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

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

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

SGD (выбор объекта)

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

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

Координатный метод

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

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

SGD (батч.)

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

i_k ↑ случайный индекс батча от 1 до n

Координатный спуск

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

j_k ↑ случайный индекс коорд. от 1 до d

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

Док-во сходимости

Предположения:

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

Док-во:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$\gamma = \frac{1}{L}$ и затухание регуляризации

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

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

⊕ итерации дешевле в d раз (для квадратичных задач), чем у GD

⊖ кол-во итераций возросло в d раз

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

КБШ $\|a+b\|_2^2 \le 2\|a\|_2^2 + 2\|b\|_2^2$, где a сост =

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

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

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

см. как в коорд. методе

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

L- гладкость

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

Подставляем $(****)$ в $(***)$

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

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

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

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

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

как в пред. задаче

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

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

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

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

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

$$\ominus\; \mathbb{E}\left[\|h^{k-1} - h^{k-1}_{(i_k)}\|_2^2 \,|\, x^k\right] + \frac{1}{d}\|\nabla f(x^k)\|_2^2$$

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

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

как в координатном

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

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

как в пред. задаче

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

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: (◇)

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

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

Складываем с множителем M:

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

M: $\quad 1-\frac{1}{d}+\frac{2d\gamma^2}{M} = 1-\frac{1}{2d} \Rightarrow \boxed{M = 4d^2\gamma^2}$

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

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

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

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

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

Сходимость SEGA:

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

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

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

⊖ не лучше в теории, чем коорд. спуск

⊕ может быть лучше на практике