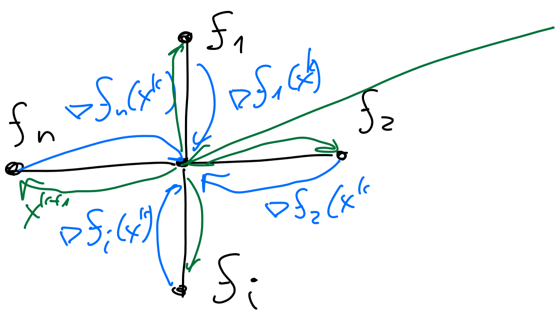
$$\min_{x\in\mathbb{R}^d}\left[f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)\right]$$

хранятся на локальных устройствах.

$$f(x) = \frac{1}{N}\sum_{i=1}^{N} l\left(g(a_i, x), b_i\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n} f_i(x) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{m}\sum_{j=1}^{m} l\left(g(a_j^i, x), b\right)\right]$$



$\nabla f(x^k)$

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

$\boxed{\text{коммуникации}}$

Опр. Стох. оператор $Q(x)$ несмещенный компрессор, если

$\forall x\in\mathbb{R}^d \hookrightarrow \mathbb{E}[Q(x)] = x \qquad \mathbb{E}\left[\|Q(x)\|_2^2\right] = \omega\|x\|_2^2$

$\omega \geq 1$

Примеры

- Rand K оператор (случайный выбор коорд.)

$$\text{Rand}\,K(x) = \frac{d}{K}\sum_{i\in S}[x]_i e_i$$

← базисный вектор

S — случ. набор из K коорд.

$$\mathbb{E}[Q(x)] = \begin{pmatrix}\mathbb{E}[Q(x)]_1\\ \vdots\\ \mathbb{E}[Q(x)]_d\end{pmatrix} = x$$

$$\mathbb{E}[Q(x)]_i = \frac{d}{K}\left(\underset{\frac{K}{d}}{p}[x]_i + (1-p)\,0\right) = [x]_i$$

$$\begin{pmatrix}1\\2\\3\end{pmatrix} \xrightarrow{\text{Rand}1} \underset{1/3}{\begin{pmatrix}1\\0\\0\end{pmatrix}} \quad \underset{1/3}{\begin{pmatrix}0\\2\\0\end{pmatrix}} \quad \underset{1/3}{\begin{pmatrix}0\\0\\3\end{pmatrix}}$$

$$\frac{1}{3}\begin{pmatrix}1\\0\\0\end{pmatrix} + \frac{1}{3}\begin{pmatrix}0\\2\\0\end{pmatrix} + \frac{1}{3}\begin{pmatrix}0\\0\\3\end{pmatrix} = \frac{1}{3}\begin{pmatrix}1\\2\\3\end{pmatrix}$$

$$\mathbb{E}\left[\|Q(x)\|_2^2\right] = \mathbb{E}\left[\sum_{i=1}^{d} [Q(x)]_i^2\right]$$

$$= \sum_{i=1}^{d} \mathbb{E}\left[Q(x)\right]_i^2 = \begin{cases} \frac{d^2}{k^2} x_i^2 & p = \frac{k}{d} \\ 0 & 1-p \end{cases}$$

$$= \frac{d^2}{k^2} \sum_{i=1}^{d} \frac{k}{d} x_i^2$$

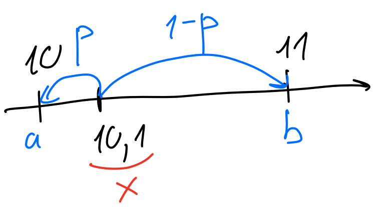
$$= \frac{d}{k} \underbrace{\sum x_i^2}_{\|x\|_2^2} = \frac{d}{k} \|x\|_2^2 \quad \Rightarrow \quad \omega = \frac{d}{k}$$

- Трехуровневая L_2-квантизация

$$[Q(x)]_i = \|x\|_2 \, \text{sign}(x_i) \, \xi_i$$

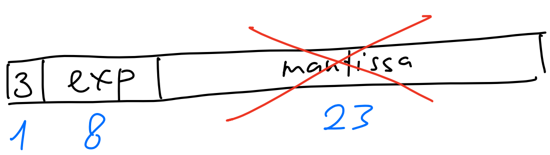
$$\xi_i = \begin{cases} 1 & p = \frac{|x_i|}{\|x\|_2} \\ 0 & 1-p \end{cases}$$

- Округление



p больше до ближайшего числа

$$p = \frac{b-x}{b-a} = \frac{9}{10}$$



- **Градиентный спуск с сжатием**

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

все f_i считают $\nabla f_i(x^k)$

пересылают $Q(\nabla f_i(x^k))$

сервер $\frac{1}{n}\sum Q(\nabla f_i(x^k))$

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

пересылает x^{k+1}

пересылает $Q\left(\frac{1}{n}\sum Q(\nabla f_i(x^k))\right)$

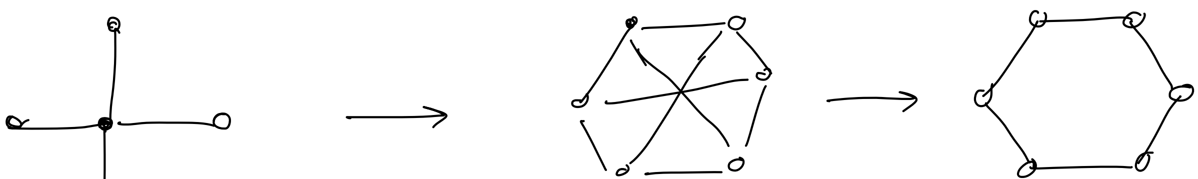
все f_i считают $x^{k+1} = x^k - \gamma Q\left(\frac{1}{n}\sum Q(\nabla f_i(x^k))\right)$

Теорема о сходимости f_i — L-гладкие и μ-сильно вып.

$$\mathbb{E}\left[\|x^k - x^*\|_2^2\right] = O\left(\underbrace{(1-\gamma\mu)^k \|x^0 - x^*\|_2^2}_{\text{линейная сход.}} + \underbrace{\gamma \cdot \frac{\omega}{n\mu} \cdot \frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(x^*)\|_2^2}_{\text{сходимость к окрестности}}\right)$$

$$\text{при } \overbrace{\gamma \le \frac{1}{L}\left(1+\frac{\omega}{n}\right)^{-1}}^{\text{хуже, чем } \frac{1}{L}}$$

- **Обучение устройств**

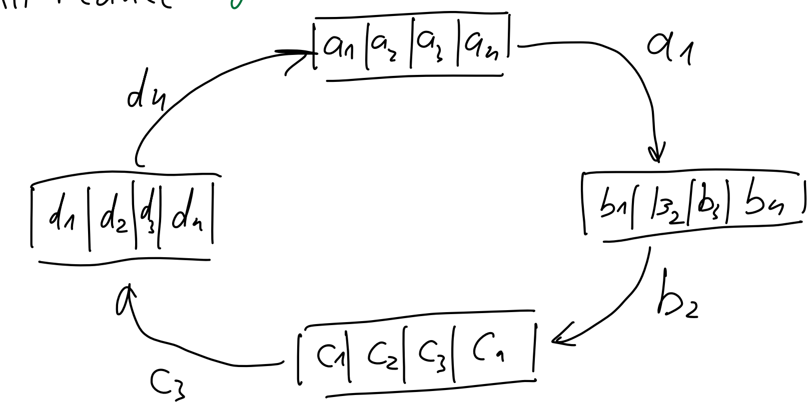


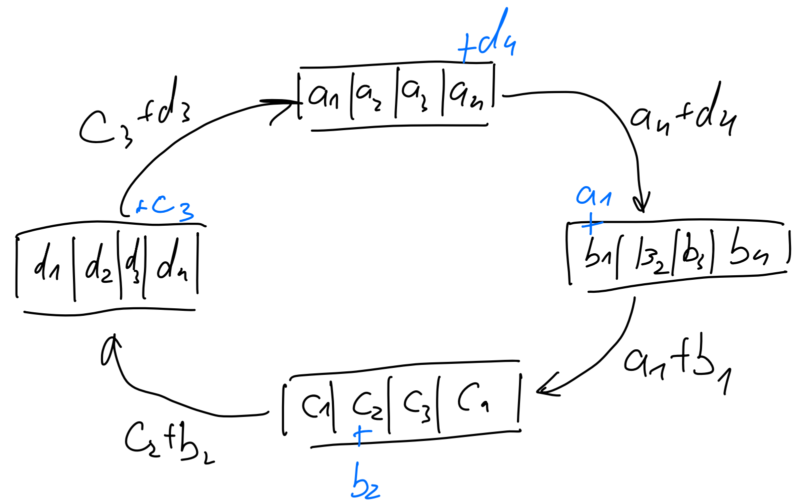
теор. постр. — реальность — то, что используем

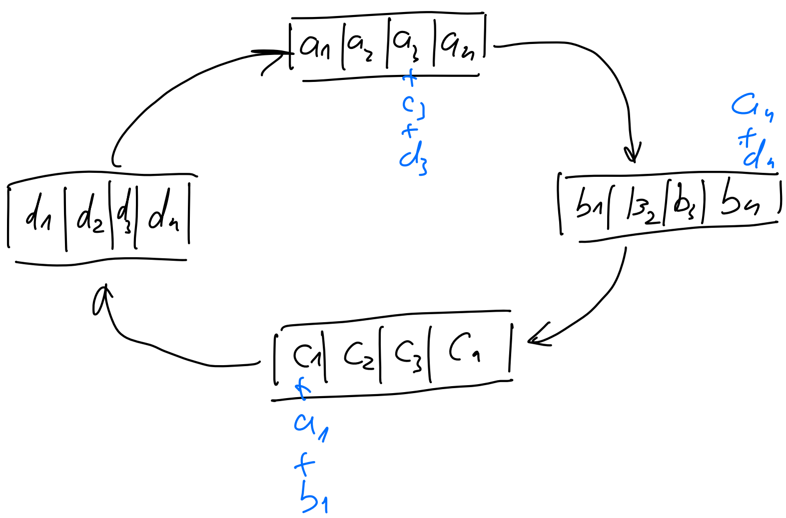
n устройств

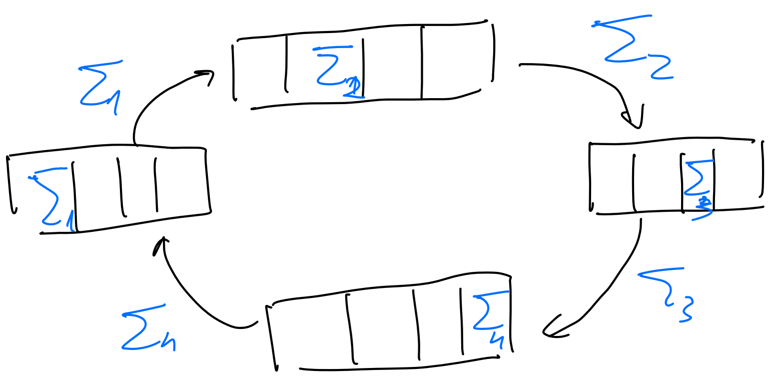
стоимость обучения $\sim n$ ← это плохо!

Ring All Reduce в узлах скорость









сложность $O(1)$

Пример Перестановочный оператор сжатия

$d = qn \quad q \in \mathbb{N}, \quad \pi = (\pi_1 \dots \pi_d)$ - перестановка $\{1 .. d\}$

на i-устройстве $Q_i(x) = n \cdot \sum\limits_{j=q(i-1)+1}^{q \cdot i} x_{\pi_j} e_{\pi_j}$

$d = 6 \quad n = 3$

$\pi = (5 \; 4 \; 2 \; 3 \; 6 \; 1)$

1 устр-ве: 5 и 4 коорд.

2: 2 и 3 коорд

3: 6 и 1 коорд

все устр-ва перечисляют разные коорд.

Опр. Оператор $C(x)$ смещенная компрессия, если

$\forall x \in \mathbb{R}^d \hookrightarrow \quad \mathbb{E}\left[\|C(x) - x\|_2^2\right] \le \left(1 - \frac{1}{\delta}\right)\|x\|_2^2 \quad \delta \geq 1.$

Пример

- „Жадный": выбор координат

$$\mathrm{Top}K(x) = \sum_{i=d-k+1}^{d} x_{(i)} e_{(i)}$$

k - коорд., но максимальных по модулю

$\{x_{(i)}\}$ - сорт. коорд. $\quad |x_{(1)}| \le |x_{(2)}| \le \dots \le |x_{(d)}|$

- Градиентный спуск с сжатием

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

расходится на кв. задачах: $n = 3 \quad d = 3$

$f_1(x) = \langle a; x \rangle^2 + \frac{1}{4}\|x\|_2^2 \qquad f_2(x) = \langle b; x \rangle^2 + \frac{1}{4}\|x\|^2$

$f_3(x) = \langle c; x \rangle^2 + \frac{1}{4}\|x\|_2^2 \qquad a = \begin{pmatrix} -3 \\ 2 \\ 2 \end{pmatrix} \quad b = \begin{pmatrix} 2 \\ -3 \\ 2 \end{pmatrix} \quad c = \begin{pmatrix} 2 \\ 2 \\ -3 \end{pmatrix}$

$$x^* = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \qquad x^0 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

$$\nabla f_1(x^0) = \frac{1}{2}\begin{pmatrix} -11 \\ 9 \\ 9 \end{pmatrix} \qquad \nabla f_2(x^0) = \frac{1}{2}\begin{pmatrix} 9 \\ -11 \\ 9 \end{pmatrix} \qquad \nabla f_3(x^0) = \frac{1}{2}\begin{pmatrix} 9 \\ 9 \\ -11 \end{pmatrix}$$

$$\frac{1}{3}\sum_{i=1}^{3} f_i(x^0) = \frac{1}{6}\begin{pmatrix} 7 \\ 7 \\ 7 \end{pmatrix} > 0$$

$C = \mathrm{Top}1$

$$\frac{1}{3}\sum_{i=1}^{3} C(\nabla f_i(x^0)) = \frac{1}{6}\begin{pmatrix} -11 \\ -11 \\ -11 \end{pmatrix} < 0$$

разные стороны

- Техника компенсации ошибки

для каждого уст-ва e_i^k — память

$$e_i^0 = 0$$

$$e_i^1 = \nabla f_i(x^0) - C(\nabla f_i(x^0))$$

след. посылка: $C(e_i^1 + \nabla f_i(x^1))$

$$e_i^2 = \underbrace{e_i^1 + \nabla f_i(x^1)}_{\text{хотел}} - \underbrace{C(e_i^1 + \nabla f_i(x^1))}_{\text{послал}}$$

$$\boxed{\begin{array}{l} \text{посылка: } C(e_i^k + \nabla f_i(x^k)) \\ e_i^{k+1} = e_i^k + \nabla f_i(x^k) - C(e_i^k + \nabla f_i(x^k)) \end{array}}$$