Introduction

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

Communication compression is a key strategy to speed up distributed optimization. While existing decentralized algorithms with compression mainly focus on DGD-type algorithm, in this work, we introduce the first primal-dual decentralized optimization algorithm with compression, LEAD. LEAD achieves better convergence, stability, and heterogeneous data supports. LEAD recovers NIDS when there is no compression.

Algorithm

\[ \begin{align*}
\text{Algorithm} & \quad X^k := \left[ \begin{array}{c}
\left( x_1^k \right)^T \\ldots \\left( x_n^k \right)^T 
\end{array} \right] \in \mathbb{R}^{n \times d}, \quad \nabla f(Y^k) := \left[ \begin{array}{c}
\left( \nabla f_i(x_1^k) \right)^T \\ldots \\left( \nabla f_i(x_n^k) \right)^T 
\end{array} \right] \in \mathbb{R}^{n \times d} \\
\text{Symmetric } W & \in \mathbb{R}^{n \times n} \text{ encodes the communication network.} \\
-1 \leq \lambda_{\min}(W) & \leq \lambda_{\max}(W) \leq \cdots \leq \lambda_{n}(W) = 1. 
\end{align*} \]

Keys: (1) Gradient correction: the dual variable D corrects the non-zero gradient at the convergence. (2) Difference compression on Y reduces the compression error and H tracks the dynamic of Y using compressed information from Q. (3) The inexact dual update implicitly compensates the compression error locally in each agent.

\[ \begin{align*}
X^{k+1} & = X^k - \eta \nabla f(X^k; \xi) - \eta D^k \\
Y^{k+1} & = \text{CompressionProcedure}(Y^k) \\
D^{k+1} & = D^k - \frac{1}{2} (I - W) Y^{k+1} = \frac{1}{2} (\hat{Y}^k - Y^k) \\
X^{k} & = X^k - \eta \nabla f(X^k; \xi) - \eta D^{k+1} \\
\text{CompressionProcedure} & \quad Q^k = \text{Compress}(Y^k - H^k) \quad \text{Compression} \\
& \quad \hat{Y}^k = H^k + \bar{Q}^k \\
& \quad H^{k+1} = (1 - \alpha) H^k + \alpha \hat{Y}^k \\
& \quad H_{\text{comm}}^{k+1} = (1 - \alpha) H_{\text{comm}}^k + \alpha \hat{Y}_{\text{comm}}^k \\
\end{align*} \]

Theory

**Assumptions**

- Compression: \[ E(\mathbf{Q}(x) = x, E\|x - Q(x)\|_2^2 \leq C\|x\|_2^2 \text{ for some } C \geq 0. \]
- \( f_i() \) is \( \mu \)-strongly convex and \( L \)-smooth:
  \[ f_i(x) \geq f_i(y) + \langle \nabla f_i(y), x - y \rangle + \frac{\mu}{2} \| x - y \|_2^2, \]
  \[ f_i(x) \leq f_i(y) + \langle \nabla f_i(y), x - y \rangle + \frac{L}{2} \| x - y \|_2^2. \]
- Gradient: \[ E_x \nabla f_i(x; \xi) = \nabla f_i(x), \quad E_x \| \nabla f_i(x; \xi) - \nabla f_i(x) \|_2^2 \leq \sigma^2. \]

Define the conditional numbers on the objectives and the communication graph \( W \):

\[ k_i = \frac{1}{\lambda_{\max}(1 - W)} \quad k_g = \frac{\lambda_{\min}(1 - W)}{\lambda_{\min}(1 - W) - \lambda_{\max}(W)} \]

**Convergence properties**

**Theorem (Complexity bounds when \( \sigma = 0 \)).**

- LEAD converges to the \( \epsilon \)-accurate solution with the iteration complexity:
  \[ O\left( (1 + C)(k_i + k_g) + C k_i k_g \log \frac{1}{\epsilon} \right). \]
- When \( C = 0 \) (i.e., no compression) or \( C \leq \frac{k_i}{k_i + k_g} \), the iteration complexity is
  \[ O\left( k_i + k_g \log \frac{1}{\epsilon} \right). \]

This recovers the convergence rate of NIDS [LSY19].

- With \( C = 0 \) (or \( C \leq \frac{k_i}{k_i + k_g} \)) and fully connected communication graph (i.e., \( W = W_0 \)), the iteration complexity is
  \[ O(\epsilon \log \frac{1}{\epsilon}). \]

This recovers the convergence rate of gradient descent.

**Theorem (Error bound when \( \sigma > 0 \)).**

\[ \frac{1}{n} \sum_{i=1}^{n} \| x^* - x^k \|_2^2 \leq O\left( \frac{1}{\epsilon} \right) \]

Experiment

Setup: 8 machines connected in a ring topology networks. The mixing weight is set as 1/3. We use the stochastic b-bits quantization (b=2) method for compression.

Conclusion

- LEAD converges fast and supports heterogeneous data well.
- LEAD supports unbiased compression of arbitrary precision.
- LEAD is robust to parameter setting and is easy to tune.

Reference

Zhi Li, Wei, Shi and Ming Yan. A Decentralized Proximal-Gradient Method With Network Independent Step-Sizes and Separated Convergence Rates, 2019 IEEE Transactions on Signal Processing