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

- SGD

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

случайно нез. и равн.

Проблема: сходимость к окрестности

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

$\neq 0$

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

уходим из решения

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

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

$x^k \to x^*$

- SAGA

y_i^k — послед. памяти

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

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

полный градиент, но запоздалый

SAGA

2) $\boxed{x^{k+1} = x^k - \gamma \cdot g^k}$

$$g^k = \frac{1}{n} \sum y_i^k + \left(1 - \frac{1}{n}\right) \nabla f_{i_k}(x^k) - \left(1 - \frac{1}{n}\right) \cdot \frac{1}{n} \sum y_i^{k-1}$$

чуть-чуть по другому:

$$\boxed{g^k = \frac{1}{n}\sum y_i^{k-1} + \nabla f_{i_k}(x^k) - y_{i_k}^{k-1}}$$

Док-во сходимости.

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

для SGD вида: $x^{k+1} = x^k - \gamma g^k$

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

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

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

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

для SGD (пр. семинар) да

для метода 1) ?

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

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

правило g_i^k

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

$p_1\ (i = i_k) = \frac{1}{n}$ $\qquad 1 - p_1$

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

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

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

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

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

см SGD и выше

$$= \cancel{\frac{1}{n}\sum y_i^{k-1}} + \nabla f(x^k) - \cancel{\frac{1}{n}\sum y_i^{k-1}}$$

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

Возвращаемся с $\mathbb{E}\left[g^k \mid x^k\right]$

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

μ-сильная выпуклость

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

Теперь $\|g^k\|_2^2$

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

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

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

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

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

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

$$\mathbb{E}\left[\|\xi - \mathbb{E}[\xi \mid x^k]\|_2^2 \mid x^k\right] = \mathbb{D}\left[\|\xi\|_2^2 \mid x^k\right]$$
$$\le \mathbb{E}\left[\|\xi\|_2^2 \mid x^k\right]$$

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

L - гладкость f_{i_k}

$$\le 4L\,\mathbb{E}\left[f_{i_k}(x^k) - f_{i_k}(x^*) - \langle \nabla f_{i_k}(x^*); x^k - x^* \rangle \mid x^k\right]$$
$$+ 2\mathbb{E}\left[\| y_{i_k}^{k-1} - \nabla f_{i_k}(x^*)\|_2^2 \mid x^k\right]$$
$$= 4L\left[\frac{1}{n}\sum\left(f_i(x^k) - f_i(x^*)\right) - \langle \frac{1}{n}\sum \nabla f_i(x^*); x^k - x^* \rangle\right]$$
$$+ 2\mathbb{E}\left[\| y_{i_k}^{k-1} - \nabla f_{i_k}(x^*)\|_2^2 \mid x^k\right]$$
$$= 4L\left(f(x^k) - f(x^*) - \langle \cancel{\nabla f(x^*)}^{\,0}; x^k - x^* \rangle\right)$$
$$+ 2\mathbb{E}\left[\| y_{i_k}^{k-1} - \nabla f_{i_k}(x^*)\|_2^2 \mid x^k\right]$$

Вернемся и подставим:

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

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

$$
\begin{aligned}
\mathbb{E}\left[\|x^{k+1}-x^*\|_2^2\right] \le (1-\gamma\mu)\,&\mathbb{E}\left[\|x^k-x^*\|_2^2\right] \\
&- 2\gamma\,\mathbb{E}\left[f(x^k)-f(x^*)\right] \\
&+ \gamma^2\cdot 4L\,\mathbb{E}\left[f(x^k)-f(x^*)\right] \\
&+ \gamma^2\cdot 2\mathbb{E}\left[\frac{1}{n}\sum\|y_i^{k-1}-\nabla f_i(x^*)\|_2^2\right]
\end{aligned}
$$

$y_i^k \approx \nabla f_i(\tilde{x}^k)$, если $x^k \to x^*$, то $\tilde{x}^k \to x^*$

старой

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

Разберёмся с последним членом:

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

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

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

L - гладкость f_i

$$
\le \mathbb{E}\left[\frac{1}{n}\cdot 2L\left(f_i(x^k)-f_i(x^*)-\langle\nabla f_i(x^*), x^k-x^*\rangle\right) + \left(1-\frac{1}{n}\right)\|y_i^{k-1}-\nabla f_i(x^*)\|_2^2\right]
$$

Суммируем по $i = 1..n$

$$\mathbb{E}\Big[\underbrace{\frac{1}{n}\sum_{i=1}^{n}\|y_i^k - \nabla f_i(x^*)\|_2^2}_{\sigma_k^2}\Big] =$$

$$\leq \mathbb{E}\Big[\frac{2L}{n}\cdot\frac{1}{n}\sum\big(f_i(x^k) - f_i(x^*)\big) - \big\langle \underbrace{\cancel{\frac{1}{n}\sum\nabla f_i(x^*)}}_{0} ; x^k - x^*\big\rangle\Big]$$

$$+ \Big(1 - \frac{1}{n}\Big)\mathbb{E}\Big[\underbrace{\frac{1}{n}\sum\|y_i^{k-1} - \nabla f_i(x^*)\|_2^2}_{\sigma_{k-1}^2}\Big]$$

Итого:

$$\mathbb{E}[\sigma_k^2] \leq \Big(1 - \frac{1}{n}\Big)\mathbb{E}[\sigma_{k-1}^2] + \frac{2L}{n}\big(f(x^k) - f(x^*)\big)$$

$+$

$\sigma_k^2 \to 0$

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

$\|x^k - x^*\|_2^2 \to 0$

Создаем общий критерий сходимости:

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

подберем потом

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

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

$$\gamma - 4L\gamma^2 - 4L\gamma^2 = \gamma - 8L\gamma^2 \ge 0$$

$$\Rightarrow \boxed{\gamma \le \frac{1}{8L}}$$

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

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

<u>линейная</u> сходимость по $x^k \to x^*$, $\sigma_k^2 \to 0$

$\gamma = \frac{1}{8L}$ и запустить рек.

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

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

Оценки сходимости:

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

для GD (полный градиент

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

⊕ сходимость, как у GD сразу (с точн. до $n \log \frac{1}{\varepsilon}$)

⊕ итерация стоит $O(1)$ у SAGA

у GD стоит $O(n)$

⊕ сходимость к точному решению

⊖ $O(nd)$ лишней памяти

- SVRG

$$\boxed{\begin{aligned} x^{k+1} &= x^k - \gamma g^k \\ g^k &= \nabla f_{i_k}(x^k) - \nabla f_{i_k}(w^k) + \nabla f(w^k) \end{aligned}}$$

w^k – обновляется редко

$$w^k = \begin{cases} x^k & \text{раз в } t \text{ итераций} \\ w^{k-1} & \text{иначе} \end{cases}$$

Сходимость:

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

Физика: $x^k \to x^* \quad (w^k \to x^*)$

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

$$\underbrace{\nabla f_{i_k}(x^*) - \nabla f_{i_k}(x^*)}_{\to 0}$$

$\Rightarrow \quad g^k \to 0$

(+) плюсы SAGA

(+) память $O(d)$

(−) подсчёт полного градиента

- **SARAH** (модификация идеи SVRG)

$$\boxed{\begin{aligned} x^{k+1} &= x^k - \gamma g^k \\ g^k &= \nabla f_{i_k}(x^k) - \nabla f_{i_k}(x^{k-1}) + g^{k-1} \\ g^k &\text{ — редко брать, как } \nabla f(x^k) \end{aligned}}$$

"смещённая" версия SVRG

g^k — смещённый (у SAGA, SVRG — был несмещённый)

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

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

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

$$\neq 0$$

Сходимость:

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

(+) лучше на практике, чем SVRG

(−) вычисляет полный градиент