Методы малоранговых разложений в распределенном и федеративном обучении

Ребриков Алексей
rebrikov.av@phystech.edu

Зыль Александр

Безносиков Александр
beznosikov.an@phystech.edu

Abstract

Подходы распределенного и федеративного обучения становятся все более популярными в обучении современных SOTA моделей машинного обучения. При этом на первый план выходит вопрос организации эффективных коммуникаций, так как процесс передачи информации занимает слишком много времени даже в случае кластерных вычислений. Из-за этого может теряться смысл в распределении/распараллеливании процесса обучения. Одной из ключевой техник борьбы с коммуникационными затратами является использование сжатий передаваемой информации. На данный момент в литературе предлагаются различные техники сжатия ([Beznosikov et al., 2023], [Alistarh et al., 2017], [Horvóth et al., 2022]), но потенциал в этом вопросе явно не исчерпан. В частности, довольно большой потенциал кроется в малоранговых разложениях [Gundersen, 2019]. В рамках проекта предлагается сконструировать операторы сжатия на основе данных разложений и встроить в методы распределенной оптимизации [Richtárik et al., 2021].

Keywords сжатие информации · малоранговые разложения · распределенное обучение · федеративное обучение

1 Введение

Цель данного исследования заключается в разработке и анализе методов малоранговых разложений для сжатия информации в контексте распределенного и федеративного обучения. Мотивация исследования проистекает из растущей потребности в эффективных методах обучения для современных масштабных моделей машинного обучения, где коммуникационные затраты становятся критическим барьером для эффективности. Объектом исследования являются операторы сжатия, основанные на малоранговых разложениях, и их интеграция в методы распределенной оптимизации.

Проводится обзор существующей литературы и анализируются последние достижения в области сжатия информации для распределенного обучения. В частности, рассматриваются существующие техники сжатия, такие как предложенные в работах [Beznosikov et al., 2023], [Alistarh et al., 2017], и [Horvóth et al., 2022], а также исследуется потенциал малоранговых разложений.

Задачами проекта являются разработка операторов сжатия на основе малоранговых разложений, их интеграция в алгоритмы распределенной оптимизации и оценка влияния на эффективность обучения. Предлагаемое решение предполагает новизну в виде конкретной реализации сжатия, которая потенциально позволяет уменьшить коммуникационные затраты без значительной потери качества обучения.

Цель эксперимента состоит в демонстрации эффективности предлагаемых методов на реальных наборах данных и в различных условиях обучения, оценке улучшения скорости и качества обучения.

2 Определение оптимизационной задачи и ее решение

Для достижения высоких результатов современные модели машинного обучения тренируются на больших наборах данных, что часто требует обширного числа обучаемых параметров. Рассматриваем задачи оптимизации вида

$$\min_{x \in \mathbb{R}^d} \left\{ f(x) \coloneqq \frac{1}{n} \sum_{i=1}^{n} f_i(x) \right\}, \tag{1}$$

где $x \in \mathbb{R}^d$ представляет параметры модели, n — количество работников/устройств, а $f_i(x)$ — функции потерь модели x на данных, хранимых на устройстве i. Функция потерь $f_i : \mathbb{R}^d \to \mathbb{R}$ часто имеет вид

$$f_i(x) \coloneqq \mathrm{E}_{\xi \sim \mathcal{P}_i} \left[f_\xi(x) \right],$$

где $\mathcal{P}_i$ обозначает распределение данных обучения, принадлежащих работнику i.

2.1 Распределенная оптимизация

Основой для решения задачи (1) является распределенный градиентный спуск (GD), выполняющий обновления по формуле

$$x^{k+1} = x^k - \frac{\eta^k}{n} \sum_{i=1}^{n} \nabla f_i(x^k),$$

где $\eta^k > 0$ — шаг. Для решения проблем коммуникации в распределенных системах были предложены улучшения, сокращающую размер передаваемых сообщений с помощью операторов сжатия.

2.2 Оператор сжатия

Под оператором сжатия имеется ввиду (возможно стохастическое) отображение$\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$ с некоторыми ограничениями. Обычно в литературе упоминаются несмещённые операторы сжатия $\mathcal{C}$ с ограниченным вторым моментом, т.е.

Definition 1 Пусть $\zeta \geq 1$. Будем говорить что $\mathcal{C} \in \mathbb{U}(\zeta)$ если $\mathcal{C}$ несмещённый (т.е., $\mathrm{E}\left[\mathcal{C}(x)\right] = x \; \forall x$) и если второй момент ограничен

$$\mathrm{E}\left[\left\| \mathcal{C}(x) \right\|_2^2 \right] \leq \zeta \left\| x \right\|_2^2, \qquad \forall x \in \mathbb{R}^d. \tag{2}$$

Далее в работе рассматривается конструирование операторов сжатия на основе малоранговых разложений.

Список литературы

Aleksandr Beznosikov, Samuel Horváth, Peter Richtárik, and Mher Safaryan. On biased compression for distributed learning. Journal of Machine Learning Research, 24(276):1–50, 2023.

Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. Qsgd: Communication-efficient sgd via gradient quantization and encoding. Advances in neural information processing systems, 30, 2017.

Samuel Horvóth, Chen-Yu Ho, Ludovit Horvath, Atal Narayan Sahu, Marco Canini, and Peter Richtárik. Natural compression for distributed deep learning. In Mathematical and Scientific Machine Learning, pages 129–141. PMLR, 2022.

Gregory Gundersen. Randomized singular value decomposition, 2019. URL `https://gregorygundersen.com/blog/2019/01/17/randomized-svd/`.

Peter Richtárik, Igor Sokolov, and Ilyas Fatkhullin. Ef21: A new, simpler, theoretically better, and practically faster error feedback. Advances in Neural Information Processing Systems, 34:4384–4396, 2021.