Современные Безградиентные Рандомизированные Алгоритмы Выпуклой Оптимизации

Лобанов Александр Владимирович

Московский физико-технический институт
Сколковский институт науки и технологий
Институт системного программирования им. В. П. Иванникова РАН

lobbsasha@mail.ru

27 июня 2024

Безградиентные алгоритмы

Рассматривается задача выпуклой оптимизации

$$\min_{x \in Q \subseteq \mathbb{R}^d} f(x)$$

Вопрос

Когда следует использовать безградиентные алгоритмы?

Безградиентные алгоритмы

Рассматривается задача выпуклой оптимизации

$$\min_{x \in Q \subseteq \mathbb{R}^d} f(x)$$

Вопрос

Когда следует использовать безградиентные алгоритмы?

Подходы для создания рандомизированных безградиентных методов

- **Негладкий случай**
  - Схема сглаживания с l_1 рандомизацией
  - Схема сглаживания с l_2 рандомизацией
- **Гладкий случай**
  - l_1 рандомизация
  - l_2 рандомизация
- **Случай с повышенной гладкостью**
  - Рандомизация с помощью ядра

Безградиентные алгоритмы

Рассматривается задача выпуклой оптимизации

$$\min_{x \in Q \subseteq \mathbb{R}^d} f(x)$$

Вопрос

Когда следует использовать безградиентные алгоритмы?

Критерии оптимальности

1. Число оракульных вызовов: T
2. Число последовательных итераций метода: N
3. Максимально допустимый уровень шума Δ

Мотивация для поиска максимального уровня шума

Рис.: Экономия ресурсов

Рис.: Устойчивость к атакам

Рис.: Конфиденциальность

1. Экономия ресурсов. Чем точнее вычисляется значение объективной функции, тем дороже этот процесс.
2. Устойчивость к атакам. Повышение максимального уровня шума делает алгоритм более устойчивым к враждебным атакам.
3. Конфиденциальность. Некоторые компании из-за секретности не могут передать всю информацию.

Содержание

Содержание

Постановка задачи

Рассматривается стохастическая негладкая выпуклая задача оптимизации

$$\min_{x \in Q \subseteq \mathbb{R}^d} f(x) := \mathbb{E}_\xi \left[f(x, \xi) \right]$$

Безградиентный оракул

Предполагается, что безградиентный оракул возвращает значение функции $f(x)$, возможно, с некоторым враждебным шумом $\delta(x)$:

$$f_\delta(x) := f(x) + \delta(x)$$

Предположение 1. (Липшицева непрерывная функция).

Функция $f(x, \xi)$ является M-Липшицевой непрерывной функцией в l_p-норме, то есть для всех $x, y \in Q$ имеем

$$|f(y, \xi) - f(x, \xi)| \leq M(\xi)\|y - x\|_p.$$

Более того, существует положительная константа M, которая определяется следующим образом: $\mathbb{E}\left[M^2(\xi)\right] \leq M^2$. В частности, для $p = 2$ используем обозначение M_2 для константы Липшица.

Предположение 2. (Ограниченность шума).

Для всех $x \in Q$ выполняется $|\delta(x)| \leq \Delta$, где Δ – уровень неточности (шума).

Основная идея

Негладкая задача

$$\min_{x \in Q \subseteq \mathbb{R}^d} f(x)$$

Гладкая задача

$$\min_{x \in Q \subseteq \mathbb{R}^d} f_\gamma(x)$$

Схемы сглаживания [1, 2]

Definition

Пусть l_p-шар определяется, как $B_p^d(r) := \{x \in \mathbb{R} : \|x\|_p \leq r\}$. Тогда гладкая аппроксимация негладкой функции $f(x)$ выглядит следующим образом

$$f_\gamma(x) := \mathbb{E}_{\tilde{e}} \left[f(x + \gamma \tilde{e}) \right],$$

где $\gamma > 0$, $\tilde{e}$ — случайный вектор, равномерно распределенный на $B_p^d(1)$
(далее ограничимся рассмотрением случаев $p = 1$ и $p = 2$).

Свойства функции $f_\gamma(x)$

Случай, когда $\tilde{e} \in RB_2^d(1)$

1. $f(x) \leq f_\gamma(x) \leq f(x) + \gamma M_2$;
2. f_γ — M-Липшицева:
$$|f_\gamma(y) - f_\gamma(x)| \leq M\|y - x\|_p,$$
3. f_γ имеет $L_{f_\gamma} = \frac{\sqrt{d}M}{\gamma}$-Липшицевый градиент:
$$\|\nabla f_\gamma(y) - \nabla f_\gamma(x)\|_q \leq L_{f_\gamma}\|y - x\|_p.$$

где q такой, что $1/p + 1/q = 1$.

Первое свойство

$$f_\gamma(x) = \mathbb{E}_{\tilde{e}}\left[f(x + \gamma\tilde{e})\right] \geq \mathbb{E}_{\tilde{e}}\left[f(x) + \langle \nabla f(x), \gamma\tilde{e}\rangle\right] = \mathbb{E}_{\tilde{e}}\left[f(x)\right] = f(x).$$
$$|f_\gamma(x) - f(x)| = |\mathbb{E}_{\tilde{e}}\left[f(x + \gamma\tilde{e})\right] - f(x)| \leq \mathbb{E}_{\tilde{e}}\left[|f(x + \gamma\tilde{e}) - f(x)|\right] \leq \gamma M_2 \mathbb{E}_{\tilde{e}}\left[\|\tilde{e}\|_2\right] \leq \gamma M_2.$$

Свойства функции $f_\gamma(x)$

Случай, когда $\tilde{e} \in RB_2^d(1)$

1. $f(x) \leq f_\gamma(x) \leq f(x) + \gamma M_2$;
2. f_γ — M-Липшицева:
$$|f_\gamma(y) - f_\gamma(x)| \leq M\|y - x\|_p,$$
3. f_γ имеет $L_{f_\gamma} = \frac{\sqrt{d}M}{\gamma}$-Липшицевый градиент:
$$\|\nabla f_\gamma(y) - \nabla f_\gamma(x)\|_q \leq L_{f_\gamma}\|y - x\|_p.$$

где q такой, что $1/p + 1/q = 1$.

Второе свойство

$$|f_\gamma(y) - f_\gamma(x)| = |\mathbb{E}_{\tilde{e}}\left[f(y + \gamma\tilde{e}) - f(x + \gamma\tilde{e})\right]| \leq \mathbb{E}_{\tilde{e}}\left[|f(y + \gamma\tilde{e}) - f(x + \gamma\tilde{e})|\right] \leq M\|y - x\|_p$$

Свойства функции $f_\gamma(x)$

Случай, когда $\tilde{e} \in RB_1^d(1)$

1. $f(x) \leq f_\gamma(x) \leq f(x) + \frac{2}{\sqrt{d}}\gamma M_2$;
2. f_γ — M-Липшицева:
$$|f_\gamma(y) - f_\gamma(x)| \leq M\|y - x\|_p,$$
3. f_γ имеет $L_{f_\gamma} = \frac{dM}{2\gamma}$-Липшицевый градиент:
$$\|\nabla f_\gamma(y) - \nabla f_\gamma(x)\|_q \leq L_{f_\gamma}\|y - x\|_p.$$

где q такой, что $1/p + 1/q = 1$.

Первое свойство

$$f_\gamma(x) = \mathbb{E}_{\tilde{e}}\left[f(x + \gamma\tilde{e})\right] \geq \mathbb{E}_{\tilde{e}}\left[f(x) + \langle \nabla f(x), \gamma\tilde{e}\rangle\right] = \mathbb{E}_{\tilde{e}}\left[f(x)\right] = f(x).$$
$$|f_\gamma(x) - f(x)| = |\mathbb{E}_{\tilde{e}}\left[f(x + \gamma\tilde{e})\right] - f(x)| \leq \mathbb{E}_{\tilde{e}}\left[|f(x + \gamma\tilde{e}) - f(x)|\right]$$
$$\leq \gamma M_2 \mathbb{E}_{\tilde{e}}\left[\|\tilde{e}\|_2\right] \leq \frac{2}{\sqrt{d}}\gamma M_2.$$

Свойства функции $f_\gamma(x)$

Случай, когда $\tilde{e} \in RB_1^d(1)$

1. $f(x) \leq f_\gamma(x) \leq f(x) + \frac{2}{\sqrt{d}}\gamma M_2$;
2. f_γ — M-Липшицева:
$$|f_\gamma(y) - f_\gamma(x)| \leq M\|y - x\|_p,$$
3. f_γ имеет $L_{f_\gamma} = \frac{dM}{2\gamma}$-Липшицевый градиент:
$$\|\nabla f_\gamma(y) - \nabla f_\gamma(x)\|_q \leq L_{f_\gamma}\|y - x\|_p.$$

где q такой, что $1/p + 1/q = 1$.

Лемма 1 из [2] для первого свойства

Пусть $q \in [1, \infty)$ и пусть $v \in RB_1^d(1)$. Тогда

$$\mathbb{E}[\|v\|_q] \leq \frac{q d^{\frac{1}{q}}}{d+1}$$

Свойства функции $f_\gamma(x)$

Случай, когда $\tilde{e} \in RB_1^d(1)$

1. $f(x) \leq f_\gamma(x) \leq f(x) + \frac{2}{\sqrt{d}}\gamma M_2$;
2. f_γ — M-Липшицева:
$$|f_\gamma(y) - f_\gamma(x)| \leq M\|y - x\|_p,$$
3. f_γ имеет $L_{f_\gamma} = \frac{dM}{2\gamma}$-Липшицевый градиент:
$$\|\nabla f_\gamma(y) - \nabla f_\gamma(x)\|_q \leq L_{f_\gamma}\|y - x\|_p.$$

где q такой, что $1/p + 1/q = 1$.

Второе свойство

$$|f_\gamma(y) - f_\gamma(x)| = |\mathbb{E}_{\tilde{e}}\left[f(y + \gamma\tilde{e}) - f(x + \gamma\tilde{e})\right]| \leq \mathbb{E}_{\tilde{e}}\left[|f(y + \gamma\tilde{e}) - f(x + \gamma\tilde{e})|\right] \leq M\|y - x\|_p$$

Основная идея

Негладкая задача

$$\min_{x \in Q \subseteq \mathbb{R}^d} f(x)$$

Гладкая задача

$$\min_{x \in Q \subseteq \mathbb{R}^d} f_\gamma(x)$$

Связь между задачами

Если имеем $\frac{\varepsilon}{2}$-точность для функции $f_\gamma(x)$, то имеем ε-точность для функции $f(x)$:

$$\begin{aligned} f(x^{N+1}) - f(x_*) &\overset{①}{\leq} f_\gamma(x^{N+1}) - f(x_*) \overset{②}{\leq} f_\gamma(x^{N+1}) - f_\gamma(x_*) + \gamma M_2 \\ &\leq f_\gamma(x^{N+1}) - f_\gamma(x_*(\gamma)) + \gamma M_2 \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon, \end{aligned}$$

Обобщение формулы Стокса [2]

$$\nabla f_\gamma(x) = \mathbb{E}_{\tilde{e}}[\nabla f(x + \gamma\tilde{e})] = \frac{\mathrm{Vol}_{d-1}(\partial D)}{\mathrm{Vol}_d(D)} \cdot \mathbb{E}_e[f(x + \gamma e)n(e)]$$

Рандомизация с двухточечной обратной связью

l_1-рандомизация

$$\nabla f_\gamma(x, \xi, e) = \frac{d}{2\gamma}(f_\delta(x + \gamma e, \xi) - f_\delta(x - \gamma e, \xi))\mathrm{sign}(e), \quad (e \in RS_1^d(1))$$

l_2-рандомизация

$$\nabla f_\gamma(x, \xi, e) = \frac{d}{2\gamma}(f_\delta(x + \gamma e, \xi) - f_\delta(x - \gamma e, \xi))e, \quad (e \in RS_2^d(1))$$

Свойства ∇f_γ для l_1-рандомизации

1. Если $\Delta = 0$, то оценки будут несмещенными

$$\mathbb{E}_{e,\xi}\left[\nabla f_\gamma(x,\xi,e)\right] = \nabla f_\gamma(x)$$

2. ("смещение") при $\Delta > 0$

$$\mathbb{E}_{e,\xi}\langle[\nabla f_\gamma(x,\xi,e)] - \nabla f_\gamma(x), r\rangle \lesssim \frac{d\Delta R}{\gamma}, \quad \forall r : \|r\|_2 \leq R.$$

3. (оценка второго момента)

$$\mathbb{E}_{e,\xi}\left[\|\nabla f_\gamma(x,\xi,e)\|_q^2\right] \leq \kappa(p,d)\left(M_2^2 + \frac{d^2\Delta^2}{12(1+\sqrt{2})^2\gamma^2}\right),$$

где $1/p + 1/q = 1$ и $\kappa(p,d) = 48(1+\sqrt{2})^2 d^{2-\frac{2}{p}}$.

Свойства ∇f_γ для l_2-рандомизации

1. Если $\Delta = 0$, то оценки будут несмещенными

$$\mathbb{E}_{e,\xi}\left[\nabla f_\gamma(x,\xi,e)\right] = \nabla f_\gamma(x)$$

2. ("смещение") при $\Delta > 0$

$$\mathbb{E}_{e,\xi}\langle[\nabla f_\gamma(x,\xi,e)] - \nabla f_\gamma(x), r\rangle \lesssim \frac{\sqrt{d}\Delta R}{\gamma}, \quad \forall r : \|r\|_2 \leq R.$$

3. (оценка второго момента)

$$\mathbb{E}_{e,\xi}\left[\|\nabla f_\gamma(x,\xi,e)\|_q^2\right] \leq \kappa(p,d)\left(M_2^2 + \frac{d^2\Delta^2}{\sqrt{2}\gamma^2}\right),$$

где $1/p + 1/q = 1$ и $\kappa(p,d) = \sqrt{2}\min\{q, \ln d\}\, d^{2-\frac{2}{p}}$.

Основная идея

Негладкая задача

$$\min_{x \in Q \subseteq \mathbb{R}^d} f(x)$$

Гладкая задача

$$\min_{x \in Q \subseteq \mathbb{R}^d} f_\gamma(x)$$

Связь между задачами

Если имеем $\frac{\varepsilon}{2}$-точность для функции $f_\gamma(x)$, то имеем ε-точность для функции $f(x)$:

$$\begin{aligned} f(x^{N+1}) - f(x_*) &\overset{①}{\leq} f_\gamma(x^{N+1}) - f(x_*) \overset{②}{\leq} f_\gamma(x^{N+1}) - f_\gamma(x_*) + \gamma M_2 \\ &\leq f_\gamma(x^{N+1}) - f_\gamma(x_*(\gamma)) + \gamma M_2 \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon, \end{aligned}$$

Параметры

$$\boxed{\gamma = \frac{\varepsilon}{2M_2}} \qquad \boxed{L_{f_\gamma} = \frac{\sqrt{d}M}{\gamma} = \frac{2\sqrt{d}MM_2}{\varepsilon}} \qquad \boxed{\sigma^2 \leq 2\sqrt{2}\min\{q, \ln d\}\, d^{2-\frac{2}{p}} M_2^2}$$

Федеративное обучение

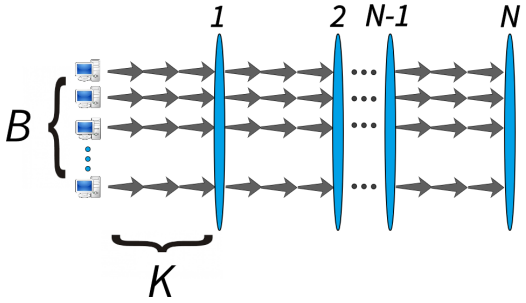


Рис.: Архитектура федеративного обучения

Algorithm	$\mathbb{E}\left[f(\cdot)\right] - f^* \lesssim \ldots$	Reference
Mb-Ac-SGD	$\frac{LR^2}{N^2} + \frac{\sigma R}{\sqrt{BNK}}$	(Woodworth et al., 2021) [3]
SM-Ac-SGD	$\frac{LR^2}{N^2K^2} + \frac{\sigma R}{\sqrt{NK}}$	(Woodworth et al., 2021) [3]
Local-AC-CA	$\frac{LR^2}{N^2K^2} + \frac{\sigma R}{\sqrt{BNK}}$	(Woodworth et al., 2020) [4]
FedAc	$\frac{LR^2}{N^2K} + \frac{\sigma R}{\sqrt{BNK}} + \min\left\{\frac{L^{\frac{1}{3}}\sigma^{\frac{2}{3}}R^{\frac{4}{3}}}{NK^{\frac{1}{3}}}, \frac{L^{\frac{1}{2}}\sigma^{\frac{1}{2}}R^{\frac{3}{2}}}{NK^{\frac{1}{4}}}\right\}$	(Yuan, Ma, 2020) [5]
Mb-SMP	$\max\left\{\frac{LR^2}{N}, \frac{\sigma R}{\sqrt{BNK}}\right\}$	
SM-SMP	$\max\left\{\frac{LR^2}{NK}, \frac{\sigma R}{\sqrt{NK}}\right\}$	

Таблица: **Результаты скорости сходимости.**
Обозначение: $R : \|x^0 - x_*\|$; B: число рабочих (компьютеров); K: число локальных обновлений; N: число коммуникационных раундов; L: гладкость.

Theorem

Схема сглаживания, применяемая к негладкой задаче, обеспечивает сходимость Minibatch Accelerated SGD [3]. Другими словами, для достижения ε точности решения негладкой задачи необходимо проделать NK итераций с максимально допустимым уровнем шума Δ и общим числом вызова безградиентного оракула T в соответствии с выбранным методом и схемой сглаживания:

- *l_1-рандомизация*

$$N = O\left(\frac{d^{1/4}\sqrt{MM_2}R}{\varepsilon}\right); \quad K = 1; \quad B = O\left(\frac{\kappa(p,d)dM_2^2R^2}{KN\varepsilon^2}\right);$$

$$T = \tilde{O}\left(\frac{\kappa(p,d)dM_2^2R^2}{\varepsilon^2}\right) = \begin{cases} O\left(\frac{dM_2^2R^2}{\varepsilon^2}\right), & p = 2 \ (q = 2) \\ O\left(\frac{M_2^2R^2}{\varepsilon^2}\right), & p = 1 \ (q = \infty). \end{cases}$$

Theorem

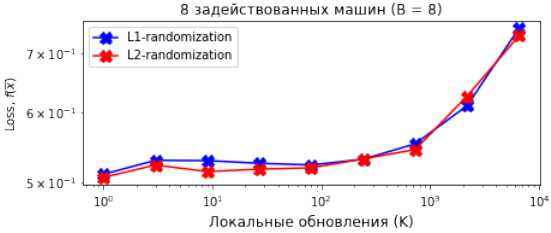
Схема сглаживания, применяемая к негладкой задаче, обеспечивает сходимость Minibatch Accelerated SGD [3]. Другими словами, для достижения ε точности решения негладкой задачи необходимо проделать NK итераций с максимально допустимым уровнем шума Δ и общим числом вызова безградиентного оракула T в соответствии с выбранным методом и схемой сглаживания:

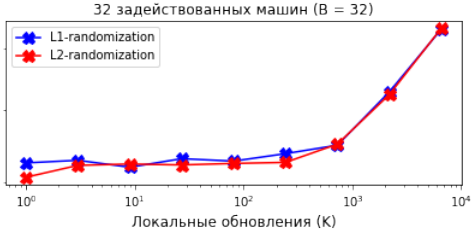
- *l_2-рандомизация*

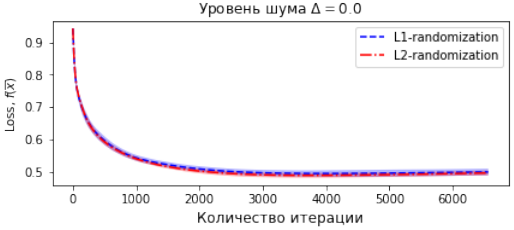
$$N = O\left(\frac{d^{1/4}\sqrt{MM_2}R}{\varepsilon}\right); \quad K = 1; \quad B = O\left(\frac{\kappa(p,d)dM_2^2R^2}{KN\varepsilon^2}\right);$$

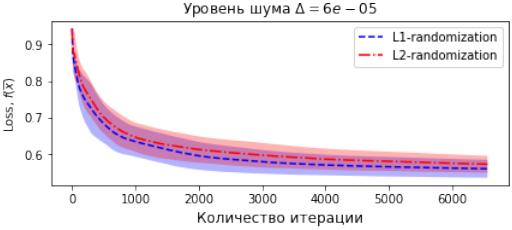
$$T = \tilde{O}\left(\frac{\kappa(p,d)dM_2^2R^2}{\varepsilon^2}\right) = \begin{cases} \tilde{O}\left(\frac{dM_2^2R^2}{\varepsilon^2}\right), & p = 2 \ (q = 2) \\ \tilde{O}\left(\frac{(\ln d)M_2^2R^2}{\varepsilon^2}\right), & p = 1 \ (q = \infty). \end{cases}$$

Эксперименты









Содержание

Постановка задачи

Рассматривается стандартная стохастическая выпуклая задача оптимизации

$$f^* = \min_{x \in \mathbb{R}^d} \{f(x) := \mathbb{E}\left[f(x, \xi)\right]\},$$

где $f : \mathbb{R}^d \to \mathbb{R}$ – гладкая выпуклая функция, которую мы хотими минимизировать на $\mathbb{R}^d$. Данная постановка задачи является общей, поэтому, чтобы определить конкретный класс задач (который будем рассматривать), вводится некоторые предположения о целевой функции и градиентном оракуле.

Предположения о целевой функции

Предположение 1. (Выпуклость)

Почти для каждого ξ, $f(x, \xi)$ – неотрицательная и выпуклая функция относительно x, т.е.

$$\forall x, y, \xi \quad f(y, \xi) \geq f(x, \xi) + \langle \nabla f(x, \xi), y - x \rangle .$$

Предположение 2. (Гладкость)

Почти для каждого ξ, $f(x, \xi)$ – неотрицательная, L-гладкая функция относительно x, т.е.

$$\forall x, y, \xi \quad f(y, \xi) \leq f(x, \xi) + \langle \nabla f(x, \xi), y - x \rangle + \frac{L}{2} \|y - x\|^2 .$$

Предположение 3.

Функция $f(x)$ – выпуклая и меет минимальное значение $f^* = \min_x f(x)$, которое достигается в точке x^* с $\|x^*\| \leq R$.

Предположения о градиентном оракуле

Определение 1. (Градиентный оракул)

Отображение $\mathbf{g} : \mathbb{R}^d \times \mathcal{D} \to \mathbb{R}^d$ такое что

$$\mathbf{g}(x, \xi) = \nabla f(x, \xi) + \mathbf{b}(x),$$

где $\mathbf{b} : \mathbb{R}^d \to \mathbb{R}^d$ такой что $\forall x \in \mathbb{R}^d$: $\|\mathbf{b}(x)\|^2 \leq \zeta^2$.

Предположения о градиентном оракуле

Определение 1. (Градиентный оракул)

Отображение $\mathbf{g} : \mathbb{R}^d \times \mathcal{D} \to \mathbb{R}^d$ такое что

$$\mathbf{g}(x, \xi) = \nabla f(x, \xi) + \mathbf{b}(x),$$

где $\mathbf{b} : \mathbb{R}^d \to \mathbb{R}^d$ такой что $\forall x \in \mathbb{R}^d$: $\|\mathbf{b}(x)\|^2 \leq \zeta^2$.

Предположение 4. (Ограниченность градиентного шума)

Существует параметр $\sigma_*^2 \geq 0$ такой что $\forall x \in \mathbb{R}^d$

$$\mathbb{E}\left[\|\nabla f(x^*, \xi) - \nabla f(x^*)\|^2\right] \leq \sigma_*^2.$$

Предположения о градиентном оракуле

Определение 1. (Градиентный оракул)

Отображение $\mathbf{g} : \mathbb{R}^d \times \mathcal{D} \to \mathbb{R}^d$ такое что

$$\mathbf{g}(x, \xi) = \nabla f(x, \xi) + \mathbf{b}(x),$$

где $\mathbf{b} : \mathbb{R}^d \to \mathbb{R}^d$ такой что $\forall x \in \mathbb{R}^d$: $\|\mathbf{b}(x)\|^2 \leq \zeta^2$.

Предположение 4. (Ограниченность градиентного шума)

Существует параметр $\sigma_*^2 \geq 0$ такой что $\forall x \in \mathbb{R}^d$

$$\mathbb{E}\left[\|\nabla f(x^*, \xi) - \nabla f(x^*)\|^2\right] \leq \sigma_*^2.$$

Сходимость алгоритма AC-SA из [6] для перепараметризованной задачи

$$\mathbb{E}\left[f(x_N^{ag}) - f^*\right] \leq c \cdot \left(\frac{LR^2}{N^2} + \frac{LR^2}{BN} + \sqrt{\frac{LR^2 f^*}{BN}}\right).$$

Theorem (Сходимость Смещенного AC-SA метода)

Пусть f удовлетворяет Предположениям 1–3 и градиентный оракул из Определения 1 удовлетворяет Предположению 4, тогда смещенный AC-SA алгоритм гарантирует сходимость с универсальной константой c

$$\mathbb{E}\left[f(x_N^{ag}) - f^*\right] \leq c \cdot \left(\frac{LR^2}{N^2} + \frac{LR^2}{BN} + \frac{\sigma_* R}{\sqrt{BN}} + \zeta R + \frac{\zeta^2}{2L} N\right).$$

Algorithm 1 Accelareted Zero-Order Stochastic Gradient Descent (AZO-SGD)

Input: Start point $x_0^{ag} = x_0 \in \mathbb{R}^d$, maximum number of iterations $N \in \mathbb{Z}_+$.
Let stepsize $\eta_k > 0$, parameters $\beta_k, \tau > 0$, batch size $B \in \mathbb{Z}_+$.

1: **for** $k = 0, ..., N-1$ **do**
2: $\beta_k = 1 + \frac{k}{6}$ and $\eta_k = \eta(k+1)$ for $\gamma = \min\left\{\frac{1}{12L}, \frac{B}{24L(N+1)}, \sqrt{\frac{BR^2}{Lf^*N^3}}\right\}$
3: $x_k^{md} = \beta_k^{-1} x_k + (1 - \beta_k^{-1}) x_k^{ag}$
4: Sample $\{e_1, ..., e_B\}$ and $\{\xi_1, ..., \xi_B\}$ independently
5: Define $\mathbf{g}_k = \frac{1}{B}\sum_{i=1}^{B} \mathbf{g}(x_k^{md}, \xi_i, e_i)$ using (3)
6: $\tilde{x}_{k+1} = x_k - \eta_k \mathbf{g}_k$
7: $x_{k+1} = \min\left\{1, \frac{R}{\|\tilde{x}_{k+1}\|}\right\} \tilde{x}_{k+1}$
8: $x_{k+1}^{ag} = \beta_k^{-1} x_{k+1} + (1 - \beta_k^{-1}) x_{k+1}^{ag}$
9: **end for**

Output: x_N^{ag}.

Смещение аппроксимации градиента

$$
\begin{aligned}
\|\mathbb{E}\left[\mathbf{g}(x_k,\xi,e)\right] - \nabla f(x_k)\| &= \left\| \mathbb{E}\left[\frac{d}{2\tau}\left(f_\delta(x_k+\tau e,\xi) - f_\delta(x_k-\tau e)\right)e\right] - \nabla f(x_k)\right\| \\
&\overset{①}{=} \left\| \mathbb{E}\left[\frac{d}{\tau}\left(f(x_k+\tau e,\xi) + \delta(x_k+\tau e)\right)e\right] - \nabla f(x_k)\right\| \\
&\overset{②}{\leq} \left\| \mathbb{E}\left[\frac{d}{\tau} f(x_k+\tau e,\xi)e\right] - \nabla f(x_k)\right\| + \frac{d\Delta}{\tau} \\
&\overset{③}{=} \|\mathbb{E}\left[\nabla f(x_k+\tau u,\xi)\right] - \nabla f(x_k)\| + \frac{d\Delta}{\tau} \\
&= \sup_{z\in S_2^d(1)} \mathbb{E}\left[|\nabla_z f(x_k+\tau u) - \nabla_z f(x_k)|\right] + \frac{d\Delta}{\tau} \\
&\leq L\tau\mathbb{E}\left[\|u\|\right] + \frac{d\Delta}{\tau} \leq L\tau + \frac{d\Delta}{\tau}.
\end{aligned}
$$

Второй момент аппроксимации градиента

$$
\begin{aligned}
\mathbb{E}\left[\|\mathbf{g}(x^*,\xi,e)\|^2\right] &= \frac{d^2}{4\tau^2}\mathbb{E}\left[\|(f_\delta(x^*+\tau e,\xi)-f_\delta(x^*-\tau e,\xi))\,e\|^2\right] \\
&= \frac{d^2}{4\tau^2}\mathbb{E}\left[(f(x^*+\tau e,\xi)-f(x^*-\tau e,\xi)+\delta(x^*+\tau e)-\delta(x^*-\tau e))^2\right] \\
&\le \frac{d^2}{2\tau^2}\left(\mathbb{E}\left[(f(x^*+\tau e,\xi)-f(x^*-\tau e,\xi))^2\right]+2\Delta^2\right) \\
&\le \frac{d^2}{2\tau^2}\left(\frac{\tau^2}{d}\mathbb{E}\left[\|\nabla f(x^*+\tau e,\xi)+\nabla f(x^*-\tau e,\xi)\|^2\right]+2\Delta^2\right) \\
&= \frac{d^2}{2\tau^2}\left(\frac{\tau^2}{d}\mathbb{E}\left[\|\nabla f(x^*+\tau e,\xi)+\nabla f(x^*-\tau e,\xi)\pm 2\nabla f(x^*,\xi)\|^2\right]+2\Delta^2\right) \\
&\le 4d\,\|\nabla f(x^*,\xi)\|^2+4dL^2\tau^2\mathbb{E}\left[\|e\|^2\right]+\frac{d^2\Delta^2}{\tau^2} \\
&\le 4d\sigma_*^2+4dL^2\tau^2\mathbb{E}\left[\|e\|^2\right]+\frac{d^2\Delta^2}{\tau^2}.
\end{aligned}
$$

Сходимость безградиентного алгоритма

$$\mathbb{E}\left[f(x_N^{ag}) - f^*\right] \lesssim \underbrace{\frac{LR^2}{N^2}}_{①} + \underbrace{\frac{LR^2}{BN}}_{②} + \underbrace{\frac{\sqrt{d}\sigma_* R}{\sqrt{BN}}}_{③} + \underbrace{\frac{\sqrt{d}L\tau R}{\sqrt{BN}}}_{④} + \underbrace{\frac{d\Delta R}{\tau\sqrt{BN}}}_{⑤}$$
$$+ \underbrace{L\tau R}_{⑥} + \underbrace{\frac{d\Delta R}{\tau}}_{⑦} + \underbrace{L\tau^2 N}_{⑧} + \underbrace{\frac{d^2\Delta^2 N}{\tau^2 L}}_{⑨}.$$

Theorem

Пусть f удовлетворяет Предположениям 1–3 и аппроксимация градиента с параметром $\tau \leq \frac{\varepsilon}{LR}$ удовлетворяет Предположению 4, тогда Accelareted Zero-Order Stochastic Gradient Descent (AZO-SGD) Method (см. Алгоритм 1) достигает ε-точности: $\mathbb{E}\left[f(x_N^{ag}) - f^*\right] \leq \varepsilon$ после

$$N = \mathcal{O}\left(\sqrt{\frac{LR^2}{\varepsilon}}\right), \quad T = \max\left\{\mathcal{O}\left(\frac{LR^2}{\varepsilon}\right), \mathcal{O}\left(\frac{d\sigma_*^2 R^2}{\varepsilon^2}\right)\right\}$$

числа итераций, общего числа обращений к безградиентному оракулу и при

$$\Delta \leq \frac{\varepsilon^2}{dLR^2}$$

максимально допустимом уровне враждебного шума.

Обобщение на задачу с ограничением и расширение класса функций (L_p норма)

Замечание 1.

Используя работу [7] (а именно, Алгоритм 2) в качестве основы, и аналогичным образом обобщая результаты сходимости (Следствие 1, [7]) на случай с градиентным оракулом, можно получить безградиентный алгоритм для более общего класса задач (L_p-норма и наличие ограничений). В этом случае, параметры (число последовательных итераций N, максимально допустимый уровень враждебного шума Δ, параметр сглаживания τ) останутся такими же, за исключением общего числа обращений к безградиентному оракулу: пусть дано $1/p + 1/q = 1$, тогда

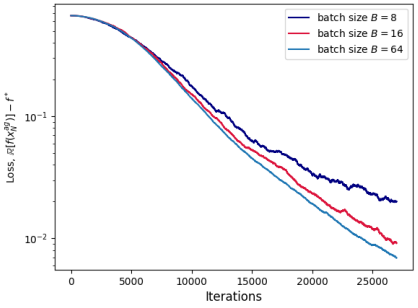
$$T = N \cdot B = \max\left\{\mathcal{O}\left(\frac{LR^2}{\varepsilon}\right), \mathcal{O}\left(\frac{\min\{q, \ln d\} d^{2-\frac{2}{p}} \sigma_*^2 R^2}{\varepsilon^2}\right)\right\}.$$

Эксперименты

Мы рассматриваем следующую задачу оптимизации:

$$\min_{x \in \mathbb{R}^d} f(x) := \frac{1}{m} \sum_{i=1}^{m} (l_i(x))^2,$$

где $l(x) = Ax - b$ – система m линейных уравнений в условии перепараметризации ($d > m$), $A \in \mathbb{R}^{m \times d}$, $x, b \in \mathbb{R}^d$. Данная задача является выпуклой стохастической задачей оптимизации, также известной как the Empirical Risk Minimization problem, где $\xi = i$ – одна из m линейных уравнений.



Содержание

Предположения на целевую функцию

Постановка задачи

Мы рассматриваем стандартную выпуклую задачу оптимизации:

$$\min_{x \in Q \subseteq \mathbb{R}^d} f(x)$$

где $f : \mathbb{R}^d \to \mathbb{R}$ выпуклая функция, которую хотим минимизировать на выпуклом мн-ве Q.

Предположение. (Повышенная гладкость функции)

Пусть l обозначает максимальное целое число строго меньше, чем β. Пусть $\mathcal{F}_\beta(L)$ обозначает множество всех функций $f : \mathbb{R}^d \to \mathbb{R}$, которые дифференцируемы l раз и для всех $x, z \in Q$ удовлетворяет условию Гельдера:

$$\left| f(z) - \sum_{0 \le |n| \le l} \frac{1}{n!} D^n f(x)(z-x)^n \right| \le L_\beta \|z - x\|_2^\beta,$$

где $L_\beta > 0$, $n = (n_1, ..., n_d)$ – мультиидекс, $n_i \ge 0$ – целые, $n! = n_1! \cdots n_d!$, $|n| = n_1 + \cdots + n_d$, и $\forall v = (v_1, ..., v_d) \in \mathbb{R}^d$, а также $D^n f(x) v^n = \frac{\partial^{|n|} f(x)}{\partial^{n_1} x_1 \cdots \partial^{n_d} x_d} v_1^{n_1} \cdots v_d^{n_d}$.

Предположения на оракул нулевого порядка

Оракул нулевого порядка

Мы предполагаем, что оракул $\tilde{f}$ может возвращать только значение целевой функции $f(x)$ в запрашиваемой точке x с некоторым стохастическим шумом ξ:

$$\tilde{f} = f(x) + \xi.$$

Предположение. (Стохастический шум)

Мы предполагаем, что выполняется следующее

- $\xi_1 \neq \xi_2$ такой что $\mathbb{E}[\xi_1^2] \leq \Delta^2$ и $\mathbb{E}[\xi_2^2] \leq \Delta^2$, $\Delta \geq 0$ – это уровень шума;
- случайные величины ξ_1 и ξ_2 являются независимыми от $\mathbf{e}$ и r.

Градиентная аппроксимация с одноточечной обратной связью

$$\mathbf{g}(x, \mathbf{e}) = d\frac{f(x + hr\mathbf{e}) + \xi_1 - f(x - hr\mathbf{e}) - \xi_2}{2h} K(r)\mathbf{e}.$$

Предположения на градиентную аппроксимацию

Определение. (Ядерная аппроксимация)

$$\mathbf{g}(x, \mathbf{e}) = d\frac{f(x + hr\mathbf{e}) + \xi_1 - f(x - hr\mathbf{e}) - \xi_2}{2h}K(r)\mathbf{e}.$$

где $h > 0$ – параметр сглаживания, $\mathbf{e} \in S_2^d(1)$ – равномерно распределенный вектор на единичной евклидовой сфере, r – случайный вектор равномерно распределенный на отрезке $r \in [0, 1]$.

Предположения на градиентную аппроксимацию

Определение. (Ядерная аппроксимация)

$$\mathbf{g}(x, \mathbf{e}) = d\frac{f(x + hr\mathbf{e}) + \xi_1 - f(x - hr\mathbf{e}) - \xi_2}{2h} K(r)\mathbf{e}.$$

где $h > 0$ – параметр сглаживания, $\mathbf{e} \in S_2^d(1)$ – равномерно распределенный вектор на единичной евклидовой сфере, r – случайный вектор равномерно распределенный на отрезке $r \in [0, 1]$.

Предположение. (Ядерная функция)

Пусть $K : [-1, 1] \to \mathbb{R}$ – Ядерная функция, которая удовлетворяет

$$\mathbb{E}[K(u)] = 0,\ \mathbb{E}[uK(u)] = 1,$$
$$\mathbb{E}[u^j K(u)] = 0,\ j = 2, ..., l,\ \mathbb{E}[|u|^\beta |K(u)|] < \infty.$$

Предположения на градиентную аппроксимацию

Определение. (Ядерная аппроксимация)

$$\mathbf{g}(x, \mathbf{e}) = d\frac{f(x + hr\mathbf{e}) + \xi_1 - f(x - hr\mathbf{e}) - \xi_2}{2h} K(r)\mathbf{e}.$$

где $h > 0$ – параметр сглаживания, $\mathbf{e} \in S_2^d(1)$ – равномерно распределенный вектор на единичной евклидовой сфере, r – случайный вектор равномерно распределенный на отрезке $r \in [0, 1]$.

Предположение. (Ядерная функция)

Пусть $K : [-1, 1] \to \mathbb{R}$ – Ядерная функция, которая удовлетворяет

$$\mathbb{E}[K(u)] = 0,\ \mathbb{E}[uK(u)] = 1,$$
$$\mathbb{E}[u^j K(u)] = 0,\ j = 2, ..., l,\ \mathbb{E}[|u|^\beta |K(u)|] < \infty.$$

Примеры Ядерной функции

Примером таких ядер является взвешенная сумма полиномов Лежандра.

Что было сделано ранее?

References	Iteration Complexity	Maximum Noise Level
Bach, Perchet (2016) [1]	$\mathcal{O}\left(\frac{d^{2+\frac{2}{\beta-1}}\Delta^2}{\varepsilon^{2+\frac{2}{\beta-1}}}\right)$	✗
Novitskii, Gasnikov (2020) [2]	$\tilde{\mathcal{O}}\left(\frac{d^{1+\frac{1}{\beta-1}}\Delta^2}{\varepsilon^{2+\frac{2}{\beta-1}}}\right)$	✗
Akhavan, Chzhen, Pontil, Tsybakov (2023) [3]	$\tilde{\mathcal{O}}\left(\frac{d^2\Delta^2}{\varepsilon^{2+\frac{2}{\beta-1}}}\right)$	✗

Is estimation on iteration complexity unimprovable?

What is the maximum noise level that can be taken?

Постановка задачи

Мы переформулируем исходную задачу оптимизации следующим образом:

$$f^* = \min_{x \in Q \subseteq \mathbb{R}^d} \{f(x) := \mathbb{E}[f(x, \xi)]\}.$$

Предположение (Выпуклость)

Функция f – выпукла, если

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle, \quad \forall x, y \in Q.$$

Предположение (L-гладкость)

Функция f – L-гладкая, если

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2, \quad \forall x, y \in Q.$$

Определение (Смещенный градиентный оракул)

Отображение $\mathbf{g} : \mathbb{R}^d \times \mathcal{D} \to \mathbb{R}^d$ такое что

$$\mathbf{g}(x, \xi) = \nabla f(x, \xi) + \mathbf{b}(x)$$

для смещения $\mathbf{b} : \mathbb{R}^d \to \mathbb{R}^d$ и несмещенного стох. градиента $\mathbb{E}\left[\nabla f(x, \xi)\right] = \nabla f(x)$.

Предположение (Ограниченное смещение)

Существует константы $\delta \geq 0$ такие что $\forall x \in \mathbb{R}^d$

$$\|\mathbf{b}(x)\| = \|\mathbb{E}\left[\mathbf{g}(x, \xi)\right] - \nabla f(x)\| \leq \delta.$$

Предположение (Ограниченный шум)

Существует константы $\rho, \sigma^2 \geq 0$ такие что обобщенное условие сильного роста выполняется $\forall x \in \mathbb{R}^d$

$$\mathbb{E}\left[\|\mathbf{g}(x, \xi)\|^2\right] \leq \rho \|\nabla f(x)\|^2 + \sigma^2.$$

Сходимость ускоренного алгоритма [8]

$$\mathbb{E}\left[f(x_N)\right] - f^* \lesssim \frac{\rho^2 L R^2}{N^2} + \frac{N\sigma^2}{\rho^2 L}$$

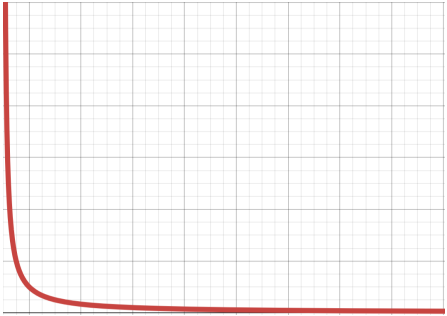


Рис.: Случай без смещения

Сходимость ускоренного алгоритма (со смещенным градиентным оракулом)

$$\mathbb{E}\left[f(x_N)\right] - f^* \lesssim \frac{\rho_B^2 L R^2}{N^2} + \frac{N\sigma^2}{\rho_B^2 LB} + \underbrace{\delta\tilde{R}}_{③} + \underbrace{\frac{N}{L}\delta^2}_{④}.$$

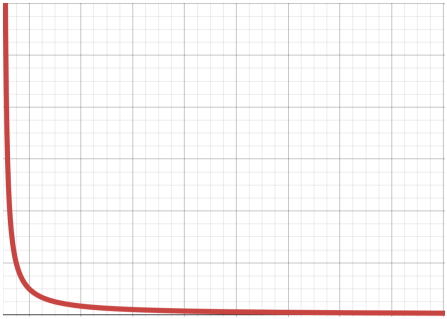


Рис.: Случай без смещения

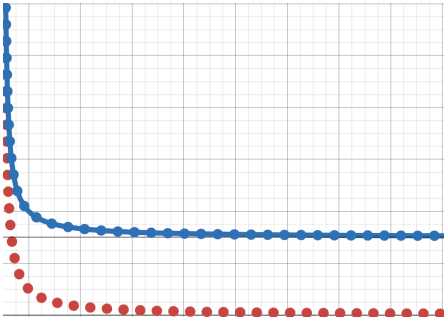


Рис.: Случай со смещением

Algorithm 1 Zero-Order Accelerated Stochastic Gradient Descent

Input: iteration number N, batch size B, Kernel $K : [-1, 1] \to \mathbb{R}$, step size η, smoothing parameter h, $x_0 = y_0 = z_0 \in \mathbb{R}^d$, $\alpha_0 = \gamma_0 = 0$.

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

1. Sample vectors $\mathbf{e}_1, \mathbf{e}_2 ..., \mathbf{e}_B$ uniformly distributed on the unit sphere $S_2^d(1)$ and scalars $r_1, r_2, ..., r_B$ uniformly distributed on the interval [-1, 1] independently
2. Define $\mathbf{g}(x_k, \mathbf{e}_i) = d\frac{\tilde{f}(x_k + hr_i\mathbf{e}_i) - \tilde{f}(x_k - hr_i\mathbf{e}_i)}{2h} K(r_i)\mathbf{e}_i$ via (6)
3. Calculate $\mathbf{g}_k = \frac{1}{B}\sum_{i=1}^{B} \mathbf{g}(x_k, \mathbf{e}_i)$
4. $x_{k+1} \leftarrow y_k - \eta \mathbf{g}_k$
5. $z_{k+1} \leftarrow z_k - \gamma_k \eta \mathbf{g}_k$
6. $y_{k+1} \leftarrow \alpha_{k+1} z_{k+1} + (1 - \alpha_{k+1}) x_{k+1}$

end for

Return: x_N

Нахождение смещения градиентной аппроксимации

Смещение градиентной аппроксимации [9]

$$\|\mathbf{b}(x)\| = \|\mathbb{E}\left[\mathbf{g}(x_k, \mathbf{e})\right] - \nabla f(x_k)\| \lesssim \kappa_\beta L h^{\beta-1}.$$

Второй момент градиентной аппроксимации [9]

$$\mathbb{E}\left[\|\mathbf{g}(x_k, \mathbf{e})\|^2\right] \leq \underbrace{4d\kappa}_{\rho} \|\nabla f(x_k)\|^2 + \underbrace{4d\kappa L^2 h^2 + \frac{\kappa d^2 \Delta^2}{h^2}}_{\sigma^2}$$

Сходимость безградиентного алгоритма

$$\mathbb{E}\left[f(x_N)\right] - f^* \lesssim \underbrace{\frac{\rho_B^2 L R^2}{N^2}}_{①} + \underbrace{\frac{N d\kappa L^2 h^2}{\rho_B^2 L B}}_{②} + \underbrace{\frac{N\kappa d^2 \Delta^2}{h^2 \rho_B^2 L B}}_{③} + \underbrace{\tilde{R}\kappa_\beta L_\beta h^{\beta-1}}_{④} + \underbrace{\frac{N\kappa_\beta^2 L_\beta^2 h^{2(\beta-1)}}{L}}_{⑤}.$$

Основная теорема

Theorem 4.1 (Convergence results) *Let the function f satisfy Assumption 2.2 and the gradient approximation $\mathbf{g}(x, \mathbf{e})$ of (7) satisfies Assumptions 2.3 and 2.4, then Zero-Order Accelerated Stochastic Gradient Descent (see Algorithm 1) with $\rho_B = \max\{1, \frac{4d\kappa}{B}\}$, and with the chosen algorithm parameters:*

$$\gamma_k = \frac{\rho_B^{-1} + \sqrt{\rho_B^{-2} + 4\gamma_{k-1}^2}}{2}; \quad a_{k+1} = \gamma_k \sqrt{\eta \rho_B}; \quad \alpha_k = \frac{\gamma_k \eta}{\gamma_k \eta + a_k^2}; \quad \eta = \frac{1}{\rho_B L}$$

converges to the desired ε accuracy, $\mathbb{E}\left[f(x_N)\right] - f^* \leq \varepsilon$

- *in the case $B \in [1, 4d\kappa]$, $h \lesssim \varepsilon^{3/4}$ and $\beta \geq \frac{7}{3}$ after*

$$N = \mathcal{O}\left(\sqrt{\frac{d^2 L R^2}{B^2 \varepsilon}}\right); \qquad T = N \cdot B = \mathcal{O}\left(\sqrt{\frac{d^2 L R^2}{\varepsilon}}\right)$$

number of iterations and gradient-free oracle calls, respectively, at

$$\Delta \lesssim \frac{\varepsilon^{3/2}}{\sqrt{d}} \quad \text{maximum noise level;}$$

- *in the case $B > 4d\kappa$ and $h \lesssim \varepsilon^{1/(\beta-1)}$ after*

$$N = \mathcal{O}\left(\sqrt{\frac{L R^2}{\varepsilon}}\right); \quad T = N \cdot B = \max\left\{\mathcal{O}\left(\sqrt{\frac{d^2 L R^2}{\varepsilon}}\right), \mathcal{O}\left(\frac{d^2 \Delta^2}{\varepsilon^{2 + \frac{2}{\beta - 1}}}\right)\right\}$$

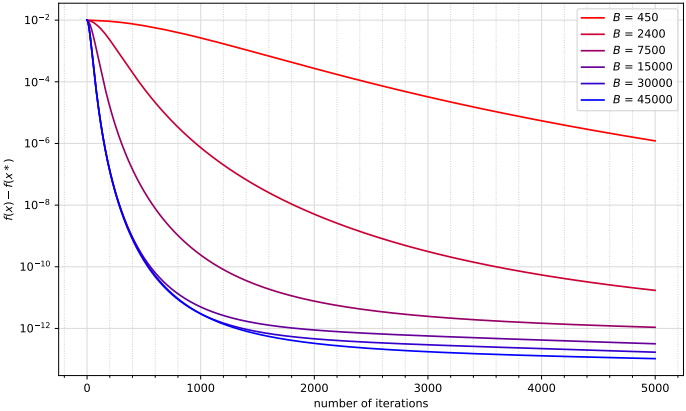
number of iterations and gradient-free oracle calls, respectively, at

$$\Delta \lesssim \frac{\varepsilon^{\frac{3\beta+1}{4(\beta-1)}}}{d} B^{1/2} \quad \text{maximum noise level;}$$

Эксперименты

Функция для задачи минимизации

$$f(w) := -y \log \left[\frac{1}{1 + \exp\left(-w^\mathsf{T} X\right)} \right] + (1 - y) \log \left[1 - \frac{1}{1 + \exp\left(-w^\mathsf{T} X\right)} \right].$$



Содержание

Постановка задачи

В данной работе рассматривается решение стандартной общей оптимизационной задачи в следующем виде:

$$\min_{x \in \mathbb{R}^d} \{f(x) := \mathbb{E}_{\xi \sim \mathcal{D}} f_\xi(x)\}, \tag{1}$$

где $f : \mathbb{R}^d \to \mathbb{R}$ - возможно невыпуклая, возможно стохастическая функция.

Такая конфигурация задачи охватывает широкий спектр приложений в сценариях машинного обучения, таких как минимизация эмпирического риска (ERM), где $\mathcal{D}$ обозначает распределение по обучающим точкам данных, а $f_\xi(x)$ представляет собой потери модели x по точке данных ξ.

Оракул высокого порядка

- $G_{x_k,i} = \nabla^i f(x_k)$ например, в Тензорных методах [10—12]

Оракул высокого порядка

- $G_{x_k,i} = \nabla^i f(x_k)$ например, в Тензорных методах [10—12]

Оракул первого порядка

- $g = \nabla f(x_k)$ например, в методах типа GD/SGD [13—15]
- $g = \nabla f(x_k) + b(x_k)$ например, в смещенных методах типа GD/SGD [16—18]
- $g = \nabla_{i_k} f(x_k)$ например, в методах типа CD [19—21]

Background

Оракул высокого порядка

- $G_{x_k,i} = \nabla^i f(x_k)$ например, в Тензорных методах [10—12]

Оракул первого порядка

- $g = \nabla f(x_k)$ например, в методах типа GD/SGD [13—15]
- $g = \nabla f(x_k) + b(x_k)$ например, в смещенных методах типа GD/SGD [16—18]
- $g = \nabla_{i_k} f(x_k)$ например, в методах типа CD [19—21]

Оракул нулевого порядка

- $\tilde{f} = f(x_k)$ например, в безградиентных алгоритмах [22, 23]
- $\tilde{f} = f(x_k) + \delta(x_k)$ например, в безградиентных алгоритмах [24, 25]
- $\tilde{f} = f(x_k) + \xi$ например, в безградиентных алгоритмах [26, 27]

A.1 Perfect coffee for everyone

As already demonstrated in Section 1 (Introduction) with the example of chocolate, the deterministic concept of the Order Oracle has many potential applications. However, this research was initiated due to a challenge one of the co-authors faced during the realization of *a startup: the creation of an ideal coffee machine that can make the perfect drink for each customer*. This startup has just started its life cycle. At the moment we have designed a coffee machine that is functioning at the testing stage (see the photo of the machine in Figure 3 and the 3D model in Figure 4).

Figure 3: Smart coffee machine.

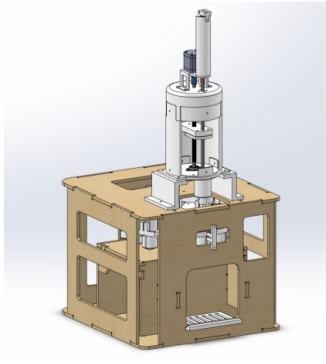


Figure 4: 3D model of a smart coffee machine.

Рис.: Шоколад Valio

Такие компании, как Valio (см. рисунок) и SberAI, уже активно привлекают ИИ к созданию десертов. Фотография взята с сайта компании Официальный сайт.

Концепция оракула приближенная к реальности

The Order Oracle

Мы рассматриваем концепцию оракула нулевого порядка (*Order Oracle*) для решения задачи (1):

$$\phi(x, y) = \text{sign}\left[f(x) - f(y) + \delta(x, y)\right], \tag{2}$$

где $|\delta(x, y)| \leq \Delta$ – некоторый ограниченный шум.

Концепция оракула приближенная к реальности

The Order Oracle

Мы рассматриваем концепцию оракула нулевого порядка (*Order Oracle*) для решения задачи (1):

$$\phi(x, y) = \text{sign}\left[f(x) - f(y) + \delta(x, y)\right], \tag{2}$$

где $|\delta(x, y)| \leq \Delta$ – некоторый ограниченный шум.

Q: Какие методы дружат с этим оракулом?

A: Методы решения задачи линейного поиска (метод золотого сечения (GRM) и др.)

Концепция оракула приближенная к реальности

The Order Oracle

Мы рассматриваем концепцию оракула нулевого порядка (*Order Oracle*) для решения задачи (1):

$$\phi(x, y) = \text{sign}\left[f(x) - f(y) + \delta(x, y)\right], \tag{2}$$

где $|\delta(x, y)| \leq \Delta$ – некоторый ограниченный шум.

Q: Какие методы дружат с этим оракулом?

A: Методы решения задачи линейного поиска (метод золотого сечения (GRM) и др.)

Проблема: эти методы решают одномерную задачу оптимизации

Как интегрировать этот оракул в многомерную оптимизацию?

Метод координатного спуска (CDM)

На каждой итерации CDM, заданный размером шага $\zeta_k > 0$ и начальной точкой $x_0 \in \mathbb{R}^d$, работает следующим образом:

$$x_{k+1} = x_k - \underbrace{\zeta_k \nabla_{i_k} f(x_k)}_{\eta_k} \mathbf{e}_{i_k},$$

где i_k - индекс координат, взятый из $[d]$.

Как интегрировать этот оракул в многомерную оптимизацию?

Метод координатного спуска (CDM)

На каждой итерации CDM, заданный размером шага $\zeta_k > 0$ и начальной точкой $x_0 \in \mathbb{R}^d$, работает следующим образом:

$$x_{k+1} = x_k - \underbrace{\zeta_k \nabla_{i_k} f(x_k)}_{\eta_k} \mathbf{e}_{i_k},$$

где i_k - индекс координат, взятый из $[d]$.

Предположения (Гладкость)

Функция $f : \mathbb{R}^d \to \mathbb{R}$ является L-координатно-липшицевой для $L_1, L_2, ..., L_d > 0$, если для любых $i \in [d]$, $x \in \mathbb{R}^d$ и $h \in \mathbb{R}$ выполняется следующее неравенство:

$$|\nabla_i f(x + h\mathbf{e}_i) - \nabla_i f(x)| \leq L_i |h|.$$

Как интегрировать этот оракул в многомерную оптимизацию?

Метод координатного спуска (CDM)

На каждой итерации CDM, заданный размером шага $\zeta_k > 0$ и начальной точкой $x_0 \in \mathbb{R}^d$, работает следующим образом:

$$x_{k+1} = x_k - \underbrace{\zeta_k \nabla_{i_k} f(x_k)}_{\eta_k} \mathbf{e}_{i_k},$$

где i_k - индекс координат, взятый из $[d]$.

Предположение ($\mu_{1-\alpha}$ (сильной) выпуклости)

Функция $f : \mathbb{R}^d \to \mathbb{R}$ является $\mu_{1-\alpha} \geq 0$ сильно выпуклой относительно нормы $\|\cdot\|_{[1-\alpha]}$, если для любых x, y $\mathbb{R}^d$ выполняется следующее неравенство:

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu_{1-\alpha}}{2} \|y - x\|^2_{[1-\alpha]},$$

где $\|\cdot\|_{[\alpha]} := \sqrt{\sum_{i=1}^d L_i^\alpha x_i^2}$.

Algorithm 1 Random Coordinate Descent with Order Oracle (OrderRCD)

Input: $x_0 \in \mathbb{R}^d$, random generator $\mathcal{R}_\alpha(L)$
for $k = 0$ **to** $N - 1$ **do**
1. choose active coordinate $i_k = \mathcal{R}_\alpha(L)$
2. compute $\eta_k = \operatorname{argmin}_\eta \{f(x_k + \eta \mathbf{e}_{i_k})\}$ via (GRM)
3. $x_{k+1} \leftarrow x_k + \eta_k \mathbf{e}_{i_k}$

end for
Return: x_N

$$p_\alpha(i) = L_i^\alpha / S_\alpha, \quad i \in [d];$$

$$\mathbb{E}\left[f(x_N)\right] - f(x^*) \leq \left(1 - \frac{\mu_{1-\alpha}}{S_\alpha}\right)^N F_0.$$

Результаты сходимости

Table 1: Comparison of oracle complexity for the methods proposed in this work with SOTA methods both in the coordinate descent class and in the Order Oracle concept (2). Notation: $F_0 = f(x_0) - f(x^*)$; $R = \|x_0 - x^*\|$; $R_{[1-\alpha]} = R_{1-\alpha}(x_0) = \sup_{x \in \mathbb{R}^d : f(x) \le f(x_0)} \|x - x^*\|_{[1-\alpha]}$; $\varepsilon =$ desired accuracy of problem solving.

Reference	Nesterov (2012)	Gorbunov et al. (2019)	Saha et al. (2021)	Tang et al. (2023)	This paper	
Non-convex	✗	$\mathcal{O}\left(\frac{dS_\alpha F_0}{\varepsilon^2}\right)$	✗	$\mathcal{O}\left(\frac{dLF_0}{\varepsilon^2}\right)$	$\tilde{\mathcal{O}}\left(\frac{S_\alpha F_0}{\varepsilon^2}\right)$	
Convex	$\mathcal{O}\left(\frac{S_\alpha R^2_{[1-\alpha]}}{\varepsilon}\right)$	$\mathcal{O}\left(\frac{dS_\alpha R^2_{[1-\alpha]}}{\varepsilon}\log\frac{1}{\varepsilon}\right)$	$\mathcal{O}\left(\frac{dLR^2}{\varepsilon}\right)$	✗	$\tilde{\mathcal{O}}\left(\frac{S_\alpha R^2_{[1-\alpha]}}{\varepsilon}\right)$	
Strongly convex	$\mathcal{O}\left(\frac{S_\alpha}{\mu_{1-\alpha}}\log\frac{1}{\varepsilon}\right)$	$\mathcal{O}\left(\frac{S_\alpha}{\mu_{1-\alpha}}\log\frac{1}{\varepsilon}\right)$	$\mathcal{O}\left(d\frac{L}{\mu}\log\frac{1}{\varepsilon}\right)$	✗	$\tilde{\mathcal{O}}\left(\frac{S_\alpha}{\mu_{1-\alpha}}\log\frac{1}{\varepsilon}\right)$	$\tilde{\mathcal{O}}\left(\frac{S_{\alpha/2}}{\sqrt{\mu_{1-\alpha}}}\log\frac{1}{\varepsilon}\right)$
Order Oracle?	✗	✓	✓	✓	✓	✓
Acceleration?	✗	✗	✗	✗	✗	✓

Algorithm 2 Accelerated Coordinate Descent Method with Order Oracle (OrderACDM)

Input: $x_0 = z_0 \in \mathbb{R}^d$, $\mathcal{R}_\alpha(L)$, $A_0 = 0$, $B_0 = 1$, $\beta = \frac{\alpha}{2}$
for $k = 0$ **to** $N - 1$ **do**

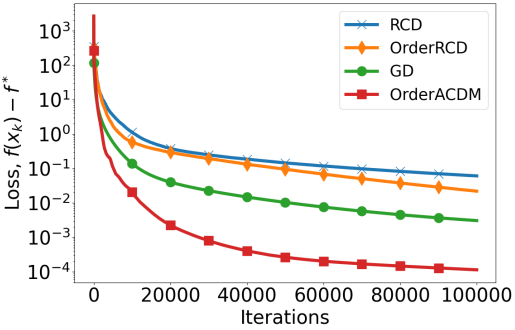
1. choose active coordinate $i_k = \mathcal{R}_\beta(L)$
2. find parameter a_{k+1} from $a_{k+1}^2 S_\beta^2 = A_{k+1} B_{k+1}$, where $A_{k+1} = A_k + a_{k+1}$ and $B_{k+1} = B_k + \mu_{1-\alpha} a_{k+1}$
3. $\alpha_k \leftarrow \frac{a_{k+1}}{A_{k+1}}$
4. $\beta_k \leftarrow \frac{\mu_{1-\alpha} a_{k+1}}{B_{k+1}}$
5. $y_k \leftarrow \frac{(1-\alpha_k)x_k + \alpha_k(1-\beta_k)z_k}{1-\alpha_k\beta_k}$
6. compute $\eta_k = \operatorname{argmin}_\eta \{f(y_k + \eta \mathbf{e}_{i_k})\}$ via (GRM)
7. $x_{k+1} \leftarrow y_k + \eta_k \mathbf{e}_{i_k}$
8. $w_k \leftarrow (1-\beta_k)z_k + \beta_k y_k + \frac{a_{k+1} L_{i_k}^\alpha}{B_{k+1} p_\beta(i)} \eta_k \mathbf{e}_{i_k}$
9. compute $\zeta_k = \operatorname{argmin}_\zeta \{f(w_k + \zeta \mathbf{e}_{i_k})\}$ via (GRM)
10. $z_{k+1} \leftarrow w_k + \zeta_k \mathbf{e}_{i_k}$

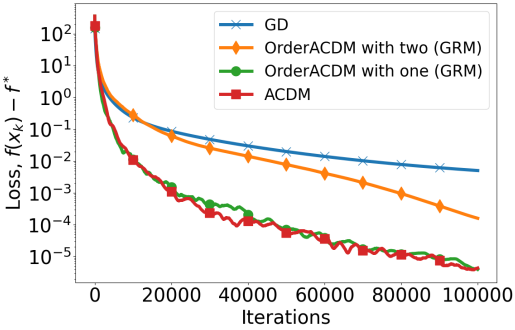
end for
Return: x_N

$$p_\beta(i) = L_i^\beta / S_\beta, \quad i \in [d];$$

$$\mathbb{E}\left[f(x_N)\right] - f(x^*) \leq \left(1 - \frac{\sqrt{\mu_{1-\alpha}}}{S_{\alpha/2}}\right)^N F_0.$$

Эксперименты





Спасибо за внимание!

Рис.: Связаться со мной

Reference I

[1] *Gasnikov A., Novitskii A., Novitskii V., Abdukhakimov F., Kamzolov D., Beznosikov A., Takáč M., Dvurechensky P., Gu B.* The Power of First-Order Smooth Optimization for Black-Box Non-Smooth Problems // 2022. URL: `https://arxiv.org/pdf/2201.12289.pdf`.

[2] Arya Akhavan и др. "A gradient estimator via L1-randomization for online zero-order optimization with two point feedback". В: *arXiv preprint arXiv:2205.13910* (2022).

[3] *Woodworth B., Bullins B., Shamir O., Srebro N.* The Min-Max Complexity of Distributed Stochastic Convex Optimization with Intermittent Communication. // Proceedings of Machine Learning Research. 2021. V. 134. P. 1–52.

[4] *Woodworth B., Patel K. K., Stich S. U., Dai Z., Bullins B., McMahan H. B., Shamir O., Srebro N.* Is Local SGD Better than Minibatch SGD? // Proceedings of the 37th International Conference on Machine Learning, PMLR. 2020. V. 119. P. 10334–10343.

[5] *Yuan H., Ma T.* Federated Accelerated Stochastic Gradient Descent. // Advances in Neural Information Processing Systems. 2020. V. 33. P. 5332–5344.

Reference II

[6] Blake E Woodworth и Nathan Srebro. "An even more optimal stochastic optimization algorithm: minibatching and interpolation learning". В: *Advances in Neural Information Processing Systems* 34 (2021), с. 7333—7345.

[7] Sasila Ilandarideva и др. "Accelerated stochastic approximation with state-dependent noise". В: *arXiv preprint arXiv:2307.01497* (2023).

[8] Sharan Vaswani, Francis Bach и Mark Schmidt. "Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron". В: *The 22nd international conference on artificial intelligence and statistics*. PMLR. 2019, с. 1195—1204.

[9] Arya Akhavan и др. "Gradient-free optimization of highly smooth functions: improved analysis and a new algorithm". В: *arXiv preprint arXiv:2306.02159* (2023).

[10] Bo Jiang, Haoyue Wang и Shuzhong Zhang. "An optimal high-order tensor method for convex optimization". В: *Conference on Learning Theory*. PMLR. 2019, с. 1799—1801.

[11] Yurii Nesterov. "Implementable tensor methods in unconstrained convex optimization". В: *Mathematical Programming* 186 (2021), с. 157—183.

Reference III

[12] Artem Agafonov и др. “Inexact tensor methods and their application to stochastic convex optimization”. В: *Optimization Methods and Software* (2023), с. 1—42.

[13] Eduard Gorbunov, Marina Danilova и Alexander Gasnikov. “Stochastic optimization with heavy-tailed noise via accelerated gradient clipping”. В: *Advances in Neural Information Processing Systems* 33 (2020), с. 15042—15053.

[14] Mert Gurbuzbalaban, Umut Simsekli и Lingjiong Zhu. “The heavy-tail phenomenon in SGD”. В: *International Conference on Machine Learning*. PMLR. 2021, с. 3964—3975.

[15] Feihu Huang и др. “Accelerated zeroth-order and first-order momentum methods from mini to minimax optimization”. В: *The Journal of Machine Learning Research* 23.1 (2022), с. 1616—1685.

[16] Ahmad Ajalloeian и Sebastian U Stich. “On the convergence of SGD with biased gradients”. В: *arXiv preprint arXiv:2008.00051* (2020).

[17] Ahmad Ajalloeian и Sebastian U Stich. “On the convergence of SGD with biased gradients”. В: *arXiv preprint arXiv:2008.00051* (2020).

Reference IV

[18] Margalit R Glasgow, Honglin Yuan и Tengyu Ma. "Sharp bounds for federated averaging (local SGD) and continuous perspective". В: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2022, с. 9050—9090.

[19] Yu Nesterov. "Efficiency of coordinate descent methods on huge-scale optimization problems". В: *SIAM Journal on Optimization* 22.2 (2012), с. 341—362.

[20] Yin Tat Lee и Aaron Sidford. "Efficient accelerated coordinate descent methods and faster algorithms for solving linear systems". В: *2013 ieee 54th annual symposium on foundations of computer science*. IEEE. 2013, с. 147—156.

[21] Paul Mangold и др. "High-Dimensional Private Empirical Risk Minimization by Greedy Coordinate Descent". В: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2023, с. 4894—4916.

[22] Ohad Shamir. "An optimal algorithm for bandit and zero-order convex optimization with two-point feedback". В: *The Journal of Machine Learning Research* 18.1 (2017), с. 1703—1713.

Reference V

[23] Alexander Gasnikov и др. “The power of first-order smooth optimization for black-box non-smooth problems”. В: *International Conference on Machine Learning*. PMLR. 2022, с. 7241—7265.

[24] Aleksandr Lobanov и др. “Gradient-Free Federated Learning Methods with l_1 and l_2-Randomization for Non-Smooth Convex Stochastic Optimization Problems”. В: *arXiv preprint arXiv:2211.10783* (2022).

[25] Nikita Kornilov и др. “Accelerated Zeroth-order Method for Non-Smooth Stochastic Convex Optimization Problem with Infinite Variance”. В: *Thirty-seventh Conference on Neural Information Processing Systems*. 2023.

[26] Francis Bach и Vianney Perchet. “Highly-smooth zero-th order online optimization”. В: *Conference on Learning Theory*. PMLR. 2016, с. 257—283.

[27] Arya Akhavan, Massimiliano Pontil и Alexandre Tsybakov. “Distributed zero-order optimization under adversarial noise”. В: *Advances in Neural Information Processing Systems* 34 (2021), с. 10209—10220.