$$\min_{x \in \mathbb{R}^d} f(x)$$

$$\Downarrow$$

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

<u>Пример:</u>

ф. потерь　　　　　　　　объект　　метка

$$\min_{x \in \mathbb{R}^d} \mathbb{E}_{\xi \sim D}\left[l\left(g(x, \xi_a), \xi_b\right)\right] \qquad (1)$$

(ξ_a, ξ_b)　　　　модель　　　веса модели

не знаем

Можно считать не $\nabla f(x)$, а $\nabla_x f(x, \xi)$

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

<u>Предположение:</u>

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

Часто имеем дело с <u>оффлайн</u> постановкой:

$$\min_{x \in \mathbb{R}^d} \left[\frac{1}{n} \sum_{i=1}^{n} l\left(g(x, \xi_a^i), \xi_b^i\right)\right] \qquad (2)$$

у нас есть n объектов
с известн. ответами

(2) Монте-Карло приближение (1)

для (2) считать $\nabla f(x)$, но не хочу

- дорого
- добавить случайности/колебаний для того, чтобы быть ближе к (1)

Поэтому будем использовать $\nabla f(x, \xi^i)$

Если i выбирается равномерно, то

$$\mathbb{E}_i\left[\nabla f(x, \xi^i)\right] = \sum_{j=1}^{n} \underbrace{\mathbb{P}\{i=j\}}_{=\frac{1}{n}} \nabla_x f(x, \xi^j)$$

$$= \frac{1}{n}\sum_{j=1}^{n} \nabla_x \left[\ell(g(x, \xi_a^j), \xi_b^j)\right]$$

$$= \nabla_x \left[\frac{1}{n}\sum_{j=1}^{n} \ell(g(x, \xi_a^j), \xi_b^j)\right]$$

$$= \nabla f(x)$$

Метод стох градиента

Алгоритм 1 Стохастический градиентный спуск (SGD)

Вход: размеры шагов $\{\gamma_k\}_{k=0} > 0$, стартовая точка $x^0 \in \mathbb{R}^d$, количество итераций K

1: **for** $k = 0, 1, \ldots, K-1$ **do**

2: Сгенерировать независимо ξ^k

3: Вычислить стохастический градиент $\nabla f(x^k, \xi^k)$

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

5: **end for**

Выход: x^K

несмещ $\nabla f(x)$

Справка

$$\mathbb{E}\left[\,\cdot\,|\,x^k\right] = \mathbb{E}\left[\,\cdot\,|\,\mathcal{F}_k\right]$$

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

Суть — „фиксирует" всю случайность до текущей итерации

tower property

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

<u>Док-во сходимости:</u>

- f — L-гладкой, μ-сильно-выпуклой
- предположение на случайность

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

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

Док-во:

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

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

$$\mathbb{E}[\,\cdot \mid x^k]$$

это не с. в. если $\mathbb{E}[\;\mid x^k]$

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

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

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

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

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

$$= \langle \nabla f(x^k); x^k - x^* \rangle$$

(по предпол. о несмещ. $\nabla f(x,\xi)$ и нез. ξ^k)

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

$$-2\gamma \langle \nabla f(x^k); x^k - x^* \rangle \quad +$$

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

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

$$\mathbb{E}\left[\|\nabla f(x^k,\xi^k) - \nabla f(x^k)\|_2^2 + \|\nabla f(x^k)\|_2^2 \cancel{- 2\langle \nabla f(x^k); \nabla f(x^k,\xi^k) - \nabla f(x^k)\rangle} \mid x^k\right] \to 0$$

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

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

(по 2 предположению)

$$\leq \sigma^2 + \|\nabla f(x^k)\|_2^2$$

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

$$-2\gamma \langle \nabla f(x^k); x^k - x^* \rangle \quad +$$

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

$$+$$

μ - сильная вып. L - гладкость

$$\leq \|x^k - x^*\|_2^2$$
$$- \mu\gamma \|x^k - x^*\|_2^2 - 2\gamma (f(x^k) - f(x^*))$$
$$+ \gamma^2 \cdot 2L (f(x^k) - f(x^*)) + \gamma^2\sigma^2$$

$$= (1 - \gamma\mu) \|x^k - x^*\|_2^2$$
$$- \gamma (1 - \gamma L) (f(x^k) - f(x^*))$$
$$+ \gamma^2 \sigma^2$$

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

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

$\mathbb{E}$ полное $\quad \mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$

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

Теорема сходимость SGD в случае ограниченной дисперсии

Пусть задача безусловной стохастической оптимизации с L-гладкой, μ-сильно выпуклой целевой функцией f решается с помощью SGD с $\gamma_k \leq \frac{1}{L}$ в условиях несмещенности и ограниченности дисперсии стохастического градиента. Тогда справедлива следующая оценка сходимости

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

Запустим рекурсию

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

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

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

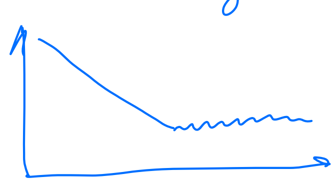
$$\ldots$$

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

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

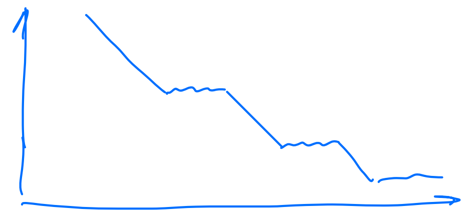
$$\mathbb{E}\left[\|x^{k+1}-x^*\|_2^2\right] \le \underbrace{(1-\gamma\mu)^{k+1}\,\mathbb{E}\left[\|x^0-x^*\|_2^2\right]}_{\text{линейная сходимость к решению}}+\underbrace{\frac{\gamma\sigma^2}{\mu}}_{\text{окрестность}}$$

Характерная сходимость:

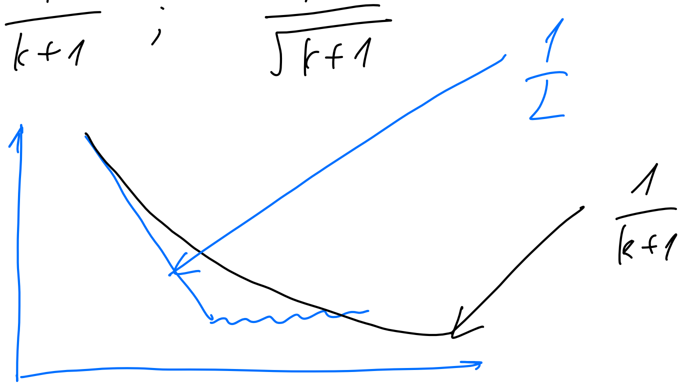


Борьба за окрестность:

- рестарт метода с меньшим шагом



- $\gamma^k = \frac{1}{k+1}$; $\frac{1}{\sqrt{k+1}}$



- не так, не σ^2

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

$|S^k| = b$

батч, набор из нескольких ξ

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

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

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

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

0 (ξ, η - независимы)

$$\leq \underbrace{\frac{1}{b^2} \sum_{\xi \in S^k} \sigma^2}_{\frac{\sigma^2 \cdot b}{b^2} = \frac{\sigma^2}{b}}$$

Pytorch: HB Polyak $\Rightarrow$ Stoch HB Polyak

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

$$x^k \approx x^{k-1}$$

есть эффект батчирования (доп. для стох. HB)

для метода Нестерова:

$$\mathbb{E}\left[\|x^{k+1} - x^*\|_2^2\right] \leq \underbrace{\left(1 - \sqrt{\frac{\mu}{L}}\right)^{k+1}}_{\text{ускорение}} \mathbb{E}\left[\|x^0 - x^*\|_2^2\right] + \underbrace{\frac{\sigma^2}{\mu^2 k}}_{\text{сублинейн. сход. из-за стох.}}$$

для град. спуска

$$\mathbb{E}\left[\|x^{k+1} - x^*\|_2^2\right] \leq \left(1 - \frac{\mu}{L}\right)^{k+1} \mathbb{E}\left[\|x^0 - x^*\|_2^2\right] + \frac{\sigma^2}{\mu^2 k}$$