• <u>Задача стох. оптимизации</u>

$$\min_{x\in\mathbb{R}^d} \left[f(x) := \mathbb{E}_{\xi\sim D}\left[f(x,\xi)\right] \right]$$

• <u>Задача ML:</u>

природа

модель

$$f(x) := \mathbb{E}_{\xi\sim D}\left[l\left(g(x,\xi_a),\ \xi_b\right) \right]$$

(ξ_a, ξ_b) функция потерь веса модели объект метка

Вычислить f и ∇f невозможно

• <u>Что делать?</u>

1) <u>Онлайн подход:</u>

$\nabla f(x,\xi)$ – градиент по точке данных

$$\nabla f(x,\xi) = \nabla_x\, l\left(g(x,\xi_a),\xi_b\right)$$

$$\mathbb{E}_{\xi\sim D}\left[\nabla f(x,\xi)\right] = \nabla f(x)$$

2) <u>Оффлайн постановка:</u> у нас есть выборка $\{\xi_i\}_{i=1}^n$

аппроксимируем $\mathbb{E}_{\xi\sim D}$:

$$\min_{x\in\mathbb{R}^d} \left[f(x) = \frac{1}{n}\sum_{i=1}^n l\left(g(x,\xi_{i,a}),\ \xi_{i,b}\right)\right]$$

можно считать ∇f

⊖ дорого

⊖ переобучение при использ. ∇f

⊕ будем вычислять градиент по части выборки (батч)

- <u>Стохастический градиентный спуск</u>

$$x^{k+1} = x^k - \gamma \nabla f(x^k, \xi^k)$$

объект выборки
(генерируется независимо
и равномерно)

- Независимость: $\mathbb{E}_\xi[\nabla f(x^k, \xi^k)]$

- Равномерность: $\mathbb{E}_\xi[\nabla f(x^k, \xi^k)] =$ (оффлайн пост.)

$$= \sum_{i=1}^{n} \underbrace{P\{\xi^k = \xi_i\}}_{\text{равен}} \nabla f(x^k, \xi_i) =$$

$$= \sum_{i=1}^{n} \frac{1}{n} \nabla f(x^k, \xi_i) = \frac{1}{n} \sum \nabla f(x^k, \xi_i) = \nabla f(x^k)$$

- <u>условное математическое ожидание</u>

$$\mathbb{E}[\cdot \mid x^k] = \mathbb{E}[\cdot \mid \mathcal{F}_k]$$

$\mathcal{F}_k$ – σ-алгебра, порожд. $x^0, \xi^0, \dots \xi^{k-1}$

$$\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[X]$$

- f – μ-сильно выпукла
- $f(\cdot, \xi)$ – L-гладкий
- $\mathbb{E}_\xi[\nabla f(x, \xi)] = \nabla f(x)$

Док-во сходимости:

$$\|x^{k+1}-x^*\|_2^2 = \|x^k - \gamma\nabla f(x^k,\xi^k) - x^*\|_2^2$$

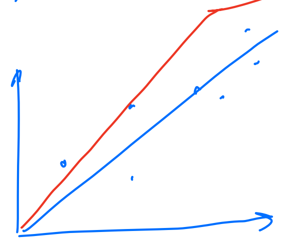
$$= \|x^k - x^*\|_2^2 - 2\gamma\langle\nabla f(x^k,\xi^k); x^k - x^*\rangle$$

$$+\gamma^2\|\nabla f(x^k,\xi^k)\|_2^2$$

$\|\nabla f(x^k,\xi^k)\|_2^2$:

$$\|\nabla f(x^k,\xi^k)\|_2^2 = \|\nabla f(x^k,\xi^k) - \nabla f(x^*,\xi^k) + \underbrace{\nabla f(x^*,\xi^k)}_{\neq 0}\|_2^2$$

$$\nabla f(x^*) = 0 \Rightarrow \cancel{\nabla f(x^*,\xi^k) = 0}$$



КБШ

$$\le 2\|\nabla f(x^k,\xi^k) - \nabla f(x^*,\xi^k)\|_2^2 + 2\|\nabla f(x^*,\xi^k)\|_2^2$$

L- гладкость $f(\ ,\xi)$

$$\le 4L\left(f(x^k,\xi^k) - f(x^*,\xi^k) + \underbrace{\langle\nabla f(x^*,\xi^k); x^k - x^*\rangle}_{\neq 0}\right)$$

$$+2\|\nabla f(x^*,\xi^k)\|_2^2$$

Подставляем:

$$\|x^{k+1}-x^*\|_2^2 \le \|x^k - x^*\|_2^2 - 2\gamma\langle\nabla f(x^k,\xi^k); x^k - x^*\rangle$$

$$+\gamma^2\cdot 4L\left(f(x^k,\xi^k) - f(x^*,\xi^k) + \langle\nabla f(x^*,\xi^k); x^k - x^*\rangle\right)$$

$$+2\gamma^2\|\nabla f(x^*,\xi^k)\|_2^2$$

Берем м.о.

$$\mathbb{E}\left[\|x^{k+1}-x^*\|_2^2\right] \le \mathbb{E}\left[\|x^k-x^*\|_2^2\right]$$
$$-2\gamma\,\mathbb{E}\left[\langle \nabla f(x^k,\xi^k), x^k-x^*\rangle\right]$$
$$+\gamma^2\cdot 4L\,\mathbb{E}\Big[f(x^k,\xi^k)-f(x^*,\xi^k)$$
$$+\langle \nabla f(x^*,\xi^k); x^k-x^*\rangle\Big]$$
$$+2\gamma^2\,\mathbb{E}\left[\|\nabla f(x^*,\xi^k)\|_2^2\right]$$

Tower property $\mathbb{E}[\ ] = \mathbb{E}[\mathbb{E}[\ |x^k]]$

$$\mathbb{E}\left[\|x^{k+1}-x^*\|_2^2\right] \le \mathbb{E}\left[\|x^k-x^*\|_2^2\right]$$
$$-2\gamma\,\mathbb{E}\left[\mathbb{E}\left[\langle \nabla f(x^k,\xi^k); x^k-x^*\rangle | x^k\right]\right]$$
$$+4\gamma^2 L\,\mathbb{E}\Big[\mathbb{E}\Big[f(x^k,\xi^k)-f(x^*,\xi^k)$$
$$-\langle \nabla f(x^*,\xi^k); x^k-x^*\rangle | x^k\Big]\Big]$$
$$+2\gamma^2\,\mathbb{E}\left[\mathbb{E}\left[\|\nabla f(x^*,\xi^k)\|_2^2 | x^k\right]\right]$$

независимость ξ

$$= \mathbb{E}\left[\|x^k-x^*\|_2^2\right]$$
$$-2\gamma\,\mathbb{E}\left[\langle \mathbb{E}[\nabla f(x^k,\xi^k)|x^k]; x^k-x^*\rangle\right]$$
$$+4\gamma^2 L\,\mathbb{E}\Big[\mathbb{E}[f(x^k,\xi^k)|x^k]-\mathbb{E}[f(x^*,\xi^k)|x^k]$$
$$+\langle \mathbb{E}[\nabla f(x^*,\xi^k)|x^k], x^k-x^*\rangle\Big]$$
$$+2\gamma^2\,\mathbb{E}\left[\mathbb{E}\left[\|\nabla f(x^*,\xi^k)\|_2^2 | x^k\right]\right]$$

равномерность по ξ

$$= \mathbb{E}\left[\|x^k - x^*\|_2^2\right]$$

$$- 2\gamma\, \mathbb{E}\left[\langle \nabla f(x^k);\ x^k - x^* \rangle\right]$$

$$+ 4\gamma^2 L\, \mathbb{E}\left[f(x^k) - f(x^*)\right]$$

$$+ 4\gamma^2 L\, \mathbb{E}\left[\langle \cancel{\nabla f(x^*)}^{\,0};\ x^k - x^* \rangle\right]$$

$$+ 2\gamma^2\, \cancel{\mathbb{E}}\Big[\underbrace{\frac{1}{n}\sum_{i=1}^{n} \|\nabla f(x^*, \xi_i)\|_2^2}_{\sigma_*^2}\Big]$$

Получим:

$$\mathbb{E}\left[\|x^{k+1} - x^*\|_2^2\right] \le \mathbb{E}\left[\|x^k - x^*\|_2^2\right]$$

$$- 2\gamma\, \mathbb{E}\left[\langle \nabla f(x^k);\ x^k - x^* \rangle\right]$$

$$+ 4\gamma^2 L\, \mathbb{E}\left[f(x^k) - f(x^*)\right]$$

$$+ 2\gamma^2 \sigma_*^2$$

μ-сильная выпуклость f:

$$\mathbb{E}\left[\|x^{k+1} - x^*\|_2^2\right] \le \mathbb{E}\left[\|x^k - x^*\|_2^2\right]$$

$$- 2\gamma\, \mathbb{E}\left[\frac{\mu}{2}\|x^k - x^*\|_2^2 + f(x^k) - f(x^*)\right]$$

$$+ 4\gamma^2 L\, \mathbb{E}\left[f(x^k) - f(x^*)\right]$$

$$+ 2\gamma^2 \sigma_*^2$$

$$= (1-\gamma\mu)\,\mathbb{E}\|x^k - x^*\|_2^2 + 2\gamma\underline{(2\gamma L - 1)}\,\underbrace{\mathbb{E}\left[f(x^k) - f(x^*)\right]}_{\geq 0} + 2\gamma^2\sigma_*^2$$

$\gamma \leq \frac{1}{2L}$

$$\leq (1-\gamma\mu)\,\mathbb{E}\left[\|x^k - x^*\|_2^2\right] + 2\gamma^2\sigma_*^2$$

$$\boxed{\underbrace{\mathbb{E}\left[\|x^{k+1} - x^*\|_2^2\right]}_{R_{k+1}^2} \leq (1-\gamma\mu)\,\underbrace{\mathbb{E}\left[\|x^k - x^*\|_2^2\right]}_{R_k^2} + 2\gamma^2\sigma_*^2}$$

Запускаем рекурсию

$$R_{k+1}^2 \leq (1-\gamma\mu)R_k^2 + 2\gamma^2\sigma_*^2$$

$$\leq (1-\gamma\mu)\left[(1-\gamma\mu)R_{k-1}^2 + 2\gamma^2\sigma_*^2\right] + 2\gamma^2\sigma_*^2$$

$$= (1-\gamma\mu)^2 R_{k-1}^2 + 2\gamma^2\sigma_*^2\left[1 + (1-\gamma\mu)\right]$$

$$\ldots$$

$$\leq (1-\gamma\mu)^{k+1} R_0^2 + 2\gamma^2\sigma_*^2 \sum_{i=0}^{k} (1-\gamma\mu)^i$$

$$\leq (1-\gamma\mu)^{k+1} R_0^2 + 2\gamma^2\sigma_*^2 \underbrace{\sum_{i=0}^{\infty} (1-\gamma\mu)^i}_{\frac{1}{\gamma\mu}}$$

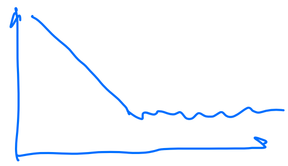
$$= (1-\gamma\mu)^{k+1} R_0^2 + \frac{2\gamma\sigma_*^2}{\mu} \qquad \blacksquare$$

Сходимость SGD с постоян. шагом

$$\boxed{\mathbb{E}\left[\|x^{k+1}-x^*\|_2^2\right] \le (1-\gamma\mu)^{k+1}\,\mathbb{E}\left[\|x^0-x^*\|_2^2\right] + \frac{2\gamma\sigma_*^2}{\mu}}$$

⊕ сходимость линейная, как у GD

⊖ до окр. решения: $\sim\gamma$, $\sim\sigma_*^2$



⊕ простой метод

- Как бороться со сход. к окрестности?

1) γ уменьшить

$$\gamma \to \gamma^k \approx \frac{1}{k}\ ;\ \frac{1}{\sqrt{k}}$$

⊕ решается вопрос сходимости

⊖ портит линейную сходимость (можно исправить с помощью доп. проц.)

2) σ_*^2 уменьшить

$$\nabla f(x^k, \xi^k) \longrightarrow \frac{1}{b}\sum_{\xi \in S_k} \nabla f(x^k, \xi)$$

S_k — набор объектов из об. выборки размера b

- $\mathbb{E}_{S^k}\left[\frac{1}{b}\sum \nabla f(x^k,\xi)\right] =$

$$= \frac{1}{b}\sum \mathbb{E}_\xi\left[\nabla f(x^k,\xi)\right] = \frac{1}{b}\sum \nabla f(x^k) = \nabla f(x^k)$$

$$\bullet\ \mathbb{E}_{S^k}\left[\left\| \frac{1}{b}\sum_{\xi\in S^k} \nabla f(x^*,\xi)\right\|_2^2\right] =$$

$$= \frac{1}{b^2}\,\mathbb{E}_{S^k}\left[\sum_{\xi\in S^k}\|\nabla f(x^*,\xi)\|_2^2\right]$$

$$+ \frac{1}{b^2}\,\mathbb{E}_{S^k}\left[\sum_{\substack{\xi\neq\eta \\ \xi,\eta\in S^k}} \langle \nabla f(x^*,\xi), \nabla f(x^*,\eta)\rangle\right]$$

$$= \frac{1}{b^2}\sum \mathbb{E}_{\xi}\|\nabla f(x^*,\xi)\|_2^2 \leftarrow \frac{1}{b}\cdot \underbrace{\frac{1}{n}\sum_{i=1}^{n}\|\nabla f(x^*,\xi_i)\|_2^2}_{\sigma_*^2} = \frac{\sigma_*^2}{b}$$

$$+ \frac{1}{b^2}\,\mathbb{E}_{S^k}\left[\sum_{\substack{\xi\neq\eta \\ \xi,\eta\in S^k}} \langle \nabla f(x^*,\underbrace{\xi), \nabla f(x^*,\eta}_{\text{независимы}})\rangle\right]$$

$$\mathbb{E}\langle \nabla f(x^*,\xi),\ \nabla f(x^*,\eta)\rangle$$
$$\parallel$$
$$\langle \mathbb{E}\,\nabla f(x^*,\xi);\ \mathbb{E}\,\nabla f(x^*,\eta)\rangle$$
$$\parallel$$
$$\langle \nabla f(x^*); \nabla f(x^*)\rangle = 0$$

$$= \boxed{\frac{\sigma_*^2}{b}} \leftarrow \text{эффект батчирования}$$

⊕ окрестность уменьш. в b раз

⊖ стоимость возрастает