Адаптивное сжатие в распределенной оптимизации

Фанис Адикович Хафизов
Научный руководитель: к.ф.-м.н. А. Н. Безносиков

Кафедра интеллектуальных систем ФПМИ МФТИ
Специализация: Интеллектуальный анализ данных
Направление: 03.03.01 Прикладные математика и физика

2024

Цель исследования

Проблема: Современные нейросети требуют больших вычислительных мощностей, из-за чего приходится прибегать к распределенным методам обучения.
Одной из главных проблем является скорость передачи данных между устройствами.

Цель: Предложить новый способ сжатия градиентов для более эффективной коммуникации устройств.

Решение: Предлагается семейство смещенных операторов сжатия, использующие показатели важности весов.

Литература

- Aleksandr Beznosikov, Samuel Horváth, Peter Richtárik, Mher Safaryan. On Biased Compression for Distributed Learning. 2024
- Nam Nguyen, Deanna Needell, Tina Woolf. Linear Convergence of Stochastic Iterative Greedy Algorithms with Sparse Constraints. 2014

Постановка задачи

Ставится задача распределенной оптимизации.

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

где n – количество устройств, x – обучаемые параметры, $f_i(x)$ – функция потерь для i-го устройства.

Требуется сократить количество передаваемой информации, не сильно потеряв в скорости обучения.

Исследуемый метод

Для решения задачи распределенной оптимизации используется градиентный спуск со сжатием (DCGD)

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

где η – размер шага, $\mathcal{C}_i^k$ – оператор сжатия на k-й итерации i-го устройства.

Для случая одного устройства итерация метода запишется как

$$x^{k+1} = x^k - \eta \mathcal{C}^k(\nabla f(x^k)).$$

Исследуемый метод

Для еще большего сокращения размера передаваемой информации добавим оператор квантизации $\mathcal{Q}$.
Метод DCGD тогда перепишется:

$$x^{k+1} = x^k - \frac{\eta}{n}\sum_{i=1}^{n} \mathcal{Q}\left(\mathcal{C}_i^k(\nabla f_i(x^k))\right).$$

Для одного устройства:

$$x^{k+1} = x^k - \eta\mathcal{Q}\left(\mathcal{C}^k(\nabla f(x^k))\right).$$

В качестве оператора $\mathcal{Q}$ выбрано округление до ближайшей степени 2.

Базовое решение

В качестве базового решения рассматриваются операторы сжатия:

- RandK:

$$\mathcal{C}(x) := \frac{d}{k} \sum_{i \in S} x_i e_i,$$

где $S \subseteq [d]$ – случайный поднабор индексов размера k.

- TopK:

$$\mathcal{C}(x) := \frac{d}{k} \sum_{i=d-k+1}^{d} x_{(i)} e_{(i)},$$

где координаты расположены по неубыванию модуля $|x_{(1)}| \leq |x_{(2)}| \leq \dots \leq |x_{(d)}|$.

Предлагаемые операторы сжатия

Предлагается определить вектор важности $w \in [0,1]^d$ и на основе него построить семейство операторов сжатия $\mathcal{C}(x, w)$.

Для нахождения вектора w решается задача оптимизации

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

где $\Delta_d = \left\{ x \in \mathbb{R}^d \mid \sum_{i=1}^{d} x_i = 1, x_i \geq 0 \ \forall i = \overline{1, d} \right\}$ – вероятностный симплекс, $\odot$ – поэлементное умножение.

Задача нахождения w^k решается методом зеркального спуска на вероятностном симплексе.

Примеры операторов сжатия с важностью

- MD Stochastic

$$\mathcal{C}(x, w) := \frac{d}{k} \sum_{i \in S} x_i e_i,$$

где $S \subseteq [d]$ – случайный поднабор индексов размера k, вероятность выбора i-й компоненты равна w_i.

- MD Greedy

$$\mathcal{C}(x, w) := \frac{d}{k} \sum_{i=d-k+1}^{d} x_{(i)} e_{(i)},$$

где координаты расположены по неубыванию w: $w_{(1)} \leq w_{(2)} \leq \cdots \leq w_{(d)}$.

Примеры операторов сжатия с важностью

- MD Greedy Weighted

$$\mathcal{C}(x, w) := \frac{d}{k} \sum_{i=d-k+1}^{d} w_{(i)} x_{(i)} e_{(i)},$$

где координаты расположены по неубыванию w: $w_{(1)} \leq w_{(2)} \leq \cdots \leq w_{(d)}$.

- MD Weighted Greedy

$$\mathcal{C}(x, w) := \frac{d}{k} \sum_{i=d-k+1}^{d} w_{(i)} x_{(i)} e_{(i)},$$

где координаты расположены по неубыванию взвешенных модулей: $|w_{(1)} x_{(1)}| \leq |w_{(2)} x_{(2)}| \leq \cdots \leq |w_{(d)} x_{(d)}|$.

Эксперимент. Логистическая регрессия

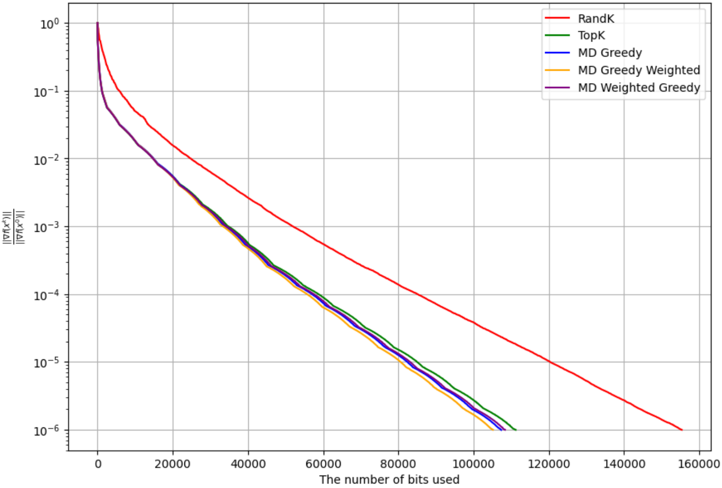

Рис.: Сравнение операторов сжатия, задача логистической регрессии на датасете mushrooms, доля передаваемых компонент $k/d = 0.2$

Эксперимент. Классификация изображений

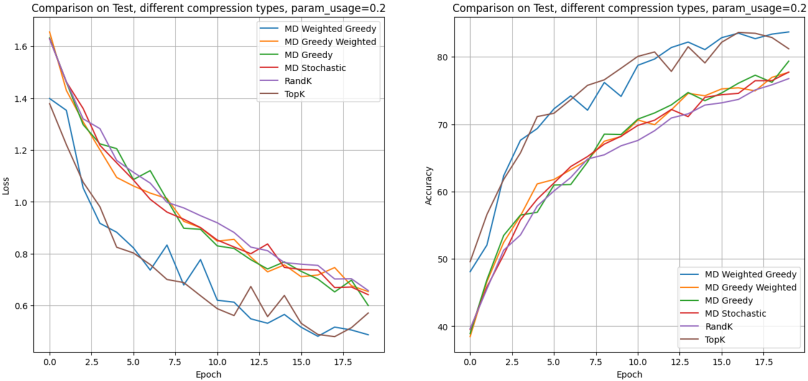

Рис.: Сравнение операторов сжатия, задача классификации изображений на датасете CIFAR10, модель ResNet18, доля передаваемых компонент $k/d = 0.2$

Заключение

Имеющиеся результаты

1. Предложено семейство операторов сжатия, использующие вектор важности. Из него приведены 4 примера операторов.
2. Вычислительные эксперименты показали, что предложенные операторы работают не хуже RandK, оператор MD Weighted Greedy работает на уровне TopK.

Будущая работа

1. Продолжать эксперименты, получить сходимость лучше TopK.
2. Развить теорию для описанного семейства смещенных операторов.