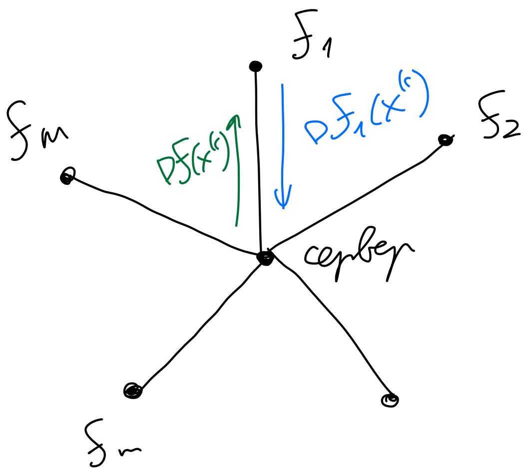
Задача распредел. оптимизации (ERM распред.):

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

данные на уст-ве m

- f_m – на устр-ве m (∇f_m, f_m можно вычислять только на m устр-ве)



Распределённый градиентный спуск

1) $\nabla f_m(x^k)$ на своем уст-ве
2) устр-ва $\nabla f_m(x^k)$ на сервер
3) сервер получает посылки и усредняет $\frac{1}{M} \sum_{m=1}^{M} \nabla f_m(x^k) = \nabla f(x^k)$
4) сервер посылает $\nabla f(x^k)$ на все устройства
5) устр-ва делают $x^{k+1} = x^k - \gamma \nabla f(x^k)$

Проблема: коммуникации – дорого и мешают полной параллелизации

Подход N1 – Локальный

Общаться каждую итерацию = дорого
Давайте общаться не каждую итерацию

Parallel GD/FedAvg/Local GD:

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

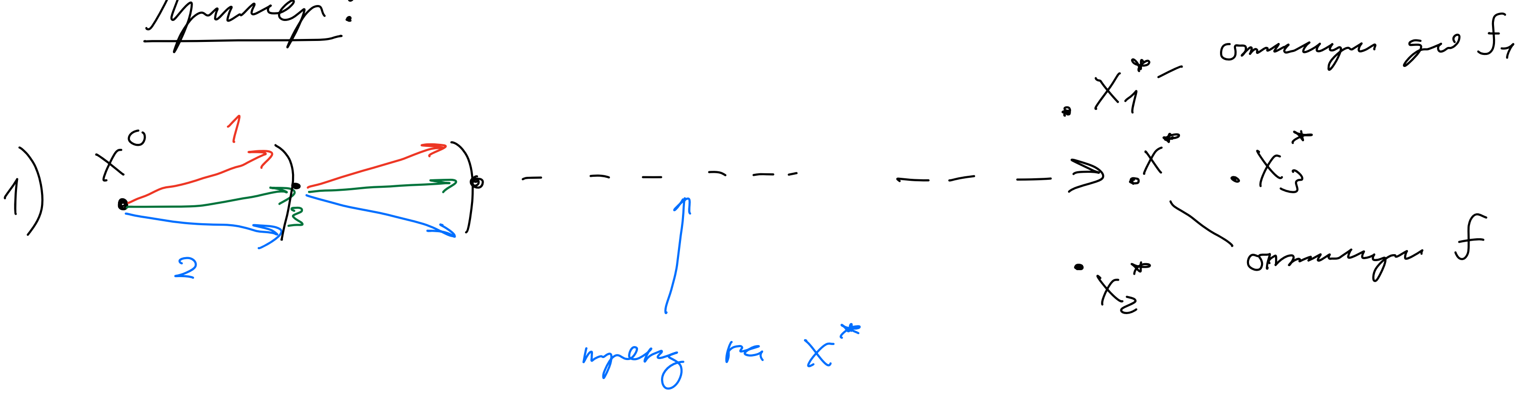
раз t итераций $x^k = \frac{1}{M} \sum_{m=1}^{M} x_m^k$, $x_m^k = x^k$

(+) снижение коммуникац. затрат

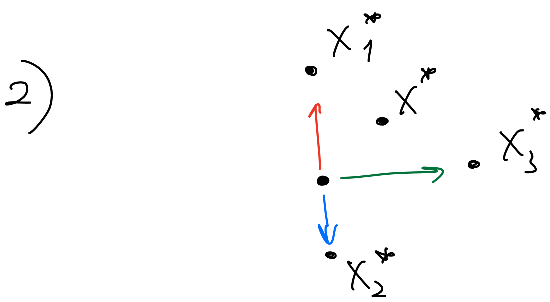
(−) метод сходится только до окрестности

Пример:

1)



2)



нет тренда, все смотрят в свои стороны

чем больше локальных шагов, тем хуже

Теория

кол-во лок. шагов

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

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

сход. до окр.

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

Хотим метод, чтобы сходился при фикс. шаге?

Идея: локальные шаги = переобучение

регуляризация = защиту от переобучения

локальный шаг: минимизация $f_m(x)$

$\Downarrow$

референсная точка

минимизация $F_m(x) = f_m(x) + \frac{\lambda}{2}\|x - v\|_2^2$

Fed Prox:

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

$$\vartheta^{k+1} = \vartheta^k$$

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

FedProx:

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

Scaffold

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

⊕ плюс локального GD

⊕ сходимость до решения

Проблема локальных методов:

они не лучше, чем GD (см. нижние оценки)

Но случае похожести данных (f_m) локальные методы доказуемо могут давать улучшение

• Как описать похожесть?

1) $\|\nabla f_m(x) - \nabla f(x)\| \leq \Delta$

для квадр. задачи: $\|A_m x - Ax\| \sim \|A_m - A\| \|x\|$

x не ограничено, т.е. Δ не сущ.

2) $\|\nabla^2 f_m(x) - \nabla^2 f(x)\| \leq \delta$

для квадр: $\|A_m - A\| \leq \delta$

не нер-во Хефдинга

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

L — константа Липшица градиента f

N — кол-во данных в f_m

Снова метод зеркального спуска:

$$x^{k+1} = \arg\min_{x \in \mathbb{R}^d} \{ \gamma \langle \nabla f(x^k); x \rangle + V(x, x^k) \}$$

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

для задачи похожести в лок. методах

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

Решение argmin на устройстве с f_1

Всего 1 коммуникация за итерацию

Сходимость метода зеркального спуска:

$\gamma = 1$

сильная выпуклость f

$$\underbrace{V(x^*, x^k)}_{\sim \varepsilon} \leq \left(1 - \frac{\mu}{\mu + 2\delta}\right)^k V(x^*, x^0)$$

похожесть данных

для достижения точности ε нужно

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

جسو GD

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

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