UDC 519.86

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

K. ACHARYA, T. KHARISOV, and A. BEZNOSIKOV, Moscow Institute of Physics and Technology

1 ABSTRACT

The proposed research entails a theoretical analysis of the convergence rate and efficiency of a novel distributed optimization method, which incorporates independent segmentation of gradient coordinates ($PermK$) followed by a greedy coordinate selection process ($TopK$) for each gradient segment. Our findings indicate that the new method attains comparable results to state-of-the-art techniques, such as $MARINA - PermK$ [4] and $EF - TopK$ [1], in terms of zero-variance and general variance regimes, respectively. Additionally, the experimental performance of our approach is demonstrated through its application to quadratic problems and computer vision models.

2 PROBLEM STATEMENT AND CURRENT SOLUTIONS

This paper considers 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, updating $x^{k+1} = x^k - \frac{\eta^k}{n}\sum_{i=1}^{n}\nabla f_i\left(x^k\right)$, where $\eta^k > 0$ is a step size.

In order to minimize the communication overhead between devices, we propose transmitting only a portion of the gradient, rather than the entire gradient. This can be achieved using approaches such as the practical, greedy method ($TopK$) and the somewhat unconventional random method ($PermK$), as described in the work by Szlendak [4]. In the same study, a comparison was made between the $TopK$ and $PermK$ algorithms for a quadratic optimization problem involving non-convex f_i functions. The results showed that the $PermK$ compressor resulted in faster convergence with a larger number of devices, while $TopK$ performed better with a smaller number of devices.

3 THEORETICAL RESULTS

We say $C \in \mathbb{B}^3(\delta)$ for some $\delta > 1$ if $\mathrm{E}\left[\|C(x) - x\|_2^2\right] \le \left(1 - \frac{1}{\delta}\right)\|x\|_2^2, \quad \forall x \in \mathbb{R}^d$.

Lemma 3.1. *For $TopK - PermK$ it is proven that $(1 - \frac{1}{\delta}) = \frac{d-k}{d}$, which is the same as for $TopK$*

From this lemma and Theorem 16 about Error Feedback [2] we say the convergence rates of these algorithms with Error Feedback are the same.

Authors' address: K. Acharya, kirillacharya7@gmail.com; T. Kharisov, tkharisov7@gmail.com; A. Beznosikov, beznosikov.an@phystech.edu, Moscow Institute of Physics and Technology .

4 EXPERIMENTS

In this section we reproduced [Compressed_SGD_PyTorch] the Experiment 5.1 from [4]. We considered a synthetic (strongly convex) quadratic function $f = \sum_{i=1}^{n} f_i$ composed of nonconvex quadratics

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

where $b_i \in \mathbb{R}^d$, $A_i \in \mathbb{R}^{d\times d}$, and $A_i = A_i^\top$. The Algorithm 1 from [4] generates λ-strongly convex f, where $\lambda = 1e-6$, and dimension $d = 1000$ are fixed. After we generated optimization tasks with the number of nodes $n \in \{10, 50, 100\}$ and noise scale $ns \in \{0, 0.05, 0.1, 0.2, 0.8\}$. We compared two versions of the novel algorithm: with Error Feedback (*Biased*) and multiplied by the number of workers n (pseudounbiased, for simplicity we call it *Unbiased*) with the classic *TopK* with the Error Feedback, which was chosen to be EF21 [3] algorithm. The plots provided on Fig. 1 for each compressor are the ones with the best convergence rate.

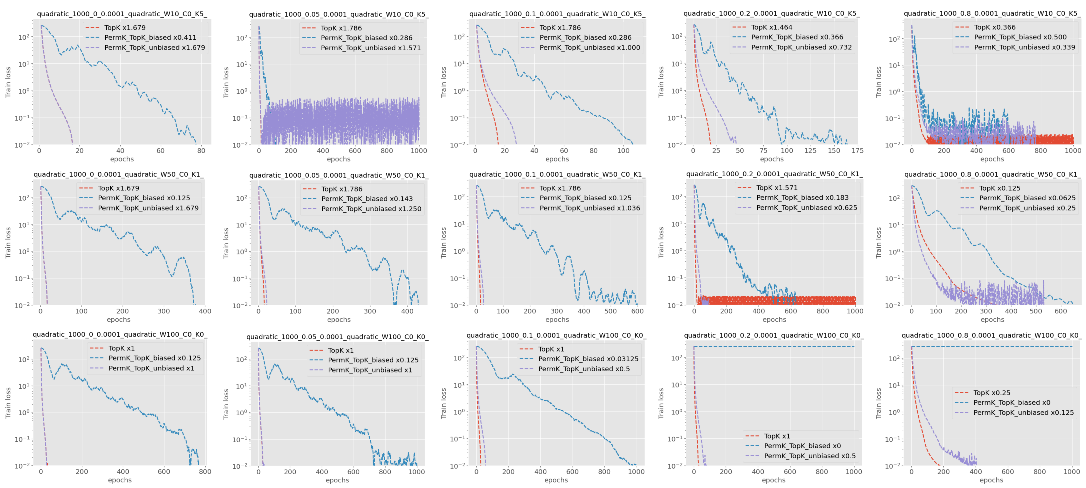

Fig. 1. Comparison of algorithms with EF21 on the Quadratic optimization problem. Each row corresponds to a fixed number of nodes; each column corresponds to a fixed noise scale. In the legends there are compressor names and fine-tuned multiplicity factors

We see that *Unbiased* version performs not worse than *TopK* in low-variance regime, which reproduces theoretical dependencies. The aim for the further experiments is to investigate the behaviour of our sparsification method with the setting with larger n parameter, e.g. $n \in \{1000, 10000\}$

REFERENCES

[1] Dan Alistarh, Torsten Hoefler, Mikael Johansson, Sarit Khirirat, Nikola Konstantinov, and Cédric Renggli. 2018. The Convergence of Sparsified Gradient Methods. arXiv:1809.10505 [cs.LG]

[2] Aleksandr Beznosikov, Samuel Horváth, Peter Richtárik, and Mher Safaryan. 2020. On Biased Compression for Distributed Learning. *CoRR* abs/2002.12410 (2020). arXiv:2002.12410 https://arxiv.org/abs/2002.12410

[3] Peter Richtárik, Igor Sokolov, and Ilyas Fatkhullin. 2021. EF21: A New, Simpler, Theoretically Better, and Practically Faster Error Feedback. arXiv:2106.05203 [cs.LG]

[4] Rafal Szlendak, Alexander Tyurin, and Peter Richtárik. 2021. Permutation Compressors for Provably Faster Distributed Nonconvex Optimization. *CoRR* abs/2110.03300 (2021). arXiv:2110.03300 https://arxiv.org/abs/2110.03300