ICML 2019

Decentralized Stochastic Optimization and Gossip Algorithms with Compressed Communication



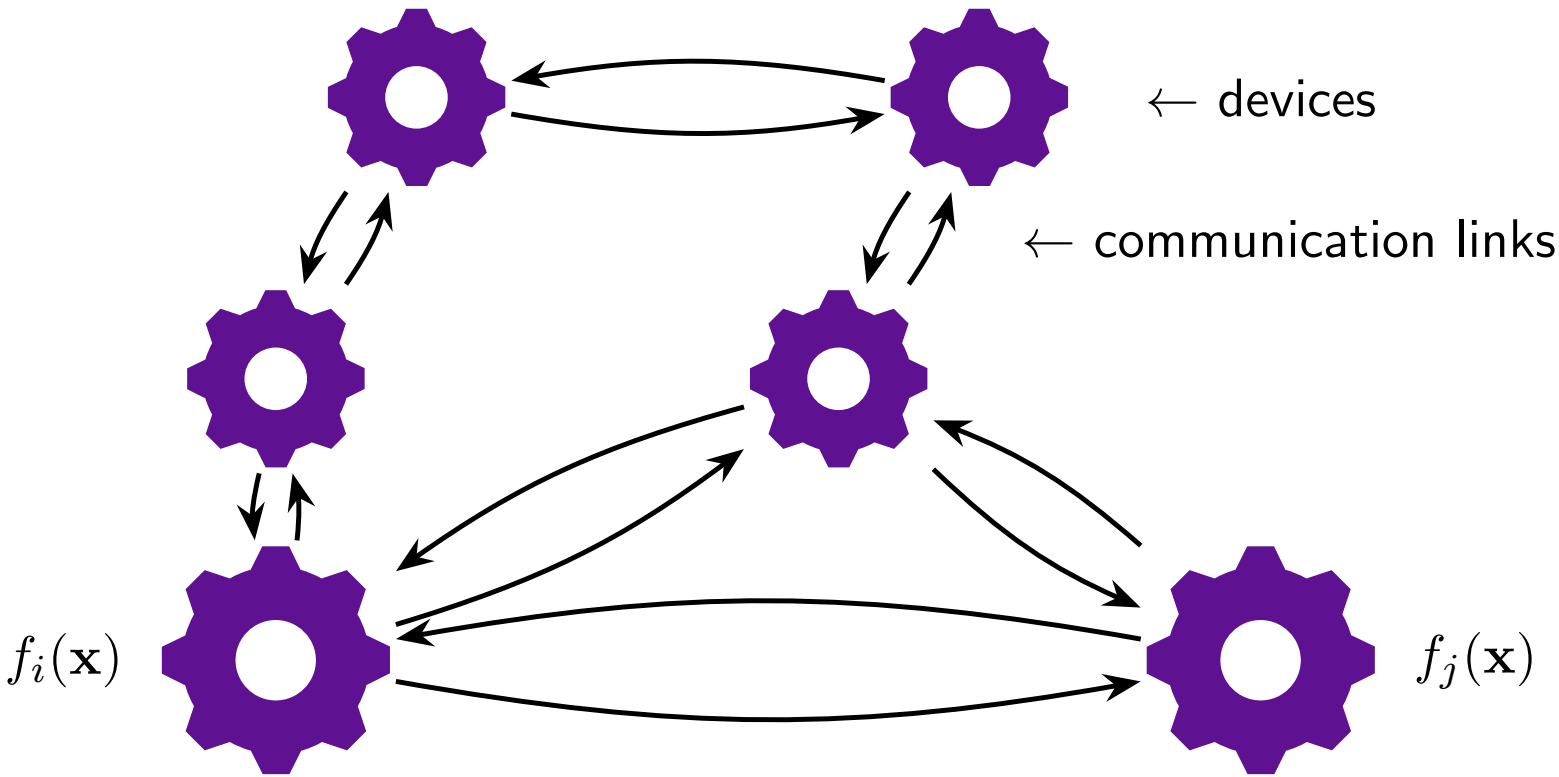
Anastasia Koloskova, Sebastian U. Stich, Martin Jaggi

EPFL, Switzerland
mlo.epfl.ch

June 11, 2019

Decentralized Stochastic Optimization

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



each device has oracle access to stochastic gradients

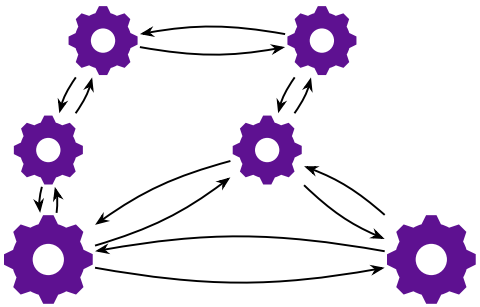
$$\mathbf{g}_i(\mathbf{x}),\ \mathbb{E}\mathbf{g}_i(\mathbf{x}) = \nabla f_i(\mathbf{x}),\ \mathrm{Var}[\mathbf{g}_i] \leq \sigma_i^2$$

Decentralized Stochastic Optimization

Applications: servers, mobile devices, sensors, hospitals, ...

Advantages:
- no central coordinator
- local communication vs. all-reduce
- data distributed (storage & privacy aspects)

This work: bandwidth restricted setting
where communication is a bottleneck



Data Compression for Efficient Communication

Communication Compression:
Compress models/model updates before sending over the network.

This work:
Arbitrary compressors, supporting the main SOTA techniques!

General Compressor: $Q\colon \mathbb{R}^d \to \mathbb{R}^d$ **can be biased!**

$$\mathbb{E}_Q \left\|\mathbf{x} - Q(\mathbf{x})\right\|^2 \leq (1-\delta)\left\|\mathbf{x}\right\|^2 \qquad \forall \mathbf{x} \in \mathbb{R}^d$$

Examples: Quantization, rounding, sign, top-k, rank-k

Main Contribution: CHOCO-SGD

We propose CHOCO-SGD: a decentralized SGD algorithm with communication compression.

Main result: CHOCO-SGD converges at the rate

$$f(\bar{\mathbf{x}}_T) - f^\star = \mathcal{O}\left(\underbrace{\frac{\bar{\sigma}^2}{\mu n T}}_{\substack{\text{linear speedup} \\ \text{matches centralized baseline}}} + \underbrace{\frac{1}{\mu^2 \delta^2 \rho^4 T^2}}_{\substack{\text{higher order term, accounting} \\ \text{for topology and compression}}} \right)$$

f μ-strong convex, variance $\bar{\sigma} = \frac{1}{n}\sigma_i^2$, spectral gap of topology $\rho > 0$

- first scheme with linear speedup for arbitrary compressors
- improves over previous approach **[Tang et al., Neurips 18]**

Key Technique: CHOCO-Gossip

We propose CHOCO-Gossip: a new algorithm with communication compression for the average consensus problem:

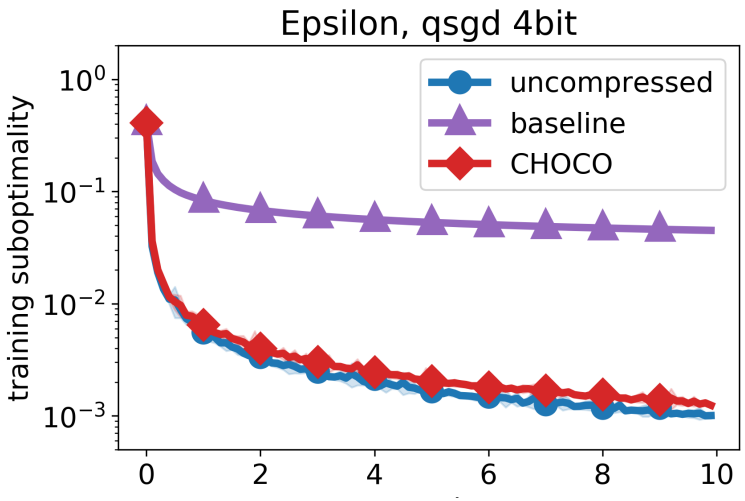
$$\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i$$

classic gossip averaging **[Xiao & Boyd, 04]** + compression with error feedback **[Stich et al., NeurIPS 18]**

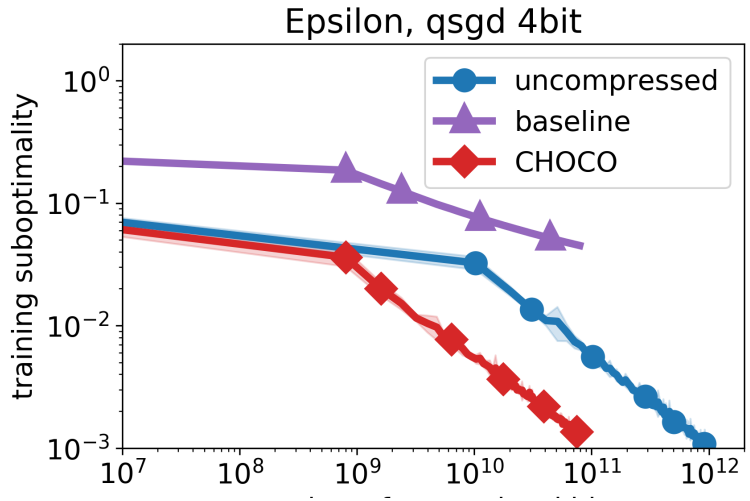
- linear convergence for arbitrary compressors
- all previous gossip schemes with compression did not converge linearly (or not at all) for arbitrary compressors

Experimental Results

Example: quantization to 4bits



epochs



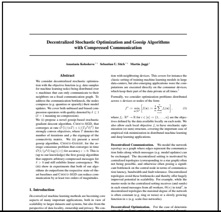
transmitted data

Logistic regression on `epsilon` dataset, ring topology with $n = 9$ nodes.

Summary

+ compression with error feedback gives drastic reduction in communication, without hurting the convergence
+ first compressed gossip scheme that converges at linear rate
+ first decentralized SGD with compressed communication that converges for arbitrary compression (without hampering the rate)

Compression **for free**, by enabling error feedback in the decentralized setting



Poster #197