Координатные методы распределенной оптимизации в условиях гомогенности данных.

Обзор литературы

Подготовила: Алимаскина Екатерина

Участники проекта: Максимов Роман, Алимаскина Екатерина

Руководитель: Былинкин Дмитрий, МФТИ

11.02.2025

«Распределенной оптимизации»

Algorithm Централизованный GD

Вход: Размер шага $\gamma > 0$, стартовая точка $w_0 \in \mathbb{R}^d$

for $k = 0, 1, \dots, K-1$ **do**

  Отправить w_k всем рабочим

  for $i = 1, \dots, n$ параллельно **do**

    Принять w_k от мастера

    Вычислить градиент $\nabla f_m(w_k)$ в точке w_k

    Отправить $\nabla f_m(w_k)$ мастеру

  end for

  Принять $\nabla f_m(w_k)$ от всех рабочих

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

  $w_{k+1} = w_k - \gamma \nabla f(w_k)$

end for

Выход: w^K

- Один мастер, много рабочих
- Узкое место – коммуникация. Хочется делать ее дешевле и реже.

«Распределенной оптимизации»

Оптимальный алгоритм одновременно и по числу коммуникаций и по кол-ву вызовов локальных градиентов

$$\min_{x\in\mathbb{R}^d} r(x) = \underbrace{f_1(x)}_{:=q(x)} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}[f_i(x) - f_1(x)]}_{:=p(x)}$$

Algorithm 1 Accelerated Extragradient

1: **Input:** $x^0 = x_f^0 \in \mathbb{R}^d$
2: **Parameters:** $\tau \in (0,1), \eta, \theta, \alpha > 0, K \in \{1, 2, \ldots\}$
3: **for** $k = 0, 1, 2, \ldots, K-1$ **do**
4: $x_g^k = \tau x^k + (1-\tau)x_f^k$
5: $x_f^{k+1} \approx \arg\min_{x\in\mathbb{R}^d}\left[A_\theta^k(x) := p(x_g^k) + \langle\nabla p(x_g^k), x - x_g^k\rangle + \frac{1}{2\theta}\|x - x_g^k\|^2 + q(x)\right]$
6: $x^{k+1} = x^k + \eta\alpha(x_f^{k+1} - x^k) - \eta\nabla r(x_f^{k+1})$
7: **end for**
8: **Output:** x^K

Optimal Gradient Sliding and its Application to Distributed Optimization Under Similarity

Dmitry Kovalev
KAUST
dakovalev1@gmail.com

Aleksandr Beznosikov
MIPT
anbeznosikov@gmail.com

Ekaterina Borodich
MIPT
borodich.ed@phystech.edu

Alexander Gasnikov
MIPT
gasnikov@yandex.ru

Gesualdo Scutari
Purdue University
gscutari@purdue.edu

Reference	Communication complexity	Local gradient complexity
This paper	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}\log\frac{1}{\varepsilon}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}\right)$

r(x) – выпуклая
q(x) – гладкая, выпуклая
p(x) – гладкая, возможно невыпуклая

https://arxiv.org/pdf/2205.15136

«Гомогенности данных»

Function similarity («похожесть»)

Assumption 6. *$f_1(x) \dots, f_n(x)$ are δ-related: $\|\nabla^2 f_i(x) - \nabla^2 f(x)\| \leq \delta$, for all i and $x \in \mathbb{R}^d$, and some $\delta > 0$.*

Откуда мы берем такое предположение?

Если мы случайно и одинаково распределяем данные по устройствам, то это вполне естественно.

Как это использовать?

Из Assumption 6 следует, что $\|\nabla^2 p\| \leq \delta$, то есть p имеет L_p- липшицев градиент, где $L_p = \delta$.

Но может так случиться, что разные куски данных похожи неодинаково!

«Координатные»

$$\min_{x,y} f(x,y)$$

$$\begin{cases} \mathcal{O}\left(\sqrt{L/\mu_x}\log\frac{1}{\epsilon}\right) \text{ calculations of } \nabla_x f \\ \mathcal{O}\left(\sqrt{L/\mu_y}\log\frac{1}{\epsilon}\right) \text{ calculations of } \nabla_y f \end{cases}$$

Литература:

https://arxiv.org/pdf/1212.0873

https://arxiv.org/pdf/2212.14439