1. Problem

We are solving the distributed optimization problem:

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

Communication is the bottleneck in both directions!

2. Technical Setup

\[D_1\]
\[D_2\]
\[D_3\]

Communication is the bottleneck in both directions!

3. Main Baseline: Vanilla GD Method

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

Recap: Communication Complexity of GD:

\[d \times O \left( \frac{\mu}{\beta} \log \frac{1}{\epsilon} \right)\]

Recap: Convergence of GD

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

\[T = \frac{2L}{\mu} \log \left( \frac{1}{\varepsilon} \right) \times \left( \frac{1}{\alpha} \log \frac{1}{\epsilon} \right)\]

\[O(\frac{L}{\mu} \log \frac{1}{\epsilon})\] iterations.

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

Algorithm 1 EF21-P + DIANA

1: Parameters: learning rates $\gamma > 0$ (for learning the model) and $\beta > 0$ (for learning the gradient shifts); initial model $x^0 \in \mathbb{R}^d$ (stored on the server and the workers); initial gradient shifts $h^0, \ldots, h^T \in \mathbb{R}^d$ (stored on the server).
2: for $t = 0, 1, \ldots, T - 1$ do
3: for $i = 1, \ldots, n$ in parallel do
4: $m_i \leftarrow \frac{1}{n} \sum_{t=0}^{T} \nabla f_i(w_t)$
5: send compressed message $m_i$ to the server
6: $h^{t+1} \leftarrow h^t + \beta m_i$
7: end for
8: $m^t \leftarrow \frac{1}{n} \sum_{i=1}^{n} m_i$
9: $h^{t+1} \leftarrow h^t + \beta m^t$
10: $p^{t+1} \leftarrow \gamma h^{t+1} + m^t$
11: $p^{t+1} \leftarrow p^{t+1} - \gamma \nabla f_i(w_t)$
12: broadcast compressed message $p^{t+1}$ to all $n$ workers
13: $w_t^{t+1} \leftarrow p^{t+1} + m_t$
14: end for
15: end for
16: $w_T^{t+1} \leftarrow w_T^{t+1} + \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(x^T)$
17: end for
18: end for

Communication Complexity of EF21-P + DIANA

\[K \times O \left( \left( \frac{\mu}{\alpha \mu} + \omega + \frac{\omega L \max \mu}{n \mu} \right) \log \frac{1}{\varepsilon} \right)\]

where $\omega$ and $\xi$ are compression parameters.

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.

GD

Communication Complexity of EF21-P + DIANA

\[K \times O \left( \left( \frac{\mu}{\alpha \mu} + \omega + \frac{\omega L \max \mu}{n \mu} \right) \log \frac{1}{\varepsilon} \right)\]

where $\omega$ and $\xi$ 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 ($\varepsilon$ is small or the # of workers $n_1$ 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.

6. Main Tools: Compression Operators

Unbiased compressor

\[E[\|x\|] = x \quad E[\|x\| - x]^2 \leq \omega \|x\|^2\]

Example: RandK

\[K = 1 \quad \frac{2}{\alpha} < \frac{1}{\omega}\]

Biased compressor

\[E[\|x\|] \leq (1 - \alpha) \|x\|^2 \quad \alpha \in (0, 1)\]

Example: TopK

\[K = 1 \quad \frac{2}{\alpha} < \frac{1}{\omega}\]

7. New SOTA in the Convex Setting

For method comparison, we use the notation $\omega_{\varepsilon} \equiv \omega_\varepsilon \equiv 1/\alpha - 1$.

\[\text{Unbiased} \quad \text{Bias} \quad \omega_{\varepsilon} \equiv \omega_{\varepsilon} \equiv 1/\alpha - 1\]

\[\text{Method} \quad \text{# Communication Rounds} \quad \Omega \left( 1 + \frac{\omega_{\varepsilon}}{\mu} \right)\]

EF

\[\text{Seide et al., 2014}\]

\[\text{DIANA} \quad \text{Mishchenko et al., 2019}\]

\[\text{Dana et al., 2020}\]

\[\text{Philippenko & Dieuleveut, 2020}\]

\[\text{MCM} \quad \text{Philippenko & Dieuleveut, 2021}\]

\[\text{EF21-P + DIANA (new)} \quad \Omega \left( \sqrt{\frac{\omega_{\varepsilon}^2 + \omega_{\varepsilon}}{n_1}} + \frac{\omega_{\varepsilon}}{n_1} \right)\]

\[\text{EF21-P + DCGD (new)} \quad \omega_{\varepsilon} \equiv \omega_{\varepsilon} \equiv 1/\alpha - 1\]

The relationship between $\omega_{\varepsilon}$ and $\omega_{\varepsilon}$ in EF21-P + DIANA is linear!

8. New SOTA in the Non-Convex Setting

\[\text{MCM} \quad \text{Philippenko & Dieuleveut, 2021}\]

\[\text{CD-Adam (Wang et al., 2022)} \quad \sqrt{\frac{\omega_{\varepsilon}^2 + \omega_{\varepsilon}}{n_1}} \leq \omega_{\varepsilon}\]

\[\text{EF21-BC (Fakhtkhullin et al., 2021)} \quad \frac{\omega_{\varepsilon}}{n_1} \leq \omega_{\varepsilon}\]

\[\text{EF21-P + DCGD (new)} \quad \frac{\omega_{\varepsilon}}{n_1} + \frac{\omega_{\varepsilon}}{n_1} \leq \omega_{\varepsilon}\]

(Strong-growth assumption)

References

[1] Konstantin Mishchenko, Eduard Gorbunov, Martin Takáč, Peter Richtárik

Distributed learning with compressed gradient differences

arXiv:1901.09269