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

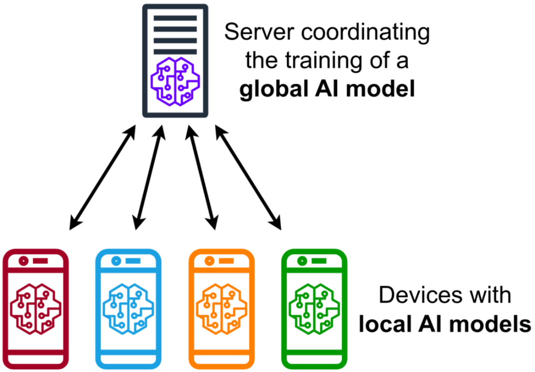
T. Kharisov[1] K. Acharya[1] A. Beznosikov[1]

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

Science Practice Conference, May 2 2023

Table of Contents

Federated learning



Credit: *Wikipedia*

Figure: Federated Learning scheme

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

Brief introduction

- Distributed optimization methods/machine learning methods require the 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, the novel method of combining the greedy approach of Top-k and the random approach of Perm-k algorithms for better performance is introduced.

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 $\pi = (\pi_1, \dots, \pi_d)$ be a random permutation of $\{1, \dots, d\}$. Then for all $x \in \mathbb{R}^d$ and each $i \in \{1, 2, \dots, n\}$ we define

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

Table of Contents

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]

Error feedback proof

Lemma 22 [Beznosikov et al. 2020]

$\eta^k \leq \frac{1}{14(2\delta+B)L}, \forall k \geq 0$ and $\left\{\left(\eta^k\right)^2\right\}_{k\geq 0}$ – 2δ-slow decreasing. Then

$$\mathrm{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} e_i^{k+1}\right\|_2^2\right] \leq \frac{(1-1/\delta)}{49L(2\delta+B)}\sum_{j=0}^{k}\left[\left(1-\frac{1}{4\delta}\right)^{k-j}\left(f\left(x^j\right)-f\left(x^*\right)\right)\right]+$$

$$+\eta^k \frac{2(\delta-1)}{7L}\left(2D+\frac{C}{2\delta+B}\right).$$

Furthermore, for any 4δ-slow increasing non-negative sequence $\left\{w^k\right\}_{k\geq 0}$ it holds:

$$3L\cdot\sum_{k=0}^{K} w^k \cdot \mathrm{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} e_i^k\right\|_2^2\right] \leq$$

$$\leq \frac{1}{4}\sum_{k=0}^{K} w^k\left(\mathrm{E}\left[f\left(x^k\right)\right]-f\left(x_*\right)\right)+\left(3\delta D+\frac{3C}{4}\right)\sum_{k=0}^{K} w^k\eta^k.$$

Proof [Beznosikov et al. 2020]

$$\mathrm{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} e_i^{k+1}\right\|_2^2\right] \overset{Jensen}{\leq} \frac{1}{n}\mathrm{E}\left[\sum_{i=1}^{n}\left\|e_i^{k+1}\right\|_2^2\right] =$$

$$= \frac{1}{n}\mathrm{E}\left[\sum_{i=1}^{n}\left\|e_i^k + \eta^k g_i^k - \tilde{g}_i^k\right\|_2^2\right] =$$

$$= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[\left\|e_i^k + \eta^k g_i^k - \mathcal{C}\left(e_i^k + \eta^k g_i^k\right)\right\|_2^2\right] \overset{\mathbb{B}(\delta)}{\leq}$$

$$\overset{\mathbb{B}(\delta)}{\leq} \frac{1 - 1/\delta}{n}\sum_{i=1}^{n}\mathrm{E}_{\nabla}\left[\left\|e_i^k + \eta^k g_i^k\right\|_2^2\right] =$$

$$= \frac{1 - 1/\delta}{n}\sum_{i=1}^{n}\mathrm{E}_{\nabla}\left[\left\|e_i^k + \eta^k \nabla f_i\left(x^k\right) + \eta^k \xi_i^k\right\|_2^2\right]$$

Proof [NEW]

$$\mathrm{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} e_i^{k+1}\right\|_2^2\right] \overset{definition}{=}$$

$$= \mathrm{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} e_i^{k} + \eta^k g_i^k - \mathcal{C}\left(e_i^k + \eta^k g_i^k\right)\right\|_2^2\right] =$$

$$= \mathrm{E}\left[\left\|\frac{1}{n}\left(\sum_{i=1}^{n} \mathcal{C}\left(e_i^k + \eta^k g_i^k\right)\right) - \frac{1}{n}\left(\sum_{i=1}^{n} e_i^k + \eta^k g_i^k\right)\right\|_2^2\right] \overset{y_i := e_i^k + \eta^k g_i^k}{=}$$

$$= \mathrm{E}\left[\left\|\frac{1}{n}\left(\sum_{i=1}^{n} \mathcal{C}\left(y_i\right)\right) - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)\right\|_2^2\right] \overset{?}{\leq} (1-\delta)\left\|\sum_{i=1}^{n} y_i\right\|^2$$

Optimization problem

Main lemma

$$\mathrm{E}\left[\left\|\frac{1}{n}\left(\sum_{i=1}^{n} TopK\left(Perm_i\left(y_i\right)\right)\right)-\frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)\right\|_2^2\right] \overset{?}{\leq} (1-\delta)\left\|\sum_{i=1}^{n} y_i\right\|^2$$

Main lemma (simple case)

Lets assume, that $\forall i, j \nabla f_i(x) = \nabla f_j(x) = y$.
Then the following implies:

$$\mathbb{E}\|\frac{1}{n}\sum_{i}^{n} Top_k(Perm_i(y)) - y\|^2 \leq (1-\frac{nk}{d})\|y\|^2$$

This is n times better than better than $(1-\frac{k}{d})$. [Beznosikov et al. 2020]

Proof. 1. Lets proof the inequality itself. Fix the norm $\|y\|^2 = const$. Without loss of generality, y coordinates are in increasing order $y_1 < y_2 < \ldots < y_d$.

$$\mathbb{E}\|\frac{1}{n}\sum_i^n Top_k(Perm_i(y)) - y\|^2 = \frac{1}{\#\sigma}\sum_\sigma \frac{1}{n^2}\|\sum_i^n Top_k(Perm_i(y)) - ny\|^2 = \frac{1}{\#\sigma}\frac{1}{n^2}\sum_\sigma\sum_j^d (nI^\sigma_{y_j}y_j - ny_j)^2$$

Where $I^\sigma_{y_j} = 1$ if y_j is chosen by at least one Top_k in the σ permutation of $Perm$-s, and 0 otherwise. Then this for each j in fixed permutation σ we have:

$$(nI^\sigma_{y_j} - n)^2 = \begin{cases} 0 & \text{if } y_j \text{ is not chosen} \\ n^2, & \text{otherwise} \end{cases}$$

For each $1 \leq j \leq d$ let $p_j = \sum_\sigma (nI^\sigma_{y_j} - ny_j)^2$. Then it is clear that $\forall i < j \; p_i \geq p_j$, because greater the value of y_j more often it is chosen by Top_k.

$$\sum_\sigma\sum_j^d (nI^\sigma_{y_j}y_j - ny_j)^2 = \sum_j^d\sum_\sigma (nI^\sigma_{y_j}y_j - ny_j)^2 = \sum_j^d p_j y_j^2$$

Lets look at $y_i < y_{i+1}$. If we increase y_i^2 by ε and decrease y_{i+1}^2 by the same value, then $(y_i^2+\varepsilon)+(y_{i+1}^2-\varepsilon) = y_i^2 + y_{i+1}^2$. But $p_i(y_i^2+\varepsilon) + p_{i+1}(y_{i+1}^2-\varepsilon) \geq p_i y_i^2 + p_{i+1}y_{i+1}^2$. If $0 < y_i < y_{i+1}$ or $y_i < y_{i+1} < 0$, then moving y_i and y_{i+1} towards each other only increases variance value. So we can move all coordinates, until we have $y_1 = y_2 = \ldots = y_m < 0 < y_{m+1} = \ldots = y_d$.
Without loss of generality, $\|y_1\|^2 \leq \|y_{m+1}\|^2$.

2. It stands that the case, when $y_1 = y_2 = \ldots = y_d$ is optimal for maximizing the variance with $\|y\|^2 = const$. In that case, it can be easily observed, that the constant is $1 - \frac{nk}{d}$.

$$\mathbb{E}\|\frac{1}{n}\sum_i^n Top_k(Perm_i(y)) - y\|^2 = \sum_j^{\text{not chosen } d-nk \text{ coordinates}} y_j^2 = (d-nk)y_1^2 = (1 - \frac{nk}{d})\|y\|^2$$

□

Hessian variance

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.

- Theoretical proof will help us to find the best μ parameter depending on other hyperparameters, including $L_\pm$.
- The following attempts were to estimate the communication complexity of the algorithm on the problem where $L_\pm = 0$.

General case

Main lemma

$$\mathrm{E}\left[\left\|\frac{1}{n}\left(\sum_{i=1}^{n} TopK\left(Perm_i\left(y_i\right)\right)\right) - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)\right\|_2^2\right] \overset{?}{\leq} (1-\delta)\left\|\sum_{i=1}^{n} y_i\right\|^2$$

Gradient's coefficients

In case when $\nabla f(x) = y$, assuming that $y_1 < y_2 < \ldots < y_d$. We want to count the probability P_i - that y_i is chosen by $TopK(Perm(y))$ in all possible permutations.

$$P_i = \frac{\sum_{j=0}^{k-1} \mathbb{C}_{i-1}^{j} \cdot \mathbb{C}_{d-i}^{\frac{d}{n}-1-j}}{\mathbb{C}_{d-1}^{\frac{d}{n}-1}}$$

Further considerations

- We can try solving this quadratic optimization problem by using numerical solvers. This will help us to find the dependence on $L_{\pm}$, n, k, d.
  Bottleneck: exponential complexity with increase of n and d!
- We may test the hypothesis, that inequation's optimum is the case of equal gradients. For that $TopK(Perm_i) = Perm_i$ case can be checked. Also numerical results will give insights.
- Easier cases can be checked.
  For example, when $\forall i, j \, \|\nabla f_i - \nabla f_j\|^2 \leq \mathcal{C}$ for $\mathcal{C} \in \mathbb{R}^+$

Table of Contents

Observable compressors and EFs

TopK + Error Feedback

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

Unbiased TopK-PermK

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

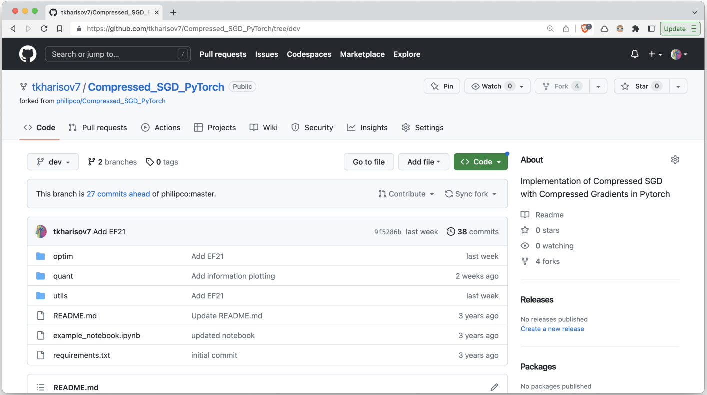
Biased TopK-PermK + Error Feedback

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

Classic Error Feedback

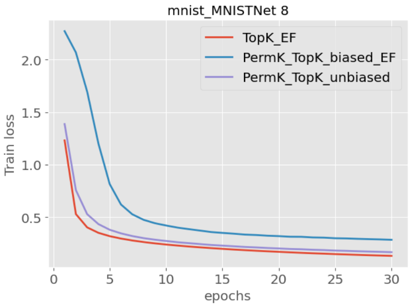
$$e_i^{k+1} = e_i^k + \nabla f_i(x^k) - C(Q_i^k(e_i^k + \nabla f_i(x^k)))$$

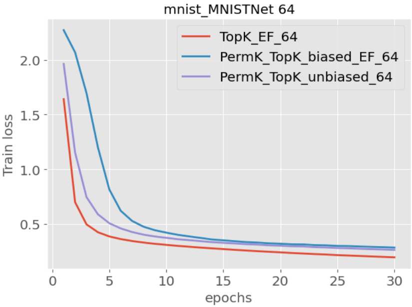
Framework implementation



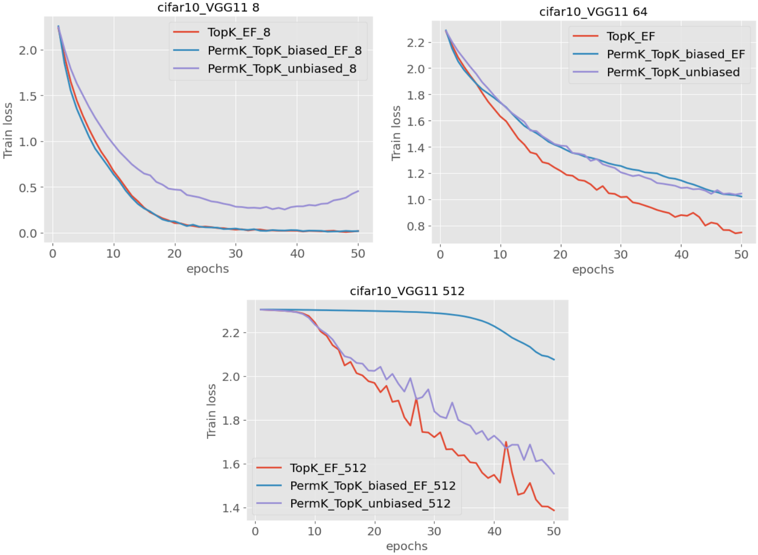
We have developed framework for our optimization problems and algorithm. Implementation is based on Horthath et al. 2020.

MnistNet comparison

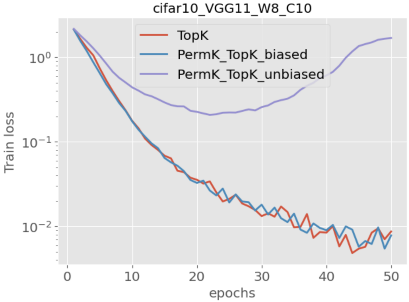




CIFAR-10 + VGG11



Dataset mix-up (common 10%)



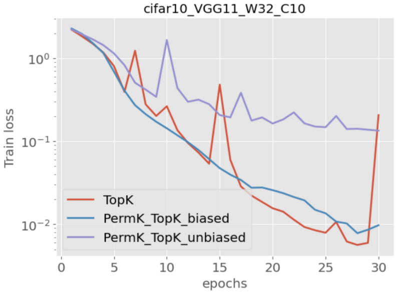


Table of Contents

Results

1. It was rigorously proven, that in $L_{\pm} = 0$ regime TopK+PermK combined with the classic Error Feedback achieves n times better convergence results than TopK with Error Feedback.
2. A theoretical search of convergence rate in a general regime is equivalent to the quadratic optimization task with $d * n$ variables and $d! * n$ inequalities.
3. Experiments prove that with an increase in datasets correlation, TopK-PermK perfomance enjoys a similar convergence rate as TopK-EF.
4. With the increase of nodes number TopK-PermK without Error Feedback shows better perfomance than one with Error Feedback.

References

Rafał Szlendak and Alexander Tyurin and Peter Richtárik (2021)
Permutation Compressors for Provably Faster Distributed Nonconvex Optimization
ICLR 2022, poster session

Aleksandr Beznosikov and Samuel Horváth and Peter Richtárik and Mher Safaryan (2022)
On Biased Compression for Distributed Learning
CoRR abs/2002.12410, arXiv:2002.1241

Horváth, Samuel and Richtárik, Peter (2020)
A Better Alternative to Error Feedback for Communication-Efficient Distributed Learning
arXiv preprint, arXiv:2006.11077

Table of Contents

Q&A

Your questions, please!