
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. 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

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

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 $f_i : \mathbb{R}^d \rightarrow \mathbb{R}$ часто имеет вид

$$f_i(x) := \mathbb{E}_{\xi \sim \mathcal{P}_i} [f_\xi(x)],$$

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 $\mathcal{C} : \mathbb{R}^d \rightarrow \mathbb{R}^d$ 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 $\mathcal{C} \in \mathcal{U}(\zeta)$ если \mathcal{C} несмещённый (т.е., $\mathbb{E}[\mathcal{C}(x)] = x \ \forall x$) и если второй момент ограничен

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

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

3 Вычислительный эксперимент

Целью данного эксперимента является сравнение эффективности известных и новых малоранговых операторов сжатия градиентов в контексте распределенного обучения с применением алгоритма EF21 [Richtárik et al., 2021]. Исследование направлено на оценку влияния использования различных операторов сжатия на скорость и качество обучения моделей машинного обучения.

3.1 Описание базового набора данных

В эксперименте используется набор данных Mashrooms, взятый из источника [Chang and Lin, 2011]. Данный набор содержит 8124 записи, разделенные на 2 класса, с 112 признаками каждая.

3.2 План эксперимента

Основной план эксперимента заключается в построении графика, который демонстрирует зависимость точности модели от объема переданной информации. Такой подход позволит количественно оценить, насколько эффективно операторы сжатия уменьшают объем необходимых данных без значительной потери в качестве обучения.

3.3 Предварительный отчёт

Ожидается, что при применении операторов сжатия без алгоритма EF21 модель может вообще не сходиться, а после применения будет так же сходиться, но с большим числом итераций. А вот если смотреть на зависимость от количества переданной информации, то ожидается более быстрая сходимость при применении операторов сжатия и EF21.

3.4 Теория

Algorithm 1 EF21 (Multiple nodes)

```

1: Input: starting point  $x^0 \in \mathbb{R}^d$ ;  $g_i^0 = \mathcal{C}(\nabla f_i(x^0))$  for  $i = 1, \dots, n$  (known by nodes and the master); learning rate
    $\gamma > 0$ ;  $g^0 = \frac{1}{n} \sum_{i=1}^n g_i^0$  (known by master)
2: for  $t = 0, 1, 2, \dots, T - 1$  do
3:   Master computes  $x^{t+1} = x^t - \gamma g^t$  and broadcasts  $x^{t+1}$  to all nodes
4:   for all nodes  $i = 1, \dots, n$  in parallel do
5:     Compress  $c_i^t = \mathcal{C}(\nabla f_i(x^{t+1}) - g_i^t)$  and send  $c_i^t$  to the master
6:     Update local state  $g_i^{t+1} = g_i^t + \mathcal{C}(\nabla f_i(x^{t+1}) - g_i^t)$ 
7:   end for
8:   Master computes  $g^{t+1} = \frac{1}{n} \sum_{i=1}^n g_i^{t+1}$  via  $g^{t+1} = g^t + \frac{1}{n} \sum_{i=1}^n c_i^t$ 
9: end for

```

Richtárik et al. [2021]

Algorithm 2 Алгоритм сжатия данных с использованием HOSVD

```

1: Input: vector  $x \in \mathbb{R}^d$ .
2: Reshape the vector  $x$  into a tensor  $\mathcal{T}(x) = \text{reshape}(x, \text{dims})$ .
3: Apply HOSVD to the tensor  $\mathcal{T}(x)$  to obtain its decomposition.
4: Truncate the ranks of the tensor decomposition to  $(r_1, r_2, \dots, r_k)$ .
5: Send the compressed tensor to the master.
6: Master reconstructs the full tensor from its decomposition.
7: Master reshapes the tensor back into the vector form.

```

5

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 Horváth, 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.
- 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.
- Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27, 2011.

1. Цитаты в аннотации делают, но не очень часто.
Уточните у руководителя, если Александр советует делать так,
то ок. Но нужно уточнить

2. Не нашел обзора, где он проводится? Видимо,
его нужно добавить сюда

3. Что-то сломалось здесь

4. Mushrooms

5. Написано хорошо, но мало.

Не нашел:

- обзора литературы
- результатов экспериментов
- теории (секция 3.4 не в счет - приведение алгоритмов без комментариев не считается.
И обычно теорию ставят до эксперимента)
- Выводов и заключения