

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

2025

Адаптивное сжатие в распределенной оптимизации

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: Предлагаются семейство смещенных операторов сжатия, использующие показатели важности весов, и схема компенсации ошибок.

- ▶ Aleksandr Beznosikov, Samuel Horváth, Peter Richtárik, Mher Safaryan. On Biased Compression for Distributed Learning. 2024
- ▶ Peter Richtárik, Igor Sokolov, Ilyas Fatkhullin. EF21: A New, Simpler, Theoretically Better, and Practically Faster Error Feedback. 2021

Задача распределенной оптимизации

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

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.

Требования на f_i

Предположение

Функция f называется μ -сильно выпуклой, если для любых $x, y \in \mathbb{R}^d$ выполняется следующее неравенство:

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|_2^2, \quad (2)$$

где $\mu > 0$ — константа сильной выпуклости.

Предположение

Функция f называется L -гладкой, если для любых $x, y \in \mathbb{R}^d$ выполняется следующее неравенство:

$$\|\nabla f(y) - \nabla f(x)\|_2 \leq L \|y - x\|_2, \quad (3)$$

где $L > 0$ — константа гладкости.

Распределенный градиентный спуск со сжатием (DCGD)

$$x^{k+1} = x^k - \frac{\eta}{n} \sum_{i=1}^n \mathcal{C}_i^k(\nabla f_i(x^k)) \quad (4)$$

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

$$x^{k+1} = x^k - \eta \mathcal{C}^k(\nabla f(x^k)). \quad (5)$$

Новые операторы сжатия

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

Для нахождения вектора w решается задача оптимизации

$$w^k = \arg \min_{w \in Q} f(x^k - \eta w \odot \nabla f(x^k)), \quad (6)$$

где Q – ограниченное множество, \odot – поэлементное умножение. В качестве Q рассмотрены следующие варианты:

- ▶ $\Delta_{d-1} = \{x \in \mathbb{R}^d \mid \|x\|_1 = d, x_i \geq 0, i = \overline{1, d}\}$ – симплекс размерности $d - 1$;
- ▶ $[a, b]^d$ – куб со сторонами длины $b - a$.

Примеры операторов сжатия с важностью

- ▶ **Важностное прореживание.**

$$\mathcal{C}(\nabla f(x)) = \sum_{i=1}^k \nabla_{(i)} f(x) e_{(i)}, \quad (7)$$

где координаты расположены по убыванию значений важности $w_{(1)} \geq w_{(2)} \geq \dots \geq w_{(d)}$.

- ▶ **Рандомизированное важностное прореживание**

$$\mathcal{C}(\nabla f(x)) = \sum_{i \in S} \nabla_i f(x) e_i, \quad (8)$$

где S — множество индексов, выбранных случайно с вероятностью, пропорциональной значениям важности w .

Примеры операторов сжатия с важностью

- ▶ Важностное прореживание с перевзвешиванием

$$\mathcal{C}(\nabla f(x)) = \sum_{i=1}^k w_{(i)} \nabla_{(i)} f(x) e_{(i)}, \quad (9)$$

где координаты расположены по убыванию значений $|w_{(1)} \nabla_{(1)} f(x)| \geq \dots \geq |w_{(d)} \nabla_{(d)} f(x)|$.

- ▶ Рандомизированное важностное прореживание с перевзвешиванием

$$\mathcal{C}(\nabla f(x)) = \sum_{i \in S} w_i \nabla_i f(x) e_i, \quad (10)$$

где S — множество индексов, выбранных случайно с вероятностью, пропорциональной значениям важности w .

Сходимость DCGD с оператором $ImpK$ (9)

Теорема (Хафизов Ф. А., 2025)

Пусть $Q = [1, 2]^d, \gamma \leq \frac{1}{L}$. Тогда

$$f(x^T) - f^* \leq \left(1 - \mu\gamma(1 - L\gamma)\frac{k}{d}\right)^T (f(x^0) - f^*). \quad (11)$$

Следствие

При выборе $\gamma = \frac{1}{2L}$ для достижения точности $\varepsilon > 0$ по функции требуется

$$T \geq \frac{4L}{\mu} \frac{d}{k} \log \frac{f(x^0) - f^*}{\varepsilon} \quad (12)$$

итераций.

Схема компенсации ошибок SCAM

Алгоритм 1 SCAM (Одно устройство)

Ввод: стартовая точка x^0 , шаг обучения γ , количество итераций T , начальная ошибка $\varepsilon^0 = 0$.

for $t = 0, 1, \dots, T - 1$ **do**

$$g^t = \nabla f(x^t)$$

$$c = \mathcal{C}^t(\varepsilon^t + g^t)$$

$$\tilde{g}^t = \mathcal{C}^t \left(\sum_{i=1}^d I\{c_i \neq 0\} g_i^t e_i \right)$$

$$\varepsilon^t = \varepsilon^{t-1} + g^t - \tilde{g}^t$$

$$x^{t+1} = x^t - \gamma \tilde{g}^t$$

end for

Вывод: x^k

Сходимость SCAM с оператором *TopK*

Предположение

В схеме SCAM TopK частично сохраняет норму градиента:

$$\|\tilde{g}^t\|_2^2 \geq \delta \|g^t\|_2^2. \quad (13)$$

Теорема (Хафизов Ф. А., 2025)

Пусть $\gamma \leq \frac{2}{L}$. Тогда для любого $T \geq 1$ верно:

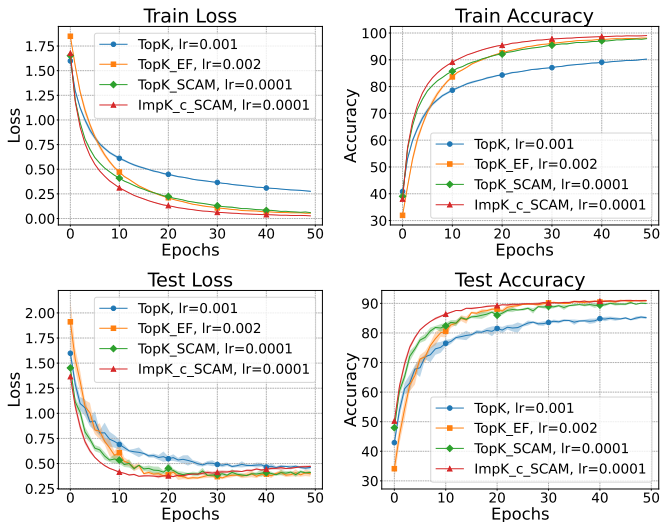
$$f(x^T) - f^* \leq (1 - \mu\gamma(2 - L\gamma)\delta)^T (f(x^0) - f^*). \quad (14)$$

Следствие

Пусть выбрано значение шага $\gamma = \frac{1}{L}$, $\delta \simeq \frac{k}{d}$. Тогда для достижения точности $\varepsilon > 0$ по функции требуется

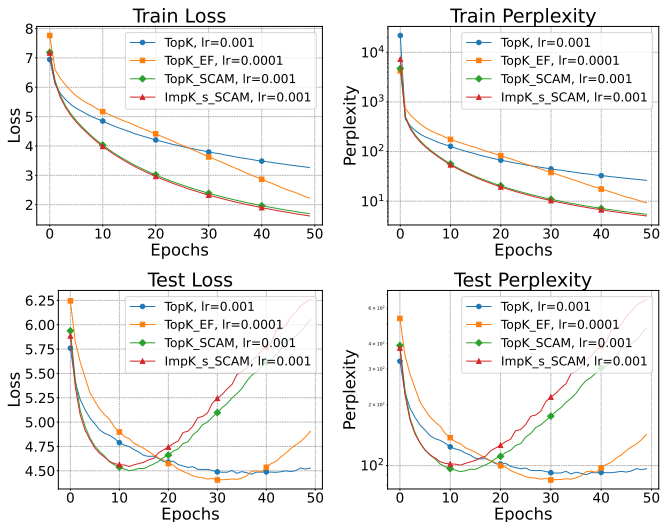
$$T \simeq \frac{Ld}{\mu k} \log \left(\frac{f(x^0) - f^*}{\varepsilon} \right). \quad (15)$$

Эксперимент. Классификация изображений



Сравнение сходимости предложенного метода SCAM с $ImpK_c$ и вариациями $TopK$ в процессе обучения ResNet-18 на CIFAR-10.

Эксперимент. Генерация текста



Сравнение сходимости предложенного метода SCAM с $ImpK_s$ и вариациями $TopK$ в процессе обучения GPT-2 на WikiText2.

1. Предложено семейство операторов сжатия, использующие вектор важности.
2. Для одного оператора сжатия по важности получена оценка сходимости.
3. Предложена схема компенсации ошибки SCAM.
4. Получена теоретическая оценка сходимости для SCAM с оператором *TopK*.
5. Вычислительные эксперименты показали превосходство схемы SCAM с *ImpK* над остальными методами.