A Stochastic Approximation Framework for Communication Efficient Decentralized Optimization

Hoi-To Wai

Department of Systems Engineering & Engineering Management,
The Chinese University of Hong Kong, Hong Kong SAR of China.

Joint work with Haoming Liu (PKU), Chung-yiu Yau (UMN)

Seminar at NUS @ November, 2025

Acknowledgement: #MMT-p5-23 of SHIAE, CUHK.

Table of Contents

Introduction

Consider the distributed optimization problem over n agents:

$$\min_{\mathbf{x}\in\mathbb{R}^d}\left[F(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x})\right], \quad f_i(\mathbf{x}) := \mathbb{E}_{\xi_i\sim\mathcal{D}_i}[\ell(\mathbf{x};\xi_i)] \tag{1}$$

where $f_i : \mathbb{R}^d \to \mathbb{R}$ is the local objective function satisfying:

- $f_i(\mathbf{x})$ is continuously differentiable (possibly non-convex).
- $f_i(\mathbf{x}) > -\infty, \ \forall \mathbf{x} \in \mathbb{R}^d$.
- $\mathcal{D}_i$ is the distribution of local data at agent i — If $\mathcal{D}_i \approx \mathcal{D}$ for all $i \Rightarrow$ **homogeneous data**, see [Li and Wai, 2025]. Otherwise, **heterogeneous data**.

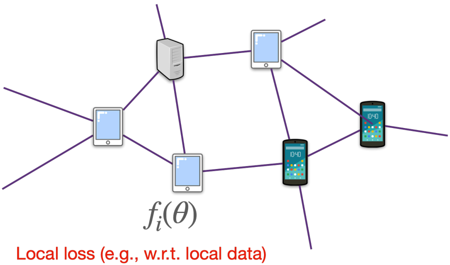
Applications:

- Large-scale machine learning (ML): [Lian et al., 2017], [Li et al., 2020], [Yuan et al., 2022].
- Signal processing (SP) on wireless sensor networks: [Mateos et al., 2010], [Dimakis et al., 2010], [Xiao et al., 2007].
- and many others...

Challenges: Distributed Training on Unreliable Networks

Setup:

- Devices or agents communicate on $G = (V, E)$.
- Each agent can only access local objective function $f_i(x)$ and its stochastic gradient.
- Communication network can be *unreliable*, e.g., for IoT devices on wireless networks.



Unreliable Networks can be subject to

(a) Link failures. (b) Bandwidth constraints. (c) Communication Noise.

$\Rightarrow$ Requires 'robust' decentralized optimization algorithms that can handle

(a) Time-varying graph. (b) Communication compression. (c) Noisy communication.

Prior Works

Decentralized methods

- Primal-only algorithms — 'ad-hocly' mimicking GD

$$\mathbf{x}_i^{t+1} = \mathbf{x}_i^t - \eta \boldsymbol{g}^t \quad \text{with} \quad \boldsymbol{g}^t \approx \nabla F((1/n)\textstyle\sum_{i=1}^n \mathbf{x}_i^t)$$

  e.g., Decentralized Gradient (DGD) [Nedic and Ozdaglar, 2009], Gradient Tracking (GT) [Qu and Li, 2017], EXTRA [Shi et al., 2015], DIGing [Nedic et al., 2017], etc.
- Primal-dual algorithms — treating (1) as constrained problem and solve its dual, e.g., Prox-PDA [Hong et al., 2017], GPDA [Yi et al., 2021], etc.

Decentralized methods with bandwidth limitation

- Compression with error-feedback:
  - deterministic gradients – [Reisizadeh et al., 2019, Magnússon et al., 2020, Liu and Li, 2021, Zhao et al., 2022].
  - stochastic gradients – [Koloskova et al., 2019a,c, Yau and Wai, 2023, Xie et al., 2024].
- Reduce comm. frequency: [Stich, 2018, Yu et al., 2019, Basu et al., 2020].
- Adaptive finite-bit quantizer: [Michelusi et al., 2022, Nassif et al., 2024].

Prior Works & This Talk

Prior Works	No Bdd Hete.	TV	Comp.	Noisy Comm.	Rate (Noiseless)	Rate (Noisy)
DSGD [Koloskova et al., 2020]	✗	✓	✗	✗	$\mathcal{O}(\bar{\sigma}/\sqrt{nT})$	/
GT [Liu et al., 2024]	✓	✗	✗	✗	$\mathcal{O}(\bar{\sigma}/\sqrt{nT})$	/
CHOCO-SGD [Koloskova et al., 2019a]	✗	✗	✓	✗	$\mathcal{O}(\bar{\sigma}/\sqrt{nT})$	/
DIMIX [Reisizadeh et al., 2023]	✗	✓	✓	✓	/	$\mathcal{O}(T^{-1/3}\ln T)$
CP-SGD [Xie et al., 2024]	✓	✗	✓	✗	$\mathcal{O}(\bar{\sigma}/\sqrt{nT})$	/
Decen-Scaffnew [Mishchenko et al., 2022]	✓	✗	✗	✗	$\mathcal{O}(\bar{\sigma}/\sqrt{nT})$	/
FSPDA-STORM (This Talk)	✓	✓	✓[†]	✗	$\mathcal{O}(\sigma^{2/3}/T^{2/3})$	/
TiCoPD (This Talk)	✓	✓	✓	✓	$\mathcal{O}(\bar{\sigma}/\sqrt{nT})$	$\mathcal{O}(n^{-1/2}T^{-1/3})$

Highlights of This Talk:

- Stochastic Approximation (SA) for decentralized optimization over *time-varying graphs* $\Rightarrow$ **Fully Stochastic Primal-dual Algorithm** (FSPDA) framework.
- **FSPDA-STORM** with variance reduction [Cutkosky and Orabona, 2019] — achieves $\mathcal{O}(1/T^{2/3})$ convergence rate to stationarity on time varying graphs.
- **Two-timescale compressed primal-dual** (TiCoPD) inspired by nonlinear gossip [Mathkar and Borkar, 2016] — with support for *compressed communication*, *noisy communication*, while being robust to *link failures*.

Setup and Notations

- The undirected connected graph $\mathcal{G} = (V, E)$ describes the communication network between the n agents
- The edges can be encoded via the incidence matrix $\mathbf{A} \in \{0, \pm 1\}^{|E| \times n}$:

$$\mathbf{A} = \begin{bmatrix} \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 1 & \cdots & -1 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$

$$A_{ki} = 1, A_{kj} = -1, \Rightarrow k = (i, j) \in E. \; A_{kl} = 0, \forall l \neq i, j.$$

- Constraint $\mathbf{x}_i = \mathbf{x}_j, \; \forall (i, j) \in E$ can be replaced by $\bar{\mathbf{A}}\mathbf{X} = \mathbf{0}$, $\mathbf{X} = [\mathbf{x}_1; \cdots \mathbf{x}_n]$, with $\bar{\mathbf{A}} = \mathbf{A} \otimes \mathbf{I}_d$.

Observation. problem (1) is equivalent to its constrained form:

$$\min_{\mathbf{x} \in \mathbb{R}^d} F(\mathbf{x}) := \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{x}) \Longleftrightarrow \min_{\mathbf{X} \in \mathbb{R}^{nd}} \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{x}_i) \;\; \text{s.t.} \;\; \bar{\mathbf{A}}\mathbf{X} = \mathbf{0} \qquad (2)$$

Time Varying Graph Model

Constrained problem (2) assumes a fixed graph $\mathcal{G}$ specified by the incidence matrix $\bar{\mathbf{A}}$.

$$\min_{\mathbf{X}\in\mathbb{R}^{nd}} \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}_i) \quad \text{s.t.} \quad \bar{\mathbf{A}}\mathbf{X} = \mathbf{0}$$

- To model *time varying (particularly, random) graphs*, replace $\bar{\mathbf{A}}\mathbf{X} = \mathbf{0}$ by **stochastic equality constraint** $\mathbb{E}[\bar{\mathbf{A}}(\xi_a)]\mathbf{X} = \mathbf{0}$, where $\bar{\mathbf{A}}(\xi_a)$ is the incidence matrix of the graph under realization ξ_a,

$$\bar{\mathbf{A}}(\xi_a) = \mathbf{I}(\xi_a)\bar{\mathbf{A}} \quad \text{with} \quad \mathbf{I}(\xi_a) \in \{0,1\}^{|E|\times|E|} \text{ is a diagonal matrix.}$$

- Graph induced by $\bar{\mathbf{A}}(\xi_a)$ is a random subgraph of $\mathcal{G}$.
- Model *random link failures* in the communication network and it can introduce random sparsified compression.

FSPDA Framework

Key step: with the new consensus constraint $\mathbb{E}[\bar{\mathbf{A}}(\xi_a)]\mathbf{X} = \mathbf{0}$, we consider the sampled augmented Lagrangian function with $\rho > 0$,

$$\mathcal{L}(\mathbf{X}, \boldsymbol{\lambda}; \xi) := \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}_i; \xi_i) + \boldsymbol{\lambda}^\top \bar{\mathbf{A}}(\xi_a)\mathbf{X} + \frac{\rho}{2}\|\bar{\mathbf{A}}(\xi_a)\mathbf{X}\|^2,$$

where $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_i)_{i \in E} \in \mathbb{R}^{|E|d}$ is the Lagrange multiplier.

Fact: Saddle points of $\mathbb{E}_\xi[\mathcal{L}(\mathbf{X}, \boldsymbol{\lambda}; \xi)]$ correspond to the **stationary points of** (2).

$\Longrightarrow$ **FSPDA Framework:** Primal-dual stochastic approximation (SA) leveraging the stochastic equality constraint.

FSPDA with Plain SA

- Applying primal-dual stochastic gradient descent-ascent on $\mathcal{L}(\mathbf{X}, \boldsymbol{\lambda}; \xi)$:

FSPDA-SA: at iteration k, we draw ξ_a^{k+1} to determine the graph realization and ξ_i^{k+1} to compute stochastic gradient at agent i,

$$\begin{cases} \mathbf{x}_i^{k+1} = \mathbf{x}_i^k - \alpha(\nabla f_i(\mathbf{x}_i^k; \xi_i^{k+1}) + \tilde{\lambda}_i^k + \rho \sum_{j \in \mathcal{N}_i(\xi_a^{k+1})} (\mathbf{x}_j^k - \mathbf{x}_i^k)) \\ \tilde{\lambda}_i^{k+1} = \tilde{\lambda}_i^k + \eta \sum_{j \in \mathcal{N}_i(\xi_a^{k+1})} (\mathbf{x}_j^k - \mathbf{x}_i^k) \end{cases}$$

- $\Longrightarrow$ an SA scheme for finding the saddle point of $\mathbb{E}_\xi[\mathcal{L}(\mathbf{X}, \boldsymbol{\lambda}; \xi)]$.
- At iteration k, it requires agent i to acquire $\{\mathbf{x}_j^k : j \in \mathcal{N}_i(\xi_a^{k+1})\}$ from its neighbors in the sampled graph.
- **Main Feature:** handle *time-varying graphs* and *stochastic gradients* in a unified & fully stochastic setting.

FSPDA-STORM Algorithm

- Incorporate the STORM estimator [Cutkosky and Orabona, 2019] to estimate $\nabla_{\mathbf{x}_i}\mathbb{E}_\xi[\mathcal{L}(\mathbf{X}^t,\boldsymbol{\lambda}^t;\xi)] \approx \mathbf{m}_x^t$ and $\nabla_{\boldsymbol{\lambda}}\mathbb{E}_\xi[\mathcal{L}(\mathbf{X}^t,\boldsymbol{\lambda}^t;\xi)] \approx \mathbf{m}_\lambda^t$ by

$$\begin{aligned} \mathbf{m}_x^{t+1} &= \nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}^{t+1},\boldsymbol{\lambda}^{t+1};\xi^{t+1}) + (1-a_x)(\mathbf{m}_x^t - \nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}^t,\boldsymbol{\lambda}^t;\xi^{t+1})), \\ \mathbf{m}_\lambda^{t+1} &= \nabla_{\boldsymbol{\lambda}}\mathcal{L}(\mathbf{x}^{t+1},\boldsymbol{\lambda}^{t+1};\xi^{t+1}) + (1-a_\lambda)(\mathbf{m}_\lambda^t - \nabla_{\boldsymbol{\lambda}}\mathcal{L}(\mathbf{x}^t,\boldsymbol{\lambda}^t;\xi^{t+1})). \end{aligned} \tag{3}$$

- **Insight**: the zero-mean control-variate term $(\mathbf{m}_x^t - \nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}^t,\boldsymbol{\lambda}^t;\xi^{t+1}))$ reduces the variance of the stochastic gradient estimator.

FSPDA-STORM: at iteration k, we draw ξ_a^{k+1} and ξ_i^{k+1}. At agent i,

$$\mathbf{x}_i^{k+1} = \mathbf{x}_i^k - \alpha\,\mathbf{m}_{x,i}^t, \quad \tilde{\lambda}_i^{k+1} = \tilde{\lambda}_i^k + \eta\,\mathbf{m}_{\lambda,i}^t$$

- Similar communication requirement as FSPDA-SA.
- The first decentralized algorithm for time varying graph to incorporate variance reduction (with provable acceleration - to be shown next).

Assumptions

A1 - Lipschitz Continuous Gradient

Each f_i is L-smooth, i.e., $\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\| \ \forall\ \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.

(A1') For FSPDA-STORM, we also need each f_i to be mean-square smooth, i.e., $\mathbb{E}_\xi \|\nabla f_i(\mathbf{x}; \xi) - \nabla f_i(\mathbf{y}; \xi)\| \leq L\|\mathbf{x} - \mathbf{y}\| \ \forall\ \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.

- smooth but possibly non-convex objective function.

A2 - Graph Spectrum

Let $\mathbf{K} := (\mathbf{I}_n - \mathbf{1}\mathbf{1}^\top/n) \otimes \mathbf{I}_d$, there exists $\rho_{\max} \geq \rho_{\min} > 0$, $\tilde{\rho}_{\max} \geq \tilde{\rho}_{\min} > 0$ such that $\rho_{\min}\mathbf{K} \preceq \bar{\mathbf{A}}^\top \mathbf{R} \bar{\mathbf{A}} \preceq \rho_{\max}\mathbf{K}$, $\tilde{\rho}_{\min}\mathbf{K} \preceq \bar{\mathbf{A}}^\top \bar{\mathbf{A}} \preceq \tilde{\rho}_{\max}\mathbf{K}$.

- captures spectral gap of the Laplacian, it holds $\mathbf{A}^\top \mathbf{A} - \tilde{\rho}_{\min}\mathbf{K} \succeq 0$ if G is connected.
- further, if $\mathrm{diag}(\mathbf{R}) > \mathbf{0}$ (each edge is selected with > 0 prob.), then $\mathbf{A}^\top \mathbf{R} \mathbf{A} - \rho_{\min}\mathbf{K} \succeq 0$.

Assumptions (cont'd)

A3 - Stochastic Gradient

For any $i \in [n]$ and fixed $\mathbf{y} \in \mathbb{R}^d$, there exists $\sigma_i \geq 0$, $\mathbb{E}[\|\nabla f_i(\mathbf{y}; \xi_i) - \nabla f_i(\mathbf{y})\|^2] \leq \sigma_i^2$.

- For simplicity, define $\bar{\sigma}^2 = 1/n \sum_{i=1}^{n} \sigma_i^2$.

A4 - Random Graph Variance

For any fixed $\mathbf{X} \in \mathbb{R}^{nd}$, $\mathbb{E}\left[\|\bar{\mathbf{A}}^\top \bar{\mathbf{A}}(\xi_a)\mathbf{X} - \bar{\mathbf{A}}^\top \mathbf{R}\bar{\mathbf{A}}\mathbf{X}\|^2\right] \leq \sigma_A^2 \|\bar{\mathbf{A}}\mathbf{X}\|_{\mathbf{R}}^2$.

- the upper bound holds as long as each edge is selected with a positive probability, in particular, $\sigma_A^2 = \mathbb{E}\left[\|\bar{\mathbf{A}}^\top(\xi_a)\mathbf{R}^{-1} - \bar{\mathbf{A}}^\top\|^2 \|\mathbf{R}\|\right]$.

Convergence of FSPDA-STORM

Theorem. Under A0, A1', A2, A3, A4, there exists step sizes and parameters α, η, γ such that for any $T \geq 1$, it holds

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}}^t)\|^2\right] \lesssim \frac{\mathbb{E}\left[F_0\right] - f^\star}{T\alpha} + \frac{a_x^2 \bar{\sigma}}{\alpha}$$

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\|\mathbf{X}^t\|_{\mathbf{K}}^2\right] \lesssim \frac{\mathbb{E}\left[F_0\right] - f^\star}{T\mathsf{a}\gamma} + \frac{a_x^2 \bar{\sigma}}{\mathsf{a}\gamma}$$

where a is a free quantity.

- Setting $\alpha = \mathcal{O}(\bar{\sigma}^{-2/3}T^{-1/3})$, $\eta = \mathcal{O}(n)$, $\gamma = \mathcal{O}(T^{-1/3})$, $\beta = \mathcal{O}(n^{-1}T^{-2/3})$, $a_x = \mathcal{O}(\bar{\sigma}^{-4/3}T^{-2/3})$, $a_\lambda = \mathcal{O}(T^{-1/3})$ (+ other constants)

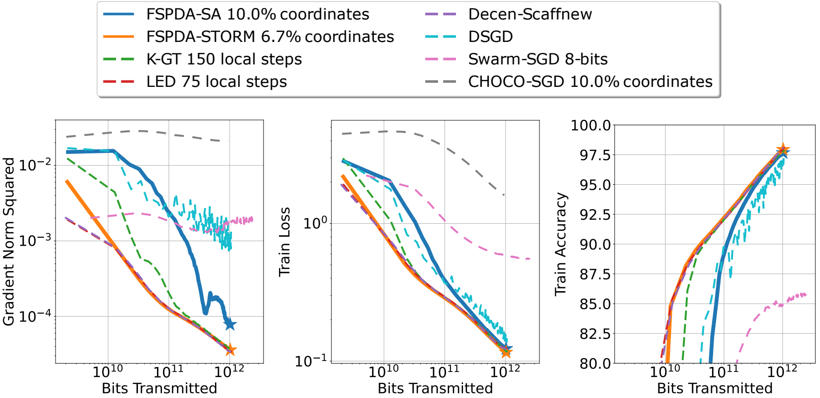
$$\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}^{\mathsf{R}})\|^2] = \boxed{\mathcal{O}\left(\bar{\sigma}^{2/3}/T^{2/3}\right)} \quad \text{with} \quad \mathsf{R} \sim \mathcal{U}\{1, \ldots, T\}.$$

- First to achieve $\mathcal{O}(1/T^{2/3})$ rate for decentralized stochastic non-convex opt. over time-varying graphs. (**Open Problem:** linear speedup?)
- **Extension:** Last-iterate convergence under PL condition, FSPDA-SA/STORM can also support *asynchronous updates*, etc.

Experiment Setup

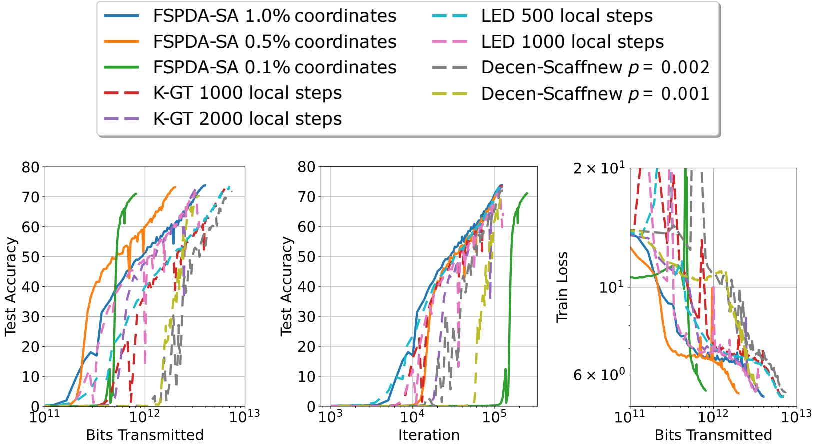
- Comparison to algorithms focusing on *'optimal' communication complexity*.
    - Decen-Scaffnew [Mishchenko et al., 2022].
    - LED [Alghunaim, 2024].
    - K-GT [Liu et al., 2024].
- Algorithms are implemented on PyTorch with MPI interface to simulate the distributed setting & the hyper-params are hand-tuned to achieve their respective best performance.
- Servers with Intel Xeon Gold 6148 CPU (for regression problem) and 8 $\times$ NVIDIA RTX 3090 GPU (for NN training problems).

Numerical Experiments: Feedforward NN on MNIST



- Decen-Scaffnew, K-GT, LED achieve 'optimal communication complexity' by communicating on a fixed graph every R iterations.
- FSPDA-STORM achieves similar performance to them while communicating on a random, time-varying graph at every iteration.

Numerical Experiments: Resnet-50 Training on CIFAR-10



- FSPDA-SA also outperforms baselines for large-scale problems[1].

[1]For Resnet-50 training, we found that FSPDA-STORM does not perform as well as FSPDA-SA. It is suspected that the mean-square smoothness property does not hold well in this setting.

TiCoPD algorithm – Development

Question: can we incorporate *compressed communication* and *noisy communication* into the primal-dual framework?

For simplicity, consider the case with deterministic graph and objective functions. The augmented Lagrangian function for (2) is

$$\mathcal{L}(\mathbf{X}, \boldsymbol{\lambda}) := \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{X}_i) + \boldsymbol{\lambda}^\top \bar{\mathbf{A}} \mathbf{X} + \frac{\rho}{2} \|\bar{\mathbf{A}} \mathbf{X}\|^2,$$

Issue: Apply primal-dual gradient descent-ascent ⇒ GPDA [Yi et al., 2021], but

$$\nabla_{\mathbf{x}_i} [(1/2) \|\bar{\mathbf{A}} \mathbf{X}\|^2] = \textstyle\sum_{j \in \mathcal{N}_i} (\mathbf{x}_j - \mathbf{x}_i) \leftarrow \text{require tx. of } d\text{-dim. vectors}$$

We develop **TiCoPD** to reduce bandwidth and communication overhead by:

- **Step 1**: A majorization-minimization step on $\mathcal{L}(\mathbf{X}, \lambda)$ that introduces a surrogate variable to separate the communication step from the optimization step.
- **Step 2**: A two-timescale update that incorporates the nonlinearly compressed update of the surrogate variable.

TiCoPD algorithm – Development

Question: can we incorporate *compressed communication* and *noisy communication* into the primal-dual framework?

For simplicity, consider the case with deterministic graph and objective functions. The augmented Lagrangian function for (2) is

$$\mathcal{L}(\mathbf{X}, \boldsymbol{\lambda}) := \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{X}_i) + \boldsymbol{\lambda}^\top \bar{\mathbf{A}}\mathbf{X} + \frac{\rho}{2}\|\bar{\mathbf{A}}\mathbf{X}\|^2,$$

Issue: Apply primal-dual gradient descent-ascent ⇒ GPDA [Yi et al., 2021], but

$$\nabla_{\mathbf{x}_i}[(1/2)\|\bar{\mathbf{A}}\mathbf{X}\|^2] = \textstyle\sum_{j \in \mathcal{N}_i}(\mathbf{x}_j - \mathbf{x}_i) \leftarrow \text{require tx. of } d\text{-dim. vectors}$$

We develop **TiCoPD** to reduce bandwidth and communication overhead by:

- **Step 1**: A majorization-minimization step on $\mathcal{L}(\mathbf{X}, \lambda)$ that introduces a surrogate variable to separate the communication step from the optimization step.
- **Step 2**: A two-timescale update that incorporates the nonlinearly compressed update of the surrogate variable.

Step 1: Majorization-Minimization

- Introducing a *surrogate variable* $\{\hat{\mathbf{X}}^t\}_{t\geq 0}$ that approximates $\hat{\mathbf{X}}^t \approx \mathbf{X}^t$.

  (to be discussed later) agent i shall acquire the neighbors' surrogate variables $(\hat{\mathbf{X}}_j^t)_{j\in\mathcal{N}_i^t}$ with **compressed communication**; see step 2.
- Consider the **majorization** anchored on $\hat{\mathbf{X}}$ – set $M \geq \|\bar{\mathbf{A}}^\top \bar{\mathbf{A}}\|_2$,

$$\|\bar{\mathbf{A}}\mathbf{X}\|^2 \leq \|\bar{\mathbf{A}}\hat{\mathbf{X}}^t\|^2 + 2(\mathbf{X} - \hat{\mathbf{X}}^t)^\top \bar{\mathbf{A}}^\top \bar{\mathbf{A}}\hat{\mathbf{X}}^t + M\|\mathbf{X} - \hat{\mathbf{X}}^t\|^2,$$

- The $\mathbf{X}$-update can be computed using the following **minimization** step:

$$\begin{aligned}\mathbf{X}^{t+1} &= \arg\min_{\mathbf{X}\in\mathbb{R}^{nd}} \nabla\mathbf{f}(\mathbf{X}^t)^\top(\mathbf{X} - \mathbf{X}^t) + \mathbf{X}^\top\bar{\mathbf{A}}^\top\boldsymbol{\lambda}^t + \frac{\theta}{2}\|\bar{\mathbf{A}}\hat{\mathbf{X}}^t\|^2 \\ &\qquad + \theta\mathbf{X}^\top\bar{\mathbf{A}}^\top\bar{\mathbf{A}}\hat{\mathbf{X}}^t + \frac{\theta M}{2}\|\mathbf{X} - \hat{\mathbf{X}}^t\|^2 + \frac{1}{2\widetilde{\alpha}}\|\mathbf{X} - \mathbf{X}^t\|^2 \\ &= \beta\mathbf{X}^t + (1-\beta)\hat{\mathbf{X}}^t - \alpha\big(\nabla\mathbf{f}(\mathbf{X}^t) + \bar{\mathbf{A}}^\top\boldsymbol{\lambda}^t + \theta\bar{\mathbf{A}}^\top\bar{\mathbf{A}}\hat{\mathbf{X}}^t\big),\end{aligned}$$

where $\nabla\mathbf{f}(\mathbf{X}^t) = [\nabla f_1(\mathbf{X}_1^t)^\top \cdots \nabla f_n(\mathbf{X}_n^t)^\top]^\top$, $\widetilde{\alpha} = \frac{\alpha}{1-\alpha\theta M}$, and $\beta = \frac{\alpha}{\widetilde{\alpha}}$.

Step 2: Two-timescale Updates

- Update of $\hat{\mathbf{X}}^t$ needs to be communication efficient and satisfies $\hat{\mathbf{X}}^t \approx \mathbf{X}^t$.

A0 - Noisy & Contractive Compression Operator

The compression operator $Q : \mathbb{R}^{nd} \times \Omega_i \to \mathbb{R}^{nd}$ satisfies:

$$Q(\mathbf{x}; \xi_q) = \hat{Q}(\mathbf{x}; \xi_q) + \mathbf{w}, \;\; \mathbf{w} \in \mathbb{R}^d \text{ is zero mean with variance } \sigma_\xi^2.$$

where $\hat{Q}(\mathbf{x}; \xi_q)$ is a **contractive** operator satisfying:

$$\mathbb{E}\left[\|\hat{Q}(\mathbf{x}; \xi_q) - \mathbf{x}\|^2 \right] \leq (1-\delta)^2 \|\mathbf{x}\|^2, \;\forall\, \mathbf{x} \in \mathbb{R}^{nd},$$

- **Example**: randomized quantization communicating bits over erroneous channels.
- Let $\gamma \in (0, 1]$. Through the lower level iteration

$$\hat{\mathbf{X}}^{t,k+1} = \hat{\mathbf{X}}^{t,k} + \gamma \boxed{Q(\mathbf{X}^t - \hat{\mathbf{X}}^{t,k}; \xi^{t,k+1})}, \;\forall\, k \geq 0,$$

  it holds $\hat{\mathbf{X}}^{t,k} \overset{k\to\infty}{\to} \mathbf{X}^t$, where k denotes the contraction iteration index.
- The iteration exhibits a fast convergence rate ($>$ MM updates) $\Rightarrow$ one step of the lower level update per PD iteration is enough.

TiCoPD Algorithm – Deterministic Version

- Replace $\mathbf{X}^t$ with the surrogate variable $\hat{\mathbf{X}}^t$ in $\boldsymbol{\lambda}$-subproblem, it holds:

$$\boldsymbol{\lambda}^{t+1} = \arg\min_{\boldsymbol{\lambda}\in\mathbb{R}^{|E|d}} \left\{ -\boldsymbol{\lambda}^\top \bar{\mathbf{A}}\hat{\mathbf{X}}^t + \frac{1}{2\eta}\|\boldsymbol{\lambda}-\boldsymbol{\lambda}^t\|^2 \right\} = \boldsymbol{\lambda}^t + \eta\bar{\mathbf{A}}\hat{\mathbf{X}}^t.$$

- Substituting $\widetilde{\boldsymbol{\lambda}}^t = \bar{\mathbf{A}}^\top\boldsymbol{\lambda}^t \in \mathbb{R}^{nd}$, we yield the TiCoPD algorithm:

$$\begin{cases} \mathbf{x}_i^{k+1} = \mathbf{x}_i^k - \alpha(\nabla f_i(\mathbf{x}_i^k) + \tilde{\lambda}_i^k + \rho\sum_{j\in\mathcal{N}_i}(\hat{\mathbf{x}}_j^k - \hat{\mathbf{x}}_i^k)) \\ \tilde{\lambda}_i^{k+1} = \tilde{\lambda}_i^k + \eta\sum_{j\in\mathcal{N}_i}(\hat{\mathbf{x}}_j^k - \hat{\mathbf{x}}_i^k) \end{cases}$$

$$\left\{ \hat{\mathbf{x}}_i^{k+1} = \hat{\mathbf{x}}_i^k + \gamma\, Q\left(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k, \xi_q^{k+1}\right) \right.$$

- *Bandwidth requirement*: agents only need to encode and transmit the **compressed** (and noisy) version of differences $\mathbf{X}^{t+1} - \hat{\mathbf{X}}^t$.

TiCoPD Algorithm – Stochastic Version

- Similar to FSPDA, working with a *sampled version* of the augmented Lagrangian function to derive:

$$
\begin{cases}
\mathbf{x}_i^{k+1} = \mathbf{x}_i^k - \alpha(\nabla \ell_i(\mathbf{x}_i^k; \xi_i^{k+1}) + \tilde{\lambda}_i^k + \rho \sum_{j \in \mathcal{N}_i(\xi_a)} (\hat{\mathbf{x}}_j^k - \hat{\mathbf{x}}_i^k)) \\
\tilde{\lambda}_i^{k+1} = \tilde{\lambda}_i^k + \eta \sum_{j \in \mathcal{N}_i(\xi_a)} (\hat{\mathbf{x}}_j^k - \hat{\mathbf{x}}_i^k)
\end{cases}
$$

$$
\left\{ \hat{\mathbf{x}}_i^{k+1} = \hat{\mathbf{x}}_i^k + \gamma \, Q \left(\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k, \xi_q^{k+1} \right) \right.
$$

- Set $\mathbf{R} = \mathbb{E}[\mathbf{I}(\xi_a)]$ as the matrix of expected edge selection probabilities.
- Naturally leads to a decentralized algorithm with support for **(a)** random graph, **(b)** compressed communication, and **(c)** noisy communication.

Convergence of TiCoPD

Theorem. Under A0-4, there exists step sizes and parameters α, η, γ such that for any $T \geq 1$, it holds

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(\bar{\mathbf{x}}^t)\|^2\right] \lesssim \frac{\mathbb{E}\left[F_0\right] - f^\star}{\alpha T} + \alpha\bar{\sigma}^2 + \frac{\gamma^2\sigma_\xi^2}{\alpha}$$

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\mathbf{X}^t\|_{\mathbf{K}}^2\right] \lesssim \frac{\mathbb{E}\left[F_0\right] - f^\star}{\alpha T\theta \mathsf{a}} + \frac{\alpha\bar{\sigma}^2}{\theta \mathsf{a}} + \frac{\gamma^2\sigma_\xi^2}{\alpha\theta \mathsf{a}}$$

where a is a free quantity.

- Do not require diminishing step size nor bounded gradient heterogeneity.
- **Noiseless Comm.** **($\sigma_\xi = 0$)**: $\alpha = \mathcal{O}(1/\sqrt{T})$, $\theta = \mathcal{O}(\sqrt{T})$, $\gamma = 1$, $\mathsf{a} = \mathcal{O}(1/\sqrt{T})$,

$$\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}^{\mathsf{R}})\|^2] = \mathcal{O}\left(\sqrt{\bar{\sigma}^2/(nT)}\right) \quad \text{with } \mathsf{R} \sim \mathcal{U}\{1,\ldots,T\}.$$

- **Noisy Comm.** **($\sigma_\xi > 0$)**: $\alpha = \mathcal{O}(T^{-\frac{2}{3}}), \theta = \mathcal{O}(T^{\frac{1}{3}}), \mathsf{a} = \mathcal{O}(T^{-\frac{1}{3}}), \gamma = \mathcal{O}(T^{-\frac{1}{3}})$,

$$\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}^{\mathsf{R}})\|^2] = \mathcal{O}\left((1+\sigma_\xi^2)/T^{\frac{1}{3}}\right) \quad \text{with } \mathsf{R} \sim \mathcal{U}\{1,\ldots,T\}.$$

Insights from Convergence Analysis (for TiCoPD)

- **Challenge**: Due to the time-varying nature of $\bar{\mathbf{A}}(\xi_a^t)$, we cannot directly use the Augmented Lagrangian function as Lyapunov function.

- Define $\mathbf{v}^t = \alpha\widetilde{\lambda}^t + \alpha\nabla\mathbf{f}((\mathbf{1}_n \otimes \mathbf{I}_d)\bar{\mathbf{x}}^t)$, serving as a 'gradient tracker'.
- Consider for some $\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d}, \mathtt{e} > 0$,

$$F_t = f(\bar{\mathbf{x}}^t) + \mathtt{a}\|\mathbf{X}^t\|^2_{\widetilde{\mathbf{K}}} + \mathtt{b}\|\mathbf{v}^t\|^2_{\widetilde{\mathbf{Q}}+\mathtt{c}\widetilde{\mathbf{K}}} + \mathtt{d}\left\langle \mathbf{X}^t \mid \mathbf{v}^t \right\rangle_{\widetilde{\mathbf{K}}} + \mathtt{e}\|\hat{\mathbf{X}}^t - \mathbf{X}^t\|^2.$$

- **Fact 1**: with appropriate a, b, c, d, e, it can be shown that

$$F_t \geq f(\bar{\mathbf{x}}^t) + \|\mathbf{v}^t\|^2_{(\mathtt{b}\cdot\tilde{\rho}_{\max}^{-1} + \mathtt{bc} - \frac{\mathtt{d}^2}{4\mathtt{a}})\mathbf{K}} + \mathtt{e}\|\hat{\mathbf{X}}^t - \mathbf{X}^t\|^2.$$

- **Fact 2**: with appropriate a, b, c, d, e, it can be shown that

$$\mathbb{E}F_{t+1} \leq \mathbb{E}F_t - \bar{\omega}_f\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^t)\right\|^2 + \alpha^2\bar{\omega}_\sigma\bar{\sigma}^2 - \bar{\omega}_x\mathbb{E}\|\mathbf{X}^t\|^2_{\mathbf{K}} + 8\mathtt{a}\gamma^2\sigma_\xi^2\frac{\rho_{\max}}{\rho_{\min}}, \quad (4)$$

with $\bar{\omega}_f, \bar{\omega}_x > 0$. Telescoping with (4) yields the convergence results.

Insights from Convergence Analysis (for TiCoPD)

- **Challenge**: Due to the time-varying nature of $\bar{\mathbf{A}}(\xi_a^t)$, we cannot directly use the Augmented Lagrangian function as Lyapunov function.
- Define $\mathbf{v}^t = \alpha\widetilde{\lambda}^t + \alpha\nabla\mathbf{f}((\mathbf{1}_n \otimes \mathbf{I}_d)\bar{\mathbf{x}}^t)$, serving as a 'gradient tracker'.
- Consider for some $\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d}, \mathtt{e} > 0$,

$$F_t = f(\bar{\mathbf{x}}^t) + \mathtt{a}\|\mathbf{X}^t\|_{\widetilde{\mathbf{K}}}^2 + \mathtt{b}\|\mathbf{v}^t\|_{\widetilde{\mathbf{Q}}+\mathtt{c}\widetilde{\mathbf{K}}}^2 + \mathtt{d}\left\langle \mathbf{X}^t \mid \mathbf{v}^t \right\rangle_{\widetilde{\mathbf{K}}} + \mathtt{e}\|\hat{\mathbf{X}}^t - \mathbf{X}^t\|^2.$$

- **Fact 1**: with appropriate $\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d}, \mathtt{e}$, it can be shown that

$$F_t \geq f(\bar{\mathbf{x}}^t) + \|\mathbf{v}^t\|^2_{(\mathtt{b}\cdot\tilde{\rho}_{\max}^{-1} + \mathtt{bc} - \frac{\mathtt{d}^2}{4\mathtt{a}})\mathbf{K}} + \mathtt{e}\|\hat{\mathbf{X}}^t - \mathbf{X}^t\|^2.$$

- **Fact 2**: with appropriate $\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d}, \mathtt{e}$, it can be shown that

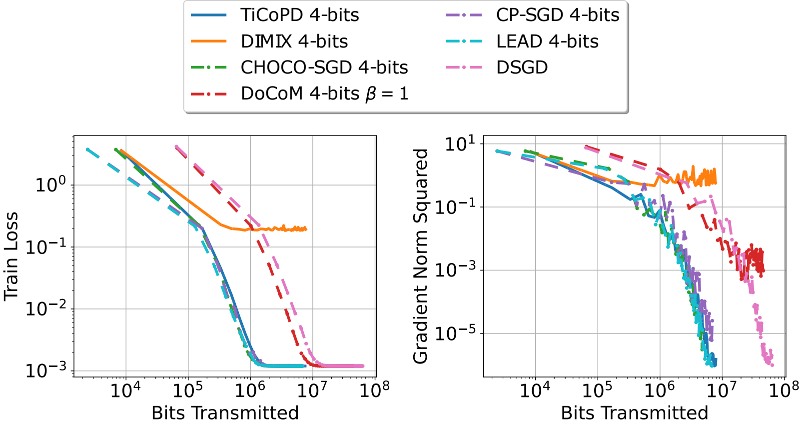
$$\begin{aligned} \mathbb{E}F_{t+1} \leq \mathbb{E}F_t - \bar{\omega}_f \mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^t)\right\|^2 + \alpha^2\bar{\omega}_\sigma\bar{\sigma}^2 \\ - \bar{\omega}_x\mathbb{E}\|\mathbf{X}^t\|_{\mathbf{K}}^2 + 8\mathtt{a}\gamma^2\sigma_\xi^2\frac{\rho_{\max}}{\rho_{\min}}, \end{aligned} \tag{4}$$

  with $\bar{\omega}_f, \bar{\omega}_x > 0$. Telescoping with (4) yields the convergence results.

Experiment Setup

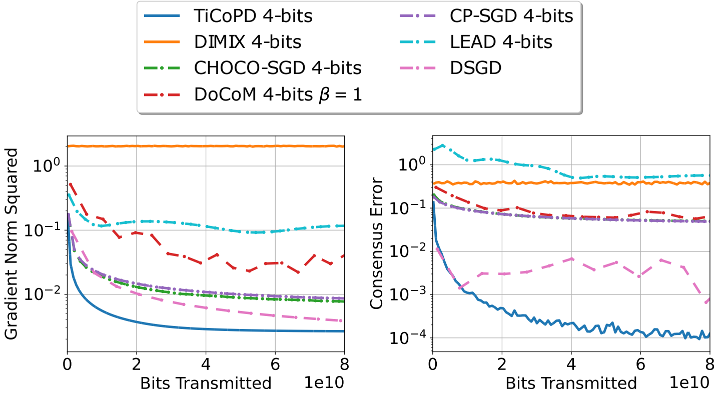
- Comparison to algorithms with communication compression via sparsification or quantization.
  - CHOCO-SGD Koloskova et al. [2019b].
  - CP-SGD Xie et al. [2024].
  - DIMIX Reisizadeh et al. [2023].
  - LEAD Liu and Li [2021].
  - DoCoM Yau and Wai [2023].
- Algorithms are implemented on PyTorch with MPI interface to simulate the distributed setting & the hyper-params are hand-tuned to achieve their respective best performance.
- Noisy communication and link failures are simulated by adding noise to received message and random edge selection, respectively.
- Servers with Intel Xeon Gold 6148 CPU (for regression problem) and 8 $\times$ NVIDIA RTX 3090 GPU (for NN training problems).

Experiment on Synthetic Data – Linear Regression



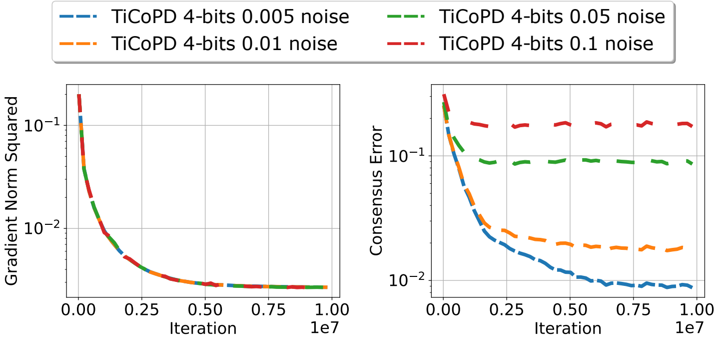
- Observe that TiCoPD outperforms most SOTA algorithms in terms of balancing between consensus error and training loss.

Experiment on Synthetic Data – Linear Model with Sigmoid Loss



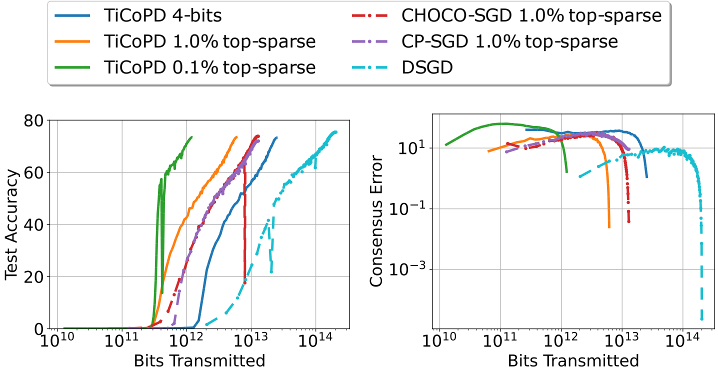
- For non-convex losses, algorithms such as LEAD [Liu and Li, 2021] fail to achieve on-par performance with TiCoPD.

Experiment on Synthetic Data – Linear Model with Sigmoid Loss



- Convergence of TiCoPD is robust to noise level in communication channels.

Experiment on ImageNet Data – Training ResNet-50



- TiCoPD achieves ~ 50 times saving in communication complexity over classical DSGD communicating on time varying graphs.

Conclusions

- **FSPDA** framework for decentralized optimization on time-varying graphs.
- **FSPDA-STORM** for fast decentralized optimization.
  - Combines **variance-reduction** with stochastic approximation $\Longrightarrow$ fast convergence.
- **TiCoPD** for decentralized optimization with support for compression.
  - Combines **majorization-minimization**, **two-time-scale iteration** methods for compressed distributied problem on the basis of primal dual algorithms.
- **Ongoing work:** strategic communication in decentralized optimization with compression and time-varying graphs.

Thank you. Comments are welcomed!

• C.-Y. Yau, H. Liu, **HT**, A Stochastic Approximation Approach for Efficient Decentralized Optimization on Random Networks, 2025 (under review). arxiv.org/abs/2410.18774

• H. Liu, C.-Y. Yau, **HT**, Decentralized Stochastic Optimization over Unreliable Networks via Two-timescales Updates, IEEE TSP, 2025 (accepted). arxiv.org/abs/2502.08964

References I

Sulaiman A Alghunaim. Local exact-diffusion for decentralized optimization and learning. *IEEE Transactions on Automatic Control*, 2024.

Debraj Basu, Deepesh Data, Can Karakus, and Suhas N Diggavi. Qsparse-local-sgd: Distributed sgd with quantization, sparsification, and local computations. *IEEE Journal on Selected Areas in Information Theory*, 1(1):217–226, 2020.

Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex sgd. *Advances in neural information processing systems*, 32, 2019.

Alexandros G Dimakis, Soummya Kar, José MF Moura, Michael G Rabbat, and Anna Scaglione. Gossip algorithms for distributed signal processing. *Proceedings of the IEEE*, 98(11):1847–1864, 2010.

Mingyi Hong, Davood Hajinezhad, and Ming-Min Zhao. Prox-pda: The proximal primal-dual algorithm for fast distributed nonconvex optimization and learning over networks. In *International Conference on Machine Learning*, pages 1529–1538. PMLR, 2017.

Anastasia Koloskova, Sebastian Stich, and Martin Jaggi. Decentralized stochastic optimization and gossip algorithms with compressed communication. In *International Conference on Machine Learning*, pages 3478–3487. PMLR, 2019a.

Anastasia Koloskova, Sebastian U. Stich, and Martin Jaggi. Decentralized stochastic optimization and gossip algorithms with compressed communication. In *ICML 2019 - Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 3479–3487. PMLR, 2019b. URL http://proceedings.mlr.press/v97/koloskova19a.html.

Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian Stich. A unified theory of decentralized sgd with changing topology and local updates. In *International conference on machine learning*, pages 5381–5393. PMLR, 2020.

Anastasiia Koloskova, Tao Lin, Sebastian Urban Stich, and Martin Jaggi. Decentralized deep learning with arbitrary communication compression. In *Proceedings of the 8th International Conference on Learning Representations*, 2019c.

Qiang Li and Hoi-To Wai. Tighter analysis for decentralized stochastic gradient method: Impact of data homogeneity. *IEEE Transactions on Automatic Control*, 2025.

Xiuxian Li, Xinlei Yi, and Lihua Xie. Distributed online optimization for multi-agent networks with coupled inequality constraints. *IEEE Transactions on Automatic Control*, 66(8):3575–3591, 2020.

Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. *Advances in neural information processing systems*, 30, 2017.

References II

Xiaorui Liu and Yao Li. Linear convergent decentralized optimization with compression. In *International Conference on Learning Representations*, 2021.

Yue Liu, Tao Lin, Anastasia Koloskova, and Sebastian U Stich. Decentralized gradient tracking with local steps. *Optimization Methods and Software*, pages 1–28, 2024.

Sindri Magnússon, Hossein Shokri-Ghadikolaei, and Na Li. On maintaining linear convergence of distributed learning and optimization under limited communication. *IEEE Transactions on Signal Processing*, 68:6101–6116, 2020.

Gonzalo Mateos, Juan Andrés Bazerque, and Georgios B Giannakis. Distributed sparse linear regression. *IEEE Transactions on Signal Processing*, 58(10):5262–5276, 2010.

Adwaitvedant S Mathkar and Vivek S Borkar. Nonlinear gossip. *SIAM Journal on Control and Optimization*, 54(3): 1535–1557, 2016.

Nicolò Michelusi, Gesualdo Scutari, and Chang-Shen Lee. Finite-bit quantization for distributed algorithms with linear convergence. *IEEE Transactions on Information Theory*, 68(11):7254–7280, 2022.

Konstantin Mishchenko, Grigory Malinovsky, Sebastian Stich, and Peter Richtárik. Proxskip: Yes! local gradient steps provably lead to communication acceleration! finally! In *International Conference on Machine Learning*, pages 15750–15769. PMLR, 2022.

Roula Nassif, Stefan Vlaski, Marco Carpentiero, Vincenzo Matta, and Ali H Sayed. Differential error feedback for communication-efficient decentralized optimization. In *2024 IEEE 13rd Sensor Array and Multichannel Signal Processing Workshop (SAM)*, pages 1–5. IEEE, 2024.

Angelia Nedic and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009.

Angelia Nedic, Alex Olshevsky, and Wei Shi. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27(4):2597–2633, 2017.

Guannan Qu and Na Li. Harnessing smoothness to accelerate distributed optimization. *IEEE Transactions on Control of Network Systems*, 5(3):1245–1260, 2017.

Amirhossein Reisizadeh, Aryan Mokhtari, Hamed Hassani, and Ramtin Pedarsani. An exact quantized decentralized gradient descent algorithm. *IEEE Transactions on Signal Processing*, 67(19):4934–4947, 2019.

Hadi Reisizadeh, Behrouz Touri, and Soheil Mohajer. Dimix: Diminishing mixing for sloppy agents. *SIAM Journal on Optimization*, 33(2):978–1005, 2023.

References III

Wei Shi, Qing Ling, Gang Wu, and Wotao Yin. Extra: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2):944–966, 2015.

Sebastian U Stich. Local sgd converges fast and communicates little. *arXiv preprint arXiv:1805.09767*, 2018.

Lin Xiao, Stephen Boyd, and Seung-Jean Kim. Distributed average consensus with least-mean-square deviation. *Journal of parallel and distributed computing*, 67(1):33–46, 2007.

Antai Xie, Xinlei Yi, Xiaofan Wang, Ming Cao, and Xiaoqiang Ren. A communication-efficient stochastic gradient descent algorithm for distributed nonconvex optimization. *arXiv preprint arXiv:2403.01322*, 2024.

Chung-Yiu Yau and Hoi To Wai. Docom: Compressed decentralized optimization with near-optimal sample complexity. *Transactions on Machine Learning Research*, 2023.

Xinlei Yi, Shengjun Zhang, Tao Yang, Tianyou Chai, and Karl H Johansson. Linear convergence of first-and zeroth-order primal–dual algorithms for distributed nonconvex optimization. *IEEE Transactions on Automatic Control*, 67(8): 4194–4201, 2021.

Hao Yu, Sen Yang, and Shenghuo Zhu. Parallel restarted sgd with faster convergence and less communication: Demystifying why model averaging works for deep learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 5693–5700, 2019.

Binhang Yuan, Yongjun He, Jared Davis, Tianyi Zhang, Tri Dao, Beidi Chen, Percy S Liang, Christopher Re, and Ce Zhang. Decentralized training of foundation models in heterogeneous environments. *Advances in Neural Information Processing Systems*, 35:25464–25477, 2022.

Haoyu Zhao, Boyue Li, Zhize Li, Peter Richtárik, and Yuejie Chi. Beer: Fast $o(1/t)$ rate for decentralized nonconvex optimization with communication compression. *Advances in Neural Information Processing Systems*, 35:31653–31667, 2022.