Методы малоранговых разложений в распределенном и федеративном обучении

Алексей Витальевич Ребриков

Московский физико-технический институт

Курс: Автоматизация научных исследований (практика, В. В. Стрижов)/Группа 105

Эксперт: к.ф.-м.н. А. Н. Безносиков

Консультант: А. В. Зыль

2024

Цель исследования

Решается задача

распределённой оптимизации с использованием EF21

$$\min_{x \in \mathbb{R}^d} \left\{ f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x) \right\} \qquad x^{k+1} = x^k - \frac{\eta^k}{n} \sum_{i=1}^{n} \nabla f_i(x^k)$$

Основная идея

сжатие градиентов с использованием малоранговых разложений (например HOSVD)

Цель

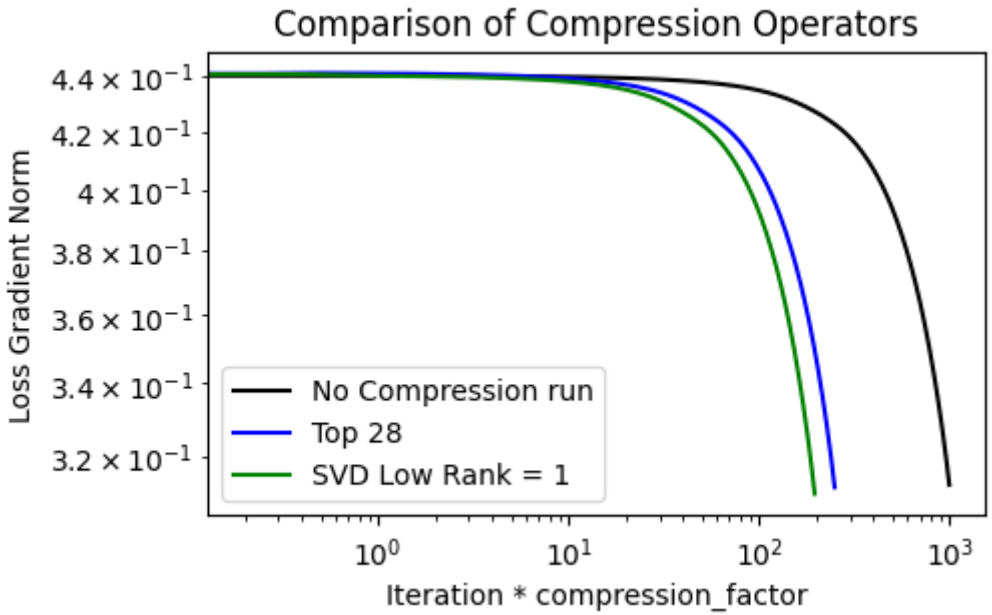
сравнить известные операторы сжатия с новыми

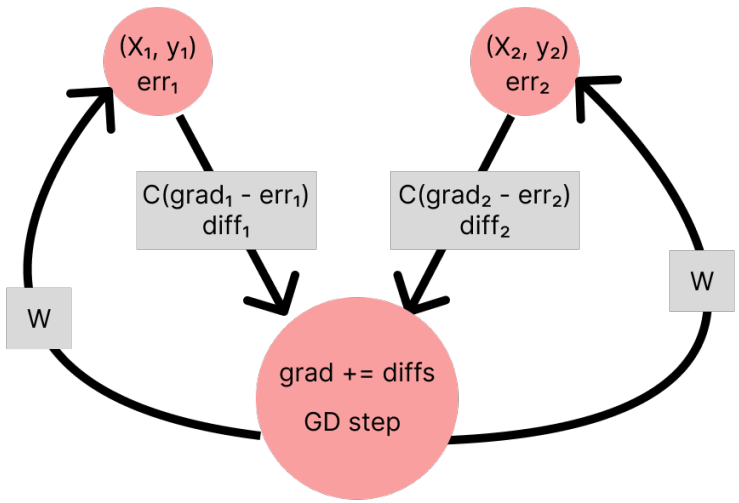
Доклад с одним слайдом

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

Ключевые пункты:

1. Данные распределяются.
2. Веса модели общие.
3. Передача градиентов сжимается.
4. Сжимается не сам градиент, а его разность. (EF21)





Литература

Alistarh, D., Hoefler, T., Johansson, M., Konstantinov, N., Khirirat, S., and Renggli, C. (2018).

The convergence of sparsified gradient methods.

Advances in Neural Information Processing Systems, 31.

Beznosikov, A., Horváth, S., Richtárik, P., and Safaryan, M. (2023).

On biased compression for distributed learning.

Journal of Machine Learning Research, 24(276):1–50.

De Lathauwer, L., De Moor, B., and Vandewalle, J. (2000).

A multilinear singular value decomposition.

SIAM journal on Matrix Analysis and Applications, 21(4):1253–1278.

Richtárik, P., Sokolov, I., and Fatkhullin, I. (2021).

Ef21: A new, simpler, theoretically better, and practically faster error feedback.

Advances in Neural Information Processing Systems, 34:4384–4396.

Algorithm EF21 (Multiple nodes)

Input: starting point $x^0 \in \mathbb{R}^d$; $g_i^0 = \mathcal{C}(\nabla f_i(x^0))$ for $i = 1, \ldots, n$ (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)
for $t = 0, 1, 2, \ldots, T-1$ **do**
    Master computes $x^{t+1} = x^t - \gamma g^t$ and broadcasts x^{t+1} to all nodes
    for all nodes $i = 1, \ldots, n$ **in parallel do**
        Compress $c_i^t = \mathcal{C}(\nabla f_i(x^{t+1}) - g_i^t)$ and send c_i^t to the master
        Update local state $g_i^{t+1} = g_i^t + \mathcal{C}(\nabla f_i(x^{t+1}) - g_i^t)$
    end for
    Master computes $g^{t+1} = \frac{1}{n}\sum_{i=1}^n g_i^{t+1}$ via $g^{t+1} = g^t + \frac{1}{n}\sum_{i=1}^n c_i^t$
end for

[Richtárik et al., 2021]

Оператор сжатия HOSVD

Определение

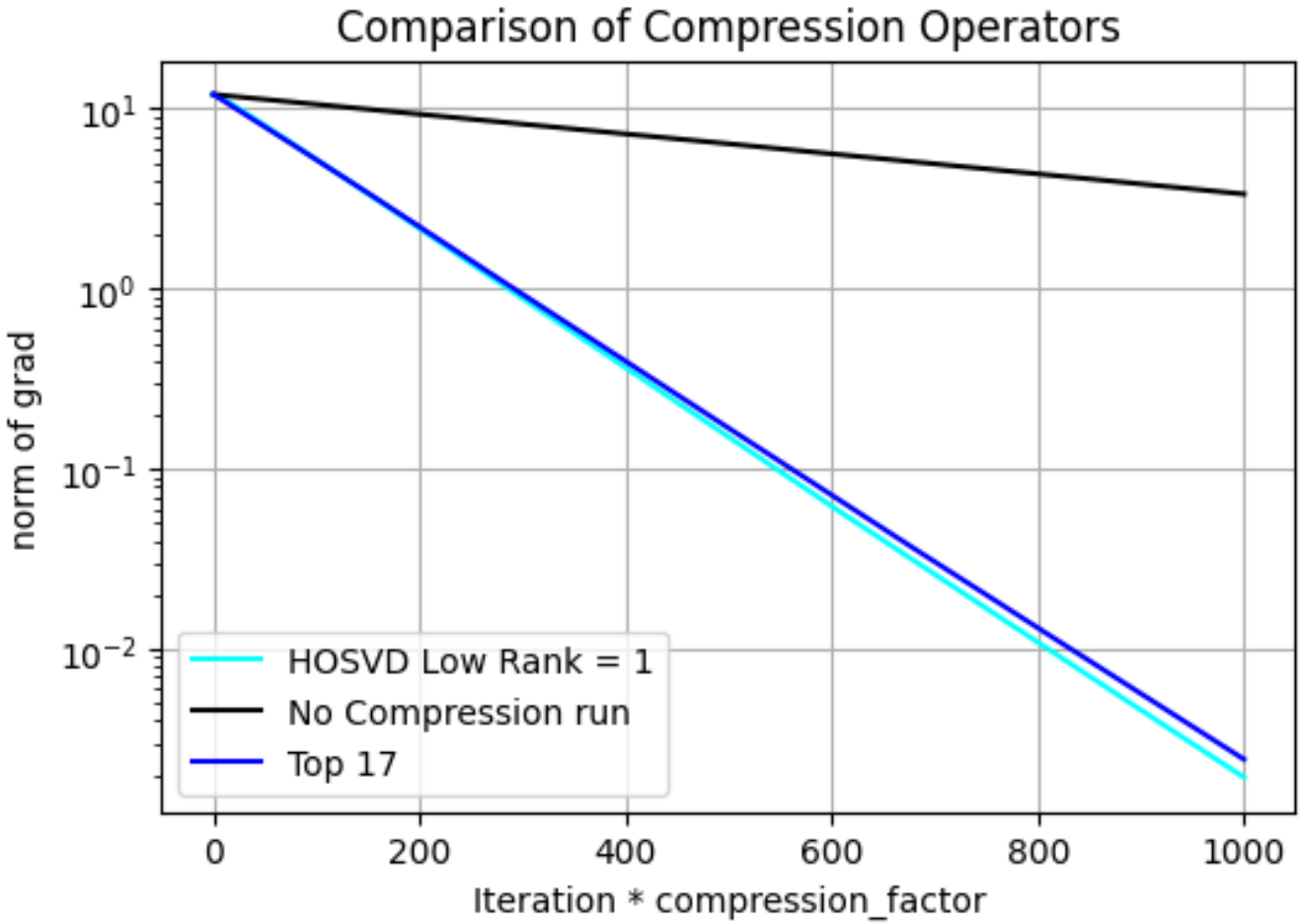
Под оператором сжатия понимается (возможно стохастическое) отображение $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$, которое удовлетворяет определённым ограничениям на изменение информации.

Algorithm 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 $(r_1, r_2, \ldots, r_k)$.
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.

Вычислительный эксперимент

Сравнение нормы градиента от величины пропорциональной переданной информации



HOSVD оператор работает наравне с используемым top-k

Заключение

Результаты

- предложен новый метод сжатия,
- проведено сравнение с Top-k [Alistarh et al., 2018]

Будущаяя работа

- перенести код на PyTorch,
- применить для сложных сетей (свёрточных),
- исследовать другие разложения (TT, Tucker, etc.)