Задача стохастической оптимизации

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

Пример: в ML

лосс, модель

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

(ξ_a,ξ_b) — природа (неизв.) — веса — объект — метка

$\Downarrow$

Вычислить f, ∇f невозможно

Что делать?

1) Онлайн подход:

$$\nabla f(x,\xi) = \nabla_x l\left(g(x,\xi_a),\xi_b\right) \leftarrow \text{градиент по опред. сэмплу}$$

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

2) Оффлайн подход:

есть выборка $\{\xi_i\}_{i=1}^n$

Аппроксимация $\mathbb{E}_{\xi\sim D}$ по Монте-Карло

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

это другая задача: $\bar f \approx f$ при большом n

можно считать $\nabla \bar f$:

⊖ дорого

⊖ переобучение ($\bar f \neq f$)

⊕ вычислять не полный град., а град. по части выборки — батч

Стохастический градиентный спуск

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

по 1ому объекту выборки

ξ^k – независимо и равномерно

- Независимость:

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

- Равномерность:

$$\mathbb{E}_\xi [\nabla f(x^k, \xi^k)] = \text{(определение мат. ожидания)}$$

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

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

Условное математическое ожидание

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

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

фиксируем все, что происходило до x^k (включительно)

tower property:

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

Предположения

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

Док-во сходимости:

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

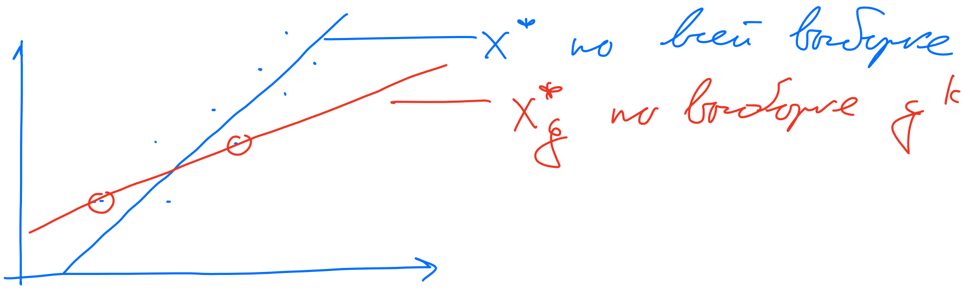
Полное м.о. от обеих частей:

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

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

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

$$
\nabla f(x^*) = 0 \not\Rightarrow \nabla f(x^*,\xi^k) = 0
$$



КБШ

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

L - гладкость $f(\ ,\xi)$

$$\leq 4L\,\mathbb{E}\Big[f(x^k,\xi^k) - f(x^*,\xi^k) + \underbrace{\langle \nabla f(x^*,\xi^k); x^* - x^k\rangle}_{\neq 0}\Big]$$
$$+ 2\,\mathbb{E}\big[\|\nabla f(x^*,\xi^k)\|_2^2\big] \ominus$$

Tower property: $\mathbb{E}\big[\mathbb{E}[\ \mid x^k]\big] = \mathbb{E}[\ ]$

$$\mathbb{E}\big[\mathbb{E}[f(x^k,\xi^k)\mid x^k]\big] = \mathbb{E}[f(x^k)]$$

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

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

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

$$= \mathbb{E}\big[\langle \overset{0}{\cancel{\nabla f(x^*)}}; x^* - x^k\rangle\big] = 0$$

соединяем

$$\ominus\ 4L\,\mathbb{E}[f(x^k) - f(x^*)] + 2\,\mathbb{E}\big[\|\nabla f(x^*,\xi^k)\|_2^2\big]$$

$$\mathbb{E}\big[\|\nabla f(x^*,\xi^k)\|_2^2\big] = \text{равномерность}$$

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

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

Подставим (**) в (*)

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

(***)

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

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

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-сильная выпуклость для f

$$\mathbb{E}\left[\|x^{k+1}-x^*\|_2^2\right] \leq \mathbb{E}\left[\|x^k-x^*\|_2^2\right]-2\gamma\mathbb{E}\left[\frac{\mu}{2}\|x^k-x^*\|_2^2+f(x^k)-f(x^*)\right]$$
$$+\gamma^2\cdot\left(4L\,\mathbb{E}\left[f(x^k)-f(x^*)\right]+2\sigma_*^2\right)$$
$$=(1-\gamma\mu)\mathbb{E}\left[\|x^k-x^*\|_2^2\right]$$
$$-2\gamma(1-2\gamma L)\underbrace{\mathbb{E}\left[f(x^k)-f(x^*)\right]}_{\geq 0}$$
$$+2\gamma^2\sigma_*^2$$

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

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

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

то же самое было в GD, но без $\mathbb{E}$ и без

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

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

Заметим рекурсию:

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

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

$$\ldots$$

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

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

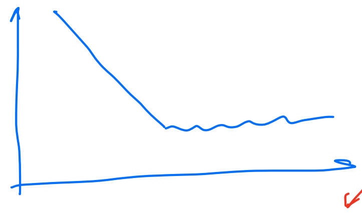
$$\le (1-\gamma\mu)^{k+1} R_0^2 + \frac{2\gamma\sigma_*^2}{\mu} \qquad \blacksquare$$

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

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



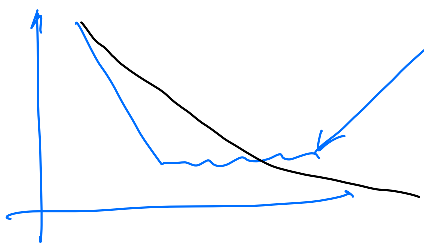
влияет

⊕ простота + ML-метрики

Как бороться со сходимостью к окрестности?

1) γ уменьшить : ⊕ сходимость глубже
⊖ медленнее

$\gamma \to \gamma_k \sim \frac{1}{k}\ ;\ \frac{1}{\sqrt{k}}$: ⊕ сходимость до решения
⊖ вообще пропадает лин. сход.



2) σ_*^2 уменьшить

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

S^k – батч, набор объектов из об. выборки, размера b
(все объекты берутся независимо и равномерно)

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

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

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

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

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

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

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

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

ξ, η — независимые

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

$$= \boxed{\frac{\sigma_*^2}{b}!}$$ ← эффект батчирования

(+) скорость уменьшилась в b раз

(−) стоимость растет

На практике:

- $\nabla f_1 \to \nabla f_2 \to \nabla f_3 \dots \to \nabla f_n$

  не стох. метод.

- замиксовать данные 1 раз $\to$ $\nabla f_1 \dots \nabla f_n$ более равномерно
- миксовать каждую эпоху $\to$ еще меньше смещения

техники shuffling