Accelerated Decentralized SGD

Chebykin Simon

Scientific adviser - Aleksandr Beznosikov

May 5, 2024

Table of Contents

Table of Contents

We study the distributed stochastic optimization problem

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

where the components $f_i : R^d \to R$ are distributed among n nodes and are given in stochastic form:

$$f_i(x) := \mathbb{E}_{\xi_i \sim D_i} F_i(x_i, \xi_i) \tag{2}$$

- Privacy
  - f_i are stored locally
- Scalability
  - What if we add more devices?
- Acceleration

Under assumptions:

- F is L-smooth and f is μ-convex
- W, communication matrix is doubly stochastic, and mixes (p, c).
- Bounded noise:

$$\sum \mathbb{E}\|\nabla F_i(x_i, \xi_i) - \nabla f(x_i)\|^2 < \sigma^2$$

The main result of [1] is the following

Theorem. The problem (1) can be solved (in terms of weighted sum of function error and argument error) in

$$\tilde{\mathcal{O}}\left(\frac{\sigma^2}{\mu n \varepsilon} + \frac{\sqrt{L}\sigma}{\mu\sqrt{p}c\sqrt{\varepsilon}} + \frac{L}{\mu p c}\log\frac{1}{\varepsilon}\right) \quad \text{iterations}\,. \tag{3}$$

The main result of [2] is the following

Theorem. The problem (1) can be solved (in terms of weighted sum of function error and argument error) in

$$\tilde{\mathcal{O}}\left(\frac{\sigma^2}{\mu n \varepsilon} + \frac{\sqrt{L}(\zeta\tau + \sigma\sqrt{p\tau}}{\mu p c\sqrt{\varepsilon}} + \frac{L\tau}{\mu p}\log\frac{1}{\varepsilon}\right) \quad \text{iterations}\,. \tag{4}$$

Table of Contents

Bibliography I

[1] Sebastian U. Stich Anastasia Koloskova Tao Lin. *An Improved Analysis of Gradient Tracking for Decentralized Machine Learning.* 2022. arXiv: 2202.03836.

[2] Anastasia Koloskova Nicolas Loizou Sadra Boreiri Martin Jaggi Sebastian U. Stich. *A Unified Theory of Decentralized SGD with Changing Topology and Local Updates.* 2020. arXiv: 2003.10422.