Адаптивное сжатие в распределенной оптимизации

A Preprint

Хафизов Фанис Адикович
Физтех-школа Прикладной Математики и Информатики
Московский Физико-Технический Институт
г. Долгопрудный
khafizov.fa@phystech.edu

Безносиков Александр Николаевич
Московский Физико-Технический Институт
https://anbeznosikov.github.io/

Abstract

В данной работе рассматривается проблема распределённого обучения больших моделей (например, современных нейросетей), когда вычисления необходимо распараллеливать между несколькими устройствами. Основная сложность в таких системах заключается в высокой стоимости коммуникации при передаче больших объёмов градиентов. Мы предлагаем семейство операторов адаптивного сжатия, которые учитывают важность координат и тем самым снижают трафик, сохраняя качество сходимости. В экспериментальной части показано, что предлагаемые операторы могут работать не хуже классических вариантов `RandK` и `TopK`, а в ряде случаев достигают сопоставимого качества с `TopK`.

Keywords Распределённая оптимизация · Сжатие градиентов · Нейронные сети · Метод стохастического градиента

1 Введение

Современные нейросетевые архитектуры требуют значительных вычислительных мощностей и объёмов памяти. Для ускорения обучения такие модели обычно тренируют на нескольких устройствах (GPU/TPU и т. д.), каждый из которых вычисляет градиенты на своей части данных. В итоге возникает проблема обмена градиентами между устройствами, так как полная пересылка всех координат может стать узким местом.

Цель данной работы — предложить новый подход к адаптивному сжатию градиентов, основанный на идее взвешивания координат по их важности, и проверить его эффективность на задачах логистической регрессии и классификации изображений нейросетью.

2 Связанные работы

В литературе известно несколько типов операторов сжатия:

- `RandK` — случайный выбор k координат из d;
- `TopK` — выбор k крупнейших по модулю координат.

Эти методы существенно сокращают объём передаваемых данных, однако могут приводить к смещённым оценкам градиента и замедлять сходимость. Поэтому активно исследуются смещённые операторы

сжатия, которые, с одной стороны, уменьшают трафик, а с другой — обладают хорошими свойствами сходимости при грамотном учёте структуры данных.

3 Постановка задачи

Рассмотрим задачу минимизации функции

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

где n — число устройств (воркеров), $f_i(x)$ — функция потерь на i-м устройстве. Типичная схема распределённого градиентного спуска (DCGD) с учётом сжатия выглядит так:

$$x^{k+1} = x^k - \frac{\eta}{n}\sum_{i=1}^{n} \mathcal{Q}\big(\mathcal{C}_i^k(\nabla f_i(x^k))\big), \tag{2}$$

где $\mathcal{C}_i^k$ — оператор сжатия на i-м устройстве в момент времени k, а $\mathcal{Q}$ — опциональный оператор квантизации (например, округление до ближайшей степени двойки).

4 Предлагаемый метод

4.1 Идея важности координат

Ключевая идея — ввести вектор важности $w \in [0,1]^d$, который присваивает каждой координате некоторый «вес»:

$$w^k = \arg\min_{w\in\Delta_d} f\big(x^k - \eta\,(w \odot \nabla f(x^k))\big),$$

где $\Delta_d = \{\, w \in \mathbb{R}^d \,|\, w_i \geq 0, \sum_i w_i = 1\}$ и $\odot$ обозначает поэлементное умножение. Оптимизацию по w можно проводить методом зеркального спуска. После нахождения w^k оператор сжатия учитывает не просто значение координаты, а её «важность».

4.2 Примеры операторов

Мы предлагаем несколько вариантов на базе w:

- MD Stochastic: случайный отбор k координат с вероятностью w_i;
- MD Greedy: отбор k координат с наибольшими w_i;
- MD Greedy Weighted: масштабирование координат x_i дополнительно на w_i;
- MD Weighted Greedy: выбор по наибольшим значениям $|w_i\, x_i|$.

В результате получается смещённая оценка градиента, однако весовые коэффициенты w помогают выделять более важные координаты.

5 Эксперименты

Мы провели эксперименты на задаче логистической регрессии (`mushrooms` датасет) и на задаче классификации изображений (`CIFAR10` с моделью ResNet18).

5.1 Логистическая регрессия

На Рис. 1 приведены результаты для доли передаваемых координат $k/d = 0.2$. Виден выигрыш операторов, учитывающих важность, по сравнению с `RandK`.

5.2 Классификация (ResNet18 на CIFAR10)

На Рис. 2 показаны результаты обучения ResNet18. Предлагаемые методы демонстрируют качество, близкое к `TopK`, но требуют меньше объёма передачи данных по сравнению с полной пересылкой.

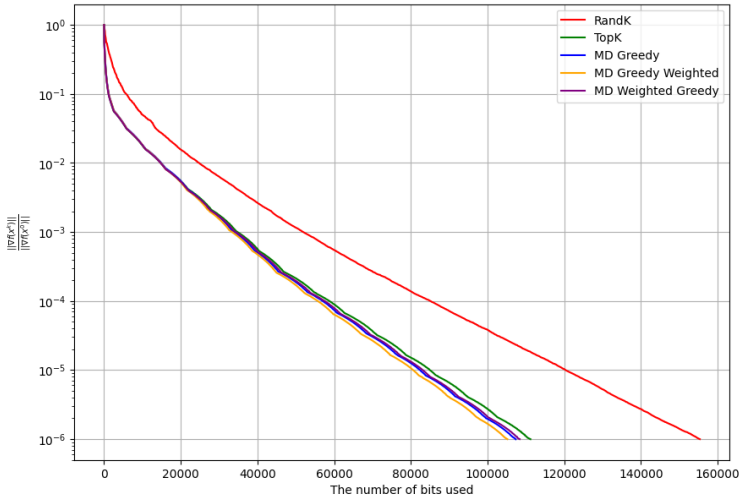

Рис. 1: Сходимость алгоритма логистической регрессии (датасет mushrooms).

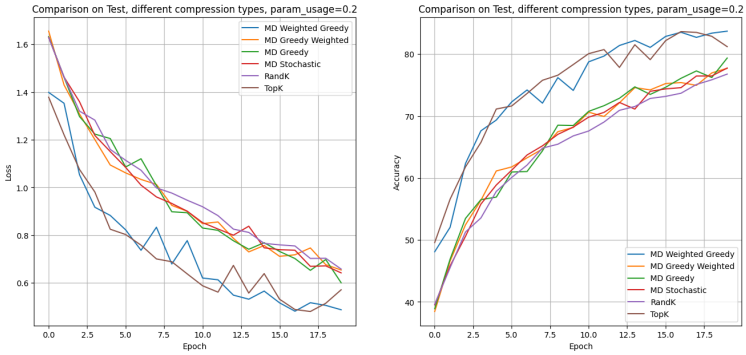

Рис. 2: Сходимость при обучении ResNet18 на CIFAR10 с долей $k/d = 0.2$.

6 Выводы

- Введено семейство операторов сжатия, использующих вектор важности w.
- Эксперименты показывают, что данные операторы, особенно `MD Weighted Greedy`, дают сравнимую сходимость с классическим `TopK`.
- В дальнейших исследованиях планируется:
  1. Улучшать качество сходимости за счёт дополнительной теоретической настройки параметров;
  2. Расширять эксперименты на более сложных нейросетевых архитектурах.