Задача стохастической оптимизации

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

Пример из ML

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

- L — ф. потерь
- x — веса
- $\xi = (\xi_a,\xi_b)$
- D — природа (неизв.)
- g — модель
- ξ_a — объект
- ξ_b — метка

⇓

f, ∇f вычислить нельзя

Что делать?

1) Онлайн подход

$$\nabla_x f(x,\xi) = \nabla_x L(g(x,\xi_a),\xi_b) \quad \leftarrow \text{градиент на точке данных}$$

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

2) Оффлайн подход

есть выборка $\{\xi_i\}_{i=1}^n$

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

$\hat f \neq f$ — другая задача

$\hat f \approx f$ — при больших n (аппроксимация $\mathbb{E}_{\xi\sim D}$ по Монте-Карло)

можно считать только $\nabla \hat f$:

⊖ дорого ⊖ переобучение ($\hat f \neq f$)

⊕ вычислять не полный $\nabla \hat f$, а град. по части выборки

Стохастический градиентный спуск

$$x^{k+1} = x^k - \gamma \underbrace{\nabla f(x^k, \xi^k)}_{\text{стохаст. град.}}$$

ξ^k – независимы и равномерно

по 1-му аргументу об. выпуклая

- Независимость

$$\mathbb{E}_\xi [\nabla f(x, \xi)]$$

- Равномерность

$$\mathbb{E}_\xi [\nabla f(x, \xi)] = \text{(определение матожидания)}$$

$$= \sum_{i=1}^n \mathbb{P}\{\xi = \xi_i\} \nabla f(x, \xi_i) = \text{(равномерность)}$$

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

Условное математическое ожидание

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

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

фиксирует все, что зависит до x^k (включительно)

tower property:

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

Док-во сходимости:

Предположения:

- f – μ-сильно выпуклая
- $f(\cdot, \xi)$ – L-гладкая (можно $L = L_{max}$ по выборке)
- $\mathbb{E}_\xi[\nabla f(x, \xi)] = \nabla f(x)$ $\quad \mathbb{E}_\xi[f(x, \xi)] = f(x)$

Док-во:

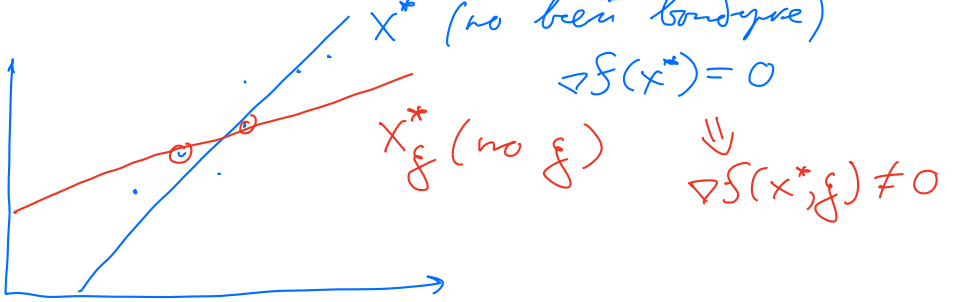
$$\|x^{k+1} - x^*\|_2^2 = \|x^k - \gamma \nabla f(x^k, \xi^k) - x^*\|_2^2$$
$$= \|x^k - x^*\|_2^2 - 2\gamma \langle \nabla f(x^k, \xi^k); x^k - x^* \rangle + \gamma^2 \|\nabla f(x^k, \xi^k)\|_2^2$$

Только м.о. от обеих частей:

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

$\mathbb{E}\left[\|\nabla f(x^k, \xi^k)\|_2^2\right]$:

$$\mathbb{E}\left[\|\nabla f(x^k, \xi^k)\|_2^2\right] = \mathbb{E}\left[\|\nabla f(x^k, \xi^k) - \nabla f(x^*, \xi^k) + \nabla f(x^*, \xi^k)\|_2^2\right]$$



КБШ $\quad \|a+b\|_2^2 \le 2\|a\|_2^2 + 2\|b\|_2^2$

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

L - smoothness of $f(\cdot, \xi)$

$$\leq 4L\,\mathbb{E}\left[f(x^k, \xi^k) - f(x^*, \xi^k) + \langle \nabla f(x^*, \xi^k), x^* - x^k \rangle \right]$$

$$+ 2\,\mathbb{E}\left[\|\nabla f(x^*, \xi^k)\|_2^2 \right] \quad (\ast)$$

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

$$\mathbb{E}\left[\mathbb{E}\left[f(x^k, \xi^k) | x^k \right]\right] = \mathbb{E}\left[f(x^k) \right]$$

$$\mathbb{E}\left[f(x^*, \xi^k) \right] = f(x^*)$$

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

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

$$= \mathbb{E}\left[\langle \underbrace{\nabla f(x^*)}_{=0}, x^k - x^* \rangle \right] = 0$$

$$\mathbb{E}\left[\mathbb{E}\left[\|\nabla f(x^*, \xi^k)\|_2^2 | x^k \right]\right]$$

определение $=$ $\underbrace{\frac{1}{n}\sum_{i=1}^{n} \|\nabla f(x^*, \xi_i)\|_2^2}_{\sigma_*^2}$

описание $=$ $\underbrace{\mathbb{E}_{\xi \sim D} \|\nabla f(x^*, \xi)\|_2^2}_{\sigma_*^2}$

Следовательно

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

Подставим $(**)$ в $(*)$

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

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

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

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

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

Подставляем $(****)$ в $(***)$

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

μ - сильно выпуклость

$$\mathbb{E}\left[\|x^{k+1}-x^*\|_2^2\right] \leq \mathbb{E}\left[\|x^k-x^*\|_2^2\right]-2\gamma\,\mathbb{E}\left[f(x^k)-f(x^*)+\frac{\mu}{2}\|x^k-x^*\|_2^2\right] + \gamma^2\cdot\left(4L\,\mathbb{E}\left[f(x^k)-f(x^*)\right]+2\sigma_*^2\right)$$

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

$$-2\gamma(1-2\gamma L)\underbrace{\mathbb{E}[f(x^k)-f(x^*)]}_{\geq 0} + 2\gamma^2\sigma_*^2$$

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

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

$$\boxed{\mathbb{E}[\|x^{k+1}-x^*\|_2^2] \leq \underbrace{(1-\gamma\mu)\mathbb{E}[\|x^k-x^*\|_2^2]}_{\text{то же самое, что и для GD}} + \underbrace{2\gamma^2\sigma_*^2}_{\text{не было} + \mathbb{E}}}$$

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

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

Запустим рекурсию:

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

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

$$\dots$$

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

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

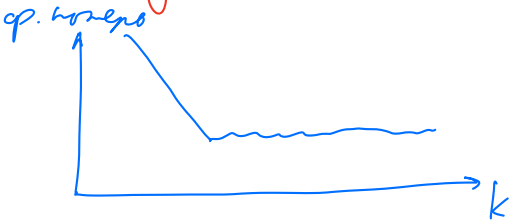
$$\leq (1-\gamma\mu)^{k+1}R_0^2 + \frac{2\gamma\sigma_*^2}{\mu} \quad \blacksquare$$

Сходимость SGD с постоянным шагом

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

⊕ сходимость линейная, как у GD

⊖ сходимость до окрестности: $\sim\gamma$, $\sim\sigma_*^2$

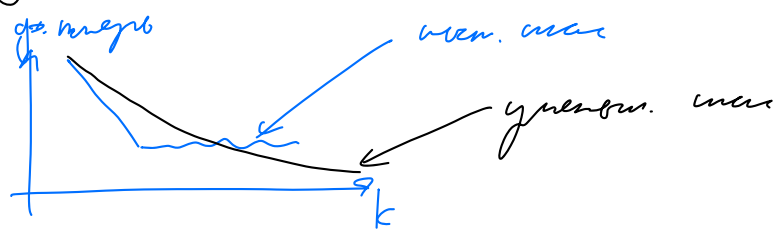


⊕ простота + при хорошей настройке хорошие ML-модели

Как бороться со сходимостью к окрестности?

1) γ уменьшать: ⊕ сходимость глубже
⊖ медленно

$\gamma \to \gamma_k \sim \frac{1}{k},\ \frac{1}{\sqrt{k}}$: ⊕ сходимость к решению
⊖ требуется линейный осд.



2) σ_*^2 уменьшать

$$\nabla f(x,\xi) \to \frac{1}{b}\sum_{\xi\in S}\nabla f(x,\xi)$$

S – батч (набор объектов и меток) размера b

(все точки делаем нд. и равномерно)

$$\bullet\ \mathbb{E}_S\left[\frac{1}{b}\sum_{\xi\in S}\nabla f(x,\xi)\right] =$$

$$= \frac{1}{b}\sum_{\xi\in S}\mathbb{E}_\xi\left[\nabla f(x,\xi)\right] = \text{тк и равн.}$$

$$= \frac{1}{b}\sum_{\xi\in S}\nabla f(x) = \frac{1}{b}\cdot b\,\nabla f(x) = \nabla f(x)$$

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

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

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

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

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

ξ, η — независимо

$$\mathbb{E}_\xi\left[\langle\nabla f(x^*,\xi); \nabla f(x^*,\eta)\rangle\right]$$

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

$$= \langle\underbrace{\nabla f(x^*)}_{=0}, \nabla f(x^*,\eta)\rangle = 0$$

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

⊕ отслеживать улучшение в раз

⊖ стимулирует ритм

На практике:

- $\nabla f_1 \to \nabla f_2 \to \dots \to \nabla f_n$ (и снова к ∇f_1)

не стох метод

shuffling:

- замиксовать данные 1 раз → $f_1 \dots f_n$ более равномерные
- миксовать каждую эпоху → еще стохастичнее