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

$$f(x) = \frac{1}{2} x^T \underset{\substack{\text{симм} \\ \mathbb{R}^{d\times d}}}{A} x - b^T x \qquad \nabla f(x) = \underbrace{\underbrace{\underbrace{Ax}_{d^2 \text{ операций}} - b}_{+d \text{ операций}}}_{O(d^2)}$$

$$f(x) = \frac{1}{2n} \sum_{i=1}^{n} \left(a_i^T x - b_i\right)^2 \leftarrow \text{лин. регрессия}$$

SGD (выбора объекта)

$$\nabla f_i(x) = \nabla\left(\frac{1}{2}\left(a_i^T x - b_i\right)^2\right) = \left(a_i^T x - b_i\right) a_i$$

градиент $\nabla f_i(x)$ дешевле в n раз по сравнению с $\nabla f(x)$

кол-во данных

Координатный (выбор признаков)

$$\nabla f_{(j)}(x) = \left(\frac{1}{2n} \sum_{i=1}^{n} \left(a_{i1} x_1 + \ldots + a_{ij} x_j + \ldots - b_i\right)^2\right)'_j = \frac{1}{n} \sum_{i=1}^{n} \left(a_i^T x - b_i\right) a_{ij}$$

градиент / произв. $\nabla f_{(j)}(x)$ дешевле в d раз по сравнению с $\nabla f(x)$

SGD (выбор объекта)

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

i_k — случ. индекс от 1 до n

Координатный спуск

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

j_k — случ. индекс коорд. от 1 до d

$$\nabla f_{(j_k)}(x^k) = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ f'_{(j_k)}(x^k) \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \langle \nabla f(x^k); e_{j_k} \rangle e_{j_k}$$

Док-во сходимости:

Предположения:

- f — μ-сильно выпукла
- f — L-гладкая

Док-во:

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

$\mathbb{E}[\nabla f_{(j_k)}(x^k); x^k-x^*\rangle]$ + tower property

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

$$\mathbb{E}\left[\nabla f_{(j_k)}(x^k) \mid x^k\right] = \sum_{i=1}^{d} \underbrace{\mathbb{P}\{j_k=i\}}_{\frac{1}{d}} \nabla f_{(i)}(x^k)$$

$$= \frac{1}{d}\underbrace{\sum_{i=1}^{d} \nabla f_{(i)}(x^k)}_{\nabla f(x^k)}$$

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

$$\ominus\ \mathbb{E}\left[\frac{1}{d}\langle \nabla f(x^k); x^k-x^*\rangle\right] \qquad (**)$$

Подставим (**) в (*)

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

$\mathbb{E}\left[\|\nabla f_{(j_k)}(x^k)\|_2^2\right]$ + tower property

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

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

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

$$= \frac{1}{d} \mathbb{E}\left[\|\nabla f(x^k)\|_2^2\right] \qquad (****)$$

Подставляем $(****)$ в $(***)$

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

L – гладкость, μ-сильная выпуклость

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

$$= \left(1 - \frac{\gamma\mu}{d}\right) \mathbb{E}\left[\|x^k-x^*\|_2^2\right] - \frac{2\gamma}{d}(1-\gamma L)\, \mathbb{E}\left[\underbrace{f(x^k) - f(x^*)}_{\ge 0}\right]$$

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

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

$\gamma = \frac{1}{L}$ и запускаем рекурсию

Сходимость координатного спуска

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

для достижения точности ε нужно

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

⊕ итерация спуска в d раз дешевле (для квадр. задач) по ср. с GD

⊖ кол-во итераций возросло в d раз

$$\min_{x \in \mathbb{R}^d} \left[x_1^2 + x_2^2 + \ldots + x_d^2\right]$$

SAGA

$$y_i^k = \begin{cases} \nabla f_{i_k}(x^k) & i = i_k \\ y_i^{k-1} & i \ne i_k \end{cases}$$

1) $x^{k+1} = x^k - \gamma \cdot \frac{1}{n} \sum_{i=1}^{n} y_i^k$

2) $x^{k+1} = x^k - \gamma \cdot \left(\frac{1}{n} \sum_{i=1}^{n} y_i^{k-1} + \nabla f_{i_k}(x^k) - y_{i_k}^{k-1}\right)$

SAGA

почему так

SEGA

h^k – вектор памяти

$$\boxed{h_{(i)}^k = \begin{cases} \langle \nabla f(x^k); e_{i_k} \rangle & i = i_k \\ h_{(i)}^{k-1} & i \ne i_k \end{cases}}$$

1) $x^{k+1} = x^k - \gamma h^k$

2) $x^{k+1} = x^k - \gamma \cdot \left(h^{k-1} + d\left(\nabla f_{(i_k)}(x^k) - h_{(i_k)}^{k-1}\right)\right)$

несмещ

SEGA

Док-во сходимости:

Предположения:

- f – μ-сильно выпуклая
- f – L-гладкая

Док-во: $x^{k+1} = x^k - \gamma g^k$

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

$\mathbb{E}\left[\langle g^k; x^k - x^*\rangle\right]$ + tower property

$$\mathbb{E}\left[\mathbb{E}\left[\langle g^k; x^k - x^*\rangle \mid x^k\right]\right] =$$

$$\mathbb{E}\left[g^k \mid x^k\right] \boxed{=} \quad 1)\ g^k = h^k$$

$$= \mathbb{E}\left[h^k \mid x^k\right] = \begin{pmatrix} \dots \\ \mathbb{E}\left[h^k_{(i)} \mid x^k\right] \\ \dots \end{pmatrix} = \begin{pmatrix} \dots \\ \frac{1}{d}\langle \nabla f(x^k); e_i\rangle + \left(1 - \frac{1}{d}\right) h^{k-1}_{(i)} \\ \dots \end{pmatrix}$$

$$= \frac{1}{d}\underbrace{\begin{pmatrix} \dots \\ \langle \nabla f(x^k); e_i\rangle \\ \dots \end{pmatrix}}_{\nabla f(x^k)} + \left(1 - \frac{1}{d}\right)\underbrace{\begin{pmatrix} \dots \\ h^{k-1}_{(i)} \\ \dots \end{pmatrix}}_{h^{k-1}}$$

$$= \frac{1}{d}\nabla f(x^k) + \left(1 - \frac{1}{d}\right) h^{k-1} \neq \nabla f(x^k)$$

$$\boxed{=} \quad 2)\ g^k = h^{k-1} + d\left(\nabla f_{(i_k)}(x^k) - h^{k-1}_{(i_k)}\right)$$

$$= \mathbb{E}\left[h^{k-1} + d\left(\nabla f_{(i_k)}(x^k) - h^{k-1}_{(i_k)}\right) \mid x^k\right]$$

$$= h^{k-1} + d \cdot \mathbb{E}\Big[\underbrace{\nabla f_{(i_k)}(x^k) - h^{k-1}_{(i_k)}}_{\text{уже было в коорд. методе}} \mid x^k\Big]$$

$$= h^{k-1} + d \cdot \frac{1}{d}\left(\nabla f(x^k) - h^{k-1}\right) = \boxed{\nabla f(x^k)} \qquad (**)$$

Подставляем (**) в (*):

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

μ-сильная выпуклость

$$\mathbb{E}\left[\|x^{k+1}-x^*\|_2^2\right] = \mathbb{E}\left[\|x^k-x^*\|_2^2\right] - 2\gamma\,\mathbb{E}\left[\frac{\mu}{2}\|x^k-x^*\|_2^2 + f(x^k) - f(x^*)\right] + \gamma^2\,\mathbb{E}\left[\|g^k\|_2^2\right] \qquad (***)$$

$\mathbb{E}\left[\|g^k\|_2^2\right]$ + tower property

$$\mathbb{E}\left[\|g^k\|_2^2 \mid x^k\right] = \mathbb{E}\left[\left\| \underline{h^{k-1}} + d\left(\underline{\underline{\nabla f_{(i_k)}(x^k)}} - \underline{h^{k-1}_{(i_k)}}\right)\right\|_2^2 \mid x^k\right]$$

КБШ $\|a+b\|_2^2 \le 2\|a\|_2^2 + 2\|b\|_2^2$

$$\le 2\,\mathbb{E}\left[\| d\,\nabla f_{(i_k)}(x^k)\|_2^2 \mid x^k\right] + 2\,\mathbb{E}\left[\| \underline{h^{k-1}} - d\,h^{k-1}_{(i_k)}\|_2^2 \mid x^k\right]$$

$\mathbb{E}_{i_k}$

$\mathbb{E}\xi^2 \ge \mathbb{D}\xi$

$$\le 2d^2\,\mathbb{E}\left[\|\nabla f_{(i_k)}(x^k)\|_2^2 \mid x^k\right] + 2d^2\,\mathbb{E}\left[\|h^{k-1}_{(i_k)}\|_2^2 \mid x^k\right]$$

см. коорд. спуск

$$= 2d^{\cancel{2}} \cdot \frac{1}{\cancel{d}} \|\nabla f(x^k) - \nabla f(x^*)\|_2^2 + 2d^{\cancel{2}} \cdot \frac{1}{\cancel{d}} \|h^{k-1}\|_2^2$$

L - гладкость

$$\le 4Ld\left(f(x^k) - f(x^*)\right) + 2d\,\|h^{k-1}\|_2^2 \qquad (****)$$

Подставляем $(****)$ в $(***)$

$$\mathbb{E}\left[\|x^{k+1} - x^*\|_2^2\right] = \mathbb{E}\left[\|x^k - x^*\|_2^2\right]$$
$$- 2\gamma\,\mathbb{E}\left[\tfrac{\mu}{2}\|x^k - x^*\|_2^2 + f(x^k) - f(x^*)\right]$$
$$+ \gamma^2\,\mathbb{E}\left[4Ld\left(f(x^k) - f(x^*)\right) + 2d\,\|h^{k-1}\|_2^2\right]$$
$$= (1 - \gamma\mu)\,\mathbb{E}\left[\|x^k - x^*\|_2^2\right]$$
$$- (2\gamma - 4\gamma^2 Ld)\,\mathbb{E}\left[f(x^k) - f(x^*)\right]$$
$$+ 2d\gamma^2\,\mathbb{E}\left[\|h^{k-1}\|_2^2\right]$$

$\mathbb{E}\left[\|h^k\|_2^2\right]$ + tower property

$$\mathbb{E}\left[\|h^k\|_2^2 \mid x^k\right] = \mathbb{E}\left[\|\underline{h^{k-1}} + \underline{\underline{\nabla f_{(i_k)}(x^k)}} - \underline{h^{k-1}_{(i_k)}}\|_2^2 \mid x^k\right]$$

$\|a+b\|_2^2 = \|a\|_2^2 + \|b\|_2^2 + 2\langle a; b\rangle$

$$= \mathbb{E}\left[\|h^{k-1} - h^{k-1}_{(i_k)}\|_2^2 \mid x^k\right] + \mathbb{E}\left[\|\nabla f_{(i_k)}(x^k)\|_2^2 \mid x^k\right]$$
$$+ 2\,\underbrace{\mathbb{E}\left[\langle \nabla f_{(i_k)}(x^k); h^{k-1} - h^{k-1}_{(i_k)}\rangle \mid x^k\right]}_{\langle \nabla f_{(i_k)}(x^k);\, h^{k-1}_{(i_k)} - h^{k-1}_{(i_k)}\rangle = 0}$$

$$= \mathbb{E}\left[\|h^{k-1} - h^{k-1}_{(i_k)}\|_2^2 \mid x^k\right] + \mathbb{E}\left[\|\nabla f_{(i_k)}(x^k)\|_2^2 \mid x^k\right]$$
$$= \|h^{k-1}\|_2^2 + \mathbb{E}\left[\|h^{k-1}_{(i_k)}\|_2^2 \mid x^k\right] - 2\,\mathbb{E}\left[\underbrace{\langle h^{k-1}_{(i_k)}; h^{k-1}\rangle}_{\langle h^{k-1}_{(i_k)};\, h^{k-1}_{(i_k)}\rangle \,=\, \|h^{k-1}_{(i_k)}\|_2^2} \mid x^k\right]$$
$$+ \mathbb{E}\left[\|\nabla f_{(i_k)}(x^k)\|_2^2 \mid x^k\right]$$

$$= \|h^{k-1}\|_2^2 - \mathbb{E}\left[\|h^{k-1}_{(i_k)}\|_2^2 / x^k\right] + \mathbb{E}\left[\|\nabla f_{(i_k)}(x^k)\|_2^2 / x^k\right]$$

как в коорд. методе

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

L- гладкость

$$\leq \left(1-\frac{1}{d}\right)\|h^{k-1}\|_2^2 + \frac{2L}{d}\left(f(x^k) - f(x^*)\right)$$

Итого 2 нер:

$$\mathbb{E}\left[\|x^{k+1}-x^*\|_2^2\right] \leq (1-\gamma\mu)\,\mathbb{E}\left[\|x^k-x^*\|_2^2\right] - (2\gamma - 4\gamma^2 L d)\,\mathbb{E}\left[f(x^k)-f(x^*)\right] + 2d\gamma^2\,\mathbb{E}\left[\|h^{k-1}\|_2^2\right]$$

$$\mathbb{E}\left[\|h^k\|_2^2\right] \leq \left(1-\frac{1}{d}\right)\mathbb{E}\left[\|h^{k-1}\|_2^2\right] + \frac{2L}{d}\,\mathbb{E}\left[f(x^k)-f(x^*)\right]$$

Складываем с множителем M:

$$\mathbb{E}\left[\|x^{k+1}-x^*\|_2^2 + M\|h^k\|_2^2\right]$$
$$\leq (1-\gamma\mu)\,\mathbb{E}\left[\|x^k-x^*\|_2^2\right] - (2\gamma - 4\gamma^2 L d)\,\mathbb{E}\left[f(x^k)-f(x^*)\right] + 2d\gamma^2\,\mathbb{E}\left[\|h^{k-1}\|_2^2\right] + \left(1-\frac{1}{d}\right)\mathbb{E}\left[M\|h^{k-1}\|_2^2\right] + \frac{2LM}{d}\,\mathbb{E}\left[f(x^k)-f(x^*)\right]$$
$$= (1-\gamma\mu)\,\mathbb{E}\left[\|x^k-x^*\|_2^2\right] - \left(2\gamma - 4\gamma^2 L d - \frac{2LM}{d}\right)\mathbb{E}\left[f(x^k)-f(x^*)\right] + \left(1-\frac{1}{d}+\frac{2d\gamma^2}{M}\right)\mathbb{E}\left[M\|h^{k-1}\|_2^2\right]$$

$M: \left(1 - \frac{1}{d} + \frac{2d\gamma^2}{M}\right) = 1 - \frac{1}{2d} \Rightarrow \boxed{M = 4d^2\gamma^2}$

$\gamma: \quad 2\gamma - 4\gamma^2 L d - \frac{2LM}{d} \geq 0$

$2\gamma - 4\gamma^2 L d - 8L\gamma^2 d \geq 0 \qquad 1 - 6\gamma L d \geq 0$

$\Downarrow$

$\boxed{\gamma \leq \frac{1}{6Ld}}$

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

$$\leq \max\left\{(1-\gamma\mu); \left(1-\frac{1}{2d}\right)\right\}\mathbb{E}\left[\|x^k - x^*\|_2^2 + M\|h^{k-1}\|_2^2\right]$$

Сходимость SEGA:

$$\boxed{\mathbb{E}\left[\|x^{k+1} - x^*\|_2^2 + M\|h^k\|_2^2\right] \leq \max\left\{(1-\gamma\mu); \left(1-\frac{1}{2d}\right)\right\}^{k+1}\mathbb{E}\left[\|x^0 - x^*\|_2^2 + M\|h^{-1}\|_2^2\right]}$$

До точности ε $\left(\gamma = \frac{1}{6Ld}\right)$

$O\left(\frac{dL}{\mu}\right)\left(\log\frac{1}{\varepsilon}\right)$ итераций.

⊙ не лучше в теории, чем коорд спуск

⊕ может быть лучше на практике