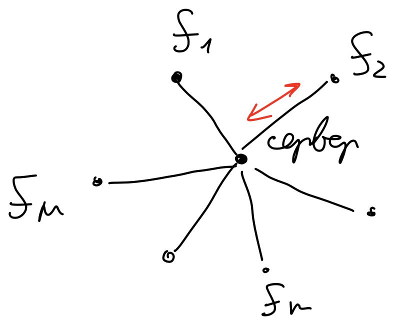
Распределенная задача:

$$\min_{x\in\mathbb{R}^d}\left[f(x) := \frac{1}{M}\sum_{m=1}^{M} f_m(x) = \frac{1}{M}\sum_{m=1}^{M} \frac{1}{n_m}\sum_{i=1}^{n_m} l(g(x,a_i^m), b_i^m)\right]$$

Модель „устр-ва - сервер":



1) x^k от сервера к устр-вам
2) $\nabla f_m(x^k)$ на устр-вах
3) $\nabla f_m(x^k)$ на сервер
4) $\nabla f(x^k) = \frac{1}{M}\sum \nabla f_m(x^k)$ на сер
5) $x^{k+1} = x^k - \gamma \nabla f(x^k)$ на сервер

общение

Проблема: ускорение в M раз не достигается

⊖ затраты на общение

⊖ несинхронность устройств

Сжатие - способ борьбы за скорость общения

Опр. Стохастический оператор $Q(x)$ будем называть несмещ. сжатием, если $\forall x\in\mathbb{R}^d$

$$\mathbb{E}[Q(x)] = x \qquad \mathbb{E}\left[\|Q(x)\|_2^2\right] \le \omega \|x\|^2$$

≥ 1

Примеры:

- Тожд. оператор $Q(x) = x$ $\quad \omega = 1$
- Выбор координат случайной

для несмещенности

базисный вектор

$$\mathrm{Rand}k(x) = \frac{d}{k}\sum_{i\in S}[x]_i e_i$$

$|S| = k$ $\quad S$ — подмножество индексов $[d]$

$$\mathbb{E}\left[\operatorname{Rand}k(x)\right] = \begin{pmatrix} \dots \\ \mathbb{E}\left[(\operatorname{Rand}k(x))_{(i)}\right] \\ \dots \end{pmatrix} =$$

$$= \begin{pmatrix} \cancel{\frac{k}{d}} \cdot \cancel{\frac{d}{k}} \cdot x_{(i)} + \left(1 - \frac{k}{d}\right) \cdot 0 \end{pmatrix}$$

p

$$= \begin{pmatrix} \dots \\ x_{(i)} \\ \dots \end{pmatrix}$$

$$\mathbb{E}\left[\|\operatorname{Rand}k(x)\|_2^2\right] = \mathbb{E}\left[\sum_{i=1}^{d} (\operatorname{Rand}k(x))_{(i)}^2\right]$$

$$= \sum_{i=1}^{d} \mathbb{E}\left[(\operatorname{Rand}k(x))_{(i)}^2\right]$$

$$= \sum_{i=1}^{d} \frac{k}{d} \cdot \left(\frac{d}{k}\right)^2 (x)_{(i)}^2 + 0$$

$$= \frac{d}{k} \underbrace{\sum_{i=1}^{d} (x)_{(i)}^2}_{\|x\|_2^2} = \frac{d}{k} \|x\|_2^2 \Rightarrow \omega = \frac{d}{k}$$

- Трехуровневая квантизация

посчитали знак

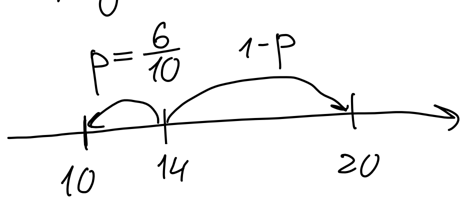
$$(Q(x))_{(i)} = \|x\|_2 \operatorname{sign}(x_{(i)}) \xi_i$$

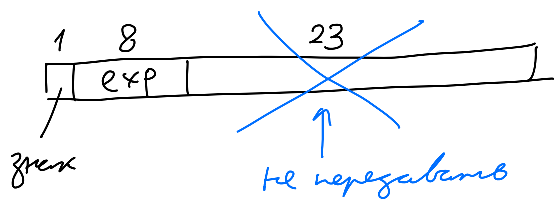
посчитали норму

$$\xi_i = \begin{cases} 1 & \text{с вер. } \frac{|x_i|}{\|x\|_2} \\ 0 & \text{иначе} \end{cases}$$

$$\omega = \sqrt{d}$$

- Округление (рандомизированное)





= округление к степени двойки

Распределенный градиентный спуск с сжатием:

$$\boxed{x^{k+1} = x^k - \gamma \cdot \frac{1}{M}\sum_{m=1}^{M} Q(\nabla f_m(x^k))}$$

при пересылке от устройств к серверу градиент сжимается

Q

сжатие от сервера к устройствам

не будем рассматр. $\rightarrow$

Док-во сходимости:

- f_m – L-гладкие
- f – μ-сильно выпуклая

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

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

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

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

$$\mathbb{E}\left[g^k \mid x^k\right] = \mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M} Q(\nabla f_m(x^k)) \mid x^k\right]$$

$$= \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}\left[Q(\nabla f_m(x^k)) \mid x^k\right]$$

несмещенность

$$= \frac{1}{M}\sum_{m=1}^{M} \nabla f_m(x^k) = \nabla f(x^k)$$

$$\ominus \langle \nabla f(x^k); x^k - x^*\rangle \qquad (**)$$

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

$$\mathbb{E}\left[\|g^k\|_2^2 \mid x^k\right] = \mathbb{E}\left[\left\| \frac{1}{M}\sum_{m=1}^{M} Q(\nabla f_m(x^k))\right\|_2^2 \mid x^k\right]$$

$$= \frac{1}{M^2}\sum_{m=1}^{M} \mathbb{E}\left[\| Q(\nabla f_m(x^k))\|_2^2 \mid x^k\right]$$

$$+ \frac{1}{M^2}\sum_{m\neq l} \mathbb{E}\left[\underbrace{\langle Q_m(\nabla f_m(x^k)); Q_l(\nabla f_l(x^k))}_{\text{независимы}} \rangle \mid x^k\right]$$

Йенсен:

$$\leq \frac{1}{M}\sum \mathbb{E}\left[\| Q(\nabla f_m(x^k))\|_2^2 \mid x^k\right]$$

$$\leq \frac{\omega}{M}\sum \|\nabla f_m(x^k)\|_2^2$$

$$= \frac{1}{M^2}\sum_{m=1}^{M} \mathbb{E}\left[\| Q(\nabla f_m(x^k))\|_2^2 \mid x^k\right]$$

$$+ \frac{1}{M^2}\sum_{m\neq l} \langle \mathbb{E}\left[Q_m(\nabla f_m(x^k))\right]; \mathbb{E}\left[Q_l(\nabla f_l(x^k))\right]\rangle$$

$$= \frac{\omega}{M^2}\sum_{m=1}^{M} \|\nabla f_m(x^k)\|_2^2$$

$$+ \frac{1}{M^2}\sum_{m\neq l} \langle \nabla f_m(x^k); \nabla f_l(x^k)\rangle$$

$$\left\|\frac{1}{M}\sum \nabla f_m(x^k)\right\|_2^2 = \frac{1}{M^2}\sum \|\nabla f_m(x^k)\|_2^2 + \frac{1}{M^2}\sum_{m\neq l} \langle \nabla f_m(x^k); \nabla f_l(x^k)\rangle$$

$$= \underbrace{\frac{\omega-1}{M^2}\sum_{m=1}^{M} \|\nabla f_m(x^k)\|_2^2 + \|\nabla f(x^k)\|_2^2}_{} \qquad (***)$$

сравни с красным выше (есть доп. эффект $\frac{1}{M}$ от нез. сжат. не упр.-вах)

Подставляем (**), (***) в (*)

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

$$+\gamma^2\, \mathbb{E}\left[\frac{\omega-1}{M^2}\sum_{m=1}^{M} \|\nabla f_m(x^k)\|_2^2 + \|\nabla f(x^k)\|_2^2\right]$$

$\pm \nabla f_m(x^*)$ КБШ $\|a+b\|_2^2 \le 2\|a\|_2^2 + 2\|b\|_2^2$

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

$$+\gamma^2\,\mathbb{E}\left[\frac{2(\omega-1)}{M^2}\sum_{m=1}^{M}\left(\|\nabla f_m(x^k) - \nabla f_m(x^*)\|_2^2 + \|\nabla f_m(x^*)\|_2^2\right)\right]$$

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

μ - сильная выпуклость и L - гладкость

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

$$+\gamma^2\,\mathbb{E}\left[\frac{2(\omega-1)}{M^2}\sum_{m=1}^{M}\left(2L\left(f_m(x^k) - f_m(x^*) + \langle \nabla f_m(x^*); x^k - x^*\rangle\right) + \|\nabla f_m(x^*)\|_2^2\right)\right]$$

$$+\gamma^2\,\mathbb{E}\left[2L\left(f(x^k) - f(x^*)\right)\right]$$

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

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

$$+\gamma^2\cdot\frac{2(\omega-1)}{M^2}\sum_{m=1}^{M}\|\nabla f_m(x^*)\|_2^2$$

$$+\gamma^2\,\mathbb{E}\left[2L\left(f(x^k) - f(x^*)\right)\right]$$

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

$$-2\gamma\left(1 - \gamma L\cdot\frac{2(\omega-1)}{M} - \gamma L\right)\mathbb{E}\left[f(x^k) - f(x^*)\right]$$

$$+\gamma^2\cdot\frac{2(\omega-1)}{M^2}\sum_{m=1}^{M}\|\nabla f_m(x^*)\|_2^2$$

$$\gamma \le \frac{1}{L\left(\frac{2(\omega-1)}{M} + 1\right)}$$

$$\le (1-\gamma\mu)\,\mathbb{E}\left[\|x^k - x^*\|_2^2\right] + \gamma^2 \cdot \frac{2(\omega-1)}{M^2}\sum_{m=1}^{M}\|\nabla f_m(x^*)\|_2^2$$

Сходимость распр. град. спуска с сжатием:

$$\boxed{\mathbb{E}\left[\|x^{k+1} - x^*\|_2^2\right] \le (1-\gamma\mu)\,\mathbb{E}\left[\|x^k - x^*\|_2^2\right] + \gamma^2 \cdot \frac{2(\omega-1)}{M^2}\sum_{m=1}^{M}\|\nabla f_m(x^*)\|_2^2}$$

⊖ сходимость, как у SGD, до окрестности

$$x^{k+1} = x^k - \gamma \cdot \frac{1}{M}\sum Q(\nabla f_m(x^k)) \qquad x^k \to x^*$$

$$\underbrace{Q(\underbrace{\nabla f_m(x^*)}_{\neq 0})}_{\text{передаем что-то } \neq 0}$$

⊖ шаг стал меньше, чем у GD

метод сходится дольше по итерациям

⊕ по количеству передаваемой информации

] $\|\nabla f_m(x^*)\| = 0$, тогда итерац. сложность $\gamma = \frac{1}{L\left(1+\frac{2(\omega-1)}{M}\right)}$

$$O\left(\frac{L}{\mu}\left(1+\frac{2(\omega-1)}{M}\right)\log\frac{1}{\varepsilon}\right) \text{ итераций}$$

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

Rand k $\omega = \frac{d}{k}$, а степень сжатия $= \frac{d}{k}$

для метода с Rand k кол-во серв. итераций.

$$O\left(\frac{L}{\mu}\left(\frac{1}{\omega} + \frac{2}{M}\right)\log\frac{1}{\varepsilon}\right)$$

может быть лучше, чем 1

$\omega \sim M$

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