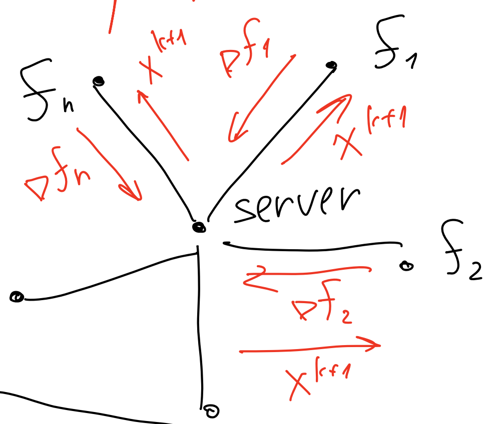
$$\min_{x\in\mathbb{R}^d} f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

распределение между устройствами



$\nabla f \leftarrow$ коммуникац.

server: $x^{k+1} = x^k - \gamma \cdot \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(x^k)$

идеал: ускорить подсч. ∇f в n раз, но в жизни все хуже!

$$\frac{T_{\nabla f\ local}}{n} + T_{comm.} \ ?\ T_{\nabla f\ local}$$

Сжатие информации (от устр. к серверу)

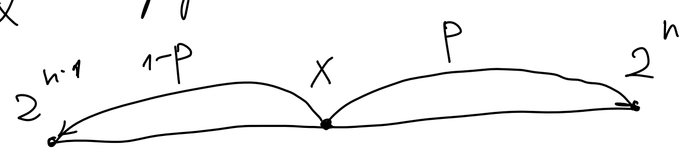
$$x^{k+1} = x^k - \gamma \cdot \frac{1}{n}\sum_{i=1}^{n} Q(\nabla f_i(x^k))$$

Q - оператор сжатия

Пример

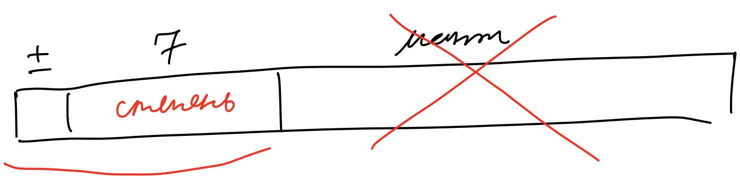
1) Q - случайный выбор координат

2) Q - округление



$p \sim 2^n - x$

$1-p \sim x - 2^{n-1}$



$$\mathbb{E}\, g^k = \nabla f(x^k)$$

$$\mathbb{E}\left[\frac{1}{n}\sum Q(\nabla f_i(x^k))\right] = \frac{1}{n}\sum \nabla f_i(x^k) = \nabla f(x^k)$$

1) Q – несмещ. функ от функц

2) $\mathbb{E}\, Q(x) = x$

3) $\mathbb{E}\|Q(x)-x\|^2 \leq \omega \|x\|^2 \quad \omega > 0 \quad \forall x \in \mathbb{R}^d$

Док-во: L – гладкие f_i, f – μ-сильно вып.

$$\mathbb{E}\left[\|x^{k+1}-x^*\|^2\right] \leq (1-\mu\gamma)\,\mathbb{E}\left[\|x^k-x^*\|^2\right] - \gamma\,\mathbb{E}\left[f(x^k)-f(x^*)\right] + \gamma^2\,\mathbb{E}\left[\left\|\frac{1}{n}\sum Q(\nabla f_i(x^k))\right\|^2\right] \ominus$$

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

$$= \mathbb{E}\left[\left\|\frac{1}{n}\sum\left(Q(\nabla f_i(x^k)) - \nabla f_i(x^k)\right)\right\|^2\right] + \mathbb{E}\left[\|\nabla f(x^k)\|^2\right] + 2\,\mathbb{E}\left[\left\langle \frac{1}{n}\sum\left(\underbrace{\mathbb{E}_k\left[Q(\nabla f_i(x^k))\right]}_{\nabla f_i(x^k)} - \nabla f_i(x^k)\right); \nabla f(x^k)\right\rangle\right]$$

$\to 0$

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

$$= \frac{1}{n^2}\,\mathbb{E}\left[\sum_{i=1}^n \|Q(\nabla f_i(x^k)) - \nabla f_i(x^k)\|^2\right] + \frac{1}{n^2}\,\mathbb{E}\left[\sum_{i\neq j}\left\langle \underbrace{\mathbb{E}Q}_{Q_i}(\nabla f_i(x^k)) - \nabla f_i(x^k); \underbrace{\mathbb{E}Q}_{Q_j}(\nabla f_j(x^k)) - \nabla f_j(x^k)\right\rangle\right]$$

$\to 0$

$$+\mathbb{E}\left[\|\nabla f(x^k)\|^2\right]$$

по опред. смещения

$$\leq \frac{1}{n^2}\sum_{i=1}^{n}\omega\,\mathbb{E}\left[\|\nabla f_i(x^k)\|^2\right]+\mathbb{E}\left[\|\nabla f(x^k)\|^2\right]$$

$$=\frac{\omega}{n}\cdot\frac{1}{n}\sum\mathbb{E}\left[\|\nabla f_i(x^k)\|^2\right]+\mathbb{E}\left[\|\nabla f(x^k)\|^2\right]$$

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

$\nabla f(x^*)=0$ но $\nabla f_i(x^*)\neq 0$

$$+\mathbb{E}\left[\|\nabla f(x^k)\|^2\right]$$

$$\leq\frac{2\omega}{n}\cdot\frac{1}{n}\sum\mathbb{E}\|\nabla f_i(x^k)-\nabla f_i(x^*)\|^2$$

$$+\frac{2\omega}{n}\cdot\frac{1}{n}\sum\|\nabla f_i(x^*)\|^2$$

$$+\mathbb{E}\left[\|\nabla f(x^k)\|^2\right]$$

L-гладкость f_i

$$\leq\frac{2\omega}{n}\cdot\frac{1}{n}\sum_{i=1}^{n}2L\mathbb{E}\left(f_i(x^k)-f_i(x^*)-\langle\nabla f_i(x^*);x^k-x^*\rangle\right)$$

$$+\frac{2\omega}{n}\cdot\frac{1}{n}\sum\|\nabla f_i(x^*)\|^2+\mathbb{E}\left[2L\left(f(x^k)-f(x^*)\right)\right]$$

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

$$+\frac{2\omega}{n}\cdot\underbrace{\frac{1}{n}\sum\|\nabla f_i(x^*)\|^2}_{\sigma_*^2}$$

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

$$-1+2L\gamma\left(1+\frac{2\omega}{n}\right)=0 \qquad \gamma=\frac{1}{4L\left(1+\frac{2\omega}{n}\right)}$$

$$\mathbb{E}\left[\|x^{k+1}-x^*\|^2\right] \le (1-\mu\gamma)\,\mathbb{E}\left[\|x^k-x^*\|^2\right]$$
$$+\gamma^2\cdot\frac{2\omega}{n}\sigma_*^2$$
$$\le \ldots \text{ (см. SGD)}$$
$$\le (1-\mu\gamma)^{k+1}\|x^0-x^*\|^2$$
$$+\frac{2\omega}{n}\sigma_*^2\cdot\frac{\gamma^2}{\mu\gamma} = \frac{2\omega\sigma_*^2\gamma}{n\mu}$$

1) $\sigma_*^2=0$ линейная $\qquad \gamma=\frac{1}{4L\left(1+\frac{2\omega}{n}\right)}$

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

число ит. $\geq$ число итер для GD

$$O\left(\frac{L}{\mu}\underbrace{\left(\frac{1}{\beta}+\frac{\omega}{\beta n}\right)}_{=1 \text{ для } GD}\log\frac{1}{\varepsilon}\right)$$ информации

$\frac{1}{\beta}+\frac{\omega}{\beta n}<1$ то есть выигрыш

$\omega = \frac{d}{k}$ k - число выбр. коорд.

Q - случ. выбор коорд.

$\beta = \frac{d}{k}$ $\omega \leq \beta$

$\Downarrow$

$$\frac{1}{\beta}+\frac{\omega}{\beta n} = \frac{k}{d}+\frac{1}{n} \leq 1$$

$\uparrow$ при малых k

2) $\sigma_*^2 \neq 0$ в общем случае

огр. знамен. $\sim \frac{\omega}{n}\sigma_*^2$

DIANA

h_i^k - доп. послед. для каждой машины i

$h_i^0 = \nabla f_i(x^0)$ $\longrightarrow$ на сервер

$\Delta_i^k = \nabla f_i(x^k) - h_i^k$

$\tilde{\Delta}_i^k = Q(\Delta_i^k)$

$h_i^{k+1} = h_i^k + \alpha\tilde{\Delta}_i^k$

→ на сервер $\tilde{\Delta}_i^k$

$$x^{k+1} = x^k - \gamma \cdot \frac{1}{n} \sum \left(h_i^k + \tilde{\Delta}_i^k\right)$$

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

Q — зачем случайное?

Q — огр. / жадное

Примеры:

1) Q — Top k — выбираем k наиб. по $|.|$ коорд.

2) Q — округление без несм.

3) $Q(x)$ для $x_{(i)}$ оставляем $\mathrm{sign}(x_{(i)})$ и от $\|x\|$

$$Q(x) = \begin{cases} \mathrm{sign}(x_{(1)}) \|x\| \\ \vdots \\ \mathrm{sign}(x_{(n)}) \|x\| \end{cases}$$

$Q(x)$ – разложение с помощью SVD

$$x^{k+1} = x^k - \gamma \frac{1}{3} \sum_{i=1}^{3} Q(\nabla f_i(x^k)) \qquad Q = \mathrm{Top}\ 1$$

$$d = 3$$

$$f_1(x) = \left\langle \begin{pmatrix} -3 \\ 2 \\ 2 \end{pmatrix}; \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \right\rangle^2 + \frac{1}{4}\|x\|^2$$

$$f_2(x) = \left\langle \begin{pmatrix} 2 \\ -3 \\ 2 \end{pmatrix} \dots \right. + \dots$$

$$f_3(x) = \left\langle \begin{pmatrix} 2 \\ 2 \\ -3 \end{pmatrix} \dots \right. + \dots$$

$$x^0 = (t, t, t)$$

$$\nabla f_1(x^0) = \frac{t}{2}\begin{pmatrix} -11 \\ 9 \\ 9 \end{pmatrix} \qquad \nabla f_2(x^0) = \frac{t}{2}\begin{pmatrix} 9 \\ -11 \\ 9 \end{pmatrix}$$

$$\nabla f_3(x^0) = \frac{t}{2}\begin{pmatrix} 9 \\ 9 \\ -11 \end{pmatrix}$$

$$\nabla f_1 + \nabla f_2 + \nabla f_3 = \frac{t}{2}\begin{pmatrix} 7 \\ 7 \\ 7 \end{pmatrix}$$

$$Q(\nabla f_1) + Q(\nabla f_2) + Q(\nabla f_3) = \frac{t}{2}\begin{pmatrix} -11 \\ -11 \\ -11 \end{pmatrix}$$

метод расходится

EF $\qquad e_i^{0} = 0$

$$\tilde{\Delta}_i^k = C\left(e_i^k + \gamma \nabla f_i(x^k)\right)$$

$$e_i^{k+1} = e_i^k + \gamma \nabla f_i(x^k) - \tilde{\Delta}_i^k$$

$$x^{k+1} = x^k - \frac{1}{n}\sum \tilde{\Delta}_i^k$$

$\mathbb{E}\, Q(x) \neq x \qquad \mathbb{E}\|Q(x) - x\|^2 \le \left(1 - \frac{1}{\delta}\right)\|x\|^2$

δ для Top k

$= \frac{d}{k}$

$\beta = \frac{d}{k}$

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

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

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

при исп. несмещ. компрес. / сжатий нет теор. улучш. сход. с точки зрения итераций