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

$$\Downarrow$$

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

<u>Пример</u>

ф. потерь

веса модели

метка

(*) $$\min_{x\in\mathbb{R}^d} \mathbb{E}_{(a,b)\sim\mathcal{D}}\left[\,l(g(x,a);b)\,\right]$$

неизвестно (задача ML стр. D с помощью g)

модель

объекты обуч. выборки

Проблема: D не знаем, не можем f, ∇f

Можем семплировать из D

- $\xi = (a,b)$ приходит в режиме <u>онлайн</u>

$$\nabla f(x,\xi) = \nabla_x\left[\,l(g(x,a),b)\,\right]$$

предположение:

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

размер выборки

- классика - <u>оффлайн</u> дата: $\{a_i, b_i\}_{i=1}^n$ из D

(**) $$\min_{x\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n l(g(x,a_i),b_i)$$ ← ERM (мин. эмп. риска)

(∗∗) с ростом n приближает (∗)

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

(Монте-Карло апр.)

Даже имея всю выборку град не считается

- долго и дорого
- обобщ способность и стохастика

$$\nabla f(x, \xi) = \nabla f_{\xi_i}(x) = \nabla_x \left[l(g(x, a_i), b_i) \right]$$

равн. из [1..n]

(выбираем один из 1...n объектов)

$$\mathbb{E}_\xi \left[\nabla f(x, \xi) \right] = \mathbb{E}_\xi \left[\nabla_x l(g(x, a_\xi), b_\xi) \right]$$

$$= \sum_{i=1}^{n} \underset{\substack{\| \\ p_i}}{\frac{1}{n}} \nabla_x l(g(x, a_i); b_i)$$

ξ дискр. с.в. [1..n]

$p = \frac{1}{n}$

$$= \nabla f(x)$$

град ERM

Метод со стох. град.

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

Усл. м.о.

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

σ-алгебра, порожд. $x^0;\ \xi^0, \xi^1 \dots \xi^{k-1}$

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

Сходимость:

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

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

$\gamma_k = \gamma$

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

с.в.

$\mathbb{E}[\,\cdot\,|x^k]$ с.в. с.в.

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

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

$$\mathbb{E}\left[\nabla f(x^k, \xi^k) \mid x^k\right] \underset{\uparrow}{=} \mathbb{E}_{\xi^k}\left[\nabla f(x^k, \xi^k)\right] = \nabla f(x^k)$$

нез. ξ^k от x^k

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

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

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

$$\underset{\uparrow}{\leq} \sigma^2 + \|\nabla f(x^k)\|_2^2$$

об огр. дисп.

$$\mathbb{E}\left[\|x^{k+1} - x^*\|_2^2 \mid x^k\right] \leq \|x^k - x^*\|_2^2 - 2\gamma\langle \nabla f(x^k); x^k - x^*\rangle + \gamma^2\|\nabla f(x^k)\|_2^2 + \underbrace{\gamma^2\sigma^2}_{\text{новое}}$$

L – гладкость, μ – сильн. вып. f

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

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

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

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

$\mathbb{E}[\mathbb{E}[X|Y]]$ от условного к полному

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

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 $\gamma_k \le \frac{1}{L}$ в условиях несмещенности и ограниченности дисперсии стохастического градиента. Тогда справедлива следующая оценка сходимости

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

Запускаем рекурсию $\gamma_k = \gamma$

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

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

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

$$\ldots$$

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

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

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

$$= \underbrace{(1-\gamma\mu)^k}_{\text{сход. град. спуска}} R_0 + \underbrace{\frac{\gamma\sigma^2}{\mu}}_{\text{окрестность сходимости}}$$

Борьба с огр. сходимости:

- $\gamma^k = \frac{1}{k+1}$ $\quad \gamma^k = \frac{1}{\sqrt{k+1}}$
- уменьшить σ^2

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

где S^k — нез. набор индексов размера b

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

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

$$+ \sum_{\substack{\xi \ne \eta \\ \in S}} \mathbb{E} \langle \nabla f(x, \xi) - \nabla f(x); \nabla f(x, \eta) - \nabla f(x) \rangle \Big)$$

$\mathbb{E}_\xi$ $\quad \mathbb{E}_\eta$

$$\le \frac{1}{b^2} \left(b \cdot \sigma^2 + 0 \right) = \boxed{\frac{\sigma^2}{b}}$$

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

можем добиться вот такого с некоторым подбором γ^k:

1) $\gamma^k = \frac{1}{L}$ $\quad$ 2) $\gamma^k \sim \frac{1}{k}$

$$R_k \le (1-\gamma\mu)^k R_0 + \frac{\sigma^2 \gamma}{\mu b}$$

$$\gamma = \min\left(\frac{1}{L}; \frac{1}{\mu k}\ln(\ldots)\right)$$

$$\le \left(1-\frac{\mu}{L}\right)^k R_0 + \exp\left(-\cancel{\frac{\mu k}{\mu k}} \ln(\ldots)\right) R_0$$

$$+ \frac{\sigma^2}{\mu^2 b k}$$

$\frac{\sigma^2}{\mu^2 b k R_0}$

Accel SGD:
(Nesterov)

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

ускорение

ускорение сход. к нулю

при фикс. γ — сходимость к окр.:

у град. спуска

$\gamma \nabla f(x^k)$

$x^k \to x^*$

$\nabla f(x^k) \to 0$

$\gamma \nabla f(x^k) \to 0$

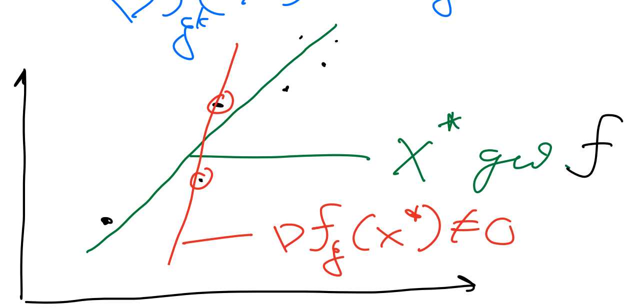
метод остан.

у SGD:

$\gamma \nabla f_{\xi^k}(x^k)$

$x^k \to x^*$

$\nabla f_{\xi^k}(x^k) \to \nabla f_{\xi^k}(x^*) \ne 0$



„Другой" стох. градиент:

$$x^{k+1} = x^k - \gamma g^k$$ „стох град"

по $g^k \to \nabla f(x^*) = 0$ $\quad x^k \to x^*$

по возможности

$$\mathbb{E}[g^k | x^k] = \nabla f(x^k)$$

в онлайн постановке невозможно

а вот оффлайн:

$$f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x)$$

- y_i^k – память: $y_i^0 = 0$ можно
- кладем в $y_{\xi^k}^k = \nabla f_{\xi^k}(x^k)$
  ост. y_i^k не меняем
- $\frac{1}{n} \sum_{i=1}^{n} y_i^k$ — запаздыв. градиент $\approx \nabla f(x^k)$

Алгоритм 2 SAGA

Вход: размер шага $\gamma > 0$, стартовая точка $x^0 \in \mathbb{R}^d$, значения памяти $y_i^0 = 0$ для всех $i \in [n]$, количество итераций K

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

2: Сгенерировать независимо i_k

3: Вычислить $g^k = \nabla f_{i_k}(x^k) - y_{i_k}^k + \frac{1}{n} \sum_{j=1}^{n} y_j^k$

4: Обновить $y_i^{k+1} = \begin{cases} \nabla f_i(x^k), & \text{если } i = i_k \\ y_i^k, & \text{иначе} \end{cases}$

5: $x^{k+1} = x^k - \gamma g^k$

6: **end for**

Выход: x^K
