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

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

A - данные (объекты)
b - лейблы

Стохастика по объектам

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

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

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

выбираем i - номер сэмплов

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

Стохастика по признакам

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

производная по j признаку (x_j)

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

выбираем j - номер коорд.

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

полный градиент $\nabla f(x)=\underbrace{A^TA}_{d^2n}x-\underbrace{A^Tb}$

- dn^2+nd одн. раз.
- d^2+d многоразово

$$\nabla f_i(x)=a_i^Ta_i x-a_i^Tb$$

n^2d и nd одноразово

- $\frac{d^2}{n}+\frac{d}{n}$ многоразовых (n раз дешевле)

n раз дешевле

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

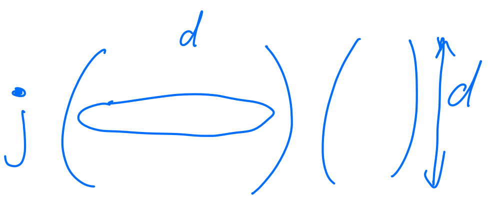
$n^2d + nd$ одноразово

- $d+1$ многоразовых операций

$$\nabla f(x)=\underbrace{A^TA}x-A^Tb$$

d раз дешевле

$\nabla f_{(j)}=$



Мотивация №2:

нет градиента

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

SGD:

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

случайный батч

Координатный спуск:

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

случайная координата

Док-во сходимости:

- f − L-гладкая
- f − μ-сильно выпуклая

$$\|x^{k+1} - x^*\|_2^2 = \|x^k - \gamma \nabla f_{(j_k)}(x^k) - x^*\|_2^2$$

$$= \|x^k - x^*\|_2^2 - 2\gamma \langle \nabla f_{(j_k)}(x^k); x^k - x^* \rangle + \gamma^2 \|\nabla f_{(j_k)}(x^k)\|_2^2$$

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

$\mathbb{E}[\;\;|x^k] = \mathbb{E}_k$ tower property

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

j_k – независимы и равномерно

$$\mathbb{E}_k\left[\nabla f_{(j_k)}(x^k)\right] = \sum_{j=1}^{d} \overset{=p}{\frac{1}{d}} \nabla f_{(j)}(x^k)$$

$$= \frac{1}{d} \underbrace{\boxed{\sum_{j=1}^{d} \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \nabla f_{(j)}(x^k) \\ 0 \\ \vdots \\ 0 \end{pmatrix}}}_{=\nabla f(x^k)} = \frac{1}{d} \nabla f(x^k)$$

$$\mathbb{E}_k\left[\|\nabla f_{(j_k)}(x^k)\|_2^2\right] = \mathbb{E}_k\left[\sum_{i=1}^{d} \nabla f^2_{(j_k),(i)}(x^k)\right]$$

$$\nabla f_{(j_k)} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \nabla f_{(j_k)}(x^k) \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

i ая коорд.

$$= \sum_{i=1}^{d} \mathbb{E}_k\left[\nabla f^2_{(j_k),(i)}(x^k)\right]$$

$$= \sum_{i=1}^{d} \frac{1}{d} \nabla f^2_{(i)}(x^k) + \left(1 - \frac{1}{d}\right) \cdot 0$$

$$= \frac{1}{d} \underbrace{\boxed{\sum_{i=1}^{d} \nabla f^2_{(i)}(x^k)}}_{=\|\nabla f(x^k)\|_2^2}$$

$$= \frac{1}{d} \|\nabla f(x^k)\|_2^2$$

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

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

L - гладкость $\quad \|\nabla f(x^k)\|_2^2 = \|\nabla f(x^k) - \nabla f(x^*)\|_2^2 \le 2L\left(f(x^k) - f(x^*)\right)$

μ - сильная выпуклость $\quad -\langle \nabla f(x^k); x^k - x^* \rangle \le -\frac{\mu}{2}\|x^k - x^*\|_2^2 + f(x^*) - f(x^k)$

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

$$+ \underbrace{\frac{2\gamma}{d}}_{>0}\underbrace{(\gamma L - 1)}_{\le 0?}\underbrace{\left(f(x^k) - f(x^*)\right)}_{\ge 0}$$

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

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

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

$$\le \left(1 - \frac{\gamma\mu}{d}\right)^{k+1}\|x^0 - x^*\|_2^2$$

$(1 - x) \le \exp(-x) \qquad 1 - x + \frac{x^2}{2} \qquad x \le 1$

$$\le \exp\left(-\frac{\gamma\mu k}{d}\right)\|x^0 - x^*\|_2^2$$

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

$$= \exp\left(-\frac{\mu k}{L d}\right)\|x^0 - x^*\|_2^2$$

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

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

В градиентном спуске было $k = \frac{L}{\mu} \log \dots$

SVRG гарант.

$$k = O\left(\left(n + \frac{L}{\mu}\right) \log \frac{\|x^0 - x^*\|_2^2}{\varepsilon}\right)$$

при этом итерация дешевле n раз

$\boxed{\text{SVRG – теор. дешевле в } O(n) \text{ раз.}}$

Координатный метод

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

итерация дешевле в d раз

$\boxed{\text{Коорд. метод – не дешевле}}$

Пример почему выигрываем d

$$f(x) = x_1^2 + (x_2 - 1)^2 + (x_3 - 2)^2 + \dots$$

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

$$\|\nabla f_{(j)}(x) - \qquad \|_2^2 =$$

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

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

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

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

$$\le \left(\frac{1}{2\tau}\right)^2 \cdot \left| \frac{L}{2} \| x+\tau e_j - (x-\tau e_j) \|^2 \right|^2 =$$

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

- Модель ошибки

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

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

Ошибка не стохастическая

$\Downarrow$

не усредняется

$\Downarrow$

сходимость к окрестности $\boxed{\tau^2 L^2 + \frac{\Delta^2}{\tau^2}}$