3PC: Three Point Compressors for Communication-Efficient Distributed Training and a Better Theory for Lazy Aggregation

Peter Richtárik[1] Igor Sokolov[1] Ilyas Fatkhulin[2] Elnur Gasanov[1] Zhize Li[1] Eduard Gorbunov[3]

[1]KAUST [2]ETH Zurich [3]MIPT

The problem

Nonconvex *distributed* optimization problem:

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

- n – number of clients
- $f_i(x)$ – smooth local loss function, i.e., $\|\nabla f_i(x) - \nabla f_i(y)\| \le L_i\|x-y\|$ for all $x, y \in \mathbb{R}^d$, $f^{\inf} := \inf_{x\in\mathbb{R}^d} f(x) > -\infty$

Goal: find $\hat{x}$ such that $\mathbb{E}\left[\|\nabla f(\hat{x})\|^2\right] \le \varepsilon^2$

Compressed learning

Contractive compressor: a (possibly randomized) map $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$ is called a *contractive compressor*, if there exists a constant $0 < \alpha \le 1$:

$$\mathbb{E}\left[\|\mathcal{C}(x) - x\|^2\right] \le (1-\alpha)\,\|x\|^2, \qquad \forall x \in \mathbb{R}^d.$$

Top-k (greedy) sparsification operator is defined via

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

where $\left|x_{(1)}\right| \le \left|x_{(2)}\right| \le \cdots \le \left|x_{(d)}\right|$. Then $\alpha = \frac{k}{d}$.

Error feedback with contractive compressor

◇ Motivation for error feedback – the method of type

$$x^{t+1} = x^t - \gamma\frac{1}{n}\sum_{i=1}^n g_i^t,$$
$$g_i^t = \mathcal{C}\left(\nabla f_i(x^t)\right)$$

- **may diverge** [1] for a biased compressor $\mathcal{C}$ and $n > 1$.

◇ Original error feedback (EF) [1]

- bounded gradients $\|\nabla f_i(x)\| \le G$
- not optimal complexity $\mathcal{O}\left({}^1/_{\varepsilon^3}\right)$

◇ Modern error feedback (EF21) [2]:

- simple analysis
- optimal complexity $\mathcal{O}\left({}^1/_{\varepsilon^2}\right)$
- better in practice

Lazy aggregation

◇ Motivation for LAG [3]: reduce communication by sending gradients only when they change significantly:

$$g_i^t = \begin{cases} \nabla f_i\left(x^t\right) & \text{if } \left\|g_i^{t-1} - \nabla f_i\left(x^t\right)\right\|^2 > \zeta\left\|\nabla f_i\left(x^t\right) - \nabla f_i\left(x^{t-1}\right)\right\|^2 \\ g_i^{t-1} & \text{otherwise,} \end{cases}$$

where $\zeta > 0$ is the trigger.

- not optimal complexity $\mathcal{O}\left({}^1/_{\varepsilon^3}\right)$
- difficult analysis

References

[1] Seide, F., Fu, H., Droppo, J., Li, G., Yu, D. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech DNNs. Interspeech, 2014.

[2] P. Richtárik, I. Sokolov, I. Fatkhullin. EF21: A new, simpler, theoretically better, and practically faster error feedback. NeurIPS'21, arXiv:2106.05203, 2021.

[3] Chen, T., Giannakis, G., Sun, T., Yin, W. LAG: Lazily aggregated gradient for communication-efficient distributed learning. NeurIPS'18.

[4] Gorbunov, E., Burlachenko, K. P., Li, Z., and Richtárik, P. MARINA: Faster non-convex distributed learning with compression. ICML'21

[5] Sun, J., Chen, T., Giannakis, G., and Yang, Z. Communication-efficient distributed learning via lazily aggregated quantized gradients. NeurIPS'19.

[6] Ghadikolaei, H. S., Stich, S., and Jaggi, M. LENA: Communication-efficient distributed learning with self- triggered gradient uploads. AISTATS'21

[7] Szlendak R, Tyurin A, Richtárik P. Permutation compressors for provably faster distributed nonconvex optimization. arXiv preprint arXiv:2110.03300, 2021

Table 1:Summary of the methods fitting our general 3PC framework. For each method we give the formula for the 3PC compressor $\mathcal{C}_{h,y}(x)$, its parameters A, B, and the ratio ${}^B/_A$ appearing in the convergence rate. Notation: α = parameter of the contractive compressor $\mathcal{C}$, ω = parameter of the unbiased compressor $\mathcal{Q}$, A_1, B_1 = parameters of three points compressor $\mathcal{C}^1_{h,y}(x)$, $\bar{\alpha} = 1 - (1-\alpha_1)(1-\alpha_2)$, where α_1, α_2 are the parameters of the contractive compressors $\mathcal{C}_1, \mathcal{C}_2$, respectively.

Variant of 3PC	Citation	$\mathcal{C}_{h,y}(x) =$	A	B	$\frac{B}{A}$
EF21	[2]	$h + \mathcal{C}(x-h)$	$1-\sqrt{1-\alpha}$	$\frac{1-\alpha}{1-\sqrt{1-\alpha}}$	$\mathcal{O}\left(\frac{1-\alpha}{\alpha^2}\right)$
LAG	[3]	$\begin{cases} x, & \text{if } \|x-h\|^2 > \zeta\|x-y\|^2, \\ h, & \text{otherwise} \end{cases}$	1	ζ	$\mathcal{O}(\zeta)$
CLAG	NEW	$\begin{cases} h + \mathcal{C}(x-h), & \text{if } \|x-h\|^2 > \zeta\|x-y\|^2, \\ h, & \text{otherwise} \end{cases}$	$1-\sqrt{1-\alpha}$	$\max\left\{\frac{1-\alpha}{1-\sqrt{1-\alpha}}, \zeta\right\}$	$\mathcal{O}\left(\max\left\{\frac{1-\alpha}{\alpha^2}, \frac{\zeta}{\alpha}\right\}\right)$
3PCv1	NEW	$y + \mathcal{C}(x-y)$	1	$1-\alpha$	$1-\alpha$
3PCv2	NEW	$b + \mathcal{C}\left(x-b\right)$, where $b = h + \mathcal{Q}(x-y)$	α	$(1-\alpha)\omega$	$\frac{(1-\alpha)\omega}{\alpha}$
3PCv3	NEW	$b + \mathcal{C}\left(x-b\right)$, where $b = \mathcal{C}^1_{h,y}(x)$	$1-(1-\alpha)(1-A_1)$	$(1-\alpha)B_1$	$\frac{(1-\alpha)B_1}{1-(1-\alpha)(1-A_1)}$
3PCv4	NEW	$b + \mathcal{C}_1\left(x-b\right)$, where $b = h + \mathcal{C}_2(x-h)$	$1-\sqrt{1-\bar{\alpha}}$	$\frac{1-\bar{\alpha}}{1-\sqrt{1-\bar{\alpha}}}$	$\mathcal{O}\left(\frac{1-\bar{\alpha}}{\bar{\alpha}^2}\right)$
3PCv5	NEW	$\begin{cases} x, & \text{w.p. } p \\ h + \mathcal{C}(x-y), & \text{w.p. } 1-p \end{cases}$	$1-\sqrt{1-p}$	$\frac{(1-p)(1-\alpha)}{1-\sqrt{1-p}}$	$\mathcal{O}\left(\frac{(1-p)(1-\alpha)}{p^2}\right)$
MARINA	[4]	N/A	p	$\frac{(1-p)\omega}{n}$	$\frac{(1-p)\omega}{np}$

Main contribution

We propose Three Point Compressor (3PC) – a general concept unifying contractive compression and lazy aggregation.

1. Three point compressor (3PC)

3PC. We say that a (possibly randomized) map

$$\mathcal{C}_{h,y}(x) : \underbrace{\mathbb{R}^d}_{h\in} \times \underbrace{\mathbb{R}^d}_{y\in} \times \underbrace{\mathbb{R}^d}_{x\in} \to \mathbb{R}^d$$

is a three point compressor (3PC) if there exist constants $0 < A \le 1$ and $B \ge 0$ such that the following relation holds for all $x, y, h \in \mathbb{R}^d$

$$\mathbb{E}\left[\|\mathcal{C}_{h,y}(x) - x\|^2\right] \le (1-A)\|h-y\|^2 + B\|x-y\|^2. \qquad (1)$$

The vectors $y \in \mathbb{R}^d$ and $h \in \mathbb{R}^d$ are parameters defining the compressor.

2. Distributed compressed GD with 3PC

Algorithm 1.

- Server broadcasts g^t to the workers; workers compute $x^{t+1} = x^t - \gamma g^t$
- Workers apply 3PC $g_i^{t+1} = \mathcal{C}_{g_i^t, \nabla f_i(x^t)}(\nabla f_i(x^{t+1}))$ and send the result to the server
- Server aggregates received messages $g^{t+1} = \frac{1}{n}\sum_{i=1}^n g_i^{t+1}$

3. Special cases

◇ GD: if we do not employ any compression, i.e., if we set

$$\mathcal{C}_{h,y}(x) \equiv x,$$

then Algorithm 1 reduces to vanilla GD and (1) holds with $B = 1$ and $A = 0$.

◇ EF21 [2]: let $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$ be a contractive compressor and

$$\mathcal{C}_{h,y}(x) := h + \mathcal{C}(x-h).$$

Then, Algorithm 1 reduces to EF21 and (1) holds with $A := 1-(1-\alpha)(1+s)$ and $B := (1-\alpha)\left(1+s^{-1}\right)$, where $s > 0$ satisfies $(1-\alpha)(1+s) < 1$.

◇ LAG [3] and CLAG: let $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$ be a contractive compressor. Choose a trigger $\zeta > 0$, and define

$$\mathcal{C}_{h,y}(x) := \begin{cases} h + \mathcal{C}(x-h), & \text{if } \|x-h\|^2 > \zeta\|x-y\|^2, \\ h, & \text{otherwise,} \end{cases}$$

Then, Algorithm 1 reduces to CLAG and (1) holds with $A := 1-(1-\alpha)(1+s)$ and $B := \max\left\{(1-\alpha)\left(1+s^{-1}\right), \zeta\right\}$, where $s > 0$ satisfies $(1-\alpha)(1+s) < 1$. If $\mathcal{C}(x) \equiv 0$ $(\alpha = 1)$, we recover LAG.

◇ In Table 1 we summarize several further 3PC compressors and the new algorithms they lead to (e.g., 3PCv1 — 3PCv5).

4. Main result

Assumption 1. The functions $f_1, \ldots, f_n : \mathbb{R}^d \to \mathbb{R}$ are differentiable. Moreover, there exists $f^{\inf} \in \mathbb{R}$ such that $f(x) \ge f^{\inf}$ for all $x \in \mathbb{R}^d$.

Assumption 2. The function $f : \mathbb{R}^d \to \mathbb{R}$ is L_--smooth, i.e., it is differentiable and its gradient satisfies

$$\|\nabla f(x) - \nabla f(y)\| \le L_-\|x-y\| \quad \forall x, y \in \mathbb{R}^d.$$

Assumption 3. There is a constant $L_+ > 0$ such that $\frac{1}{n}\sum_{i=1}^n\|\nabla f_i(x) - \nabla f_i(y)\|^2 \le L_+^2\|x-y\|^2$ for all $x, y \in \mathbb{R}^d$. Let L_+ be the smallest such number. It is easy to see that $L_- \le L_+$.

Theorem 1

Let Assumptions 1-3 hold. Assume that the stepsize γ of the 3PC method satisfies $0 \le \gamma \le {}^1/_{M_1}$, where $M_1 = L_- + L_+\sqrt{{}^B/_A}$. Then, for any $T \ge 1$ we have

$$\mathbb{E}\left[\|\nabla f(\hat{x}^T)\|^2\right] \le \frac{2\Delta^0}{\gamma T} + \frac{\mathbb{E}\left[G^0\right]}{AT},$$

where $\hat{x}^T$ is sampled uniformly at random from the points $\{x^0, x^1, \ldots, x^{T-1}\}$ produced by 3PC , $\Delta^0 := f(x^0) - f^{\inf}$, and $G^0 := \frac{1}{n}\sum_{i=1}^n\left\|g_i^0 - \nabla f_i(x^0)\right\|^2$.

Corollary 1

Let the assumptions of Theorem 1 hold and choose the stepsize $\gamma = \frac{1}{L_- + L_+\sqrt{{}^B/_A}}$. Then, to achieve $\mathbb{E}\left[\|\nabla f(\hat{x}^T)\|^2\right] \le \varepsilon^2$ for some $\varepsilon > 0$, the 3PC method requires

$$T = \mathcal{O}\left(\frac{\Delta^0\left(L_- + L_+\sqrt{{}^B/_A}\right)}{\varepsilon^2} + \frac{\mathbb{E}\left[G^0\right]}{A\varepsilon^2}\right)$$

iterations (=communication rounds).

◇ Initialization with $g_i^0 = \nabla f_i(x^0)$ implies $G^0 = 0$ and

$$T = \mathcal{O}\left(\frac{\Delta^0\left(L_- + L_+\sqrt{{}^B/_A}\right)}{\varepsilon^2}\right)$$

◇ The smaller ${}^B/_A$, the better

◇ We also have the results under the Polyak-Łojasiewicz (PŁ) condition

5. Comparison of methods with lazy aggregation

Table 2:Comparison of existing and proposed theoretically-supported methods employing lazy aggregation. In the rates for our methods, $M_1 = L_- + L_+\sqrt{{}^B/_A}$ and $M_2 = \max\left\{L_- + L_+\sqrt{{}^{2B}/_A}, {}^A/_{2\mu}\right\}$.

Method	Simple method?	Uses a contractive compressor $\mathcal{C}$?	Strongly convex rate	PŁ nonconvex rate	General nonconvex rate
LAG [3]	✓	✗	linear	✗	✗
LAQ [5]	✗	✓	linear	✗	✗
LENA [6]	✓	✓	$\mathcal{O}(G^4/T^2\mu^2)$	$\mathcal{O}(G^4/T^2\mu^2)$	$\mathcal{O}(G^{4/3}/T^{2/3})$
LAG (NEW)	✓	✗	$\mathcal{O}(\exp(-T\mu/M_2))$	$\mathcal{O}(\exp(-T\mu/M_2))$	$\mathcal{O}(M_1/T)$
CLAG (NEW)	✓	✓	$\mathcal{O}(\exp(-T\mu/M_2))$	$\mathcal{O}(\exp(-T\mu/M_2))$	$\mathcal{O}(M_1/T)$

6. Experiments

◇ Training of the **autoencoder model**

$$\min_{D\in\mathbb{R}^{d_f\times d_e}, E\in\mathbb{R}^{d_e\times d_f}}\left[f(D,E) := \frac{1}{n}\sum_{i=1}^n\|DEa_i - a_i\|^2\right],$$

where a_i are flattened representations of images with $d_f = 784$, D and E are learnable parameters.

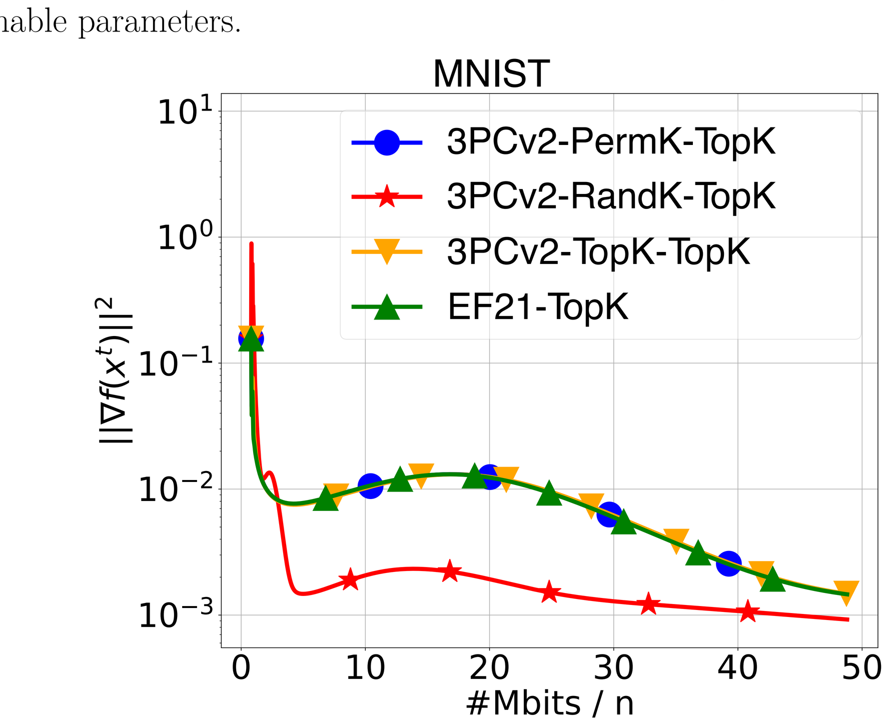

Figure 1:Number of clients $n = 100$, compression level $K = 251$.

◇ **Logistic regression problem** with a non-convex regularizer

$$\min_{x\in\mathbb{R}^d}\left[f(x) := \frac{1}{n}\sum_{i=1}^n\log(1 + e^{-y_i a_i^\top x}) + \lambda\sum_{j=1}^d\frac{x_j^2}{1+x_j^2}\right],$$

where $a_i \in \mathbb{R}^d$, $y_i \in \{-1, 1\}$ are the training data and labels, and $\lambda = 0.1$.

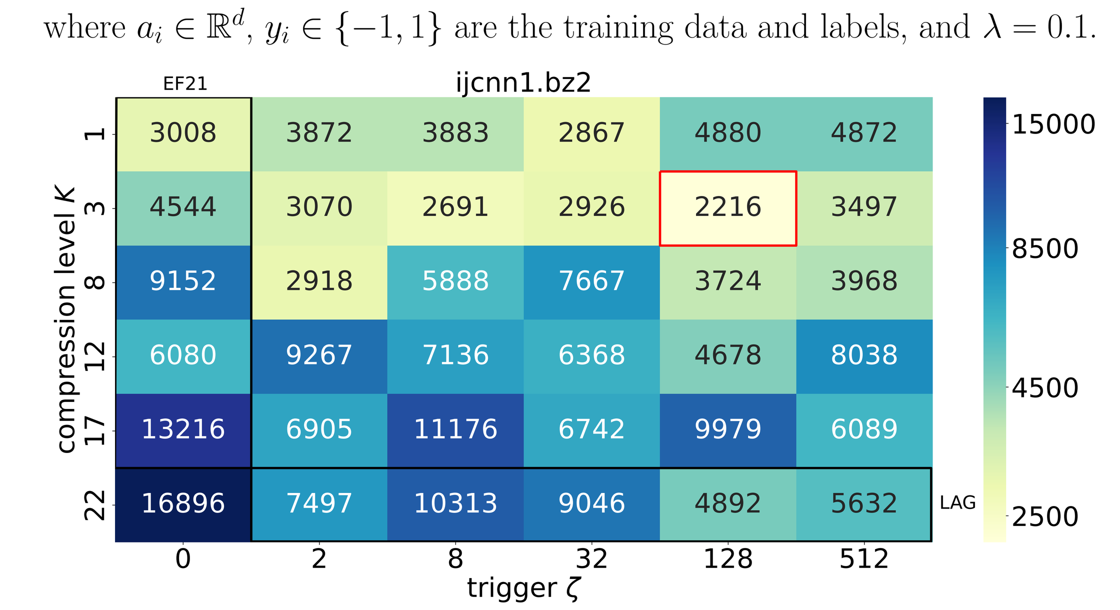

Figure 2:Number of clients $n = 20$. The red-contoured cell indicates the experiment with the smallest communication cost.

◇ Synthetic **quadratic problem**

$$\min_{x\in\mathbb{R}^d}\left[f_i(x) = \frac{1}{n}\sum_{i=1}^n\left(\frac{1}{2}x^\top\mathbf{A}_i x - x^\top b_i\right)\right],$$

where $\mathbf{A}_i \in \mathbb{R}^{d\times d}$, $b_i \in \mathbb{R}^d$, and $\mathbf{A}_i = \mathbf{A}_i^\top$ is the training data that belongs to the device/worker i. In all experiments, we fix $d = 1000$. We refer to the quantity $L_\pm^2 \ge 0$ by the name *Hessian variance* [7], which is defined as

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

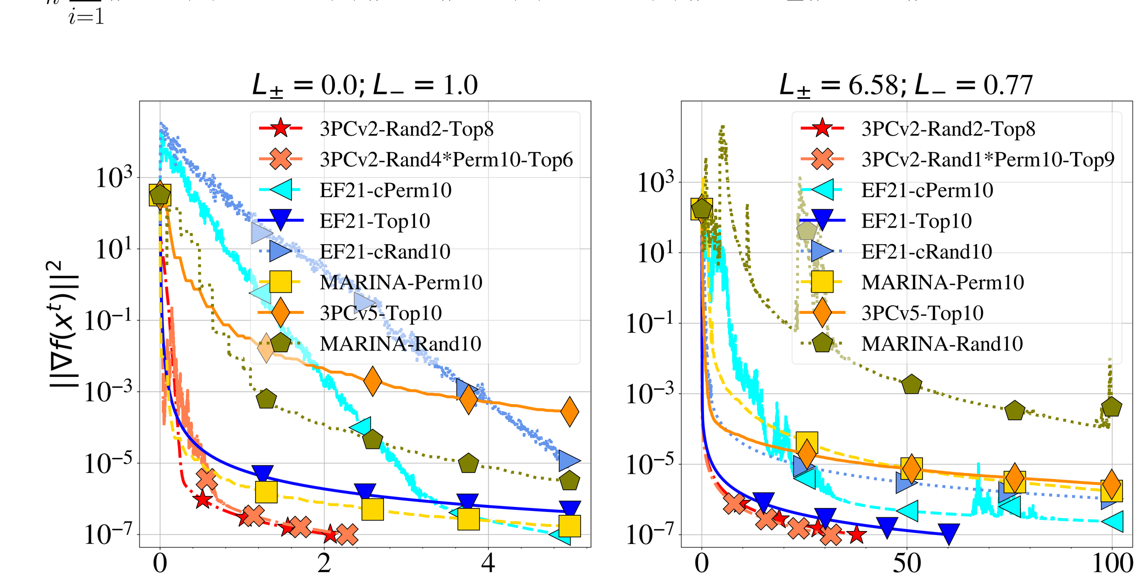

Figure 3:Number of clients $n = 100$.