

# Стохастическая оптимизация (продолжение).

## Координатный спуск

### Методы оптимизации

Александр Безносиков

Московский физико-технический институт

7 декабря 2023



## В прошлый раз

- Рассматривали постановку вида (оффлайн, ERM):

$$\min_{x \in \mathbb{R}^d} \left[ f(x) = \frac{1}{n} \sum_{i=1}^n f_i(x) \right].$$

- Предполагаем, что вызывать полный градиент дорого, но можно, генерируя равномерно и независимо  $i_k$ , получить

$$\mathbb{E}_{i_k} [\nabla f_{i_k}(x^k)] = \nabla f(x^k).$$

## В прошлый раз

- Рассматривали постановку вида (оффлайн, ERM):

$$\min_{x \in \mathbb{R}^d} \left[ f(x) = \frac{1}{n} \sum_{i=1}^n f_i(x) \right].$$

- Предполагаем, что вызывать полный градиент дорого, но можно, генерируя равномерно и независимо  $i_k$ , получить

$$\mathbb{E}_{i_k} [\nabla f_{i_k}(x^k)] = \nabla f(x^k).$$

- Идея – взять метод на подобию SGD:

$$x^{k+1} = x^k - \gamma g^k,$$

где

$$g^k \rightarrow \nabla f(x^*) = 0, \quad \text{при} \quad x^k \rightarrow x^*.$$

# SAGA

Уже знакомы

---

## Алгоритм 1 SAGA

---

**Вход:** размер шага  $\gamma > 0$ , стартовая точка  $x^0 \in \mathbb{R}^d$ , значения памяти  $y_i^0 = 0$  для всех  $i \in [n]$ , количество итераций  $K$

- 1: **for**  $k = 0, 1, \dots, K - 1$  **do**
- 2:   Сгенерировать независимо  $i_k$
- 3:   Вычислить  $g^k = \nabla f_{i_k}(x^k) - y_{i_k}^k + \frac{1}{n} \sum_{j=1}^n y_j^k$
- 4:   Обновить  $y_i^{k+1} = \begin{cases} \nabla f_i(x^k), & \text{если } i = i_k \\ y_i^k, & \text{иначе} \end{cases}$
- 5:    $x^{k+1} = x^k - \gamma g^k$
- 6: **end for**

**Выход:**  $x^K$

---

# SAGA

- Идея – если я считал когда-то градиент для  $f_i$ , то зачем его забывать? Сохраним!

# SAGA

- Идея – если я считал когда-то градиент для  $f_i$ , то зачем его забывать? Сохраним!
- $\frac{1}{n} \sum_{j=1}^n y_j^k$  – «запаздывающая» версия  $\nabla f(x^k)$ .

# SAGA

- Идея – если я считал когда-то градиент для  $f_i$ , то зачем его забывать? Сохраним!
- $\frac{1}{n} \sum_{j=1}^n y_j^k$  – «запаздывающая» версия  $\nabla f(x^k)$ .
- $\mathbb{E} [g^k \mid x^k] = \nabla f(x^k)$ .

# SAGA

- Идея – если я считал когда-то градиент для  $f_i$ , то зачем его забывать? Сохраним!
- $\frac{1}{n} \sum_{j=1}^n y_j^k$  – «запаздывающая» версия  $\nabla f(x^k)$ .
- $\mathbb{E}[g^k | x^k] = \nabla f(x^k)$ .
- При  $x^k \rightarrow x^*$  имеем, что  $y_j^k \rightarrow \nabla f_j(x^*)$ , и  $\frac{1}{n} \sum_{j=1}^n y_j^k \rightarrow \nabla f(x^*) = 0$ .  
А значит  $g^k \rightarrow 0$ .



# SAGA

- Идея – если я считал когда-то градиент для  $f_i$ , то зачем его забывать? Сохраним!
- $\frac{1}{n} \sum_{j=1}^n y_j^k$  – «запаздывающая» версия  $\nabla f(x^k)$ .
- $\mathbb{E}[g^k | x^k] = \nabla f(x^k)$ .
- При  $x^k \rightarrow x^*$  имеем, что  $y_j^k \rightarrow \nabla f_j(x^*)$ , и  $\frac{1}{n} \sum_{j=1}^n y_j^k \rightarrow \nabla f(x^*) = 0$ .  
А значит  $g^k \rightarrow 0$ .
- Из минусов: лишняя  $\mathcal{O}(nd)$  память.

# SAGA: доказательство

- Все  $f_i$  являются  $L$ -гладкими и выпуклыми, а  $f - \mu$  - сильно выпуклой.

# SAGA: доказательство

- Все  $f_i$  являются  $L$ -гладкими и выпуклыми, а  $f - \mu$  - сильно выпуклой.
- Уже привычно:

$$\|x^{k+1} - x^*\|_2^2 = \|x^k - x^*\|_2^2 - 2\gamma \langle g^k, x^k - x^* \rangle + \gamma^2 \|g^k - \nabla f(x^*)\|_2^2.$$

# SAGA: доказательство

- Все  $f_i$  являются  $L$ -гладкими и выпуклыми, а  $f - \mu$  - сильно выпуклой.
- Уже привычно:

$$\|x^{k+1} - x^*\|_2^2 = \|x^k - x^*\|_2^2 - 2\gamma \langle g^k, x^k - x^* \rangle + \gamma^2 \|g^k - \nabla f(x^*)\|_2^2.$$

- Берем условное мат.ожидание по случайности только на итерации  $k$ :

$$\begin{aligned} \mathbb{E} \left[ \|x^{k+1} - x^*\|_2^2 \mid x^k \right] &= \|x^k - x^*\|_2^2 - 2\gamma \langle \mathbb{E} [g^k \mid x^k], x^k - x^* \rangle \\ &\quad + \gamma^2 \mathbb{E} \left[ \|g^k - \nabla f(x^*)\|_2^2 \mid x^k \right]. \end{aligned}$$

# SAGA: доказательство

- Работаем с  $\mathbb{E} [g^k \mid x^k]$ :

$$\begin{aligned}\mathbb{E} [g^k \mid x^k] &= \mathbb{E} \left[ \nabla f_{i_k}(x^k) - y_{i_k}^k + \frac{1}{n} \sum_{j=1}^n y_j^k \mid x^k \right] \\ &= \mathbb{E} [\nabla f_{i_k}(x^k) - y_{i_k}^k \mid x^k] + \frac{1}{n} \sum_{j=1}^n y_j^k \\ &= \frac{1}{n} \sum_{j=1}^n [\nabla f_j(x^k) - y_j^k] + \frac{1}{n} \sum_{j=1}^n y_j^k \\ &= \nabla f(x^k)\end{aligned}$$

## SAGA: доказательство

- Теперь работаем с  $\mathbb{E} [\|g^k - \nabla f(x^*)\|_2^2 \mid x^k]$ :

$$\begin{aligned}\mathbb{E} [\|g^k - \nabla f(x^*)\|_2^2 \mid x^k] &= \mathbb{E} \left[ \left\| \nabla f_{i_k}(x^k) - y_{i_k}^k + \frac{1}{n} \sum_{j=1}^n y_j^k - \nabla f(x^*) \right\|_2^2 \mid x^k \right] \\ &= \mathbb{E} \left[ \left\| \nabla f_{i_k}(x^k) - \nabla f_{i_k}(x^*) + \nabla f_{i_k}(x^*) - y_{i_k}^k \right. \right. \\ &\quad \left. \left. + \frac{1}{n} \sum_{j=1}^n y_j^k - \nabla f(x^*) \right\|_2^2 \mid x^k \right] \\ &\leq 2\mathbb{E} \left[ \left\| \nabla f_{i_k}(x^k) - \nabla f_{i_k}(x^*) \right\|_2^2 \mid x^k \right] \\ &\quad + 2\mathbb{E} \left[ \left\| \nabla f_{i_k}(x^*) - y_{i_k}^k + \frac{1}{n} \sum_{j=1}^n y_j^k - \nabla f(x^*) \right\|_2^2 \mid x^k \right]\end{aligned}$$

## SAGA: доказательство

- Пользуемся тем, что  $\mathbb{D}\xi \leq \mathbb{E}[\xi^2]$ :

$$\begin{aligned}\mathbb{E} \left[ \|g^k - \nabla f(x^*)\|_2^2 \mid x^k \right] &\leq 2\mathbb{E} \left[ \|\nabla f_{i_k}(x^k) - \nabla f_{i_k}(x^*)\|_2^2 \mid x^k \right] \\ &\quad + 2\mathbb{E} \left[ \left\| \nabla f_{i_k}(x^*) - y_{i_k}^k + \frac{1}{n} \sum_{j=1}^n y_j^k - \nabla f(x^*) \right\|_2^2 \mid x^k \right] \\ &\leq 2\mathbb{E} \left[ \|\nabla f_{i_k}(x^k) - \nabla f_{i_k}(x^*)\|_2^2 \mid x^k \right] \\ &\quad + 2\mathbb{E} \left[ \|\nabla f_{i_k}(x^*) - y_{i_k}^k\|_2^2 \mid x^k \right]\end{aligned}$$

## SAGA: доказательство

- Берем мат.ожидание, пользуемся гладкостью (с выпуклостью):

$$\begin{aligned}\mathbb{E} \left[ \|g^k - \nabla f(x^*)\|_2^2 \mid x^k \right] &\leq 2\mathbb{E} \left[ \|\nabla f_{i_k}(x^k) - \nabla f_{i_k}(x^*)\|_2^2 \mid x^k \right] \\ &\quad + 2\mathbb{E} \left[ \|\nabla f_{i_k}(x^*) - y_{i_k}^k\|_2^2 \mid x^k \right] \\ &\leq 4L \cdot \frac{1}{n} \sum_{i=1}^n (f_i(x^k) - f_i(x^*) - \langle \nabla f_i(x^k), x^k - x^* \rangle) \\ &\quad + 2 \cdot \frac{1}{n} \sum_{i=1}^n \|\nabla f_i(x^*) - y_i^k\|_2^2 \\ &= 4L \cdot (f(x^k) - f(x^*)) \\ &\quad + 2 \cdot \frac{1}{n} \sum_{i=1}^n \|\nabla f_i(x^*) - y_i^k\|_2^2\end{aligned}$$



# SAGA: доказательство

- Промежуточный итог:

$$\begin{aligned}\mathbb{E} \left[ \|x^{k+1} - x^*\|_2^2 \mid x^k \right] &= \|x^k - x^*\|_2^2 - 2\gamma \langle \mathbb{E} [g^k \mid x^k], x^k - x^* \rangle \\ &\quad + \gamma^2 \mathbb{E} \left[ \|g^k - \nabla f(x^*)\|_2^2 \mid x^k \right].\end{aligned}$$

$$\mathbb{E} [g^k \mid x^k] = \nabla f(x^k)$$

$$\mathbb{E} \left[ \|g^k - \nabla f(x^*)\|_2^2 \mid x^k \right] \leq 4L \cdot (f(x^k) - f(x^*)) + 2 \cdot \frac{1}{n} \sum_{i=1}^n \|\nabla f_i(x^*) - y_i^k\|_2^2$$

## SAGA: доказательство

- Промежуточный итог:

$$\begin{aligned}\mathbb{E} \left[ \|x^{k+1} - x^*\|_2^2 \mid x^k \right] &= \|x^k - x^*\|_2^2 - 2\gamma \langle \mathbb{E} [g^k \mid x^k], x^k - x^* \rangle \\ &\quad + \gamma^2 \mathbb{E} \left[ \|g^k - \nabla f(x^*)\|_2^2 \mid x^k \right].\end{aligned}$$

$$\mathbb{E} [g^k \mid x^k] = \nabla f(x^k)$$

$$\mathbb{E} \left[ \|g^k - \nabla f(x^*)\|_2^2 \mid x^k \right] \leq 4L \cdot (f(x^k) - f(x^*)) + 2 \cdot \frac{1}{n} \sum_{i=1}^n \|\nabla f_i(x^*) - y_i^k\|_2^2$$

- Собираем вместе:

$$\begin{aligned}\mathbb{E} \left[ \|x^{k+1} - x^*\|_2^2 \mid x^k \right] &\leq \|x^k - x^*\|_2^2 - 2\gamma \langle \nabla f(x^k), x^k - x^* \rangle \\ &\quad + \gamma^2 \left( 4L \cdot (f(x^k) - f(x^*)) + 2 \cdot \frac{1}{n} \sum_{i=1}^n \|\nabla f_i(x^*) - y_i^k\|_2^2 \right)\end{aligned}$$

# SAGA: доказательство

- Сильная выпуклость функции  $f$ :

$$\mathbb{E} \left[ \|x^{k+1} - x^*\|_2^2 \mid x^k \right] \leq (1 - \mu\gamma) \|x^k - x^*\|_2^2 - 2\gamma(1 - 2\gamma L)(f(x^k) - f(x^*)) \\ + 2\gamma^2 \cdot \frac{1}{n} \sum_{i=1}^n \|\nabla f_i(x^*) - y_i^k\|_2^2.$$

# SAGA: доказательство

- Сильная выпуклость функции  $f$ :

$$\mathbb{E} \left[ \|x^{k+1} - x^*\|_2^2 \mid x^k \right] \leq (1 - \mu\gamma) \|x^k - x^*\|_2^2 - 2\gamma(1 - 2\gamma L)(f(x^k) - f(x^*)) \\ + 2\gamma^2 \cdot \frac{1}{n} \sum_{i=1}^n \|\nabla f_i(x^*) - y_i^k\|_2^2.$$

- Более формально пришли к тому, что если  $y_i^k \rightarrow \nabla f_i(x^*)$ , то дисперсия «убивается», а значит будет линейная сходимость. Покажем, как это можно строго оформить.

## SAGA: доказательство

- Рассмотрим, как ведет себя  $\frac{1}{n} \sum_{i=1}^n \|\nabla f_i(x^*) - y_i^k\|_2^2$ :

$$\begin{aligned}\mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|y_i^{k+1} - \nabla f_i(x^*)\|_2^2 \mid x^k \right] &= \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[ \|y_i^{k+1} - \nabla f_i(x^*)\|_2^2 \mid x^k \right] \\ &= \left(1 - \frac{1}{n}\right) \cdot \frac{1}{n} \sum_{i=1}^n \|y_i^k - \nabla f_i(x^*)\|_2^2 \\ &\quad + \frac{1}{n} \cdot \frac{1}{n} \sum_{i=1}^n \|f_i(x^k) - \nabla f_i(x^*)\|_2^2 \\ &\leq \left(1 - \frac{1}{n}\right) \cdot \frac{1}{n} \sum_{i=1}^n \|y_i^k - \nabla f_i(x^*)\|_2^2 \\ &\quad + \frac{1}{n} \cdot 2L(f(x^k) - f(x^*)).\end{aligned}$$

## SAGA: доказательство

- Итого (здесь сразу накинута полное математическое ожидание):

$$\begin{aligned}\mathbb{E} \left[ \|x^{k+1} - x^*\|_2^2 \right] &\leq (1 - \mu\gamma) \mathbb{E} \left[ \|x^k - x^*\|_2^2 \right] - 2\gamma(1 - 2\gamma L) \mathbb{E} \left[ f(x^k) - f(x^*) \right] \\ &\quad + 2\gamma^2 \cdot \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|\nabla f_i(x^*) - y_i^k\|_2^2 \right]\end{aligned}$$

$$\begin{aligned}\mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|y_i^{k+1} - \nabla f_i(x^*)\|_2^2 \right] &\leq \left( 1 - \frac{1}{n} \right) \cdot \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|y_i^k - \nabla f_i(x^*)\|_2^2 \right] \\ &\quad + \frac{1}{n} \cdot 2L \mathbb{E} \left[ f(x^k) - f(x^*) \right].\end{aligned}$$

## SAGA: доказательство

- Итого (здесь сразу накинута полное математическое ожидание):

$$\begin{aligned}\mathbb{E} \left[ \|x^{k+1} - x^*\|_2^2 \right] &\leq (1 - \mu\gamma) \mathbb{E} \left[ \|x^k - x^*\|_2^2 \right] - 2\gamma(1 - 2\gamma L) \mathbb{E} \left[ f(x^k) - f(x^*) \right] \\ &\quad + 2\gamma^2 \cdot \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|\nabla f_i(x^*) - y_i^k\|_2^2 \right]\end{aligned}$$

$$\begin{aligned}\mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|y_i^{k+1} - \nabla f_i(x^*)\|_2^2 \right] &\leq \left( 1 - \frac{1}{n} \right) \cdot \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \|y_i^k - \nabla f_i(x^*)\|_2^2 \right] \\ &\quad + \frac{1}{n} \cdot 2L \mathbb{E} \left[ f(x^k) - f(x^*) \right].\end{aligned}$$

- Получилось две «сходящиеся» последовательности, осталось их аккуратно «сшить».

# SAGA: доказательство

- Пусть  $M > 0$ :

$$\begin{aligned} & \mathbb{E} \left[ \|x^{k+1} - x^*\|_2^2 + M\gamma^2 \cdot \frac{1}{n} \sum_{i=1}^n \|y_i^{k+1} - \nabla f_i(x^*)\|_2^2 \right] \\ & \leq (1 - \mu\gamma) \mathbb{E} \left[ \|x^k - x^*\|_2^2 \right] \\ & \quad + \left( 1 + \frac{2}{M} - \frac{1}{n} \right) \mathbb{E} \left[ M\gamma^2 \cdot \frac{1}{n} \sum_{i=1}^n \|y_i^k - \nabla f_i(x^*)\|_2^2 \right] \\ & \quad - 2\gamma \left( 1 - 2\gamma L - \frac{\gamma ML}{n} \right) \mathbb{E} \left[ f(x^k) - f(x^*) \right] \end{aligned}$$



## SAGA: доказательство

- Возьмем  $M = 4n$ :

$$\begin{aligned} & \mathbb{E} \left[ \|x^{k+1} - x^*\|_2^2 + 4n\gamma^2 \cdot \frac{1}{n} \sum_{i=1}^n \|y_i^{k+1} - \nabla f_i(x^*)\|_2^2 \right] \\ & \leq (1 - \mu\gamma) \mathbb{E} \left[ \|x^k - x^*\|_2^2 \right] \\ & \quad + \left( 1 - \frac{1}{2n} \right) \mathbb{E} \left[ 4n\gamma^2 \cdot \frac{1}{n} \sum_{i=1}^n \|y_i^k - \nabla f_i(x^*)\|_2^2 \right] \\ & \quad - 2\gamma (1 - 6\gamma L) \mathbb{E} \left[ f(x^k) - f(x^*) \right] \end{aligned}$$

## SAGA: доказательство

- Возьмем  $M = 4n$ :

$$\begin{aligned} & \mathbb{E} \left[ \|x^{k+1} - x^*\|_2^2 + 4n\gamma^2 \cdot \frac{1}{n} \sum_{i=1}^n \|y_i^{k+1} - \nabla f_i(x^*)\|_2^2 \right] \\ & \leq (1 - \mu\gamma) \mathbb{E} \left[ \|x^k - x^*\|_2^2 \right] \\ & \quad + \left( 1 - \frac{1}{2n} \right) \mathbb{E} \left[ 4n\gamma^2 \cdot \frac{1}{n} \sum_{i=1}^n \|y_i^k - \nabla f_i(x^*)\|_2^2 \right] \\ & \quad - 2\gamma (1 - 6\gamma L) \mathbb{E} \left[ f(x^k) - f(x^*) \right] \end{aligned}$$

- Теперь  $\gamma \leq \frac{1}{6L}$ :

$$\begin{aligned} & \mathbb{E} \left[ \|x^{k+1} - x^*\|_2^2 + 4n\gamma^2 \cdot \frac{1}{n} \sum_{i=1}^n \|y_i^{k+1} - \nabla f_i(x^*)\|_2^2 \right] \\ & \leq \max \left\{ (1 - \mu\gamma); \left( 1 - \frac{1}{2n} \right) \right\} \mathbb{E} \left[ \|x^k - x^*\|_2^2 + 4n\gamma^2 \cdot \frac{1}{n} \sum_{i=1}^n \|y_i^k - \nabla f_i(x^*)\|_2^2 \right] \end{aligned}$$

# SAGA: сходимость

- Получили сходимость, но по необычному критерию. Суть критерия в отражении физики, как сходимости  $x^k \rightarrow x^*$ , так и  $y_i^k \rightarrow \nabla f_i(x^*)$ , что и закладывали в метод.

## Теорема сходимость SAGA

Пусть задача безусловной стохастической оптимизации вида конечной суммы с  $L$ -гладкими, выпуклыми функциями  $f_i$  и  $\mu$ -сильно выпуклой целевой функцией  $f$  решается с помощью SAGA с  $\gamma \leq \frac{1}{6L}$ . Тогда справедлива следующая оценка сходимости

$$\mathbb{E}[V_k] \leq \max \left\{ (1 - \mu\gamma); \left(1 - \frac{1}{2n}\right) \right\}^k \mathbb{E}[V_0],$$

где  $V_k = \|x^k - x^*\|_2^2 + 4n\gamma^2 \cdot \frac{1}{n} \sum_{i=1}^n \|y_i^k - \nabla f_i(x^*)\|_2^2$ .

# SAGA: сходимость

- Получили сходимость, но по необычному критерию. Суть критерия в отражении физики, как сходимости  $x^k \rightarrow x^*$ , так и  $y_i^k \rightarrow \nabla f_i(x^*)$ , что и закладывали в метод.

## Теорема сходимость SAGA

Пусть задача безусловной стохастической оптимизации вида конечной суммы с  $L$ -гладкими, выпуклыми функциями  $f_i$  и  $\mu$ -сильно выпуклой целевой функцией  $f$  решается с помощью SAGA с  $\gamma \leq \frac{1}{6L}$ . Тогда справедлива следующая оценка сходимости

$$\mathbb{E}[V_k] \leq \max \left\{ (1 - \mu\gamma); \left(1 - \frac{1}{2n}\right) \right\}^k \mathbb{E}[V_0],$$

где  $V_k = \|x^k - x^*\|_2^2 + 4n\gamma^2 \cdot \frac{1}{n} \sum_{i=1}^n \|y_i^k - \nabla f_i(x^*)\|_2^2$ .

- Легко заметить, что из сходимости по  $\mathbb{E}[V_k]$  следует и сходимость по  $\mathbb{E}[\|x^k - x^*\|_2^2]$ :  $\mathbb{E}[\|x^k - x^*\|_2^2] \leq \mathbb{E}[V_k]$

## SAGA: сходимость

- **Вопрос:** почему не взять  $M$  огромным, тогда сходимость будет лучше?

# SAGA: сходимость

- **Вопрос:** почему не взять  $M$  огромным, тогда сходимость будет лучше?  $M$  еще влияет на критерий сходимости, который так же будет расти с ростом  $M$ . При этом сходимости лучше, чем  $(1 - \frac{1}{n})$  не добиться.

# SAGA: сходимость

- **Вопрос:** почему не взять  $M$  огромным, тогда сходимость будет лучше?  $M$  еще влияет на критерий сходимости, который так же будет расти с ростом  $M$ . При этом сходимости лучше, чем  $(1 - \frac{1}{n})$  не добиться.
- Получаем следующую (уже анонсированную в прошлый раз) оценку на число итераций для достижения точности  $\varepsilon$ :

$$\mathcal{O} \left( \left[ n + \frac{L}{\mu} \right] \log \frac{1}{\varepsilon} \right)$$

# SAGA: сходимость

- **Вопрос:** почему не взять  $M$  огромным, тогда сходимость будет лучше?  $M$  еще влияет на критерий сходимости, который так же будет расти с ростом  $M$ . При этом сходимости лучше, чем  $(1 - \frac{1}{n})$  не добиться.
- Получаем следующую (уже анонсированную в прошлый раз) оценку на число итераций для достижения точности  $\varepsilon$ :

$$\mathcal{O}\left(\left[n + \frac{L}{\mu}\right] \log \frac{1}{\varepsilon}\right)$$

- У классического градиентного спуска оценка:

$$\mathcal{O}\left(\frac{L}{\mu} \log \frac{1}{\varepsilon}\right),$$

но оракульная сложность (подсчет градиентов  $\nabla f_i$ ) у градиентного спуска в  $n$  раз больше.



# Поворот на 90 градусов

- Простой пример:

$$\min_{x \in \mathbb{R}^d} \left[ f(x) = \frac{1}{n} \sum_{i=1}^n \|a_i^T x - b_i\|_2^2 \right],$$

где  $\{a_i, b_i\}_{i=1}^n$  – обучающая выборка.

# Поворот на 90 градусов

- Простой пример:

$$\min_{x \in \mathbb{R}^d} \left[ f(x) = \frac{1}{n} \sum_{i=1}^n \|a_i^T x - b_i\|_2^2 \right],$$

где  $\{a_i, b_i\}_{i=1}^n$  – обучающая выборка.

- До этого, мы брали не всю выборку для подсчета градиента, чтобы быть более эффективными. Т.е. использовали только часть строк матрицы  $A$ , составленной из  $a_i$  **Вопрос:** а как по-другому можно добиться эффективности?

# Поворот на 90 градусов

- Простой пример:

$$\min_{x \in \mathbb{R}^d} \left[ f(x) = \frac{1}{n} \sum_{i=1}^n \|a_i^T x - b_i\|_2^2 \right],$$

где  $\{a_i, b_i\}_{i=1}^n$  – обучающая выборка.

- До этого, мы брали не всю выборку для подсчета градиента, чтобы быть более эффективными. Т.е. использовали только часть строк матрицы  $A$ , составленной из  $a_i$  **Вопрос:** а как по-другому можно добиться эффективности? если до этого был выбор строк матрицы  $A$ , то теперь можно попробовать как-то завязаться на столбцы. **Вопрос:** а что означает «выбор столбцов»?

# Поворот на 90 градусов

- Простой пример:

$$\min_{x \in \mathbb{R}^d} \left[ f(x) = \frac{1}{n} \sum_{i=1}^n \|a_i^T x - b_i\|_2^2 \right],$$

где  $\{a_i, b_i\}_{i=1}^n$  – обучающая выборка.

- До этого, мы брали не всю выборку для подсчета градиента, чтобы быть более эффективными. Т.е. использовали только часть строк матрицы  $A$ , составленной из  $a_i$  **Вопрос:** а как по-другому можно добиться эффективности? если до этого был выбор строк матрицы  $A$ , то теперь можно попробовать как-то завязаться на столбцы. **Вопрос:** а что означает «выбор столбцов»? Выбор координат вектора  $x$ .

# Производная по направлению

- Часто для более сложных задач к подсчету производных по координатам/направлениям прибегают не из-за удешевления процесса, а из-за доступности только оракула нулевого порядка (значения функции). Потому что производную по направлению  $e \in \{e \in \mathbb{R}^d \mid \|e\|_2 \leq 1\}$  можно аппроксимировать через конечную разность:

$$[\nabla f(x)]_i \approx \frac{f(x + \tau e) - f(x - \tau e)}{2\tau}$$

(таким образом можно «собрать» и весь «градиент»).

# Координатный метод

---

## Алгоритм 2 Координатный метод

---

**Вход:** размер шага  $\gamma > 0$ , стартовая точка  $x^0 \in \mathbb{R}^d$ , значения памяти  $y_i^0 = 0$  для всех  $i \in [n]$ , количество итераций  $K$

- 1: **for**  $k = 0, 1, \dots, K - 1$  **do**
- 2:     Сгенерировать независимо  $i_k$  из  $[d]$
- 3:     Вычислить  $[\nabla f(x^k)]_{i_k}$
- 4:      $x^{k+1} = x^k - \gamma \cdot d[\nabla f(x^k)]_{i_k} e_{i_k}$
- 5: **end for**

**Выход:**  $x^K$

---

Здесь  $e_{i_k}$  –  $i$ -ый базисный вектор

# Координатный метод

---

## Алгоритм 3 Координатный метод

---

**Вход:** размер шага  $\gamma > 0$ , стартовая точка  $x^0 \in \mathbb{R}^d$ , значения памяти  $y_i^0 = 0$  для всех  $i \in [n]$ , количество итераций  $K$

- 1: **for**  $k = 0, 1, \dots, K - 1$  **do**
- 2:     Сгенерировать независимо  $i_k$  из  $[d]$
- 3:     Вычислить  $[\nabla f(x^k)]_{i_k}$
- 4:      $x^{k+1} = x^k - \gamma \cdot d[\nabla f(x^k)]_{i_k} e_{i_k}$
- 5: **end for**

**Выход:**  $x^K$

---

Здесь  $e_{i_k}$  –  $i$ -ый базисный вектор

- Зачем в шаге метода есть домножение на  $d$ ?

# Координатный метод

---

## Алгоритм 4 Координатный метод

---

**Вход:** размер шага  $\gamma > 0$ , стартовая точка  $x^0 \in \mathbb{R}^d$ , значения памяти  $y_i^0 = 0$  для всех  $i \in [n]$ , количество итераций  $K$

- 1: **for**  $k = 0, 1, \dots, K - 1$  **do**
- 2:   Сгенерировать независимо  $i_k$  из  $[d]$
- 3:   Вычислить  $[\nabla f(x^k)]_{i_k}$
- 4:    $x^{k+1} = x^k - \gamma \cdot d[\nabla f(x^k)]_{i_k} e_{i_k}$
- 5: **end for**

**Выход:**  $x^K$

---

Здесь  $e_{i_k}$  –  $i$ -ый базисный вектор

- Зачем в шаге метода есть домножение на  $d$ ? Для несмещенности того, что мы используем вместо градиента.



# Координатный метод: доказательство

- $f$  является  $L$ -гладкой и  $\mu$  - сильно выпуклой.

# Координатный метод: доказательство

- $f$  является  $L$ -гладкой и  $\mu$  - сильно выпуклой.
- Уже привычно:

$$\begin{aligned}\|x^{k+1} - x^*\|_2^2 &= \|x^k - x^*\|_2^2 - 2\gamma \langle d[\nabla f(x^k)]_{i_k} e_{i_k}, x^k - x^* \rangle \\ &\quad + \gamma^2 \|d[\nabla f(x^k)]_{i_k} e_{i_k}\|_2^2.\end{aligned}$$

# Координатный метод: доказательство

- $f$  является  $L$ -гладкой и  $\mu$  - сильно выпуклой.
- Уже привычно:

$$\begin{aligned}\|x^{k+1} - x^*\|_2^2 &= \|x^k - x^*\|_2^2 - 2\gamma \langle d[\nabla f(x^k)]_{i_k} e_{i_k}, x^k - x^* \rangle \\ &\quad + \gamma^2 \|d[\nabla f(x^k)]_{i_k} e_{i_k}\|_2^2.\end{aligned}$$

- Берем условное мат.ожидание по случайности только на итерации  $k$ :

$$\begin{aligned}\mathbb{E} \left[ \|x^{k+1} - x^*\|_2^2 \mid x^k \right] &= \|x^k - x^*\|_2^2 - 2\gamma \langle \mathbb{E} \left[ d[\nabla f(x^k)]_{i_k} e_{i_k} \mid x^k \right], x^k - x^* \rangle \\ &\quad + \gamma^2 \mathbb{E} \left[ \|d[\nabla f(x^k)]_{i_k} e_{i_k}\|_2^2 \mid x^k \right].\end{aligned}$$

# Координатный метод: доказательство

- Работаем с  $\mathbb{E} \left[ d[\nabla f(x^k)]_{i_k} e_{i_k} \mid x^k \right]$ :

$$\begin{aligned} \mathbb{E} \left[ d[\nabla f(x^k)]_{i_k} e_{i_k} \mid x^k \right] &= \frac{1}{d} \sum_{j=1}^d d[\nabla f(x^k)]_j e_j \\ &= \nabla f(x^k) \end{aligned}$$

# Координатный метод: доказательство

- Теперь работаем с  $\mathbb{E} [\|d[\nabla f(x^k)]\|_2^2 \mid x^k]$ :

$$\begin{aligned}\mathbb{E} [\|d[\nabla f(x^k)]\|_2^2 \mid x^k] &= \mathbb{E} [\|d[\nabla f(x^k)]\|_2^2 \mid x^k] \\ &= d^2 \mathbb{E} [\|[\nabla f(x^k)]\|_2^2 \mid x^k] \\ &= d^2 \cdot \frac{1}{d} \sum_{j=1}^d \|[\nabla f(x^k)]_j e_j\|_2^2 \\ &= d \|\nabla f(x^k)\|_2^2\end{aligned}$$

# Координатный метод: доказательство

- Промежуточный итог:

$$\begin{aligned}\mathbb{E} \left[ \|x^{k+1} - x^*\|_2^2 \mid x^k \right] &= \|x^k - x^*\|_2^2 - 2\gamma \langle \mathbb{E} \left[ d[\nabla f(x^k)]_{i_k} e_{i_k} \mid x^k \right], x^k - x^* \rangle \\ &\quad + \gamma^2 \mathbb{E} \left[ \|d[\nabla f(x^k)]_{i_k} e_{i_k}\|_2^2 \mid x^k \right].\end{aligned}$$

$$\mathbb{E} \left[ d[\nabla f(x^k)]_{i_k} e_{i_k} \mid x^k \right] = \nabla f(x^k)$$

$$\mathbb{E} \left[ \|d[\nabla f(x^k)]_{i_k} e_{i_k}\|_2^2 \mid x^k \right] = d \|\nabla f(x^k)\|_2^2$$

# Координатный метод: доказательство

- Промежуточный итог:

$$\mathbb{E} \left[ \|x^{k+1} - x^*\|_2^2 \mid x^k \right] = \|x^k - x^*\|_2^2 - 2\gamma \langle \mathbb{E} \left[ d[\nabla f(x^k)]_{i_k} e_{i_k} \mid x^k \right], x^k - x^* \rangle + \gamma^2 \mathbb{E} \left[ \|d[\nabla f(x^k)]_{i_k} e_{i_k}\|_2^2 \mid x^k \right].$$

$$\mathbb{E} \left[ d[\nabla f(x^k)]_{i_k} e_{i_k} \mid x^k \right] = \nabla f(x^k)$$

$$\mathbb{E} \left[ \|d[\nabla f(x^k)]_{i_k} e_{i_k}\|_2^2 \mid x^k \right] = d \|\nabla f(x^k)\|_2^2$$

- Собираем вместе:

$$\mathbb{E} \left[ \|x^{k+1} - x^*\|_2^2 \mid x^k \right] \leq \|x^k - x^*\|_2^2 - 2\gamma \langle \nabla f(x^k), x^k - x^* \rangle + d\gamma^2 \|\nabla f(x^k)\|_2^2.$$

# Координатный метод: доказательство

- Сильная выпуклость и гладкость функции  $f$ :

$$\mathbb{E} \left[ \|x^{k+1} - x^*\|_2^2 \mid x^k \right] \leq (1 - \mu\gamma) \|x^k - x^*\|_2^2 - 2\gamma(1 - d\gamma L)(f(x^k) - f(x^*)).$$



# Координатный метод: доказательство

- Сильная выпуклость и гладкость функции  $f$ :

$$\mathbb{E} \left[ \|x^{k+1} - x^*\|_2^2 \mid x^k \right] \leq (1 - \mu\gamma) \|x^k - x^*\|_2^2 - 2\gamma(1 - d\gamma L)(f(x^k) - f(x^*)).$$

- Пусть  $\gamma \leq \frac{1}{dL}$ :

$$\mathbb{E} \left[ \|x^{k+1} - x^*\|_2^2 \mid x^k \right] \leq (1 - \mu\gamma) \|x^k - x^*\|_2^2.$$

# Координатный метод: сходимость

## Теорема сходимость SAGA

Пусть задача безусловной оптимизации с  $L$ -гладкой и  $\mu$ -сильно выпуклой целевой функцией  $f$  решается с помощью координатного метода с  $\gamma \leq \frac{1}{dL}$ . Тогда справедлива следующая оценка сходимости

$$\mathbb{E} \left[ \|x^k - x^*\|_2^2 \right] \leq (1 - \mu\gamma)^k \mathbb{E} \left[ \|x^0 - x^*\|_2^2 \right]$$

# Координатный метод: сходимость

- Подставив  $\gamma = \frac{1}{dL}$ , получаем следующую итерационную сложность

$$\mathcal{O} \left( \frac{dL}{\mu} \log \frac{1}{\varepsilon} \right).$$

# Координатный метод: сходимость

- Подставив  $\gamma = \frac{1}{dL}$ , получаем следующую итерационную сложность

$$\mathcal{O}\left(\frac{dL}{\mu} \log \frac{1}{\varepsilon}\right).$$

**Вопрос:** есть ли улучшения по сравнению с обычным градиентным спуском?

# Координатный метод: сходимость

- Подставив  $\gamma = \frac{1}{dL}$ , получаем следующую итерационную сложность

$$\mathcal{O}\left(\frac{dL}{\mu} \log \frac{1}{\varepsilon}\right).$$

**Вопрос:** есть ли улучшения по сравнению с обычным градиентным спуском? В общем случае нет. Это доказуемо так.

## Координатный метод: сходимость

- Подставив  $\gamma = \frac{1}{dL}$ , получаем следующую итерационную сложность

$$\mathcal{O} \left( \frac{dL}{\mu} \log \frac{1}{\varepsilon} \right).$$

**Вопрос:** есть ли улучшения по сравнению с обычным градиентным спуском? В общем случае нет. Это доказуемо так.

- Если есть дополнительная информация о задаче (например, свойства констант Липшица градиента по направлению), то улучшения можно получить.

# Координатный метод: сходимость

- Подставив  $\gamma = \frac{1}{dL}$ , получаем следующую итерационную сложность

$$\mathcal{O}\left(\frac{dL}{\mu} \log \frac{1}{\varepsilon}\right).$$

**Вопрос:** есть ли улучшения по сравнению с обычным градиентным спуском? В общем случае нет. Это доказуемо так.

- Если есть дополнительная информация о задаче (например, свойства констант Липшица градиента по направлению), то улучшения можно получить.
- Еще координатный метод часто хорошо себя проявляет на практике.

## Координатный метод: сходимость

- Подставив  $\gamma = \frac{1}{dL}$ , получаем следующую итерационную сложность

$$\mathcal{O}\left(\frac{dL}{\mu} \log \frac{1}{\varepsilon}\right).$$

**Вопрос:** есть ли улучшения по сравнению с обычным градиентным спуском? В общем случае нет. Это доказуемо так.

- Если есть дополнительная информация о задаче (например, свойства констант Липшица градиента по направлению), то улучшения можно получить.
- Еще координатный метод часто хорошо себя проявляет на практике.
- Результат обобщается и на случай выбора нескольких координат.
- Возможно ускорение с помощью двух моментумов.