УДК 519.853.62, 519.676, 519.853.3

Распределенная оптимизация с композитом в условиях гомогенности данных

Е.А. Алимаскина[1], ***Р.А. Максимов***[1], ***Д.А. Быстров***[1]

Д. А. Былинкин[1,2]

[1]Московский физико-технический институт (национальный исследовательский университет)

[2]Институт системного программирования РАН

Современная оптимизация активно применяется в задачах федеративного обучения, где целевая функция имеет вид:

$$h(x) = \frac{1}{|M_f|} \sum_{m \in M_f} h_m(x)$$

В такой постановке основное узкое место — не вычислительная, а коммуникационная сложность: передача градиентов между сервером, на котором хранится $h_1(x)$, и клиентами, отвечающими за $(h - h_1)(x)$, является дорогой операцией. Ключевым понятием оказывается *похожесть* (гомогенность) данных, которую по [1] можно формализовать как

$$\|\nabla^2 h_1(x) - \nabla^2 h(x)\| \leq \delta_h$$

Общепринято считать, что данные на сервере каким-то образом отражают распределение данных в узлах. Существует множество подходов к учёту этой гомогенности (см., например, [3, 4, 5, 6]), однако сочетание похожести и композитной структуры целевой функции пока открытая и интересная проблема.

На практике сервер нередко по-разному аппроксимирует клиентские выборки в различных «модах». Например, рентгеновских снимков здоровых пациентов гораздо больше, чем изображений редких патологий. Это естественно приводит к композитной постановке:

$$h(x) = \frac{1}{|M_f|} \sum_{m \in M_f} f_m(x) + \frac{1}{|M_g|} \sum_{m \in M_g} g_m(x)$$

отражающей структуру данных, состоящую из «частых» (f_m) и «редких» (g_m) режимов. Если группа $\{f_m\}_{m \in M_f}$ близка к серверным данным $\delta_f << \delta_g$, к ней можно обращаться значительно реже, тем самым сокращая объем передаваемой информации. Поскольку размер множества M_g обычно существенно меньше, дополнительное уменьшение числа запросов к M_f ещё сильнее снижает коммуникационные издержки.

Текущие методы с оптимальными оценками, такие как Accelerated ExtraGradient [2] зависят от $\max\{\delta_f, \delta_g\}$. Наша работа предлагает учитывать неоднородность данных:

$$\|\nabla^2 f_1(x) - \nabla^2 f(x)\| \leq \delta_f,$$
$$\|\nabla^2 g_1(x) - \nabla^2 g(x)\| \leq \delta_g,$$

и использовать смещенный аппроксиматор градиента:

$$e_t = \begin{cases} p \cdot \nabla(h - h_1)(x_t), & p \\ e_{t-1} + p\left[\nabla(g - g_1)(x_t) - \nabla(g - g_1)(x_{t-1})\right], & 1 - p \end{cases}$$

Идея использования данного аппроксиматора заключается в том, что с вероятностью $1 - p$ мы пропускаем коммуникацию по f, что снижает количество обращений к "частым" клиентским данным и уменьшает коммуникационные затраты при сохранении достаточной точности оценки полного градиента.

Нам удалось решить задачу разделения коммуникационных сложностей в выпуклой постановке, что представляет собой важный результат, поскольку это минимально необходимая структура, при которой возможно использовать идею близости гессианов.

Также в ходе экспериментального исследования была проведена серия обучений: от простой полносвязной линейной модели на датасете MNIST до значительно более глубокой и архитектурно сложной модели ResNet, обученной на датасете CIFAR. В перспективе планируется расширение экспериментов на трансформерные архитектуры, основанное на предположении, что на более глубоких слоях блоков Self-Attention формируются более однородные представления.

В рамках доклада будут представлены теоретические гарантии сходимости, полученные для выпуклого случая. Мы рассмотрим основные элементы доказательства, а также ключевые технические леммы, лежащие в основе результата. Кроме того, обсудим переход к сильно выпуклой постановке: сформулируем основные трудности и возможные направления дальнейшего анализа, включая перспективы получения линейной скорости сходимости.

Литература

1. *Hadrien Hendrikx,* [et. al] *(2002).* Statistically Preconditioned Accelerated Gradient Method for Distributed Optimization. //arXiv:2002.10726
2. *Kovalev D.,* [et. al] *(2022).* Optimal Gradient Sliding and its Application to Distributed Optimization Under Similarity // arXiv:2205.15136
3. *Karimireddy,* [et. al] *(2020).* Scaffold: Stochastic controlled averaging for federated learning. // In International conference on machine learning, pp. 5132–5143. PMLR
4. *Luo, R.,* [et. al] Revisiting localsgd and scaffold: Improved rates and missing analysis. // arXiv preprint arXiv:2501.04443, 2025.
5. *Khaled A., Jin C. (2022).* Faster federated optimization under second-order similarity. // arXiv preprint arXiv:2209.02257
6. *Bylinkin D., Beznosikov A. (2024).* Accelerated Methods with Compressed Communications for Distributed Optimization Problems under Data Similarity // arXiv:2412.16414