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

ф. потерь

веса

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

(ξ_a,ξ_b) природа (неизв.) модель объект метка

$\Rightarrow$ $f, \nabla f$ вычислить нельзя

Что делать?

1) Онлайн метод

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

градиент по точке данных

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

2) Оффлайн подход

есть выборка $\{\xi_i\}_{i=1}^n$

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

$\tilde f$ — Монте-Карло оценка исходной ф. f

$\tilde f \neq f$ — другая задача

$\tilde{f} \approx f$ — при большом n (аппроксимация быстра)

$\sim \frac{1}{\sqrt{n}}$

можно считать $\nabla \tilde{f}$:

⊖ дорого

⊖ переобучение ($\tilde{f} \neq f$)

⊕ вычислять не ∇f, а град. по части выборки

Стохастический градиентный спуск

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

по 1ому объекту об. выборки

ξ^k — независимо из D/ равномерно

- независимость

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

- равномерно

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

(оффлайн постановка)

$\tilde{f}$ для простоты $\tilde{f} \Rightarrow f$

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

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

Условное математическое ожидание

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

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

характеризует всю случайность, которая произошла до x^k (включительно)

tower property:

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

Док-во сходимости:

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

Док-во:

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

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

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

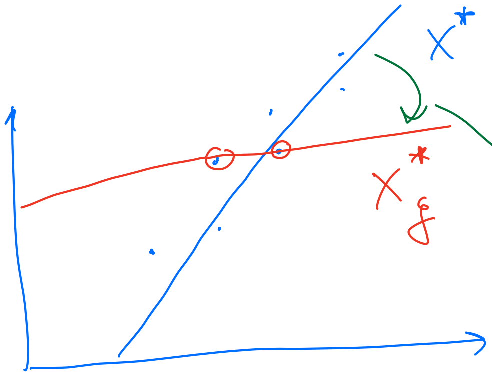
Полное м.о. от обеих частей:

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

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

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

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



x^* (по всей выборке $\nabla f(x^*)$)

$\nabla f(x^*,\xi) \neq 0$

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

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

L-гладкость $f(\cdot,\xi)$

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

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

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

$$= \mathbb{E}\Big[\big\langle \underbrace{\mathbb{E}[\nabla f(x^*, \xi^k) \mid x^k]}_{\nabla f(x^*) = 0};\ x^k - x^* \big\rangle \Big] = 0$$

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

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

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

Подставим $(**)$ & $(*)$

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

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

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

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

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

Подставим (****) в (***)

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

μ - сильной выпуклости

$$
\begin{aligned}
\mathbb{E}\left[\|x^{k+1}-x^*\|_2^2\right] &\le \mathbb{E}\left[\|x^k-x^*\|_2^2\right] \\
&\quad - 2\gamma\, \mathbb{E}\left[f(x^k)-f(x^*)+\frac{\mu}{2}\|x^k-x^*\|_2^2\right] \\
&\quad + 4L\gamma^2 \mathbb{E}\left[f(x^k)-f(x^*)\right] + 2\gamma^2\sigma_*^2 \\
&\le (1-\gamma\mu)\,\mathbb{E}\left[\|x^k-x^*\|_2^2\right] \\
&\quad - 2\gamma\underbrace{(1-2\gamma L)}_{\geq 0}\,\mathbb{E}\underbrace{\left[f(x^k)-f(x^*)\right]}_{\geq 0} \\
&\quad + 2\gamma^2\sigma_*^2
\end{aligned}
$$

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

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

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

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

Запустим рекурсию:

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

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

. . . . .

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

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

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

Сходимость SGD с постоянным шагом

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

⊕ сходимость линейная, как у GD

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

⊕ простота + хорошие ML-метрики

Как бороться со сходимостью к окрестности?

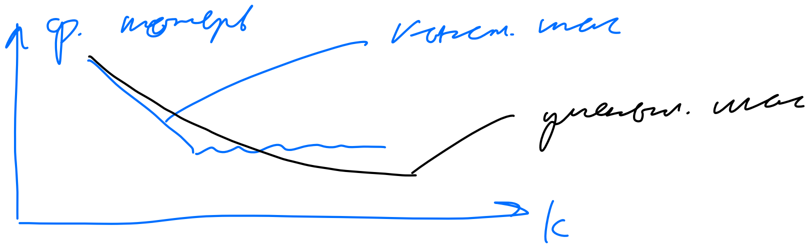
1) γ уменьшаем:

⊕ сходимость лучше

⊖ медленнее

$$\gamma \to \gamma_k \sim \frac{1}{k}\,;\ \frac{1}{\sqrt{k}}$$

без линейной сход.



2) σ_*^2 уменьшить

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

S – батч (набор объектов и меток) размера b

$$\sigma_*^2 \longrightarrow \frac{\sigma_*^2}{b}$$

эффект батчирования

⊕ окрестность уменьшается

⊖ стоимость вычислений растет