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

Идея:

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

$$g^k \to 0 \ (\nabla f(x^*)) \qquad x^k \to x^*$$

$$\nabla f_i(x^k) \not\to 0$$

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

Алгоритм 1 SAGA

Вход: размер шага $\gamma > 0$, стартовая точка $x^0 \in \mathbb{R}^d$, значения памяти $y_i^0 = 0$ для всех $i \in [n]$, количество итераций K

1: **for** $k = 0, 1, \ldots, K-1$ **do**

2: Сгенерировать независимо i_k

3: Вычислить $g^k = \nabla f_{i_k}(x^k) - y_{i_k}^k + \frac{1}{n} \sum\limits_{j=1}^{n} y_j^k$

4: Обновить $y_i^{k+1} = \begin{cases} \nabla f_i(x^k), & \text{если } i = i_k \\ y_i^k, & \text{иначе} \end{cases}$

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

6: **end for**

Выход: x^K

Суть:

y_j^k – для каждого $j = 1..n$

– хранит инф. о град ∇f_j

$\frac{1}{n} \sum_{j=1}^{n} y_j^k$ – „запазд." градиент

g^k? да, можно

$$\mathbb{E}\left[\frac{1}{n} \sum_{j=1}^{n} y_j^k \,\middle|\, x^k \right] = \mathbb{E}_{i_k}\left[\frac{1}{n} \sum_{j=1}^{n} y_j^k \right]$$

$$= \frac{1}{n} \sum_{j=1}^{n} \left[\frac{1}{n} \nabla f_j(x^k) + \left(1 - \frac{1}{n}\right) y_j^{k-1} \right]$$

$$= \frac{1}{n^2}\sum_{j=1}^{n} \nabla f_j(x^k) + \left(1-\frac{1}{n}\right)\cdot\frac{1}{n}\sum y_j^{k-1}$$

$$= \frac{1}{n}\nabla f(x^k) + \dots \neq \nabla f(x^k)$$

- g^k – смещ. ($\mathbb{E} \neq \nabla f(x^k)$)
- это неплохо (но мы исправим)

Поэтому в SAGA

$$g^k = \nabla f_{i_k}(x^k) - y_{i_k}^k + \frac{1}{n}\sum_{j=1}^{n} y_j^k$$

$\frac{1}{n}$ ———— $\frac{1}{n}$ ———— в пред. версии

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

$$= \sum_{j=1}^{n}\frac{1}{n}\nabla f_j(x^k) - \cancel{\sum_{j=1}^{n}\frac{1}{n} y_j^k} + \cancel{\frac{1}{n}\sum_{j=1}^{n} y_j^k}$$

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

Фишка:

$$g^k \xrightarrow{?} 0 \qquad x^k \to x^*$$

$$\nabla f_{i_k}(x^k) - y_{i_k}^k + \frac{1}{n}\sum y_j^k$$

$$x^k \to x^* \qquad y_j^k \to \nabla f_j(x^*)$$

$$\cancel{\nabla f_{i_k}(x^*)} - \cancel{\nabla f_{i_k}(x^*)} + \underbrace{\frac{1}{n}\sum \nabla f_j(x^*)}_{\nabla f(x^*) \to 0}$$

Док-во сходимости:

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

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

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

Посчитаем $$\mathbb{E}\left[\|g^k\|_2^2 \mid x^k\right] =$$

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

$$\left(\mathbb{E}\left[\ \mid x^k\right] = \mathbb{E}_{i_k}\right.$$

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

$$= \mathbb{E}_{i_k}\Big[\| \nabla f_{i_k}(x^k) - y_{i_k}^k + \underbrace{\frac{1}{n}\sum y_j^k - \underbrace{\nabla f(x^*)}_{\frac{1}{n}\sum \nabla f_j(x^*)}}_{\frac{1}{n}\sum (y_j^k - \nabla f_j(x^*))}\|_2^2\Big]$$

$$= \mathbb{E}_{i_k}\Big[\| \nabla f_{i_k}(x^k) - \nabla f_{i_k}(x^*) - y_{i_k}^k + \nabla f_{i_k}(x^*)$$

$$+ \frac{1}{n}\sum y_j^k - \nabla f_j(x^*)\|_2^2$$

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

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

$$+ 2\,\mathbb{E}_{i_k}\left[\| y_{i_k}^k - \nabla f_{i_k}(x^*) - \frac{1}{n}\sum (y_j^k - \nabla f_j(x^*))\|_2^2\right]$$

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

$$\mathbb{E}\|\xi - \mathbb{E}\xi\|_2^2 \le \mathbb{E}\|\xi\|^2$$

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

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

$$= 2\sum_{j=1}^{n} \frac{1}{n}\|\nabla f_j(x^k) - \nabla f_j(x^*)\|_2^2$$

$$+ 2\sum_{j=1}^{n} \frac{1}{n}\| y_j^k - \nabla f_j(x^*)\|_2^2$$

L - гладкость f_j $\qquad \nabla f(x^*)=0$

$$\le 2\cdot\frac{1}{n}\sum_{j=1}^{n} 2L\left(f_j(x^k)-f_j(x^*)-\langle \nabla f_j(x^*); x^k-x^*\rangle\right)$$
$$+\,2\sum_{j=1}^{n}\frac{1}{n}\|y_j^k-\nabla f_j(x^*)\|_2^2$$

$$= 4L\left(f(x^k)-f(x^*)\right)$$
$$+\,2\sum_{j=1}^{n}\frac{1}{n}\|y_j^k-\nabla f_j(x^*)\|_2^2$$

Возвращаемся:

$$\mathbb{E}\left[\|x^{k+1}-x^*\|_2^2 \mid x^k\right] = \|x^k-x^*\|_2^2-2\gamma\langle \nabla f(x^k); x^k-x^*\rangle \quad \oplus$$
$$+\,4L\gamma^2\left(f(x^k)-f(x^*)\right) \quad \oplus$$
$$+\,2\gamma^2\cdot\frac{1}{n}\sum\|y_j^k-\nabla f_j(x^*)\|_2^2 \quad ?$$

$y_j^k \to \nabla f_j(x^*)$

как ведет себя $\frac{1}{n}\sum\|y_j^k-\nabla f_j(x^*)\|_2^2$:

$$\underbrace{\mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{n}\|y_j^{k+1}-\nabla f_j(x^*)\|_2^2 \mid x^k\right]}_{\sigma_{k+1}^2}$$
$$= \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[\|y_j^{k+1}-\nabla f_j(x^*)\|_2^2 \mid x^k\right]$$
$$= \frac{1}{n}\sum_{j=1}^{n}\ \frac{1}{n}\cdot\|\nabla f_j(x^k)-\nabla f_j(x^*)\|_2^2 + \left(1-\frac{1}{n}\right)\cdot\|y_j^k-\nabla f_j(x^*)\|_2^2$$
$$= \frac{1}{n}\cdot\frac{1}{n}\sum\|\nabla f_j(x^k)-\nabla f_j(x^*)\|_2^2$$
$$+\left(1-\frac{1}{n}\right)\cdot\underbrace{\frac{1}{n}\sum\|y_j^k-\nabla f_j(x^*)\|_2^2}_{\sigma_k^2}$$

$$\mathbb{E}\left[\sigma_{k+1}^2 \mid x^k\right] = \left(1-\frac{1}{n}\right)\sigma_k^2 + \frac{1}{n}\cdot\frac{1}{n}\sum_{j=1}^{n}\|\nabla f_j(x^k) - \nabla f_j(x^*)\|_2^2$$

L - гладкость

$$\mathbb{E}\left[\sigma_{k+1}^2 \mid x^k\right] = \left(1-\frac{1}{n}\right)\sigma_k^2 + \frac{1}{n}\cdot 2L\left(f(x^k) - f(x^*)\right)$$

Итого:

μ-сильная выпуклость

1) $$\mathbb{E}\left[\|x^{k+1} - x^*\|_2^2 \mid x^k\right] \le \|x^k - x^*\|_2^2 - 2\gamma\langle \nabla f(x^k); x^k - x^*\rangle + 4L\gamma^2\left(f(x^k) - f(x^*)\right) + 2\gamma^2\cdot\sigma_k^2$$

$$\le (1-\gamma\mu)\|x^k - x^*\|_2^2 + 2\gamma(2L\gamma - 1)\left(f(x^k) - f(x^*)\right) + 2\gamma^2\sigma_k^2 \quad ?$$

лин. сход

$\gamma \le \dots$ уберем

2) $$\mathbb{E}\left[\sigma_{k+1}^2 \mid x^k\right] = \left(1-\frac{1}{n}\right)\sigma_k^2 + \frac{1}{n}\cdot 2L\left(f(x^k) - f(x^*)\right) \quad ?$$

лин. сход.

Суммируем с множителем $M > 0$

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

$$+ M \cdot 2L \cdot \frac{1}{n}\left(f(x^k) - f(x^*)\right)$$

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

$$1-\frac{1}{n}+\frac{2\gamma^2}{M} \le 1-\frac{1}{2n} \qquad \boxed{M = 4\gamma^2 n}$$

$$= (1-\gamma\mu)\|x^k-x^*\|_2^2$$
$$+ \left(1-\frac{1}{2n}\right)\cdot M\sigma_k^2$$
$$+ \underbrace{\left(2\gamma(2L\gamma-1)+8\gamma^2 L\right)}\ \underbrace{\left(f(x^k)-f(x^*)\right)}_{\ge 0}$$

$$2\gamma(6L\gamma-1)\le 0 \qquad \boxed{\gamma \le \frac{1}{6L}}$$

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

Теорема сходимость SAGA

Пусть задача безусловной стохастической оптимизации вида конечной суммы с L-гладкими, выпуклыми функциями f_i и μ-сильно выпуклой целевой функцией f решается с помощью SAGA с $\gamma \le \frac{1}{6L}$. Тогда справедлива следующая оценка сходимости

$$\mathbb{E}\left[\|x^k-x^*\|_2^2\right] \le \mathbb{E}\left[V_k\right] \le \max\left\{(1-\mu\gamma);\left(1-\frac{1}{2n}\right)\right\}^k \mathbb{E}\left[V_0\right],$$

где $V_k = \|x^k - x^*\|_2^2 + 4n\gamma^2 \cdot \frac{1}{n}\sum_{i=1}^{n}\|y_i^k - \nabla f_i(x^*)\|_2^2$.

$$\|x^k-x^*\|_2^2 \le V_k$$

Оценка на число итераций:

$$\gamma = \frac{1}{6L}$$

$$\max\left\{\left(1 - \frac{\mu}{6L}\right); \left(1 - \frac{1}{2n}\right)\right\}^k$$

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

для град спуска

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

n f_i за итерацию

град спуск:

$$O\left(n\frac{L}{\mu}\log\frac{1}{\varepsilon}\right)$$ f_i выз

SAGA:

$$O\left(\left(\frac{L}{\mu} + n\right)\log\frac{1}{\varepsilon}\right)$$ f_i выз

⊕ значительно дешевле, чем град спуск

⊖ $O(nd)$ памяти

Алгоритм 2 SVRG

Вход: размер шага $\gamma > 0$, стартовая точка $x^0 \in \mathbb{R}^d$, количество итераций в эпохе K, количество эпох S

for $s = 0, 1, \ldots, S-1$ **do**
  Обновить $w^s = x^{s-1,K}$
  Посчитать и сохранить $\nabla f(w^s)$
  for $k = 0, 1, \ldots, K-1$ **do**
    $x^{s,k+1} = x^{s,k} - \gamma g^k$
    Сгенерировать независимо i_k
    Вычислить $g^{k+1} = \nabla f_{i_k}(x^{s,k+1}) - \nabla f_{i_k}(w^s) + \nabla f(w^s)$
  end for
end for

Выход: $x^{S-1,K}$

Идея:

$\nabla f(w^s)$ – град. в референсной точке

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

$x^k \to x^*$ $\quad \nabla f_{i_k}(x^*) - \nabla f_{i_k}(x^*) + \nabla f(x^*) \to$

$w \to x^*$ $\quad \to \nabla f(x^*) \to 0$

Сложность:

$$O\left(\left(n + \frac{L}{\mu}\right) \log \frac{1}{\varepsilon}\right) \quad f_i \text{ вых}$$

⊕ нет памяти $O(nd)$, только $O(d)$

⊖ подсчет полного градиента

Ускорение:

- 2 моментума

$$O\left(\sqrt{n \frac{L}{\mu}} \log \frac{1}{\varepsilon}\right) \quad f_i \text{ вых}$$

Нестеров:

$$O\left(n \sqrt{\frac{L}{\mu}} \log \frac{1}{\varepsilon}\right)$$

Алгоритм 3 SARAH

Вход: размер шага $\gamma > 0$, стартовая точка $x^0 \in \mathbb{R}^d$, количество итераций в эпохе K, количество эпох S

1: **for** $s = 0, 1, \ldots, S-1$ **do**
2: Посчитать $g^0 = \nabla f(x^{s-1,K})$
3: **for** $k = 0, 1, \ldots, K-1$ **do**
4: $x^{s,k+1} = x^{s,k} - \gamma g^k$
5: Сгенерировать независимо i_k
6: Вычислить $g^{k+1} = \nabla f_{i_k}(x^{s,k+1}) - \nabla f_{i_k}(x^{s,k}) + g^k$
7: **end for**
8: **end for**

Выход: $x^{S-1,K}$

<u>Идея</u>:

Более плавный SVRG

$\nabla f(w) \longrightarrow g^k$ — от же рев $\approx \nabla f(x^k)$

SVRG

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

$$= \nabla f(x^{k+1}) - \nabla f(x^k) + g^k \neq 0$$

⊕ сходится плавнее, чем SVRG

⊖ смещённые градиенты