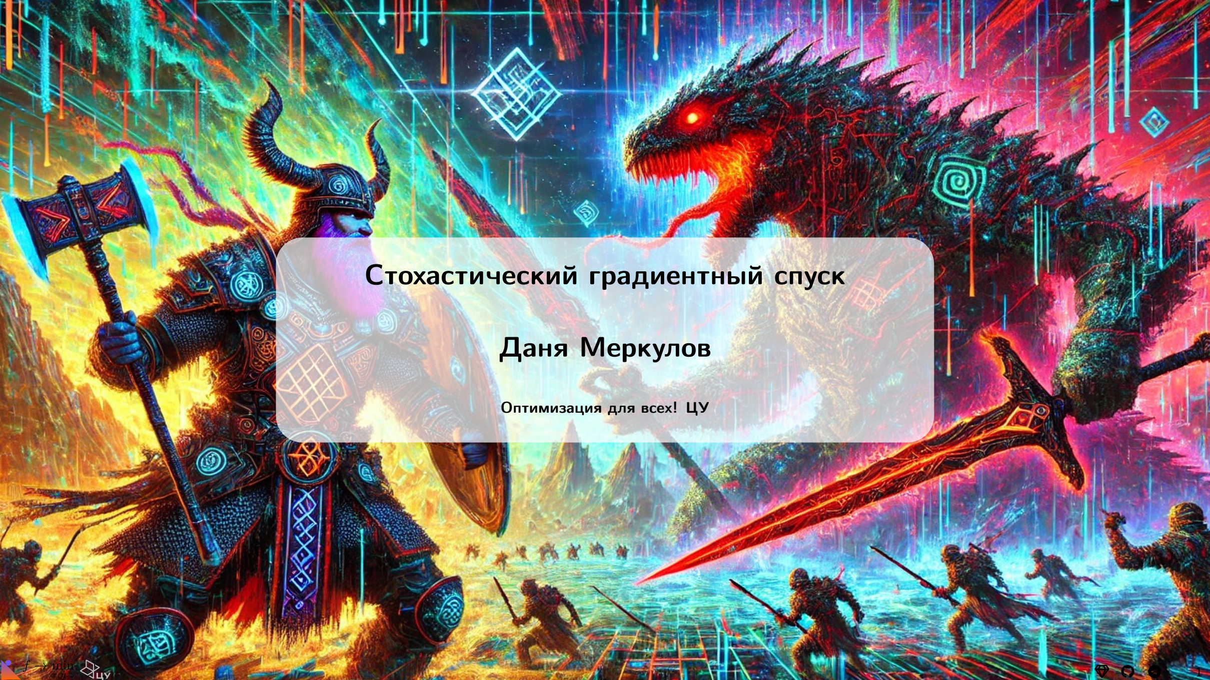


Стохастический градиентный спуск

Даня Меркулов

Оптимизация для всех! ЦУ

Задача с конечной суммой

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:

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

- Сходимость с постоянным α 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:

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

- Сходимость с постоянным α 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:

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

- Сходимость с постоянным α 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:

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

- Сходимость с постоянным α 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:

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

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 ∇f является липшицевым, то мы получаем:

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

Результаты для градиентного спуска

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

Если ∇f является липшицевым, то мы получаем:

Предположение	Детерминированный градиентный спуск	Стохастический градиентный спуск
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 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 ∇f является липшицевым, то мы получаем:

Предположение	Детерминированный градиентный спуск	Стохастический градиентный спуск
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 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 ∇f является липшицевым, то мы получаем:

Предположение	Детерминированный градиентный спуск	Стохастический градиентный спуск
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 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 ∇f является липшицевым, то мы получаем:

Предположение	Детерминированный градиентный спуск	Стохастический градиентный спуск
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 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 несмещенную аппроксимацию градиента с ограниченной дисперсией.

Результаты для градиентного спуска

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

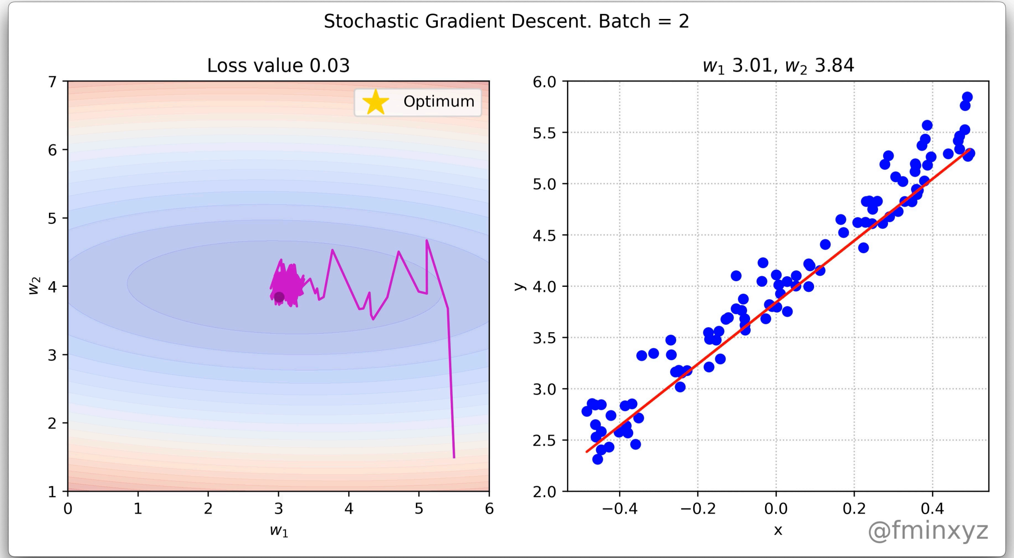
Если ∇f является липшицевым, то мы получаем:

Предположение	Детерминированный градиентный спуск	Стохастический градиентный спуск
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 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 несмещенную 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).

Стохастический градиентный спуск (SGD)

Типичное поведение



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 несмещенную оценку градиента: $\mathbb{E}[\nabla f_{i_k}(x_k)] = \nabla f(x_k)$:

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

Гладкий PL-случай с постоянным шагом

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

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):

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

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

Гладкий PL-случай с постоянным шагом

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):

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

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

Гладкий PL-случай с постоянным шагом

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):

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

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

Вычтем f^*

Гладкий PL-случай с постоянным шагом

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):

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

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

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

Гладкий PL-случай с постоянным шагом

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):

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

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

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

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

Гладкий PL-случай с постоянным шагом

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):

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

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

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

Вычтем f^*

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

Переставляем

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

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

Гладкий PL-случай с постоянным шагом

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):

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

Сходимость. Гладкий PL-случай.

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.

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.

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

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

Сходимость. Гладкий PL-случай.

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

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

$$\overset{(2k+1)^2<(2k+2)^2=4(k+1)^2}{\leq} \frac{k^2}{(k+1)^2}[f(x_k) - f^*] + \frac{L\sigma^2}{2\mu^2(k+1)^2}$$

Сходимость. Гладкий PL-случай.

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

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

$$\underset{(2k+1)^2<(2k+2)^2=4(k+1)^2}{\leq} \frac{k^2}{(k+1)^2}[f(x_k) - f^*] + \frac{L\sigma^2}{2\mu^2(k+1)^2}$$

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

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

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

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

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

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

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

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

$$
\begin{aligned}
\delta_f(i+1) &\leq \delta_f(i) + \frac{L\sigma^2}{2\mu^2} \\
\sum_{i=0}^{k} \left[\delta_f(i+1) - \delta_f(i)\right] &\leq \sum_{i=0}^{k} \frac{L\sigma^2}{2\mu^2} \\
\delta_f(k+1) - \delta_f(0) &\leq \frac{L\sigma^2(k+1)}{2\mu^2} \\
(k+1)^2 \mathbb{E}[f(x_{k+1}) - f^*] &\leq \frac{L\sigma^2(k+1)}{2\mu^2}
\end{aligned}
$$

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

$$
\begin{aligned}
\delta_f(i+1) &\leq \delta_f(i) + \frac{L\sigma^2}{2\mu^2} \\
\sum_{i=0}^{k} \left[\delta_f(i+1) - \delta_f(i)\right] &\leq \sum_{i=0}^{k} \frac{L\sigma^2}{2\mu^2} \\
\delta_f(k+1) - \delta_f(0) &\leq \frac{L\sigma^2(k+1)}{2\mu^2} \\
(k+1)^2 \mathbb{E}[f(x_{k+1}) - f^*] &\leq \frac{L\sigma^2(k+1)}{2\mu^2} \\
\mathbb{E}[f(x_k) - f^*] &\leq \frac{L\sigma^2}{2\mu^2 k}
\end{aligned}
$$

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*

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

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

Гладкий выпуклый случай с постоянным шагом

Пусть f — выпуклая функция (не обязательно гладкая), а дисперсия стохастического градиента ограничена $\mathbb{E}\left[\|\nabla f_{i_k}(x_k)\|^2\right] \le \sigma^2 \quad \forall k$. Если SGD использует постоянный шаг $\alpha_t \equiv \alpha > 0$, то для любого $k \ge 1$

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

где $\bar{x}_k = \frac{1}{k}\sum_{t=0}^{k-1} x_t$.

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

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

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

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

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

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

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

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

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

4. Суммируем (телескопируем) по $t = 0, \ldots, k-1$:

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

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

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

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

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

Мини-батч SGD

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.

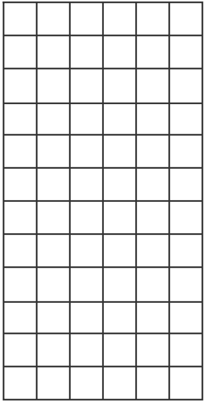
Влияние ошибки на скорость сходимости

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

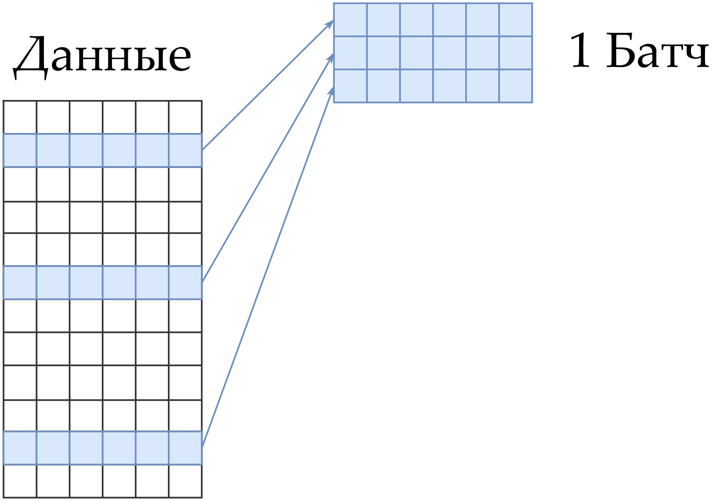
$$
f(x_{k+1}) \leq f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2 + \frac{1}{2L}\|e_k\|^2.
$$

Идея SGD и батчей

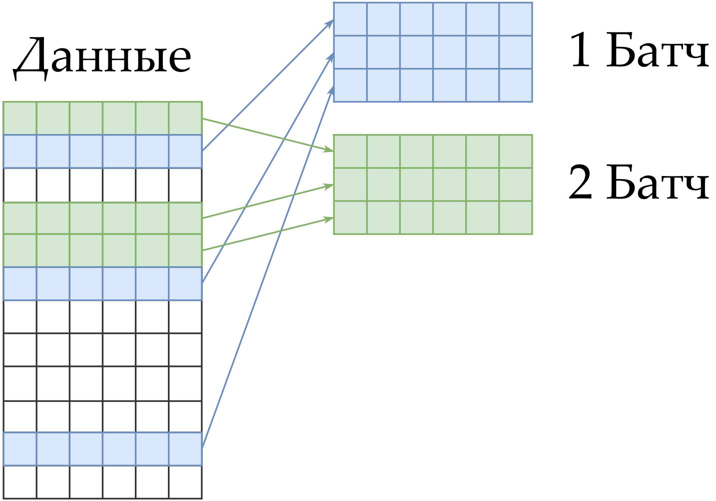
Данные



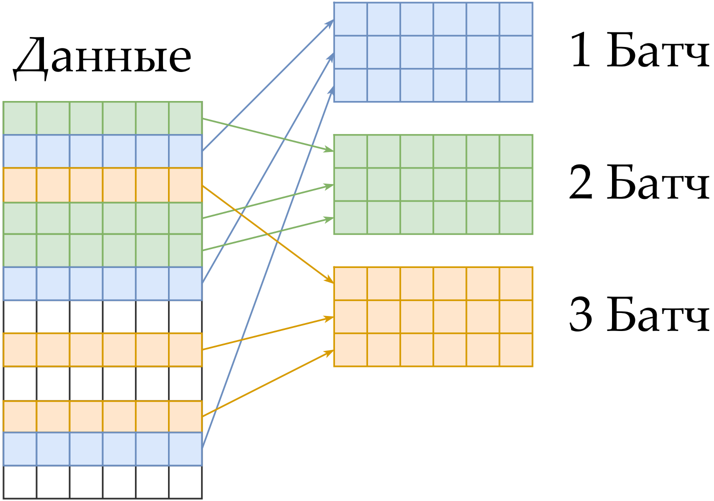
Идея SGD и батчей



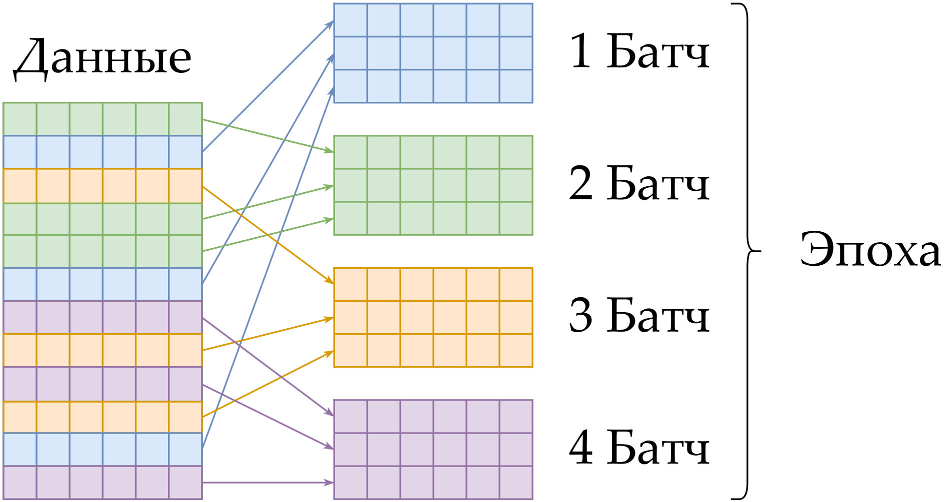
Идея SGD и батчей



Идея SGD и батчей

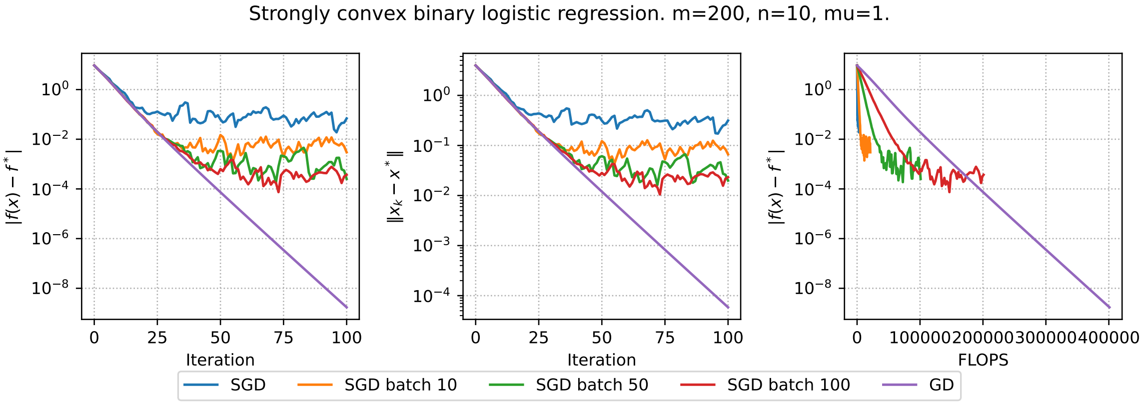


Идея SGD и батчей



Основная проблема SGD

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



Основные результаты сходимости SGD

Пусть f - L-гладкая μ-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 μ-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 μ-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:

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

Заключения

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

Заключения

* 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.

Заключения

- 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

Заключения

- 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.