A Better Alternative to Error-Feedback for Communication-Efficient Distributed Learning

Samuel Horváth and Peter Richtárik

King Abdullah University of Science and Technology

Problem setup

We consider distributed optimization problems of the form

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

- n is the number of nodes,
- $f_i : \mathbb{R}^d \to \mathbb{R}$ is a smooth differentiable loss function composed of data stored on worker i.

Communication Bottleneck

In distributed training, model updates (or gradient vectors) have to be exchanged in each iteration. Due to the size of the communicated messages for commonly considered deep models (Alistarh et al., 2016), this represents significant bottleneck of the whole optimization procedure. To reduce the amount of data that has to be transmitted, *communication compression* is one of the popular approaches. Considering both practice and theory, compression operators can be split into two groups: biased and unbiased.

Definition ("Unbiased")

A randomized mapping $C : \mathbb{R}^d \to \mathbb{R}^d$ is an *unbiased compression operator (unbiased compressor)* if there exists $\delta \geq 1$ such that

$$\mathrm{E}[C(x)] = x, \quad \mathrm{E}\|C(x)\|^2 \leq \delta \|x\|^2, \quad \forall x \in \mathbb{R}^d.$$

If this holds, we will for simplicity write $C \in \mathbb{U}(\delta)$.

Definition ("Biased")

A (possibly) randomized mapping $C : \mathbb{R}^d \to \mathbb{R}^d$ is a *general compression operator (general compressor)* if there exists $\lambda > 0$ and $\delta \geq 1$ such that

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

If this holds, we will for simplicity write $C \in \mathbb{C}(\delta)$.

Induced Compressor vs. Error-Feedback (EF) [Example]

Assumptions:
Biased and Unbiased Compressors: $C_1 \in \mathbb{C}(\delta)$ and $C_2 \in \mathbb{U}(\delta)$.
f is over-parametrized (Vaswani et al., 2019) and μ-quasi convex, i.e.

$$f^\star \geq f(x) + \langle \nabla f(x), x^\star - x\rangle + \frac{\mu}{2}\|x^\star - x\|^2, \quad \forall x \in \mathbb{R}^d.$$

where $f^\star$ is the optimal solution of f and $f^\star = f_i^\star, \ \forall i \in [n]$.
f_i's are L-smooth.

Construction [on worker i]:

Induced Compressor:
- obtain $g_i^k [\mathrm{E}[g_i^k] = \nabla f_i(x^k)]$
- $\Delta_i^k = \eta^k[h_i^k + C_1^k(g_i^k - h_i^k)]$, $h_i^k = C_1^k(g_i^k)$
- send Δ_i^k to master
- [no need to keep track of errors]

Error-Feedback:
- obtain $g_i^k [\mathrm{E}[g_i^k] = \nabla f_i(x^k)]$
- $\Delta_i^k = C_1^k(\eta_k g_i^k + e_i^k)$
- send Δ_i^k to master
- $e_i^{k+1} = \eta^k g_i^k + e_i^k - \Delta_i^k$

Convergence Rates $[\mathrm{E}[f(\bar{x}^T) - f^\star]]$:

Induced Compressor: $$O\left(\left(\frac{\delta - 1}{n} + 1\right) L \exp\left[-\frac{\mu T}{4\delta_n L}\right]\right)$$

Error-Feedback: $$O\left(\delta L \exp\left[-\frac{\mu T}{4\delta_n L}\right]\right)$$

Contributions

- **Induced Compressor.** When used with EF framework, biased compressors (e.g., Top-K) can often achieve superior performance when compared to their unbiased counterparts (e.g., Rand-K), which is attributed to their low empirical variance. Our key contribution is the development of a simple but remarkably effective alternative (described above), which we argue leads to better and more versatile methods both in theory and practice.
- **Better Theory for DCSGD.** We provide a new and tighter theoretical analysis of DCSGD under weaker assumptions.
- **Partial Participation.** We extend our results to obtain the first convergence guarantee for partial participation with arbitrary distributions over nodes.

Motivational Example

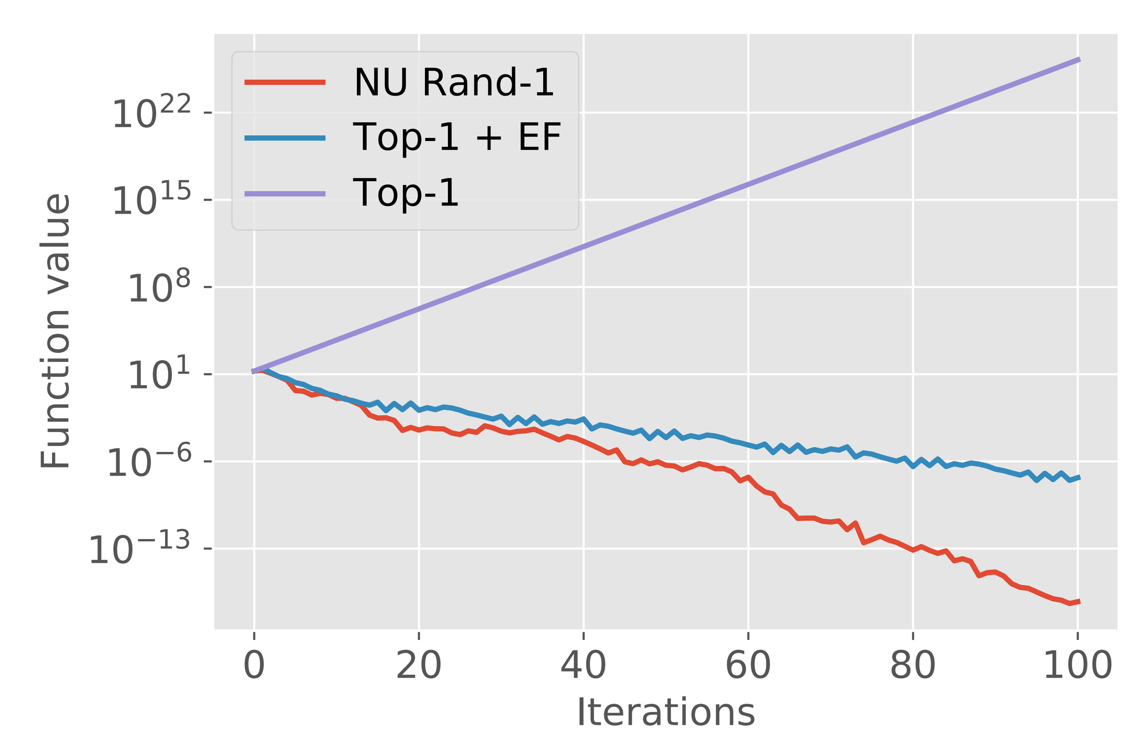
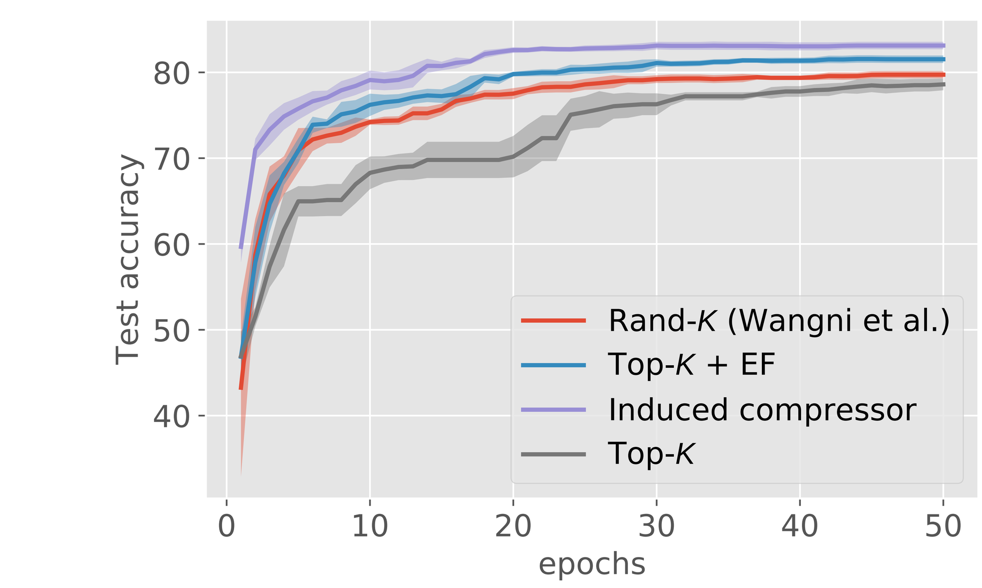


Figure: Comparison of Top-1 (+ EF) and NU Rand-1 on Example 1 from Beznosikov et al., 2020.

Numerical Experiments



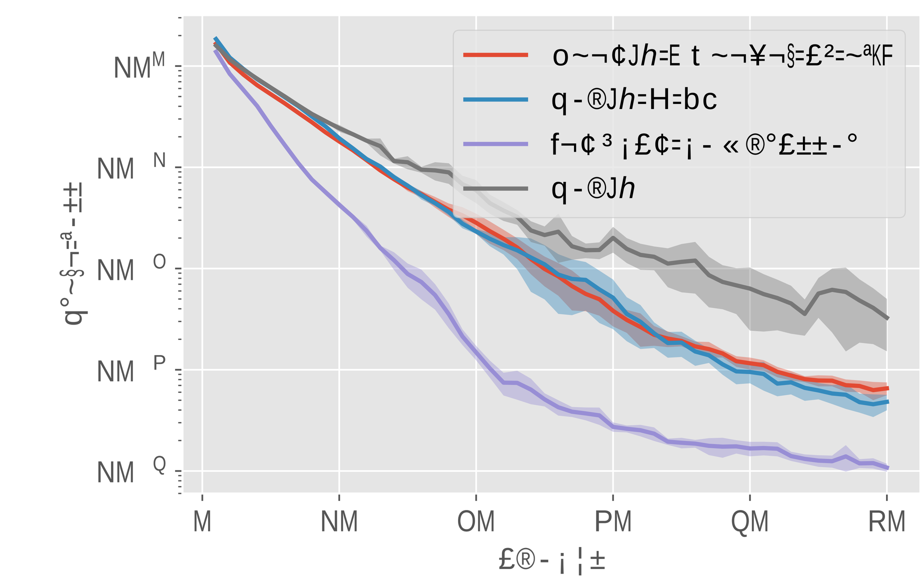


Figure: : Comparison of different sparsification techniques on CIFAR10 with Resnet18.

Contact Information

- Paper: https://arxiv.org/pdf/2002.05359.pdf
- Email: samuel.horvath@kaust.edu.sa

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

NEURAL INFORMATION PROCESSING SYSTEMS