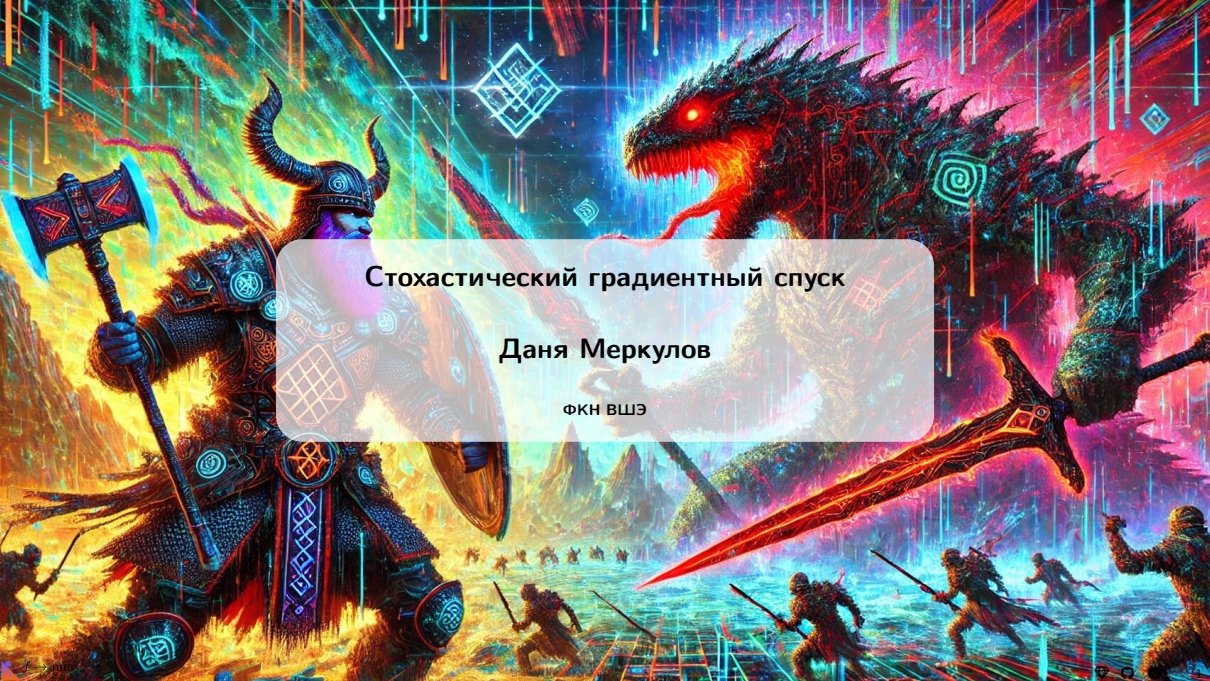


Стохастический градиентный спуск

Даня Меркулов

ФКН ВШЭ

Задача минимизации конечной суммы

От градиентного спуска к стохастическому

finite-sum
finite sample
average

Рассмотрим классическую задачу минимизации среднего по конечной выборке:

$$\min_{x \in \mathbb{R}^p} f(x) = \min_{x \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} f_i(x)$$

Градиентный спуск записывается следующим образом:

$$x_{k+1} = x_k - \frac{\alpha_k}{n} \sum_{i=1}^{n} \nabla f_i(x_k)$$

$$\nabla f(x) = \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(x)$$ (GD)

- Сходимость гарантируется при постоянном шаге или линейном поиске.

От градиентного спуска к стохастическому

Рассмотрим классическую задачу минимизации среднего по конечной выборке:

$$\min_{x \in \mathbb{R}^p} f(x) = \min_{x \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} f_i(x)$$

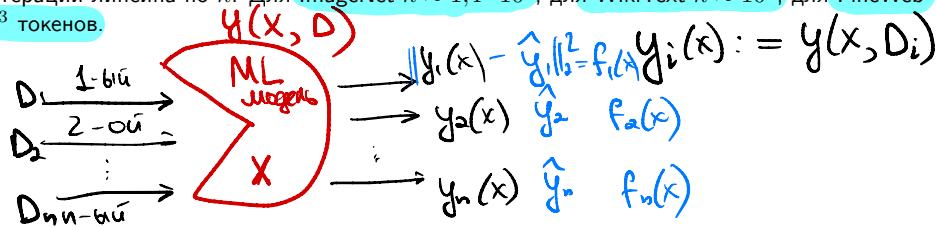
часто кол-во объектов в обуч. выборке

обучаемые параметры модели

Градиентный спуск записывается следующим образом:

$$x_{k+1} = x_k - \frac{\alpha_k}{n} \sum_{i=1}^{n} \nabla f_i(x_k) \qquad \text{(GD)}$$

- Сходимость гарантируется при постоянном шаге или линейном поиске.
- Стоимость итерации линейна по n. Для ImageNet $n \approx 1{,}4 \cdot 10^7$, для WikiText $n \approx 10^8$, для FineWeb $n \approx 1{,}5 \cdot 10^{13}$ токенов.



От градиентного спуска к стохастическому

Рассмотрим классическую задачу минимизации среднего по конечной выборке:

$$\min_{x \in \mathbb{R}^p} f(x) = \min_{x \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} f_i(x)$$

Градиентный спуск записывается следующим образом:

$$x_{k+1} = x_k - \frac{\alpha_k}{n} \sum_{i=1}^{n} \nabla f_i(x_k) \tag{GD}$$

- Сходимость гарантируется при постоянном шаге или линейном поиске.
- Стоимость итерации линейна по n. Для ImageNet $n \approx 1{,}4 \cdot 10^7$, для WikiText $n \approx 10^8$, для FineWeb $n \approx 1{,}5 \cdot 10^{13}$ токенов.

От градиентного спуска к стохастическому

Рассмотрим классическую задачу минимизации среднего по конечной выборке:

$$\min_{x \in \mathbb{R}^p} f(x) = \min_{x \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} f_i(x)$$

Градиентный спуск записывается следующим образом:

$$x_{k+1} = x_k - \frac{\alpha_k}{n} \sum_{i=1}^{n} \nabla f_i(x_k)$$

$i_k \in (1 \dots n)$

(GD)

- Сходимость гарантируется при постоянном шаге или линейном поиске.
- Стоимость итерации линейна по n. Для ImageNet $n \approx 1{,}4 \cdot 10^7$, для WikiText $n \approx 10^8$, для FineWeb $n \approx 1{,}5 \cdot 10^{13}$ токенов.

Перейдём от полного градиента к его несмещённой оценке, случайно выбирая индекс i_k на каждой итерации равновероятно:

$$x_{k+1} = x_k - \alpha_k \nabla f_{i_k}(x_k)$$

стох. градиент

(SGD)

При $p(i_k = i) = \frac{1}{n}$ стохастический градиент является несмещённой оценкой:

$$\mathbb{E}[\nabla f_{i_k}(x)] = \sum_{i=1}^{n} p(i_k = i) \nabla f_i(x) = \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(x) = \nabla f(x)$$

Стоимость итерации и общая стоимость метода

раньше: итерация $\frac{1}{n}\sum \nabla f_i(x)$

теперь: $\nabla f_4(x)$

Стохастические итерации в n раз дешевле, но сколько итераций нужно для достижения точности ε?

в n раз дешевле!

Стоимость итерации и общая стоимость метода

Стохастические итерации в n раз дешевле, но сколько итераций нужно для достижения точности ε?

Если ∇f липшицев, то имеем:

Предположение	GD	SGD
PL	$\mathcal{O}(\log(1/\varepsilon))$	$\mathcal{O}(1/\varepsilon)$
Выпуклый	$\mathcal{O}(1/\varepsilon)$	$\mathcal{O}(1/\varepsilon^2)$
Невыпуклый	$\mathcal{O}(1/\varepsilon)$	$\mathcal{O}(1/\varepsilon^2)$

Стоимость итерации и общая стоимость метода

Стохастические итерации в n раз дешевле, но сколько итераций нужно для достижения точности ε?

Если ∇f липшицев, то имеем:

Предположение	GD	SGD
PL	$\mathcal{O}\left(\log(1/\varepsilon)\right)$	$\mathcal{O}\left(1/\varepsilon\right)$
Выпуклый	$\mathcal{O}\left(1/\varepsilon\right)$	$\mathcal{O}\left(1/\varepsilon^2\right)$
Невыпуклый	$\mathcal{O}\left(1/\varepsilon\right)$	$\mathcal{O}\left(1/\varepsilon^2\right)$

- Стохастический метод имеет дешёвые итерации, но медленную сходимость.

Стоимость итерации и общая стоимость метода

Стохастические итерации в n раз дешевле, но сколько итераций нужно для достижения точности ε?

Если ∇f липшицев, то имеем:

Предположение	GD	SGD
PL	$\mathcal{O}\left(\log(1/\varepsilon)\right)$	$\mathcal{O}\left(1/\varepsilon\right)$
Выпуклый	$\mathcal{O}\left(1/\varepsilon\right)$	$\mathcal{O}\left(1/\varepsilon^2\right)$
Невыпуклый	$\mathcal{O}\left(1/\varepsilon\right)$	$\mathcal{O}\left(1/\varepsilon^2\right)$

- Стохастический метод имеет дешёвые итерации, но медленную сходимость.
  - Сублинейная сходимость даже в сильно выпуклом случае.

Стоимость итерации и общая стоимость метода

Стохастические итерации в n раз дешевле, но сколько итераций нужно для достижения точности ε?

Если ∇f липшицев, то имеем:

Предположение	GD	SGD
PL	$\mathcal{O}\left(\log(1/\varepsilon)\right)$	$\mathcal{O}\left(1/\varepsilon\right)$
Выпуклый	$\mathcal{O}\left(1/\varepsilon\right)$	$\mathcal{O}\left(1/\varepsilon^2\right)$
Невыпуклый	$\mathcal{O}\left(1/\varepsilon\right)$	$\mathcal{O}\left(1/\varepsilon^2\right)$

SGD – оптимальный (нижние оценки совпадают с верхними)

- Стохастический метод имеет дешёвые итерации, но медленную сходимость.
  - Сублинейная сходимость даже в сильно выпуклом случае.
  - Эти оценки не могут быть улучшены при стандартных предположениях: оракул возвращает несмещённую оценку градиента с ограниченной дисперсией.

Стоимость итерации и общая стоимость метода

Стохастические итерации в n раз дешевле, но сколько итераций нужно для достижения точности ε?

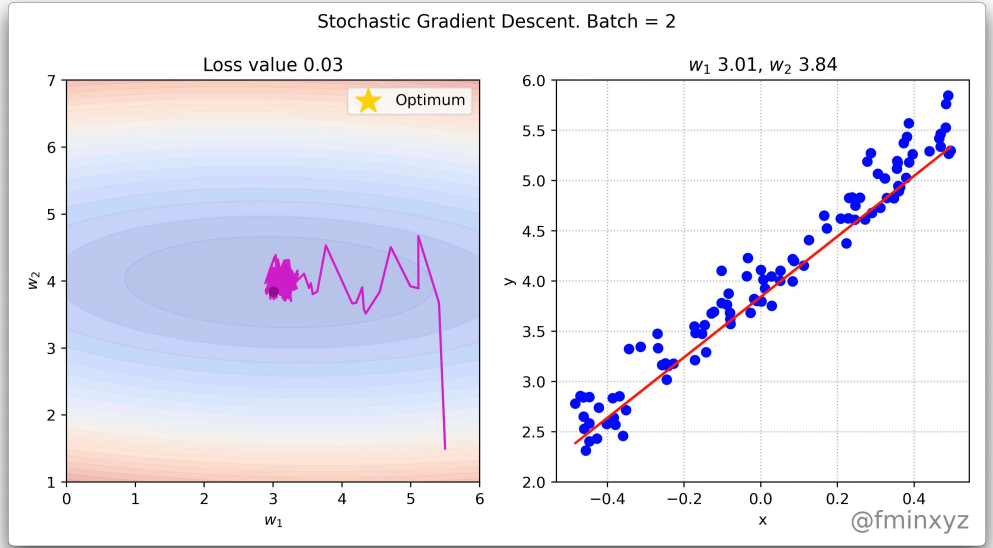
Если ∇f липшицев, то имеем:

Предположение	GD	SGD
PL	$\mathcal{O}\left(\log(1/\varepsilon)\right)$	$\mathcal{O}\left(1/\varepsilon\right)$
Выпуклый	$\mathcal{O}\left(1/\varepsilon\right)$	$\mathcal{O}\left(1/\varepsilon^2\right)$
Невыпуклый	$\mathcal{O}\left(1/\varepsilon\right)$	$\mathcal{O}\left(1/\varepsilon^2\right)$

- Стохастический метод имеет дешёвые итерации, но медленную сходимость.
  - Сублинейная сходимость даже в сильно выпуклом случае.
  - Эти оценки не могут быть улучшены при стандартных предположениях: оракул возвращает несмещённую оценку градиента с ограниченной дисперсией.
- Моментные и квазиньютоновские методы **не улучшают** скорость в стохастическом случае. Они могут улучшить лишь константы, поскольку узким местом становится дисперсия, а не число обусловленности.

Стохастический градиентный спуск (SGD)

Типичное поведение SGD



Базовое неравенство сходимости

Из липшицевости градиента следует:

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2$$

липшицевы параболы

$x_{k+1} = x_k - \alpha_k \cdot g_k$

$g_k = \nabla f_{i_k}(x_k)$

стох градиент

Базовое неравенство сходимости

Из липшицевости градиента следует:

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2$$

Подставляя (SGD):

$$f(x_{k+1}) \leq f(x_k) - \alpha_k \langle \nabla f(x_k), \nabla f_{i_k}(x_k) \rangle + \alpha_k^2 \frac{L}{2}\|\nabla f_{i_k}(x_k)\|^2$$

Базовое неравенство сходимости

Из липшицевости градиента следует:

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2$$

Подставляя (SGD):

$$f(x_{k+1}) \leq f(x_k) - \alpha_k \langle \nabla f(x_k), \nabla f_{i_k}(x_k) \rangle + \alpha_k^2 \frac{L}{2}\|\nabla f_{i_k}(x_k)\|^2$$

$\mathbb{E}_{i_k}(\,|x_k)$

Возьмём математическое ожидание по i_k:

$$\mathbb{E}[f(x_{k+1})] \leq \mathbb{E}\left[f(x_k) - \alpha_k \langle \nabla f(x_k), \nabla f_{i_k}(x_k) \rangle + \alpha_k^2 \frac{L}{2}\|\nabla f_{i_k}(x_k)\|^2\right]$$

Базовое неравенство сходимости

Из липшицевости градиента следует:

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2$$

Подставляя (SGD):

$$f(x_{k+1}) \leq f(x_k) - \alpha_k \langle \nabla f(x_k), \nabla f_{i_k}(x_k) \rangle + \alpha_k^2 \frac{L}{2}\|\nabla f_{i_k}(x_k)\|^2$$

Возьмём математическое ожидание по i_k:

$$\mathbb{E}[f(x_{k+1})] \leq \mathbb{E}\left[f(x_k) - \alpha_k \langle \nabla f(x_k), \nabla f_{i_k}(x_k) \rangle + \alpha_k^2 \frac{L}{2}\|\nabla f_{i_k}(x_k)\|^2\right]$$

По линейности математического ожидания:

$$\mathbb{E}[f(x_{k+1})] \leq f(x_k) - \alpha_k \langle \nabla f(x_k), \mathbb{E}[\nabla f_{i_k}(x_k)] \rangle + \alpha_k^2 \frac{L}{2}\mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2]$$

$\nabla f(x_k)$

Базовое неравенство сходимости

Из липшицевости градиента следует:

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2$$

Подставляя (SGD):

$$f(x_{k+1}) \leq f(x_k) - \alpha_k \langle \nabla f(x_k), \nabla f_{i_k}(x_k) \rangle + \alpha_k^2 \frac{L}{2}\|\nabla f_{i_k}(x_k)\|^2$$

Возьмём математическое ожидание по i_k:

$$\mathbb{E}[f(x_{k+1})] \leq \mathbb{E}\left[f(x_k) - \alpha_k \langle \nabla f(x_k), \nabla f_{i_k}(x_k) \rangle + \alpha_k^2 \frac{L}{2}\|\nabla f_{i_k}(x_k)\|^2\right]$$

По линейности математического ожидания:

$$\mathbb{E}[f(x_{k+1})] \leq f(x_k) - \alpha_k \langle \nabla f(x_k), \mathbb{E}[\nabla f_{i_k}(x_k)] \rangle + \alpha_k^2 \frac{L}{2}\mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2]$$

Поскольку равномерная выборка даёт несмещённую оценку $\mathbb{E}[\nabla f_{i_k}(x_k)] = \nabla f(x_k)$:

$$\mathbb{E}[f(x_{k+1})] \leq f(x_k) - \alpha_k \|\nabla f(x_k)\|^2 + \alpha_k^2 \frac{L}{2}\mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2] \tag{1}$$

Сходимость SGD

Гладкий PL-случай. Постоянный шаг

Theorem

Пусть f — L-гладкая функция, удовлетворяющая условию Поляка — Лоясиевича (PL) с константой $\mu > 0$, а дисперсия стохастического градиента ограничена: $\mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2] \leq \sigma^2$. Тогда SGD с постоянным шагом $\alpha < \frac{1}{2\mu}$ гарантирует

$$\mathbb{E}[f(x_k) - f^*] \leq (1 - 2\alpha\mu)^k [f(x_0) - f^*] + \frac{L\sigma^2\alpha}{4\mu}.$$

Доказательство

Начнём с неравенства (1):

$$\mathbb{E}[f(x_{k+1})] \leq f(x_k) - \alpha_k \|\nabla f(x_k)\|^2 + \alpha_k^2 \frac{L}{2} \mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2]$$

Гладкий PL-случай. Постоянный шаг

Theorem

Пусть f — L-гладкая функция, удовлетворяющая условию Поляка — Лоясиевича (PL) с константой $\mu > 0$, а дисперсия стохастического градиента ограничена: $\mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2] \leq \sigma^2$. Тогда SGD с постоянным шагом $\alpha < \frac{1}{2\mu}$ гарантирует

$$\mathbb{E}[f(x_k) - f^*] \leq (1 - 2\alpha\mu)^k [f(x_0) - f^*] + \frac{L\sigma^2\alpha}{4\mu}.$$

Доказательство

Начнём с неравенства (1):

$$\mathbb{E}[f(x_{k+1})] \leq f(x_k) - \alpha_k \|\nabla f(x_k)\|^2 + \alpha_k^2 \frac{L}{2} \mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2]$$

PL: $\|\nabla f(x_k)\|^2 \geq 2\mu(f(x_k) - f^*)$

$$-\|\nabla f(x_k)\|^2 \leq -2\mu\left(f(x_k) - f^*\right)$$

Гладкий PL-случай. Постоянный шаг

Theorem

Пусть f — L-гладкая функция, удовлетворяющая условию Поляка — Лоясиевича (PL) с константой $\mu > 0$, а дисперсия стохастического градиента ограничена: $\mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2] \leq \sigma^2$. Тогда SGD с постоянным шагом $\alpha < \frac{1}{2\mu}$ гарантирует

$$\mathbb{E}[f(x_k) - f^*] \leq (1 - 2\alpha\mu)^k [f(x_0) - f^*] + \frac{L\sigma^2\alpha}{4\mu}.$$

Доказательство

Начнём с неравенства (1):

$$\mathbb{E}[f(x_{k+1})] \leq f(x_k) - \alpha_k \|\nabla f(x_k)\|^2 + \alpha_k^2 \frac{L}{2} \mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2]$$

$$\text{PL: } \|\nabla f(x_k)\|^2 \geq 2\mu(f(x_k) - f^*) \quad \leq f(x_k) - 2\alpha_k\mu(f(x_k) - f^*) + \alpha_k^2 \frac{L}{2} \mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2]$$

Гладкий PL-случай. Постоянный шаг

Theorem

Пусть f — L-гладкая функция, удовлетворяющая условию Поляка — Лоясиевича (PL) с константой $\mu > 0$, а дисперсия стохастического градиента ограничена: $\mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2] \leq \sigma^2$. Тогда SGD с постоянным шагом $\alpha < \frac{1}{2\mu}$ гарантирует

$$\mathbb{E}[f(x_k) - f^*] \leq (1 - 2\alpha\mu)^k [f(x_0) - f^*] + \frac{L\sigma^2\alpha}{4\mu}.$$

Доказательство

Начнём с неравенства (1):

$$\mathbb{E}[f(x_{k+1})] \leq f(x_k) - \alpha_k \|\nabla f(x_k)\|^2 + \alpha_k^2 \frac{L}{2} \mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2]$$

PL: $\|\nabla f(x_k)\|^2 \geq 2\mu(f(x_k) - f^*)$

$$\leq f(x_k) - 2\alpha_k\mu(f(x_k) - f^*) + \alpha_k^2 \frac{L}{2} \mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2]$$

Вычитаем f^*

Гладкий PL-случай. Постоянный шаг

Theorem

Пусть f — L-гладкая функция, удовлетворяющая условию Поляка — Лоясиевича (PL) с константой $\mu > 0$, а дисперсия стохастического градиента ограничена: $\mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2] \le \sigma^2$. Тогда SGD с постоянным шагом $\alpha < \frac{1}{2\mu}$ гарантирует

$$\mathbb{E}[f(x_k) - f^*] \le (1 - 2\alpha\mu)^k [f(x_0) - f^*] + \frac{L\sigma^2\alpha}{4\mu}.$$

Доказательство

Начнём с неравенства (1):

$$\mathbb{E}[f(x_{k+1})] \le f(x_k) - \alpha_k \|\nabla f(x_k)\|^2 + \alpha_k^2 \frac{L}{2} \mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2]$$

PL: $\|\nabla f(x_k)\|^2 \ge 2\mu(f(x_k) - f^*)$

$$\le f(x_k) - 2\alpha_k\mu(f(x_k) - f^*) + \alpha_k^2 \frac{L}{2} \mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2]$$

Вычитаем f^*

$$\mathbb{E}[f(x_{k+1})] - f^* \le (1 - 2\alpha_k\mu)[f(x_k) - f^*] + \alpha_k^2 \frac{L}{2} \mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2]$$

Гладкий PL-случай. Постоянный шаг

Theorem

Пусть f — L-гладкая функция, удовлетворяющая условию Поляка — Лоясиевича (PL) с константой $\mu > 0$, а дисперсия стохастического градиента ограничена: $\mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2] \le \sigma^2$. Тогда SGD с постоянным шагом $\alpha < \frac{1}{2\mu}$ гарантирует

$$\mathbb{E}[f(x_k) - f^*] \le (1 - 2\alpha\mu)^k [f(x_0) - f^*] + \frac{L\sigma^2\alpha}{4\mu}.$$

Доказательство

Начнём с неравенства (1):

$$\mathbb{E}[f(x_{k+1})] \le f(x_k) - \alpha_k \|\nabla f(x_k)\|^2 + \alpha_k^2 \frac{L}{2} \mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2]$$

PL: $\|\nabla f(x_k)\|^2 \ge 2\mu(f(x_k) - f^*)$

$$\le f(x_k) - 2\alpha_k\mu(f(x_k) - f^*) + \alpha_k^2 \frac{L}{2} \mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2]$$

Вычитаем f^*

$$\mathbb{E}[f(x_{k+1})] - f^* \le (1 - 2\alpha_k\mu)[f(x_k) - f^*] + \alpha_k^2 \frac{L}{2} \mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2]$$

Огр. дисперсия

Гладкий PL-случай. Постоянный шаг

Theorem

Пусть f — L-гладкая функция, удовлетворяющая условию Поляка — Лоясиевича (PL) с константой $\mu > 0$, а дисперсия стохастического градиента ограничена: $\mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2] \le \sigma^2$. Тогда SGD с постоянным шагом $\alpha < \frac{1}{2\mu}$ гарантирует

$$\mathbb{E}[f(x_k) - f^*] \le (1 - 2\alpha\mu)^k [f(x_0) - f^*] + \frac{L\sigma^2\alpha}{4\mu}.$$

Доказательство

Начнём с неравенства (1):

$$\begin{aligned}
\mathbb{E}[f(x_{k+1})] &\le f(x_k) - \alpha_k \|\nabla f(x_k)\|^2 + \alpha_k^2 \frac{L}{2} \mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2] \\
\text{PL: } \|\nabla f(x_k)\|^2 \ge 2\mu(f(x_k) - f^*) \quad &\le f(x_k) - 2\alpha_k\mu(f(x_k) - f^*) + \alpha_k^2 \frac{L}{2} \mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2] \\
\text{Вычитаем } f^* \quad \mathbb{E}[f(x_{k+1})] - f^* &\le (1 - 2\alpha_k\mu)[f(x_k) - f^*] + \alpha_k^2 \frac{L}{2} \mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2] \\
\text{Огр. дисперсия} \quad &\le (1 - 2\alpha_k\mu)[f(x_k) - f^*] + \frac{L\sigma^2\alpha_k^2}{2}.
\end{aligned}$$

Гладкий PL-случай. Постоянный шаг

Theorem

Пусть f — L-гладкая функция, удовлетворяющая условию Поляка — Лоясиевича (PL) с константой $\mu > 0$, а дисперсия стохастического градиента ограничена: $\mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2] \le \sigma^2$. Тогда SGD с постоянным шагом $\alpha < \frac{1}{2\mu}$ гарантирует

$$\mathbb{E}[f(x_k) - f^*] \le (1 - 2\alpha\mu)^k [f(x_0) - f^*] + \frac{L\sigma^2\alpha}{4\mu}.$$

Доказательство

Начнём с неравенства (1):

$$\begin{aligned}
\mathbb{E}[f(x_{k+1})] &\le f(x_k) - \alpha_k \|\nabla f(x_k)\|^2 + \alpha_k^2 \frac{L}{2} \mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2] \\
\text{PL: } \|\nabla f(x_k)\|^2 \ge 2\mu(f(x_k) - f^*) \quad &\le f(x_k) - 2\alpha_k\mu(f(x_k) - f^*) + \alpha_k^2 \frac{L}{2} \mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2] \\
\text{Вычитаем } f^* \quad \mathbb{E}[f(x_{k+1})] - f^* &\le (1 - 2\alpha_k\mu)[f(x_k) - f^*] + \alpha_k^2 \frac{L}{2} \mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2] \\
\text{Огр. дисперсия} \quad &\le (1 - 2\alpha_k\mu)[f(x_k) - f^*] + \frac{L\sigma^2\alpha_k^2}{2}.
\end{aligned}$$

Гладкий PL-случай. Постоянный шаг

Theorem

Пусть f — L-гладкая функция, удовлетворяющая условию Поляка — Лоясиевича (PL) с константой $\mu > 0$, а дисперсия стохастического градиента ограничена: $\mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2] \leq \sigma^2$. Тогда SGD с постоянным шагом $\alpha < \frac{1}{2\mu}$ гарантирует

$$\mathbb{E}[f(x_k) - f^*] \leq (1 - 2\alpha\mu)^k [f(x_0) - f^*] + \frac{L\sigma^2\alpha}{4\mu}.$$

Доказательство

Начнём с неравенства (1):

$$\mathbb{E}[f(x_{k+1})] \leq f(x_k) - \alpha_k \|\nabla f(x_k)\|^2 + \alpha_k^2 \frac{L}{2} \mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2]$$

PL: $\|\nabla f(x_k)\|^2 \geq 2\mu(f(x_k) - f^*)$

$$\leq f(x_k) - 2\alpha_k\mu(f(x_k) - f^*) + \alpha_k^2 \frac{L}{2} \mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2]$$

Вычитаем f^*

$$\mathbb{E}[f(x_{k+1})] - f^* \leq (1 - 2\alpha_k\mu)[f(x_k) - f^*] + \alpha_k^2 \frac{L}{2} \mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2]$$

Огр. дисперсия

$$\leq (1 - 2\alpha_k\mu)[f(x_k) - f^*] + \frac{L\sigma^2\alpha_k^2}{2}.$$

Гладкий PL-случай. Убывающий шаг

Theorem

Пусть f — L-гладкая, удовлетворяет PL с константой $\mu > 0$, а дисперсия стохастического градиента ограничена: $\mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2] \leq \sigma^2$. Тогда SGD с убывающим шагом $\alpha_k = \frac{2k+1}{2\mu(k+1)^2}$ гарантирует

$$\mathbb{E}[f(x_k) - f^*] \leq \frac{L\sigma^2}{2\mu^2 k}.$$

$\alpha \sim \frac{1}{k}$

Доказательство

1. Применяя стратегию убывающего шага $\alpha_k = \frac{2k+1}{2\mu(k+1)^2}$ к рекуррентному неравенству из предыдущей теоремы:

$$\mathbb{E}[f(x_k) - f^*] \leq \left(1 - 2\alpha_k\mu\right)^k \left(f(x_0) - f^*\right) + \frac{L\sigma^2\alpha}{4\mu}$$

$$1 - 2\alpha_k\mu = 1 - \frac{2(2k+1)}{2(k+1)^2} = \frac{2k^2 + 4k + 2 - 4k - 2}{2(k+1)^2} = \left(\frac{k}{k+1}\right)^2$$

$$\frac{\alpha^2}{2} = \frac{(2k+1)^2}{2\mu^2(k+1)^4}$$

Гладкий PL-случай. Убывающий шаг

Theorem

Пусть f — L-гладкая, удовлетворяет PL с константой $\mu > 0$, а дисперсия стохастического градиента ограничена: $\mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2] \leq \sigma^2$. Тогда SGD с убывающим шагом $\alpha_k = \frac{2k+1}{2\mu(k+1)^2}$ гарантирует

$$\mathbb{E}[f(x_k) - f^*] \leq \frac{L\sigma^2}{2\mu^2 k}.$$

Доказательство

1. Применяя стратегию убывающего шага $\alpha_k = \frac{2k+1}{2\mu(k+1)^2}$ к рекуррентному неравенству из предыдущей теоремы:

$$1-2\alpha_k\mu = \frac{k^2}{(k+1)^2}$$

Гладкий PL-случай. Убывающий шаг

Theorem

Пусть f — L-гладкая, удовлетворяет PL с константой $\mu > 0$, а дисперсия стохастического градиента ограничена: $\mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2] \le \sigma^2$. Тогда SGD с убывающим шагом $\alpha_k = \frac{2k+1}{2\mu(k+1)^2}$ гарантирует

$$\mathbb{E}[f(x_k) - f^*] \le \frac{L\sigma^2}{2\mu^2 k}.$$

Доказательство

1. Применяя стратегию убывающего шага $\alpha_k = \frac{2k+1}{2\mu(k+1)^2}$ к рекуррентному неравенству из предыдущей теоремы:

$$1-2\alpha_k\mu=\frac{k^2}{(k+1)^2} \qquad \mathbb{E}[f(x_{k+1}) - f^*] \le \frac{k^2}{(k+1)^2}[f(x_k) - f^*] + \frac{L\sigma^2(2k+1)^2}{8\mu^2(k+1)^4}$$

Гладкий PL-случай. Убывающий шаг

Theorem

Пусть f — L-гладкая, удовлетворяет PL с константой $\mu > 0$, а дисперсия стохастического градиента ограничена: $\mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2] \leq \sigma^2$. Тогда SGD с убывающим шагом $\alpha_k = \frac{2k+1}{2\mu(k+1)^2}$ гарантирует

$$\mathbb{E}[f(x_k) - f^*] \leq \frac{L\sigma^2}{2\mu^2 k}.$$

Доказательство

1. Применяя стратегию убывающего шага $\alpha_k = \frac{2k+1}{2\mu(k+1)^2}$ к рекуррентному неравенству из предыдущей теоремы:

$$\mathbb{E}[f(x_{k+1}) - f^*] \leq \frac{k^2}{(k+1)^2}[f(x_k) - f^*] + \frac{L\sigma^2(2k+1)^2}{8\mu^2(k+1)^4}$$

$$\leq \frac{k^2}{(k+1)^2}[f(x_k) - f^*] + \frac{L\sigma^2}{2\mu^2(k+1)^2}$$

Гладкий PL-случай. Убывающий шаг

Theorem

Пусть f — L-гладкая, удовлетворяет PL с константой $\mu > 0$, а дисперсия стохастического градиента ограничена: $\mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2] \le \sigma^2$. Тогда SGD с убывающим шагом $\alpha_k = \frac{2k+1}{2\mu(k+1)^2}$ гарантирует

$$\mathbb{E}[f(x_k) - f^*] \le \frac{L\sigma^2}{2\mu^2 k}.$$

Доказательство

1. Применяя стратегию убывающего шага $\alpha_k = \frac{2k+1}{2\mu(k+1)^2}$ к рекуррентному неравенству из предыдущей теоремы:

$$\mathbb{E}[f(x_{k+1}) - f^*] \le \frac{k^2}{(k+1)^2}[f(x_k) - f^*] + \frac{L\sigma^2(2k+1)^2}{8\mu^2(k+1)^4}$$

($1 - 2\alpha_k\mu = \frac{k^2}{(k+1)^2}$)

$$\le \frac{k^2}{(k+1)^2}[f(x_k) - f^*] + \frac{L\sigma^2}{2\mu^2(k+1)^2}$$

($(2k+1)^2 < (2k+2)^2 = 4(k+1)^2$)

2. Домножая обе части на $(k+1)^2$ и обозначая $\delta_f(k) \equiv k^2\mathbb{E}[f(x_k) - f^*]$:

$$\delta_f(k+1) \le \delta_f(k) + \frac{L\sigma^2}{2\mu^2}.$$

Гладкий PL-случай. Убывающий шаг

3. Суммируя предыдущее неравенство от $i = 0$ до k и используя $\delta_f(0) = 0$:

что и даёт заявленную скорость.

Гладкий PL-случай. Убывающий шаг

3. Суммируя предыдущее неравенство от $i = 0$ до k и используя $\delta_f(0) = 0$:

$$\delta_f(k+1) \leq \frac{L\sigma^2(k+1)}{2\mu^2}$$

что и даёт заявленную скорость.

Гладкий PL-случай. Убывающий шаг

3. Суммируя предыдущее неравенство от $i = 0$ до k и используя $\delta_f(0) = 0$:

$$\delta_f(k+1) \leq \frac{L\sigma^2(k+1)}{2\mu^2}$$

$$(k+1)^2\mathbb{E}[f(x_{k+1}) - f^*] \leq \frac{L\sigma^2(k+1)}{2\mu^2}$$

что и даёт заявленную скорость.

Гладкий PL-случай. Убывающий шаг

3. Суммируя предыдущее неравенство от $i = 0$ до k и используя $\delta_f(0) = 0$:

$$\delta_f(k+1) \leq \frac{L\sigma^2(k+1)}{2\mu^2}$$

$$(k+1)^2\mathbb{E}[f(x_{k+1}) - f^*] \leq \frac{L\sigma^2(k+1)}{2\mu^2}$$

$$\boxed{\mathbb{E}[f(x_k) - f^*] \leq \frac{L\sigma^2}{2\mu^2 k}}$$

что и даёт заявленную скорость.

Гладкий PL-случай. Убывающий шаг

3. Суммируя предыдущее неравенство от $i = 0$ до k и используя $\delta_f(0) = 0$:

$$\delta_f(k+1) \leq \frac{L\sigma^2(k+1)}{2\mu^2}$$

$$(k+1)^2\mathbb{E}[f(x_{k+1}) - f^*] \leq \frac{L\sigma^2(k+1)}{2\mu^2}$$

$$\mathbb{E}[f(x_k) - f^*] \leq \frac{L\sigma^2}{2\mu^2 k}$$

что и даёт заявленную скорость.

Гладкий PL-случай. Убывающий шаг

3. Суммируя предыдущее неравенство от $i = 0$ до k и используя $\delta_f(0) = 0$:

$$\delta_f(k+1) \leq \frac{L\sigma^2(k+1)}{2\mu^2}$$

$$(k+1)^2\mathbb{E}[f(x_{k+1}) - f^*] \leq \frac{L\sigma^2(k+1)}{2\mu^2}$$

$$\mathbb{E}[f(x_k) - f^*] \leq \frac{L\sigma^2}{2\mu^2 k}$$

что и даёт заявленную скорость.

- С постоянным шагом получаем линейную сходимость, но лишь до уровня шума $\frac{L\sigma^2\alpha}{4\mu}$.

Гладкий PL-случай. Убывающий шаг

3. Суммируя предыдущее неравенство от $i = 0$ до k и используя $\delta_f(0) = 0$:

$$\delta_f(k+1) \leq \frac{L\sigma^2(k+1)}{2\mu^2}$$

$$(k+1)^2\mathbb{E}[f(x_{k+1}) - f^*] \leq \frac{L\sigma^2(k+1)}{2\mu^2}$$

$$\mathbb{E}[f(x_k) - f^*] \leq \frac{L\sigma^2}{2\mu^2 k}$$

что и даёт заявленную скорость.

- С постоянным шагом получаем линейную сходимость, но лишь до уровня шума $\frac{L\sigma^2\alpha}{4\mu}$.
- С убывающим шагом сходимся к точному решению, но сублинейно: $\mathcal{O}(1/k)$ вместо $\mathcal{O}(\log(1/\varepsilon))$ у GD.

Гладкий выпуклый случай. Постоянный шаг

Theorem

Пусть f — выпуклая функция, а дисперсия стохастического градиента ограничена: $\mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2] \leq \sigma^2$ для всех k. Если SGD использует постоянный шаг $\alpha_k \equiv \alpha > 0$, то для любого $k \geq 1$

$$\mathbb{E}[f(\bar{x}_k) - f^*] \leq \frac{\|x_0 - x^*\|^2}{2\alpha k} + \frac{\alpha \sigma^2}{2}$$

где $\bar{x}_k = \frac{1}{k}\sum_{t=0}^{k-1} x_t$ — усреднённая по итерациям точка.

При выборе $\alpha = \frac{\|x_0 - x^*\|}{\sigma\sqrt{k}}$ получаем

$$\mathbb{E}[f(\bar{x}_k) - f^*] \leq \frac{\|x_0 - x^*\|\,\sigma}{\sqrt{k}} = \mathcal{O}\Big(\tfrac{1}{\sqrt{k}}\Big).$$

Гладкий выпуклый случай. Постоянный шаг

Доказательство

1. Начнём с разложения квадрата расстояния до минимума:

$$\|x_{k+1} - x^*\|^2 = \|x_k - \alpha \nabla f_{i_k}(x_k) - x^*\|^2 = \|x_k - x^*\|^2 - 2\alpha \langle \nabla f_{i_k}(x_k), x_k - x^* \rangle + \alpha^2 \|\nabla f_{i_k}(x_k)\|^2.$$

метод (SGD)

Гладкий выпуклый случай. Постоянный шаг

Доказательство

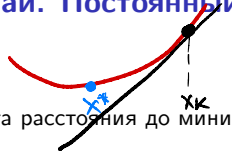
1. Начнём с разложения квадрата расстояния до минимума:

$$\|x_{k+1} - x^*\|^2 = \|x_k - \alpha \nabla f_{i_k}(x_k) - x^*\|^2 = \|x_k - x^*\|^2 - 2\alpha \langle \nabla f_{i_k}(x_k), x_k - x^* \rangle + \alpha^2 \|\nabla f_{i_k}(x_k)\|^2.$$

2. Возьмём условное математическое ожидание по i_k (обозначим $\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot \mid x_k]$). Используем несмещённость $\mathbb{E}_k[\nabla f_{i_k}(x_k)] = \nabla f(x_k)$, ограниченность дисперсии $\mathbb{E}_k[\|\nabla f_{i_k}(x_k)\|^2] \leq \sigma^2$ и выпуклость f (которая даёт $\langle \nabla f(x_k), x_k - x^* \rangle \geq f(x_k) - f^*$):

$$\begin{aligned} \mathbb{E}_k[\|x_{k+1} - x^*\|^2] &= \|x_k - x^*\|^2 - 2\alpha \langle \nabla f(x_k), x_k - x^* \rangle + \alpha^2 \mathbb{E}_k[\|\nabla f_{i_k}(x_k)\|^2] \\ &\leq \|x_k - x^*\|^2 - 2\alpha (f(x_k) - f^*) + \alpha^2 \sigma^2. \end{aligned}$$

Гладкий выпуклый случай. Постоянный шаг



$$f(x^*) \geq f(x_k) + \langle \nabla f(x_k), x^* - x_k \rangle$$

Доказательство

1. Начнём с разложения квадрата расстояния до минимума:

$$\|x_{k+1} - x^*\|^2 = \|x_k - \alpha \nabla f_{i_k}(x_k) - x^*\|^2 = \|x_k - x^*\|^2 - 2\alpha \langle \nabla f_{i_k}(x_k), x_k - x^* \rangle + \alpha^2 \|\nabla f_{i_k}(x_k)\|^2.$$

2. Возьмём условное математическое ожидание по i_k (обозначим $\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot \mid x_k]$). Используем несмещённость $\mathbb{E}_k[\nabla f_{i_k}(x_k)] = \nabla f(x_k)$, ограниченность дисперсии $\mathbb{E}_k[\|\nabla f_{i_k}(x_k)\|^2] \leq \sigma^2$ и выпуклость f (которая даёт $\langle \nabla f(x_k), x_k - x^* \rangle \geq f(x_k) - f^*$):

$$\begin{aligned} \mathbb{E}_k[\|x_{k+1} - x^*\|^2] &= \|x_k - x^*\|^2 - 2\alpha \langle \nabla f(x_k), x_k - x^* \rangle + \alpha^2 \mathbb{E}_k[\|\nabla f_{i_k}(x_k)\|^2] \\ &\leq \|x_k - x^*\|^2 - 2\alpha (f(x_k) - f^*) + \alpha^2 \sigma^2. \end{aligned}$$

3. Перенесём член с $f(x_k)$ в левую часть и возьмём полное математическое ожидание:

$$2\alpha\, \mathbb{E}[f(x_k) - f^*] \leq \mathbb{E}[\|x_k - x^*\|^2] - \mathbb{E}[\|x_{k+1} - x^*\|^2] + \alpha^2 \sigma^2.$$

Гладкий выпуклый случай. Постоянный шаг

4. Просуммируем (телескопически) по $t = 0, \dots, k-1$:

$$\begin{aligned}\sum_{t=0}^{k-1} 2\alpha\, \mathbb{E}[f(x_t) - f^*] &\leq \sum_{t=0}^{k-1} (\mathbb{E}[\|x_t - x^*\|^2] - \mathbb{E}[\|x_{t+1} - x^*\|^2]) + k\alpha^2\sigma^2 \\ &= \mathbb{E}[\|x_0 - x^*\|^2] - \mathbb{E}[\|x_k - x^*\|^2] + k\alpha^2\sigma^2 \\ &\leq \|x_0 - x^*\|^2 + k\alpha^2\sigma^2.\end{aligned}$$

Гладкий выпуклый случай. Постоянный шаг

4. Просуммируем (телескопически) по $t = 0, \dots, k-1$:

$$\begin{aligned}\sum_{t=0}^{k-1} 2\alpha\, \mathbb{E}[f(x_t) - f^*] &\leq \sum_{t=0}^{k-1} (\mathbb{E}[\|x_t - x^*\|^2] - \mathbb{E}[\|x_{t+1} - x^*\|^2]) + k\alpha^2\sigma^2 \\ &= \mathbb{E}[\|x_0 - x^*\|^2] - \mathbb{E}[\|x_k - x^*\|^2] + k\alpha^2\sigma^2 \\ &\leq \|x_0 - x^*\|^2 + k\alpha^2\sigma^2.\end{aligned}$$

5. Разделим на $2\alpha k$:

$$f(\bar{x}_k) - f^* \leq \frac{1}{k}\sum_{t=0}^{k-1} \mathbb{E}[f(x_t) - f^*] \leq \frac{\|x_0 - x^*\|^2}{2\alpha k} + \frac{\alpha\sigma^2}{2}.$$

Гладкий выпуклый случай. Постоянный шаг

4. Просуммируем (телескопически) по $t = 0, \dots, k-1$:

$$\begin{aligned}\sum_{t=0}^{k-1} 2\alpha\, \mathbb{E}[f(x_t) - f^*] &\leq \sum_{t=0}^{k-1} (\mathbb{E}[\|x_t - x^*\|^2] - \mathbb{E}[\|x_{t+1} - x^*\|^2]) + k\alpha^2\sigma^2 \\ &= \mathbb{E}[\|x_0 - x^*\|^2] - \mathbb{E}[\|x_k - x^*\|^2] + k\alpha^2\sigma^2 \\ &\leq \|x_0 - x^*\|^2 + k\alpha^2\sigma^2.\end{aligned}$$

5. Разделим на $2\alpha k$:

$$\frac{1}{k}\sum_{t=0}^{k-1} \mathbb{E}[f(x_t) - f^*] \leq \frac{\|x_0 - x^*\|^2}{2\alpha k} + \frac{\alpha\sigma^2}{2}.$$

6. Применим неравенство Йенсена для усреднённой точки $\bar{x}_k = \frac{1}{k}\sum_{t=0}^{k-1} x_t$, используя выпуклость f:

$$\mathbb{E}[f(\bar{x}_k) - f^*] \leq \frac{1}{k}\sum_{t=0}^{k-1} \mathbb{E}[f(x_t) - f^*] \leq \frac{\|x_0 - x^*\|^2}{2\alpha k} + \frac{\alpha\sigma^2}{2}.$$

Гладкий выпуклый случай. Убывающий шаг

Theorem

При тех же предположениях, но со спадом шага $\alpha_k = \frac{\alpha_0}{\sqrt{k+1}}$, $0 < \alpha_0 \leq \frac{1}{4L}$,

$$\boxed{\mathbb{E}[f(\bar{x}_k) - f^*] \leq \frac{5\|x_0 - x^*\|^2}{4\alpha_0\sqrt{k}} + 5\alpha_0\sigma^2\frac{\log(k+1)}{\sqrt{k}}} = \mathcal{O}\left(\frac{\log k}{\sqrt{k}}\right).$$

Гладкий выпуклый случай. Убывающий шаг

Theorem

При тех же предположениях, но со спадом шага $\alpha_k = \frac{\alpha_0}{\sqrt{k+1}}$, $0 < \alpha_0 \leq \frac{1}{4L}$,

$$\boxed{\mathbb{E}[f(\bar{x}_k) - f^*] \leq \frac{5\|x_0 - x^*\|^2}{4\alpha_0\sqrt{k}} + 5\alpha_0\sigma^2\frac{\log(k+1)}{\sqrt{k}}} = \mathcal{O}\left(\frac{\log k}{\sqrt{k}}\right).$$

- С постоянным шагом α получаем «остаточный» член $\frac{\alpha\sigma^2}{2}$ — SGD «зависает» в окрестности оптимума.

Гладкий выпуклый случай. Убывающий шаг

$\sum \alpha_k \uparrow\uparrow \quad \sum \alpha_k^2 \downarrow\downarrow$

Theorem

При тех же предположениях, но со спадом шага $\alpha_k = \frac{\alpha_0}{\sqrt{k+1}}$, $0 < \alpha_0 \leq \frac{1}{4L}$,

$$\boxed{\mathbb{E}[f(\bar{x}_k) - f^*] \leq \frac{5\|x_0 - x^*\|^2}{4\alpha_0\sqrt{k}} + 5\alpha_0\sigma^2\frac{\log(k+1)}{\sqrt{k}}} = \mathcal{O}\Big(\frac{\log k}{\sqrt{k}}\Big).$$

- С постоянным шагом α получаем «остаточный» член $\frac{\alpha\sigma^2}{2}$ — SGD «зависает» в окрестности оптимума.
- С убывающим шагом сходимся к точному решению, но скорость замедляется до $\mathcal{O}(\log k/\sqrt{k})$.

Мини-батч SGD

Мини-батч: компромисс между GD и SGD

Детерминированный метод использует все n градиентов:

$$\nabla f(x_k) = \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(x_k).$$

Стохастический метод аппроксимирует это одним сэмплом:

$$\nabla f_{i_k}(x_k) \approx \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(x_k).$$

Мини-батч: компромисс между GD и SGD

Детерминированный метод использует все n градиентов:

$$\nabla f(x_k) = \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(x_k).$$

Стохастический метод аппроксимирует это одним сэмплом:

$$\nabla f_{i_k}(x_k) \approx \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(x_k).$$

Распространённый вариант — использовать выборку B_k («мини-батч»):

$$\frac{1}{|B_k|} \sum_{i \in B_k} \nabla f_i(x_k) \approx \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(x_k),$$

что особенно удобно для векторизации и параллелизации на GPU/TPU.

Например, при 16 ядрах ставим $|B_k| = 16$ и считаем 16 градиентов одновременно.

Мини-батч как ГС с ошибкой

стох. градиент: $g_k = \frac{1}{B} \sum_{i \in B_k} \nabla f_i(x_k)$

SGD с мини-батчем B_k использует итерации:

$$x_{k+1} = x_k - \alpha_k \left(\frac{1}{|B_k|} \sum_{i \in B_k} \nabla f_i(x_k) \right).$$

Мини-батч как ГС с ошибкой

SGD с мини-батчем B_k использует итерации:

$$x_{k+1} = x_k - \alpha_k \left(\frac{1}{|B_k|} \sum_{i \in B_k} \nabla f_i(x_k) \right).$$

Рассмотрим это как «градиентный метод с ошибкой»:

$$x_{k+1} = x_k - \alpha_k (\nabla f(x_k) + e_k),$$

где e_k — разность между приближённым и истинным градиентами.

Мини-батч как ГС с ошибкой

SGD с мини-батчем B_k использует итерации:

$$x_{k+1} = x_k - \alpha_k \left(\frac{1}{|B_k|} \sum_{i \in B_k} \nabla f_i(x_k) \right).$$

Рассмотрим это как «градиентный метод с ошибкой»:

$$x_{k+1} = x_k - \alpha_k (\nabla f(x_k) + e_k),$$

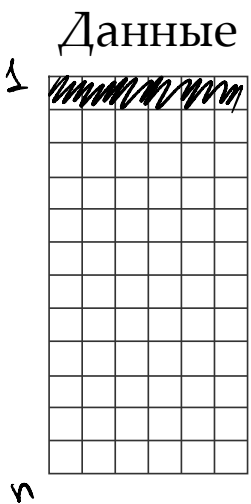
где e_k — разность между приближённым и истинным градиентами.

При $\alpha_k = \frac{1}{L}$, используя лемму спуска:

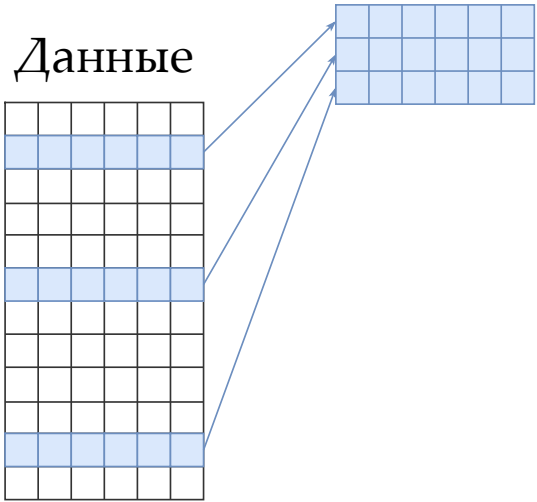
$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L} \|\nabla f(x_k)\|^2 + \frac{1}{2L} \|e_k\|^2,$$

для любой ошибки e_k. Чем меньше дисперсия ошибки e_k, тем ближе шаг к шагу обычного градиентного спуска.

Идея SGD и батчей



Идея SGD и батчей

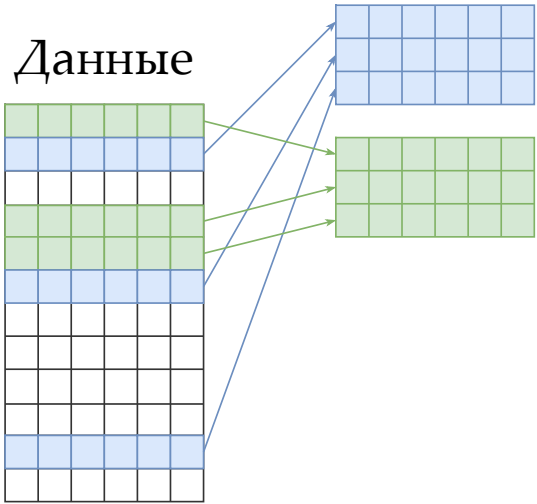


$b = 3$

посчитали g_1

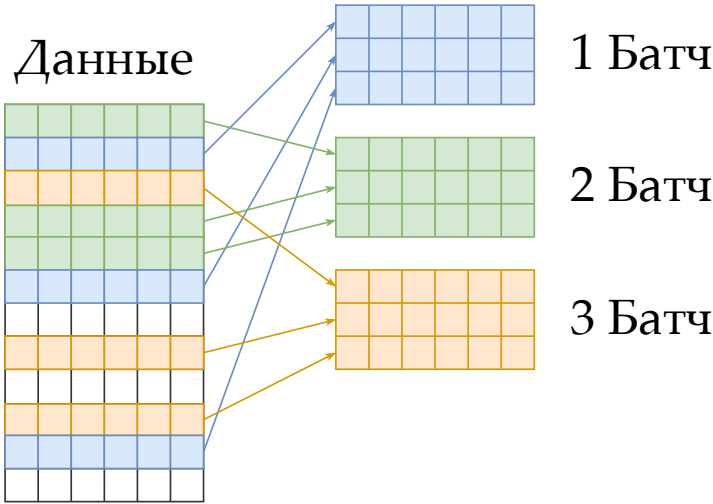
$x_1 = x_0 - \alpha g_1$

Идея SGD и батчей

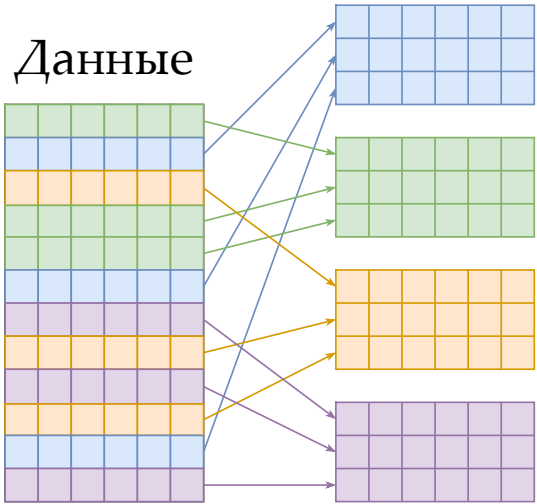


$$x_2 = x_1 - d g_2$$

Идея SGD и батчей



Идея SGD и батчей



① в ↑↑ # шаг в эпохе ↓↓

② в ↑↑ время шага const *

* при наличии дост. памяти

③ в ↑↑ время эпохи ↓↓

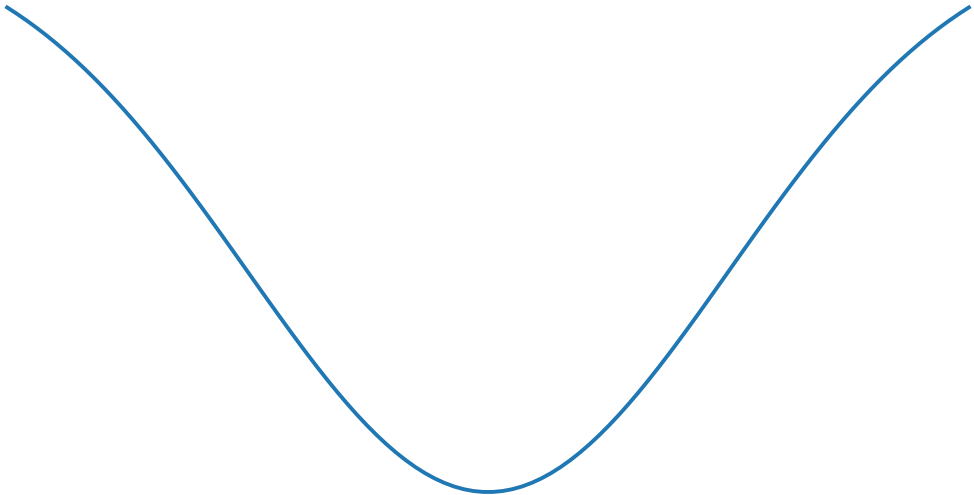
Эпоха

④ в ↑↑ SGD ↓ GD

Поведение SGD на практике

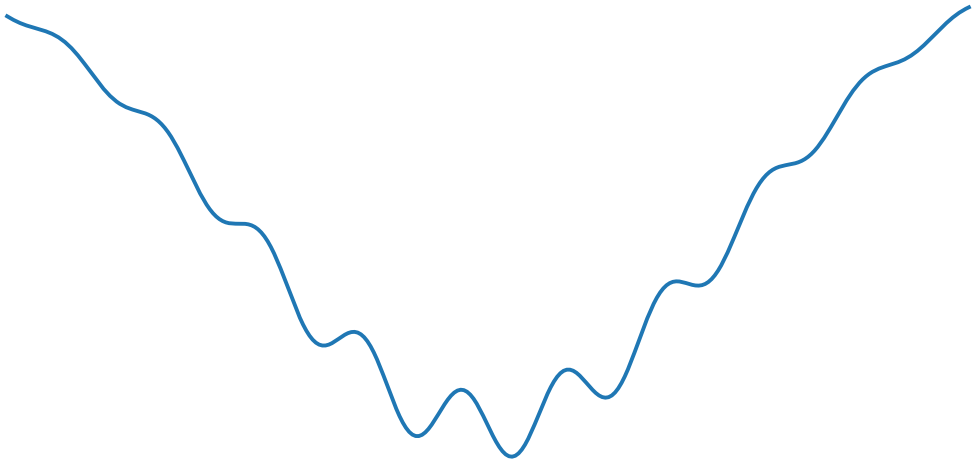
Детерминированный против стохастического

Градиентный спуск сходится к локальному минимуму



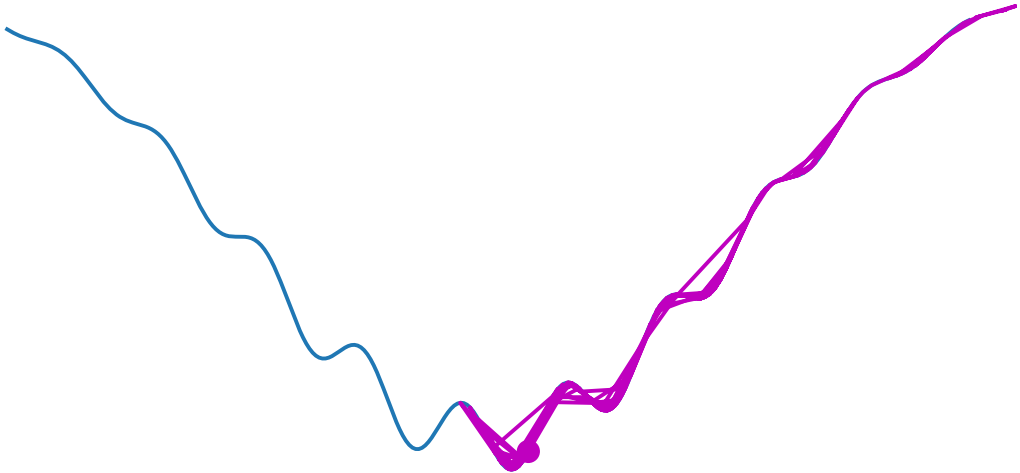
Детерминированный против стохастического

Градиентный спуск
сходится к локальному минимуму



SGD помогает убежать из локальных минимумов

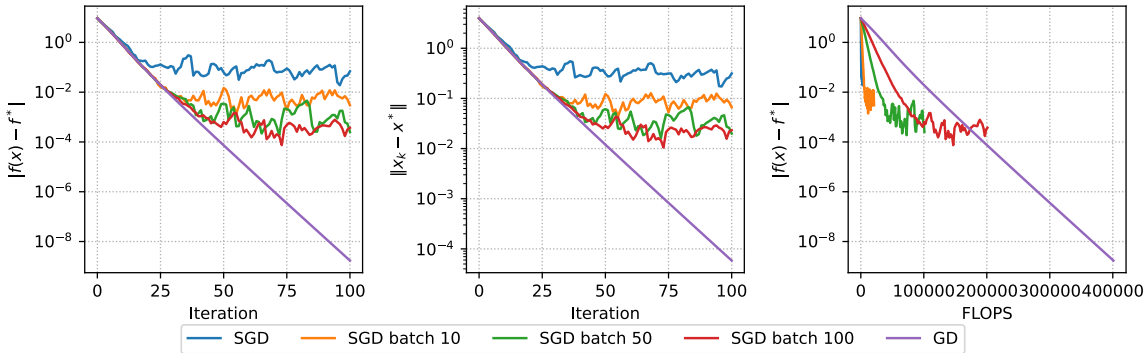
Стохастический градиентный спуск
выпрыгивает из локальных минимумов



Главная проблема SGD

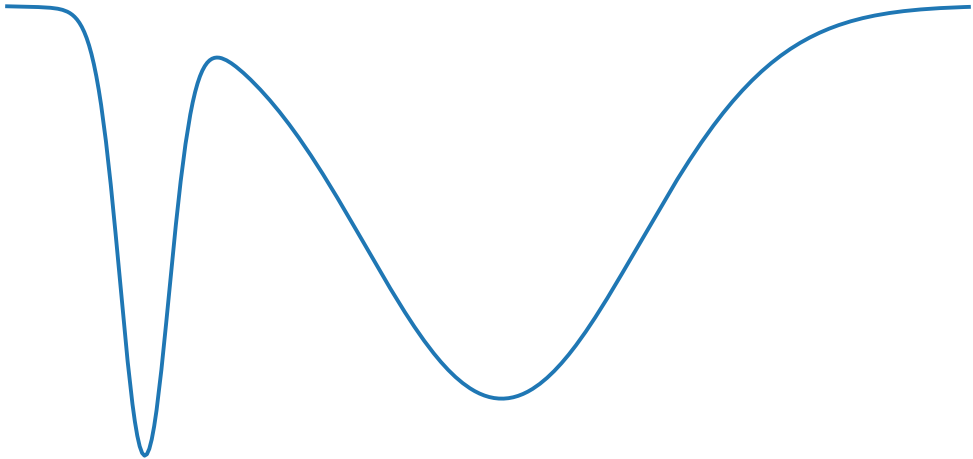
$$f(x) = \frac{\mu}{2}\|x\|_2^2 + \frac{1}{m}\sum_{i=1}^{m}\log(1+\exp(-y_i\langle a_i, x\rangle)) \to \min_{x\in\mathbb{R}^n}$$

Strongly convex binary logistic regression. m=200, n=10, mu=1.



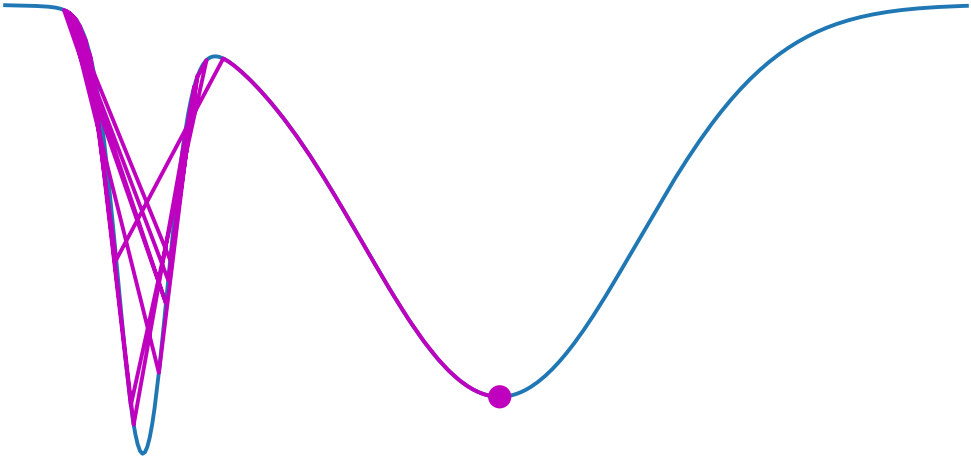
GD сходится к ближайшему минимуму

Градиентный спуск с маленьким шагом
сходится в узкий локальный минимум



SGD «расходится» вблизи минимума

Градиентный спуск с большим шагом
избегает узкого локального минимума



Сводные результаты и переход к редукции дисперсии

Основные результаты сходимости SGD

Пусть f — L-гладкая μ-сильно выпуклая функция, а дисперсия стохастического градиента конечна: $\mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2] \leq \sigma^2$. Тогда SGD с постоянным шагом $\alpha < \frac{1}{2\mu}$ гарантирует

$$\mathbb{E}[f(x_{k+1}) - f^*] \leq (1 - 2\alpha\mu)^k [f(x_0) - f^*] + \frac{L\sigma^2\alpha}{4\mu}.$$

Основные результаты сходимости SGD

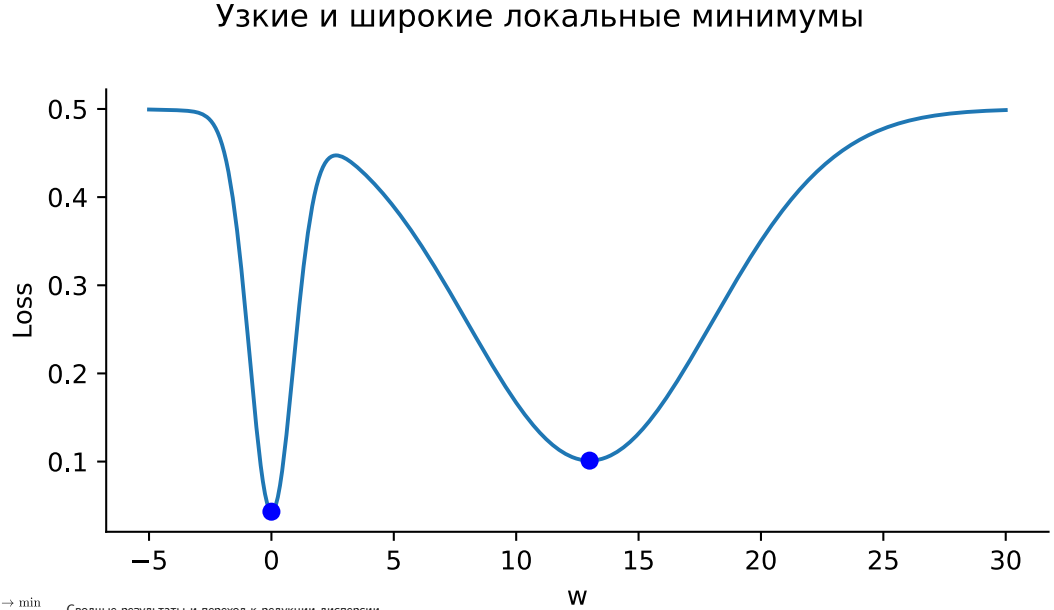
Пусть f — L-гладкая μ-сильно выпуклая функция, а дисперсия стохастического градиента конечна: $\mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2] \leq \sigma^2$. Тогда SGD с постоянным шагом $\alpha < \frac{1}{2\mu}$ гарантирует

$$\mathbb{E}[f(x_{k+1}) - f^*] \leq (1 - 2\alpha\mu)^k [f(x_0) - f^*] + \frac{L\sigma^2\alpha}{4\mu}.$$

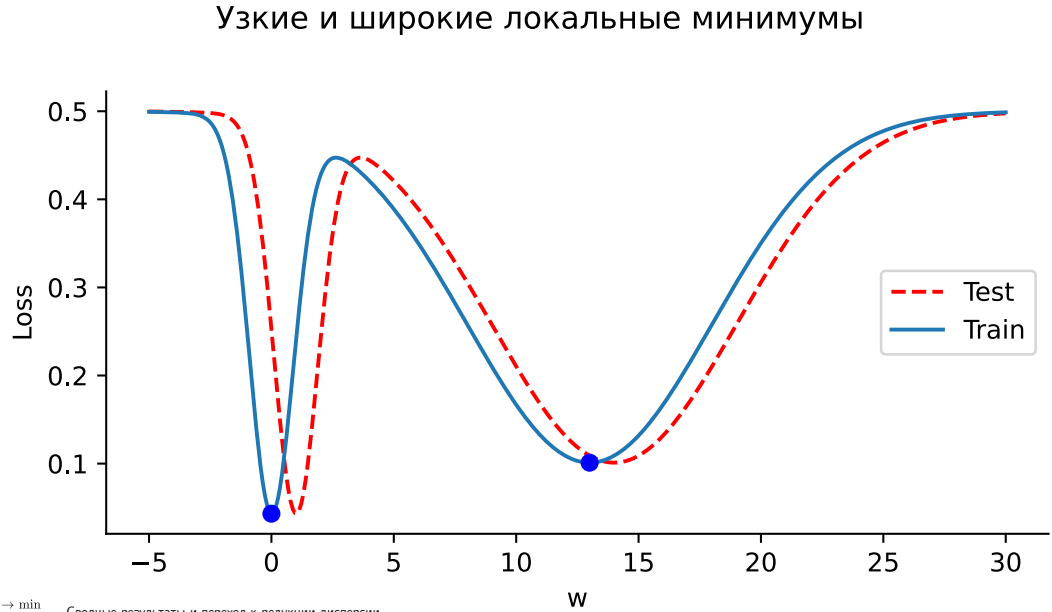
Пусть f — L-гладкая μ-сильно выпуклая функция, а дисперсия стохастического градиента конечна: $\mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2] \leq \sigma^2$. Тогда SGD с убывающим шагом $\alpha_k = \frac{2k+1}{2\mu(k+1)^2}$ гарантирует

$$\mathbb{E}[f(x_{k+1}) - f^*] \leq \frac{L\sigma^2}{2\mu^2(k+1)}.$$

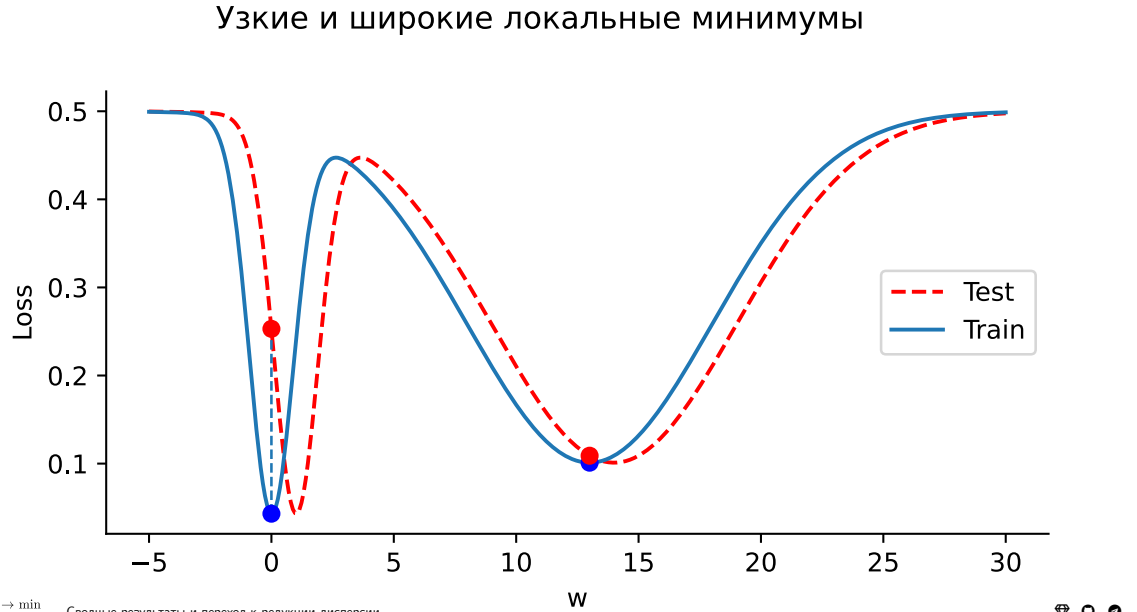
Ширина локальных минимумов



Ширина локальных минимумов



Ширина локальных минимумов



Итоги

- **SGD** — дешёвые итерации ($\mathcal{O}(1)$ вместо $\mathcal{O}(n)$), но медленная сублинейная сходимость $\mathcal{O}(1/k)$ вместо линейной у GD.

Итоги

- **SGD** — дешёвые итерации ($\mathcal{O}(1)$ вместо $\mathcal{O}(n)$), но медленная сублинейная сходимость $\mathcal{O}(1/k)$ вместо линейной у GD.
- **Постоянный шаг**: SGD не сходится к точному решению — остаётся «окрестность» размера $\frac{L\sigma^2\alpha}{4\mu}$.

Итоги

- **SGD** — дешёвые итерации ($\mathcal{O}(1)$ вместо $\mathcal{O}(n)$), но медленная сублинейная сходимость $\mathcal{O}(1/k)$ вместо линейной у GD.
- **Постоянный шаг**: SGD не сходится к точному решению — остаётся «окрестность» размера $\frac{L\sigma^2\alpha}{4\mu}$.
- **Убывающий шаг**: сходимость к точному решению, но сублинейно.

Итоги

- **SGD** — дешёвые итерации ($\mathcal{O}(1)$ вместо $\mathcal{O}(n)$), но медленная сублинейная сходимость $\mathcal{O}(1/k)$ вместо линейной у GD.
- **Постоянный шаг**: SGD не сходится к точному решению — остаётся «окрестность» размера $\frac{L\sigma^2\alpha}{4\mu}$.
- **Убывающий шаг**: сходимость к точному решению, но сублинейно.
- **Мини-батч** — компромисс между GD и SGD; удобен для GPU-параллелизации.

Итоги

- **SGD** — дешёвые итерации ($\mathcal{O}(1)$ вместо $\mathcal{O}(n)$), но медленная сублинейная сходимость $\mathcal{O}(1/k)$ вместо линейной у GD.
- **Постоянный шаг**: SGD не сходится к точному решению — остаётся «окрестность» размера $\frac{L\sigma^2\alpha}{4\mu}$.
- **Убывающий шаг**: сходимость к точному решению, но сублинейно.
- **Мини-батч** — компромисс между GD и SGD; удобен для GPU-параллелизации.
- SGD помогает убежать из «узких» локальных минимумов — это связано с лучшей обобщающей способностью в глубоком обучении.

Итоги

- **SGD** — дешёвые итерации ($\mathcal{O}(1)$ вместо $\mathcal{O}(n)$), но медленная сублинейная сходимость $\mathcal{O}(1/k)$ вместо линейной у GD.
- **Постоянный шаг**: SGD не сходится к точному решению — остаётся «окрестность» размера $\frac{L\sigma^2\alpha}{4\mu}$.
- **Убывающий шаг**: сходимость к точному решению, но сублинейно.
- **Мини-батч** — компромисс между GD и SGD; удобен для GPU-параллелизации.
- SGD помогает убежать из «узких» локальных минимумов — это связано с лучшей обобщающей способностью в глубоком обучении.
- Моментные методы и ускорения **не улучшают** скорость SGD: узкое место — дисперсия, а не число обусловленности.

Итоги

- **SGD** — дешёвые итерации ($\mathcal{O}(1)$ вместо $\mathcal{O}(n)$), но медленная сублинейная сходимость $\mathcal{O}(1/k)$ вместо линейной у GD.
- **Постоянный шаг**: SGD не сходится к точному решению — остаётся «окрестность» размера $\frac{L\sigma^2\alpha}{4\mu}$.
- **Убывающий шаг**: сходимость к точному решению, но сублинейно.
- **Мини-батч** — компромисс между GD и SGD; удобен для GPU-параллелизации.
- SGD помогает убежать из «узких» локальных минимумов — это связано с лучшей обобщающей способностью в глубоком обучении.
- Моментные методы и ускорения **не улучшают** скорость SGD: узкое место — дисперсия, а не число обусловленности.

Итоги

- **SGD** — дешёвые итерации ($\mathcal{O}(1)$ вместо $\mathcal{O}(n)$), но медленная сублинейная сходимость $\mathcal{O}(1/k)$ вместо линейной у GD.
- **Постоянный шаг**: SGD не сходится к точному решению — остаётся «окрестность» размера $\frac{L\sigma^2\alpha}{4\mu}$.
- **Убывающий шаг**: сходимость к точному решению, но сублинейно.
- **Мини-батч** — компромисс между GD и SGD; удобен для GPU-параллелизации.
- SGD помогает убежать из «узких» локальных минимумов — это связано с лучшей обобщающей способностью в глубоком обучении.
- Моментные методы и ускорения **не улучшают** скорость SGD: узкое место — дисперсия, а не число обусловленности.

Дальше: можно ли достичь линейной сходимости как у GD при стоимости итерации как у SGD? Да — **методы редукции дисперсии** (SAG, SVRG, SAGA), которым будет посвящена следующая лекция.