EF21-P and Friends: Improved Theoretical Communication Complexity for Distributed Optimization with Bidirectional Compression

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

Kaja Gruntkowska   Alexander Tyurin   Peter Richtárik
KAUST, Saudi Arabia

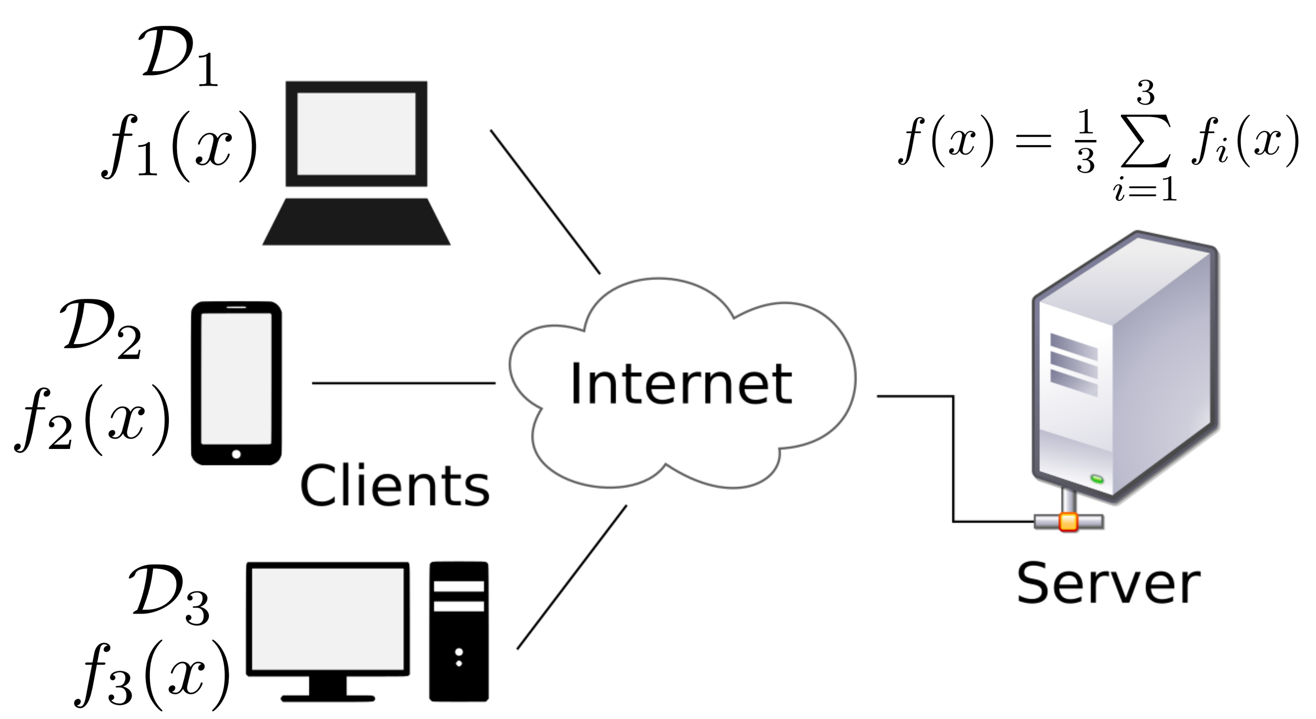
ICML 40 Years
International Conference On Machine Learning

1. Problem

We are solving the distributed optimization problem:

$$\underbrace{\min_{x\in\mathbb{R}^d}}_{\text{\# model parametes / features}} f(x) \stackrel{\text{def}}{=} \frac{1}{\overset{\text{\# devices / machines}}{n}} \sum_{i=1}^{n} \underbrace{f_i(x)}_{\text{The functions } f_i \text{ can be arbitrarily heterogeneous/different}}$$

2 .Technical Setup



Communication is the bottleneck in **both directions**!

3. Main Baseline: Vanilla GD Method

$$x^{t+1} = x^t - \gamma\nabla f(x^t) = x^t - \gamma\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x^t)$$

Recap: Communication Complexity of GD:

$$\underbrace{d}_{\text{\# of coordinates}} \times \underbrace{\mathcal{O}\left(\frac{L}{\mu}\log\frac{1}{\varepsilon}\right)}_{\text{\# of GD iterations}}$$

Recap: Convergence of GD

GD returns an ε-solution (i.e., $\|x^T - x^*\|^2 \leq \varepsilon$) after

$$T = \frac{2L}{\mu}\log\left(\frac{\|x^0-x^*\|^2}{\varepsilon}\right) = \boxed{\mathcal{O}\left(\frac{L}{\mu}\log\frac{1}{\varepsilon}\right)} \text{ iterations.}$$

4. New SOTA: EF21-P + DIANA and EF21-P + DCGD

Algorithm 1 EF21-P + DIANA

```
1: Parameters: learning rates γ > 0 (for learning the model) and β > 0 (for learning the gradient shifts);
   initial model x^0 ∈ R^d (stored on the server and the workers); initial gradient shifts h_1^0, ..., h_n^0 ∈ R^d
   (stored on the workers); average of the initial gradient shifts h^0 = (1/n) Σ_{i=1}^n h_i^0 (stored on the server);
   initial model shift w^0 = x^0 ∈ R^d (stored on the server and the workers)
2: for t = 0, 1, ..., T − 1 do
3:    for i = 1, ..., n in parallel do
4:       m_i^t = C_i^D(∇f_i(w^t) − h_i^t)
5:       Send compressed message m_i^t to the server
6:       h_i^{t+1} = h_i^t + β m_i^t
7:    end for
8:    m^t = (1/n) Σ_{i=1}^n m_i^t
9:    h^{t+1} = h^t + β m^t
10:   g^t = h^t + m^t
11:   x^{t+1} = x^t − γ g^t
12:   p^{t+1} = C^P(x^{t+1} − w^t)
13:   w^{t+1} = w^t + p^{t+1}
14:   Broadcast compressed message p^{t+1} to all n workers
15:   for i = 1, ..., n in parallel do
16:      w^{t+1} = w^t + p^{t+1}
17:   end for
18: end for
```

Lines 3–11: **DIANA** *(idea from [1])*

Lines 12–17: **EF21-P** *(new idea)*

5. Contributions

1. **EF21-P + DIANA** provides **new state-of-the-art convergence rate** for distributed optimization in the **strongly convex** and **general convex** regimes.

2. **EF21-P + DIANA** is the **first method** supporting bidirectional compression whose server-to-workers and workers-to-server communication complexity is **no worse (can be much better!)** than that of vanilla **GD** in the **strongly convex** and **general convex** regimes:

Communication Complexity of EF21-P + DIANA

$$\underbrace{K}_{\text{\# of coordinates}} \times \underbrace{\mathcal{O}\left(\left(\frac{L}{\alpha\mu} + \omega + \frac{\omega L_{\max}}{n\mu}\right)\log\frac{1}{\varepsilon}\right)}_{\text{\# of EF21-P + DIANA iterations}}$$

where ω and α are compression parameters.

3. In the **nonconvex regime**, **EF21-P + DCGD** provides the **new state-of-the-art convergence rate** in the *low accuracy regimes* (ε is small or the # of workers n is large). We provide examples of optimization problems where **EF21-P + DCGD** achieves **new state-of-the-art convergence rate** even in the *high accuracy regime.*

References

[1] Konstantin Mishchenko, Eduard Gorbunov, Martin Takáč, Peter Richtárik
Distributed learning with compressed gradient differences
arXiv:1901.09269

6. Main Tools: Compression Operators

Unbiased compressor

$$\mathbb{E}\left[\mathcal{C}(x)\right] = x,\quad \mathbb{E}\left[\|\mathcal{C}(x)-x\|^2\right] \leq \omega\|x\|^2$$

Example: *RandK* $K = 1$

$$\omega = \frac{d}{K} - 1 = \frac{4}{1} - 1 = 3$$

$$\begin{bmatrix}-3\\2\\6\\-8\end{bmatrix} \xrightarrow{\mathcal{C}} \underbrace{4}_{d/K} \times \begin{bmatrix}0\\2\\0\\0\end{bmatrix}$$

Random entry

Biased compressor

$$\mathbb{E}\left[\|\mathcal{C}(x)-x\|^2\right] \leq (1-\alpha)\|x\|^2 \qquad \alpha \in (0,1]$$

Example: *TopK* $K = 1$

$$\alpha = \frac{K}{d} = \frac{1}{4}$$

$$\begin{bmatrix}-3\\2\\6\\-8\end{bmatrix} \xrightarrow{\mathcal{C}} \begin{bmatrix}0\\0\\0\\-8\end{bmatrix}$$

Largest entry (in magnitude)

7. New SOTA in the Convex Setting

For method comparison, we use the notation $\omega_w \equiv \omega,\ \omega_s \equiv 1/\alpha - 1$

Method	# Communication Rounds
EF (No server-to-worker compression!) (Seide et al., 2014)	$\Omega\left((1+\omega_{\rm w})\frac{L_{\max}}{\mu}\right)$
DIANA (No server-to-worker compression!) (Mishchenko et al., 2019)	$\left(1+\frac{\omega_{\rm w}}{n}\right)\frac{L_{\max}}{\mu} + \omega_{\rm w}$
Dore, Artemis (Liu et al., 2020) (Philippenko & Dieuleveut, 2020)	$\Omega\left(\frac{\omega_{\rm s}\omega_{\rm w}}{n}\frac{L_{\max}}{\mu}\right)$
MCM (Philippenko & Dieuleveut, 2021)	$\Omega\left(\left(\omega_{\rm s}^{3/2} + \frac{\omega_{\rm s}\omega_{\rm w}^{1/2}}{\sqrt{n}} + \frac{\omega_{\rm w}}{n}\right)\frac{L_{\max}}{\mu}\right)$
EF21-P + DIANA (new)	$(1+\omega_{\rm s})\frac{L}{\mu} + \frac{\omega_{\rm w}}{n}\frac{L_{\max}}{\mu} + \omega_{\rm w}$

The relationship between ω_s and $\omega_{\rm w}$ in EF21-P + DIANA is **linear**!

8. New SOTA in the Non-Convex Setting

MCM (Philippenko & Dieuleveut, 2021)	$\frac{\omega_{\rm s}^{3/2}}{\varepsilon} + \frac{\omega_{\rm s}\omega_{\rm w}^{1/2}}{\sqrt{n\varepsilon}} + \frac{\omega_{\rm w}}{n\varepsilon}$
CD-Adam (Wang et al., 2022)	$\frac{\sqrt{d}\max\{\omega_{\rm s},\omega_{\rm w}\}^4}{\varepsilon^2}$
EF21-BC (Fatkhullin et al., 2021)	$\frac{\omega_{\rm w}\omega_{\rm s}}{\varepsilon}$
EF21-P + DCGD (new)	$\frac{\omega_{\rm w}}{n\varepsilon^2} + \frac{\omega_{\rm s}}{\varepsilon}$
EF21-P + DCGD (new)	$\frac{D\omega_{\rm w}}{n\varepsilon} + \frac{\omega_{\rm s}}{\varepsilon}$ (strong-growth assumption)