• Проблема распред. оптимизации — эффективные коммуникации

На прошлых занятиях эффективность за счет пересылки меньше, не каждую итерацию

Ортогональный подход — коммун. по полной, но реже

Parallel GD/Fed Avg/Local GD:

$$x_m^{k+1} = x_m^k - \gamma \nabla f_m(x_m^k)$$ ← локальный шаг по f_m

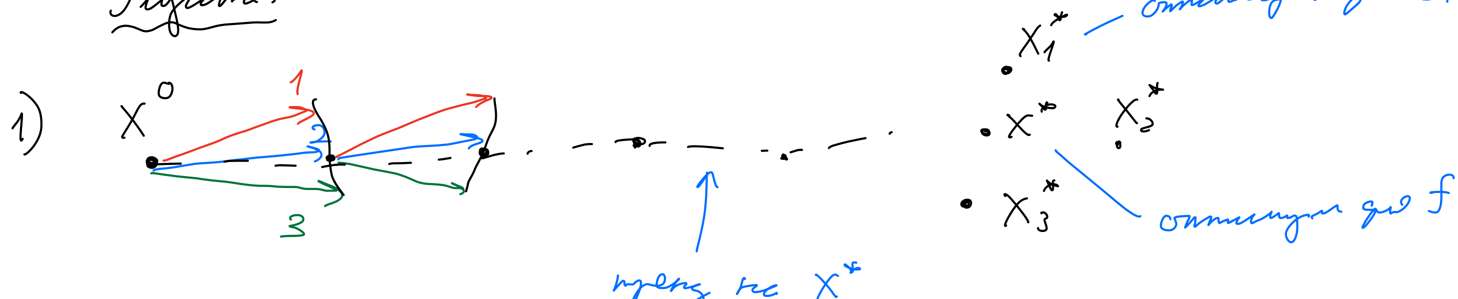
раз t итераций $x^k = \frac{1}{M}\sum_{m=1}^{M} x_m^k$, $x_m^k = x^k$ ← усреднение

⊕ снижение коммун. затрат

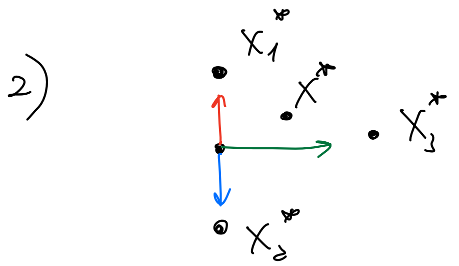
⊖ метод сходится только до окрестности

Рисунок:

1)



2)



нет тренда, все идут в свою сторону

чем больше локальных шагов, тем хуже

Теорема

$$\|\bar{x}^k - x^*\|_2^2 \le (1-\gamma\mu)^k \|x^0 - x^*\|_2^2 + \frac{\gamma \sigma_{opt}^2 (t-1)}{\mu}$$

$\bar{x}^k = \frac{1}{M}\sum_{m=1}^{M} x_m^k$

$(t-1)$ — кол-во лок. шагов

$\sigma_{opt}^2 = \frac{1}{M}\sum_{m=1}^{M} \|\nabla f_m(x^*)\|_2^2$

сходимость к окр.

Хотим метод, который сходится при фикс. шаге

Идея: локальные шаги = переобучение

регуляризация = защита от переобучения

Локальный: минимизация $f_m(x)$

⇓ регуляр.

минимизация $\tilde{f}_m(x) = f_m(x) + \frac{\lambda}{2}\|x - v\|_2^2$ (референсная точка: v)

FedProx:

$$x_m^{k+1} = x_m^k - \gamma\left(\nabla f_m(x_m^k) + \lambda(x_m^k - v^k)\right)$$

$$v^{k+1} = v^k$$

раз t итераций: $x^k = \frac{1}{M}\sum_{m=1}^{M} x_m^k \quad x_m^k = x^k \quad v^k = x^k$

FedProx:

$$x_m^{k+1} = x_m^k - \gamma\left(\nabla f_m(x_m^k) + \lambda x_m^k - \lambda v^k\right)$$

Scaffold:

$$x_m^{k+1} = x_m^k - \gamma\left(\nabla f_m(x_m^k) \underbrace{- c_m^k + c^k}_{\text{редукция дисперсии}}\right)$$

⊕ плюс локального GD

⊕ сходимость до решения

Проблема локальных методов:

они не лучше, чем GD

Но в случае, когда данные локальные (f_m) похожи, методы доказуемо могут быть лучше

- Как описать неопределённость?

1) $\|\nabla f_m(x) - \nabla f(x)\|_2 \le \Delta \quad \forall x$

для квадр. задачи: $\|A_m x - A x\|_2 \sim \|A_m - A\|_2 \|x\|_2$ $x \in \mathbb{R}^d$, но Δ не сущ.

2) $\|\nabla^2 f_m(x) - \nabla^2 f(x)\|_2 \le \delta \quad \forall x$

для квадрат: $\|A_m - A\|_2 \le \delta$

Оценим δ:

- $f(x) = \frac{1}{M}\sum_{m=1}^{M} f_m(x) = \frac{1}{M}\sum_{m=1}^{M}\left[\frac{1}{N}\sum_{i=1}^{N} \underbrace{L(g(x, a_i^m), b_i^m)}_{\ell(x, c_i^m)}\right]$ (для всех m и i)

- выборка на f_m берется равномерно из всей выборки в f

- нер-во Хефдинга (матричное)

$\{X_i\}_{i=1}^N$ – i.i.d. матрицы размера d

$X_i^T = X_i \qquad \mathbb{E}[X_i] = 0 \qquad X_i^2 \preceq A^2$

тогда с вероятностью $1-p$

$$\left\|\frac{1}{N}\sum X_i\right\|_2 \le \sqrt{\frac{8\|A\|_2^2}{N}\ln\frac{d}{p}}$$

↑ плохой: $\left\|\frac{1}{N}\sum X_i\right\|_2 \le \frac{1}{N}\sum\|X_i\|_2 \le \frac{1}{N}\sum\|A\|_2 = \|A\|_2$ (треуг.)

- Хефдинг для гессианов

$$f_m(x) = \frac{1}{N}\sum_{i=1}^{N}\ell(x, c_i^m)$$

$\Rightarrow \|\underbrace{\nabla^2 f_m(x) - \nabla^2 f(x)}\|_2$

- $\mathbb{E} = 0$ (из-за равномерности)
- ℓ – L-гладкая, то $\nabla^2\ell \preceq L I$

- $X_i = \nabla^2 f_i - \nabla^2 f$
- $\mathbb{E} X_i = 0, \quad X_i^2 \le A^2 = 4L^2 \cdot I$

$$\Downarrow$$

$$\| \nabla^2 f_m(x) - \nabla^2 f(x) \|_2 \le \delta \sim \frac{L}{\sqrt{N}}$$

δ меньше, если выборка на ус-ве больше

- Как в методе достичь точности?

Метод зеркального спуска:

$$x^{k+1} = \operatorname{argmin}_{x \in \mathbb{R}^d} \{ \gamma \langle \nabla f(x^k), x \rangle + V(x, x^k) \}$$

$$V(x,y) = \varphi(x) - \varphi(y) - \langle \nabla \varphi(y), x - y \rangle$$

для задачи с точностью мех. функции:

$$\varphi(x) = f_1(x) + \frac{\delta}{2} \|x\|_2^2$$

argmin становится подзадач. = нужен солвер

решение argmin на устройстве с f_1

всего 1 коммуникация за итерацию

Сходимость

$\gamma = 1$

$$V(x^*, x^k) \le \left(1 - \frac{\mu}{\mu + 2\delta}\right)^k V(x^*, x^0)$$

μ — сильная выпуклость f

$\sim \varepsilon$

для достижения точности ε нужно

$$k = \frac{2\delta + \mu}{\mu} \log \frac{V(x^*, x^0)}{\varepsilon}$$

для GD

$$k = \frac{L}{\mu} \log \frac{\|x^0 - x^*\|_2^2}{\varepsilon}$$

$$\delta \sim \frac{L}{\sqrt{N}}$$

⊕ с точки зрения памяти может быть сильное улучшение

⊕ с точки зрения локальных вычислений
- на сервере (f_1) $\frac{L}{\mu} \log \frac{1}{\varepsilon}$ (как у GD)
- на ост. уз-ах $\frac{\delta}{\mu} \log \frac{1}{\varepsilon}$ (лучше чем у GD)

⊕ вычисления на сервере

NB достаточно, что f_1 и f δ-похожи

⊖ открытый вопрос см. в презентации.