Robust and Communication-Efficient Collaborative Learning

Amirhossein Reisizadeh[1]

Hossein Taheri[1] Aryan Mokhtari[2] Hamed Hassani[3] Ramtin Pedarsani[1]

1 UCSB 2 UT Austin 3 UPenn

Collaborative Learning

Collaborative learning: A task of learning a common objective among multiple computing agents without any central node and by using on-device computation and local communication among agents.

- In context of machine learning and optimization
- Applications in
  - Distributed deep learning
  - Industrial IoT
  - Smart Healthcare

We consider the *decentralized* implementation:

- general data-parallel setting
- the data is distributed across different computing nodes
- local computation
- communication among neighbors

Challenges in Decentralized Implementation

Straggling nodes

Nodes randomly slow down in their local computation.

Communication load

Message passing algorithm induces large communication overhead.

Our Goal

To develop decentralized optimization methods while addressing the above two challenges, i.e. *robust* and *communication-efficient*.

Problem Setup

- Stochastic learning model $\min_{\mathbf{x}} \mathbb{E}_{\theta\sim\mathcal{P}}[\ell(\mathbf{x},\theta)]$
- Empirical risk model $\min_{\mathbf{x}} L_N(\mathbf{x}) := \min_{\mathbf{x}} \frac{1}{N}\sum_{k=1}^{N} \ell(\mathbf{x},\theta_k)$
- Collaborative learning model
  - A network of n nodes, weight matrix W
  - Local loss for node i $\quad f_i(\mathbf{x}) := \frac{1}{m}\sum_{j=1}^{m} \ell(\mathbf{x},\theta_i^j)$
  - Global loss $\min_{\mathbf{x}} f(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}) = \min_{\mathbf{x}} \frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m} \ell(\mathbf{x},\theta_i^j)$

Our Proposal: `QuanTimed-DSGD`

At iteration t and node i:

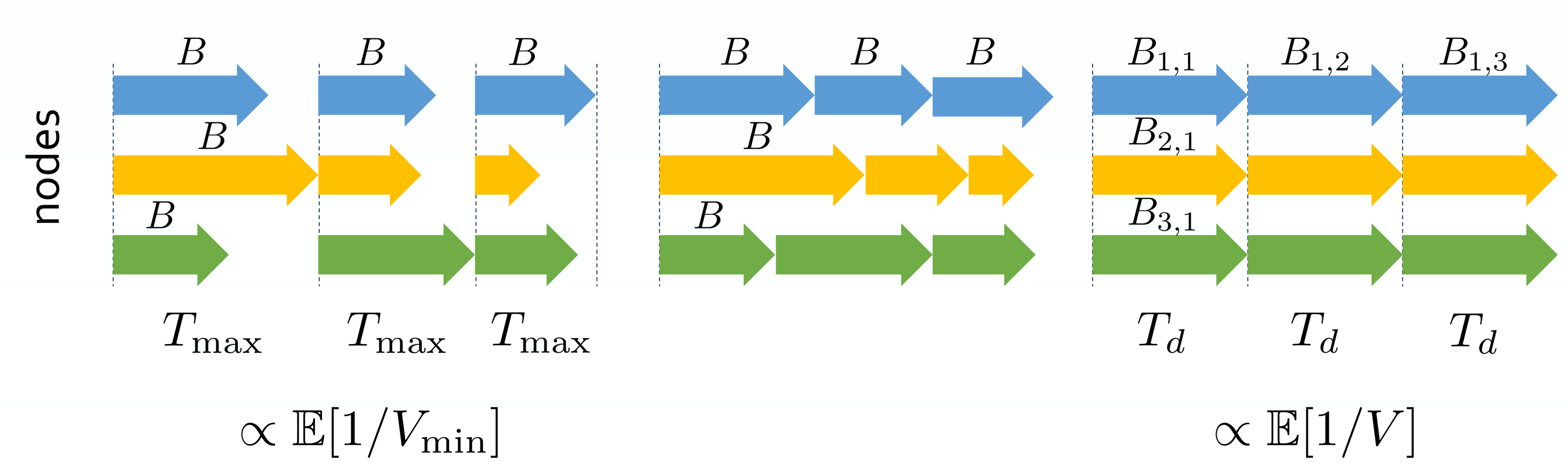
- *Deadline-Based Gradient Computation*
  - A deadline T_d is fixed
  - Node i computes gradients on (random) sample subset $\mathcal{S}_{i,t}$

$$\widetilde{\nabla} f_i(\mathbf{x}_{i,t}) = \frac{1}{|\mathcal{S}_{i,t}|}\sum_{\theta\in\mathcal{S}_{i,t}} \nabla\ell(\mathbf{x}_{i,t};\theta)$$

  - Computation time: random speed $V_{i,t} \sim F_V \Rightarrow |\mathcal{S}_{i,t}| = T_d V_{i,t}$
- *Quantized Message-Passing*
  - Nodes exchange quantized models $\mathbf{z}_{i,t} = Q(\mathbf{x}_{i,t})$
- *Update*

$$\mathbf{x}_{i,t+1} = (1-\varepsilon+\varepsilon w_{ii})\mathbf{x}_{i,t} + \varepsilon\sum_{j\in\mathcal{N}_i} w_{ij}\mathbf{z}_{j,t} - \alpha\varepsilon\widetilde{\nabla} f_i(\mathbf{x}_{i,t})$$

✓ Iteration time implication:



`QuanTimed-DSGD` in Theory

Assumptions.

A1. Weight matrix W is doubly stochastic.

A2. Random quantizer $Q(.)$ is unbiased & variance-bounded.

A3. Loss function ℓ is K-smooth.

A4. Stochastic gradients $\nabla\ell(\mathbf{x};\theta)$ are unbiased & variance-bounded.

Convergence for non-convex losses

- ✓ Assumptions A1-4
- ✓ Large enough iterations T
- ✓ Pick step-sizes $\alpha = T^{-1/6}$ and $\varepsilon = T^{-1/3}$

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}||\nabla f(\bar{\mathbf{x}}_t)||^2 = \mathcal{O}\left(\frac{1}{T^{1/3}}\right)$$

$$\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}||\bar{\mathbf{x}}_t - \mathbf{x}_{i,t}||^2 = \mathcal{O}\left(\frac{1}{T^{1/3}}\right)$$

A5. Loss function ℓ is μ-strongly convex.

Convergence for strongly-convex losses

- ✓ Assumptions A1-5
- ✓ Pick $\delta \in (0,1/2)$
- ✓ Large enough iterations T
- ✓ Pick step-sizes $\alpha = T^{-\delta/2}$ and $\varepsilon = T^{-3\delta/2}$

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}||\mathbf{x}_{i,t} - \mathbf{x}^*||^2 = \mathcal{O}\left(\frac{1}{T^{\delta}}\right)$$

`QuanTimed-DSGD` in Simulation

Binary classification over CIFAR-10