A Computation and Communication Efficient Method for Distributed Nonconvex Problems in the Partial Participation Setting

Alexander Tyurin and Peter Richtárik
KAUST, Saudi Arabia

1. Distributed Stochastic Optimization Problem

We are solving the nonconvex distributed optimization problem:

$$\min_{x\in\mathbb{R}^d} f(x) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

devices / machines

model parametes / features

Loss on local data $\mathcal{D}_i$ stored on device i

$$f_i(x) = \mathrm{E}_{\xi\sim\mathcal{D}_i}\left[f_i(x;\xi)\right]$$

The datasets $\mathcal{D}_1,\ldots,\mathcal{D}_n$ can be arbitrarily heterogeneous

2. Goal

We want to find a stationary point of the optimization problem:

Find a (possibly random) vector $x \in \mathbb{R}^d$ such that

$$\mathbb{E}\left[\|\nabla f(x)\|^2\right] \leq \varepsilon$$

3. Unbiased Compressor

$$\mathcal{C}_i \in \mathbb{U}(\omega_i) \qquad \begin{array}{c}\mathbb{E}\left[\mathcal{C}_i(v)\right] = v \quad \forall v\in\mathbb{R}^d \\ \mathbb{E}\left[\|\mathcal{C}_i(v)-v\|^2\right] \leq \omega_i\|v\|^2 \quad \forall v\in\mathbb{R}^d\end{array}$$

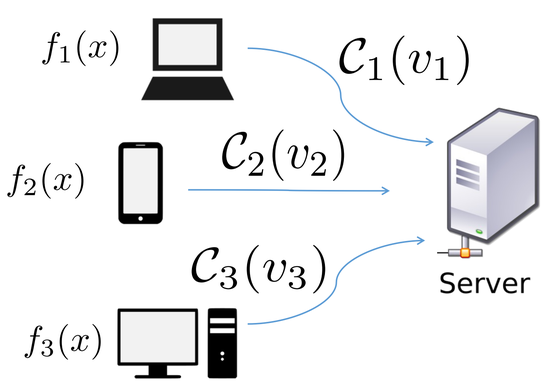
Example: *RandK* compressor

$$d=5; K=1 \qquad \begin{pmatrix}4\\-7\\2\\1\\-3\end{pmatrix} \xrightarrow{\mathcal{C}_i} 5\times\begin{pmatrix}0\\0\\2\\0\\0\end{pmatrix}$$

Random entry

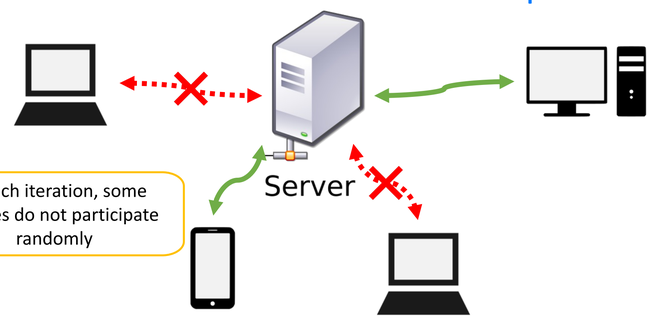
$$\omega_i = \frac{d}{K}-1 = \frac{5}{1}-1 = 4$$

4. Compressed Communication Approach



Clients send compressed vectors $\mathcal{C}_i(v_i)$ to the Server

5. Main Problem: Partial Participation



In each iteration, some devices do not participate randomly

Goal: Design a method that would have *the optimal oracle complexity* and *the state-of-the-art communication complexity* in the partial participation setting

6. New Method: DASHA-PP-MVR

$$x^{t+1} = x^t - \gamma g^t \qquad g^t = \frac{1}{n}\sum_{i=1}^{n} g_i^t$$

The server aggregates information and does the step

if i^{th} node is participating, then

$$k_i^{t+1} = \nabla f_i(x^{t+1},\xi_i^{t+1}) - \nabla f_i(x^t,\xi_i^{t+1}) - b(h_i^t - \nabla f_i(x^t,\xi_i^{t+1})),$$

$$h_i^{t+1} = h_i^t + \frac{1}{p_a}k_i^{t+1},$$

$$g_i^{t+1} = g_i^t + \mathcal{C}_i\left(\frac{1}{p_a}k_i^{t+1} - \frac{a}{p_a}(g_i^t - h_i^t)\right),$$

The nodes calculate new vectors and send the compressed vectors

else

$$h_i^{t+1} = h_i^t, \quad g_i^{t+1} = g_i^t,$$

where p_a is the probability that the node participates, and a and b are momentums.

7. Theoretical Guarantees

Corollary (Informal)**.** The communication complexity equals

$$\mathcal{O}\left(\frac{d\sigma}{\sqrt{p_a}\sqrt{n}\varepsilon} + \frac{d}{p_a\sqrt{n}\varepsilon}\right),$$

SOTA communication complexity

and the expected number of stochastic gradient calculations per node equals

$$\mathcal{O}\left(\frac{\sigma^2}{\sqrt{p_a}n\varepsilon} + \frac{\sigma}{p_a\varepsilon^{3/2}n}\right).$$

optimal oracle complexity

Our assumptions are standard:

Assumption 1. There exists $f^* \in \mathbb{R}$ such that $f(x) \geq f^*$ for all $x \in \mathbb{R}$.

Assumption 2. For all $i \in [n]$, the function f_i is L_i–smooth.

The two assumptions below are provided for the stochastic setting.

Assumption 4. For all $i \in [n]$ and for all $x \in \mathbb{R}^d$, the stochastic gradient $\nabla f_i(x;\xi)$ is unbiased and has bounded variance, i.e.,

$$\mathrm{E}_\xi\left[\nabla f_i(x;\xi)\right] = \nabla f_i(x), \qquad \text{and} \qquad \mathrm{E}_\xi\left[\|\nabla f_i(x;\xi) - \nabla f_i(x)\|^2\right] \leq \sigma^2,$$

where $\sigma^2 \geq 0$.

Assumption 5. For all $i \in [n]$ and for all $x, y \in \mathbb{R}$, the stochastic gradient $\nabla f_i(x;\xi)$ satisfies the mean-squared smoothness property, i.e.,

$$\mathrm{E}_\xi\left[\|\nabla f_i(x;\xi) - \nabla f_i(y;\xi) - (\nabla f_i(x) - \nabla f_i(y))\|^2\right] \leq L_\sigma^2\|x-y\|^2.$$

8. Comparison with Previous Methods

Table 1: Abbr.: *VR* (Variance Reduction) = Does a method have the optimal oracle complexity $\mathcal{O}\left(\frac{\sigma^2}{\varepsilon}+\frac{\sigma}{\varepsilon^{3/2}}\right)$? *PP* (Partial Participation) = Does a method support partial participation? *CC* = Does a method have the communication complexity equals to $\mathcal{O}\left(\frac{\omega}{\sqrt{n}\varepsilon}\right)$?

Method	*VR*	*PP*	*CC*	Limitations
SPIDER, SARAH, PAGE, STORM (Fang et al., 2018; Nguyen et al., 2017) (Li et al., 2021a; Cutkosky and Orabona, 2019)	✓	✗	✗	—
MARINA (Gorbunov et al., 2021)	✓	✗	✓	Suboptimal convergence rate.
FedPAGE (Zhao et al., 2021b)	✗	✗	✗	Suboptimal oracle complexity $\mathcal{O}\left(\frac{\sigma^2}{\varepsilon^2}\right)$.
FRECON (Zhao et al., 2021a)	✗	✓	✓	—
FedAvg (McMahan et al., 2017; Karimireddy et al., 2020b)	✗	✓	✗	Bounded gradients (dissimilarity) assumption of f_i.
SCAFFOLD (Karimireddy et al., 2020b)	✗	✓	✗	Suboptimal convergence rate.
MIME (Karimireddy et al., 2020a)	✗	✓	✗	Calculates full gradient. Bounded gradients (dissimilarity) assumption of f_i. Suboptimal oracle compl. $\mathcal{O}\left(1/\varepsilon^{3/2}\right)$ in the finite-sum setting.
CE-LSGD (for Partial Participation) (Patel et al., 2022) (concurrent work)	✓	✓	✗	Bounded gradients (dissimilarity) assumption of f_i. Suboptimal oracle compl. $\mathcal{O}\left(1/\varepsilon^{3/2}\right)$ in the finite-sum setting.
DASHA (Tyurin and Richtárik, 2023)	✓ or ✗	✗ or ✓	✓	—
DASHA-PP (new)	✓	✓	✓	—

9. New Methods DASHA-PP-PAGE and DASHA-PP-FINITE-MVR in the Finite-Sum Setting

In the finite-sum setting, we assume that $f_i(x) = \frac{1}{m}\sum_{j=1}^{m} f_{ij}(x)$.

Corollary (Informal)**.** The communication complexity is

$$\mathcal{O}\left(d + \frac{d}{p_a\varepsilon\sqrt{n}}\right),$$

SOTA communication complexity

and the expected number of gradient calculations per node equals

$$\mathcal{O}\left(m + \frac{\sqrt{m}}{p_a\varepsilon\sqrt{n}}\right).$$

optimal oracle complexity

Our assumptions are standard:

Assumption 1. There exists $f^* \in \mathbb{R}$ such that $f(x) \geq f^*$ for all $x \in \mathbb{R}$.

Assumption 2. For all $i \in [n]$, the function f_i is L_i–smooth.

The next assumption is used in the finite-sum setting.

Assumption 3. For all $i \in [n], j \in [m]$, the function f_{ij} is L_{ij}-smooth.

References

[1] Alexander Tyurin, Peter Richtárik
DASHA: Distributed Nonconvex Optimization with Communication Compression and Optimal Oracle Complexity
ICLR 2023

[2] Cutkosky, Ashok, and Francesco Orabona
Momentum-based variance reduction in non-convex SGD
NeurIPS 2019