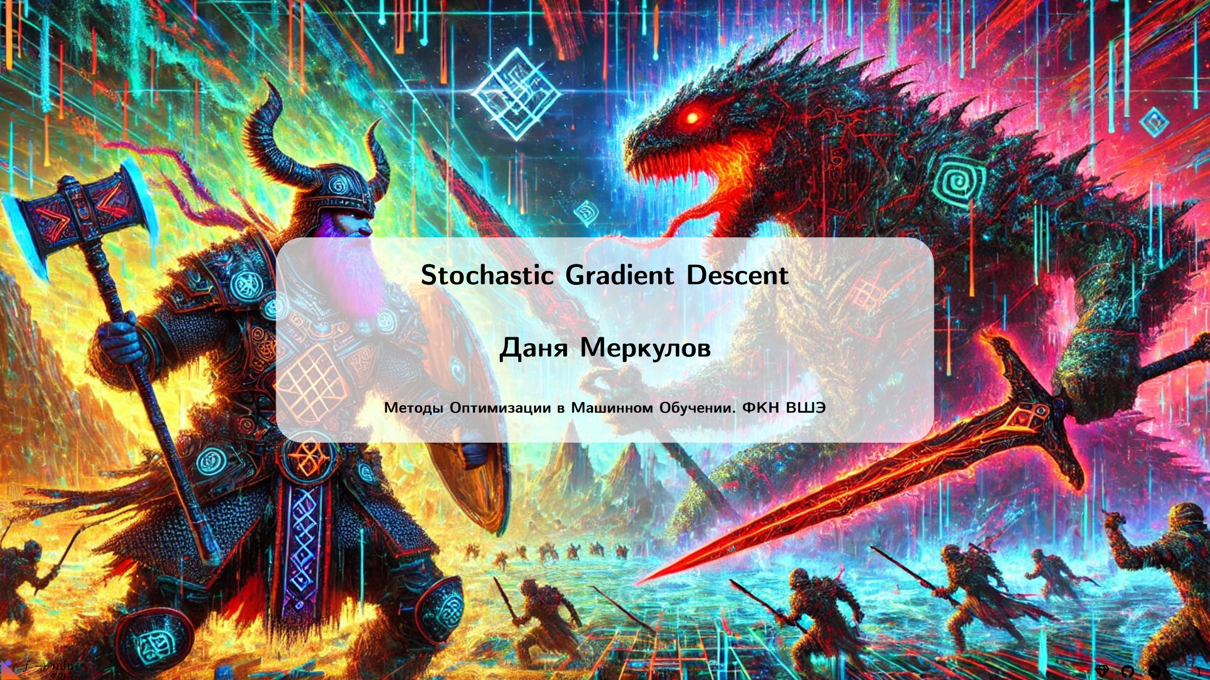


Stochastic Gradient Descent

Даня Меркулов

Методы Оптимизации в Машинном Обучении. ФКН ВШЭ

Finite-sum problem

Finite-sum problem

We consider classic finite-sample average minimization:

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

The gradient descent acts like follows:

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

- Convergence with constant α or line search.

Finite-sum problem

We consider classic finite-sample average minimization:

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

The gradient descent acts like follows:

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

- Convergence with constant α or line search.
- Iteration cost is linear in n. For ImageNet $n \approx 1.4 \cdot 10^7$, for WikiText $n \approx 10^8$. For FineWeb $n \approx 15 \cdot 10^{12}$ tokens.

Finite-sum problem

We consider classic finite-sample average minimization:

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

The gradient descent acts like follows:

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

- Convergence with constant α or line search.
- Iteration cost is linear in n. For ImageNet $n \approx 1.4 \cdot 10^7$, for WikiText $n \approx 10^8$. For FineWeb $n \approx 15 \cdot 10^{12}$ tokens.

Finite-sum problem

We consider classic finite-sample average minimization:

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

The gradient descent acts like follows:

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

- Convergence with constant α or line search.
- Iteration cost is linear in n. For ImageNet $n \approx 1.4 \cdot 10^7$, for WikiText $n \approx 10^8$. For FineWeb $n \approx 15 \cdot 10^{12}$ tokens.

Let's switch from the full gradient calculation to its unbiased estimator, when we randomly choose i_k index of point at each iteration uniformly:

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

With $p(i_k = i) = \frac{1}{n}$, the stochastic gradient is an unbiased estimate of the gradient, given by:

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

This indicates that the expected value of the stochastic gradient is equal to the actual gradient of $f(x)$.

Results for Gradient Descent

Stochastic iterations are n times faster, but how many iterations are needed?

If ∇f is Lipschitz continuous then we have:

Assumption	Deterministic Gradient Descent	Stochastic Gradient Descent
PL	$\mathcal{O}\left(\log(1/\varepsilon)\right)$	
Convex	$\mathcal{O}\left(1/\varepsilon\right)$	
Non-Convex	$\mathcal{O}\left(1/\varepsilon\right)$	

Results for Gradient Descent

Stochastic iterations are n times faster, but how many iterations are needed?

If ∇f is Lipschitz continuous then we have:

Assumption	Deterministic Gradient Descent	Stochastic Gradient Descent
PL	$\mathcal{O}\left(\log(1/\varepsilon)\right)$	$\mathcal{O}\left(1/\varepsilon\right)$
Convex	$\mathcal{O}\left(1/\varepsilon\right)$	$\mathcal{O}\left(1/\varepsilon^2\right)$
Non-Convex	$\mathcal{O}\left(1/\varepsilon\right)$	$\mathcal{O}\left(1/\varepsilon^2\right)$

- Stochastic has low iteration cost but slow convergence rate.

Results for Gradient Descent

Stochastic iterations are n times faster, but how many iterations are needed?

If ∇f is Lipschitz continuous then we have:

Assumption	Deterministic Gradient Descent	Stochastic Gradient Descent
PL	$\mathcal{O}\left(\log(1/\varepsilon)\right)$	$\mathcal{O}\left(1/\varepsilon\right)$
Convex	$\mathcal{O}\left(1/\varepsilon\right)$	$\mathcal{O}\left(1/\varepsilon^2\right)$
Non-Convex	$\mathcal{O}\left(1/\varepsilon\right)$	$\mathcal{O}\left(1/\varepsilon^2\right)$

- Stochastic has low iteration cost but slow convergence rate.
    - Sublinear rate even in strongly-convex case.

Results for Gradient Descent

Stochastic iterations are n times faster, but how many iterations are needed?

If ∇f is Lipschitz continuous then we have:

Assumption	Deterministic Gradient Descent	Stochastic Gradient Descent
PL	$\mathcal{O}\left(\log(1/\varepsilon)\right)$	$\mathcal{O}\left(1/\varepsilon\right)$
Convex	$\mathcal{O}\left(1/\varepsilon\right)$	$\mathcal{O}\left(1/\varepsilon^2\right)$
Non-Convex	$\mathcal{O}\left(1/\varepsilon\right)$	$\mathcal{O}\left(1/\varepsilon^2\right)$

- Stochastic has low iteration cost but slow convergence rate.
    - Sublinear rate even in strongly-convex case.
    - Bounds are unimprovable under standard assumptions.

Results for Gradient Descent

Stochastic iterations are n times faster, but how many iterations are needed?

If ∇f is Lipschitz continuous then we have:

Assumption	Deterministic Gradient Descent	Stochastic Gradient Descent
PL	$\mathcal{O}\left(\log(1/\varepsilon)\right)$	$\mathcal{O}\left(1/\varepsilon\right)$
Convex	$\mathcal{O}\left(1/\varepsilon\right)$	$\mathcal{O}\left(1/\varepsilon^2\right)$
Non-Convex	$\mathcal{O}\left(1/\varepsilon\right)$	$\mathcal{O}\left(1/\varepsilon^2\right)$

- Stochastic has low iteration cost but slow convergence rate.
  - Sublinear rate even in strongly-convex case.
  - Bounds are unimprovable under standard assumptions.
  - Oracle returns an unbiased gradient approximation with bounded variance.

Results for Gradient Descent

Stochastic iterations are n times faster, but how many iterations are needed?

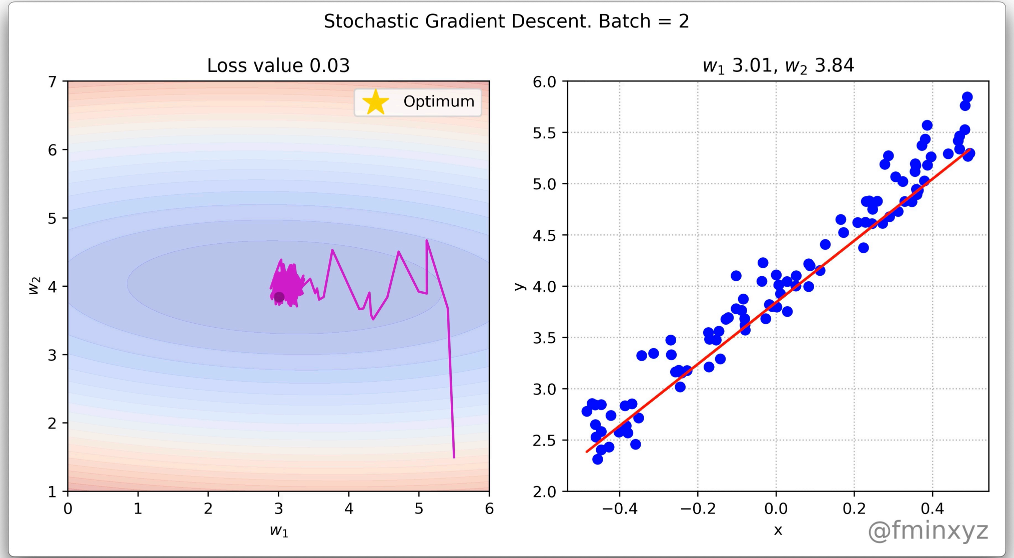
If ∇f is Lipschitz continuous then we have:

Assumption	Deterministic Gradient Descent	Stochastic Gradient Descent
PL	$\mathcal{O}\left(\log(1/\varepsilon)\right)$	$\mathcal{O}\left(1/\varepsilon\right)$
Convex	$\mathcal{O}\left(1/\varepsilon\right)$	$\mathcal{O}\left(1/\varepsilon^2\right)$
Non-Convex	$\mathcal{O}\left(1/\varepsilon\right)$	$\mathcal{O}\left(1/\varepsilon^2\right)$

- Stochastic has low iteration cost but slow convergence rate.
  - Sublinear rate even in strongly-convex case.
  - Bounds are unimprovable under standard assumptions.
  - Oracle returns an unbiased gradient approximation with bounded variance.
- Momentum and Quasi-Newton-like methods do not improve rates in stochastic case. Can only improve constant factors (bottleneck is variance, not condition number).

Stochastic Gradient Descent (SGD)

Typical behaviour



Convergence

Lipschitz continiity implies:

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

Convergence

Lipschitz continiity implies:

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

using (SGD):

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

Convergence

Lipschitz continiity implies:

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

using (SGD):

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

Now let's take expectation with respect to i_k:

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

Convergence

Lipschitz continiity implies:

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

using (SGD):

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

Now let's take expectation with respect to i_k:

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

Using linearity of expectation:

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

Convergence

Lipschitz continiity implies:

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

using (SGD):

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

Now let's take expectation with respect to i_k:

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

Using linearity of expectation:

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

Since uniform sampling implies unbiased estimate of gradient: $\mathbb{E}[\nabla f_{i_k}(x_k)] = \nabla f(x_k)$:

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

Smooth PL case with constant learning rate

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

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

We start from inequality (1):

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

Smooth PL case with constant learning rate

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

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

We start from inequality (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: $\|\nabla f(x_k)\|^2 \geq 2\mu(f(x_k) - f^*)$

Smooth PL case with constant learning rate

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

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

We start from inequality (1):

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

Smooth PL case with constant learning rate

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

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

We start from inequality (1):

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

Subtract f^*

Smooth PL case with constant learning rate

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

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

We start from inequality (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: $\|\nabla f(x_k)\|^2 \geq 2\mu(f(x_k) - f^*)$

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

Subtract f^*

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

Smooth PL case with constant learning rate

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

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

We start from inequality (1):

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

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

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

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

Smooth PL case with constant learning rate

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

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

We start from inequality (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: $\|\nabla f(x_k)\|^2 \geq 2\mu(f(x_k) - f^*)$

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

Subtract f^*

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

Rearrange

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

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

Smooth PL case with constant learning rate

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

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

We start from inequality (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: $\|\nabla f(x_k)\|^2 \geq 2\mu(f(x_k) - f^*)$

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

Subtract f^*

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

Rearrange

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

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

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

Convergence. Smooth PL case.

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. Consider **decreasing stepsize** strategy with $\alpha_k = \frac{2k+1}{2\mu(k+1)^2}$ we obtain

Convergence. Smooth PL case.

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. Consider **decreasing stepsize** strategy with $\alpha_k = \frac{2k+1}{2\mu(k+1)^2}$ we obtain

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

Convergence. Smooth PL case.

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. Consider **decreasing stepsize** strategy with $\alpha_k = \frac{2k+1}{2\mu(k+1)^2}$ we obtain

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

Convergence. Smooth PL case.

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. Consider **decreasing stepsize** strategy with $\alpha_k = \frac{2k+1}{2\mu(k+1)^2}$ we obtain

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

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

Convergence. Smooth PL case.

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. Consider **decreasing stepsize** strategy with $\alpha_k = \frac{2k+1}{2\mu(k+1)^2}$ we obtain

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

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

2. Multiplying both sides by $(k+1)^2$ and letting $\delta_f(k) \equiv k^2\mathbb{E}[f(x_k) - f^*]$ we get

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

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

Convergence. Smooth PL case.

3. Summing up previous inequality from $i = 0$ to k and using the fact that $\delta_f(0) = 0$ we get

which gives the stated rate.

Convergence. Smooth PL case.

3. Summing up previous inequality from $i = 0$ to k and using the fact that $\delta_f(0) = 0$ we get

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

which gives the stated rate.

Convergence. Smooth PL case.

3. Summing up previous inequality from $i = 0$ to k and using the fact that $\delta_f(0) = 0$ we get

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

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

which gives the stated rate.

Convergence. Smooth PL case.

3. Summing up previous inequality from $i = 0$ to k and using the fact that $\delta_f(0) = 0$ we get

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

which gives the stated rate.

Convergence. Smooth PL case.

3. Summing up previous inequality from $i = 0$ to k and using the fact that $\delta_f(0) = 0$ we get

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

which gives the stated rate.

Convergence. Smooth PL case.

3. Summing up previous inequality from $i = 0$ to k and using the fact that $\delta_f(0) = 0$ we get

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

which gives the stated rate.

Convergence. Smooth convex case (bounded variance)

Auxiliary notation

For a (possibly) non-constant stepsize sequence $(\alpha_t)_{t\geq 0}$ define the stepsize-weighted average

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

Everywhere below $f^* \equiv \min_x f(x)$ and $x^* \in \arg\min_x f(x)$.

Smooth convex case with constant learning rate

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

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

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

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

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

Smooth convex case with constant learning rate

1. Начнём с разложения квадрата расстояния до минимума:

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

Smooth convex case with constant learning rate

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

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

Smooth convex case with constant learning rate

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 член с $f(x_k)$ влево и берём полное матожидание:

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

Smooth convex case with constant learning rate

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

Smooth convex case with constant learning rate

5. Делим на $2\alpha k$:

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

Smooth convex case with constant learning rate

5. Делим на $2\alpha k$:

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

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

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

Вычитая f^* из обеих частей, получаем:

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

Smooth convex case with constant learning rate

5. Делим на $2\alpha k$:

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

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

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

Вычитая f^* из обеих частей, получаем:

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

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

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

Smooth convex case with decreasing learning rate

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

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

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

Mini-batch SGD

Mini-batch SGD

Approach 1: Control the sample size

The deterministic method uses all n gradients:

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

The stochastic method approximates this using just 1 sample:

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

A common variant is to use a larger sample B_k ("mini-batch"):

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

particularly useful for vectorization and parallelization.
For example, with 16 cores set $|B_k| = 16$ and compute 16 gradients at once.

Mini-Batching as Gradient Descent with Error

The SG method with a sample B_k ("mini-batch") uses iterations:

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

Let's view this as a "gradient method with error":

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

where e_k is the difference between the approximate and true gradient.

If you use $\alpha_k = \frac{1}{L}$, then using the descent lemma, this algorithm has:

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

for any error e_k.

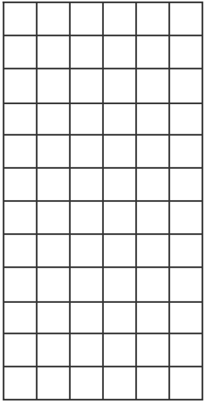
Effect of Error on Convergence Rate

Our progress bound with $\alpha_k = \frac{1}{L}$ and error in the gradient of e_k is:

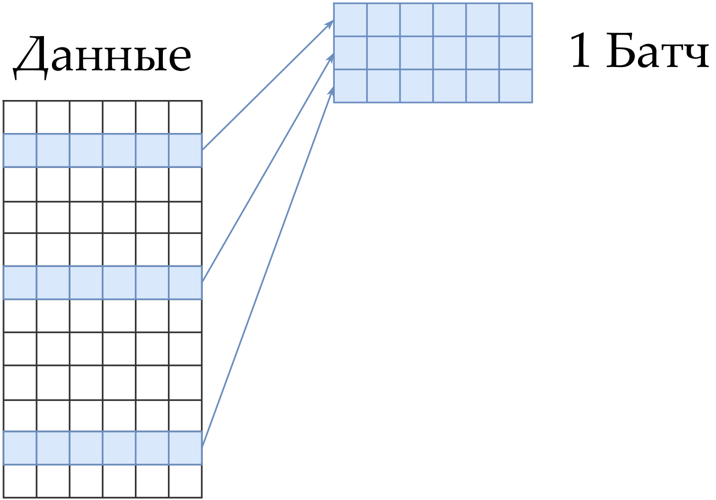
$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2 + \frac{1}{2L}\|e_k\|^2.$$

Идея SGD и батчей

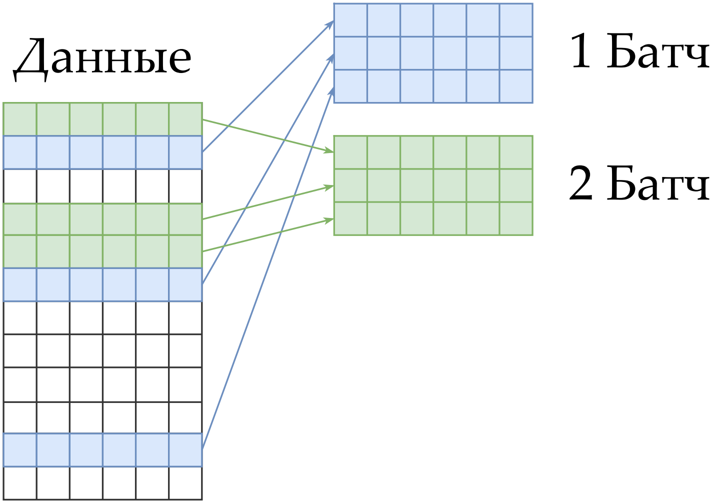
Данные



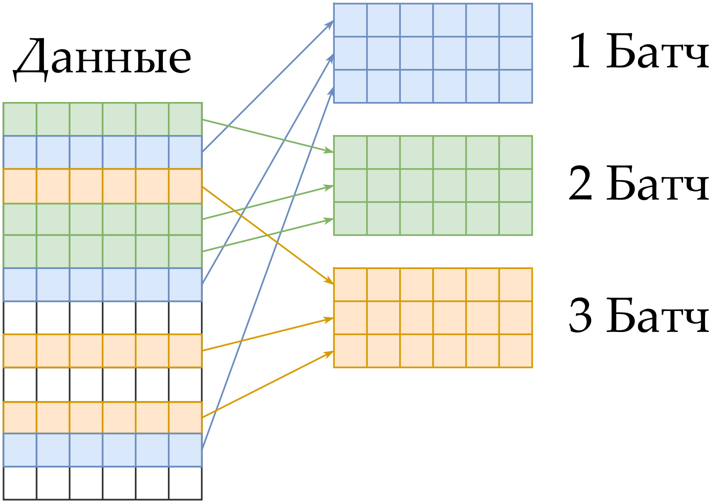
Идея SGD и батчей



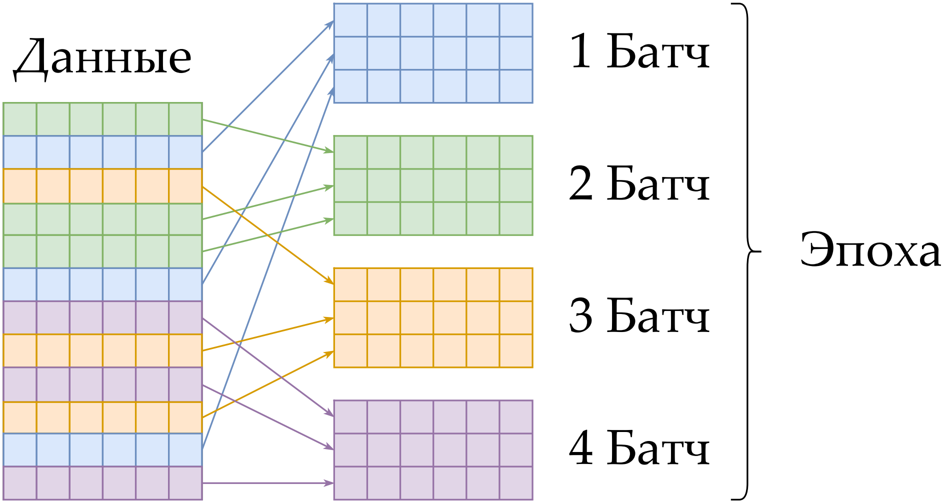
Идея SGD и батчей



Идея SGD и батчей

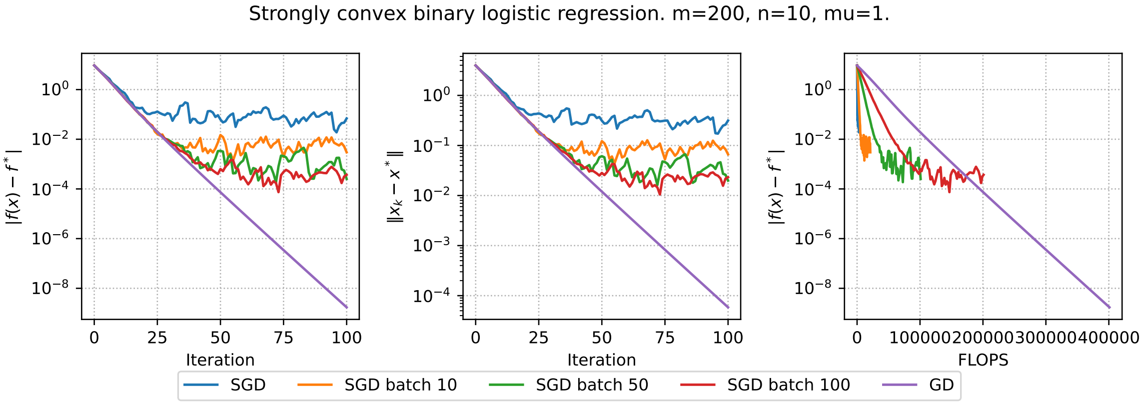


Идея SGD и батчей



Main problem of 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:

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

Conclusions

- SGD with fixed learning rate does not converge even for PL (strongly convex) case

Conclusions

- SGD with fixed learning rate does not converge even for PL (strongly convex) case
- SGD achieves sublinear convergence with rate $\mathcal{O}\left(\frac{1}{k}\right)$ for PL-case.

Conclusions

* SGD with fixed learning rate does not converge even for PL (strongly convex) case
* SGD achieves sublinear convergence with rate $\mathcal{O}\left(\frac{1}{k}\right)$ for PL-case.
* Nesterov/Polyak accelerations do not improve convergence rate

Conclusions

- SGD with fixed learning rate does not converge even for PL (strongly convex) case
- SGD achieves sublinear convergence with rate $\mathcal{O}\left(\frac{1}{k}\right)$ for PL-case.
- Nesterov/Polyak accelerations do not improve convergence rate
- Two-phase Newton-like method achieves $\mathcal{O}\left(\frac{1}{k}\right)$ without strong convexity.