Research on the combination of Top-K and Perm-K gradient sparsification algorithms for distributed setting

K. Acharya[1] T. Kharisov[1]

[1] Department of Applied Mathematics and Informatics
Moscow Institute of Physics and Technology

MIPT conference, April 5 2023

Table of Contents

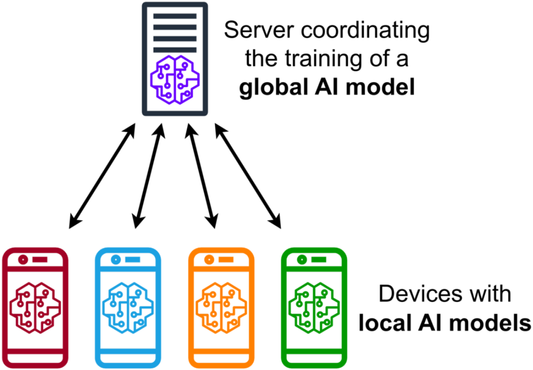
1 Introduction

2 Experiments

3 Theoretical results

4 Q&A

Federated learning



Communication cost is a bottleneck for the Federated Learning approach: worker devices use unstable and slow networks such as WIFI and Cellular.

Introduction

- Distributed optimization methods/machine learning methods require efficient organization of communications, since communications in this case very often take up most of the time of the algorithm.
- To reduce the cost of one communication, you can apply compression of the transmitted information.
- Different Techniques: Random Approaches, Greedy Approaches
- In this work, we want to combine the greedy approach of Top-k and the random approach of Perm-k algorithms for better performance

Problem statement

- We consider optimization problems of the form

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

where $x \in \mathbb{R}^d$ collects the parameters of a statistical model to be trained, n is the number of workers/devices, and $f_i(x)$ is the loss incurred by model x on data stored on worker i.

- A general baseline for solving problem is distributed gradient descent, performing updates of the form

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

where $\eta^k > 0$ is a stepsize.

Compressors review

- **Paper:** On Biased Compression for Distributed Learning (Aleksandr Beznosikov, Samuel Horváth, Peter Richtárik, Mher Safaryan)
- **Main contribution:** Distributed SGD with Biased Compression and Error Feedback Algorithm

Definition

Top-k

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

where coordinates are ordered by their magnitudes so that $|x_{(1)}| \leq |x_{(2)}| \leq \cdots \leq |x_{(d)}|$.

Definition

Error Feedback $e_i^{k+1} = e_i^k + \nabla f_i(x^k) - C(e_i^k + \nabla f_i(x^k))$

Compressors review

- **Paper:** Permutation Compressors for Provably Faster Distributed Nonconvex Optimization (Rafał Szlendak, Alexander Tyurin, Peter Richtárik)
- **Main contribution:** Construction of the new compressors based on the idea of a random permutation (Perm K).
  Provably reduce the variance caused by compression beyond what independent compressors can achieve.

Definition

(Perm K for $d \geq n$). Assume that $d \geq n$ and $d = qn$, where $q \geq 1$ is an integer. Let $\pi = (\pi_1, \ldots, \pi_d)$ be a random permutation of $\{1, \ldots, d\}$. Then for all $x \in \mathbb{R}^d$ and each $i \in \{1, 2, \ldots, n\}$ we define

$$\mathcal{C}_i(x) := n \cdot \sum_{j=q(i-1)+1}^{qi} x_{\pi_j} e_{\pi_j}.$$

Motivation

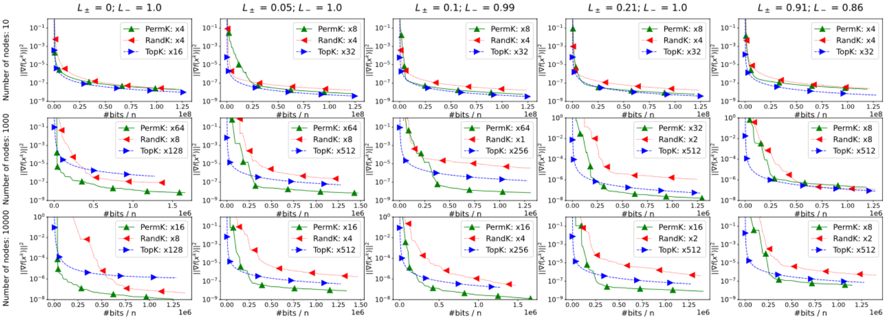


Figure 1: Comparison of algorithms on synthetic quadratic optimization tasks with nonconvex $\{f_i\}$.

On a simple quadratic optimization problem it is seen that TopK and PermK approaches compete with each other mainly depending on the number of nodes.

Table of Contents

1. Introduction

2. Experiments

3. Theoretical results

4. Q&A

Quadratic optimization problem

- First we consider the quadratic optimization problem (defined previously) with:

$$f_i(x) := \frac{1}{2}x^TA_ix - b_i^Tx, \tag{1}$$

  where $A_i \in \mathbb{R}^{d\times d}$, $b \in \mathbb{R}^d$.
- Matrix generation A_i
- Implementation of Error Feedback to the algorithms

Comparison of different EF approaches

Notations

C - *TopK* biased compressor.

Q_i^k - *PermK* unbiased compressor for i-th node on k-th step.

Error Feedback 0

$$e_i^{k+1} = e_i^k + \nabla f_i(x^k) - C(Q_i^k(e_i^k + \nabla f_i(x^k)))$$

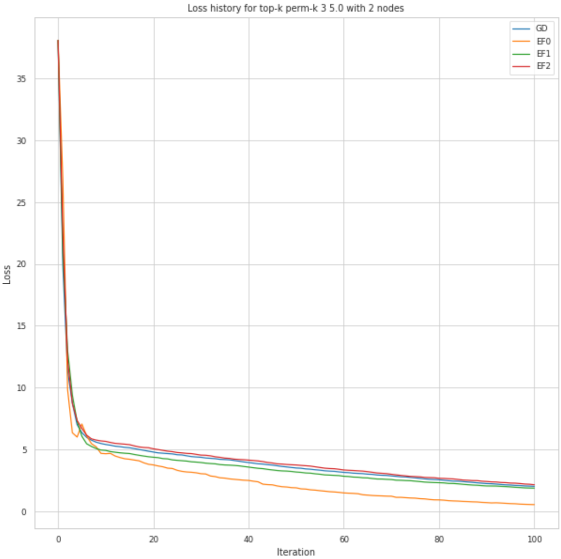
Error Feedback 1

$$e_i^{k+1} = e_i^k + Q_i^k(\nabla f_i(x^k)) - C(Q_i^k(e_i^k + \nabla f_i(x^k)))$$

Error Feedback 2

$$e_i^{k+1} = Q_i^k(e_i^k + \nabla f_i(x^k)) - C(Q_i^k(e_i^k + \nabla f_i(x^k)))$$

Comparison of different EF approaches



Observable compressors

TopK + Error Feedback

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

where coordinates are ordered by their magnitudes so that $\left|x_{(1)}\right| \leq \left|x_{(2)}\right| \leq \cdots \leq \left|x_{(d)}\right|$.

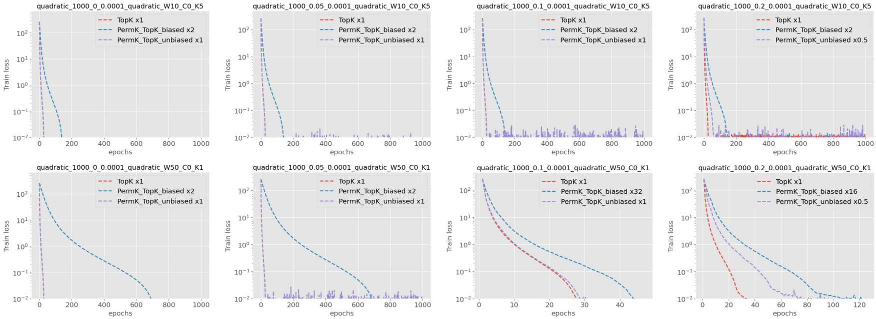
Unbiased TopK-PermK

$\mathcal{C}(x) := \mathcal{T}_k \circ \mathcal{P}_q(x) \cdot n$

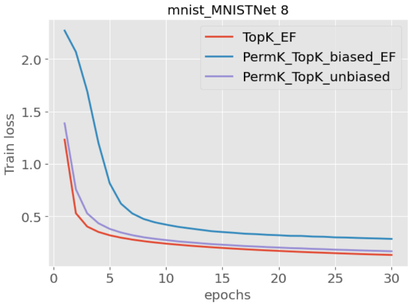
Biased TopK-PermK + Error Feedback

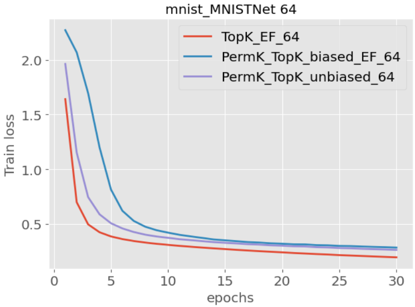
$\mathcal{C}(x) := \mathcal{T}_k \circ \mathcal{P}_q(x)$

Quadratic problem experiment reproduction

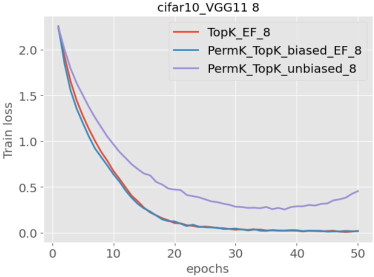


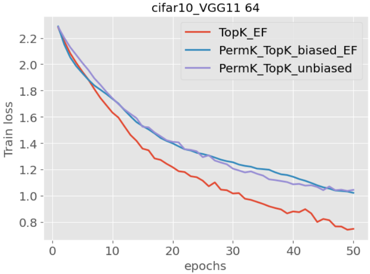
MnistNet comparison

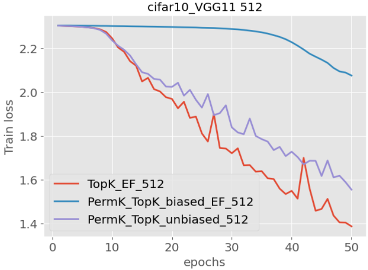




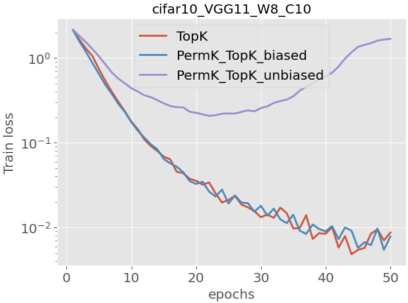
CIFAR-10 + VGG11


cifar10_VGG11 8


cifar10_VGG11 64


cifar10_VGG11 512

Dataset mix-up (common 10%)



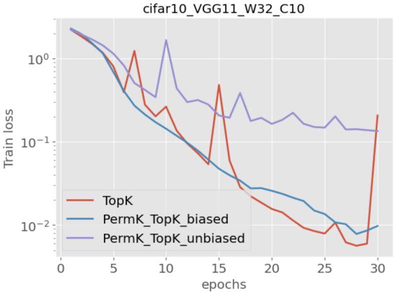


Table of Contents

1. Introduction

2. Experiments

3. Theoretical results

4. Q&A

Biased classes

Biased compressor class

We say $C \in \mathbb{B}^3(\delta)$ for some $\delta > 1$ if

$$\mathrm{E}\left[\|C(x) - x\|_2^2\right] \leq \left(1 - \frac{1}{\delta}\right) \|x\|_2^2, \quad \forall x \in \mathbb{R}^d$$

Bounds for compressors

- *TopK:* $(1 - \frac{1}{\delta}) = \frac{d-k}{d}$ [Alistarh et al., 2018a]
- *TopK-PermK:* $(1 - \frac{1}{\delta}) = \frac{d-k}{q}$, where q is PermK parameter [NEW]

For the second approach bound is worse $\frac{d}{q} = n$ times, where n is a number of nodes (workers).

Theorem

Error Feedback theorem [Beznosikov et al., 2023]

Let $\left\{x^k\right\}_{k\geq 0}$ denote the iterates of EF for solving SGD problem, where each f_i is L-smooth and μ-strongly convex. Let $x^\star$ be the minimizer of f and let $f^\star := f\left(x^\star\right)$ and

$$D := \frac{1}{n}\sum_{i=1}^{n}\left\|\nabla f_i\left(x^\star\right)\right\|_2^2 .$$

$\mathcal{O}(1)$ stepsizes & equal weights. Let, for all $k \geq 0$, the stepsizes and weights be set as $\eta^k = \eta$ and $w^k = 1$, respectively, where $\eta \leq \frac{1}{14(2\delta+B)L}$. Then

$$\mathrm{E}\left[f\left(\bar{x}^K\right)\right] - f^\star = \mathcal{O}\left(\frac{A_1}{K} + \frac{A_2}{\sqrt{K}}\right)$$

where $A_1 := L(2\delta + B)\left\|x^0 - x^\star\right\|_2^2$ and
$A_2 := \sqrt{C(1 + 1/n) + D(2B/n + 3\delta)}\left\|x^0 - x^\star\right\|_2$.

Counterexamples

Counterexample 1

Assume the vector $x \in \mathrm{R}^d$ where k coordinates equal to 1, others equal to 0.

- Then $\mathrm{P}(P(x)_i = 1) = \frac{q}{d}$.
- $\mathrm{E}\left[\| T(P(x)) - x\|_2^2\right] = k * (1 - \frac{q}{d}) = \frac{kd-q}{d}$

Counterexample 2

- The same vector x, but in case the matrices of f_i will be independent of each other.
- Then it is clear that *TopK* will be guaranteed to take these coordinates into hidden gradients and in fact will transmit complete information for convergence.
- *PermK* will spoil the picture in this case. Each of the nonzero coordinates will be taken with a probability of $1/n$ and the final convergence will be slower by about n times.

Further assumptions

- Then I would like to consider those cases when the functions on the nodes will be similar to each other. There are many ideas on how to achieve this, but it is not clear which one should be chosen.
- One of best ideas is to inspect the example when f_i Hessians are similar to each other.

Hessian variance

Let $L_{\pm} \geq 0$ be the smallest quantity such that

$$\frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(x) - \nabla f_i(y)\|^2 - \|\nabla f(x) - \nabla f(y)\|^2 \leq L_{\pm}^2 \|x - y\|^2, \quad \forall x, y \in \mathbb{R}^d.$$

We can refer to the quantity $L_{\pm}^2$ by the name Hessian variance.

Table of Contents

1. Introduction
2. Experiments
3. Theoretical results
4. Q&A

Q&A

Your questions, please!