MICHIGAN STATE UNIVERSITY

Decentralized Algorithm&Compressed SGD

Yao Li[†‡]

[†]Department of Mathematics
[‡]Department of Computational Mathematics, Science and Engineering

Outline

MICHIGAN STATE UNIVERSITY

1. Decentralized Algorithm
2. Communication Compressed SGD
3. Further Topics

Problem Description

- Problem formulation

$$\underset{x \in \mathbb{R}^d}{\text{minimize}} \quad \bar{f}(x) := \sum_{i=1}^{n} f_i(x) \qquad (1)$$
$$\text{subject to} \quad (i, j) \in \mathcal{G}$$

  where each f_i known by agent i privately is proper, convex, closed and $\mathcal{G}$ is a connected undirected graph.
- Each f_i is L-smooth, i.e. ∇f_i is L-Lipschitz continuous.
- $\mathbf{W}$ as mixing matrix encodes graph topology and communication weights.

Mixing Matrix

MICHIGAN STATE UNIVERSITY

- $\mathbf{W}$ is symmetric and satisfies
  - $\mathbf{W}\mathbf{1} = \mathbf{1}$
  - $\lambda(\mathbf{W}) \in (-1, 1]$ and $\mathbf{Null}(\mathbf{I} - \mathbf{W}) = \mathbf{span}\{\mathbf{1}\}$
- $\mathbf{W}$ can be constructed by
  - Laplacian matrix $\mathbf{L}$ of $\mathcal{G}$
  - Metropolis constant edge weight
  - Symmetric fastest distributed linear averaging problem [5].

Consensus Problem

- The consensus problem will be solved instead

$$\begin{aligned} \underset{\mathbf{x}=[x_1,\cdots,x_n]^T\in\mathbb{R}^{n\times d}}{\text{minimize}} \quad & \mathbf{f}(\mathbf{x}) := \sum_{i}^{n} f_i(x_i) \\ \text{subject to} \quad & \mathbf{W}\mathbf{x} = \mathbf{x} \end{aligned} \tag{2}$$

- The optimality implies

$$(\mathbf{W} - \mathbf{I})\mathbf{x}^* = \mathbf{0},$$

i.e., consensus $x_1^* = \cdots = x_n^*$.

DGD

MICHIGAN STATE UNIVERSITY

- Decentralized Gradient Descent (DGD) combines gossip algorithm and gradient descent (GD)

$$\mathbf{x}^{k+1} = \mathbf{W}\mathbf{x}^k - \alpha \nabla \mathbf{f}(\mathbf{x}^k), \tag{3}$$

- Equivalent to use GD to solve

$$\underset{\mathbf{x}\in\mathbb{R}^{n\times d}}{\text{minimize}} \quad \mathbf{f}(\mathbf{x}) + \frac{1}{\alpha}(\mathbf{I} - \mathbf{W})\mathbf{x} \tag{4}$$

- Fixed stepsize $\alpha \in (0, \lambda_{\min}(\mathbf{I} + \mathbf{W})/L)$ only achieves inexact linear convergence for strongly convex (SC) f_is, while diminishing stepsize can give exact convergence only at sublinear rate [6].

EXTRA and NIDS

MICHIGAN STATE UNIVERSITY

- EXTRA [4] uses one more step parameter in update.

$$\mathbf{x}^{k+2} = \frac{\mathbf{I}+\mathbf{W}}{2}\left[2\mathbf{x}^{k+1} - \mathbf{x}^{k}\right] - \alpha\nabla\mathbf{f}(\mathbf{x}^{k+1}) + \alpha\nabla\mathbf{f}(\mathbf{x}^{k}), \qquad (5)$$

  where $\alpha \in (0, \lambda_{\min}(\mathbf{I}+\mathbf{W})\mu/L^2)$ under SC assumption on $\bar{f}$.
- Exact linear convergence is comparable to centralized algorithm.
- The upper bound on α is proportional to $\frac{\mu}{L}$, much smaller than centralized one .

EXTRA and NIDS

MICHIGAN STATE UNIVERSITY

- NIDS [2] communicates gradient information, compared to EXTRA.

$$\mathbf{x}^{k+2} = \frac{\mathbf{I}+\mathbf{W}}{2}\left[2\mathbf{x}^{k+1} - \mathbf{x}^{k} - \alpha\nabla\mathbf{f}(\mathbf{x}^{k+1}) + \alpha\nabla\mathbf{f}(\mathbf{x}^{k})\right], \quad (6)$$

  where $\alpha \in (0, 2/L)$ under SC assumption on each f_is.
- The upper bound of α coincides with the centralized one and is independent of the mixing matrix.
- The assumption on f_is is stronger than strong convex assumption on $\bar{f}$.

Improvement

The improvement of EXTRA and NIDS in [1] also includes the mixing matrix, i.e., the relaxed mixing matrix can be use to accelerate algorithms.

	EXTRA [4]	EXTRA [1]	NIDS [2]	NIDS [1]
$\lambda(\mathbf{W})$	$(-1,1]$	$(-5/3,1]$	$(-1,1]$	$(-5/3,1]$
f_is	SC on $\bar{f}$	SC on $\bar{f}$	SC on f_is	SC on $\bar{f}$
$\alpha_{\max}$	$\frac{(1+\lambda_{\min}(\mathbf{W}))\mu}{L^2}$	$\frac{5+3\lambda_{\min}(\mathbf{W})}{4L}$	$\frac{2}{L}$	$\frac{2}{L}$

The gap between decentralized and centralized algorithm is closed in the aspect of linear convergence and largest stepsize.

Experiment

Linear regression with strongly convex $\bar{f}$.

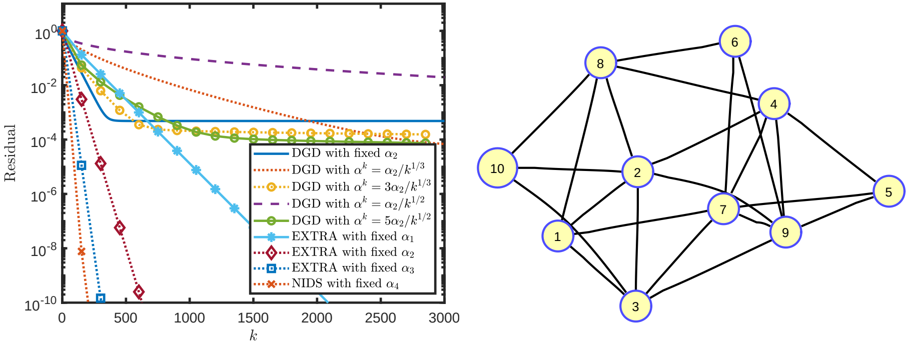

Figure 1: LHS: the error $\frac{\|\mathbf{x}^k - \mathbf{x}^*\|_F}{\|\mathbf{x}^0 - \mathbf{x}^*\|_F}$ vs iterations for DGD with different stepsizes, EXTRA with three stepsizes, and NIDS. RHS: The random network with 10 nodes.

Distributed Scheme

- The following distributed scheme is considered for problem

$$\underset{x \in \mathbb{R}^d}{\text{minimize}} \quad f(x) + R(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x) + R(x), \tag{7}$$

  where f is smooth and R is a nonsmooth regularizer.
- Each f_i is L-smooth and strongly convex in convex setting.
- $R = 0$ and f_i is L-smooth in nonconvex setting.

Distributed Scheme

MICHIGAN STATE UNIVERSITY

- Distributed Stochastic Gradient Descent (DSGD)
  - Master: receive g_is, get $\bar{g} = \frac{1}{n}\sum_{i=1}^{n} g_i$, update model parameter $x = \mathbf{prox}_{\gamma R}(x - \gamma\bar{g})$ and broadcast x.
  - Worker i: receive x, sample g_i based on local data such that $\mathbb{E}[g_i|x] = \nabla f_i(x)$ and send gradient parameter g_i.
- When bandwidth is limited, the communication dominates the convergence.
- Compressed(Quantized) low-bit parameter will be used instead.

Communication Reduction

MICHIGAN STATE
UNIVERSITY

There are mainly two kinds of methods to compress parameter:

- Deterministic Method:
  - Top-k Sparsification, e.g. $[1, 100, 1, 1, 1] \rightarrow [0, 100, 0, 0, 0]$
  - Clipping, e.g, $1.23456 \rightarrow 1.2$
  - 1-Bit Quantization, i.e., compress x into $\|x\|\mathrm{sign}(x)$
- Stochastic Method:
  - Randomized Quantization
  - P-norm Quantization
  - Randomized Sparsification

All stochastic method will generate unbiased estimator parameter, i.e., $\mathbb{E}[Q(x)] = x$.

DORE: DOuble REsidual compression SGD

DORE is proposed in [3] using stochastic method to compress the residual of parameters on both master and worker nodes.

Algorithm 1 The Proposed DORE.[1]

Input: Stepsize $\alpha, \beta, \gamma, \eta$, initialize $\mathbf{h}^0 = \mathbf{h}_i^0 = \mathbf{0}^d$, $\hat{\mathbf{x}}_i^0 = \hat{\mathbf{x}}^0$, $\forall i \in \{1, \ldots, n\}$.
for $k = 1, 2, \cdots, K-1$ **do**
  For each worker $i \in \{1, 2, \cdots, n\}$:
    Sample $\mathbf{g}_i^k$ such that $\mathbb{E}[\mathbf{g}_i^k | \hat{\mathbf{x}}_i^k] = \nabla f_i(\hat{\mathbf{x}}_i^k)$
    Gradient residual: $\Delta_i^k = \mathbf{g}_i^k - \mathbf{h}_i^k$
    Compression: $\hat{\Delta}_i^k = Q(\Delta_i^k)$
    $\mathbf{h}_i^{k+1} = \mathbf{h}_i^k + \alpha \hat{\Delta}_i^k$
    $\{ \hat{\mathbf{g}}_i^k = \mathbf{h}_i^k + \hat{\Delta}_i^k \}$
    Send $\hat{\Delta}_i^k$ to the master
    Receive $\hat{\mathbf{q}}^k$ from the master
    $\hat{\mathbf{x}}_i^{k+1} = \hat{\mathbf{x}}_i^k + \beta \hat{\mathbf{q}}^k$
  For the master:
    Receive $\{\hat{\Delta}_i^k\}$ from workers
    $\hat{\Delta}^k = 1/n \sum_i^n \hat{\Delta}_i^k$
    $\hat{\mathbf{g}}^k = \mathbf{h}^k + \hat{\Delta}^k \quad \{= 1/n \sum_i^n \hat{\mathbf{g}}_i^k \}$
    $\mathbf{x}^{k+1} = \mathbf{prox}_{\gamma R}(\hat{\mathbf{x}}^k - \gamma \hat{\mathbf{g}}^k)$
    $\mathbf{h}^{k+1} = \mathbf{h}^k + \alpha \hat{\Delta}^k$
    Model residual: $\mathbf{q}^k = \mathbf{x}^{k+1} - \hat{\mathbf{x}}^k + \eta \mathbf{e}^k$
    Compression: $\hat{\mathbf{q}}^k = Q(\mathbf{q}^k)$
    $\mathbf{e}^{k+1} = \mathbf{q}^k - \hat{\mathbf{q}}^k$
    $\hat{\mathbf{x}}^{k+1} = \hat{\mathbf{x}}^k + \beta \hat{\mathbf{q}}^k$
    Broadcast $\hat{\mathbf{q}}^k$ to workers
end for
Output: $\hat{\mathbf{x}}^K$ or any $\hat{\mathbf{x}}_i^K$

DORE: DOuble REsidual compression SGD

MICHIGAN STATE UNIVERSITY

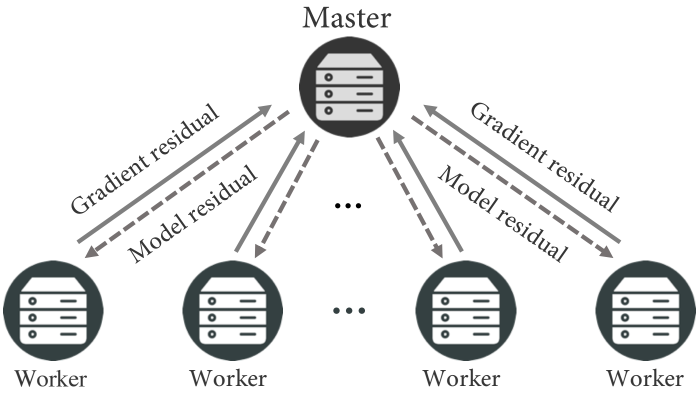

Figure 2: An illustration of DORE

Convergence Result

The following assumptions are made throughout the paper:

- Bounded variance of estimator on gradient, i.e., $\mathbb{E}[\|g_i - \nabla f_i(x)\|^2] \leq \sigma_i^2$.
- Bounded signal-to-noise factor, i.e., $\mathbb{E}[\|Q(x) - x\|^2] \leq C\|x\|^2$.

In convex setting, DORE converges to the neighborhood of optimal point linearly.
In nonconvex setting, the similar rate to the vanilla DSGD is achieved

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\nabla f(x^k)\|^2 \lesssim \frac{1}{K} + \frac{1}{\sqrt{Kn}}.$$

Experiments

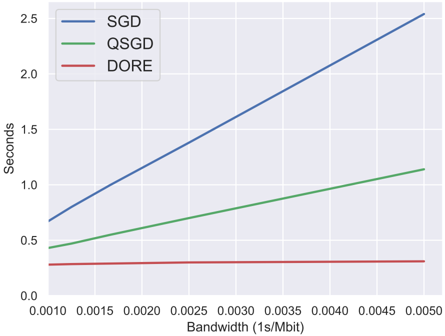

Figure 3: Per iteration time cost on Resnet18 for SGD, QSGD, and DORE. It is tested in a shared cluster environment connected by Gigabit Ethernet interface. DORE speeds up the training process significantly by mitigating the communication bottleneck.

Experiments

MICHIGAN STATE UNIVERSITY

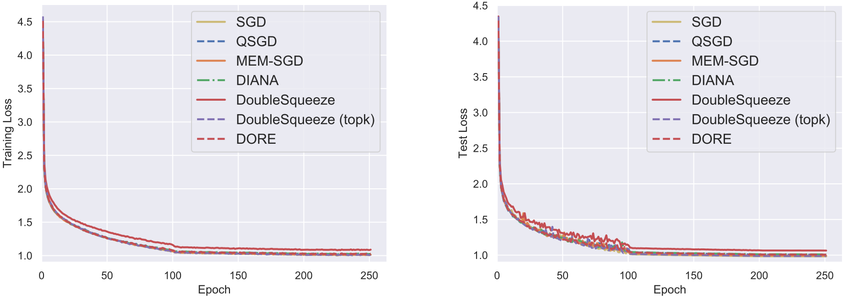

Figure 4: Resnet18 trained on CIFAR10. DORE achieves similar convergence and accuracy as most baselines. DoubeSuqeeze converges slower and suffers from the higher loss but it works well with topk compression.

Experiments

MICHIGAN STATE
UNIVERSITY

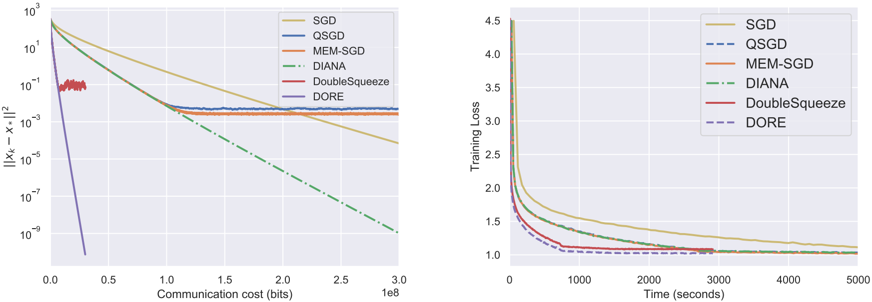

Figure 5: LHS: Linear regression on synthetic data (error vs communication cost); RHS: ResNet18 on CIFAR10 under 200 Mbps bandwidth

Further Topics

- Quantization algorithm for decentralized optimization
- Stochastic Modified Equation (SME) to study the dynamics of SGD

Reference

MICHIGAN STATE UNIVERSITY

Yao Li and Ming Yan. "On linear convergence of two decentralized algorithms". In: *arXiv preprint arXiv:1906.07225* (2019).

Zhi Li, Wei Shi, and Ming Yan. "A decentralized proximal-gradient method with network independent step-sizes and separated convergence rates". In: *IEEE Transactions on Signal Processing* 67.17 (2019), pp. 4494–4506.

Xiaorui Liu et al. "A Double Residual Compression Algorithm for Efficient Distributed Learning". In: *arXiv preprint arXiv:1910.07561* (2019).

Reference

MICHIGAN STATE
UNIVERSITY

Wei Shi et al. "Extra: An exact first-order algorithm for decentralized consensus optimization". In: *SIAM Journal on Optimization* 25.2 (2015), pp. 944–966.

Lin Xiao and Stephen Boyd. "Fast linear iterations for distributed averaging". In: *Systems & Control Letters* 53.1 (2004), pp. 65–78.

Kun Yuan, Qing Ling, and Wotao Yin. "On the convergence of decentralized gradient descent". In: *SIAM Journal on Optimization* 26.3 (2016), pp. 1835–1854.

Thank You !