ПРОГРАММИРОВАНИЕ, 2023, № 6, с. 60–74

АНАЛИЗ ДАННЫХ

УДК 519.85

БЕЗГРАДИЕНТНЫЕ АЛГОРИТМЫ ДЛЯ РЕШЕНИЯ СТОХАСТИЧЕСКИХ СЕДЛОВЫХ ЗАДАЧ ОПТИМИЗАЦИИ С УСЛОВИЕМ ПОЛЯКА–ЛОЯСИЕВИЧА

© 2023 г. С. И. Садыков[a,*] (ORCID: 0009-0008-7101-6532), А. В. Лобанов[a,b,] (ORCID: 0000-0003-1620-9581), А. М. Райгородский[a,b,c,***]**

[a]*Московский физико-технический институт, 141701, г. Долгопрудный, Институтский пер., 9, Россия*

[b]*Исследовательский центр доверенного искусственого интеллекта ИСП РАН, 109004, г. Москва, ул. Александра Солженицына, 25, Россия*

[c]*Кавказский математический центр Адыгейского гос. университета, 385000, г. Майкоп, ул. Первомайская, 208, Россия*

*****E-mail: sadykov.si@phystech.edu*

******E-mail: lobbsasha@mail.ru*

*******E-mail: mraigor@yandex.ru*

Поступила в редакцию 13.06.2023 г.
После доработки 10.07.2023 г.
Принята к публикации 20.07.2023 г.

Данная работа фокусируется на решения подкласса стохастической невыпукло-невогнутой задачи оптимизации черного ящика с седловой точкой, которая удовлетворяет условию Поляка–Лоясиевича. Для решения такой задачи мы предоставляем первый, насколько нам известно, безградиентный алгоритм, подход к созданию которого основывается на применении градиентной аппроксимации (ядерной аппроксимации) к алгоритму стохастического градиентного спуска подъема со смещенным оракулом. Мы представляем теоретические оценки, гарантирующие глобальную линейную скорость сходимости к желаемой точности. Теоретические результаты мы проверяем на модельном примере, сравнивая с алгоритмом, использующую Гауссовскую аппроксимацию.

DOI: 10.31857/S0132347423060079, **EDN:** DSVULZ

1. ВВЕДЕНИЕ

В данной работе изучается стандартная стохастическая задача оптимизации с седловой точкой, которая имеет следующий вид:

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} f(x,y) := \mathbb{E}\big[f(x,y,\xi)\big]. \tag{1.1}$$

Такая минимаксная задача широко распространена и активно исследуется в области теории игр и исследований операций, а также в современных задачах машинного обучения: генеративно-состязательные нейросети (GANs) [1], обучение с подкреплением (Reinforcement learning) [2]. В частности, стоит отметить другие приложения, включающие в себя надежную оптимизацию (Robust optimization) [3], обучение без учителя (Unsupervised learning) [4], состязательное машинное обучение (Adversarial machine learning) [5] и другие. Мы рассматриваем более узкую задачу оптимизации с седловой точкой (1.1), предполагая, что целевая функция f является не просто гладкой, а имеет повышенную гладкость, а также предполагаем, что оракул может вернуть при обращении только зашумленное значение целевой функции $\tilde{f} = f + \delta$ (шум ограничен). Такой класс задач имеет различные названия, в частности, часто упоминающее в литературе: задача “черного ящика” (Black box problem) [6]. Где в качестве черного ящика выступает тот самый оракул $\tilde{f}$, который в дальнейшем будет иметь название “оракул нулевого порядка” (zero-order oracle) [7].

Существуют различные техники решения задач черного ящика [8], основная идея которых состоит в использовании оптимального (ускоренного пробатченного) алгоритма первого порядка, заменяя истинный градиент на соответствующую градиентную аппроксимацию. Выбор градиентной аппроксимации зачастую зависит от предположений на целевую функцию, например в работах [9] и [10] используется схема сглаживания с l_1 и l_2 соответственно, поскольку предполагается, что целевая функция является негладкой. В нашем случае функция f является не только глад-

кой, но может иметь и повышенную гладкость, поэтому в качестве градиентной аппроксимации мы будем использовать ту, которая учитывает преимущество порядка гладкости [11]

$$\frac{d}{2\gamma}\left(\tilde{f}(z+\gamma r\mathbf{e},\xi)-\tilde{f}(z-\gamma r\mathbf{e},\xi)\right)K(r)\mathbf{e}. \qquad (1.2)$$

В работах [12] и [13] авторы предложили безградиентный метод для решения седловой задачи в выпукло-вогнутой настройке и для решения задачи оптимизации в (сильно) выпуклой настройке соответственно, используя ядерную аппроксимацию (1.2). Однако наибольший интерес с точки зрения приложений возникает в задачах с невыпуклой настройкой. Тогда на помощь приходит, пожалуй, одно из немногих условий позволяющих для подкласса невыпуклых задач доказать глобальную сходимость. Данное условие имеет следующее название: условие Поляка–Лоясиевича [14, 15]. Для задачи оптимизации в статье [16] авторы предложили смещенный алгоритм первого порядка biased SGD, а уже совсем недавно в [17] предложили безградиентный алгоритм, который основывается на алгоритме [16]. Стоит обратить внимание, что при условии Поляка–Лоясиевича неускоренные алгоритмы уже являются оптимальными [18], именно поэтому авторы статьи [17] основывались на смещенном SGD при создании нового безградиентного алгоритма для решения невыпуклых задач оптимизации черного ящика. Также совсем недавно был предложен алгоритм Stoc-AGDA [19] для решения стохастической задачи оптимизации с седловой точкой, удовлетворяющей условию Поляка–Лоясиевича. Но в настоящее время нет алгоритма, который решит минимаксную задачу черного ящика, целевая функция которой удовлетворяет условию Поляка–Лоясиевича.

Таким образом, мы можем сформулировать наш основной вклад в данную статью. Мы фокусируемся на решении стохастической седловой задачи оптимизации, когда вычисление градиента недоступно. Для создания безградиентного алгоритма мы обобщаем результаты сходимости алгоритма Stoc-AGDA из статьи [19] на случай со смещенным оракулом (данный результат может вызывать независимый интерес). Мы предоставляем новый безградиентный алгоритм нулевого порядка стохастического градиентного спуска подъема (Zero-Order SGDA). Мы анализируем сходимость предложенного алгоритма при различных вариантах безградиентного оракула. В качестве основных результатов мы предоставляем следующие оценки: общее число последовательных итераций N, общее число обращений к оракулу нулевого порядка T, а также максимально допустимый уровень враждебного шума Δ. С помощью практического эксперимента мы подтверждаем теоретические результаты, показывая преимущество “ядерной” аппроксимации над Гауссовской в задаче с седловой точкой в безградиентной настройке.

2. ОСНОВНЫЕ ОБОЗНАЧЕНИЯ И ПРЕДПОЛОЖЕНИЯ

Обозначения. На протяжении всей статьи мы используем следующие обозначения. Мы используем $\langle x, y\rangle := \sum_{i=1}^{d} x_i y_i$ для обозначения стандартного скалярного произведения $x, y \in \mathbb{R}^d$. Мы используем $\|\cdot\|$ для обозначения евклидовой нормы вектора $\|x\| := \left(\sum_{i=1}^{d} |x_i|^2\right)^{1/2} = \sqrt{\langle x, x\rangle}$. Для обозначения евклидова шара и сферы мы используем следующие обозначения:

$$\mathcal{B}^d := \left\{x \in \mathbb{R}^d : \|x\| \le 1\right\}$$

$$\mathcal{S}^d := \left\{x \in \mathbb{R}^d : \|x\| = 1\right\}.$$

Для любого $\beta \ge 2$ мы обозначаем через $\lfloor\beta\rfloor$ наибольшее целое число, строго меньшее β. Через $\mathcal{O}(\cdot)$ мы обозначаем верхнюю границу с точностью до константы. Также мы используем $\tilde{\mathcal{O}}(\cdot)$, чтобы скрыть логарифмический множитель.

2.1. Предположения

Для начала необходимо определить три понятия оптимальности для минимаксной задачи (1.1). Самое прямое понятие оптимальности – это глобальная минимаксная точка, в которой x^*– оптимальное решение задачи минимизации функции $\max_y f(x, y)$, а y^* – оптимальное решение для $\max_y f(x^*, y)$. Для седловой точки (x^*, y^*) x^* является оптимальным решением для $\min_x f(x, y^*)$ и y^* – оптимальным решением для $\max_y f(x^*, y)$.

Определение 1 (Глобальное решение).

1. (x^*, y^*) *глобальная минимаксная точка* (*global minimax point*), *если для любых* (x, y):

$$f(x^*, y) \le f(x^*, y^*) \le \max_{y'} f(x, y'). \qquad (2.1)$$

2. (x^*, y^*) *седловая точка* (*saddle point*), *если для любых* (x, y) :

$$f(x^*, y) \le f(x^*, y^*) \le f(x, y^*). \qquad (2.2)$$

3. (x^*, y^*) *стационарная точка* (*stationary point*), *если*:

$$\nabla_x f(x^*, y^*) = \nabla_y f(x^*, y^*) = 0. \qquad (2.3)$$

Такое определение используется в [19].

На протяжении всей статьи предполагается, что функция f в (1.1) непрерывно дифференцируема и имеет Липшицев градиент.

Предположение 1 (Градиент Липшица). *Существует положительное число $l > 0$ такое что*:

$$\|\nabla_x f(x_1, y_1) - \nabla_x f(x_2, y_2)\| \le L_2 \left[\|x_1 - x_2\| + \|y_1 - y_2\|\right],$$

$$\|\nabla_y f(x_1, y_1) - \nabla_y f(x_2, y_2)\| \le L_2 \left[\|x_1 - x_2\| + \|y_1 - y_2\|\right],$$

выполняется для всех $x_1, x_2 \in \mathbb{R}^{d_x}, y_1, y_2 \in \mathbb{R}^{d_y}$.

Такое предположение используется в [19].

Введем прямое обобщение PL условия на минимаксную задачу: функция $f(x, y)$ удовлетворяет условию PL с константой μ_x относительно x, а $-f(x, y)$ удовлетворяет условию PL с константой μ_y относительно y. Мы формально сформулируем это в следующем предположении.

Предположение 2 (Двустороннее PL условие). *Непрерывно дифференцируемая функция $f(x, y)$ удовлетворяет двустороннему условию PL, если существуют константы $\mu_x, \mu_y > 0$ такие что $\forall x \in \mathbb{R}^{d_x}$, $y \in \mathbb{R}^{d_y}$ выполняется*:

$$\|\nabla_x f(x, y)\|^2 \ge 2\mu_x [f(x, y) - \min_x f(x, y)],$$

$$\|\nabla_y f(x, y)\|^2 \ge 2\mu_y [\max_y f(x, y) - f(x, y)].$$

Такое предположение используется в [19, 17].

Для всех наших теоретических результатов мы предполагаем, что f не просто гладкая, а имеет высокий порядок гладкости.

Предположение 3 (Условие Гельдера). *Зафиксируем некоторые $\beta \ge 2$ и $L_\beta > 0$. Обозначим через $\mathcal{F}_\beta(L_\beta)$ множество всех функций $f : \mathbb{R}^d \to \mathbb{R}$, которые являются $l = \lfloor \beta \rfloor$ раз непрерывно дифференцируемы и удовлетворяют для всех x, $x' \in \mathbb{R}^{d_x}$ y, $y' \in \mathbb{R}^{d_y}$ условию Гельдера*

$$\|f^{(l)}(x, y) - f^{(l)}(x', y)\| \le L_\beta \|x - x'\|^{\beta - l},$$

$$\|f^{(l)}(x, y) - f^{(l)}(x, y')\| \le L_\beta \|y - y'\|^{\beta - l}.$$

Такое предположение используется в [20].

Теперь сформулируем стандартные предположения для смещенного градиентного оракула. Для этого введем следующее определение.

Определение 2 (Смещенный градиентный оракул). *Для отображений $\mathbf{G_x} : \mathbb{R}^{d_x + d_y} \times \mathcal{D} \to \mathbb{R}^{d_x}$, $\mathbf{G_y} : \mathbb{R}^{d_x + d_y} \times \mathcal{D} \to \mathbb{R}^{d_y}$ выполнено*:

$$\mathbf{G_x}(x, y, \xi) = \nabla_x f(x, y) + \mathbf{b_x}(x, y) + \mathbf{n_x}(x, y, \xi),$$

$$\mathbf{G_y}(x, y, \xi) = \nabla_y f(x, y) + \mathbf{b_y}(x, y) + \mathbf{n_y}(x, y, \xi),$$

где $\mathbf{b_x} : \mathbb{R}^{d_x + d_y} \to \mathbb{R}^{d_x}$, $\mathbf{b_y} : \mathbb{R}^{d_x + d_y} \to \mathbb{R}^{d_y}$ смещения (bias), $\mathbf{n_x} : \mathbb{R}^{d_x + d_y} \times \mathcal{D} \to \mathbb{R}^{d_x}$, $\mathbf{n_y} : \mathbb{R}^{d_x + d_y} \times \mathcal{D} \to \mathbb{R}^{d_y}$ нулевой средний шум (zero-mean noise), то есть $\mathbb{E}_\xi \mathbf{n_x}(x, y, \xi) = \mathbb{E}_\xi \mathbf{n_y}(x, y, \xi) = 0$, $\forall x \in \mathbb{R}^{d_x}$, $y \in \mathbb{R}^{d_y}$.

Такое определение используется в [16, 17]. А также предполагается, что этот градиентный оракул имеет ограниченные смещение и шум.

Предположение 4 ((M, σ^2)-*ограничение на шум*). *Существуют константы $M, \sigma^2 \ge 0$ такие что $\forall x \in \mathbb{R}^{d_x}$, $\forall y \in \mathbb{R}^{d_y}$*

$$\mathbb{E}_\xi \|\mathbf{n_x}(x, y, \xi)\|^2 \le M \|\nabla_x f(x, y) + \mathbf{b}_x(x, y)\|^2 + \sigma^2,$$

$$\mathbb{E}_\xi \|\mathbf{n_y}(x, y, \xi)\|^2 \le M \|\nabla_y f(x, y) + \mathbf{b}_y(x, y)\|^2 + \sigma^2$$

Такое определение используется в [16, 17].

Предположение 5 (ζ^2-*ограничение на смещение*). *Существуют константа $\zeta^2 \ge 0$ такая что $\forall x \in \mathbb{R}^{d_x}$, $\forall y \in \mathbb{R}^{d_y}$*

$$\|\mathbf{b}_x(x, y)\|^2 \le \zeta^2,$$

$$\|\mathbf{b}_y(x, y)\|^2 \le \zeta^2.$$

Такое определение используется в [16, 17]. Здесь используются общие оценки M и ζ^2 для аппроксимации по x и по y для удобства. Далее используется обозначение $d = \max(d_x, d_y)$.

3. СМЕЩЕННЫЙ СТОХАСТИЧЕСКИЙ ГРАДИЕНТНЫЙ СПУСК ПОДЪЕМ

Теперь мы можем представить алгоритм 1. Этот алгоритм является модификацией метода SGDA [19]. Основная идея данного алгоритма заключается в том, чтобы использовать "зашумленное" значение, которое возвращается градиентным оракулом (см. Определение 2) вместо истинного градиента (градиентного оракула). Кроме того, для достижения оптимальной итерационной сложности мы добавляем размер батча B.

Algorithm 1 Biased Mini-Batch Stochastic Gradient Descent Ascent (BMB-SGDA)

Вход: 2 последовательности размера шага $(\tau_{xk})_{k \ge 0}, (\tau_{yk})_{k \ge 0}$, размер батча B, $x_0 \in \mathbb{R}^{d_x}$, $y_0 \in \mathbb{R}^{d_y}$;

for $k = 0$ **to** $N - 1$ **do**

Вычислить $\mathbf{G}_{xk} = \frac{1}{B} \sum_{i=1}^{B} \tilde{\mathbf{G}}_x(x_k, y_k, \mathbf{e}_i)$

$x_{k+1} \leftarrow x_k - \tau_{xk} \mathbf{G}_{xk}$

Вычислить $\mathbf{G}_{yk} = \frac{1}{B}\sum_{i=1}^{B}\tilde{\mathbf{G}}_y(x_{k+1}, y_k, \mathbf{e}_i)$

$y_{k+1} \leftarrow y_k + \tau_{yk}\mathbf{G}_{yk}$

end for

Return: x_N, y_N

Мы хотим получить некоторый результат, говорящий о сходимости алгоритма, который основан на аппроксимации, имеющей смещение и шум. Для этого определим следующие функции:

$$g(x) = \max_y f(x,y), \quad g^* = \min_x \max_y f(x,y)$$

$$a_t = \mathbb{E}\left[g(x_t) - g^*\right], \quad b_t = \mathbb{E}[g(x_t) - f(x_t, y_t)].$$

Легко видеть, что $a_t, b_t \geq 0$. Поэтому имеет смысл минимизировать следующую величину

$$P_t = a_t + \lambda b_t. \tag{3.1}$$

При двустороннем условии PL можно показать, что функция $g(x) := \max_y f(x,y)$ удовлетворяет условию PL с μ_x (см. приложение). Более того, g имеет Липшицев градиент с константой $L := L_2 + L_2^2/\mu_y$ [21].

В следующей теореме говорится о том, что при такой постановке задачи, сходимость алгоритма линейная и можно подобрать параметр λ таким образом, что знаменатель геометрической прогрессии будет меньше единицы.

Теорема 1. *Предположим, что выполняются предположения* 1–5 *и* $f(x,y)$ *удовлетворяет двустороннему* PL*-условию с* μ_x *и* μ_y. *Определим* $P_t := a_t + \frac{1}{10}b_t$. *Если мы запустим алгоритм* 1 *с* $\tau_y^t = \tau_y = \frac{1}{(M+1)L_2}$ *и* $\tau_x^t = \tau_x = \frac{\mu_y^2\tau_y}{12L_2^2}$, *то*

$$P_t \leq (1 - \mu_x\tau_x)^t P_0 + \frac{\tau_y^2 L_2 \frac{\sigma^2}{B} + \tau_y\zeta^2}{10\mu_x\tau_x}.$$

Доказательство см. в приложении

Из теоремы 1 видно, что дисперсию шума аппроксимации (σ^2) можно уменьшить, увеличивая размер батча B, но второй момент смещения (ζ^2) так уменьшить не получится. И в целом второе слагаемое в свободном члене сложнее уменьшить хотя бы потому что перед смещением стоит размер шага в первой степени в отличие от первого слагаемого, где размер шага возводится в квадрат, благодаря чему можно уменьшить это слагаемое, уменьшая размер шага. А так как на каждой итерации решается еще и внутренняя задача максимизации, то размер шага по игрек должен быть на порядок больше, о чем прописано в условии теоремы.

4. ГЛАВНЫЙ РЕЗУЛЬТАТ

Применение стандартных методов градиентного спуск-подъема может столкнуться с проблемой невозможности получения градиента функции. В таких случаях возникает необходимость использовать безградиентные методы аппроксимации градиента. Безградиентные методы предлагают альтернативные подходы к оптимизации, которые не требуют полного вычисления градиента функции и могут быть применимы к седловым задачам.

4.1. Градиентная аппроксимация с двухточечной обратной связью

В данном разделе мы описываем наш подход к решению задачи (1.1), учитывая, что оракул градиента (см. Определение 2) не предоставляет информацию о производных целевой функции. Наш подход состоит в разработке нового алгоритма под названием ZO-BMB-SGDA, который является оптимальным безградиентным методом, учитывающим сложность оракула, сложность итерации и максимальный уровень шума. Этот алгоритм основан на методе первого порядка, в частности, на SGDA. Для достижения этой цели, мы применяем вместо градиентного оракула (см. Определение 2) аппроксимацию градиента, которая использует оракул нулевого порядка $\tilde{f}$. Этот оракул предоставляет значение целевой функции $f(x,y,\xi)$ с добавлением враждебного детерминированного шума $\delta(x,y)$, удовлетворяющего условиям $|\delta(x,y)| \leq \Delta$ и $\Delta > 0$.

$$\tilde{f}(x,y,\xi) = f(x,y,\xi) + \delta(x,y). \tag{4.1}$$

Эта концепция враждебного шума хорошо описана в [8]. Для решения данной задачи мы применяем безградиентную ядерную аппроксимацию градиента, которая представляет собой предпочтительный выбор, поскольку учитывает повышенную гладкость функции.

$$\begin{gathered}\tilde{\mathbf{G}}_{\mathbf{x}}(x,y,\xi,\mathbf{e}) = \\ = d_x\frac{\tilde{f}(x+\gamma r\mathbf{e},y,\xi) - \tilde{f}(x-\gamma r\mathbf{e},y,\xi)}{2\gamma}K(r)\mathbf{e}, \\ \tilde{\mathbf{G}}_{\mathbf{y}}(x,y,\xi,\mathbf{e}) = \\ = d_y\frac{\tilde{f}(x,y+\gamma r\mathbf{e},\xi) - \tilde{f}(x,y-\gamma r\mathbf{e},\xi)}{2\gamma}K(r)\mathbf{e},\end{gathered} \tag{4.2}$$

где $\mathbf{e}$ равномерно распределен на сфере $S_2^d(1)$, r равномерно распределен на отрезке $[-1,1]$, $\mathbf{e}$ и r

независимы, $K:[-1,1]\to\mathbb{R}$ – это ядро функции, которое удовлетворяет следующим условиям:

$$\mathbb{E}[K(u)]=0,\quad \mathbb{E}[uK(u)]=1,$$

$$\mathbb{E}[u^jK(u)]=0,\quad j=2,\ldots,p,\quad \mathbb{E}[|u|^\beta|K(u)|]<\infty.$$

В следующей теореме представлены результаты сходимости алгоритма 1 Zero-Order Biased Mini-Batch Stochastic Gradient Descent Ascent с аппроксимацией градиента (4.2) с помощью оракула нулевого порядка (4.1).

Теорема 2. *Пусть функция $f(x,y)$ удовлетворяет предположениям* 1–3 *и градиентная аппроксимация* (4.2) *удовлетворяет предположениям* 4–5 *и пусть размер шага* $\tau_y=\dfrac{1}{(M+1)L_2}$ *и* $\tau_x=\dfrac{\mu_y^2\tau_y}{12L_2^2}$, *тогда существуют параметры*

$$M=4d\beta^3\qquad \sigma^2=4d\beta^3L_2\gamma^2+\frac{d^2\Delta^2\beta^3}{\gamma^2},$$

$$\zeta^2=\beta^2\left(\frac{L_\beta}{(l-1)!}\frac{d}{d+\beta-1}\gamma^{\beta-1}+d\frac{\Delta}{\gamma}\right)^2$$

такие, что Алгоритм 1 *с параметром сглаживания* $\gamma=\mathcal{O}(d^{1/\beta}\Delta^{1/\beta})$ *достигает следующего уровня ошибки*

$$P_t=\mathcal{O}\left(\frac{1}{\mu_x\mu_y^2}d^{\frac{2(\beta-1)}{\beta}}\Delta^{\frac{2(\beta-1)}{\beta}}\right),$$

для доказательства смотри раздел B. Результат сходимости, установленный в теореме 2, демонстрирует, что алгоритм 1, использующий градиентную аппроксимацию (4.2), достигает минимальной ошибки $\mathcal{O}\left(\dfrac{1}{\mu_x\mu_y^2}d^{\frac{2(\beta-1)}{\beta}}\Delta^{\frac{2(\beta-1)}{\beta}}\right)$ с линейной скоростью сходимости. Это происходит из-за накопления состязательного шума в смещении $\mathbf{b}_x(x,y)$ и $\mathbf{b}_y(x,y)$.

Следствие 1. *Пусть функция $f(x,y)$ удовлетворяет предположениям* 1–3 *и градиентная аппроксимация* (4.2) *удовлетворяет предположениям* 4–5 *и пусть размеры шагов* $\tau_y=\dfrac{1}{(M+1)L_2}$ *и* $\tau_x=\dfrac{\mu_y^2\tau_y}{12L_2^2}$, *а параметр сглаживания имеет вид* $\gamma=\mathcal{O}\left((\mu_x\mu_y^2\varepsilon)^{\frac{1}{2(\beta-1)}}\right)$, *тогда алгоритм* 1 *достигает точности* ε *для задачи* (1.1) *со следующими параметрами*:

$$\Delta=\mathcal{O}\left((\mu_x\mu_y^2)^{\frac{\beta}{2(\beta-1)}}\varepsilon^{\frac{\beta}{2(\beta-1)}}d^{-1}\right);$$

$$N=\mathcal{O}\left(\mu_x^{-1}\mu_y^{-2}\ln\frac{1}{\varepsilon}\right);\quad T=\mathcal{O}\left(\beta^3d\mu_x^{-1}\mu_y^{-2}\ln\frac{1}{\varepsilon}\right),$$

где Δ – максимальный уровень шума, N – количество итераций и T – оракульная сложность.

4.2. Градиентная аппроксимация с одноточечной обратной связью

В такой настройке оракул нулевого порядка может иметь следующий вид

$$\tilde f(x,y,\xi)=f(x,y)+\xi, \tag{4.3}$$

а градиентная аппроксимация тогда примет следующую форму

$$\tilde{\mathbf{G}}_{\mathbf{x}}(x,y,\xi,\mathbf{e})=$$
$$=d_x\frac{f(x+\gamma r\mathbf{e},y)+\xi_1-f(x-\gamma r\mathbf{e},y)-\xi_2}{2\gamma}K(r)\mathbf{e},$$
$$\tilde{\mathbf{G}}_{\mathbf{y}}(x,y,\xi,\mathbf{e})= \tag{4.4}$$
$$=d_y\frac{\tilde f(x,y+\gamma r\mathbf{e},\xi_1)-\tilde f(x,y-\gamma r\mathbf{e},\xi_2)}{2\gamma}K(r)\mathbf{e},$$

где $\xi_1\neq\xi_2$ – это враждебные стохастические шумы такие, что $\mathbb{E}[\xi_1^2]\le\tilde\Delta^2$ и $\mathbb{E}[\xi_2^2]\le\tilde\Delta^2$, где $\tilde\Delta\ge0$. Случайные величины ξ_1 и ξ_2 независимы от $\mathbf{e}$ и r. Кроме того, для этой концепции не требуется предположение о нулевом среднем ξ_1 и ξ_2. Достаточно, чтобы $\mathbb{E}[\xi_1\mathbf{e}]=0$ и $\mathbb{E}[\xi_2\mathbf{e}]=0$. В следующей теореме представлены результаты сходимости алгоритма 1 с аппроксимацией градиента (4.4) через оракул нулевого порядка (4.3).

Теорема 3. *Пусть функция $f(x,y)$ удовлетворяет предположениям* 1–3 *и градиентная аппроксимация* (4.4) *удовлетворяет предположениям* 4–5, *пусть размеры шагов* $\tau_y=\dfrac{1}{(M+1)L_2}$ *и* $\tau_x=\dfrac{\mu_y^2\tau_y}{12L_2^2}$, *тогда существуют параметры*

$$M=4d\beta^3\qquad \sigma^2=4d\beta^3L_2\gamma^2+\frac{d^2\tilde\Delta^2\beta^3}{\gamma^2},$$

$$\zeta^2=\beta^2\left(\frac{L_\beta}{(l-1)!}\frac{d}{d+\beta-1}\gamma^{\beta-1}\right)^2$$

так что метод ZO-BMB-SGD имеет следующую скорость сходимости

$$P_t=\mathcal{O}((1-\mu_x\tau_x)^tP_0).$$

Доказательство см. в приложении *C*. Результаты теоремы 3 показывают, что алгоритм 1 с градиентной аппроксимацией (4.4) имеет линейную скорость сходимости. Также, в отличие от предыдущей теоремы 3, она не имеет ярко выраженной асимптоты. Этот эффект наблюдается потому, что концепция оракула нулевого порядка (4.3) не

предполагает накопления состязательного шума в смещении, а также уменьшает дисперсию за счет большого размера партии B.

Следствие 2. *Пусть функция $f(x,y)$ удовлетворяет предположениям* 1–3 *и градиентная аппроксимация* (4.4) *удовлетворяет предположениям* 4–5, *пусть размеры шагов* $\tau_y = \frac{1}{(M+1)L_2}$ *и* $\tau_x = \frac{\mu_y^2 \tau_y}{12L_2^2}$, *а параметр сглаживания имеет вид* $\gamma = \mathbb{O}\left((\mu_x \mu_y^2 \varepsilon)^{\frac{1}{2(\beta-1)}}\right)$, *тогда алгоритм* 1 *достигает точности* ε *для задачи* (1.1) *со следующими параметрами*:

$$\tilde{\Delta} = \mathbb{O}\left(d^{\frac{-1}{2}}(\mu_x \mu_y^2 \varepsilon)^{\frac{1}{\beta-1}}\right);$$

$$N = \mathbb{O}\left(\mu_x^{-1}\mu_y^{-2}\ln\frac{1}{\varepsilon}\right); \quad T = \mathbb{O}\left(\beta^3 d\mu_x^{-1}\mu_y^{-2}\ln\frac{1}{\varepsilon}\right),$$

где $\tilde{\Delta}$ – максимальный уровень шума, N – количество итераций и T – оракульная сложность.

5. ЭКСПЕРИМЕНТЫ

В этом разделе выполняется проверка того, согласуются ли теоретически полученные границы с числовыми характеристиками метода Zero-order Biased Mini-Batch Stochastic Gradient Descent Ascent (ZO-BMB-SGDA). В частности, сравнивается алгоритм 1 с безградиентным аналогом из [16], в котором вместо точного градиента используется аппроксимация сглаживания по Гауссу. Во всех тестах мы понимаем враждебный шум как вычислительную ошибку (мантисса). Рассмотрим стандартную задачу, удовлетворяющую условию PL. А именно решение системы p нелинейных уравнений, аналогично как в [17], только для седловых задач. Задача оптимизации (1.1) имеет следующий вид:

$$\min_{x\in Q_1\subset\mathbb{R}^{d_x}}\max_{y\in Q_2\subset\mathbb{R}^{d_y}}\{f(x,y) := \|A\sin(x) + B\sin(y) - c\|^2 - 2\|B\sin(y) - B\sin(y_0)\|^2\},$$

где множества Q_1 и Q_2 являются многомерными кубами, где каждая координата лежит в отрезке $[-100, 100]$, $A \in \mathbb{R}^{p\times d_x}$, $B \in \mathbb{R}^{p\times d_y}$. В качестве ядра $K(r)$ используются взвешенные суммы полиномов Лежандра. Например, ниже приведены следующие значения для $\beta = \{1,2,3,4,5,6\}$ [11]:

$$K_\beta(r) = 3r \quad \beta = 1,2;$$

$$K_\beta(r) = \frac{15r}{4}(5 - 7r^2) \quad \beta = 3,4;$$

$$K_\beta(r) = \frac{195r}{64}(99r^4 - 126r^2 + 35) \quad \beta = 5,6.$$

На рис. 1 представлена зависимость от количества уравнений. На каждом графике разное количество итераций для лучшей наглядности. Можно увидеть, что разработанный алгоритм сходится лучше Гауссовой аппроксимации, которая имеет следующий вид

$$\tilde{\mathbf{G}}_{\mathbf{x}}(x,y,\mathbf{u}) = \frac{\tilde{f}(x+\gamma\mathbf{u},y) - \tilde{f}(x,y)}{2\gamma}\mathbf{u},$$
$$\tilde{\mathbf{G}}_{\mathbf{y}}(x,y,\mathbf{u}) = \frac{\tilde{f}(x,y+\gamma\mathbf{u}) - \tilde{f}(x,y)}{2\gamma}\mathbf{u}, \tag{5.1}$$

где $\mathbf{u} \sim \mathcal{N}(0,1)$.

На рис. 2 представлена зависимость от размера гладкости β. Можно видеть, что при меньшем β скорость, с которой сходится алгоритм выше. Это объясняется тем, что свободный член в формуле сходимости содержит коэфициент β^3.

6. ЗАКЛЮЧЕНИЕ

В данной работе был предложен новый безградиентный алгоритм для решения стохастических невыпукло-невогнутых в общем случае задач оптимизации черного ящика с седловой точкой, удовлетворяющих условию Поляка–Лоясиевича. Данный алгоритм является надежным при различных видах враждебного шума: детерминированного и стохастического. Для создания безградиентного алгоритма мы обобщили результат сходимости Stoch-AGDA на случай со смещенным градиентным оракулом (данный результат может вызывать независимый интерес). Также мы показали, что наш алгоритм, аналогично стандартной оптимизационной настройке сходится с линейной скоростью к асимптоте, однако данную асимптоту можно регулировать, тем самым достигая желаемой точности. Наши теоретические результаты подтвердились на модельном примере, где использовался тот факт, что в качестве враждебного шума выступала машинная неточность.

ПРИЛОЖЕНИЕ

A. Вспомогательные Леммы для доказательства Теоремы 1

Пусть $\kappa_\beta = \int |u|^\beta |K(u)|du$ и положим $\kappa = \int K^2(u)du$. Тогда, если K – взвешенная сумма полиномов Лежандра, то в [11], см. Приложение A.3, доказано, что κ_β и κ не зависят от d, они зависят только от β, для $\beta \geq 1$:

$$\kappa_\beta \leq 2\sqrt{2}(\beta - 1), \tag{A.1}$$

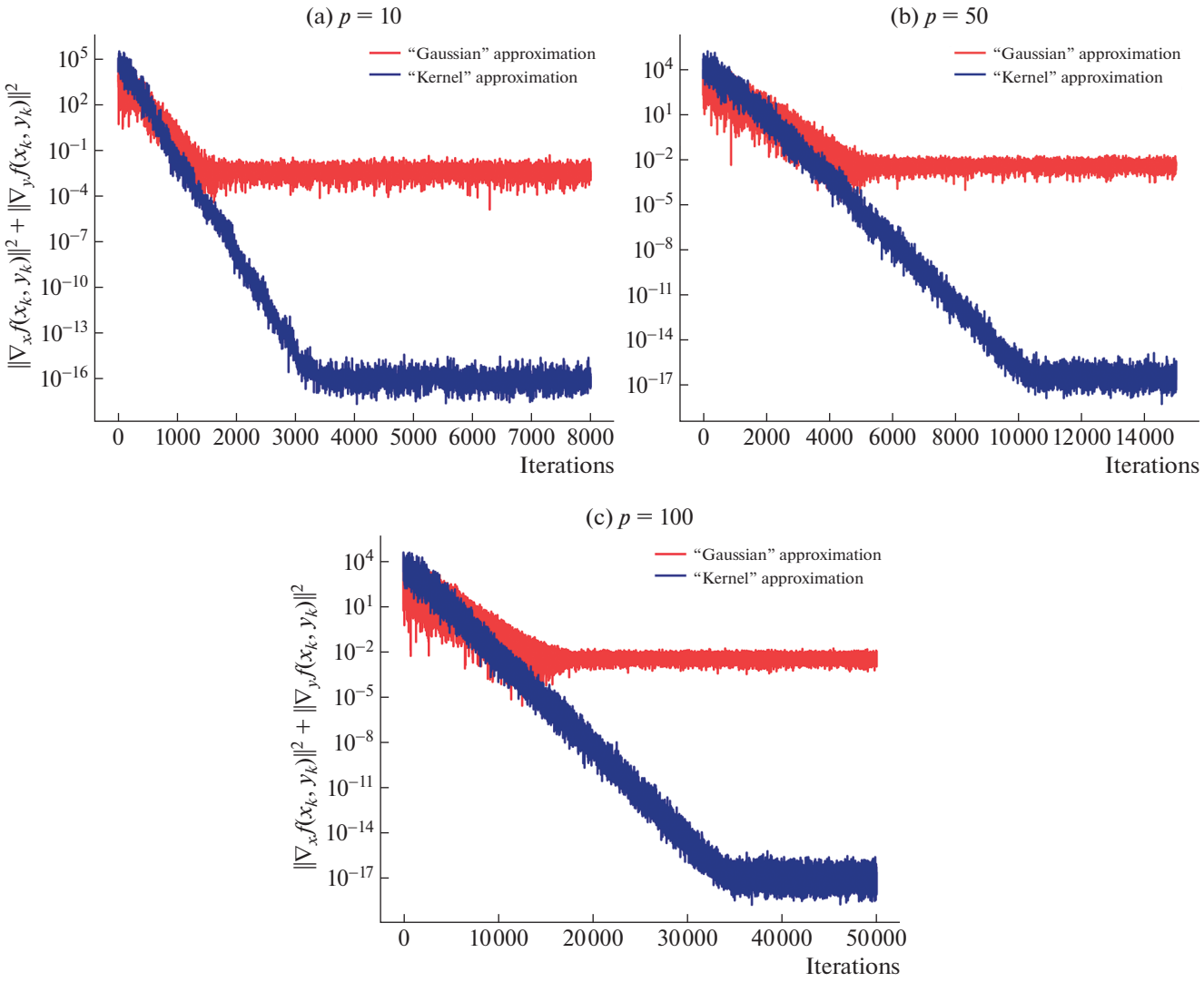

Рис. 1. Зависимость скорости сходимости от количества уравнений в системе. Параметры задачи: $d_x = 10000$, $d_y = 5000$, $\tau_x = 0.02$, $\tau_y = 0.1$, $\beta = 5$, $B = 5$, $\gamma = 0.001$.

$$\kappa_{\leq} 3\beta^3. \qquad \text{(A.2)}$$

Сначала необходимо предоставить несколько ключевых лемм.

Лемма 1 ([24]). *Если $f(\cdot)$ является L_2-гладкой и удовлетворяет условию PL с константой μ, то она также удовлетворяет условию ограниченности ошибки с μ, т.е.*

$$\|\nabla f(x)\| \geq \mu\|x_p - x\|, \quad \forall x,$$

где x_p — проекция x на оптимальное множество, она также удовлетворяет условию квадратичного роста с μ, т.е.

$$f(x) - f^* \geq \frac{\mu}{2}\|x_p - x\|^2, \quad \forall x.$$

Наоборот, если $f(\cdot)$ является L_2-гладкой и удовлетворяет условию ограниченности ошибки с константой μ, то она удовлетворяет условию PL с константой μ/L_2.

Из вышеуказанной леммы легко увидеть, что $L_2 \geq \mu$.

Лемма 2 ([21]). *В минимаксной задаче, когда $-f(x,\cdot)$ удовлетворяет условию PL с константой μ_y для любого x и f удовлетворяет предположению 1, тогда функция $g(x) := \max_y f(x,y)$ является L-гладкой с $L := L_2 + L_2^2/\mu_y$ и $\nabla g(x) = \nabla_x f(x, y^*(x))$ для любого $y^*(x) \in \arg\max_y f(x,y)$.*

Для следующей леммы необходимо расммотреть задачу $\min_x f(x)$

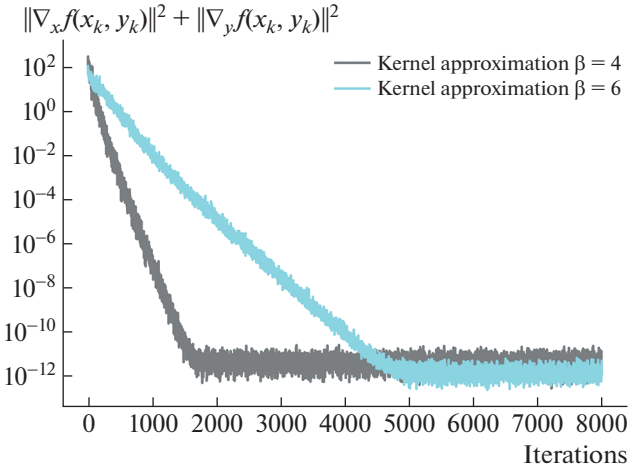

Рис. 2. Зависимость скорости сходимости от количества уравнений в системе. Параметры задачи: $d_x = 200$, $d_y = 200$, $p = 250$, $\beta = \{4, 6\}$, $B = 50$, $\gamma = 0.01$, $\tau_x = 0.04$, $\tau_y = 0.2$.

Лемма 3. *Пусть $\{x_k\}_{k\geq 0}$ обозначает количество итераций алгоритма Mini-batch SGD, сгенерированных на функции $f(\cdot)$ при предположениях 1–5. Тогда существует размер шага $\eta \leq \frac{1}{(M+1)L_2}$ такой, что он выполняется для всех $N \geq 0$*

$$\mathbb{E}[f(x_N)] - f^* \leq \leq (1-\eta\mu)^N \left(f(x_0) - f^*\right) + \frac{\zeta^2}{2\mu} + \frac{\eta L_2 \sigma^2}{2B\mu},$$

где L_2 – константа Липшица градиента такая, что $\|\nabla f(x) - \nabla f(y)\| \leq L_2\|x - y\|$.

Докозательство. В силу L_2-гладкости f и выбора размера шага $\eta \leq \frac{1}{(M+1)L_2}$ имеем

$$\begin{aligned}
\mathbb{E}[f(x_{k+1})] &\leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k\rangle + \\
&+ \frac{L_2}{2}\|x_{k+1} - x_k\|^2 \leq f(x_k) - \eta\langle \nabla f(x_k), \mathbb{E}[\mathbf{G}_k]\rangle + \\
&+ \frac{\eta^2 L_2}{2}\left(\mathbb{E}[\|\mathbf{G}_k - \mathbb{E}[\mathbf{G}_k]\|^2] + \mathbb{E}[\|\mathbb{E}[\mathbf{G}_k]\|^2]\right) = \\
&\overset{①}{=} f(x_k) - \eta\langle \nabla f(x_k), \nabla f(x_k) + \mathbf{b}(x_k)\rangle + \\
&+ \frac{\eta^2 L_2}{2}\left(\mathbb{E}[\|\mathbf{n}(x_k, \xi)\|^2] + \mathbb{E}[\|\nabla f(x_k) + \mathbf{b}(x_k)\|^2]\right) \leq \\
&\overset{②}{\leq} f(x_k) - \eta\langle \nabla f(x_k), \nabla f(x_k) + \mathbf{b}(x_k)\rangle + \\
&+ \frac{\eta^2 L_2}{2}\left((M+1)\mathbb{E}[\|\nabla f(x_k) + \mathbf{b}(x_k)\|^2] + \sigma^2\right) = \\
&= f(x_k) + \frac{\eta}{2}\left(\pm\|\nabla f(x_k)\|^2 - \right. \\
&\left. - 2\langle \nabla f(x_k), \nabla f(x_k) + \mathbf{b}(x_k)\rangle + \|\nabla f(x_k) + \mathbf{b}(x_k)\|^2\right) + \\
&+ \frac{\eta^2 L_2}{2}\sigma^2 = f(x_k) + \frac{\eta}{2}\left(-\|\nabla f(x_k)\|^2 + \|\mathbf{b}(x_k)\|^2\right) + \\
&+ \frac{\eta^2 L_2}{2}\sigma^2 \overset{②,③}{\leq} (1 - \eta\mu)\left(f(x_k) - f^*\right) + \\
&+ \frac{\eta\zeta^2}{2} + \frac{\eta^2 L_2}{2}\sigma^2 + f^*,
\end{aligned} \tag{A.3}$$

где в ① мы использовали Определение 2, в ② мы использовали Предположение 4, а в ③ мы использовали Предположение 5.

Применяя рекурсию к (A.3) и добавляя батчирование (с размером пакета B), получаем:

$$\mathbb{E}[f(x_N)] - f^* \leq \leq (1-\eta\mu)^N \left(f(x_0) - f^*\right) + \frac{\zeta^2}{2\mu} + \frac{\eta L_2 \sigma^2}{2B\mu}. \tag{A.4}$$

□

Лемма 4. *Пусть выполняются предположения 1–5 и $f(x, y)$ удовлетворяет условию двустороннего PL с μ_x и μ_y. Если мы запустим одну итерацию алгоритма 1 с $\tau_x^t = \tau_x \leq \frac{1}{(M+1)L}$ (L указано в лемме 2) и $\tau_y^t = \tau_y \leq \frac{1}{(M+1)L_2}$, то*

$$a_{t+1} + \lambda b_{t+1} \leq \max\{k_1, k_2\}(a_t + \lambda b_t) + \\ + \lambda\left(\tau_y^2 L_2 \frac{\sigma^2}{B} + \tau_y \zeta^2\right),$$

где

$$k_1 := 1 - \mu_x\tau_x[1 + \lambda(1 - \mu_y\tau_y)], \tag{A.5}$$

$$k_2 := 1 + \frac{L_2^2\tau_x}{\mu_y\lambda} - \mu_y\tau_y + \sigma^2\frac{L_2^2}{\mu_y}\tau_x - \tau_x\tau_y L_2^2\sigma^2. \tag{A.6}$$

Докозательство. Поскольку g является L-гладкой по лемме 2 и выбрав размер шага такой, что $\tau_x \leq \frac{1}{(M+1)L}$, мы имеем:

$$\begin{aligned}
\mathbb{E}[g(x_{k+1})] &\leq g(x_k) + \langle \nabla g(x_k), x_{k+1} - x_k\rangle + \\
&+ \frac{L}{2}\|x_{k+1} - x_k\|^2 \leq g(x_k) - \tau_x\langle \nabla g(x_k), \mathbb{E}[\mathbf{G}_k]\rangle + \\
&+ \frac{\tau_x^2 L}{2}\left(\mathbb{E}[\|\mathbf{G}_k - \mathbb{E}[\mathbf{G}_k]\|^2] + \mathbb{E}[\|\mathbb{E}[\mathbf{G}_k]\|^2]\right) =
\end{aligned}$$

$$\overset{①}{=} g(x_k) - \tau_x\langle\nabla g(x_k), \nabla_x f(x_k, y_k) + \mathbf{b}(x_k)\rangle + \frac{\tau_x^2 L}{2}\times$$
$$\times(\mathbb{E}[\|\mathbf{n}(x_k, y_k, \xi)\|^2] + \mathbb{E}[\|\nabla g(x_k) + \mathbf{b}_x(x_k, y_k)\|^2]) \le$$
$$\overset{②}{\le} g(x_k) - \tau_x\langle\nabla g(x_k), \nabla_x f(x_k, y_k) + \mathbf{b}_x(x_k, y_k)\rangle +$$
$$+\frac{\tau_x^2 L}{2}((M+1)\mathbb{E}[\|\nabla_x f(x_k, y_k) + \mathbf{b}_x(x_k, y_k)\|^2] + \quad \text{(A.7)}$$
$$+\sigma^2) = g(x_k) + \frac{\tau_x}{2}(\pm\|\nabla g(x_k)\|^2 -$$
$$-2\langle\nabla g(x_k), \nabla f(x_k, y_k) + \mathbf{b}_x(x_k, y_k)\rangle +$$
$$+\|\nabla_x f(x_k, y_k) + \mathbf{b}_x(x_k, y_k)\|^2) + \frac{\tau_x^2 L}{2}\sigma^2 =$$
$$= g(x_k) + \frac{\tau_x}{2}(-\|\nabla g(x_k)\|^2 +$$
$$+\|-\nabla g(x_k) + \mathbf{b}_x(x_k, y_k) + \nabla_x f(x_k, y_k)\|^2) + \frac{\tau_x^2 L}{2}\sigma^2,$$

где в ① мы использовали Определение 2, в ② мы использовали Предположение 4.

Теперь достаточно выразить $\|g(x_t)\|^2$ и $\|\nabla_x f(x_t, y_t) - \nabla g(x_t)\|^2$ через a_t и b_t. Используя лемму 2, мы имеем:

$$\|\nabla_x f(x_t, y_t) - \nabla g(x_t)\|^2 =$$
$$= \|\nabla_x f(x_t, y_t) - \nabla_x f(x_t, y^*(x_t))\|^2 \le \quad \text{(A.8)}$$
$$\le L_2^2\|y^*(x_t) - y_t\|^2$$

для любого $y^*(x_t) \in \arg\max_y f(x_t, y)$. Теперь можно зафиксировать $y^*(x_t)$ как проекцию y_t на множество $\arg\max_y f(x_t, y)$. Поскольку $-f(x_t, \cdot)$ удовлетворяет условию PL с μ_y, а лемма 1, следовательно, указывает, что функция также удовлетворяет условию квадратичного роста с μ_y, т.е.

$$\|y^*(x_t) - y_t\|^2 \le \frac{2}{\mu_y}[g(x_t) - f(x_t, y_t)], \quad \text{(A.9)}$$

вместе с (A.8), мы получаем

$$\|\nabla_x f(x_t, y_t) - \nabla g(x_t)\|^2 \le \frac{2L_2^2}{\mu_y}[g(x_t) - f(x_t, y_t)]. \quad \text{(A.10)}$$

Поскольку g удовлетворяет условию PL с μ_x,

$$\|\nabla g(x_t)\|^2 \ge 2\mu_x[g(x_t) - g^*]. \quad \text{(A.11)}$$

Взяв математическое ожидание у обеих сторон A.7 и подставляя A.10, A.11, мы получаем

$$a_{t+1} \le (1 - \tau_x\mu_x)a_t + \tau_x\frac{L_2^2}{\mu_y}b_t + \frac{\tau_x}{2}\|\mathbf{b}_x\|^2 \quad \text{(A.12)}$$

Поскольку $-f(x_{t+1}, y)$ L_2-гладкая и μ_y-PL, по неравенству (A.3) из леммы 3 при $\tau_y \le \frac{1}{(M+1)L_2}$ имеем

$$\mathbb{E}[g(x_{t+1}) - f(x_{t+1}, y_{t+1})] \le (1 - \mu_y\tau_y)\mathbb{E}[g(x_{t+1}) -$$
$$-f(x_{t+1}, y_t)] + \frac{\tau_y\zeta^2}{2} + \frac{\tau_y^2 L_2}{2}\sigma^2 \le \quad \text{(A.13)}$$
$$\le (1 - \mu_y\tau_y)\mathbb{E}[g(x_t) - f(x_t, y_t) + f(x_t, y_t) -$$
$$-f(x_{t+1}, y_t) + g(x_{t+1}) - g(x_t)] + \frac{\tau_y\zeta^2}{2} + \frac{\tau_y^2 L_2}{2}\sigma^2$$

Используя выкладки из леммы 3 можно ограничить $f(x_t, y_t) - f(x_{t+1}, y_t)$ следующим образом

$$f(x_t, y_t) - f(x_{t+1}, y_t) \le \frac{\tau_x}{2}\zeta^2 + \frac{\tau_x^2 L_2}{2}\sigma^2. \quad \text{(A.14)}$$

Также из A.12,

$$\mathbb{E}[g(x_{t+1}) - g(x_t)] \le -\tau_x\mu_x a_t + \frac{\tau_x L_2^2}{\mu_y}b_t + \frac{\tau_x}{2}\zeta^2. \quad \text{(A.15)}$$

Комбинируя (A.13), (A.14) и (A.15),

$$\mathbb{E}[g(x_{t+1}) - f(x_{t+1}, y_{t+1})] \le (1 - \mu_y\tau_y)(-\tau_x\mu_x a_t +$$
$$+\left(1 + \frac{\tau_x L_2^2}{\mu_y}\sigma^2\right)b_t) + (1 - \mu_y\tau_y)\left(\tau_x\zeta^2 + \frac{\tau_x^2 L_2}{2}\sigma^2\right) + \quad \text{(A.16)}$$
$$+\frac{\tau_y^2 L_2}{2}\sigma^2 + \frac{\tau_y\zeta^2}{2} \le (1 - \mu_y\tau_y)\times$$
$$\times\left(-\tau_x\mu_x a_t + \left(1 + \frac{\tau_x L_2^2}{\mu_y}\sigma^2\right)b_t\right) + \tau_y^2 L_2\sigma^2 + \frac{3}{4}\tau_y\zeta^2,$$

где в последнем неравенстве учитывается, что τ_x меньше чем τ_y. Даже можно предполагать, что $\tau_x \le \frac{\lambda}{2}\tau_y$. Комбинируя (A.12) и (A.16), имеем $\forall\lambda > 0$

$$a_{t+1} + \lambda b_{t+1} \le a_t[1 - \mu_x\tau_x - \lambda(1 - \mu_y\tau_y)\mu_x\tau_x] +$$
$$+\lambda b_t\left[1 + \frac{L_2^2\tau_x}{\mu_y\lambda} - \mu_y\tau_y + \frac{\tau_x L_2^2}{\mu_y}\sigma^2 - \tau_x\tau_y L_2^2\sigma^2\right] +$$
$$+\lambda\left(\tau_y^2 L_2\sigma^2 + \tau_y\zeta^2\right).$$

Добавляя батчирование (с размером батча B), получим:

$$a_{t+1} + \lambda b_{t+1} \le a_t[1 - \mu_x\tau_x - \lambda(1 - \mu_y\tau_y)\mu_x\tau_x] +$$
$$+\lambda b_t\left[1 + \frac{l^2\tau_x}{\mu_y\lambda} - \mu_y\tau_y + \frac{\tau_x L_2^2}{\mu_y}\frac{\sigma^2}{B} - \tau_x\tau_y L_2^2\frac{\sigma^2}{B}\right] + \quad \text{(A.17)}$$
$$+\lambda\left(\tau_y^2 L_2\frac{\sigma^2}{B} + \tau_y\zeta^2\right).$$

□

Доказательство теоремы 1.

Доказательство. В условиях леммы 4 $\tau_x^t = \tau_x$ и $\tau_y^t = \tau_y, \forall t$ нужно только выбрать τ_x, τ_y, λ, чтобы $k_1, k_2 < 1$. Здесь сначала выбирается $\lambda = 1/10$. Затем

$$k_1 = 1 - \mu_x[\tau_x + \lambda(1 - \mu_y\tau_y)\tau_x] \le 1 - \tau_x\mu_x. \quad \text{(A.18)}$$

Также,

$$k_2 = 1 + \frac{\tau_x L_2^2}{\mu_y\lambda} - \mu_y\tau_y + \frac{\tau_x L_2^2}{\mu_y}\frac{\sigma^2}{B} - \tau_x\tau_y L_2^2\frac{\sigma^2}{B} =$$
$$= 1 - \frac{\tau_x L_2^2}{\mu_y}\left\{\frac{\mu_y^2\tau_y}{\tau_x L_2^2} - \frac{1}{\lambda} - \frac{\sigma^2}{B}(1 - \mu_y\tau_y)\right\} \le \quad \text{(A.19)}$$
$$\le 1 - \frac{\tau_x L_2^2}{\mu_y},$$

где в последнем неравенстве подставляется λ и используется $\frac{\mu_y^2\tau_y}{\tau_x L_2^2} \ge 12$ за счет выбора τ_x. Выбирая большое B порядка d^2, можно сделать $\frac{\sigma^2}{B} \le 1$. Обратите внимание, что $\tau_x\mu_x < \frac{l^2\tau_x}{\mu_y}$, потому что $(\tau_x\mu_x)/\left(\frac{l^2\tau_x}{\mu_y}\right) = \frac{\mu_x\mu_y}{l^2} < 1$. Пусть $P_t := a_t + \frac{1}{10}b_t$, и по теореме 4,

$$P_{t+1} \le (1 - \tau_x\mu_x)P_t + \frac{1}{10}\left(\tau_y^2 L_2\frac{\sigma^2}{B} + \tau_y\zeta^2\right).$$

С помощью некоторых простых вычислений, получим:

$$P_t \le (1 - \mu_x\tau_x)^t P_0 + \frac{\tau_y^2 L_2\frac{\sigma^2}{B} + \tau_y\zeta^2}{10\mu_x\tau_x}. \quad \text{(A.20)}$$

Проверка, что $\tau_x \le \frac{1}{(M+1)L}$ осуществляется за счет того, что $\tau_x \le \frac{\mu_y^2\tau_y}{12L_2^2} \le \frac{\mu_y^2}{12(M+1)L_2^3} \le \frac{\mu_y}{2(M+1)L_2^2}$ и $L = L_2 + \frac{L_2^2}{\mu_y} \le \frac{2L_2^2}{\mu_y}$.

□

Доказательства для методов нулевого порядка.

В этом разделе мы доказываем леммы для разных случаев вида задачи. В следующих леммах мы не привязываемся к седловой задаче, а больше рассматриваем ядерную аппроксимацию градиента, поэтому для следующих лемм рассмотрим задачу $\min_{x\in\mathbb{R}} f(x)$

Лемма 5 (Сведение интеграла по области к интегралу по поверхности). *Пусть D – открытое связное подмножество $\mathbb{R}$ с кусочно-гладкой границей ∂D, ориентированное по внешней единичной нормали $\mathbf{n} = (n_1, \ldots, n_m)^\top$. Пусть f – гладкая функция в $D \cup \partial D$, тогда*

$$\int_D \nabla f(x)dx = \int_{\partial D} f(x)\mathbf{n}(x)dS(x).$$

Remark 4. *Мы ссылаемся на [25, раздел 12.3.2, определения 4 и 5] для определения кусочно-гладких поверхностей и их ориентации соответственно.*

Лемма 6. *Пусть $f : \mathbb{R}^d \to \mathbb{R}$ – непрерывно дифференцируемая функция. Пусть $r, \mathbf{h}, \mathbf{e}$ равномерно распределены на $[-1,1], \mathcal{B}_2^d$ и $\mathcal{S}^d$ соответственно. Тогда для любого $\gamma > 0$ имеем*

$$\mathbb{E}[\nabla f(x + \gamma r\mathbf{h})rK(r)] = \frac{d}{\gamma}\mathbb{E}[f(x + \gamma r\mathbf{e})\mathbf{e}K(r)].$$

Доказательство. Зафиксируем $r \in [-1,1]\backslash\{0\}$. Определим $\phi : \mathbb{R}^d \to \mathbb{R}$ как $\phi(\mathbf{h}) = f(x + \gamma r\mathbf{h})K(r)$ и заметим, что $\nabla\phi(\mathbf{h}) = \gamma r\nabla f(x + \gamma r\mathbf{h})K(r)$. Следовательно, у нас есть

$$\mathbb{E}[\nabla f(x + \gamma r\mathbf{h})K(r)|r] = \frac{1}{\gamma r}\mathbb{E}[\nabla\phi(\mathbf{h})|r] =$$
$$= \frac{d}{\gamma r}\mathbb{E}[\phi(\mathbf{e})\mathbf{e}|r] = \frac{d}{\gamma r}K(r)\mathbb{E}[f(x + \gamma r\mathbf{e})\mathbf{e}|r],$$

где второе равенство получается из теоремы 5. Доказательство завершается умножением на r с обеих сторон, использованием того факта, что r следует за непрерывным распределением, и принятием полного матожидания.

□

B. Доказательство теоремы 2

Лемма 7 (Смещение ядерной аппроксимации). *Пусть выполняются предположения* 1–3. *Пусть x_t и $\mathbf{G}(x_t, \mathbf{e})$ определены алгоритмом* 1 *в момент времени $t \ge 1$ с аппроксимацией градиента* (4.2) *для оракула нулевого порядка* (4.1). *Тогда*,

$$\|\mathbb{E}[\mathbf{G}(x_t, \xi, \mathbf{e})|x_t] - \nabla f(x_t)\| \le$$
$$\le \kappa_\beta\frac{L_\beta}{(l-1)!}\cdot\frac{d}{d+\beta-1}\gamma^{\beta-1} + \kappa_\beta d\frac{\Delta}{\gamma}, \quad \text{(B.1)}$$

где мы напоминаем, что $l = \lfloor\beta\rfloor$.

Доказательство леммы 7. Используя лемму 6, тот факт, что $\int_{-1}^{1} rK(r)dr = 1$, и вариационное представление евклидовой нормы, мы можем написать

$$\|\mathbb{E}[\mathbf{G}(x_t, \xi, \mathbf{e})|x_t] - \nabla f(x_t)\| =$$

$$= \sup_{\mathbf{v}\in\mathcal{S}^d} \mathbb{E}[(\nabla_{\mathbf{v}} f(x+\gamma r\mathbf{h},\xi) - \nabla_{\mathbf{v}} f(x,\xi) + \\ + \frac{d}{2\gamma r}(\delta(x+\gamma r\mathbf{h}) - \delta(x-\gamma r\mathbf{h})))rK(r)] \le \\ \le \sup_{\mathbf{v}\in\mathcal{S}^d} \mathbb{E}[(\nabla_{\mathbf{v}} f(x+\gamma r\mathbf{h}) - \nabla_{\mathbf{v}} f(x))rK(r)] + \kappa_\beta d\frac{\Delta}{\gamma}, \tag{B.2}$$

где мы напоминаем, что $\mathbf{h}$ равномерно распределена на $\mathcal{B}_2^d$. Так как $f(x)$ удовлетворяет условию Гельдера с константами β и L_β, то для любого $\mathbf{v}\in\mathcal{S}^d$ направленный градиент $\nabla_{\mathbf{v}} f(\cdot)$ удовлетворяет условию Гельдера с константами $\beta-1$ и L_β. Таким образом справедливо следующее разложение Тейлора

$$\nabla_{\mathbf{v}} f(x_t+\gamma r\mathbf{h}) = \nabla_{\mathbf{v}} f(x_t) + \\ + \sum_{1\le|\mathbf{m}|\le l-1} \frac{(r\gamma)^{|\mathbf{m}|}}{\mathbf{m}!} D^{\mathbf{m}}\nabla_{\mathbf{v}} f(x_t)(\mathbf{h})^{\mathbf{m}} + R(\gamma r\mathbf{h}), \tag{B.3}$$

где остаточный член $R(\cdot)$ удовлетворяет условию $|R(x)| \le \frac{L_\beta}{(l-1)!}\|x\|^{\beta-1}$.

Подставляя уравнение (B.3) в уравнение (B.2) и используя свойства "обнуления" ядра K, получаем, что

$$\|\mathbb{E}[\mathbf{G}(x_t,\xi,\mathbf{e})\,|\,x_t] - \nabla f(x_t)\| \le \\ \le \kappa_\beta\gamma^{\beta-1}\frac{L_\beta}{(l-1)!}\mathbb{E}\|\mathbf{h}\|^{\beta-1} = \\ = \kappa_\beta\gamma^{\beta-1}\frac{L_\beta}{(l-1)!}\frac{d}{d+\beta-1} + \kappa_\beta d\frac{\Delta}{\gamma},$$

где последнее равенство получается из того, что $\mathbb{E}\|\mathbf{h}\|^q = \frac{d}{d+q}$ для любого $q \ge 0$.

□

Лемма 8 (Дисперсия ядерной аппроксимации). *Пусть выполняются предположения* 1–3. *Пусть* x_t *и* $\mathbf{G}(x_t,\xi,\mathbf{e})$ *определены алгоритмом* 1 *с аппроксимацией градиента* (4.2) *для оракула нулевого порядка* (4.1). *Предположим, что* $f\in\mathcal{F}_2(L_2)$, *тогда если* $d\ge 2$

$$\mathbb{E}\|\mathbf{G}(x_t,\xi,\mathbf{e})\|^2 \le \frac{d^2\kappa}{d-1}\mathbb{E}\left[\|\nabla f(x_t)\| + L_2\gamma^2\right] + \frac{d^2\Delta^2\kappa}{\gamma^2},$$

где мы вспоминаем, что $\kappa = \int_{-1}^{1} K^2(r)dr$.

Результат леммы 8 может быть дополнительно упрощен как

$$\mathbb{E}\|\mathbf{G}(x_t,\xi,\mathbf{e})\|^2 \le 4d\kappa\mathbb{E}\|\nabla f(x_t)\|^2 + \\ + 4d\kappa L_2^2\gamma^2 + \frac{d^2\Delta^2\kappa}{\gamma^2}, \quad d\ge 2. \tag{B.4}$$

Доказательство леммы 8. Для простоты мы опускаем индекс t у всех величин. Распишем второй момент следующей величины.

$$\mathbb{E}\|\mathbf{G}(x,\xi,\mathbf{e})\|^2 = \\ = \frac{d^2}{4\gamma^2}\mathbb{E}[(f(x+\gamma r\mathbf{e},\xi) - f(x-\gamma r\mathbf{e},\xi) + \\ + (\delta(x+\gamma r\mathbf{e}) - \delta(x-\gamma r\mathbf{e})))^2K^2(r)] \le \tag{B.5}$$

$$\le \frac{d^2}{4\gamma^2}(\mathbb{E}[(f(x+\gamma r\mathbf{e}) - \\ - f(x-\gamma r\mathbf{e})^2K^2(r)] + 4\kappa\Delta^2).$$

В дальнейшем все возникающие ожидания следует понимать условно на x_t. Обратите внимание, что поскольку $\mathbb{E}[f(x+hr\mathbf{e}) - f(x-hr\mathbf{e})|r] = 0$ и $f\in\mathcal{F}_2(L_2)$, то используя неравенство Виртингера–Пуанкаре [22, 23], см. Eq. (3.1) или теорему 2 соответственно получаем

$$\mathbb{E}\left[(f(x+hr\mathbf{e}) - f(x-hr\mathbf{e}))^2\,|\,r\right] \le \\ \le \frac{h^2}{d-1}\mathbb{E}\left[\|\nabla f(x+hr\mathbf{e}) + \nabla f(x-hr\mathbf{e})\|^2\,|\,r\right]. \tag{B.6}$$

Так как $f\in\mathcal{F}_2(L_2)$, то из неравенства треугольника далее следует, что

$$\mathbb{E}\left[\|\nabla f(x+hr\mathbf{e}) + \nabla f(x-hr\mathbf{e})\|^2\,|\,r\right] \le \\ \le 4(\|\nabla f(x)\| + L_2\gamma)^2. \tag{B.7}$$

В заключение мы подставим приведенную выше оценку в уравнение (B.6) и примем во внимание уравнение (B.5).

□

Теперь мы можем вычислить шум и смещение ядерной аппроксимации:

$$M = 4d\beta^3 \qquad \sigma^2 = 4d\beta^3L_2\gamma^2 + \frac{d^2\Delta^2\beta^3}{\gamma^2} \tag{B.8}$$

$$\zeta^2 = \beta^2\left(\frac{L_\beta}{(l-1)!}\frac{d}{d+\beta-1}\gamma^{\beta-1} + d\frac{\Delta}{\gamma}\right)^2 \tag{B.9}$$

Или же более грубая оценка на смещение:

$$\zeta^2 = \beta^2\left(\frac{L_\beta}{(l-1)!}\right)^2\gamma^{2\beta-2} + \beta^2d^2\frac{\Delta^2}{\gamma^2}$$

Теперь мы можем оценить скорость сходимости для ядерной аппроксимации, подставив значения найденных констант в итоговую оценку для сходимости:

$$P_t \le (1-\mu_x\tau_x)^tP_0 + \frac{\tau_y^2L_2\frac{\sigma^2}{B} + \tau_y\zeta^2}{10\mu_x\tau_x} = (1-\mu_x\tau_x)^tP_0 +$$

$$+\frac{12}{5B}\frac{L_2^3 d\gamma^2}{\mu_x\mu_y^2}+\frac{3}{5B}\frac{L_2^2 d^2\Delta^2}{\mu_x\mu_y^2\gamma^2}+$$
$$+\frac{12}{5}\frac{L_2^2\beta^2}{\mu_x\mu_y^2}\left(\frac{L_\beta}{(l-1)!}\right)^2\gamma^{2\beta-2}+\frac{12}{5}\frac{L_2^2\beta^2 d^2\Delta^2}{\mu_x\mu_y^2\gamma^2}= \quad \text{(B.10)}$$
$$=\mathbb{O}\left(\frac{L_2^2 d\gamma^2}{B\mu_x\mu_y^2}+\frac{L_2^2\beta^2}{\mu_x\mu_y^2}\left(\frac{L_\beta}{(l-1)!}\right)^2\gamma^{2\beta-2}+\frac{L_2^2\beta^2 d^2\Delta^2}{\mu_x\mu_y^2\gamma^2}\right).$$

Здесь мы подставляем значения для $\tau_y=\frac{1}{(M+1)L_2}$ и $\tau_x=\frac{\mu_y^2\tau_y}{12L_2^2}$.

Поскольку B можно взять большим, второе и третье слагаемые отвечают за асимптоту. Находим оптимальный параметр сглаживания γ, минимизирующий последние два члена:

$$P_t=\mathbb{O}\left(\frac{L_2^2\beta^2 d^2}{\mu_x\mu_y^2}\left(\frac{L_\beta}{(l-1)!}\right)^{\frac{2}{\beta}}\left(\frac{\beta-1}{d+\beta-1}\right)^{\frac{2}{\beta}}\Delta^{\frac{2(\beta-1)}{\beta}}\right)= \quad \text{(B.11)}$$
$$=\mathbb{O}\left(\frac{1}{\mu_x\mu_y^2}d^{\frac{2(\beta-1)}{\beta}}\Delta^{\frac{2(\beta-1)}{\beta}}\right),$$

где $\gamma_k=\left(\frac{(l-1)!}{L_\beta}\frac{d+\beta-1}{\beta-1}\Delta\right)^{1/\beta}$ – параметр оптимального сглаживания. Тогда из (B.11) мы можем найти максимальный уровень шума, предполагая, что $(d\Delta)^{\frac{2(\beta-1)}{\beta}}\le\varepsilon$, для $\varepsilon>0$ тогда имеем

$$\Delta=\mathbb{O}\left((\mu_x\mu_y^2)^{\frac{\beta}{2(\beta-1)}}\varepsilon^{\frac{\beta}{2(\beta-1)}}d^{-1}\right).$$

При таком максимальном шуме $\gamma_k=\mathbb{O}\left((\mu_x\mu_y^2\varepsilon)^{\frac{1}{2(\beta-1)}}\right)$. Таким образом мы гарантируем, что второе и третье слагаемые в (B.10) меньше ε (с точностью до константы) при выбранных параметрах. Для уменьшения количества итераций, мы выберем размер батча порядка $\beta^3 d$. Определим минимальное количество итераций. Это можно сделать, решив неравенство:

$$(1-\mu_x\tau_x)^N P_0\le\varepsilon$$

Откуда мы получим минимальное число итераций

$$N\ge\frac{1}{\tau_x\mu_x}\ln\frac{P_0}{\varepsilon}=12\frac{(\beta^3 d/B+1)L_2^3}{\mu_x\mu_y^2}\ln\frac{P_0}{\varepsilon}=$$
$$=\mathbb{O}\left(\mu_x^{-1}\mu_y^{-2}\ln\frac{1}{\varepsilon}\right),$$

где во втором неравенстве мы использовали то, что $\tau_x=\frac{\mu_y^2\tau_y}{12L_2^2}$, $\tau_y=\frac{1}{(M+1)L_2}$ и $M=\mathbb{O}(\beta^3 d/B)$, $d=\max(d_x,d_y)$ При достаточно большом B порядка $\beta^3 d$ зависимость от размерности пропадает.

Оракульная сложность получается из итерационной путем домножения на размер батча, то есть:

$$T=\mathbb{O}\left(\beta^3 d\mu_x^{-1}\mu_y^{-2}\ln\frac{1}{\varepsilon}\right).$$

Таким образом все слагаемые в формуле (B.10) меньше ε.

C. Доказательство теоремы 3

Лемма 9 (Смещение ядерной аппроксимации). *Пусть выполняются предположения* 1–5. *Пусть* x_t *и* $\mathbf{G}(x_t,\xi,\mathbf{e})$ *определены алгоритмом* 1 *в момент времени* $t\ge 1$ *с аппроксимацией градиента* (4.4) *для оракула нулевого порядка* (4.3). *Тогда*,

$$\|\mathbb{E}[\mathbf{G}(x_t,\xi,\mathbf{e})\,|\,x_t]-\nabla f(x_t)\|\le$$
$$\le\kappa_\beta\frac{L_\beta}{(l-1)!}\cdot\frac{d}{d+\beta-1}\gamma^{\beta-1}, \quad \text{(C.1)}$$

где мы напоминаем, что $l=\lfloor\beta\rfloor$.

Доказательство леммы 9. Используя лемму 6, тот факт, что $\int_{-1}^{1}rK(r)dr=1$, и вариационное представление евклидовой нормы, мы можем написать

$$\|\mathbb{E}[\mathbf{G}(x_t,\xi,\mathbf{e})\,|\,x_t]-\nabla f(x_t)\|=$$
$$=\sup_{\mathbf{v}\in\mathcal{S}^d}\mathbb{E}[(\nabla_{\mathbf{v}}f(x+\gamma r\mathbf{h})-\nabla_{\mathbf{v}}f(x))rK(r)], \quad \text{(C.2)}$$

где мы напоминаем, что $\mathbf{h}$ равномерно распределена на $\mathcal{B}_2^d$. Так как $f(x)$ удовлетворяет условию Гельдера с константами β и L_β, то для любого $\mathbf{v}\in\mathcal{S}^d$ направленный градиент $\nabla_{\mathbf{v}}f(\cdot)$ удовлетворяет условию Гельдера с константами $\beta-1$ и L_β. Таким образом справедливо следующее разложение Тейлора

$$\nabla_{\mathbf{v}}f(x_t+\gamma r\mathbf{h})=\nabla_{\mathbf{v}}f(x_t)+$$
$$+\sum_{1\le|\mathbf{m}|\le l-1}\frac{(r\gamma)^{|\mathbf{m}|}}{\mathbf{m}!}D^{\mathbf{m}}\nabla_{\mathbf{v}}f(x_t)(\mathbf{h})^{\mathbf{m}}+R(\gamma r\mathbf{h}), \quad \text{(C.3)}$$

где остаточный член $R(\cdot)$ удовлетворяет условию $|R(x)|\le\frac{L_\beta}{(l-1)!}\|x\|^{\beta-1}$.

Подставляя уравнение (C.3) в уравнение (C.2) и используя свойства “обнуления” ядра K, получаем, что

$$\|\mathbb{E}[\mathbf{G}(x_t,\xi,\mathbf{e})\,|\,x_t] - \nabla f(x_t)\| \le$$
$$\le \kappa_\beta \gamma^{\beta-1}\frac{L_\beta}{(l-1)!}\mathbb{E}\|\mathbf{h}\|^{\beta-1} = \kappa_\beta\gamma^{\beta-1}\frac{L_\beta}{(l-1)!}\frac{d}{d+\beta-1},$$

где последнее равенство получается из того, что $\mathbb{E}\|\mathbf{h}\|^q = \dfrac{d}{d+q}$ для любого $q \ge 0$.

□

Лемма 10 (Дисперсия ядерной аппроксимации). *Пусть выполняются предположения* 1–3. *Пусть* x_t *и* $\mathbf{G}(x_t,\mathbf{e})$ *определены алгоритмом* 1 *с аппроксимацией градиента* (4.4) *для оракула нулевого порядка* (4.3). *Предположим, что* $f \in \mathcal{F}_2(L_2)$, *тогда если* $d \ge 2$

$$\mathbb{E}\|\mathbf{G}(x_t,\xi,\mathbf{e})\|^2 \le \frac{d^2\kappa}{d-1}\mathbb{E}\left[\|\nabla f(x_t)\| + L_2\gamma^2\right] + \frac{d^2\tilde{\Delta}^2\kappa}{\gamma^2},$$

где мы вспоминаем, что $\kappa = \int_{-1}^{1} K^2(r)dr$.

Результат леммы 10 может быть дополнительно упрощен как

$$\mathbb{E}\|\mathbf{G}(x_t,\xi,\mathbf{e})\|^2 \le 4d\kappa\mathbb{E}\|\nabla f(x_t)\|^2 + \\ + 4d\kappa L_2^2\gamma^2 + \frac{d^2\tilde{\Delta}^2\kappa}{\gamma^2}, \quad d \ge 2. \tag{C.4}$$

Доказательство леммы 10. Для простоты мы опускаем индекс t у всех величин. Распишем второй момент следующей величины.

$$\mathbb{E}\|\mathbf{G}(x,\xi,\mathbf{e})\|^2 = \frac{d^2}{4\gamma^2}\mathbb{E}\Big[(f(x+\gamma r\mathbf{e}) - f(x-\gamma r\mathbf{e}) + \\ + (\xi_1 - \xi_2))^2 K^2(r)\Big] \le \frac{d^2}{4\gamma^2}\Big(\mathbb{E}\Big[(f(x+\gamma r\mathbf{e}) - \\ - f(x-\gamma r\mathbf{e})^2 K^2(r)\Big] + 4\kappa\tilde{\Delta}^2\Big). \tag{C.5}$$

В дальнейшем все возникающие ожидания следует понимать условно на x_t. Обратите внимание, что поскольку $\mathbb{E}[f(x+hr\mathbf{e}) - f(x-hr\mathbf{e})|r] = 0$ и $f \in \mathcal{F}_2(L_2)$, то используя неравенство Виртингера–Пуанкаре [22, 23], см. Eq. (3.1) или теорему 2 соответственно получаем

$$\mathbb{E}\Big[(f(x+hr\mathbf{e}) - f(x-hr\mathbf{e}))^2\,|\,r\Big] \le \\ \le \frac{h^2}{d-1}\mathbb{E}\Big[\|\nabla f(x+hr\mathbf{e}) + \nabla f(x-hr\mathbf{e})\|^2\,|\,r\Big]. \tag{C.6}$$

Так как $f \in \mathcal{F}_2(L_2)$, то из неравенства треугольника далее следует, что

$$\mathbb{E}\Big[\|\nabla f(x+hr\mathbf{e}) + \nabla f(x-hr\mathbf{e})\|^2\,|\,r\Big] \le \\ \le 4(\|\nabla f(x)\| + L_2\gamma)^2. \tag{C.7}$$

В заключение мы подставим приведенную выше оценку в уравнение (C.6) и примем во внимание уравнение (C.5).

□

Теперь мы можем вычислить шум и смещение ядерной аппроксимации:

$$M = 4d\beta^3 \quad \sigma^2 = 4d\beta^3 L_2\gamma^2 + \frac{d^2\tilde{\Delta}^2\beta^3}{\gamma^2} \tag{C.8}$$

$$\zeta^2 = \beta^2\left(\frac{L_\beta}{(l-1)!}\frac{d}{d+\beta-1}\gamma^{\beta-1}\right)^2 \tag{C.9}$$

Теперь мы можем оценить скорость сходимости для ядерной аппроксимации, подставив значения найденных констант в итоговую оценку для сходимости:

$$P_t \le (1-\mu_x\tau_x)^t P_0 + \frac{\tau_y^2 L_2\dfrac{\sigma^2}{B} + \tau_y\zeta^2}{10\mu_x\tau_x} = (1-\mu_x\tau_x)^t P_0 + \\ + \frac{12L_2^3 d\gamma^2}{5B\mu_x\mu_y^2} + \frac{3L_2^2 d^2\tilde{\Delta}^2}{5B\mu_x\mu_y^2\gamma^2} + \frac{12L_2^2\beta^2}{5\mu_x\mu_y^2}\left(\frac{L_\beta}{(l-1)!}\right)^2\gamma^{2\beta-2} = \\ = \mathbb{O}\left(\frac{L_2^3\gamma^2}{B\mu_x\mu_y^2} + \frac{L_2^2 d\tilde{\Delta}^2}{B\mu_x\mu_y^2\gamma^2} + \frac{L_2^2\beta^2}{\mu_x\mu_y^2}\left(\frac{L_\beta}{(l-1)!}\right)^2\gamma^{2\beta-2}\right). \tag{C.10}$$

Здесь мы подставляем значения для $\tau_y = \dfrac{1}{(M+1)L_2}$ и $\tau_x = \dfrac{\mu_y^2\tau_y}{12L_2^2}$. Найдем ограничения на параметр сглаживания γ, минимизируя смещение аппроксимации. Получим оптимальный параметр $\gamma_k = \sqrt[4]{\dfrac{d\tilde{\Delta}^2}{4L_2}}$. Максимальный уровень шума найдем из последнего слагаемого в (C.10). Получим $\tilde{\Delta} = \mathbb{O}\left(d^{\frac{-1}{2}}(\mu_x\mu_y^2\varepsilon)^{\frac{1}{\beta-1}}\right)$. Тогда параметр сглаживания примет следующий вид $\gamma_k = \mathbb{O}\left((\mu_x\mu_y^2\varepsilon)^{\frac{1}{2(\beta-1)}}\right)$. При таких параметрах последнее слагаемое меньше ε.

При выборе B порядка $\beta^3 d$ первые два слагаемых в (C.10) будут меньше ε. Определим минимальное количество итераций. Это можно сделать, решив неравенство:

$$(1-\mu_x\tau_x)^N P_0 \le \varepsilon$$

Откуда мы получим минимальное число итераций

$$N \geq \frac{1}{\tau_x \mu_x} \ln \frac{P_0}{\varepsilon} =$$

$$= 12 \frac{\left(\beta^3 d/B + 1\right) L_2^3}{\mu_x \mu_y^2} \ln \frac{P_0}{\varepsilon} = \mathbb{O}\left(\mu_x^{-1} \mu_y^{-2} \ln \frac{1}{\varepsilon}\right),$$

где во втором неравенстве мы использовали то, что $\tau_x = \frac{\mu_y^2 \tau_y}{12 L_2^2}$, $\tau_y = \frac{1}{(M+1)L_2}$ и $M = \mathbb{O}(\beta^3 d/B)$, $d = \max(d_x, d_y)$. При достаточно большом B порядка $\beta^3 d$ зависимость от размерности пропадает.

Оракульная сложность получается из итерационной, путем домножения на размер батча, то есть:

$$T = \mathbb{O}\left(\beta^3 d \mu_x^{-1} \mu_y^{-2} \ln \frac{1}{\varepsilon}\right).$$

При таких параметрах алгоритм 1 с градиентной аппроксимацией (4.4) в данной модели безградиентного оракула (4.3) сходится с требуемой точностью.

ИСТОЧНИК ФИНАНСИРОВАНИЯ

Работа А.М. Райгородского в разделах 1–3 была выполнена при финансовой поддержке гранта ведущих научных школ НШ775.2022.1.1, в разделах 4–6 выполнена за счет гранта Российского научного фонда (проект № 21-71-30005), https://rscf.ru/project/21-71-30005/.

СПИСОК ЛИТЕРАТУРЫ

1. *Heaton J.* Ian Goodfellow, Yoshua Bengio, and Aaron Courville: Deep learning: The MIT Press, 2016, 800 pp, ISBN: 0262035618 // Genetic programming and evolvable machines. 2018. V. 19. № 1–2. P. 305–307.
2. *Dai B. et al.* SBEED: Convergent reinforcement learning with nonlinear function approximation // International Conference on Machine Learning. PMLR, 2018. P. 1125–1134.
3. *Namkoong H., Duchi J.C.* Variance-based regularization with convex objectives // Advances in neural information processing systems. 2017. V. 30.
4. *Xu L. et al.* Maximum margin clustering // Advances in neural information processing systems. 2004. V. 17.
5. *Sinha A. et al.* Certifying some distributional robustness with principled adversarial training // arXiv preprint arXiv:1710.10571. 2017.
6. *Audet C., Hare W.* Derivative-free and blackbox optimization. 2017.
7. *Rosenbrock H.H.* An automatic method for finding the greatest or least value of a function // The computer journal. 1960. V. 3. № 3. P. 175–184.
8. *Gasnikov A. et al.* Randomized gradient-free methods in convex optimization // arXiv preprint arXiv:2211.13566. 2022.
9. *Lobanov A. et al.* Gradient-Free Federated Learning Methods with l_1 and l_2-Randomization for Non-Smooth Convex Stochastic Optimization Problems // arXiv preprint arXiv:2211.10783. 2022.
10. *Gasnikov A. et al.* The power of first-order smooth optimization for black-box non-smooth problems // International Conference on Machine Learning. PMLR, 2022. P. 7241–7265.
11. *Bach F., Perchet V.* Highly-smooth zero-th order online optimization // Conference on Learning Theory. PMLR, 2016. P. 257–283.
12. *Beznosikov A., Novitskii V., Gasnikov A.* One-point gradient-free methods for smooth and non-smooth saddle-point problems // Mathematical Optimization Theory and Operations Research: 20th International Conference, MOTOR 2021, Irkutsk, Russia, July 5–10, 2021, Proceedings 20. Springer International Publishing, 2021. P. 144–158.
13. *Akhavan A., Pontil M., Tsybakov A.* Exploiting higher order smoothness in derivative-free optimization and continuous bandits // Advances in Neural Information Processing Systems. 2020. V. 33. P. 9017–9027.
14. *Polyak B.T.* Gradient methods for the minimisation of functionals // USSR Computational Mathematics and Mathematical Physics. 1963. V. 3. № 4. P. 864–878.
15. *Lojasiewicz S.* Une propriété topologique des sous-ensembles analytiques réels // Les équations aux dérivées partielles. 1963. V. 117. P. 87–89.
16. *Ajalloeian A., Stich S.U.* On the convergence of SGD with biased gradients // arXiv preprint arXiv:2008.00051. 2020.
17. *Lobanov A., Gasnikov A., Stonyakin F.* Highly Smoothness Zero-Order Methods for Solving Optimization Problems under PL Condition // arXiv preprint arXiv:2305.15828. 2023.
18. *Yue P., Fang C., Lin Z.* On the Lower Bound of Minimizing Polyak-Łojasiewicz functions // arXiv preprint arXiv:2212.13551. 2022.
19. *Yang J., Kiyavash N., He N.* Global convergence and variance-reduced optimization for a class of nonconvex-nonconcave minimax problems // arXiv preprint arXiv:2002.09621. 2020.
20. *Akhavan A. et al.* Gradient-free optimization of highly smooth functions: improved analysis and a new algorithm // arXiv preprint arXiv:2306.02159. 2023.
21. *Nouiehed M. et al.* Solving a class of non-convex min-max games using iterative first order methods // Advances in Neural Information Processing Systems. 2019. V. 32.
22. *Osserman R.* The isoperimetric inequality // Bulletin of the American Mathematical Society. 1978. V. 84. № 6. P. 1182–1238.
23. *Beckner W.* A generalized Poincaré inequality for Gaussian measures // Proceedings of the American Mathematical Society. 1989. V. 105. № 2. P. 397–400.
24. *Karimi H., Nutini J., Schmidt M.* Linear convergence of gradient and proximal-gradient methods under the polyak-E,ojasiewicz condition // Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Ita-

ly, September 19–23, 2016, Proceedings, Part I 16. Springer International Publishing, 2016. P. 795–811.

25. *Zorich V.A., Paniagua O.* Mathematical analysis II. Berlin : Springer, 2016. V. 220.

GRADIENT-FREE ALGORITHMS FOR SOLVING STOCHASTIC SADDLE OPTIMIZATION PROBLEMS WITH THE POLYAK–LOYASIEVICH CONDITION

S. I. Sadykov[a], A. V. Lobanov[a,b], and A. M. Raigorodskii[a,c]

[a]*Moscow Institute of Physics and Technology*
Institutskiy per., 9, Moscow region, Dolgoprudny, 141701 Russia
[b]*Trusted Artificial Intelligence Research Center of ISP RAS*
Alexander Solzhenitsyn st., 25, Moscow, 109004 Russia
[c]*Caucasian Mathematical Center of the Adyghe State University*
st. Pervomaiskaya, 208, Maykop, Republic of Adygea, 385016 Russia

This paper focuses on solving a subclass of a stochastic nonconvex-concave black box optimization problem with a saddle point that satisfies the Polyak–Loyasievich condition. To solve such a problem, we provide the first, to our knowledge, gradient-free algorithm, the approach to which is based on applying a gradient approximation (kernel approximation) to the oracle-shifted stochastic gradient descent algorithm. We present theoretical estimates that guarantee a global linear rate of convergence to the desired accuracy. We check the theoretical results on a model example, comparing with an algorithm using Gaussian approximation.