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

$$f(x) = \frac{1}{2}\|Ax - b\|_2^2$$

A – данные (объекты)
b – метки

Стохастика по объектам

$$f(x) = \frac{1}{2n}\sum_{i=1}^{n}(a_i^T x - b_i)^2 = \frac{1}{n}\sum f_i(x)$$

$$\nabla f_i(x) = a_i^T(a_i x - b_i)$$

выбираем i – номер объекта

$$\nabla f_i(x) = \begin{pmatrix} \text{все компоненты ненулевые} \end{pmatrix}$$

Стохастика по признакам

$$f(x) = \frac{1}{2n}\sum_{i=1}^{n}\left(\sum_{j=1}^{d} a_{ij}x_j - b_i\right)^2$$

продифференцируем по j признаку (по переменной x_j)

$$\nabla f_{(j)}(x) = \frac{1}{n}\sum_{i=1}^{n} a_{ij}\left(\sum_{l=1}^{d} a_{il}x_l - b_i\right)$$

выбираем j – номер коорд.

$$\nabla f_{(j)}(x) = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ j \text{ коорд} \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

полный градиент $\nabla f(x) = \underbrace{A^T A}\, x - A^T b$

- $d^2 n + nd$ единоразово
- $d^2 + d$ многоразово

$$\nabla f_i(x) = a_i^T a_i x - a_i^T b$$

$n^2 d$ и nd единоразово

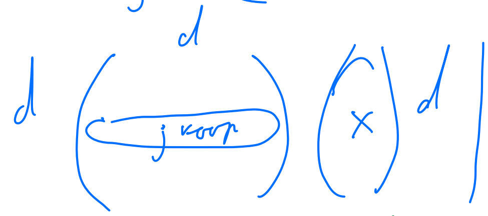
n раз дешевле вычисление на каждой итерации

$$\nabla f_{(j)}(x) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d} a_{ij}(a_{ij}x_j - b_i)$$

$n^2 d$ и nd единоразово

$$\nabla f = A^T A x - A^T b$$

$$\nabla f_{(j)} = (A^T A x - A^T b)_{(j)}$$



$d+1$ многоразово $\Rightarrow$ d раз дешевле

Мотивация №2:

нет градиента

$$\nabla f_{(j)}(x) = \frac{f(x+\tau e_j) - f(x-\tau e_j)}{2\tau}$$

SGD:

$$x^{k+1} = x^k - \gamma \nabla f_{i_k}(x^k)$$

случ. выбор объек.

Коорд. спуск

$$x^{k+1} = x^k - \gamma \nabla f_{(j_k)}(x^k)$$

случ. коорд.

Сходимость:

- f – L-гладкость
- f – μ- сильно выпуклая

Если хотим $\|x^k - x^*\|_2^2 \sim \varepsilon$

$$k = \frac{dL}{\mu} \log \frac{\|x^0 - x^*\|_2^2}{\varepsilon}$$

В градиентном спуске $\frac{L}{\mu} \log \frac{1}{\varepsilon}$

SVRG

$$k = O\left(\left(n + \frac{L}{\mu}\right)\left(\log \frac{1}{\varepsilon}\right)\right)$$

при этом итерация дешевле в n раз

SVRG в теории в $O(n)$ раз дешевле

Коорд

$$k = \frac{dL}{\mu} \log \frac{1}{\varepsilon}$$

итерация в d раз дешевле

Коорд. метод – не дешевле

Пример: $f(x) = x_1^2 + (x_2-1)^2 + \dots (x_i - i + 1)^2 + \dots$

1

коорд сепарабельны

$$\nabla f_{(j)}(x) \approx \begin{pmatrix} 0 \\ \vdots \\ \frac{f(x+\tau e_j) - f(x-\tau e_j)}{2\tau} \\ \\ 0 \end{pmatrix}$$

e_j — j базисный вектор

На сколько близка?

$$\| \nabla f_{(j)}(x) \|^2 =$$

$$= \left\| \langle \nabla f(x); e_j \rangle e_j - \frac{f(x+\tau e_j) - f(x-\tau e_j)}{2\tau} e_j \right\|_2^2$$

$$= \left| \langle \nabla f(x); e_j \rangle - \frac{f(x+\tau e_j) - f(x-\tau e_j)}{2\tau} \right|^2$$

$$= \frac{1}{(2\tau)^2} \left| \langle \nabla f(x); \underset{\pm x}{2\tau e_j} \rangle - f(x+\tau e_j) + f(x-\tau e_j) \right|^2$$

$$= \frac{1}{(2\tau)^2} \Big| \underbrace{\langle \nabla f(x); x + \tau e_j - (x - \tau e_j) \rangle - f(x+\tau e_j) + f(x-\tau e_j)}_{L\text{-гладка}} \Big|^2$$

$$\leq \frac{1}{(2\tau)^2} \left| \frac{L}{2} \| x + \tau e_j - (x - \tau e_j) \|_2^2 \right|^2$$

$$\le \frac{1}{4\tau^2}\cdot\frac{L^2}{4}(2\tau)^4 = L^2\tau^2$$

чем меньше τ тем лучше?

модель ошибки

$$\tilde f(x) = f(x) + \delta(x) \qquad |\delta(x)| < \Delta$$

$$\left\| \nabla f_j(x) - \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \dfrac{\tilde f(x+\tau e_j) - \tilde f(x-\tau e_j)}{2\tau} \\ 0 \\ \vdots \\ 0 \end{pmatrix} \right\|_2^2 \sim \underbrace{\tau^2 L^2 + \frac{\Delta^2}{\tau^2}}_{\text{борьба}}$$

Ошибка не стохастическая

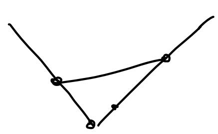
$\Downarrow$

не устранить

$\Downarrow$

сходимся окрестности $\tau^2L^2 + \dfrac{\Delta^2}{\tau^2}$

Негладкий случай



$$g(x) = \frac{f(x+\tau e) - f(x-\tau e)}{2\tau} e$$

стох градиент

$$\mathbb{E}_e[g(x)] \overset{?}{=} \nabla f(x)$$

$$\mathbb{E}_e[g(x)] = \frac{2\,\mathbb{E}[f(x+\tau e)e]}{2\tau} =$$

$$= \frac{\mathbb{E}[f(x+\tau e)e]}{\tau} \neq \nabla f(x)$$

Теорема

$$\mathbb{E}_e[g(x)] = \frac{\nabla f_\tau(x)}{d} \qquad f_\tau(x) = \mathbb{E}_u\left[f(x+\tau u)\right]$$

$e \in US_2^d(1)$ равномерно с евклидовой сферы

$u \in UB_2^d(1)$ равномерно с шара

$g(x)$ — это несмещ. стох. градиент, но для f_τ

- f_τ — гладкая, для негладкой f

  $\frac{\sqrt{d}\,M}{\tau}$ M-липшицева

- $f(x) \le f_\tau(x) \le f(x) + \tau M \quad \forall x \in \mathbb{R}^d$

  при маленьком τ мы миним. ту же функцию