - \alpha$ -PL функция, получаем

$$(8) \quad \delta_{t+1} \leq \delta_t \left(1 - \frac{\eta_t \alpha}{2} \right) + \frac{\eta_t}{2} d \left(L\kappa_\beta h_t^{\beta-1} \right)^2 + \frac{3\bar{L}\eta_t^2}{4} \kappa d \left(\frac{\sigma^2}{h_t^2} + \frac{3\bar{L}^2 d}{4} h_t^2 \right),$$

где $\delta_t = \mathbf{E}[f(\mathbf{x}_t) - f^*]$. Пусть $T_0 = \lceil 72\bar{L}d\kappa/\alpha \rceil$. Заметим, что

$$\eta_t = \begin{cases} \frac{1}{18\bar{L}d\kappa} & \text{при } t \leq T_0, \\ \frac{4}{\alpha t} & \text{при } t \geq T_0 + 1 \end{cases}$$

и

$$(9) \quad \frac{4}{(T_0 + 1)\alpha} \leq \eta_t \leq \frac{4}{T_0\alpha}, \quad \text{при } t \leq T_0.$$

Рассмотрим отдельно случаи $T \geq T_0 + 1$ и $T \leq T_0$.

Используя (8), легко проверить, что результат теоремы имеет место при $T \geq T_0 + 1$ и $T \in \{1, 2\}$.

Рассмотрим теперь случай, когда $T \geq T_0 + 1$ и $T \geq 3$, т.е. $T \geq T'_0 + 1$, где $T'_0 = \max(T_0, 2)$. При всех $t \geq T_0 + 1$ имеем $\eta_t = 4/(\alpha t)$, так что

$$\delta_{t+1} \leq \delta_t \left(1 - \frac{2}{t} \right) + \frac{d}{\alpha t} \left(2 \left(L\kappa_\beta h_t^{\beta-1} \right)^2 + \frac{3\bar{L}}{\alpha t} \kappa \left(\frac{\sigma^2}{h_t^2} + \frac{3\bar{L}^2 d}{4} h_t^2 \right) \right).$$

Подставляя сюда $h_t = \left(\frac{3\bar{L}}{\alpha t} \frac{\kappa \sigma^2}{2L^2 \kappa_\beta^2} \right)^{\frac{1}{2\beta}}$, получаем, что при $t \geq T'_0 + 1$ выполняется неравенство

$$\delta_{t+1} \leq \delta_t \left(1 - \frac{2}{t} \right) + \mathbf{A} \frac{d}{\alpha t} \left(\left(\frac{1}{\alpha t} \right)^{\frac{\beta-1}{\beta}} + \frac{d}{t} \left(\frac{1}{\alpha t} \right)^{\frac{1}{\beta}} \right).$$

Здесь и далее через $\mathbf{A}$ обозначаются положительные постоянные, не обязательно принимающие одни и те же значения и зависящие только от $L, \bar{L}, \beta$, и σ^2 . Так как $T \geq T'_0 + 1 \geq 3$, то в силу [21, лемма 32] справедлива верхняя граница

$$(10) \quad \delta_T \leq \frac{2T'_0}{T} \delta_{T'_0+1} + \mathbf{A} \frac{d}{\alpha} \left(\left(\frac{1}{\alpha T} \right)^{\frac{\beta-1}{\beta}} + \frac{d}{T} \left(\frac{1}{\alpha T} \right)^{\frac{1}{\beta}} \right).$$

Если $T_0 \leq 2$ (т.е. $T'_0 = 2$), то используя (8) при $t = 1, t = 2$, определение T_0 , соотношение (9) и определение h_t , получаем

$$\delta_{T'_0+1} = \delta_3 \leq \delta_1 + \mathbf{A} \frac{d}{\alpha} \left(\left(\frac{1}{\alpha} \right)^{\frac{\beta-1}{\beta}} + d \left(\frac{1}{\alpha} \right)^{\frac{1}{\beta}} \right).$$

Подстановка этого неравенства в (10) завершает доказательство для случая $T'_0 = 2, T \geq T'_0 + 1$.

Пусть теперь $T'_0 \geq 3$ (что влечет $T_0 = T'_0$) и $T \geq T'_0 + 1$. Из неравенств (8), (9) и определения h_t следует, что для всех $t \leq T_0$,

$$(11) \quad \delta_{t+1} \leq \delta_t \left(1 - \frac{2}{T_0 + 1} \right) + \mathbf{A} \frac{d}{\alpha T_0} \left(\left(\frac{1}{\alpha} \right)^{\frac{\beta-1}{\beta}} \left(T^{-\frac{\beta-1}{\beta}} + \frac{T^{\frac{1}{\beta}}}{T_0} \right) + \frac{d}{T_0} \left(\frac{1}{\alpha T} \right)^{\frac{1}{\beta}} \right).$$

Итерируя (11) с упрощением $1 - 2/(T_0 + 1) \leq 1$ и принимая во внимание определение T_0 и неравенство $T \geq T_0 + 1$, получаем

$$\frac{2T_0}{T} \delta_{T_0+1} \leq \frac{144\bar{L}d\kappa}{\alpha T} \delta_1 + \mathbf{A} \frac{d}{\alpha} \left(\left(\frac{1}{\alpha T} \right)^{\frac{\beta-1}{\beta}} + \frac{d}{T} \left(\frac{1}{\alpha T} \right)^{\frac{1}{\beta}} \right).$$

Подстановка этого неравенства в (10) дает:

$$\delta_T \leq \frac{144\bar{L}d\kappa}{\alpha T} \delta_1 + \mathbf{A} \frac{d}{\alpha} \left(\left(\frac{1}{\alpha T} \right)^{\frac{\beta-1}{\beta}} + \frac{d}{T} \left(\frac{1}{\alpha T} \right)^{\frac{1}{\beta}} \right),$$

что завершает доказательство для случая $T \geq T_0 + 1$.

Рассмотрим теперь случай $T \leq T_0$. В этом случае (11) выполняется при всех $t \leq T$, так что

$$\delta_{T+1} \leq \delta_1 \left(1 - \frac{2}{T_0 + 1} \right)^T + \mathbf{A} \frac{d}{\alpha} \left(\left(\frac{1}{\alpha T} \right)^{\frac{\beta-1}{\beta}} + \frac{d}{T} \left(\frac{1}{\alpha T} \right)^{\frac{1}{\beta}} \right).$$

Так как $(1 - \lambda)^T \leq \exp(-\lambda T) \leq 1/(\lambda T)$ для всех $T > 0, \lambda \in (0, 1)$, то

$$\delta_{T+1} \leq \frac{T_0 + 1}{2T} \delta_1 + \mathbf{A} \frac{d}{\alpha} \left(\left(\frac{1}{\alpha T} \right)^{\frac{\beta-1}{\beta}} + \frac{d}{T} \left(\frac{1}{\alpha T} \right)^{\frac{1}{\beta}} \right).$$

Учитывая, что $4TT_0 \geq (T+1)(T_0+1)$, окончательно приходим к неравенству

$$\begin{aligned} \delta_{T+1} &\leq \frac{2T_0}{T+1} \delta_1 + \mathbf{A} \frac{d}{\alpha} \left(\left(\frac{1}{\alpha(T+1)} \right)^{\frac{\beta-1}{\beta}} + \frac{d}{T+1} \left(\frac{1}{\alpha(T+1)} \right)^{\frac{1}{\beta}} \right) \leq \\ &\leq \frac{144\bar{L}d\kappa}{\alpha(T+1)} \delta_1 + \mathbf{A} \frac{d}{\alpha} \left(\left(\frac{1}{\alpha(T+1)} \right)^{\frac{\beta-1}{\beta}} + \frac{d}{T+1} \left(\frac{1}{\alpha(T+1)} \right)^{\frac{1}{\beta}} \right). \end{aligned}$$

Доказательство теоремы 2. Фиксируя $\mathbf{x} \in \Theta$, по лемме 5 имеем

$$\begin{aligned} f(\mathbf{x}_t) - f(\mathbf{x}) &\leq \frac{a_t - \mathbf{E}[a_{t+1}|\mathbf{x}_t]}{2\eta_t} + \frac{\|\nabla f(\mathbf{x}_t) - \mathbf{E}[\mathbf{g}_t|\mathbf{x}_t]\|^2}{\alpha} + \\ &+ \frac{\eta_t}{2} \mathbf{E}[\|\mathbf{g}_t\|^2|\mathbf{x}_t] - \frac{\alpha a_t}{4}, \end{aligned}$$

где $a_t = \|\mathbf{x}_t - \mathbf{x}\|^2$. В силу лемм 2, 3 и условия $\max_{\mathbf{x} \in \Theta} \|\nabla f(\mathbf{x})\| \leq G$ получим

$$\begin{aligned} f(\mathbf{x}_t) - f(\mathbf{x}) &\leq \frac{a_t - \mathbf{E}[a_{t+1}|\mathbf{x}_t]}{2\eta_t} + \frac{d}{\alpha} (\kappa_\beta L h_t^{\beta-1})^2 + \\ &+ \frac{3\eta_t}{4} \kappa d \left(\frac{3}{4} (d\bar{L}^2 h_t^2 + 8G^2) + \frac{\sigma^2}{h_t^2} \right) - \frac{\alpha a_t}{4}. \end{aligned}$$

Пусть $b_t = \mathbf{E}[\|\mathbf{x}_t - \mathbf{x}\|^2]$. Подставив сюда $\eta_t = 4/(\alpha(t+1))$ и взяв математическое ожидание, приходим к неравенству

$$\begin{aligned} \mathbf{E}[f(\mathbf{x}_t) - f(\mathbf{x})] &\leq \frac{\alpha}{8} ((t+1)(b_t - b_{t+1}) - 2b_t) + \\ &+ \frac{d}{\alpha} (\kappa_\beta L h_t^{\beta-1})^2 + \frac{3}{\alpha(t+1)} \kappa d \left(\frac{3}{4} (d\bar{L}^2 h_t^2 + 8G^2) + \frac{\sigma^2}{h_t^2} \right). \end{aligned}$$

Суммируя обе части этого неравенства от 1 до T и используя тот факт, что

$$\sum_{t=1}^T t((t+1)(b_t - b_{t+1}) - 2b_t) \leq 0,$$

находим

$$\begin{aligned} \sum_{t=1}^T t \mathbf{E}[f(\mathbf{x}_t) - f(\mathbf{x})] &\leq \\ &\leq \frac{d}{\alpha} \sum_{t=1}^T \left(t(\kappa_\beta L h_t^{\beta-1})^2 + \frac{3t}{t+1} \kappa \left(\frac{3}{4} (d\bar{L}^2 h_t^2 + 8G^2) + \frac{\sigma^2}{h_t^2} \right) \right). \end{aligned}$$

Подставляя сюда $h_t = \left(\frac{3}{2t} \frac{\kappa \sigma^2}{\kappa_\beta^2 L^2} \right)^{\frac{1}{2\beta}}$, приходим к неравенству

$$\begin{aligned} \sum_{t=1}^T t \mathbf{E}[f(\mathbf{x}_t) - f(\mathbf{x})] &\leq \frac{18G^2 d \kappa T}{\alpha} + \mathbf{A} \frac{d}{\alpha} \sum_{t=1}^T \left(t^{\frac{1}{\beta}} + d t^{-\frac{1}{\beta}} \right) \leq \\ &\leq \frac{18G^2 d \kappa T}{\alpha} + \mathbf{A}' \frac{d}{\alpha} \left(1 + d T^{-\frac{2}{\beta}} \right) T^{\frac{\beta+1}{\beta}}. \end{aligned}$$

Здесь через $\mathbf{A}, \mathbf{A}'$ обозначаются положительные постоянные, зависящие только от $L, \bar{L}, \beta$, и σ^2 . Разделив обе части последнего неравенства на $T(T+1)/2$ и применив неравенство Йенсена, получим результат теоремы.

СПИСОК ЛИТЕРАТУРЫ

1. *Stone C.J.* Additive regression and other nonparametric models // *Annals of Statistics*. 1985. V. 13. P. 689–705.
2. *Hastie T., Tibshirani R.* Generalized additive models // *Statistical Science*. 1986. V. 1. No. 3. P. 297–310.
3. *Wood S.N.* Generalized additive models. Chapman and Hall/CRC, 2017.
4. *Stone C.J.* Optimal rates of convergence for nonparametric estimators // *Annals of Statistics*. 1980. V. 8. P. 1348–1360.
5. *Stone C.J.* Optimal global rates of convergence for nonparametric regression // *Annals of Statistics*. 1982. V. 10. P. 1040–1053.
6. *Ibragimov I.A., Khas'minskiĭ R.Z.* Statistical estimation: Asymptotic theory. Springer, 1981.
7. *Ибрагимов И.А., Хасъминский Р.З.* О границах качества непараметрического оценивания регрессии // *Теория вероятн. и ее примен.* 1982. V. 27. No. 1. P. 81–94.
8. *Поляк Б.Т.* Градиентные методы минимизации функционалов // *Ж. вычисл. матем. и матем. физ.* 1963. V. 3. No. 4. P. 643–653.
9. *Lojasiewicz S.* A topological property of real analytic subsets // *Coll. du CNRS, Les équations aux dérivées partielles*. 1963. V. 117. No. 2. P. 87–89.
10. *Kiefer J., Wolfowitz J.* Stochastic estimation of the maximum of a regression function // *Annals of Mathematical Statistics*. 1952. V. 23. P. 462–466.
11. *Fabian V.* Stochastic approximation of minima with improved asymptotic speed // *Annals of Mathematical Statistics*. 1967. V. 38. No. 1. P. 191–200.
12. *Nemirovsky A.S., Yudin D.B.* Problem complexity and method efficiency in optimization. Wiley & Sons, 1983.
13. *Поляк Б.Т., Цыбаков А.Б.* Оптимальные порядки точности поисковых алгоритмов стохастической оптимизации // *Пробл. передачи информ.* 1990. V. 26. No. 2. P. 45–53.
14. *Jamieson K.G., Nowak R., Recht B.* Query complexity of derivative-free optimization // *Advances in Neural Information Processing Systems*. 2012. V. 26. P. 2672–2680.
15. *Shamir O.* On the complexity of bandit and derivative-free stochastic convex optimization // *Proc. 30th Annual Conference on Learning Theory*. 2013. P. 1–22.
16. *Ghadimi S., Lan G.* Stochastic first- and zeroth-order methods for nonconvex stochastic programming // *SIAM Journal on Optimization*. 2013. V. 23(4). P. 2341–2368.
17. *Nesterov Y., Spokoiny V.* Random gradient-free minimization of convex functions // *Foundations of Computational Mathematics*. 2017. V. 17. No. 2. P. 527–566.
18. *Bach F., Perchet V.* Highly-smooth zero-th order online optimization // *Proc. 29th Annual Conference on Learning Theory*. 2016.
19. *Akhavan A., Pontil M., Tsybakov A.B.* Exploiting higher order smoothness in derivative-free optimization and continuous bandits // *Advances in Neural Information Processing Systems*. 2020. V. 33. P. 9017–9027.
20. *Balasubramanian K., Ghadimi S.* Zeroth-order nonconvex stochastic optimization: Handling constraints, high dimensionality, and saddle points // *Foundations of Computational Mathematics*. 2021. P. 1–42.

21. Akhavan A., Chzhen E., Pontil M., Tsybakov A.B. Gradient-free optimization of highly smooth functions: improved analysis and a new algorithm // Journal of Machine Learning Research. 2024. V. 25. No. 370. P. 1–50.
22. Гасников А.В., Лагуновская А.А., Усманова И.Н., Федоренко Ф.А. Безградиентные прокс-методы с неточным оракулом для негладких задач выпуклой стохастической оптимизации на симплексе // Автомат. и телемех. 2016. No. 10. P. 57–77.
23. Novitskii V., Gasnikov A. Improved exploitation of higher order smoothness in derivative-free optimization // Optimization Letters. 2022. V. 16. P. 2059–2071.
24. Akhavan A., Pontil M., Tsybakov A.B. Distributed zero-order optimization under adversarial noise // Advances in Neural Information Processing Systems. 2021. V. 34. P. 10209–10220.
25. Gasnikov A.V., Lobanov A.V., Stonyakin F.S. Highly smooth zeroth-order methods for solving optimization problems under the PL condition // Computational Mathematics and Mathematical Physics. 2024. V. 64. P. 739–770.
26. Yu Q., Wang Y., Huang B., et al. Stochastic zeroth-order optimization under strong convexity and Lipschitz Hessian: Minimax sample complexity // Advances in Neural Information Processing Systems. 2024. V. 37.
27. Fokkema H., van der Hoeven D., Lattimore T., Mayo J.J. Online Newton method for bandit convex optimisation // arXiv preprint arXiv:2406.06506. 2024.
28. Karimi H., Nutini J., Schmidt M. Linear convergence of gradient and proximalgradient methods under the Polyak–Łojasiewicz condition // Machine Learning and Knowledge Discovery in Databases. 2016. P. 795–811.
29. Tsybakov A.B. Introduction to nonparametric estimation. New York: Springer, 2009.
30. Граничин О.Н. Процедура стохастической аппроксимации с возмущением на входе // Автомат. и телемех. 1992. No. 2. P. 97–104.
31. Polyak B.T., Tsybakov A.B. On stochastic approximation with arbitrary noise (the KW-case) // Advances in Soviet Mathematics, vol. 12. 1992. P. 107–113.

Статья представлена к публикации членом редколлегии П.С. Щербаковым.

Поступила в редакцию 03.03.2025

После доработки 20.05.2025

Принята к публикации 27.06.2025