Ускоренные методы нулевого порядка в гладкой выпуклой стохастической оптимизации

Хафизов Фанис
khafizov.fa@phystech.edu

Богданов Александр
bogdanov.ai@phystech.edu

Безносиков Александр
beznosikov.an@phystech.edu

Аннотация

Данная работа посвящена задаче оптимизации без доступа к градиенту целевой функции. Рассматривается безградиентный метод, требующий $\mathcal{O}(1)$ оракульных вызовов на итерацию. Применяется координатная аппроксимация в ускоренном градиентном методе и доказывается его сходимость для выпуклой задачи оптимизации с доступом к оракулу с детерминированным ограниченным по модулю шумом. В вычислительном эксперименте проводится сравнение предложенного метода с градиентным спуском и методом Нестерова с использованием той же координатной аппроксимации. Наш метод показывает результаты лучше конкурентов в случаях и детерминированного, и стохастического шума.

Ключевые слова методы нулевого порядка, координантная аппроксимация, стохастическая оптимизация

1 Введение

1.1 Мотивация

Стохастические градиентные методы являются необходимыми в решении различных оптимизационных задач. Однако в нынешних проблемах машинного обучения возникает потребность оценивать градиент, ввиду, например, дороговизны его подсчета, либо же незнания явного вида минимизируемой функции. Тогда на помощь приходят методы нулевого порядка. Так как имеется доступ только к значениям целевой функции f в различных точках $x \in \mathbb{R}^d$, то необходимо строить методы, аппроксимирующие градиент, используя конечные суммы значений целевой функции.
Одним из возможных усложнений задачи является добавление шума к оракулу: вместо $f(x)$ он будет возвращать $f_\delta(x) = f(x) + \delta(x)$. Более того, виды шума разделяются на стохастический [6] и детерминированный [5].
Другую важную роль в решении оптимизационных задач играют ускоренные методы. Они как правило имеют более быструю сходимость по сравнению со стандартными алгоритмами. Предложенный Нестеровым [7] быстрый градиентный метод является классическим примером.

1.2 Обзор литературы

Впервые метод JAGUAR аппроксимации градиента был предложен в [1], доказана сходимость для методов Франка-Вульфа и градиентного спуска для невыпуклой, выпуклой, и сильно-выпуклой задач. Также есть модификация алгоритма Франка-Вульфа для стохастического случая (рассмотрены случаи one-point feedback и two-point feedback).
В работе [2] рассмотрены ускоренные методы первого порядка в невыпуклых и сильно выпуклых задачах оптимизации, содержащих марковский шум. Там при аппроксимации градиента использовался рандомизированный размер батча. Результаты и подходы оттуда адаптированы под метод JAGUAR в данной работе.
State-of-the-art решения в области безградиентной оптимизации собраны в [3]. Представлены рандомизированные методы нулевого порядка, однако лишь для нестохастического случая.

Ускоренный безградиентный метод в гладкой выпуклой стохастической оптимизации был предложен в [4], но лишь в случае two-point feedback, который далеко не всегда реализуется на практике.

1.3 Предложения

Предлагается ускоренный метод, в качестве градиента использующий координатную аппроксимацию. Для него доказывается теорема о скорости сходимости для сильно-выпуклой гладкой функции с оракулом с детерминированным шумом. Теоретическая часть подкреплена численным экспериментом: классификация на датасете небольшого размера и минимизация квадратичной функции. Проведодится сравнение с методами градиентного спуска и Нестерова.

2 Постановка задачи

Задача подразделяется на детерминированный случай, когда f_δ зависит только от x, и стохастический, в котором есть зависимость от случайного вектора $\xi \sim \pi$.

2.1 Детерминированный случай

В этом разделе рассматривается детерминированная задача оптимизации:

$$\min_{x \in \mathbb{R}^d} f(x). \tag{1}$$

Мы полагаем, что есть доступ только к оракулу нулевого порядка, то есть, мы можем получать только значения $f(x)$, но не градиента $\nabla f(x)$. Следовательно, нужно как-то аппроксимировать градиент $\nabla f(x)$. Также предполагается, что оракул возвращает

$$f_\delta(x) = f(x) + \delta(x). \tag{2}$$

Для аппроксимации градиента используется следующая разностная схема:

$$\widetilde{\nabla} f_\delta(x) := d\frac{f_\delta(x + \tau e_i) - f_\delta(x - \tau e_i)}{2\tau} e_i, \tag{3}$$

где e_i — случайный вектор из стандартного базиса в $\mathbb{R}^d$, $i \sim U\{1, d\}$, $\tau > 0$ — достаточно мало.

2.2 Стохастический случай

В этом разделе рассматривается более общая стохастическая задача:

$$\min_{x \in \mathbb{R}^d} f(x) := \mathsf{E}_{\xi \sim \pi}[f(x, \xi)], \tag{4}$$

где ξ — случайный вектор из неизвестного распределения π. Здесь мы так же считаем, что у нас нет доступа к градиенту $\nabla f(x, \xi)$, а оракул нулевого порядка возвращает зашумленное значение функции $f_\delta(x, \xi) := f(x, \xi) + \delta(x, \xi)$.
Разделяются два вида аппроксимации градиента в стохастическом случае:

- Two-point feedback (TPF)

$$\widetilde{\nabla} f_\delta(x, \xi) := d\frac{f_\delta(x + \tau e_i, \xi) - f_\delta(x - \tau e_i, \xi)}{2\tau} e_i, \tag{5}$$

- One-point feedback (OPF)

$$\widetilde{\nabla} f_\delta(x, \xi^+, \xi^-) := d\frac{f_\delta(x + \tau e_i, \xi^+) - f_\delta(x - \tau e_i, \xi^-)}{2\tau} e_i, \tag{6}$$

где $\xi, \xi^+, \xi^- \sim \pi$, e_i — случайный базисный вектор в $\mathbb{R}^d$, $i \sim U\{1, d\}$, $\tau > 0$ — достаточно мало.
В случае TPF в обоих точках, где мы вызываем оракула, одно и то же значение ξ, что может быть тяжело реализуемо. Также TPF является частным случаем OPF ($\xi^+ = \xi^- = \xi$).

3 Основные результаты

Для доказательств сходимости нам требуется сделать несколько предположений о функции f и об оракуле f_δ.

3.1 Предположения

А 1 (Гладкость). Функция f является L-гладкой на $\mathbb{R}^d$ с константой $L > 0$, т.е.:

$$\exists L > 0 : \forall x, y \in \mathbb{R}^d \hookrightarrow \|\nabla f(x) - \nabla f(y)\| \leqslant L\|x - y\|. \tag{7}$$

А 2 (Сильная выпуклость). Функция f является μ-сильно выпуклой на $\mathbb{R}^d$, т.е.:

$$\exists \mu > 0 : \forall x, y \in \mathbb{R}^d \hookrightarrow \frac{\mu}{2}\|x - y\|^2 \leqslant f(x) - f(y) - \langle \nabla f(y), x - y\rangle. \tag{8}$$

А 3 (Ограниченность оракульного шума). Оракульный шум ограничен некоторой константой $\Delta > 0$, т.е.:

$$\exists \Delta > 0 : \forall x \in \mathbb{R}^d \hookrightarrow |\delta(x)| \leqslant \Delta. \tag{9}$$

3.2 Алгоритм

Предлагается следующая версия ускоренного градиентного метода с аппроксимацией вида (3), взятой по случайной координате.

Algorithm 1 Accelerated Gradient Descent

Require: stepsize $\gamma > 0$, momentums θ, η, β, p, number of iterations N, approximation parameter $\tau > 0$.
Initialization: choose $x^0 = x_f^0$
1: **for** $k = 0, 1, \ldots, N-1$ **do**
2: $\quad x_g^k = \theta x_f^k + (1-\theta)x^k$
3: $\quad$ Sample $i \sim U\{1, \ldots, d\}$
4: $\quad g^k = d\frac{f_\delta(x_g^k + \tau e_i) - f_\delta(x_g^k - \tau e_i)}{2\tau} e_i$
5: $\quad x_f^{k+1} = x_g^k - p\gamma g^k$
6: $\quad x^{k+1} = \eta x_f^{k+1} + (p-\eta)x_f^k + (1-p)(1-\beta)x^k + (1-p)\beta x_g^k$
7: **end for**

При доказательстве сходимости понадобится следующая лемма.

Lemma 1. Предположим А 1, А 3. Тогда для аппроксимации градиента g^k в алгоритме 1 выполняется

$$\|\nabla f(x_g^k) - \mathbb{E}g^k\|^2 \leqslant d\left(\frac{L\tau}{2} + \frac{\Delta}{\tau}\right)^2. \tag{10}$$

$$\mathbb{E}[\|\nabla f(x_g^k) - g^k\|^2] \leqslant 2d\|\nabla f(x_g^k)\|^2 + 2d^2\left(\frac{L\tau}{2} + \frac{\Delta}{\tau}\right)^2. \tag{11}$$

Сформулируем теорему о сходимости метода на описанном классе функций.

Theorem 1 (Сходимость (1) в случае детерминированного шума). Предположим А 1 - А 3. Тогда ускоренный градиентный спуск (Algorithm 1) имеет скорость сходимости на задаче (1):

$$\begin{aligned}\mathbb{E}\left[\|x^N - x^*\|^2 + \frac{6}{\mu}(f(x_f^N) - f(x^*))\right] \leqslant \exp\left(-N\sqrt{\frac{p^2\mu\gamma}{3}}\right)\left(\|x^0 - x^*\|^2 + \frac{6}{\mu}(f(x_f^0) - f(x^*))\right) + \\ + \frac{3\sqrt{3\gamma}}{\mu^{3/2}} \cdot d\left(\frac{L\tau}{2} + \frac{\Delta}{\tau}\right)^2\left(2 + \sqrt{\frac{3}{\gamma\mu}}\right),\end{aligned} \tag{12}$$

где $\gamma \in (0, \frac{3}{4L}], \beta, \theta, \eta, p$ такие, что:

$$p \simeq (2(1+\gamma L))^{-1}, \beta \simeq \sqrt{p^2\mu\gamma}, \eta \simeq \sqrt{\frac{1}{\mu\gamma}}, \theta \simeq \frac{p\eta^{-1} - 1}{\beta p\eta^{-1} - 1.} \tag{13}$$

4 Вычислительный эксперимент

В эксперименте рассматриваются две задачи: минимизация квадратичной функции и логистическая регрессия на выборке mushrooms. Ускоренный градиентный спуск сравнивается со стандартным градиентным спуском и с методом Нестерова. В качестве аппроксимации градиента используется разностная схема (3), взятая по случайно выбранному базисному вектору $e_i, i \sim U\{1, d\}$.
В ходе последующих экспериментов рассматривался детерминированный шум в виде округления до 6 знаков после запятой (шум ограничен 10^{-6} по абсолютной величине), и параметр $\tau = 10^{-4}$.

4.1 Квадратичная задача

В качестве целевой функции выступает

$$f(x) = x^T Ax - b^T x + c, \tag{14}$$

где $A \in \mathbb{S}_d$ — случайная симметричная матрица с собственными значениями на отрезке $[\mu, L]$, $b \in \mathbb{R}^d$, $c \in \mathbb{R}$ — случайные. В конкретном эксперименте были взяты $\mu = 1, L = 1000, d = 100$.
Все методы запускаются из одной и той же случайной точки $x^0 \in \mathbb{R}^d$. Сходимость рассматриваем по аргументу целевой функции $\frac{\|x^k - x^*\|}{\|x^0 - x^*\|}$(OracleCalls($k$)), где x^k — значение аргумента на k-й итерации, x^* — точка оптимума, найденная численно, x^0 — стартовая точка алгоритма, OracleCalls(k) — суммарное количество вызовов оракула за k итераций.

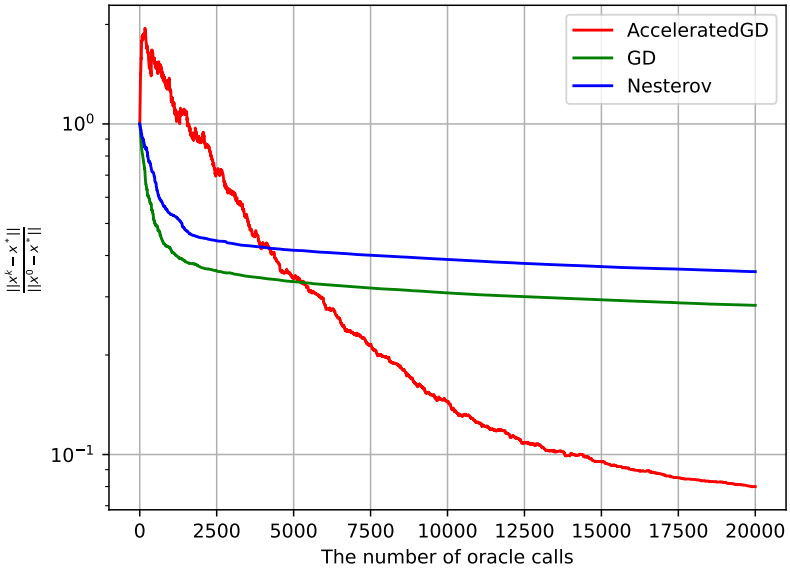

Рис. 1: Зависимость относительной ошибки $\frac{\|x^k - x^*\|}{\|x^0 - x^*\|}$ от числа оракульных вызовов для методов ускоренного градиентного спуска, градиентного спуска и Нестерова на квадратичной задаче минимизации. Детерминированный шум.

4.2 Логистическая регрессия

Оптимизируется модель логистической регрессии с L_2-регуляризацией вида

$$\min_{w \in R^d} f(w) = \frac{1}{m} \sum_{k=1}^{m} \log(1 + \exp(-y_k \cdot (Xw)_k)) + \lambda \|w\|_2^2, \tag{15}$$

где взято $\lambda = 0,1$.
В логистической регрессии используется датасет mushrooms из библиотеки LibSVM. Для него $d = 112$, $m = 8124$. В качестве осей графика применяются $\frac{\|\nabla f(x^k)\|}{\|\nabla f(x^0)\|}$(OracleCalls($k$)).

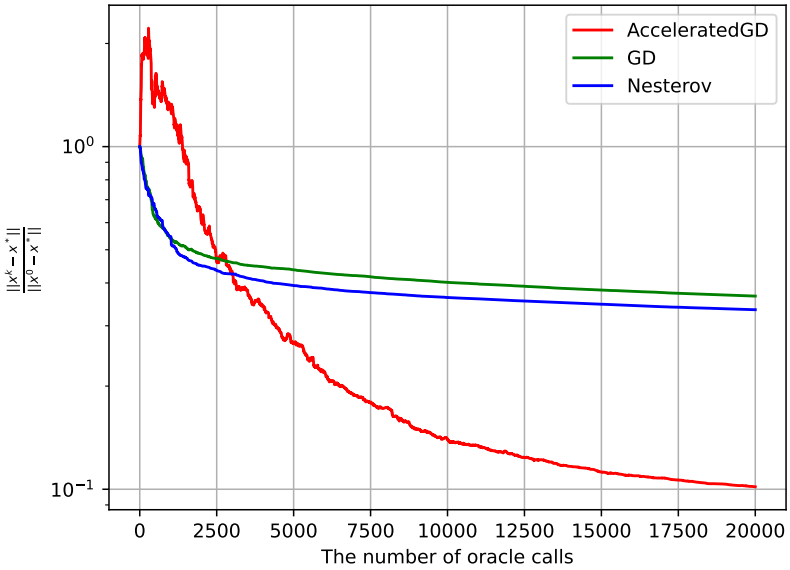

Рис. 2: Зависимость относительной ошибки $\frac{\|x^k - x^*\|}{\|x^0 - x^*\|}$ от числа оракульных вызовов для методов ускоренного градиентного спуска, градиентного спуска и Нестерова на квадратичной задаче минимизации. Стохастический шум (OPF).

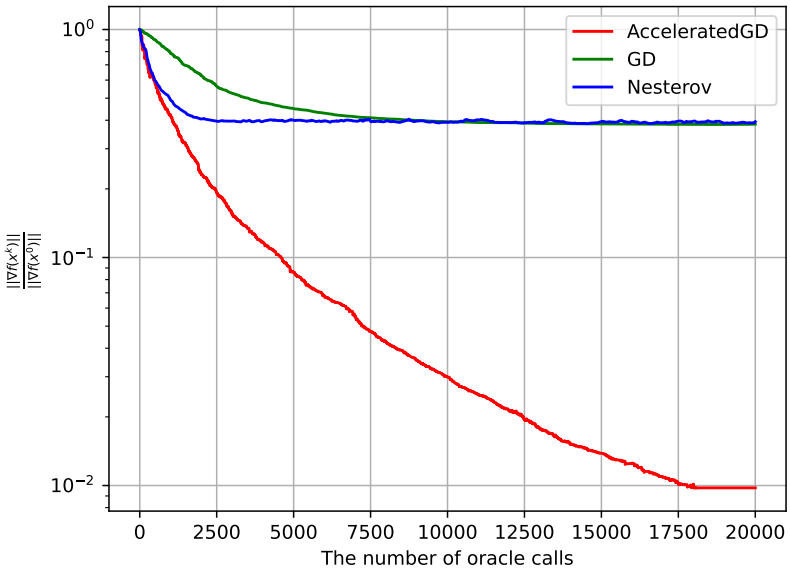

Рис. 3: Зависимость относительной ошибки $\frac{\|\nabla f(x^k)\|}{\|\nabla f(x^0)\|}$ от числа оракульных вызовов для методов ускоренного градиентного спуска, градиентного спуска и Нестерова на задаче логистической регрессии. Детерминированный шум.

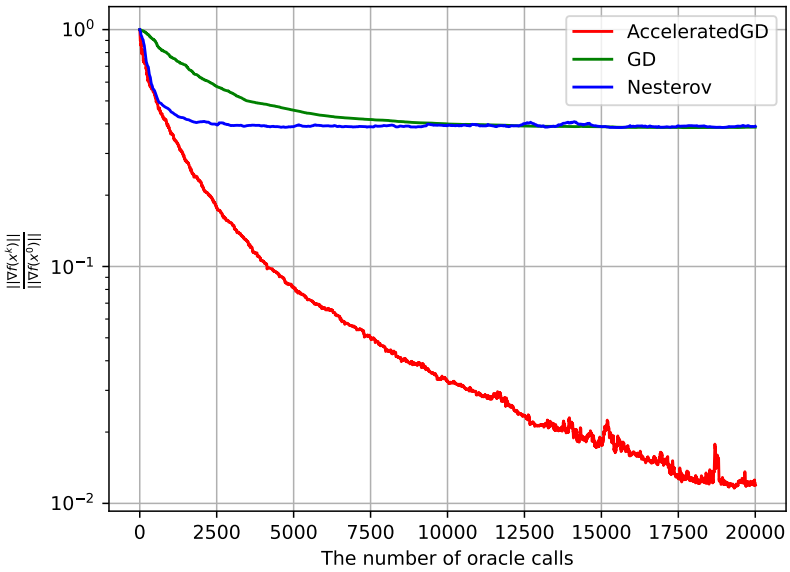

Рис. 4: Зависимость относительной ошибки $\frac{\|\nabla f(x^k)\|}{\|\nabla f(x^0)\|}$ от числа оракульных вызовов для методов ускоренного градиентного спуска, градиентного спуска и Нестерова на задаче логистической регрессии. Стохастический шум (OPF).

4.3 Отчет по эксперименту

В проведенных экспериментах ускоренный метод показал себя лучше других, хотя, например, в квадратичной задаче поначалу он ведет себя намного хуже. В задаче логистической регрессии предложенный метод сильно обходит остальные методы. Также в задаче логистической регрессии в случае детерминированного шума ускоренный метод достиг предельной точности для аппроксимации (3) и вышел на плато. То есть значения в τ-окрестности этой точки отличаются меньше, чем на величину округления. В эксперименте со стохастичнеским шумом на тех же итерациях видны скачки, что тоже свидетельствует о приближении к предельной точности решения.

5 Анализ ошибки

Посмотрим, как величина τ влияет на сходимость. С одной стороны, чем она меньше, тем более близкие точки мы берем, тем лучше линейная аппроксимации приближает функцию на этом отрезке, тем точнее должна быть оценка градиента. С другой стороны, присутствует шум и слишком близкие точки будут давать большую погрешность. Для сравнения также построим график для метода с градиентом, вычисленным аналитически, считая стоимость вычисления градиента за d оракульных вызовов. Он моделирует идеальные условия, когда $\Delta = 0, \tau \to 0$.

5.1 Квадратичная задача

Все описанные в секции вычислительного эксперимента параметры сохранены, за исключением переменного τ.

В построенном графике наблюдается, что параметр $\tau = 10^{-3}$ является оптимальным из рассмотренных. Также видно, что методы, использующие аппроксимацию градиента на небольшом количестве итераций опережают версию с честным градиентом. Обясняется это тем, что последний делает много оракульных вызовов, как следствие мало шагов, и не успевает накопить моменты.

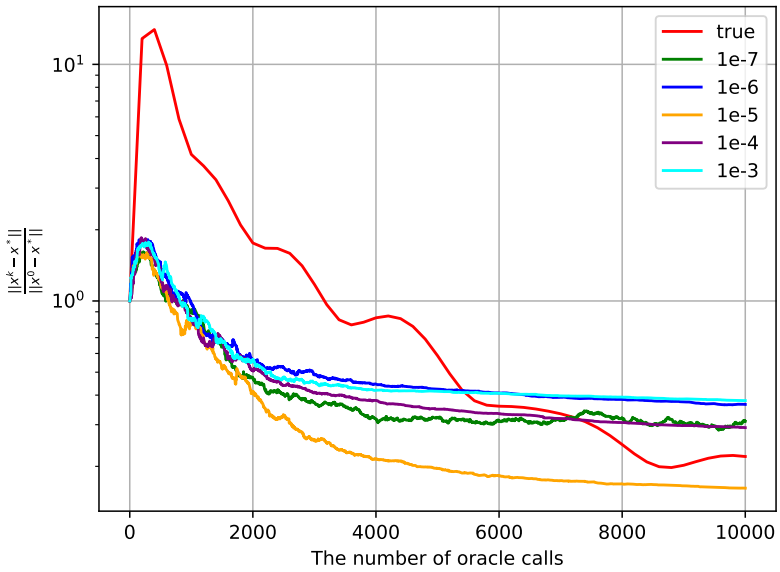

Рис. 5: Зависимость относительной ошибки $\frac{\|x^k - x^*\|}{\|x^0 - x^*\|}$ от числа оракульных вызовов для Accelerated GD с изменяющимся значением τ, а также Accelerated GD с аналитически вычисленным градиентом, квадратичная задача.

5.2 Логистическая регрессия

Условия также как в вычислительном эксперименте, за исключением изменений, описанных выше. Здесь

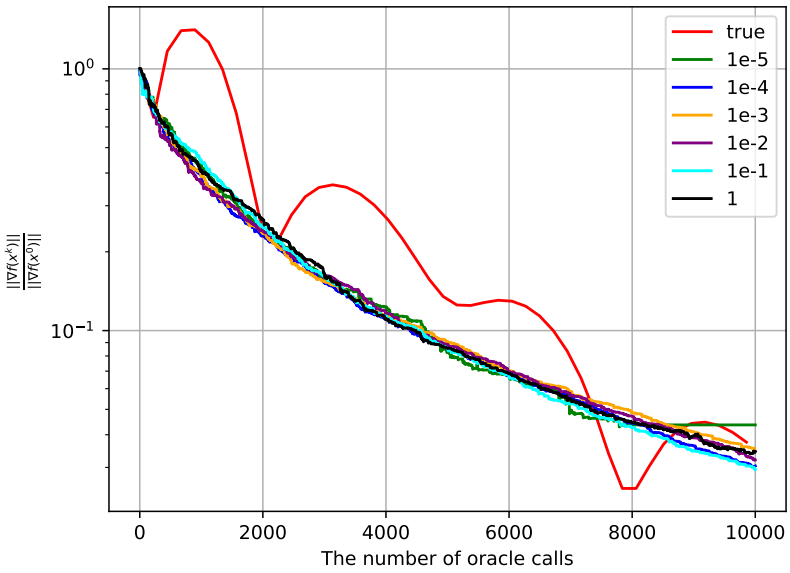

Рис. 6: Зависимость относительной ошибки $\frac{\|\nabla f(x^k)\|}{\|\nabla f(x^0)\|}$ от числа оракульных вызовов для Accelerated GD с изменяющимся значением τ, а также Accelerated GD с аналитически вычисленным градиентом, логистическая регрессия.

хорошо сходятся методы со значениями $\tau \in [10^{-4}, 10^{-1}]$. Также по сравнению с истинным градиентом, сходимость выглядит более стабильно.

Список литературы

[1] Vladislav Minashkin Aleksandr Beznosikov Alexander Gasnikov Andrey Veprikov, Alexander Bogdanov. New aspects of black box conditional gradient: Variance reduction and one point feedback, 2024.

[2] Aleksandr Beznosikov, Sergey Samsonov, Marina Sheshukova, Alexander Gasnikov, Alexey Naumov, and Eric Moulines. First order methods with markovian noise: from acceleration to variational inequalities, 2023.

[3] Alexander Gasnikov, Darina Dvinskikh, Pavel Dvurechensky, Eduard Gorbunov, Aleksander Beznosikov, and Aleksandr Lobanov. Randomized gradient-free methods in convex optimization, 2024.

[4] Eduard Gorbunov, Pavel Dvurechensky, and Alexander Gasnikov. An accelerated method for derivative-free smooth stochastic convex optimization, 2020.

[5] Aleksandr Lobanov, Anton Anikin, Alexander Gasnikov, Alexander Gornov, and Sergey Chukanov. Zero-order stochastic conditional gradient sliding method for non-smooth convex optimization, 2023.

[6] Aurelien Lucchi, Frank Proske, Antonio Orvieto, Francis Bach, and Hans Kersting. On the theoretical properties of noise correlation in stochastic optimization, 2022.

[7] Yurii Nesterov. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. Proceedings of the USSR Academy of Sciences, 269:543–547, 1983.

Lemma 2 (Lemma 1). Предположим A 1, A 3. Тогда для аппроксимации градиента g^k в алгоритме 1 выполняется

$$\|\nabla f(x_g^k) - \mathbb{E}g^k\|^2 \leqslant d\left(\frac{L\tau}{2} + \frac{\Delta}{\tau}\right)^2. \tag{16}$$

$$\mathbb{E}[\|\nabla f(x_g^k) - g^k\|^2] \leqslant 2d\|\nabla f(x_g^k)\|^2 + 2d^2\left(\frac{L\tau}{2} + \frac{\Delta}{\tau}\right)^2. \tag{17}$$

Доказательство.

$$\|\nabla f(x_g^k) - \mathbb{E}g^k\|^2 = \left\|\nabla f(x_g^k) - \mathbb{E}\left[d\frac{f(x_g^k + \tau e_i) - f(x_g^k - \tau e_i)}{2\tau}e_i + d\frac{\delta(x_g^k + \tau e_i) - \delta(x_g^k - \tau e_i)}{2\tau}e_i\right]\right\|^2 =$$

$$= \sum_{i=1}^d \left|\frac{f(x + \tau e_i) - f(x - \tau e_i)}{2\tau} - \nabla_i f(x_g^k) - \frac{\delta(x + \tau e_i) - \delta(x - \tau e_i)}{2\tau}\right|^2 \leqslant$$

$$\leqslant \sum_{i=1}^d \left(\left|\frac{1}{2\tau}\int_{-\tau}^{\tau}\langle f(x_g^k + te_i) - \nabla f(x_g^k), e_i\rangle dt\right| + \frac{\Delta}{\tau}\right)^2 \leqslant$$

$$\leqslant \sum_{i=1}^d \left(\frac{1}{2\tau}\int_{-\tau}^{\tau}\|f(x_g^k + te_i) - \nabla f(x_g^k)\|_2 \cdot \|e_i\|_2 dt + \frac{\Delta}{\tau}\right)^2 \leqslant$$

$$\leqslant \sum_{i=1}^d \left(\frac{1}{2\tau}\int_{-\tau}^{\tau} L|t|dt + \frac{\Delta}{\tau}\right)^2 = \sum_{i=1}^d \left(\frac{L\tau}{2} + \frac{\Delta}{\tau}\right)^2 = d\left(\frac{L\tau}{2} + \frac{\Delta}{\tau}\right)^2.$$

$$\mathbb{E}[\|\nabla f(x_g^k) - g^k\|^2] = \frac{1}{d}\sum_{i=1}^d \left\|\nabla f(x_g^k) - d\frac{f(x_g^k + \tau e_i) - f(x_g^k - \tau e_i)}{2\tau}e_i - d\frac{\delta(x_g^k + \tau e_i) - \delta(x_g^k - \tau e_i)}{2\tau}e_i\right\|^2 =$$

$$= \frac{1}{d}\sum_{i=1}^d \left(\sum_{j\neq i}|\nabla_j f(x_g^k)|^2 + \left|d\left(\frac{f(x_g^k + \tau e_i) - f(x_g^k - \tau e_i)}{2\tau} - \nabla_i f(x_g^k)\right) + (d-1)\nabla_i f(x_g^k) + d\frac{\delta(x_g^k + \tau e_i) - \delta(x_g^k - \tau e_i)}{2\tau}\right|^2\right) \leqslant$$

$$\leqslant \frac{1}{d}\sum_{i=1}^d \left(\|\nabla f(x_g^k)\|^2 + \left(\left|\frac{d}{2\tau}\int_{-\tau}^{\tau}\langle f(x_g^k + te_i) - \nabla f(x_g^k), e_i\rangle dt\right| + (d-1)|\nabla_i f(x_g^k)| + d\frac{\Delta}{\tau}\right)^2\right) \leqslant$$

$$\leqslant \frac{1}{d}\sum_{i=1}^d \left(\|\nabla f(x_g^k)\|^2 + \left(d\frac{L\tau}{2} + (d-1)|\nabla_i f(x_g^k)| + d\frac{\Delta}{\tau}\right)^2\right) \leqslant$$

$$\leqslant \frac{1}{d}\sum_{i=1}^d \left(\|\nabla f(x_g^k)\|^2 + 2\left(d\frac{L\tau}{2} + d\frac{\Delta}{\tau}\right)^2 + 2(d-1)^2|\nabla_i f(x_g^k)|^2\right) =$$

$$= \|\nabla f(x_g^k)\|^2 + 2d^2\left(\frac{L\tau}{2} + \frac{\Delta}{\tau}\right)^2 + 2\frac{(d-1)^2}{d}\|\nabla f(x_g^k)\|^2 \leqslant 2d\|\nabla f(x_g^k)\|^2 + 2d^2\left(\frac{L\tau}{2} + \frac{\Delta}{\tau}\right)^2.$$

□

Lemma 3. (Неравенство из доказательства Theorem 6 в работе [2]) Предположим A 1 - A 3. В Algoritm 1 коэффициенты $\gamma \in (0, \frac{3}{4L}], \beta, \theta, \eta, p$ подобраны следующим образом:

$$p \simeq (2(1 + \gamma L))^{-1}, \beta \simeq \sqrt{p^2\mu\gamma}, \eta \simeq \sqrt{\frac{1}{\mu\gamma}}, \theta \simeq \frac{p\eta^{-1} - 1}{\beta p\eta^{-1} - 1.} \tag{18}$$

Тогда справедливо неравенство:

$$\mathbb{E}[\|x^{k+1} - x^*\|^2 + 2\gamma\eta^2(f(x_f^{k+1}) - f(x^*))] \leqslant$$
$$\leqslant (1 - \beta/2)\|x^k - x^*\|^2 + (1 - p/\eta)2\gamma\eta^2(f(x_f^k) - f(x^*)) + p\eta^2\gamma^2(1 + 2p/\beta)\|\mathbb{E}[g^k] - \nabla f(x_g^k)\|^2 +$$
$$+2p^2\eta^2\gamma^2(1 + \gamma L)\mathbb{E}[\|g^k - \nabla f(x_g^k)\|^2] - p\gamma^2\eta^2(1 - 2p(1 + \gamma L))\|\nabla f(x_g^k)\|^2 \tag{19}$$

Theorem 2 (Theorem 1). Предположим A 1 - A 3. Тогда ускоренный градиентный спуск (Algorithm 1) имеет скорость сходимости на задаче (1):

$$\mathbb{E}\left[\|x^N - x^*\|^2 + \frac{6}{\mu}(f(x_f^N) - f(x^*))\right] \leqslant \exp\left(-N\sqrt{\frac{p^2\mu\gamma}{3}}\right)\left(\|x^0 - x^*\|^2 + \frac{6}{\mu}(f(x_f^0) - f(x^*))\right) + \\ + \frac{3\sqrt{3\gamma}}{\mu^{3/2}} \cdot d\left(\frac{L\tau}{2} + \frac{\Delta}{\tau}\right)^2\left(2 + \sqrt{\frac{3}{\gamma\mu}}\right), \tag{20}$$

где $\gamma \in (0, \frac{3}{4L}], \beta, \theta, \eta, p$ такие, что:

$$p \simeq (2(1+\gamma L))^{-1}, \beta \simeq \sqrt{p^2\mu\gamma}, \eta \simeq \sqrt{\frac{1}{\mu\gamma}}, \theta \simeq \frac{p\eta^{-1} - 1}{\beta p\eta^{-1} - 1.} \tag{21}$$

Доказательство. Подставим (16) и (17) в (19).

$$\mathbb{E}[\|x^{k+1} - x^*\|^2 + 2\gamma\eta^2(f(x_f^{k+1}) - f(x^*))] \leqslant (1-\beta/2)\|x^k - x^*\|^2 + (1-p/\eta)2\gamma\eta^2(f(x_f^k) - f(x^*)) +$$

$$+p\eta^2\gamma^2(1+2p/\beta)\|\mathbb{E}[g^k] - \nabla f(x_g^k)\|^2 + 2p^2\eta^2\gamma^2(1+\gamma L)\mathbb{E}[\|g^k - \nabla f(x_g^k)\|^2] - p\gamma^2\eta^2(1-2p(1+\gamma L))\|\nabla f(x_g^k)\|^2 \leqslant$$

$$\leqslant (1-\beta/2)\|x^k - x^*\|^2 + (1-p/\eta)2\gamma\eta^2(f(x_f^k) - f(x^*)) + p\eta^2\gamma^2(1+2p/\beta)d\left(\frac{L\tau}{2} + \frac{\Delta}{\tau}\right)^2 +$$

$$+2p^2\eta^2\gamma^2(1+\gamma L)\left(2d\|\nabla f(x_g^k)\|^2 + 2d^2\left(\frac{L\tau}{2} + \frac{\Delta}{\tau}\right)^2\right) - p\gamma^2\eta^2(1-2p(1+\gamma L))\|\nabla f(x_g^k)\|^2 =$$

$$= (1-\beta/2)\|x^k - x^*\|^2 + (1-p/\eta)2\gamma\eta^2(f(x_f^k) - f(x^*)) +$$

$$+p\eta^2\gamma^2 d\left(\frac{L\tau}{2} + \frac{\Delta}{\tau}\right)^2(1 + 2p/\beta + 4pd(1+\gamma L)) +$$

$$+p\gamma\eta^2\|\nabla f(x_g^k)\|^2(2p(1+\gamma L)\cdot 2d + 2p(1+\gamma L) - 1)$$

Возьмем $p = \frac{1}{2(1+\gamma L)(2d+1)}$.

$$\mathbb{E}[\|x^{k+1} - x^*\|^2 + 2\gamma\eta^2(f(x_f^{k+1}) - f(x^*))] \leqslant$$

$$\leqslant (1-\beta/2)\|x^k - x^*\|^2 + (1-p/\eta)2\gamma\eta^2(f(x_f^k) - f(x^*)) + p\eta^2\gamma^2 d\left(\frac{L\tau}{2} + \frac{\Delta}{\tau}\right)^2(1 + 2p/\beta + 4pd(1+\gamma L))$$

Теперь выберем $\beta/2 = p/\eta$, $p\eta\gamma = 3\beta/2\mu$.

$$\mathbb{E}[\|x^{k+1} - x^*\|^2 + 2\gamma\eta^2(f(x_f^{k+1}) - f(x^*))] \leqslant$$

$$\leqslant (1-\beta/2)\left(\|x^k - x^*\|^2 + 2\gamma\eta^2(f(x_f^k) - f(x^*))\right) + \frac{9\beta^2}{4\mu^2}\cdot 2(1+\gamma L)(2d+1)d\left(\frac{L\tau}{2} + \frac{\Delta}{\tau}\right)^2(1 + 2p/\beta + 4pd(1+\gamma L)) =$$

$$= (1-\beta/2)\left(\|x^k - x^*\|^2 + 2\gamma\eta^2(f(x_f^k) - f(x^*))\right) + \frac{9\beta^2}{2\mu^2}\cdot(1+\gamma L)(2d+1)d\left(\frac{L\tau}{2} + \frac{\Delta}{\tau}\right)^2\left(1 + \eta + \frac{2d}{2d+1}\right)$$

Подставляя $\beta = \sqrt{\frac{4p^2\mu\gamma}{3}}$, получим:

$$\mathbb{E}[\|x^N - x^*\|^2 + 2\gamma\eta^2(f(x_f^N) - f(x^*))] \leqslant$$

$$\leqslant \left(1 - \sqrt{\frac{p^2\mu\gamma}{3}}\right)^N\left(\|x^0 - x^*\|^2 + 2\gamma\eta^2(f(x_f^0) - f(x^*))\right) + \frac{9\beta^2}{2\mu^2}\cdot(1+\gamma L)(2d+1)d\left(\frac{L\tau}{2} + \frac{\Delta}{\tau}\right)^2(1 + \eta + 1)\cdot\frac{2}{\beta} =$$

$$= \left(1 - \sqrt{\frac{p^2\mu\gamma}{3}}\right)^N\left(\|x^0 - x^*\|^2 + 2\gamma\eta^2(f(x_f^0) - f(x^*))\right) + \frac{9}{\mu^2}\sqrt{\frac{4p^2\mu\gamma}{3}}\cdot(1+\gamma L)(2d+1)d\left(\frac{L\tau}{2} + \frac{\Delta}{\tau}\right)^2(2 + \eta) \leqslant$$

$$\leqslant \left(1 - \sqrt{\frac{p^2\mu\gamma}{3}}\right)^N \left(\|x^0 - x^*\|^2 + 2\gamma\eta^2(f(x_f^0) - f(x^*))\right) + \frac{3\sqrt{3\gamma}}{\mu^{3/2}} \cdot d\left(\frac{L\tau}{2} + \frac{\Delta}{\tau}\right)^2 (2+\eta)\,.$$

Наконец, $\eta = \sqrt{\frac{3}{\gamma\mu}}$.

$$\mathbb{E}\left[\|x^N - x^*\|^2 + \frac{6}{\mu}(f(x_f^N) - f(x^*))\right] \leqslant$$

$$\leqslant \left(1 - \sqrt{\frac{p^2\mu\gamma}{3}}\right)^N \left(\|x^0 - x^*\|^2 + \frac{6}{\mu}(f(x_f^0) - f(x^*))\right) + \frac{3\sqrt{3\gamma}}{\mu^{3/2}} \cdot d\left(\frac{L\tau}{2} + \frac{\Delta}{\tau}\right)^2 \left(2 + \sqrt{\frac{3}{\gamma\mu}}\right).$$

Воспользуемся соотношением $(1-x)^N \leqslant \exp(-Nx)$:

$$\mathbb{E}\left[\|x^N - x^*\|^2 + \frac{6}{\mu}(f(x_f^N) - f(x^*))\right] \leqslant$$

$$\leqslant \exp\left(-N\sqrt{\frac{p^2\mu\gamma}{3}}\right) \left(\|x^0 - x^*\|^2 + \frac{6}{\mu}(f(x_f^0) - f(x^*))\right) + \frac{3\sqrt{3\gamma}}{\mu^{3/2}} \cdot d\left(\frac{L\tau}{2} + \frac{\Delta}{\tau}\right)^2 \left(2 + \sqrt{\frac{3}{\gamma\mu}}\right).$$

□