



# Стохастический градиентный спуск

Даня Меркулов

Оптимизация для всех! ЦУ



## Задача с конечной суммой

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 действует следующим образом:

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

- Сходимость с постоянным  $\alpha$  или поиском по линии.

## Задача с конечной суммой

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 действует следующим образом:

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

- Сходимость с постоянным  $\alpha$  или поиском по линии.
- Стоимость итерации линейна по  $n$ . Для ImageNet  $n \approx 1.4 \cdot 10^7$ , для WikiText  $n \approx 10^8$ . Для FineWeb  $n \approx 15 \cdot 10^{12}$  токенов.

## Задача с конечной суммой

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 действует следующим образом:

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

- Сходимость с постоянным  $\alpha$  или поиском по линии.
- Стоимость итерации линейна по  $n$ . Для ImageNet  $n \approx 1.4 \cdot 10^7$ , для WikiText  $n \approx 10^8$ . Для FineWeb  $n \approx 15 \cdot 10^{12}$  токенов.

## Задача с конечной суммой

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 действует следующим образом:

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

- Сходимость с постоянным  $\alpha$  или поиском по линии.
- Стоимость итерации линейна по  $n$ . Для ImageNet  $n \approx 1.4 \cdot 10^7$ , для WikiText  $n \approx 10^8$ . Для FineWeb  $n \approx 15 \cdot 10^{12}$  токенов.

Давайте перейдем от полного вычисления градиента к его несмещенной оценке, когда мы случайным образом выбираем индекс  $i_k$  точки на каждой итерации равномерно:

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

С  $p(i_k = i) = \frac{1}{n}$ , стохастический градиент является несмещенной оценкой градиента, которая задается следующим образом:

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

## Результаты для градиентного спуска

Стохастические итерации в  $n$  раз быстрее, но сколько итераций потребуется для достижения заданной точности?

Если  $\nabla f$  является липшицевым, то мы получаем:

| Предположение | Детерминированный градиентный спуск | Стохастический градиентный спуск |
|---------------|-------------------------------------|----------------------------------|
| PL            | $\mathcal{O}(\log(1/\varepsilon))$  |                                  |
| Выпуклый      | $\mathcal{O}(1/\varepsilon)$        |                                  |
| Невыпуклый    | $\mathcal{O}(1/\varepsilon)$        |                                  |

## Результаты для градиентного спуска

Стохастические итерации в  $n$  раз быстрее, но сколько итераций потребуется для достижения заданной точности?

Если  $\nabla f$  является липшицевым, то мы получаем:

| Предположение | Детерминированный градиентный спуск | Стохастический градиентный спуск |
|---------------|-------------------------------------|----------------------------------|
| PL            | $\mathcal{O}(\log(1/\varepsilon))$  | $\mathcal{O}(1/\varepsilon)$     |
| Выпуклый      | $\mathcal{O}(1/\varepsilon)$        | $\mathcal{O}(1/\varepsilon^2)$   |
| Невыпуклый    | $\mathcal{O}(1/\varepsilon)$        | $\mathcal{O}(1/\varepsilon^2)$   |

- Стохастический градиентный спуск имеет низкую стоимость итерации, но медленную скорость сходимости.

## Результаты для градиентного спуска

Стохастические итерации в  $n$  раз быстрее, но сколько итераций потребуется для достижения заданной точности?

Если  $\nabla f$  является липшицевым, то мы получаем:

| Предположение | Детерминированный градиентный спуск | Стохастический градиентный спуск |
|---------------|-------------------------------------|----------------------------------|
| PL            | $\mathcal{O}(\log(1/\varepsilon))$  | $\mathcal{O}(1/\varepsilon)$     |
| Выпуклый      | $\mathcal{O}(1/\varepsilon)$        | $\mathcal{O}(1/\varepsilon^2)$   |
| Невыпуклый    | $\mathcal{O}(1/\varepsilon)$        | $\mathcal{O}(1/\varepsilon^2)$   |

- Стохастический градиентный спуск имеет низкую стоимость итерации, но медленную скорость сходимости.
  - Сублинейная скорость даже в сильно выпуклом случае.

## Результаты для градиентного спуска

Стохастические итерации в  $n$  раз быстрее, но сколько итераций потребуется для достижения заданной точности?

Если  $\nabla f$  является липшицевым, то мы получаем:

| Предположение | Детерминированный градиентный спуск | Стохастический градиентный спуск |
|---------------|-------------------------------------|----------------------------------|
| PL            | $\mathcal{O}(\log(1/\varepsilon))$  | $\mathcal{O}(1/\varepsilon)$     |
| Выпуклый      | $\mathcal{O}(1/\varepsilon)$        | $\mathcal{O}(1/\varepsilon^2)$   |
| Невыпуклый    | $\mathcal{O}(1/\varepsilon)$        | $\mathcal{O}(1/\varepsilon^2)$   |

- Стохастический градиентный спуск имеет низкую стоимость итерации, но медленную скорость сходимости.
  - Сублинейная скорость даже в сильно выпуклом случае.
  - Оценки скорости не могут быть улучшены при стандартных предположениях.

## Результаты для градиентного спуска

Стохастические итерации в  $n$  раз быстрее, но сколько итераций потребуется для достижения заданной точности?

Если  $\nabla f$  является липшицевым, то мы получаем:

| Предположение | Детерминированный градиентный спуск | Стохастический градиентный спуск |
|---------------|-------------------------------------|----------------------------------|
| PL            | $\mathcal{O}(\log(1/\varepsilon))$  | $\mathcal{O}(1/\varepsilon)$     |
| Выпуклый      | $\mathcal{O}(1/\varepsilon)$        | $\mathcal{O}(1/\varepsilon^2)$   |
| Невыпуклый    | $\mathcal{O}(1/\varepsilon)$        | $\mathcal{O}(1/\varepsilon^2)$   |

- Стохастический градиентный спуск имеет низкую стоимость итерации, но медленную скорость сходимости.
  - Сублинейная скорость даже в сильно выпуклом случае.
  - Оценки скорости не могут быть улучшены при стандартных предположениях.
  - Оракул возвращает несмешенную аппроксимацию градиента с ограниченной дисперсией.

## Результаты для градиентного спуска

Стохастические итерации в  $n$  раз быстрее, но сколько итераций потребуется для достижения заданной точности?

Если  $\nabla f$  является липшицевым, то мы получаем:

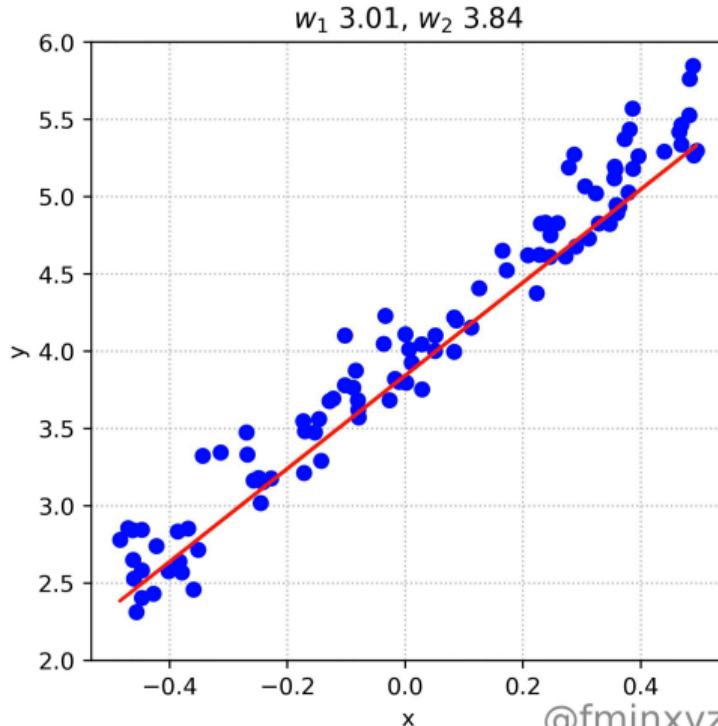
| Предположение | Детерминированный градиентный спуск | Стохастический градиентный спуск |
|---------------|-------------------------------------|----------------------------------|
| PL            | $\mathcal{O}(\log(1/\varepsilon))$  | $\mathcal{O}(1/\varepsilon)$     |
| Выпуклый      | $\mathcal{O}(1/\varepsilon)$        | $\mathcal{O}(1/\varepsilon^2)$   |
| Невыпуклый    | $\mathcal{O}(1/\varepsilon)$        | $\mathcal{O}(1/\varepsilon^2)$   |

- Стохастический градиентный спуск имеет низкую стоимость итерации, но медленную скорость сходимости.
  - Сублинейная скорость даже в сильно выпуклом случае.
  - Оценки скорости не могут быть улучшены при стандартных предположениях.
  - Оракул возвращает несмешенную аппроксимацию градиента с ограниченной дисперсией.
- Методы с моментом и квази-Ньютоновские методы не улучшают скорость в стохастическом случае, а только могут улучшить константные множители (бутылочное горлышко — дисперсия, а не число обусловленности).



## Типичное поведение

Stochastic Gradient Descent. Batch = 2



@fminxyz

## Сходимость

Липшицевость градиента означает:

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

## Сходимость

Липшицевость градиента означает:

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

Используя (SGD):

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

## Сходимость

Липшицевость градиента означает:

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

Используя (SGD):

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

Теперь возьмем матожидание по  $i_k$ :

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

## Сходимость

Липшицевость градиента означает:

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

Используя (SGD):

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

Теперь возьмем матожидание по  $i_k$ :

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

Используя линейность матожидания:

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

## Сходимость

Липшицевость градиента означает:

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

Используя (SGD):

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

Теперь возьмем матожидание по  $i_k$ :

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

Используя линейность матожидания:

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

Поскольку равномерное выборочное распределение означает несмешенную оценку градиента:  
 $\mathbb{E}[\nabla f_{i_k}(x_k)] = \nabla f(x_k)$ :

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

## Гладкий PL-случай с постоянным шагом

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

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

Начнем 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 с постоянным шагом

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

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

Начнем с неравенства (1):

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

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

## Гладкий PL-случай с постоянным шагом

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

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

Начнем с неравенства (1):

$$\begin{aligned}\mathbb{E}[f(x_{k+1})] &\leq f(x_k) - \alpha_k \|\nabla f(x_k)\|^2 + \alpha_k^2 \frac{L}{2} \mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2] \\ \text{PL: } \|\nabla f(x_k)\|^2 \geq 2\mu(f(x_k) - f^*) &\leq f(x_k) - 2\alpha_k \mu(f(x_k) - f^*) + \alpha_k^2 \frac{L}{2} \mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2]\end{aligned}$$

## Гладкий PL-случай с постоянным шагом

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

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

Начнем с неравенства (1):

$$\begin{aligned}\mathbb{E}[f(x_{k+1})] &\leq f(x_k) - \alpha_k \|\nabla f(x_k)\|^2 + \alpha_k^2 \frac{L}{2} \mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2] \\ \text{PL: } \|\nabla f(x_k)\|^2 \geq 2\mu(f(x_k) - f^*) &\leq f(x_k) - 2\alpha_k \mu(f(x_k) - f^*) + \alpha_k^2 \frac{L}{2} \mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2]\end{aligned}$$

Вычтем  $f^*$

## Гладкий PL-случай с постоянным шагом

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

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

Начнем с неравенства (1):

$$\begin{aligned}\mathbb{E}[f(x_{k+1})] &\leq f(x_k) - \alpha_k \|\nabla f(x_k)\|^2 + \alpha_k^2 \frac{L}{2} \mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2] \\ \text{PL: } \|\nabla f(x_k)\|^2 \geq 2\mu(f(x_k) - f^*) &\leq f(x_k) - 2\alpha_k \mu(f(x_k) - f^*) + \alpha_k^2 \frac{L}{2} \mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2] \\ \text{Вычтем } f^* \quad \mathbb{E}[f(x_{k+1})] - f^* &\leq (f(x_k) - f^*) - 2\alpha_k \mu(f(x_k) - f^*) + \alpha_k^2 \frac{L}{2} \mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2]\end{aligned}$$

## Гладкий PL-случай с постоянным шагом

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

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

Начнем с неравенства (1):

$$\begin{aligned}\mathbb{E}[f(x_{k+1})] &\leq f(x_k) - \alpha_k \|\nabla f(x_k)\|^2 + \alpha_k^2 \frac{L}{2} \mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2] \\ \text{PL: } \|\nabla f(x_k)\|^2 \geq 2\mu(f(x_k) - f^*) &\leq f(x_k) - 2\alpha_k \mu(f(x_k) - f^*) + \alpha_k^2 \frac{L}{2} \mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2] \\ \text{Вычтем } f^* \quad \mathbb{E}[f(x_{k+1})] - f^* &\leq (f(x_k) - f^*) - 2\alpha_k \mu(f(x_k) - f^*) + \alpha_k^2 \frac{L}{2} \mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2] \\ \text{Переставляем} \quad &\leq (1 - 2\alpha_k \mu)[f(x_k) - f^*] + \alpha_k^2 \frac{L}{2} \mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2]\end{aligned}$$

## Гладкий PL-случай с постоянным шагом

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

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

Начнем с неравенства (1):

$$\begin{aligned}\mathbb{E}[f(x_{k+1})] &\leq f(x_k) - \alpha_k \|\nabla f(x_k)\|^2 + \alpha_k^2 \frac{L}{2} \mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2] \\ \text{PL: } \|\nabla f(x_k)\|^2 \geq 2\mu(f(x_k) - f^*) &\leq f(x_k) - 2\alpha_k \mu(f(x_k) - f^*) + \alpha_k^2 \frac{L}{2} \mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2] \\ \text{Вычтем } f^* \quad \mathbb{E}[f(x_{k+1})] - f^* &\leq (f(x_k) - f^*) - 2\alpha_k \mu(f(x_k) - f^*) + \alpha_k^2 \frac{L}{2} \mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2] \\ \text{Переставляем} \quad &\leq (1 - 2\alpha_k \mu)[f(x_k) - f^*] + \alpha_k^2 \frac{L}{2} \mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2]\end{aligned}$$

Ограниченност дисперсии:  $\mathbb{E}[\|\nabla f_i(x_k)\|^2] \leq \sigma^2$

## Гладкий PL-случай с постоянным шагом

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

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

Начнем с неравенства (1):

$$\begin{aligned}\mathbb{E}[f(x_{k+1})] &\leq f(x_k) - \alpha_k \|\nabla f(x_k)\|^2 + \alpha_k^2 \frac{L}{2} \mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2] \\ \text{PL: } \|\nabla f(x_k)\|^2 \geq 2\mu(f(x_k) - f^*) &\leq f(x_k) - 2\alpha_k \mu(f(x_k) - f^*) + \alpha_k^2 \frac{L}{2} \mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2] \\ \text{Вычтем } f^* \quad \mathbb{E}[f(x_{k+1})] - f^* &\leq (f(x_k) - f^*) - 2\alpha_k \mu(f(x_k) - f^*) + \alpha_k^2 \frac{L}{2} \mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2] \\ \text{Переставляем} &\leq (1 - 2\alpha_k \mu)[f(x_k) - f^*] + \alpha_k^2 \frac{L}{2} \mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2] \\ \text{Ограниченност дисперсии: } \mathbb{E}[\|\nabla f_i(x_k)\|^2] \leq \sigma^2 &\leq (1 - 2\alpha_k \mu)[f(x_k) - f^*] + \frac{L\sigma^2\alpha_k^2}{2}.\end{aligned}$$

## Сходимость. Гладкий PL-случай.

- i** Пусть  $f$  —  $L$ -гладкая функция, удовлетворяющая условию Поляка-Лоясиевича (PL) с константой  $\mu > 0$ , а дисперсия стохастического градиента ограничена:  $\mathbb{E}[\|\nabla f_i(x_k)\|^2] \leq \sigma^2$ . Тогда стохастический градиентный спуск с убывающим шагом  $\alpha_k = \frac{2k+1}{2\mu(k+1)^2}$  гарантирует

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

1. Рассмотрим стратегию **убывающего шага** с  $\alpha_k = \frac{2k+1}{2\mu(k+1)^2}$ , тогда мы получаем

## Сходимость. Гладкий PL-случай.

- i** Пусть  $f$  —  $L$ -гладкая функция, удовлетворяющая условию Поляка-Лоясиевича (PL) с константой  $\mu > 0$ , а дисперсия стохастического градиента ограничена:  $\mathbb{E}[\|\nabla f_i(x_k)\|^2] \leq \sigma^2$ . Тогда стохастический градиентный спуск с убывающим шагом  $\alpha_k = \frac{2k+1}{2\mu(k+1)^2}$  гарантирует

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

1. Рассмотрим стратегию **убывающего шага** с  $\alpha_k = \frac{2k+1}{2\mu(k+1)^2}$ , тогда мы получаем

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

## Сходимость. Гладкий PL-случай.

- i** Пусть  $f$  —  $L$ -гладкая функция, удовлетворяющая условию Поляка-Лоясиевича (PL) с константой  $\mu > 0$ , а дисперсия стохастического градиента ограничена:  $\mathbb{E}[\|\nabla f_i(x_k)\|^2] \leq \sigma^2$ . Тогда стохастический градиентный спуск с убывающим шагом  $\alpha_k = \frac{2k+1}{2\mu(k+1)^2}$  гарантирует

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

1. Рассмотрим стратегию **убывающего шага** с  $\alpha_k = \frac{2k+1}{2\mu(k+1)^2}$ , тогда мы получаем

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

## Сходимость. Гладкий PL-случай.

- i** Пусть  $f$  —  $L$ -гладкая функция, удовлетворяющая условию Поляка-Лоясиевича (PL) с константой  $\mu > 0$ , а дисперсия стохастического градиента ограничена:  $\mathbb{E}[\|\nabla f_i(x_k)\|^2] \leq \sigma^2$ . Тогда стохастический градиентный спуск с убывающим шагом  $\alpha_k = \frac{2k+1}{2\mu(k+1)^2}$  гарантирует

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

1. Рассмотрим стратегию **убывающего шага** с  $\alpha_k = \frac{2k+1}{2\mu(k+1)^2}$ , тогда мы получаем

$$\begin{aligned} 1 - 2\alpha_k \mu &= \frac{(k+1)^2}{(k+1)^2} - \frac{2k+1}{(k+1)^2} = \frac{k^2}{(k+1)^2} & \mathbb{E}[f(x_{k+1}) - f^*] &\leq \frac{k^2}{(k+1)^2} [f(x_k) - f^*] + \frac{L\sigma^2(2k+1)^2}{8\mu^2(k+1)^4} \\ (2k+1)^2 &< (2k+2)^2 = 4(k+1)^2 & \leq \frac{k^2}{(k+1)^2} [f(x_k) - f^*] + \frac{L\sigma^2}{2\mu^2(k+1)^2} \end{aligned}$$

## Сходимость. Гладкий PL-случай.

- i** Пусть  $f$  —  $L$ -гладкая функция, удовлетворяющая условию Поляка-Лоясиевича (PL) с константой  $\mu > 0$ , а дисперсия стохастического градиента ограничена:  $\mathbb{E}[\|\nabla f_i(x_k)\|^2] \leq \sigma^2$ . Тогда стохастический градиентный спуск с убывающим шагом  $\alpha_k = \frac{2k+1}{2\mu(k+1)^2}$  гарантирует

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

1. Рассмотрим стратегию **убывающего шага** с  $\alpha_k = \frac{2k+1}{2\mu(k+1)^2}$ , тогда мы получаем

$$\begin{aligned} 1 - 2\alpha_k \mu &= \frac{(k+1)^2}{(k+1)^2} - \frac{2k+1}{(k+1)^2} = \frac{k^2}{(k+1)^2} & \mathbb{E}[f(x_{k+1}) - f^*] &\leq \frac{k^2}{(k+1)^2} [f(x_k) - f^*] + \frac{L\sigma^2(2k+1)^2}{8\mu^2(k+1)^4} \\ (2k+1)^2 &< (2k+2)^2 = 4(k+1)^2 & \leq \frac{k^2}{(k+1)^2} [f(x_k) - f^*] + \frac{L\sigma^2}{2\mu^2(k+1)^2} \end{aligned}$$

2. Умножив обе части на  $(k+1)^2$  и пусть  $\delta_f(k) \equiv k^2 \mathbb{E}[f(x_k) - f^*]$  мы получаем

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

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

## Сходимость. Гладкий PL-случай.

3. Просуммируем предыдущее неравенство от  $i = 0$  до  $k$  и используем тот факт, что  $\delta_f(0) = 0$  мы получаем

которое дает указанную скорость.

## Сходимость. Гладкий PL-случай.

3. Просуммируем предыдущее неравенство от  $i = 0$  до  $k$  и используем тот факт, что  $\delta_f(0) = 0$  мы получаем

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

которое дает указанную скорость.

## Сходимость. Гладкий PL-случай.

3. Просуммируем предыдущее неравенство от  $i = 0$  до  $k$  и используем тот факт, что  $\delta_f(0) = 0$  мы получаем

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

$$\sum_{i=0}^k [\delta_f(i+1) - \delta_f(i)] \leq \sum_{i=0}^k \frac{L\sigma^2}{2\mu^2}$$

которое дает указанную скорость.

## Сходимость. Гладкий PL-случай.

3. Просуммируем предыдущее неравенство от  $i = 0$  до  $k$  и используем тот факт, что  $\delta_f(0) = 0$  мы получаем

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

$$\sum_{i=0}^k [\delta_f(i+1) - \delta_f(i)] \leq \sum_{i=0}^k \frac{L\sigma^2}{2\mu^2}$$

$$\delta_f(k+1) - \delta_f(0) \leq \frac{L\sigma^2(k+1)}{2\mu^2}$$

которое дает указанную скорость.

## Сходимость. Гладкий PL-случай.

3. Просуммируем предыдущее неравенство от  $i = 0$  до  $k$  и используем тот факт, что  $\delta_f(0) = 0$  мы получаем

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

$$\sum_{i=0}^k [\delta_f(i+1) - \delta_f(i)] \leq \sum_{i=0}^k \frac{L\sigma^2}{2\mu^2}$$

$$\delta_f(k+1) - \delta_f(0) \leq \frac{L\sigma^2(k+1)}{2\mu^2}$$

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

которое дает указанную скорость.

## Сходимость. Гладкий PL-случай.

3. Просуммируем предыдущее неравенство от  $i = 0$  до  $k$  и используем тот факт, что  $\delta_f(0) = 0$  мы получаем

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

$$\sum_{i=0}^k [\delta_f(i+1) - \delta_f(i)] \leq \sum_{i=0}^k \frac{L\sigma^2}{2\mu^2}$$

$$\delta_f(k+1) - \delta_f(0) \leq \frac{L\sigma^2(k+1)}{2\mu^2}$$

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

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

которое дает указанную скорость.

## Сходимость. Гладкий выпуклый случай (ограниченная дисперсия)

### Вспомогательные обозначения

Для (возможно) неконстантной последовательности шагов  $(\alpha_t)_{t \geq 0}$  определим *взвешенное среднее*

$$\bar{x}_k \stackrel{\text{def}}{=} \frac{1}{\sum_{t=0}^{k-1} \alpha_t} \sum_{t=0}^{k-1} \alpha_t x_t, \quad k \geq 1.$$

Везде ниже  $f^* \equiv \min_x f(x)$  и  $x^* \in \arg \min_x f(x)$ .

## Гладкий выпуклый случай с постоянным шагом

**i** Пусть  $f$  — выпуклая функция (не обязательно гладкая), а дисперсия стохастического градиента ограничена  $\mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2] \leq \sigma^2 \quad \forall k$ . Если SGD использует постоянный шаг  $\alpha_t \equiv \alpha > 0$ , 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  $\bar{x}_k = \frac{1}{k} \sum_{t=0}^{k-1} x_t$ .

При выборе постоянного  $\alpha = \frac{\|x_0 - x^*\|}{\sigma\sqrt{k}}$  (зависящего от  $k$ ) имеем

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

## Гладкий выпуклый случай с постоянным шагом

1. Начнём с разложения квадрата расстояния до минимума:

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

## Гладкий выпуклый случай с постоянным шагом

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. Берём условное матожидание по  $i_k$  (обозначим  $\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot|x_k]$ ), используем свойство  $\mathbb{E}_k[\nabla f_{i_k}(x_k)] = \nabla f(x_k)$ , 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 с постоянным шагом

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. Берём условное матожидание по  $i_k$  (обозначим  $\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot|x_k]$ ), используем свойство  $\mathbb{E}_k[\nabla f_{i_k}(x_k)] = \nabla f(x_k)$ , ограниченность дисперсии  $\mathbb{E}_k[\|\nabla f_{i_k}(x_k)\|^2] \leq \sigma^2$  и выпукłość  $f$  (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. Переносим слагаемое с  $f(x_k)$  влево и берём полное матожидание:

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

## Гладкий выпуклый случай с постоянным шагом

- Начнём с разложения квадрата расстояния до минимума:

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

- Берём условное матожидание по  $i_k$  (обозначим  $\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot|x_k]$ ), используем свойство  $\mathbb{E}_k[\nabla f_{i_k}(x_k)] = \nabla f(x_k)$ , ограниченность дисперсии  $\mathbb{E}_k[\|\nabla f_{i_k}(x_k)\|^2] \leq \sigma^2$  и выпукłość  $f$  (которая даёт  $\langle \nabla f(x_k), x_k - x^* \rangle \geq f(x_k) - f^*$ ):

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

- Переносим слагаемое с  $f(x_k)$  влево и берём полное матожидание:

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

- Суммируем (тескопирируем) по  $t = 0, \dots, k-1$ :

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

## Гладкий выпуклый случай с постоянным шагом

5. Делим на  $2\alpha k$ :

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

## Гладкий выпуклый случай с постоянным шагом

5. Делим 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. Используя выпуклость  $f$  и неравенство Йенсена для усреднённой точки  $\bar{x}_k = \frac{1}{k} \sum_{t=0}^{k-1} x_t$ :

$$\mathbb{E}[f(\bar{x}_k)] \leq \mathbb{E} \left[ \frac{1}{k} \sum_{t=0}^{k-1} f(x_t) \right] = \frac{1}{k} \sum_{t=0}^{k-1} \mathbb{E}[f(x_t)].$$

Вычитая  $f^*$  из обеих частей, получаем:

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

## Гладкий выпуклый случай с постоянным шагом

5. Делим 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. Используя выпуклость  $f$  и неравенство Йенсена для усреднённой точки  $\bar{x}_k = \frac{1}{k} \sum_{t=0}^{k-1} x_t$ :

$$\mathbb{E}[f(\bar{x}_k)] \leq \mathbb{E} \left[ \frac{1}{k} \sum_{t=0}^{k-1} f(x_t) \right] = \frac{1}{k} \sum_{t=0}^{k-1} \mathbb{E}[f(x_t)].$$

Вычитая  $f^*$  из обеих частей, получаем:

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

7. Объединяя (5) и (6), получаем искомую оценку:

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

## Гладкий выпуклый случай с убывающим шагом

$$\alpha_k = \frac{\alpha_0}{\sqrt{k+1}}, \quad 0 < \alpha_0 \leq \frac{1}{4L}$$

**i** При тех же предположениях, но с убывающим шагом  $\alpha_k = \frac{\alpha_0}{\sqrt{k+1}}$

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



## Мини-батч SGD

Подход 1: контролировать размер выборки

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, используя только 1 выборку:

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

Распространнёный вариант — использовать большую выборку  $B_k$  ("мини-батч"):

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

особенно полезно для векторизации и параллелизации.

Например, с 16 ядрами установите  $|B_k| = 16$  и вычислите 16 градиентов одновременно.

## Мини-батч как градиентный спуск с ошибкой

Метод SG с выборкой  $B_k$  (“мини-батч”) использует итерации:

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

Посмотрим на это как на “градиентный метод с ошибкой”:

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

где  $e_k$  — разница между аппроксимированным и истинным градиентом.

Если вы используете  $\alpha_k = \frac{1}{L}$ , то используя лемму о спуске, этот алгоритм имеет:

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

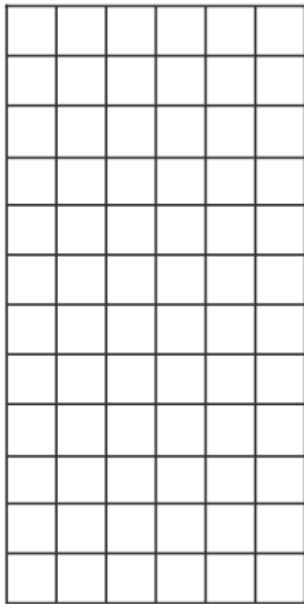
для любой ошибки  $e_k$ .

## Влияние ошибки на скорость сходимости

Оценка прогресса с  $\alpha_k = \frac{1}{L}$  и ошибкой в градиенте  $e_k$  выглядит следующим образом:

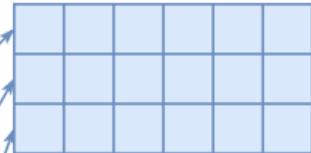
$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L} \|\nabla f(x_k)\|^2 + \frac{1}{2L} \|e_k\|^2.$$

### Данные



## Идея SGD и батчей

Данные



1 Батч

## Идея SGD и батчей



## Идея SGD и батчей



## Идея SGD и батчей



## Основная проблема SGD

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

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



## Основные результаты сходимости SGD

- Пусть  $f$  -  $L$ -гладкая  $\mu$ -сильно выпуклая функция, а дисперсия стохастического градиента конечна ( $\mathbb{E}[\|\nabla f_i(x_k)\|^2] \leq \sigma^2$ ). Тогда траектория стохастического градиентного спуска с постоянным шагом  $\alpha < \frac{1}{2\mu}$  будет гарантировать:

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

## Основные результаты сходимости SGD

- Пусть  $f$  -  $L$ -гладкая  $\mu$ -сильно выпуклая функция, а дисперсия стохастического градиента конечна ( $\mathbb{E}[\|\nabla f_i(x_k)\|^2] \leq \sigma^2$ ). Тогда траектория стохастического градиентного спуска с постоянным шагом  $\alpha < \frac{1}{2\mu}$  будет гарантировать:

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

- Пусть  $f$  -  $L$ -гладкая  $\mu$ -сильно выпуклая функция, а дисперсия стохастического градиента конечна ( $\mathbb{E}[\|\nabla f_i(x_k)\|^2] \leq \sigma^2$ ). Тогда стохастический градиентный шум с уменьшающимся шагом  $\alpha_k = \frac{2k+1}{2\mu(k+1)^2}$  будет сходиться сублинейно:

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

## Summary

- SGD с постоянным шагом не сходится даже для PL (сильно выпуклого) случая

## Summary

- SGD с постоянным шагом не сходится даже для PL (сильно выпуклого) случая
- SGD достигает сублинейной сходимости с скоростью  $\mathcal{O}\left(\frac{1}{k}\right)$  для PL-случая.

## Summary

- SGD с постоянным шагом не сходится даже для PL (сильно выпуклого) случая
- SGD достигает сублинейной сходимости с скоростью  $\mathcal{O}\left(\frac{1}{k}\right)$  для PL-случая.
- Ускорения Нестерова/Поляка не улучшают скорость сходимости

## Summary

- SGD с постоянным шагом не сходится даже для PL (сильно выпуклого) случая
- SGD достигает сублинейной сходимости с скоростью  $\mathcal{O}\left(\frac{1}{k}\right)$  для PL-случая.
- Ускорения Нестерова/Поляка не улучшают скорость сходимости
- Двухфазный Ньютоновский метод достигает  $\mathcal{O}\left(\frac{1}{k}\right)$  без сильной выпуклости.

## Стохастическая оптимизация

# Много методов

Rosenbrock Function.  
Adaptive stochastic gradient algorithms.  
Learning rate 0.003



SGD расходится с любым шагом обучения для LLS

# Оптимизация для глубокого обучения с практической точки зрения

# Как их сравнить? AlgoPerf benchmark<sup>1 2</sup>

- **AlgoPerf Benchmark:** Сравнивает алгоритмы обучения нейросетей по двум режимам:

# Как их сравнить? AlgoPerf benchmark <sup>1 2</sup>

- **AlgoPerf Benchmark:** Сравнивает алгоритмы обучения нейросетей по двум режимам:
  - **Внешняя настройка (External Tuning):** моделирует тюнинг гиперпараметров при ограниченных ресурсах (5 запусков, квазислучайный поиск). Оценка — медианное минимальное время достижения цели по 5 наборам задач.

# Как их сравнить? AlgoPerf benchmark 1 2

- **AlgoPerf Benchmark:** Сравнивает алгоритмы обучения нейросетей по двум режимам:
  - **Внешняя настройка (External Tuning):** моделирует тюнинг гиперпараметров при ограниченных ресурсах (5 запусков, квазислучайный поиск). Оценка — медианное минимальное время достижения цели по 5 наборам задач.
  - **Самонастройка (Self-Tuning):** моделирует автоматический тюнинг на одной машине (фиксированный/внутрипетлевой тюнинг, бюджет  $\times 3$ ). Оценка — медианное время выполнения по 5 наборам задач.

# Как их сравнить? AlgoPerf benchmark 1 2

- **AlgoPerf Benchmark:** Сравнивает алгоритмы обучения нейросетей по двум режимам:
  - **Внешняя настройка (External Tuning):** моделирует тюнинг гиперпараметров при ограниченных ресурсах (5 запусков, квазислучайный поиск). Оценка — медианное минимальное время достижения цели по 5 наборам задач.
  - **Самонастройка (Self-Tuning):** моделирует автоматический тюнинг на одной машине (фиксированный/внутрипетлевой тюнинг, бюджет  $\times 3$ ). Оценка — медианное время выполнения по 5 наборам задач.
- **Оценка:** результаты агрегируются с помощью профилей производительности. Профили показывают долю задач, решённых за время, не превышающее фактор  $\tau$  относительно самой быстрой посылки. Итоговый балл — нормированная площадь под кривой профиля ( $1.0 =$  самая быстрая на всех задачах).

# Как их сравнить? AlgoPerf benchmark<sup>1 2</sup>

- **AlgoPerf Benchmark:** Сравнивает алгоритмы обучения нейросетей по двум режимам:
  - **Внешняя настройка (External Tuning):** моделирует тюнинг гиперпараметров при ограниченных ресурсах (5 запусков, квазислучайный поиск). Оценка — медианное минимальное время достижения цели по 5 наборам задач.
  - **Самонастройка (Self-Tuning):** моделирует автоматический тюнинг на одной машине (фиксированный/внутрипетлевой тюнинг, бюджет  $\times 3$ ). Оценка — медианное время выполнения по 5 наборам задач.
- **Оценка:** результаты агрегируются с помощью профилей производительности. Профили показывают долю задач, решённых за время, не превышающее фактор  $\tau$  относительно самой быстрой посылки. Итоговый балл — нормированная площадь под кривой профиля ( $1.0 =$  самая быстрая на всех задачах).
- **Вычислительная стоимость:** оценка требует  $\sim 49,240$  суммарных часов на 8x NVIDIA V100 GPUs (в среднем  $\sim 3469$ ч/внешняя настройка,  $\sim 1847$ ч/самонастройка).

<sup>1</sup>Benchmarking Neural Network Training Algorithms

<sup>2</sup>Accelerating neural network training: An analysis of the AlgoPerf competition

## AlgoPerf benchmark

**Summary фиксированных базовых задач в AlgoPerf benchmark.** Функции потерь включают кросс-энтропию (CE), среднюю абсолютную ошибку (L1) и CTC-потерю (Connectionist Temporal Classification, CTC).

Дополнительные метрики оценки: индекс структурного сходства (SSIM), коэффициент ошибок (ER) и доля ошибок по словам (WER), средняя усреднённая точность (mAP) и метрика BLEU (bilingual evaluation understudy). Бюджет времени выполнения соответствует набору правил внешней настройки; набор правил самонастройки допускает обучение, в 3 раза более длительное.

| Задача                        | Датасет     | Модель      | Потери | Метрика | Целевое значение на валидации | Бюджет времени |
|-------------------------------|-------------|-------------|--------|---------|-------------------------------|----------------|
| Clickthrough rate prediction  | CRITEO 1TB  | DLRMSMALL   | CE     | CE      | 0.123735                      | 7703           |
| MRI reconstruction            | FASTMRI     | U-NET       | L1     | SSIM    | 0.7344                        | 8859           |
| Image classification          | IMAGENET    | ResNet-50   | CE     | ER      | 0.22569                       | 63,008         |
|                               |             | ViT         | CE     | ER      | 0.22691                       | 77,520         |
| Speech recognition            | LIBRISPEECH | Conformer   | CTC    | WER     | 0.085884                      | 61,068         |
|                               |             | DeepSpeech  | CTC    | WER     | 0.119936                      | 55,506         |
| Molecular property prediction | OGBG        | GNN         | CE     | mAP     | 0.28098                       | 18,477         |
| Translation                   | WMT         | Transformer | CE     | BLEU    | 30.8491                       | 48,151         |

# AlgoPerf benchmark

| Submission                  | Line | Score  |
|-----------------------------|------|--------|
| PYTORCH DISTRIBUTED SHAMPOO |      | 0.7784 |
| SCHEDULE FREE ADAMW         |      | 0.7077 |
| GENERALIZED ADAM            |      | 0.6383 |
| CYCLIC LR                   |      | 0.6301 |
| NADAMP                      |      | 0.5909 |
| BASELINE                    |      | 0.5707 |
| AMOS                        |      | 0.4918 |
| CASPR ADAPTIVE              |      | 0.4722 |
| LAWA QUEUE                  |      | 0.3699 |
| LAWA EMA                    |      | 0.3384 |
| SCHEDULE FREE PRODIGY       |      | 0      |

(a) External tuning leaderboard



(b) External tuning performance profiles

# AlgoPerf benchmark

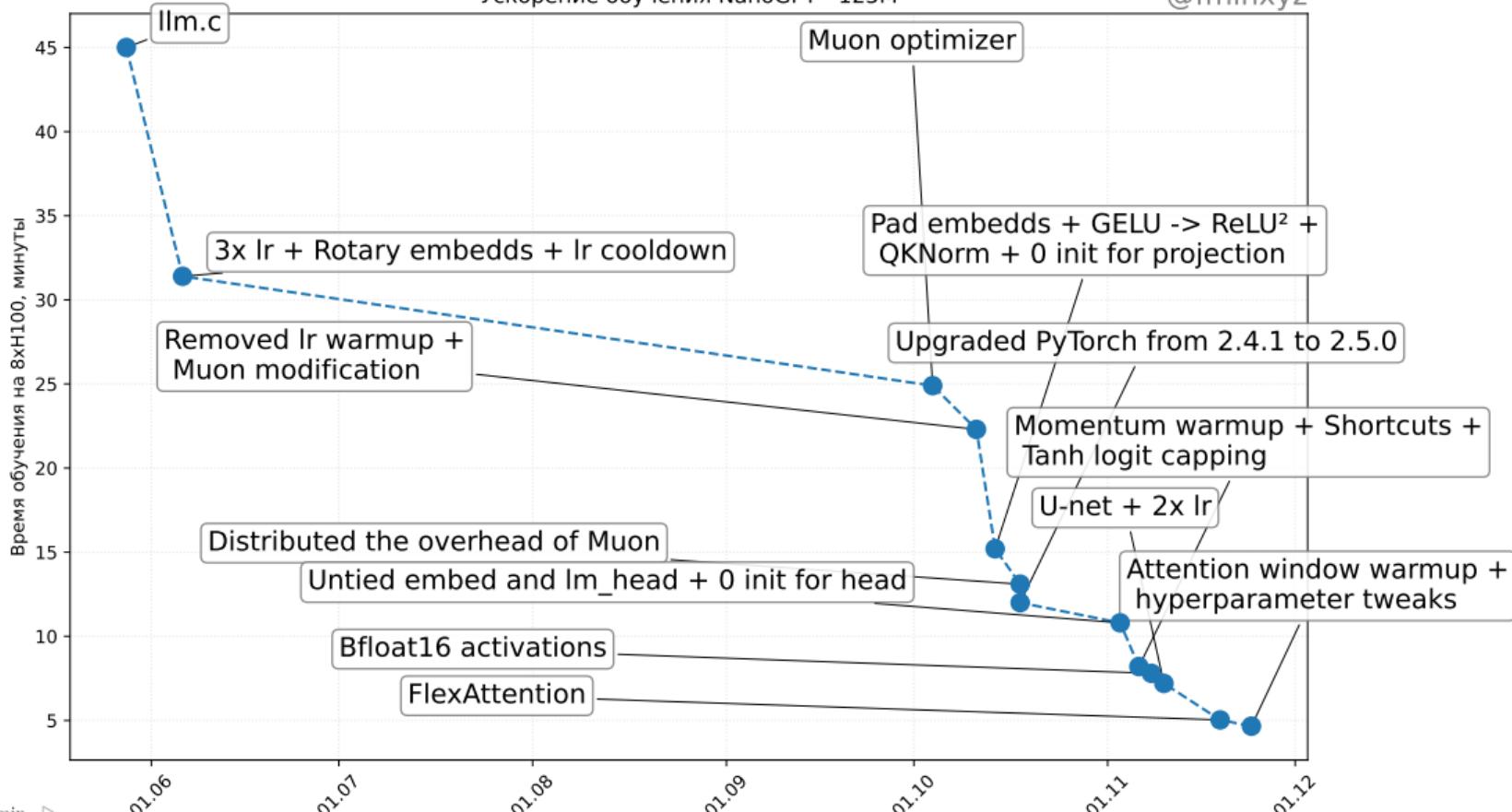


- PyTorch Distr. Shampoo
- Schedule Free AdamW
- Generalized Adam
- Cyclic LR
- NadamP
- Baseline
- Amos
- CASPR Adaptive
- Lawa Queue
- Lawa EMA
- Schedule Free Prodigy

# NanoGPT speedrun

Ускорение обучения NanoGPT - 125M

@fminxyz



## Работают ли трюки, если увеличить размер модели?

### Scaling up the NanoGPT (124M) speedrun



Рис. 3: Источник

## Работают ли трюки, если увеличить размер модели?



Рис. 4: Источник

## Неожиданные истории

## Adam работает хуже для CV, чем для LLM? <sup>3</sup>

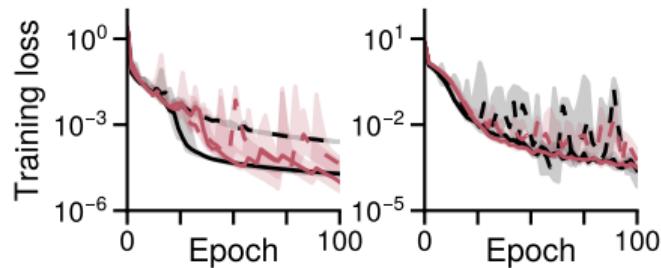


Рис. 5: CNNs on MNIST and CIFAR10



Рис. 6: Transformers on PTB, WikiText2, and SQuAD

Черные линии - SGD; красные линии - Adam.

<sup>3</sup>Linear attention is (maybe) all you need (to understand transformer optimization)



<sup>4</sup>Linear attention is (maybe) all you need (to understand transformer optimization)

## Почему Adam работает хуже для CV, чем для LLM? <sup>5</sup>

Нет! Метки имеют тяжелые хвосты!

В компьютерном зрении датасеты часто сбалансированы: 1000 котиков, 1000 песелей и т.д.  
В языковых датасетах почти всегда не так: слово *the* встречается часто, слово *tie* на порядки реже

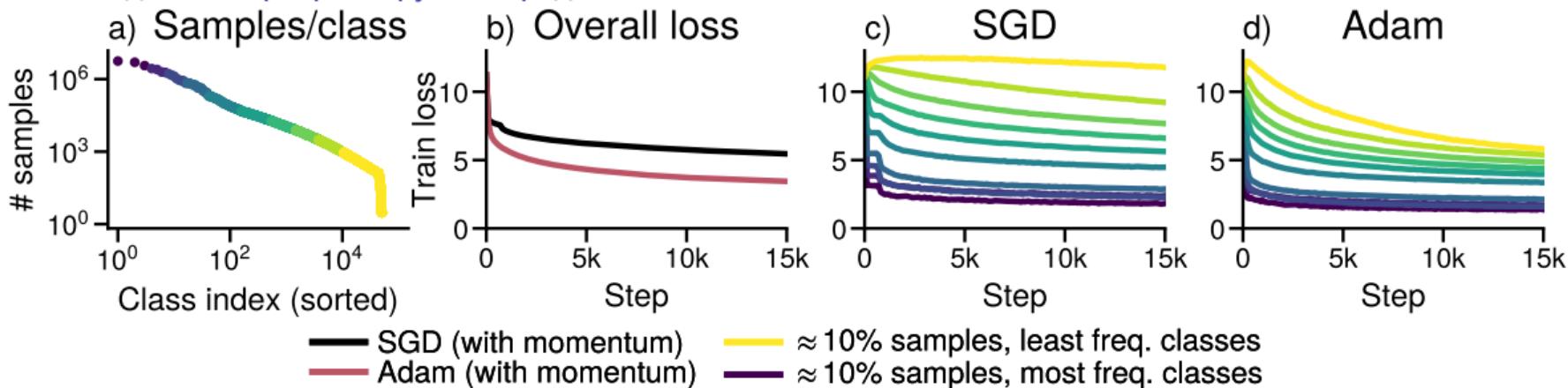


Рис. 7: Распределение частоты токенов в PTB

<sup>5</sup>Heavy-Tailed Class Imbalance and Why Adam Outperforms Gradient Descent on Language Models

# Почему Adam работает хуже для CV, чем для LLM? <sup>6</sup>

SGD медленно прогрессирует на редких классах



SGD не добивается прогресса на низкочастотных классах, в то время как Adam добивается. Обучение GPT-2 S на WikiText-103. (a) Распределение классов, отсортированных по частоте встречаемости, разбитых на группы, соответствующие  $\approx 10\%$  данных. (b) Значение функции потерь при обучении. (c, d) Значение функции потерь при обучении для каждой группы при использовании SGD и Adam.

<sup>6</sup>Heavy-Tailed Class Imbalance and Why Adam Outperforms Gradient Descent on Language Models

## Влияние инициализации<sup>7</sup>

- 💡 Правильная инициализация нейронной сети важна. Функция потерь нейронной сети является высоко невыпуклой; оптимизировать её для достижения «хорошего» решения трудно, требует тщательной настройки.
- Не инициализируйте все веса одинаково — почему?

## Влияние инициализации<sup>7</sup>



Правильная инициализация нейронной сети важна. Функция потерь нейронной сети является высоко невыпуклой; оптимизировать её для достижения «хорошего» решения трудно, требует тщательной настройки.

- Не инициализируйте все веса одинаково — почему?
- Случайная инициализация: задавайте случайно, например, из гауссовского распределения  $N(0, \sigma^2)$ , где стандартное отклонение  $\sigma$  зависит от числа нейронов в слое. Это обеспечивает нарушение симметрии. *Symmetry breaking*.

## Влияние инициализации<sup>7</sup>



Правильная инициализация нейронной сети важна. Функция потерь нейронной сети является высоко невыпуклой; оптимизировать её для достижения «хорошего» решения трудно, требует тщательной настройки.

- Не инициализируйте все веса одинаково — почему?
- Случайная инициализация: задавайте случайно, например, из гауссовского распределения  $N(0, \sigma^2)$ , где стандартное отклонение  $\sigma$  зависит от числа нейронов в слое. Это обеспечивает нарушение симметрии. *Symmetry breaking*.
- Можно найти более полезные советы здесь

<sup>7</sup>On the importance of initialization and momentum in deep learning Ilya Sutskever, James Martens, George Dahl, Geoffrey Hinton

# Влияние инициализации весов нейронной сети на сходимость методов<sup>8</sup>



Рис. 8: 22-layer ReLU net: good init converges faster

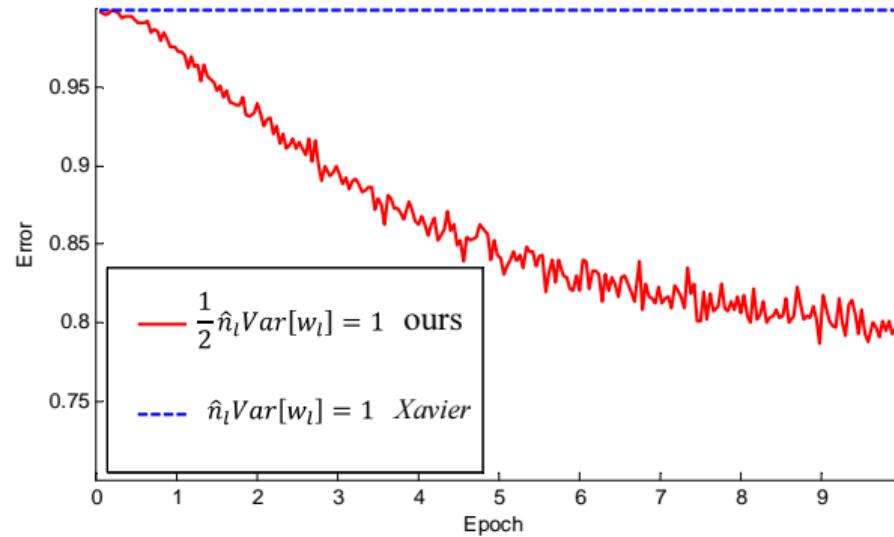


Рис. 9: 30-layer ReLU net: good init is able to converge

<sup>8</sup>Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification, Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun

## Весёлые истории

Градиентный спуск сходится к локальному минимуму



# Градиентный спуск сходится к локальному минимуму



Стохастический градиентный спуск  
выпрыгивает из локальных минимумов



## Визуализация с помощью проекции на прямую

- Обозначим начальную точку как  $w_0$ , представляющую собой веса нейронной сети при инициализации. Веса, полученные после обучения, обозначим как  $\hat{w}$ .

$$L(\alpha) = L(w_0 + \alpha w_1), \text{ where } \alpha \in [-b, b].$$

## Визуализация с помощью проекции на прямую

- Обозначим начальную точку как  $w_0$ , представляющую собой веса нейронной сети при инициализации. Веса, полученные после обучения, обозначим как  $\hat{w}$ .
- Генерируем случайный вектор такой же размерности и нормы  $w_1 \in \mathbb{R}^p$ , затем вычисляем значение функции потерь вдоль этого вектора:

$$L(\alpha) = L(w_0 + \alpha w_1), \text{ where } \alpha \in [-b, b].$$

# Проекция функции потерь нейронной сети на прямую

No Dropout

Loss surface, Line projection around the starting point



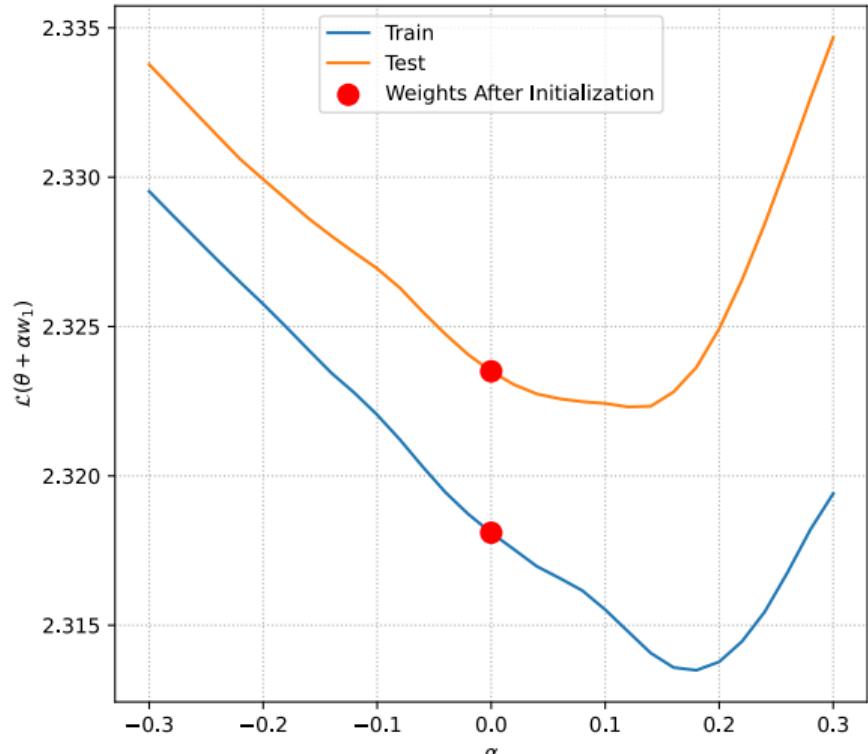
Loss surface, Line projection around the final point



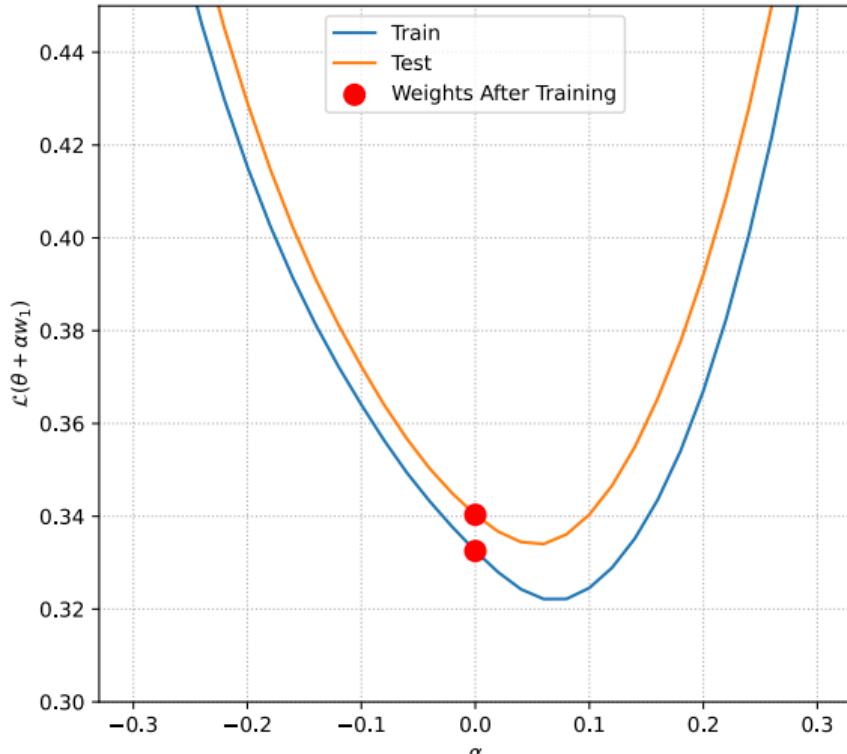
# Проекция функции потерь нейронной сети на прямую

Dropout 0.2

Loss surface, Line projection around the starting point



Loss surface, Line projection around the final point



# Проекция функции потерь нейронной сети на плоскость

- Мы можем расширить эту идею и построить проекцию поверхности потерь на плоскость, которая задается 2 случайными векторами.

$$L(\alpha, \beta) = L(w_0 + \alpha w_1 + \beta w_2), \text{ where } \alpha, \beta \in [-b, b]^2.$$

No Dropout. Plane projection of loss surface.



# Проекция функции потерь нейронной сети на плоскость

- Мы можем расширить эту идею и построить проекцию поверхности потерь на плоскость, которая задается 2 случайными векторами.
- Два случайных гауссовых вектора в пространстве большой размерности с высокой вероятностью ортогональны.

$$L(\alpha, \beta) = L(w_0 + \alpha w_1 + \beta w_2), \text{ where } \alpha, \beta \in [-b, b]^2.$$

No Dropout. Plane projection of loss surface.



Может ли быть полезно изучение таких проекций? <sup>9</sup>



Рис. 13: The loss surface of ResNet-56  
without skip connections



Рис. 14: The loss surface of ResNet-56 with skip connections

<sup>9</sup>Visualizing the Loss Landscape of Neural Nets, Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, Tom Goldstein

Может ли быть полезно изучение таких проекций, если серьезно? <sup>10</sup>



Рис. 15: Examples of a loss landscape of a typical CNN model on FashionMNIST and CIFAR10 datasets found with MPO. Loss values are color-coded according to a logarithmic scale

<sup>10</sup>Loss Landscape Sightseeing with Multi-Point Optimization, Ivan Skorokhodov, Mikhail Burtsev  
 $f \rightarrow \min_{x,y,z}$

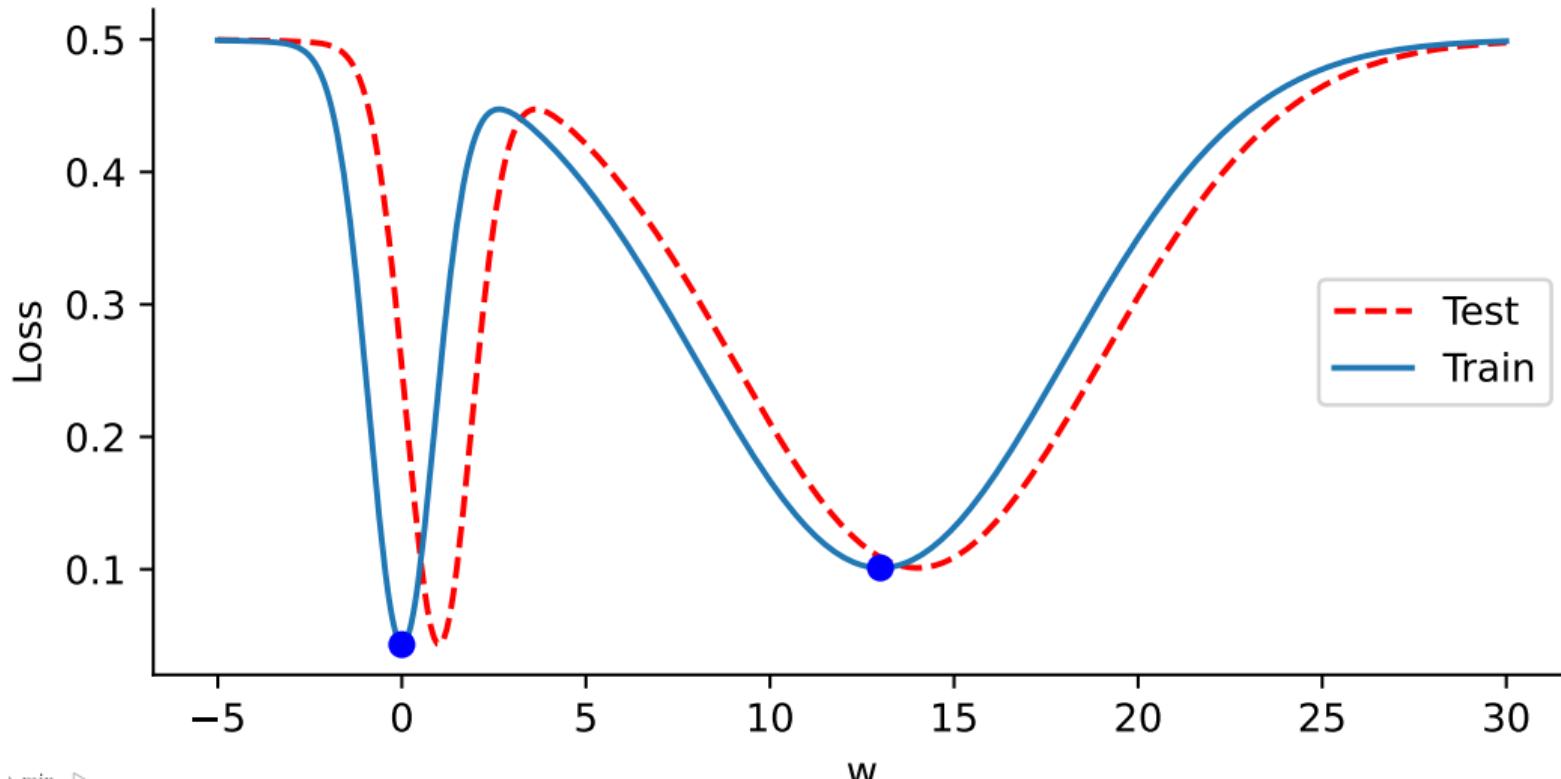
## Ширина локальных минимумов

Узкие и широкие локальные минимумы



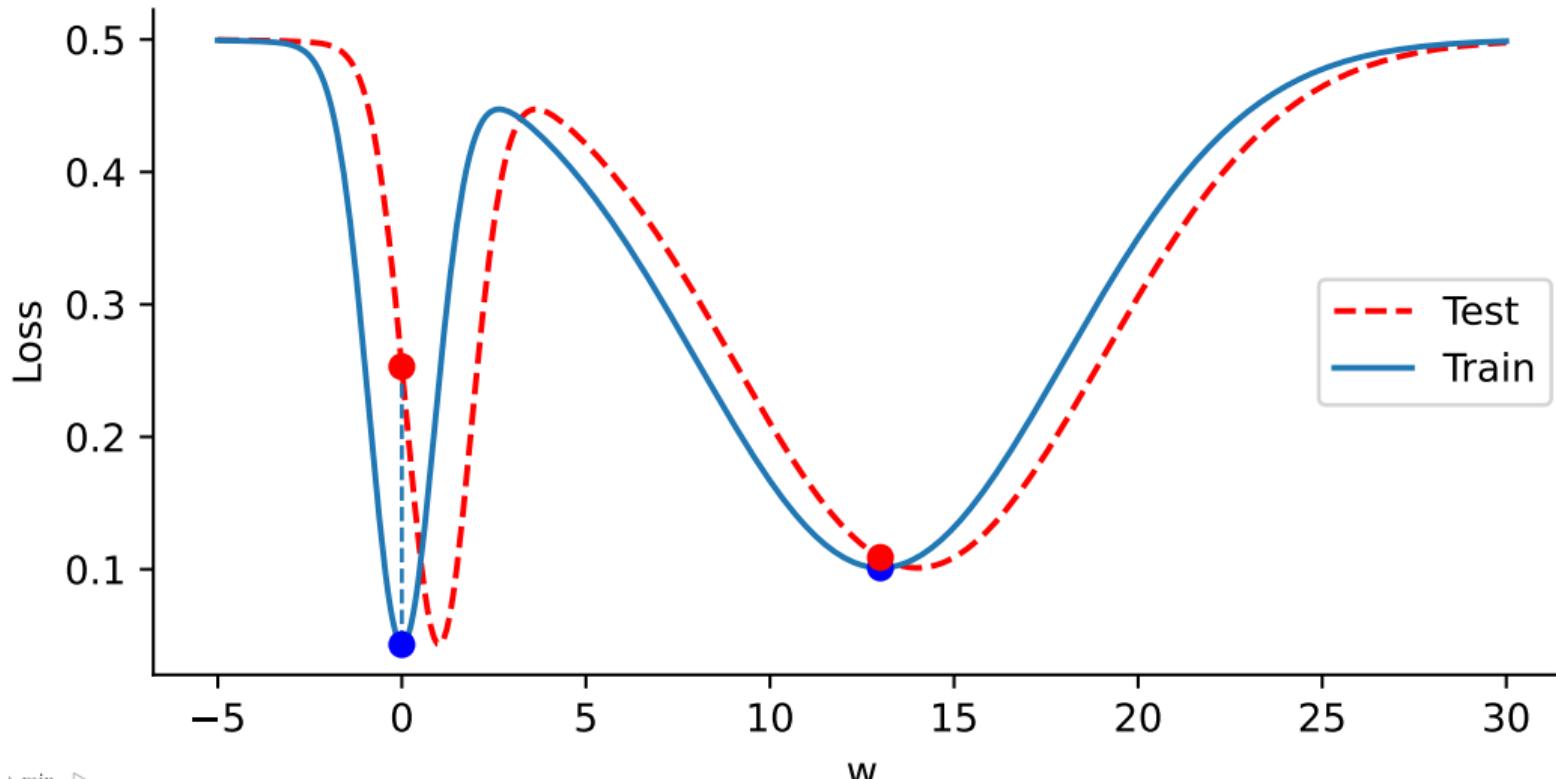
## Ширина локальных минимумов

Узкие и широкие локальные минимумы



## Ширина локальных минимумов

Узкие и широкие локальные минимумы



# Экспоненциальный шаг обучения

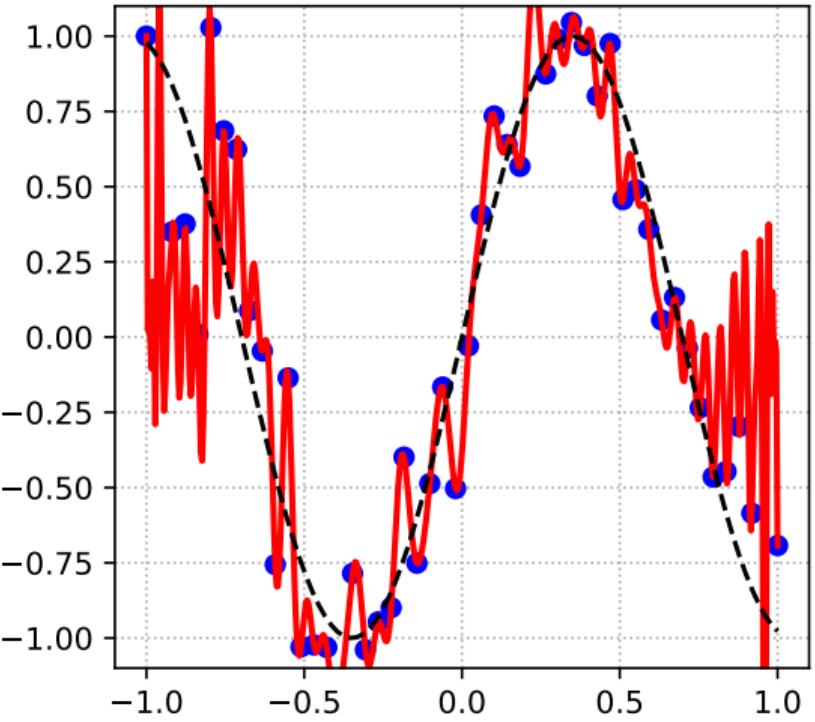
- Exponential Learning Rate Schedules for Deep Learning

## Double Descent<sup>11</sup>

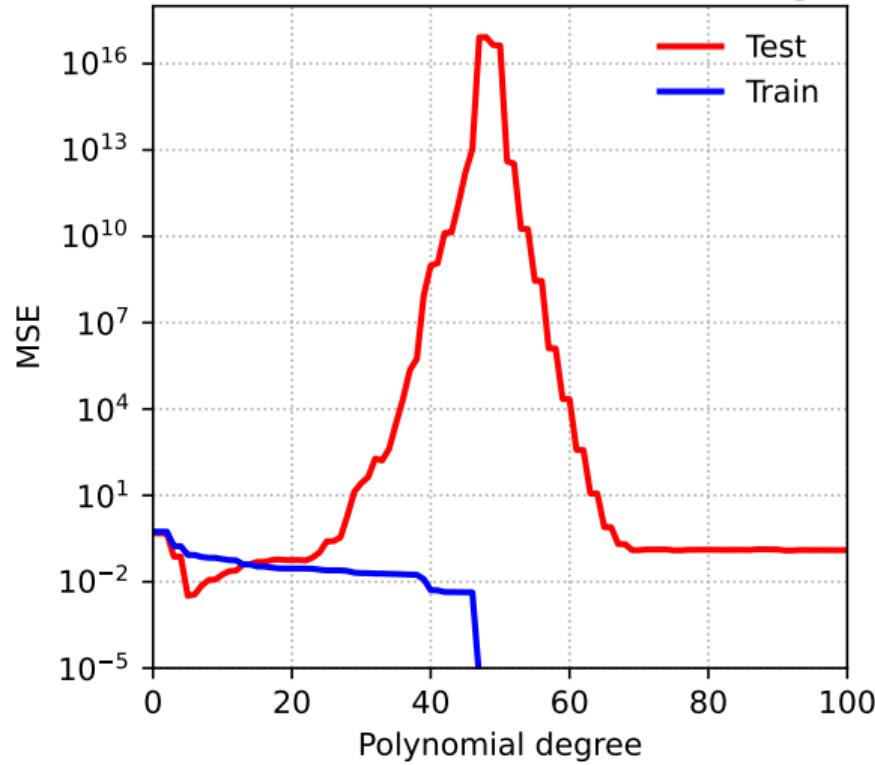
<sup>11</sup>Reconciling modern machine learning practice and the bias-variance trade-off, Mikhail Belkin, Daniel Hsu, Siyuan Ma, Soumik Mandal

## Double Descent

Polynomial Fitting



@fminxyz



Modular Division (training on 50% of data)



Рис. 16: Training transformer with 2 layers, width 128, and 4 attention heads, with a total of about  $4 \cdot 10^5$  non-embedding parameters. Reproduction of experiments (~ half an hour) is available here

- Рекомендую посмотреть лекцию Дмитрия Ветрова **Удивительные свойства функции потерь в нейронной сети** (Surprising properties of loss landscape in overparameterized models). видео, Презентация

# Grokking<sup>12</sup>

Modular Division (training on 50% of data)



Рис. 16: Training transformer with 2 layers, width 128, and 4 attention heads, with a total of about  $4 \cdot 10^5$  non-embedding parameters. Reproduction of experiments ( $\sim$  half an hour) is available here

- Рекомендую посмотреть лекцию Дмитрия Ветрова **Удивительные свойства функции потерь в нейронной сети** (Surprising properties of loss landscape in overparameterized models). видео, Презентация
- Автор канала Свидетели Градиента собирает интересные наблюдения и эксперименты про гроккинг.

Modular Division (training on 50% of data)



Рис. 16: Training transformer with 2 layers, width 128, and 4 attention heads, with a total of about  $4 \cdot 10^5$  non-embedding parameters. Reproduction of experiments ( $\sim$  half an hour) is available here

- Рекомендую посмотреть лекцию Дмитрия Ветрова **Удивительные свойства функции потерь в нейронной сети** (Surprising properties of loss landscape in overparameterized models). видео, Презентация
- Автор канала Свидетели Градиента собирает интересные наблюдения и эксперименты про гроккинг.
- Также есть видео с его докладом **Чем не является гроккинг**.

<sup>12</sup>Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets, Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, Vedant Misra