$$\min_{x\in\mathbb{R}^d}\left[f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x)\right]$$

- SGD

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

случайно
незав. и равном.

Проблема: сходимость к окрестности

$$x = x^* - \gamma \nabla f_i(x^*)$$

$\neq 0$

$\nabla f(x^*) = 0$

$\nabla f_i(x^*) \neq 0$

уходит из решения

$\nabla f_{i_k}(x^*) \neq 0$

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

заменить на g^k: $g^k \to 0$ $\quad x^k \to x^*$

- SAGA

y_i^k — последовательности „памяти"

$\nabla f_{10}(x^k)$, $\nabla f_{20}(x^{k+1})$ $\quad x^k \approx x^{k+1}$

дорого 2 град. вместо 1

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

1) $x^{k+1} = x^k - \gamma \cdot \underbrace{\frac{1}{n}\sum_{i=1}^{n} y_i^k}_{\approx \nabla f(\text{устаревшей точке})}$

SAGA

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

2)
$$\boxed{\begin{aligned} g^k &= \frac{1}{n} \sum_{i=1}^{n} y_i^{k-1} + \nabla f_{i_k}(x^k) - y_{i_k}^{k-1} \\ x^{k+1} &= x^k - \gamma g^k \end{aligned}}$$

Док-во сходимости:

- f_i – L-гладкие
- f – μ-сильно выпукла

док-во сходимости SGD $x^{k+1} = x^k - \gamma g^k$

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

$$\mathbb{E}[\quad] = \mathbb{E}\left[\mathbb{E}[\quad | x^k] \right]$$

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

$\mathbb{E}[g^k | x^k] = \nabla f(x^k)$ где SGD (гарантия несмещ.)

где лемма 1)?

$$\mathbb{E}\left[g^k | x^k \right] = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} y_i^k | x^k \right]$$

$$= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[y_i^k | x^k \right] =$$

$$= \frac{1}{n} \sum_{i=1}^{n} \frac{1}{n} \nabla f_i(x^k) + \left(1 - \frac{1}{n}\right) y_i^{k-1}$$

$$= \frac{1}{n} \cdot \underbrace{\frac{1}{n} \sum_{i=1}^{n} \nabla f_i(x^k)}_{\nabla f(x^k)} + \left(1 - \frac{1}{n}\right) \cdot \frac{1}{n} \sum_{i=1}^{n} y_i^{k-1}$$

$$= \frac{1}{n} \nabla f(x^k) + \left(1 - \frac{1}{n}\right) \cdot \frac{1}{n} \sum_{i=1}^{n} y_i^{k-1} \neq \nabla f(x^k)$$

$$\mathbb{E}[g^k | x^k] = \frac{1}{n} \sum_{i=1}^{n} y_i^{k-1} + \frac{1}{n} \mathbb{E}\left[\nabla f_{i_k}(x^k) - y_{i_k}^{k-1} | x^k\right]$$

$$= \frac{1}{n} \nabla f(x^k) + \left(1 - \frac{1}{n}\right) \cdot \frac{1}{n} \sum_{i=1}^{n} y_i^{k-1}$$

α

$$= \alpha \nabla f(x^k) + (1 - \alpha) \cdot \frac{1}{n} \sum_{i=1}^{n} y_i^{k-1}$$

$\alpha = 1$

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

для 2) несмещенность есть!

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

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

μ-сильная выпуклость

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

Работаем с $\|g^k\|_2^2$

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

g^k

$$= \mathbb{E}\left[\left\| \frac{1}{n}\sum_{i=1}^{n} y_i^{k-1} + \nabla f_{i_k}(x^k) - y_{i_k}^{k-1} - \nabla f(x^*)\right\|_2^2 \mid X^k\right]$$

$\pm \nabla f_{i_k}(x^*)$

$$= \mathbb{E}\Big[\| \nabla f_{i_k}(x^k) - \nabla f_{i_k}(x^*)$$
$$+ \nabla f_{i_k}(x^*) - y_{i_k}^{k-1} + \frac{1}{n}\sum_{i=1}^{n} y_i^{k-1} - \nabla f(x^*)\|_2^2 \mid X^k\Big]$$

KSU $\quad \|a+b\|_2^2 \leq 2\|a\|_2^2 + 2\|b\|_2^2$

$$\leq 2\mathbb{E}\left[\|\nabla f_{i_k}(x^k) - \nabla f_{i_k}(x^*)\|_2^2 \mid X^k\right]$$
$$+ 2\mathbb{E}\left[\|y_{i_k}^{k-1} - \nabla f_{i_k}(x^*) + \frac{1}{n}\sum y_i^{k-1} - \nabla f(x^*)\|_2^2 \mid X^k\right]$$

$$\mathbb{E}\left[y_{i_k}^{k-1} - \nabla f_{i_k}(x^*) \mid X^k\right] = \frac{1}{n}\sum y_i^{k-1} - \nabla f(x^*)$$

$$\mathbb{E}\left[\| g - \mathbb{E}[g \mid X^k]\|_2^2 \mid X^k\right] = \mathbb{D}\left[g \mid X^k\right]$$
$$\leq \mathbb{E}\left[\|g\|_2^2 \mid X^k\right]$$

$$\leq 2\mathbb{E}\left[\|\nabla f_{i_k}(x^k) - \nabla f_{i_k}(x^*)\|_2^2 \mid X^k\right]$$
$$+ 2\mathbb{E}\left[\|y_{i_k}^{k-1} - \nabla f_{i_k}(x^*)\|_2^2 \mid X^k\right]$$

$$= 2\cdot \frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(x^k) - \nabla f_i(x^*)\|_2^2$$
$$+ 2\cdot \frac{1}{n}\sum_{i=1}^{n} \|y_i^{k-1} - \nabla f_i(x^*)\|_2^2$$

L - гладкость

$$\leq 4L \cdot \frac{1}{n}\sum_{i=1}^{n}\left(f_i(x^k) - f_i(x^*) - \langle \nabla f_i(x^*); x^k - x^*\rangle\right)$$

$$+ 2 \cdot \frac{1}{n}\sum_{i=1}^{n}\|y_i^{k-1} - \nabla f_i(x^*)\|_2^2$$

$$= 4L\left(f(x^k) - f(x^*) - \langle \cancel{\nabla f(x^*)}^{\,0}; x^k - x^*\rangle\right)$$

$$+ 2 \cdot \frac{1}{n}\sum_{i=1}^{n}\|y_i^{k-1} - \nabla f_i(x^*)\|_2^2$$

Вернемся и соединим:

$$\mathbb{E}\left[\|x^{k+1} - x^*\|_2^2\right] \leq (1-\gamma\mu)\,\mathbb{E}\left[\|x^k - x^*\|_2^2\right]$$
$$- 2\gamma\,\mathbb{E}\left[f(x^k) - f(x^*)\right]$$
$$+ 4L\gamma^2\,\mathbb{E}\left[f(x^k) - f(x^*)\right]$$
$$+ 2\gamma^2 \cdot \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\|y_i^{k-1} - \nabla f_i(x^*)\|_2^2\right]$$

$y_i^k \approx \nabla f_i(\underset{\text{старая}}{\tilde{x}^k})$

если $x^k \to x^*$, то $\tilde{x}^k \to x^*$

$y_i^k \to \nabla f_i(x^*)$

Разберемся с последним членом:

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\|y_i^k - \nabla f_i(x^*)\|_2^2 \,\middle|\, x^k\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\|y_i^k - \nabla f_i(x^*)\|_2^2 \mid x^k\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{n}\|\nabla f_i(x^k) - \nabla f_i(x^*)\|_2^2 + \left(1-\frac{1}{n}\right)\|y_i^{k-1} - \nabla f_i(x^*)\|_2^2\right)$$

$$= \left(1-\frac{1}{n}\right)\cdot\frac{1}{n}\sum_{i=1}^{n}\|y_i^{k-1} - \nabla f_i(x^*)\|_2^2 + \frac{1}{n}\cdot\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(x^k) - \nabla f_i(x^*)\|_2^2$$

L- гладкость

$$\leq \left(1-\frac{1}{n}\right)\cdot\frac{1}{n}\sum_{i=1}^{n}\|y_i^{k-1} - \nabla f_i(x^*)\|_2^2 + \frac{1}{n}\cdot 2L\left(f(x^k) - f(x^*)\right)$$

Итого:

$$\mathbb{E}\left[\|x^{k+1} - x^*\|_2^2\right] \leq (1-\gamma\mu)\,\mathbb{E}\left[\|x^k - x^*\|_2^2\right] - 2\gamma\,\mathbb{E}\left[f(x^k) - f(x^*)\right] + 4L\gamma^2\,\mathbb{E}\left[f(x^k) - f(x^*)\right] + 2\gamma^2\cdot\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\|y_i^{k-1} - \nabla f_i(x^*)\|_2^2\right]$$

$+$ мат. ожид.

γ

$+$

$$\underbrace{\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\|y_i^k - \nabla f_i(x^*)\|_2^2\right]}_{\sigma_k^2} \leq \left(1-\frac{1}{n}\right)\cdot\underbrace{\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\|y_i^{k-1} - \nabla f_i(x^*)\|_2^2\right]}_{\sigma_{k-1}^2} + \frac{2L}{n}\cdot\mathbb{E}\left[f(x^k) - f(x^*)\right]$$

$+$ мат. ожид.

обозначение для краткости

$$\mathbb{E}\left[\|x^{k+1}-x^*\|_2^2 + M \cdot \sigma_k^2\right]$$

$\uparrow$ megfelelően

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

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

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

$$2\gamma - 4\gamma^2 L - 8\gamma^2 L \geq 0 \Rightarrow \gamma - 6\gamma^2 L \geq 0 \Rightarrow \boxed{\gamma \leq \frac{1}{6L}}$$

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

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

<u>линейная</u> сходимость $x^k \to x^*$

$\sigma_k^2 \to 0$

$\gamma = \frac{1}{6L}$ и задуманная рекурсия

$$\mathbb{E}\left[\|x^k - x^*\|_2^2 + M \cdot \sigma_k^2\right] \le \max\left\{\left(1-\frac{\mu}{8L}\right);\left(1-\frac{1}{2n}\right)\right\}^k \cdot \mathbb{E}\left[\|x^0 - x^*\|_2^2 + M\sigma_0^2\right]$$

$$\le$$

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

Оценка сходимости

$$\boxed{k = O\left(\left[\frac{L}{\mu} + n\right]\log\frac{1}{\varepsilon}\right) \text{ итераций}}$$

для GD:

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

Сложность вычислительная:

SAGA: $O\left(\left[\frac{L}{\mu} + n\right]\log\frac{1}{\varepsilon}\right)$ вызовов ∇f_i

GD: $O\left(n\frac{L}{\mu}\log\frac{1}{\varepsilon}\right)$ вызовов ∇f_i

⊕ сходимость, как у GD (по итерациям) и в O компоненты

⊕ итерация стоит $O(1)$ (у GD $O(n)$)

⊕ сходимость к точному решению

⊖ $O(nd)$ дополнит памяти

⊖ шаг у SAGA меньше $\frac{1}{L}$ (хуже, чем у GD)

- SVRG

$$\boxed{\begin{gathered} x^{k+1} = x^k - \gamma g^k \\ g^k = \nabla f_{i_k}(x^k) - \nabla f_{i_k}(w^k) + \nabla f(w^k) \\ w^k \text{ – обновляем редко} \\ w^k = \begin{cases} x^k & \text{раз в } t \text{ итераций} \\ w^{k-1} & \text{иначе} \end{cases} \end{gathered}}$$

Физика: $x^k \to x^*$, то $w^k \to x^*$

$$g^k = \underbrace{\underbrace{\nabla f_{i_k}(x^k) - \nabla f_{i_k}(w^k)}_{\nabla f_{i_k}(x^*) - \nabla f_{i_k}(x^*)} \to 0 + \nabla f(w^k)}_{g^k \to 0}$$

$$\nabla f(x^*) = 0$$

Сходимость:

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

⊕ плюсы SAGA

⊕ память $O(d)$

⊖ подсчет полного градиента

⊖ $O(1)$ стоимость итерации ≈ 3 (у SAGA $=1$)

• **SARAH** (модификация идеи SVRG)

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

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

$$g^k = \begin{cases} \nabla f(x^k) & \text{раз в } t \text{ итер.} \\ g^k & \text{иначе} \end{cases}$$

$\approx \nabla f(x^{k-1})$ смещающаяся референсная точка

"сложенная" версия SVRG

g^k — смещенный (у SAGA, SVRG был несмещ.)

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

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

$$= \nabla f(x^k) - \nabla f(x^{k-1}) + g^{k-1}$$

$\neq 0$

Сходимость:

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

⊕ лучше, „мягче", „сложн." на практике, чем SVRG

⊖ полный градиент