Стохастический градиентный спуск

x_* - реш. задачи (*) $\quad f(x) := \mathbb{E}_\xi f(x,\xi) \to \min\limits_{x\in\mathbb{R}^n} \quad (*)$

SGD: $\quad x^{k+1} = x^k - h\nabla_x f(x^k,\xi^k)$, ξ^k i.i.d.

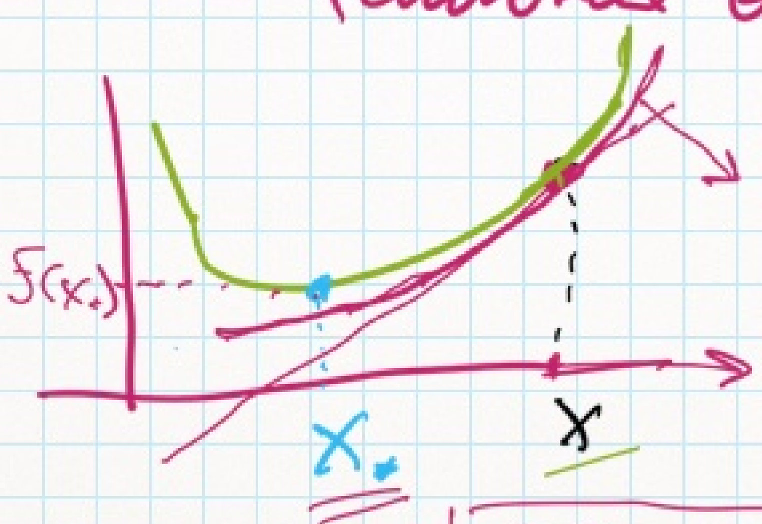
Stochastic Gradient Descent

Предположение 1. $\forall x \to \mathbb{E}_\xi \nabla_x f(x,\xi) = \nabla f(x)$

Предположение 2.

$$\mathbb{E}_\xi\left[\|\nabla f(x,\xi)\|_2^2\right] \le 2L(f(x)-f(x_*)) + \sigma^2$$

Предположение 3. (сильная выпуклость) $\quad f(x_*) \ge f(x) + \langle\nabla f(x), x_*-x\rangle + \frac{\mu}{2}\|x_*-x\|_2^2$



$\left(1-\frac{\mu}{2L}\right)$

Теорема. $h \le \frac{1}{2L}$

$$\boxed{\mathbb{E}\|x^{k+1}-x_*\|_2^2 \le (1-\mu h)^k\|x^0-x_*\|_2^2 + \frac{h\sigma^2}{\mu}}$$

$$\mathbb{E}_\xi\left[\|\nabla f(x,\xi)\|_2^2\right] \le 2L\left(f(x)-f(x_*)\right) + \sigma^2 \quad (**)$$

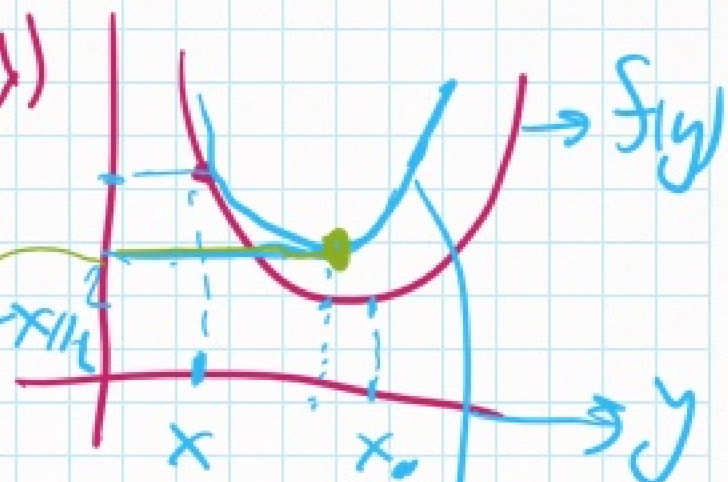
1 случай (Липш. град., нет стох.)

$$\nabla f(x,\xi) \equiv \nabla f(x)$$

$$\|\nabla f(y)-\nabla f(x)\|_2 \le L\|y-x\|_2 \quad (!)$$

$$\Downarrow$$

$$f(y) \le f(x) + \langle \nabla f(x), y-x\rangle + \frac{L}{2}\|y-x\|_2^2$$



$$f(x) + \langle \nabla f(x), y-x\rangle + \frac{L}{2}\|y-x\|_2^2 \to \min_y$$

$$\boxed{f(x) - \frac{1}{2L}\|\nabla f(x)\|_2^2}$$

$$\geq$$

$$f(x_*)$$

$$\Downarrow$$

$$\|\nabla f(x)\|_2^2 \le 2L\left(f(x)-f(x_*)\right)$$

$$\sigma^2 = 0$$

2 случай $\nabla f(x,\xi) = \nabla f(x) + \xi$, $\mathbb{E}\|\xi\|_2^2 \le \sigma^2$

$\mathbb{E}\left[\|\frac{1}{r}\sum \xi_i\|_2^2\right] \le \frac{\sigma^2}{r}$

(**) - верно; L - конст. из (!), σ^2

$$\mathbb{E}\left[\|\nabla f(x,\xi)\|_2^2\right] = \underbrace{\|\nabla f(x)\|_2^2}_{\text{не случ. вел.}} + \mathbb{E}\|\xi\|_2^2$$

$$\mathbb{E}\xi = 0 \qquad \|\nabla f(x)\|_2^2 \le 2L(f(x)-f(x_*)) \qquad \mathbb{E}\|\xi\|_2^2 = \sigma^2$$

<u>3 случай</u>

$$f(x) = \underbrace{\frac{1}{m}\sum_{i=1}^{m} f_i(x)}_{\mathbb{E} f_\xi(x)} \qquad \xi \in R[1,\dots,m]$$

$\forall i = 1,\dots,m$

$\|\nabla f_i(y) - \nabla f_i(x)\|_2 \le L\|y-x\|_2$

$$\mathbb{E}_\xi\left[\|\nabla f_\xi(x)\|_2^2\right] = \mathbb{E}_\xi \|\underbrace{\nabla f_\xi(x) - \nabla f_\xi(x_*)}_{a} + \underbrace{\nabla f_\xi(x_*)}_{b}\|_2^2$$

$$\le 2\mathbb{E}_\xi\|\nabla f_\xi(x) - \nabla f_\xi(x_*)\|_2^2 + 2\mathbb{E}\|\nabla f_\xi(x_*)\|_2^2 \le$$

$(a+b)^2 \le 2a^2 + 2b^2$

$$\le \underbrace{4L}_{2L}(f(x)-f(x_*)) + 2\mathbb{E}_\xi\left[\|\nabla f_\xi(x_*)\|_2^2\right]$$

$\nabla f_i(x_*) = 0$

интерполяция

$$\underbrace{\frac{2}{m}\sum_{i=1}^{m}\|\nabla f_i(x_*)\|_2^2}_{\sigma^2 = 0}$$

$f_i(x) = (y_i - g_i(x))^2$

$y_i = g_i(x) \quad i = \overline{1,m}$

Док-во теоремы

$h = h_k$ $\quad x^{k+1} = x^k - h \nabla_x f(x^k, \xi^k)$

ξ^k i.i.d.

Лемма

$$\mathbb{E}\|x^{k+1} - x_*\|_2^2 \le (1 - \mu h)\,\mathbb{E}\|x^k - x_*\|_2^2 - h\,\mathbb{E}\big(f(x^k) - f(x_*)\big) + h^2\sigma^2$$

$\boxed{h \le \frac{1}{2L}}$

○

(II) $$\mathbb{E}\Big[\|x^{k+1} - x_*\|_2^2 \,\Big|\, \underbrace{x^k, \dots, x^1, \xi^{k-1}, \dots, \xi^0}_{\mathcal{F}_k}\Big] = \mathbb{E}\Big[\|x^k - x_*\|_2^2 -$$

$$- 2h\langle \nabla_x f(x^k, \xi^k), x^k - x_*\rangle + h^2\|\nabla_x f(x^k, \xi^k)\|_2^2 \,\Big|\, \mathcal{F}_k\Big] =$$

$$= \|x^k - x_*\|_2^2 - 2h\langle \nabla f(x^k), x^k - x_*\rangle +$$

по м-сильной вып.

$$+ h^2\underbrace{\mathbb{E}\big[\|\nabla_x f(x^k, \xi^k)\|_2^2 \,\big|\, \mathcal{F}_k\big]}_{\mathbb{E}_{\xi^k}\left[\|\nabla_x f(x^k, \xi^k)\|_2^2\right]} \le \|x^k - x_*\|_2^2 -$$

$\mathbb{E}_{\xi^k}\left[\|\nabla_x f(x^k, \xi^k)\|_2^2\right] \to$ оцениваю по предл. 2.

фикс.

$$- 2h\Big(\frac{\mu}{2}\|x^k - x_*\|_2^2 + \big(f(x^k) - f(x_*)\big)\Big) +$$

$\to$ предл. 3

$$+ h^2\big(2L\big(f(x^k) - f(x_*)\big) + \sigma^2\big)$$

$\to$ предл 2

$$\mathbb{E}_{\mathcal{F}_k}[(\square)]$$

$$\mathbb{E}[\|x^{k+1}-x_*\|_2^2] \le (1-\mu h)\,\mathbb{E}[\|x^k-x_*\|_2^2] -$$

$$-\underbrace{2h\,(1-Lh)}\,(\mathbb{E} f(x^k) - f(x_*)) + h^2\sigma^2$$

$$h \le \frac{1}{2L} \Rightarrow (1-Lh) \ge \frac{1}{2}$$

$$\mathbb{E}[\|x^{k+1}-x_*\|_2^2] \le (1-\mu h)\,\mathbb{E}[\|x^k-x_*\|_2^2] + h^2\sigma^2$$

$$h = 1/(2L) \qquad \Downarrow \qquad h^2\sigma^2\sum_{t=0}^{\infty}(1-\mu h)^t = h^2\sigma^2\frac{1}{1-(1-\mu h)}$$

$$\boxed{\mathbb{E}\|x^{k+1}-x_*\|_2^2 \le \left(1-\frac{\mu}{2L}\right)^k\|x^0-x_*\|_2^2 + \frac{1\,\sigma^2}{2L\,\mu}}$$

Мини-батчи

$$\nabla f(x,\xi) \to \frac{1}{r}\sum_{j=1}^{r}\nabla f(x,\xi_j)$$

$$\sigma^2 \to \sigma^2/r$$

S. Stich '19

$$\mathbb{E}(f(\bar{x}^N)) - f(x_*) + \underbrace{\mu \mathbb{E}\|x^N - x_*\|_2^2}_{\text{можно не писать}} =$$

$$\bar{x}^N = \frac{1}{N}\sum_{k=0}^{N-1} x^k$$

$$= O\left(\min\left\{LR^2 \exp\left(-\frac{\mu N}{4L}\right) + \frac{\sigma^2}{\mu N}, \frac{LR^2}{N} + \frac{\sigma R}{\sqrt{N}}\right\}\right)$$

$\boxed{\sigma^2 \to \frac{\sigma^2}{r}}$

$$R = \|x^0 - x_*\|_2$$

случай $\varepsilon/2$ // $\varepsilon/2$ // $\varepsilon/2$

$$= O\left(\min\left\{LR^2 \exp\left(-\text{const}\sqrt{\frac{\mu}{L}}N\right) + \frac{\sigma^2}{\mu N}, \frac{LR^2}{N^2} + \frac{\sigma R}{\sqrt{N}}\right\}\right)$$

паралл.

$\nabla f(x)$ – L-Липш.

$\nabla f(x, \xi)$ – σ^2 дисперс.

ε – точн. по функц.

$$\left(1-\frac{\mu}{2L}\right)^N \cdot R^2 \simeq \varepsilon$$

$$N \simeq \frac{L}{\mu}\ln\left(\frac{R^2}{\varepsilon}\right)$$ – число итераций

$$\frac{h\sigma^2}{\mu} \sim \frac{\sigma^2}{L\mu} \simeq \varepsilon$$

$$\frac{\sigma^2}{L\mu r} \simeq \varepsilon \Rightarrow r \simeq \frac{\sigma^2}{L\mu\varepsilon}$$

$$x^{k+1} = x^k - h\nabla_x f(x^k, \xi^k) \qquad \underline{r \sim \frac{1}{\varepsilon}}$$

$$\frac{1}{r}\sum_{j=1}^{r} \nabla_x f(x^k, \xi^{k,j})$$ i.i.d. – лучше

$$\nabla f(x^k)$$

Ускорение: $\left(1-\sqrt{\frac{\mu}{L}}\right)^N R^2 \simeq \varepsilon \Rightarrow N \simeq \sqrt{\frac{L}{\mu}}\ln\frac{R}{\varepsilon}$

Weak growth condition

Vaswani, Bach, Schmidt, 19

$$\mathbb{E}\left[\|\nabla f(x,\xi)\|_2^2\right] \le 2\underset{\geq 1}{\rho} L\,(f(x) - f(x_*)) \quad (WGC)$$

Strong growth condition

$$\mathbb{E}\left[\|\nabla f(x,\xi)\|_2^2\right] \le \rho \|\nabla f(x)\|_2^2 + \sigma^2 \quad (SGC)$$

$\le 2\rho L(f(x) - f(x_*))$ (WGC)

Ускорение имеет место

$$\mathbb{E} f(x^k) - f(x_*) \le \left(1 - \sqrt{\frac{\mu}{\rho^2 L}}\right)^k \left(f(x^0) - f(x_*) + \frac{\mu}{2}\|x^0 - x_*\|_2^2\right)$$

$$n\langle \nabla f(x), e\rangle e = \nabla f(x, e)$$

$n = \dim x$

$$\boxed{n \frac{f(x+\tau e) - f(x)}{\tau} e}$$

$e \in S^n(1)$ – сфера, радиус 1

$$\mathbb{E}\,\nabla f(x,e) = \nabla f(x)$$

$\rho = n$ SGC

$$\left(1 - \frac{1}{n}\sqrt{\frac{\mu}{L}}\right)^N \Delta f = \varepsilon$$

$$\boxed{N \sim n\sqrt{L/\mu}\,\ln(\Delta f/\varepsilon)}$$

Fang, Fan, Friedlander ICLR'2021

$$f(x) := \frac{1}{m}\sum_{i=1}^{m} f_i(x)$$

$\nabla f_\xi(x) = \ell'(h(x)) \cdot \underbrace{\nabla h(x)}_{\leq M}$

$\|\nabla h(x)\|_2 \leq M$

$$\| \quad \text{Crowdsourcing.}$$

$$1 \quad \ell(h(x)), \quad \ell : \mathbb{R} \to \mathbb{R}_+$$

$$|\ell'(\alpha) - \ell'(\beta)| \leq L|\alpha - \beta| \quad (\infty)$$

$\boxed{\ell(0) = 0. \quad \ell(\alpha) \geq 0}$

$\ell'(0) = 0.$

$$|h(y) - h(x)| \leq M\|y - x\|_2$$

$$\|\nabla_x f(x,\xi)\|^2 = \|\nabla f_\xi(x)\|^2 \leq M^2(\ell'(h(x)))^2 =$$

$$= M^2(\ell'(h(x)) - \ell'(0))^2 \leq 2M^2(\ell(h(x)) - \ell(0)) =$$

$$= 2M^2 f_\xi(x)$$

(∞) $\boxed{\frac{1}{2L}|\ell'(h)|^2 \leq \ell(h) - \ell(0)}$

$L = 1$

$$\|\nabla_x f(x,\xi)\|_2^2 \leq 2M^2(f_\xi(x) - f(x_*)) + 2M^2 f(x_*)$$

$$\mathbb{E}\|\nabla_x f(x,\xi)\|_2^2 \leq 2M^2(f(x) - f(x_*)) + \underbrace{2M f(x_*)}_{\sigma^2}.$$