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

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

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

Scientific Practicum Conference, May 19 2023

Table of Contents

Motivation

- 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.
- Communication cost is a bottleneck for the Federated Learning approach: worker devices use unstable and slow networks such as Wi-Fi and Cellular.
- 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 $\left|x_{(1)}\right| \leq \left|x_{(2)}\right| \leq \cdots \leq \left|x_{(d)}\right|$.

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]
- *TopK-PermK:* $(1 - \frac{1}{\delta}) = \frac{d-k}{d}$ [NEW]

Proof

Proof of delta

$$E\left[\| TopK(aPermK(x)) - x\|^2\right] = E\left[\| TopK(aPermK(x))\|^2\right] +$$

$$E\left[\|x\|^2\right] - 2E[\langle TopK(aPermK(x)), x\rangle] =$$

$$= E[\langle TopK(aPermK(x)), TopK(aPermK(x))\rangle] + \|x\|^2 -$$

$$2E[\langle TopK(aPermK(x)), x\rangle] = E[< TopK(aPermK(x)), aPermK(x)\rangle] + \|x\|^2$$

$$-\frac{2}{a}E[\langle TopK(aPermK(x)), aPermK(x)\rangle] =$$

$$= \|x\|^2 - \frac{2-a}{a}E[< TopK(aPermK(x)), aPermK(x) >] \leq$$

$$\leq \|x\|^2 - \frac{kn}{d}\frac{2-a}{a}E\left[\|aPermK(x)\|^2\right] =$$

$$= \|x\|^2 - \frac{kn}{d}\frac{2-a}{a}\frac{a^2}{n}\|x\|^2 = \|x\|^2\left(1 - \frac{k(2-a)a}{d}\right) \Rightarrow \delta = \frac{d}{k(2-a)a}$$

Theorem

Error Feedback theorem [Beznosikov et al., 2023]

Let $\{x^k\}_{k\geq 0}$ 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 $f^\star := f(x^\star)$ and

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

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

Table of Contents

Quadratic optimization problem

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

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

  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:

Different Error Feedbacks

C - *TopK* biased compressor.

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

EF0: $e_i^{k+1} = e_i^k + \nabla f_i(x^k) - C(Q_i^k(e_i^k + \nabla f_i(x^k)))$

EF1: $e_i^{k+1} = e_i^k + Q_i^k(\nabla f_i(x^k)) - C(Q_i^k(e_i^k + \nabla f_i(x^k)))$

EF2: $e_i^{k+1} = Q_i^k(e_i^k + \nabla f_i(x^k)) - C(Q_i^k(e_i^k + \nabla f_i(x^k)))$

Experiment reproduction

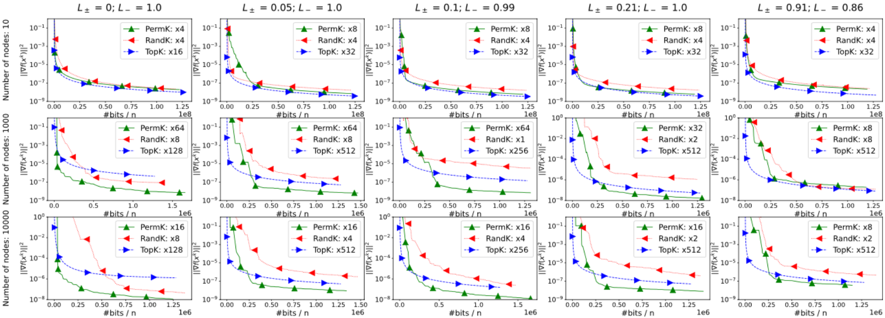


Figure 1: Comparison of algorithms on synthetic quadratic optimization tasks with nonconvex $\{f_i\}$.

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.

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

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

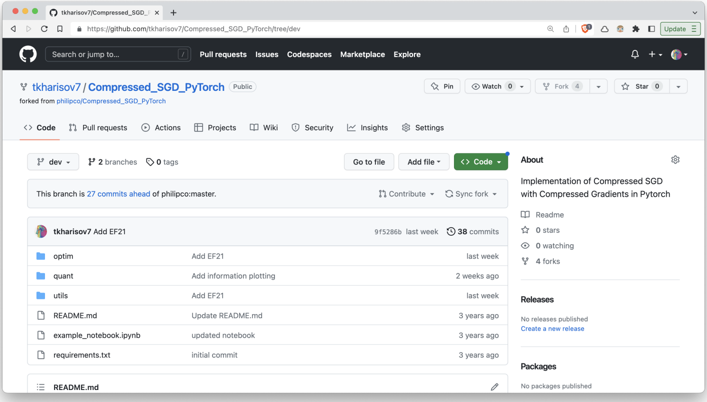
Error Feedback 21

Paper: EF21: A New, Simpler, Theoretically Better, and Practically Faster Error Feedback (Peter Richtárik, Igor Sokolov, Ilyas Fatkhullin)

EF21: SOTA Error Feedback method

```
EF21 (Multiple nodes)
Input: $x^0; g_i^0 = \mathcal{C}\left(\nabla f_i\left(x^0\right)\right)$ for $i = 1, \ldots, n$ (nodes; master);
    learning rate $\gamma > 0; g^0 = \frac{1}{n}\sum_{i=1}^n g_i^0$( known by master)
for $t = 0, 1, 2, \ldots, T-1$ do
    Master computes $x^{t+1} = x^t - \gamma g^t$ and broadcasts $x^{t+1}$ to all nodes
    for all nodes $i = 1, \ldots, n$ in parallel do
        Compress $c_i^t = \mathcal{C}\left(\nabla f_i\left(x^{t+1}\right) - g_i^t\right)$ and send $c_i^t$ to the master
        Update local state $g_i^{t+1} = g_i^t + \mathcal{C}\left(\nabla f_i\left(x^{t+1}\right) - g_i^t\right)$
    end for
    Master computes $g^{t+1} = \frac{1}{n}\sum_{i=1}^n g_i^{t+1}$ via $g^{t+1} = g^t + \frac{1}{n}\sum_{i=1}^n c_i^t$
end for
```

Framework implementation



tkharisov7 / Compressed_SGD_PyTorch Public

forked from philipco/Compressed_SGD_PyTorch

This branch is 27 commits ahead of philipco:master.

tkharisov7 Add EF21 9f5286b last week 38 commits

optim	Add EF21	last week
quant	Add information plotting	2 weeks ago
utils	Add EF21	last week
README.md	Update README.md	3 years ago
example_notebook.ipynb	updated notebook	3 years ago
requirements.txt	initial commit	3 years ago

README.md

About

Implementation of Compressed SGD with Compressed Gradients in Pytorch

- Readme
- 0 stars
- 0 watching
- 4 forks

Releases

No releases published
Create a new release

Packages

No packages published

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

Quadratic problem experiment reproduction (Without EF)

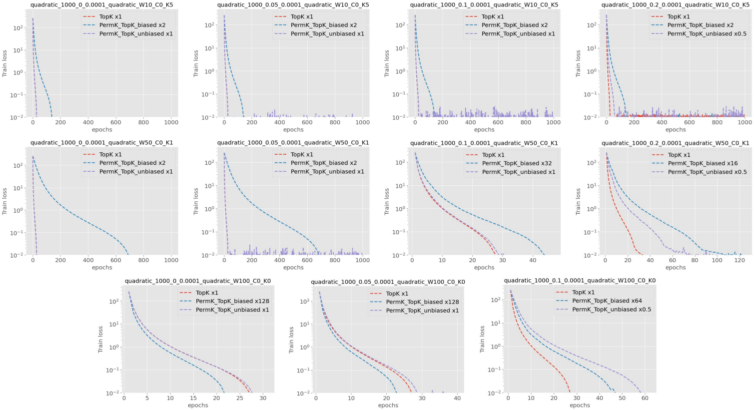
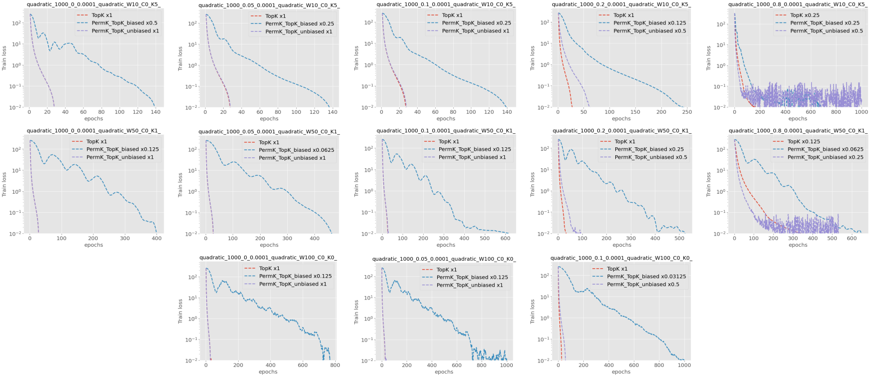


Figure: Comparison of algorithms withot EF on the Quadratic optimization problem. Each row corresponds to a fixed number of nodes; each column corresponds to a fixed noise scale. In the legends there are compressor names and multiplicity factors

Quadratic problem experiment reproduction (With EF21)



quadratic_1000_0_0.0001_quadratic_W10_C0_K5_

quadratic_1000_0.05_0.0001_quadratic_W10_C0_K5_

quadratic_1000_0.1_0.0001_quadratic_W10_C0_K5_

quadratic_1000_0.2_0.0001_quadratic_W10_C0_K5_

quadratic_1000_0.8_0.0001_quadratic_W10_C0_K5_

quadratic_1000_0_0.0001_quadratic_W50_C0_K1_

quadratic_1000_0.05_0.0001_quadratic_W50_C0_K1_

quadratic_1000_0.1_0.0001_quadratic_W50_C0_K1_

quadratic_1000_0.2_0.0001_quadratic_W50_C0_K1_

quadratic_1000_0.8_0.0001_quadratic_W50_C0_K1_

quadratic_1000_0_0.0001_quadratic_W100_C0_K0_

quadratic_1000_0.05_0.0001_quadratic_W100_C0_K0_

quadratic_1000_0.1_0.0001_quadratic_W100_C0_K0_

Figure: Comparison of algorithms with EF21 on the Quadratic optimization problem. In the legends there are compressor names and learning rates' multiplicity factors

Experiment with fine-tuned lr with EF21

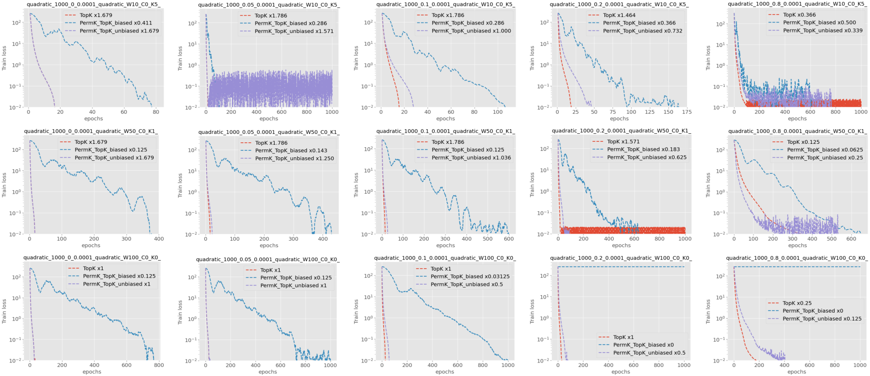


Figure: Comparison of algorithms with EF21 on the Quadratic optimization problem. In the legends there are compressor names and learning rates fine-tuned multiplicity factors

Table of Contents

Results and Further actions

Results:

- Theoretical estimation of the biased class parameter of the compressor algorithm
- Set up the experiments with the low noise scale with the average number of workers with learning rate tuning

Further actions:

- The following attempts will be to estimate the communication complexity of the algorithm on the problem where noise scale is relatively small
- Make more experiments with the low noise scale with the bigger number of workers and make more accurate learning rate tuning using more computational power

Table of Contents

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

Peter Richtárik and Igor Sokolov and Ilyas Fatkhullin (2021)
EF21: A New, Simpler, Theoretically Better, and Practically Faster Error Feedback
NeurIPS 2021