$$\min_{x \in \mathbb{R}^d} f(x) \quad \cancel{\to} \quad \nabla f(x)$$

$$\| $$

$$\mathbb{E}[f(x,\xi)] \leftarrow \mathbb{E}_D[l(g(x,\xi),b)] \to \min$$

веса — x; данные $\xi \in D$ ← не знаем

$\xi \sim D$

$\xi_1 \dots \xi_n \in D$ → заменяем средним

$$\frac{1}{n}\sum_{i=1}^{n} l(g(x;\xi_i); b_i)$$

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

- GD: $x^{k+1} = x^k - \gamma \nabla f(x^k)$ ← целое n-батчевое
- SGD: $x^{k+1} = x^k - \gamma \nabla f_{i_k}(x^k)$ ← каждую итерацию считаем ∇f_{i_k} случайного индекса i_k

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

$$\|x^{k+1} - x^*\|^2 = \|x^k - x^*\|^2 - 2\gamma_k \langle \nabla f_{i_k}(x^k); x^k - x^* \rangle + \gamma_k^2 \|\nabla f_{i_k}(x^k)\|^2$$

$\mathbb{E}$ по i_k $\qquad \mathbb{E}[\ \cdot \mid x^k]$

данные всего пред.

$$\mathbb{E}[\|x^{k+1} - x^*\|^2 \mid x^k] = \|x^k - x^*\|^2 - 2\gamma_k \mathbb{E}[\langle \nabla f_{i_k}(x^k); x^k - x^* \rangle \mid x^k] + \gamma_k^2 \mathbb{E}[\|\nabla f_{i_k}(x^k)\|^2 \mid x^k]$$

$$= \|x^k - x^*\|^2 - 2\gamma_k \langle \underline{\mathbb{E}[\nabla f_{i_k}(x^k) \mid x^k]}; x^k - x^* \rangle + \gamma_k^2 \, \mathbb{E}\left[\|\nabla f_{i_k}(x^k)\|^2 \mid x^k \right]$$

$$= \|x^k - x^*\|^2 - \underline{2\gamma_k \langle \nabla f(x^k); x^k - x^* \rangle} + \gamma_k^2 \, \mathbb{E}\left[\|\nabla f_{i_k}(x^k)\|^2 \mid x^k \right]$$

[illegible]

$$\le \|x^k - x^*\|^2 - 2\gamma_k \left(f(x^k) - f(x^*) + \frac{\mu}{2} \|x^k - x^*\|^2 \right) + \gamma^2 \, \mathbb{E}\left[\|\nabla f_{i_k}(x^k)\|^2 \mid x^k \right]$$

было GD

$$= (1 - \gamma_k \mu) \|x^k - x^*\|^2 - 2\gamma_k \left(f(x^k) - f(x^*) \right) + \gamma^2 \, \mathbb{E}\left[\|\nabla f_{i_k}(x^k)\|^2 \mid x^k \right]$$

$=$

$$\mathbb{E}\left[\|\nabla f_{i_k}(x^k)\|^2 \mid x^k \right] =$$

$$\mathbb{E}\left[\|\nabla f_{i_k}(x^k) - \nabla f(x^k) + \nabla f(x^k)\|^2 \mid x^k \right]$$

$$= \mathbb{E}\left[\underbrace{\|\nabla f_{i_k}(x^k) - \nabla f(x^k)\|^2}_{c} + \underbrace{\|\nabla f(x^k)\|^2}_{h} + 2 \langle \underbrace{\nabla f_{i_k}(x^k) - \nabla f(x^k)}_{c}; \underbrace{\nabla f(x^k)}_{h} \rangle \mid x^k \right]$$

$$= \mathbb{E}\left[\|\nabla f_{i_k}(x^k) - \nabla f(x^k)\|^2 \mid x^k \right] + \|\nabla f(x^k)\|^2$$

$$+2 \langle \mathbb{E}\left[\nabla f_{i_k}(x^k) | x^k\right] - \nabla f(x^k) \; ; \; \nabla f(x^k) \rangle$$

$$= 0$$

$$= \mathbb{E}\left[\| \nabla f_{i_k}(x^k) - \nabla f(x^k)\|^2 \,\middle|\, x^k\right] + \|\nabla f(x^k)\|^2$$ было GD

Предположение:

$$\mathbb{E}\left[\|\nabla f_i(x) - \nabla f(x)\|^2\right] \leq B\|\nabla f(x)\|^2 + \sigma^2$$

$$\mathbb{E}\left[\|\nabla f_{i_k}(x^k)\|^2 \,\middle|\, x^k\right] \leq \|\nabla f(x^k)\|^2 + B\|\nabla f(x^k)\|^2 + \sigma^2$$

$$= (1-\gamma_k\mu)\|x^k - x^*\|^2 - 2\gamma_k(f(x^k) - f(x^*)) + (1+B)\gamma_k^2\|\nabla f(x^k) - \nabla f(x^*)\|^2 + \gamma_k^2\sigma^2$$

$\nabla f(x^*) = 0$ иначе град.

$$\leq (1-\gamma_k\mu)\|x^k - x^*\|^2 - 2\gamma_k(f(x^k) - f(x^*)) + (1+B)\gamma_k^2\left(2L(f(x^k) - f(x^*))\right) + \gamma_k^2\sigma^2$$

$$-2 + (1+B)4L\gamma_k = 0$$

$$\gamma_k \leq \frac{1}{2L(1+B)}$$

$$\leq (1-\gamma_k\mu)\|x^k - x^*\|^2 + \gamma_k^2\sigma^2$$

$$\mathbb{E}\left[\|x^{k+1}-x^*\|^2 \mid x^k\right] \le (1-\gamma_k\mu)\|x^k-x^*\|^2+\gamma_k^2\sigma^2$$

$$\mathbb{E}\left[\mathbb{E}\left[\|x^{k+1}-x^*\|^2 \mid x^k\right]\right] \le (1-\gamma_k\mu)\,\mathbb{E}\left[\|x^k-x^*\|^2\right]+\gamma_k^2\sigma^2$$

tower property дважды

$$\underbrace{\mathbb{E}\left[\|x^{k+1}-x^*\|^2\right]}_{R_{k+1}^2} \le (1-\gamma_k\mu)\underbrace{\mathbb{E}\left[\|x^k-x^*\|^2\right]}_{R_k^2}+\gamma_k^2\sigma^2$$

1) $\gamma_k \equiv \gamma$ запускаем

$= (1-\gamma\mu)R_{k-1}^2+\gamma^2\sigma^2$

$$\begin{aligned}
R_{k+1}^2 &\le (1-\gamma\mu)R_k^2+\gamma^2\sigma^2\\
&\le (1-\gamma\mu)^2R_{k-1}^2+\gamma^2\sigma^2+(1-\gamma\mu)\gamma^2\sigma^2\\
&\le (1-\gamma\mu)^3R_{k-2}^2+\gamma^2\sigma^2+(1-\gamma\mu)\gamma^2\sigma^2+(1-\gamma\mu)^2\gamma^2\sigma^2\\
&\quad \ldots\\
&\le (1-\gamma\mu)^{k+1}R_0^2+\gamma^2\sigma^2\sum_{i=0}^{k}(1-\gamma\mu)^i
\end{aligned}$$

было GD

$$\le \gamma^2\sigma^2\sum_{i=0}^{\infty}(1-\gamma\mu)^i$$

$$\le \frac{\gamma^2\sigma^2}{\gamma\mu} = \frac{\gamma\sigma^2}{\mu}$$

$$R_{k+1}^2 \le \underbrace{(1-\mu\gamma)^{k+1}R_0^2}_{\text{как у GD}}+\underbrace{\frac{\gamma\sigma^2}{\mu}}_{\text{окр. решения}}$$

сход.

γ – пост.

- $\gamma = \frac{1}{K+1}$ $\quad R^2_{K+1} \le \ldots + \frac{\sigma^2}{\mu(K+1)}$ сублинейная

общее число итераций (теория)

плохо

- $\gamma_k = \frac{1}{k+1}$ (вспоминаем FW) $\rightarrow F_{k+1} \le (1-\eta_k)F_k + \eta_k^2 C$

$$R^2_{k+1} \le (1-\mu\gamma_k) R^2_k + \gamma_k^2 \sigma^2$$

Можно добиться

$$\mathbb{E}\left[\|x^{K+1} - x^*\|^2\right] = O\Big(\underbrace{\left(1-\frac{\mu}{2L}\right)^K \|x^0 - x^*\|^2}_{GD} + \underbrace{\frac{\sigma^2}{\mu^2 K}}_{\text{стох}}\Big)$$

сублинейная

нижняя $\Omega\Big(\underbrace{\left(1-\sqrt{\frac{\mu}{L}}\right)^K \|x^0 - x^*\|^2}_{\text{ускоряется}} + \underbrace{\frac{\sigma^2}{\mu^2 K}}_{\text{не ускоряется}}\Big)$

1) оптим метод - Nesterov (стох вариация)

2) стох больше (σ^2 ~~→~~ больше) - оптим метод SGD

- брать батч b

$\nabla f_{i_k}(x^k) \longrightarrow \frac{1}{b}\sum_{i\in S_k} \nabla f_i(x^k)$ $\quad S_k$ набор индексов i_k

небольш. размера $b \ll n$

i_k - независимы

$|S_k| = b$

$$\mathbb{E}\left[\frac{1}{b}\sum_{i\in S_k}\nabla f_i(x^k)\right] = \nabla f(x^k)$$

$$\mathbb{E}\left\|\frac{1}{b}\sum_{i\in S}\left(\nabla f_i(x)-\nabla f(x)\right)\right\|^2 =$$

$$= \mathbb{E}\left[\frac{1}{b^2}\left(\sum_{i\in S}\underbrace{\|\nabla f_i(x)-\nabla f(x)\|^2}_{\le B\|\nabla f(x)\|^2+\sigma^2} + \sum_{\substack{i\neq j\\ i,j\in S}}\langle \overbrace{\underbrace{\nabla f_i(x)-\nabla f(x)}_{\mathbb{E}[\quad|\text{fix } i]}}^{=0};\ \overbrace{\underbrace{\nabla f_j(x)-\nabla f(x)}_{\mathbb{E}[\quad|\text{fix } j]}}^{=0}\rangle\right)\right]$$

$$\le \frac{1}{b^2}\cdot b\left(B\|\nabla f(x)\|^2+\sigma^2\right)$$

$$\le \frac{B}{b}\|\nabla f(x)\|^2+\frac{\sigma^2}{b}$$

- Будем батч b_k $\qquad b_k\uparrow \qquad \dfrac{\sigma^2}{\gamma\mu\, b_k}$