Fully Stochastic Primal-Dual Algorithm (FSPDA) for Communication Efficient Decentralized Optimization

Chung-Yiu Yau, Haoming Liu, **Hoi-To Wai**

Department of Systems Engineering & Engineering Management, CUHK

October, 2024
Asilomar Conference

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

Distributed Optimization Problem

We are interested in tackling the optimization problem over a network of n agents/machines:

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

for differentiable and stochastic $f_i : \mathbb{R}^d \to \mathbb{R}$ such that

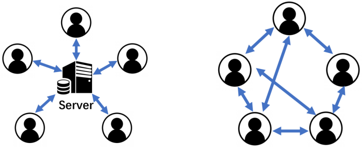
- $f_i(x) > -\infty \ \forall x \in \mathbb{R}^d$,
- $f_i(x) = \mathbb{E}_{\xi_i \sim \mathcal{D}_i}[f_i(x; \xi_i)]$ such that $\mathcal{D}_i$ is the distribution of local data,
- each agent/machine can only access its local objective function $f_i(x)$ (or $f_i(x; \xi_i)$) and its gradient.

Goal: Each agent i starts with a local iterate $x_i^0 \in \mathbb{R}^d$, finds a (**stationary**) solution $\bar{x} \in \mathbb{R}^d$ to (1) on every agent s.t. $x_1 = \cdots = x_n = \bar{x}$ (**consensus**).

Applications

- Estimation/control on wireless sensor network [Rabbat and Nowak, 2004]
  - Autonomous, self-organizing robots and drones
- Privacy-preserved machine learning
  - Data sharing is prohibited, e.g., medical records in hospitals [Warnat-Herresthal et al., 2021]
- Large-scale machine learning / deep learning
  - Accelerate training on extremely large dataset [Yuan et al., 2022], or community-based volunteer training [Ryabinin and Gusev, 2020]
  - *Decentralized* algorithm uses neighborhood (localized) communication, alleviating the communication bottleneck at server of centralized method.
  - High dimensional model (e.g., NN/LLMs with >10B parameters) necessitate **communication efficient algorithms**.

Decentralized Optimization - Setup and Notations



(a) Centralized (b) Decentralized (mutual trust)

Image credit [He et al., 2019]

- The n agents are connected via an undirected & **connected** graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$.
- The graph has a **(weighted) adjacency** $\mathbf{W} \in \mathbb{R}_+^{n \times n}$: $W_{ij} > 0$ iff $(i, j) \in \mathcal{E}$.
- The graph is endowed with an **incidence matrix** $\mathbf{A} \in \mathbb{R}^{|\mathcal{E}| \times n}$: for any $\mathbf{x} \in \mathbb{R}^n$,

$$\mathbf{A}\mathbf{x} = \mathbf{0} \Longleftrightarrow x_i = x_j, \ \forall\, i, j.$$

- The Laplacian matrix can be formed by $\mathbf{L} = \mathbf{A}^\top \mathbf{A}$ and it yields a **local difference operator** such that

$$[\mathbf{L}\mathbf{x}]_i = [\mathbf{A}^\top \mathbf{A}\mathbf{x}]_i = \sum_{j \in \mathcal{N}_i} (x_i - x_j)$$

Decentralized Optimization - Prior Arts

Primal-only methods — mimic GD/SGD,

- **Decentralized Gradient (DGD)** [Nedic and Ozdaglar, 2009]:

$$\mathbf{x}^{t+1} = \mathbf{W}\mathbf{x}^t - \gamma \nabla \mathbf{f}(\mathbf{x}^t)$$

- **Gradient Tracking (GT)** [Qu and Li, 2017]:

$$\mathbf{x}^{t+1} = \mathbf{W}\mathbf{x}^t - \gamma \mathbf{g}^t, \quad \mathbf{g}^{t+1} = \mathbf{W}\mathbf{g}^t + \nabla \mathbf{f}(\mathbf{x}^{t+1}) - \nabla \mathbf{f}(\mathbf{x}^t)$$

- Also see advanced implementation such as EXTRA [Shi et al., 2015], DIGing [Nedic et al., 2017], Directed EXTRA [Xi and Khan, 2017], analysis of DSGD in [Vlaski and Sayed, 2021] + many others

Primal-dual algorithms — take (1) as constrained opt.,

- Developed from **augmented Lagrangian** of (1) as a max-min problem

$$\max_{\lambda \in \mathbb{R}^{n \times d}} \min_{\mathbf{X} \in \mathbb{R}^{n \times d}} f(\mathbf{X}) + \beta \langle \lambda \mid \mathbf{A}\mathbf{X} \rangle + \frac{\alpha}{2} \|\mathbf{A}\mathbf{X}\|_F^2 \quad (2)$$

- Applying gradient-descent-ascent results in Prox-PDA [Hong et al., 2017], GPDA [Yi et al., 2021]; also [Mansoori and Wei, 2021, Boyd et al., 2011] + many others

Research Questions

- In practice, the graph $\mathcal{G}$ is not fixed as some links are not reliable at all times, have limited bandwidth, etc. $\Longrightarrow$ **time varying graph**
- **DGD** methods are amendable to time varying graphs, but it suffers from *slow convergence* and is prone to slow down due to *data heterogeneity*.
- **PDA** algorithms have *fast convergence*, but not for time varying graphs.
- **Wish:** a flexible algorithmic framework that enables *fast convergence* and *time varying* + *efficient communication*.

This Work[1]

- Propose a **Fully Stochastic Primal-Dual Algorithm (FSPDA)** framework based on primal-dual optimization.
- **Support random graphs** – naturally extended to have random coordinate selection for *sparsified communication*.

[1] [Yau et al., 2024] C.-Y. Yau, H. Liu, **H.-T.**, "Fully stochastic primal-dual gradient algorithm for non-convex optimization on random graphs", arXiv:2410.18774.

Fully Stochastic Primal-Dual Algorithm (FSPDA) – Development

$$\min_{x\in\mathbb{R}^d} F(x) \iff \min_{\mathbf{x}\in\mathbb{R}^{nd}} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\xi_i\sim\mathcal{D}_i}[f_i(x_i;\xi_i)] \text{ s.t. } \mathbb{E}[\tilde{\mathbf{A}}(\xi)]\mathbf{x} = \mathbf{0}$$

- **Key Idea 1**: the consensus constraint = **stochastic equality**.
- **Key Idea 2**: extended graph to control **coordinate-wise consensus**.

- Random graph encoded by a selection diagonal matrix $\mathbf{I}(\xi) \in \mathbb{R}^{Ed}$

$$\mathbf{I}(\xi)_{k,k} = \begin{cases} 1, & \text{if } e_k \in \mathcal{E}(\xi), \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

- With $\mathbb{E}[\mathbf{I}(\xi)] \succ \mathbf{0}$, the consensus constraint is equivalent to

$$\mathbf{A}\otimes\mathbf{I}\mathbf{x} = \mathbb{E}_{\xi}[\mathbf{I}(\xi)(\mathbf{A}\otimes\mathbf{I})\mathbf{x}] = \mathbf{0}$$

for $\mathbf{x} = [\mathbf{x}_1^\top \cdots \mathbf{x}_n^\top]^\top \in \mathbb{R}^{nd}$.

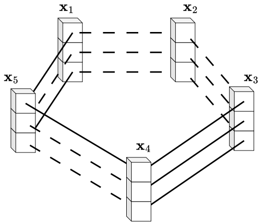


Figure: Extended graph $\mathcal{E}(\xi)$ of $\tilde{\mathbf{A}}$. Dashed line represents inactive edge.

FSPDA – Algorithm

- Set dual variable λ for $\mathbb{E}[\mathbf{A}(\xi)\mathbf{x}] = \mathbf{0}$, consider stochastic augmented Lagrangian:

$$\max_{\lambda \in \mathbb{R}^{Ed}} \min_{\mathbf{x} \in \mathbb{R}^{nd}} \mathbb{E}\left[f(\mathbf{x};\xi) + \eta \langle \lambda \mid \mathbf{A}(\xi)\mathbf{x} \rangle + \frac{\gamma}{2} \|\mathbf{A}(\xi)\mathbf{x}\|_F^2 \right]$$

- Set $\nabla \mathbf{f}(\mathbf{x};\xi) = [\nabla f_1(\mathbf{x}_1;\xi_1)^\top \cdots \nabla f_n(\mathbf{x}_n;\xi_n)^\top]^\top$, we yield a **fully stochastic algorithm** by applying SGDA to the above:

$$\mathbf{x}^{t+1} = \mathbf{x}^t - \alpha \nabla \mathbf{f}(\mathbf{x}^t;\xi^t) - \eta \widehat{\lambda}^t - \gamma \mathbf{A}^\top \mathbf{A}(\xi^t)\mathbf{x}^t \tag{4a}$$

$$\lambda^{t+1} = \lambda^t + \beta \mathbf{A}^\top \mathbf{A}(\xi^t)\mathbf{x}^t. \tag{4b}$$

- $\mathbf{A}^\top \mathbf{A}(\xi^t)\mathbf{x}^t$ consists of randomly sparsified communication on a random graph:

$$[\mathbf{A}^\top \mathbf{A}(\xi^t)\mathbf{x}^t]_i = \sum_{j \in \mathcal{N}_i(\xi^t)} \underbrace{\mathbf{I}_{ij}(\xi^t)}_{\in\{0,1\}^{d\times d},\ \text{random coordinate selection}} (\mathbf{x}_j^t - \mathbf{x}_i^t)$$

- FSPDA is naturally *decentralized* with *communication compression*.

Asynchronous FSPDA

- If we model the stochastic gradient as

  $$\nabla f_i(\mathbf{x}_i^t; \xi_i^t) = c_i(\xi_i^t)\,\bar{c}_i\,\nabla f_i(\mathbf{x}_i^t; \hat{\xi}_i^t),$$

  for the binary variable $c_i(\xi_i^t) \in \{0, 1\}$ satisfying $\mathbb{E}[c_i(\xi_i^t)] = 1/\bar{c}_i$, each agent may operate under the following **modes of operations**:
- **[Idle]** $\mathbf{x}_i^{t+1} = \mathbf{x}_i^t, \quad \widehat{\lambda}_i^{t+1} = \widehat{\lambda}_i^t$
  - After idling for τ iterations agent i will catch up by $\mathbf{x}_i^{t+\tau} = \mathbf{x}_i^t - \tau\eta\widehat{\lambda}_i^t$.
  - Idling increases the SG's variance.
- **[Local gradient steps]** $\mathbf{x}_i^{t+1} = \mathbf{x}_i^t - \alpha\nabla f_i(\mathbf{x}_i^t; \xi^t) - \eta\widehat{\lambda}_i^t, \quad \widehat{\lambda}_i^{t+1} = \widehat{\lambda}_i^t$.
  - Taking a local gradient step makes other agent idle

A1 - Lipschitz Continuous Gradient

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

A2 - Stochastic Gradient

For fixed $\mathbf{x}_i \in \mathbb{R}^d$, $\mathbb{E}_{\xi_i \sim \mathbb{P}_i}[\|\nabla f_i(\mathbf{x}_i; \xi_i) - \nabla f_i(\mathbf{x}_i)\|^2] \leq \sigma_i^2$. Set $\bar{\sigma}^2 := (1/n) \sum_{i=1}^n \sigma_i^2$.

A3 - Graph Spectrum

Let $\mathbf{K} := (\mathbf{I}_n - \mathbf{1}\mathbf{1}^\top/n) \otimes \mathbf{I}_d$ and $\mathbf{R} = \mathbb{E}[\mathbf{I}(\xi)]$, it holds

$$\rho_{\min}\mathbf{K} \preceq \mathbf{A}^\top\mathbf{R}\mathbf{A} \preceq \rho_{\max}\mathbf{K}, \quad \bar{\rho}_{\min}\mathbf{K} \preceq \mathbf{A}^\top\mathbf{A} \preceq \bar{\rho}_{\max}\mathbf{K}.$$

- A3 captures the spectral gap of the weighted Laplacian $\mathbf{A}^\top\mathbf{R}\mathbf{A}$ – satisfies $\rho_{\min} > 0$ if G is connected.

A4 - Random Graph Variance

For any $\mathbf{x} \in \mathbb{R}^{nd}$, it holds $\mathbb{E}_\xi[\|\mathbf{A}(\xi)^\top\mathbf{A}\mathbf{x} - \mathbf{A}^\top\mathbf{R}\mathbf{A}\mathbf{x}\|^2] \leq \sigma_A^2\|\mathbf{x}\|_{\mathbf{K}}^2$.

- The graph becomes more random as σ_A^2 increases.

Convergence of FSPDA

Theorem. Consider a free parameter $\mathsf{a} > 0$, suppose that

$$\gamma \leq \gamma_\infty := \frac{\rho_{\min}}{\rho_{\max}^2} \min\left\{1, \frac{\rho_{\max}}{2\sigma_A^2}\right\}, \; \eta \leq \eta_\infty = \mathcal{O}\left(\frac{\rho_{\min}^2}{\bar{\rho}_{\max}^2 \rho_{\max}^2} \gamma_\infty\right),$$

$$\alpha = \mathcal{O}\left(\frac{\gamma_\infty \rho_{\min}}{\sqrt{n}} \min\left\{\frac{\mathsf{a}}{L^2}, \eta_\infty \rho_{\min}, \sqrt{\frac{\eta_\infty \rho_{\min}}{L^2 \mathsf{a}}}\right\}\right),$$

Then

$$\begin{aligned} &\min_{t=0,\ldots,T-1} \mathbb{E}\left[\|\nabla F(\bar{\mathbf{x}}^t)\|^2\right] \leq \frac{F_0 - f_\star}{\alpha T/8} + 8\alpha \mathbb{C}_\sigma \frac{\bar{\sigma}^2}{n}, \\ &\min_{t=0,\ldots,T-1} \mathbb{E}\left[\sum_{i=1}^n \|\mathbf{x}_i^t - \bar{\mathbf{x}}^t\|^2\right] \leq \frac{F_0 - f_\star}{\mathsf{a}\gamma \rho_{\min} T/8} + \frac{8\alpha^2 \mathbb{C}_\sigma \bar{\sigma}^2}{\mathsf{a} n \gamma \rho_{\min}}, \end{aligned} \tag{5}$$

where

$$F_0 = F(\bar{\mathbf{x}}^0) + \mathsf{a}\mathcal{O}(\|\mathbf{x}^0\|_{\mathbf{K}}^2 + \eta\|\widehat{\lambda}^0\|_{\mathbf{K}}^2 + \tfrac{\alpha^2}{\eta}\|\nabla \mathbf{f}(\bar{\mathbf{x}}^0)\|_{\mathbf{K}}^2), \; \mathbb{C}_\sigma = \mathcal{O}(1 + \mathsf{a}(n^2 + \tfrac{\alpha n}{\eta \beta \rho_{\min}}))$$

- Setting $\mathsf{a} = \mathcal{O}(1/\sqrt{T})$, $\alpha = \sqrt{n/(T\sigma^2)}$ and consider $T \gg 1$ suffices to show

$$\mathbb{E}\left[\|\nabla F(\bar{\mathbf{x}}^{\mathsf{T}})\|^2\right] = \mathcal{O}\left(\bar{\sigma}/\sqrt{nT}\right) \Longrightarrow \textbf{linear speedup!}$$

Convergence of FSPDA - Transient Behavior

- From (5), we have
$$\mathbb{E}\left[\|\nabla F(\bar{\mathbf{x}}^{\mathsf{T}})\|^2\right] = \mathcal{O}\left(\bar{\sigma}/\sqrt{nT}\right)$$
- Effects of random graphs, random sparsification appears in the high-order term of $T \Rightarrow$ becomes non-dominant as $T \to \infty$.
- **[Vanishing Topology Effect]** Linear speedup is achieved after a transient time of
$$T_{\text{trans}} = \Omega\Big(\frac{\sigma_A^4}{\rho_{\min}^4} \cdot \max\left\{n^6 \rho_{\max}^2, \min\{\frac{\bar{\rho}_{\max}^4 \rho_{\max}^6}{n\bar{\sigma}\rho_{\min}^3}, \frac{n^{5/2}\bar{\rho}_{\max}^2 \rho_{\max}^4}{\bar{\sigma}\rho_{\min}^2}\}\right\}\Big)$$
(maybe improvable with a tighter analysis)
- But those effect remains dominantly in consensus error:
$$\mathbb{E}\left[\|\bar{\mathbf{x}}^{\mathsf{T}}\|_{\mathbf{K}}^2\right] = \mathcal{O}\left(n^2 \sigma_A^2 \rho_{\max}/(T\rho_{\min}^2)\right)$$

Experiments on MNIST

- We focus on SOTA that operates on **time varying / random** graphs & with communication compression via sparsification or quantization.

- *CHOCO-SGD* [Koloskova et al., 2019]: DSGD + error feedback.
- *Swarm-SGD* [Nadiradze et al., 2021]: DSGD + quantization + asynchronous optimization.

- Problem (1) defined as cross-entropy loss minimization with $n = 10$ agents.
- Model: 2-layer feedforward neural network with $d = 79510$ param.
- Dataset: MNIST with $m = 60000$ samples divided into 10 equal-sized parts:
  - uniformly at random (**shuffled** $\approx$ homogeneous data),
  - by label (**unshuffled** $\Rightarrow$ heterogeneous data).
- Graph: $\mathcal{G}$ taken as a complete graph but for each t, only 1-edge is selected for $\mathcal{G}^t$.

Experiments on MNIST

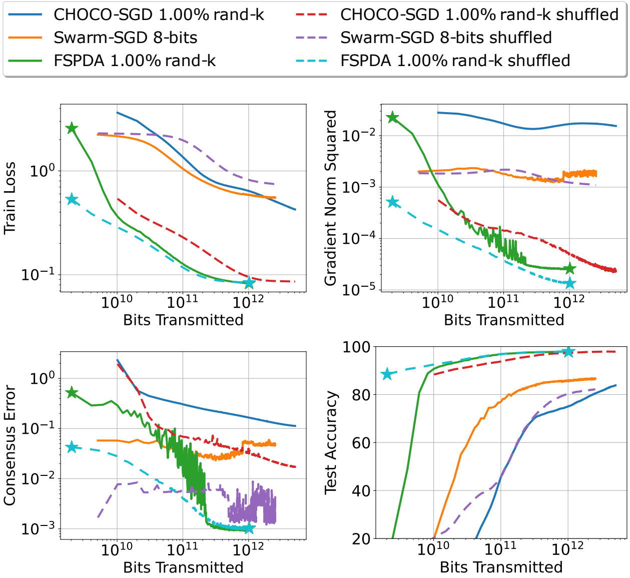


Figure: Feed-forward neural network classification training on MNIST with two levels of data heterogeneity.

Experiments on MNIST

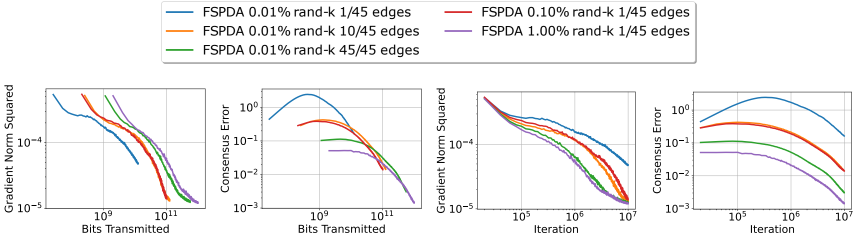


Figure: A feed-forward neural network ($d = 79510$) classification training on MNIST ($m = 60000$).

Experiments on Imagenet (10 node complete graph)

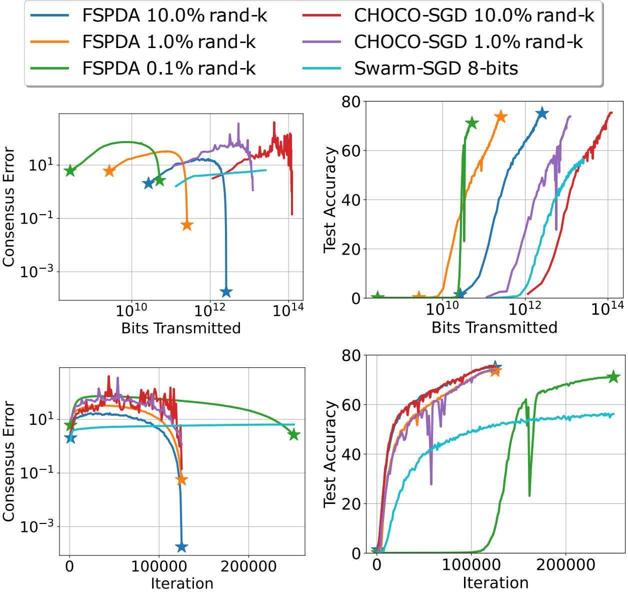


Figure: Resnet-50 classification ($d = 2.5 \times 10^7$) training on Imagenet ($m = 1.2 \times 10^6$ samples) over 100 epochs (200 epochs for FSPDA with 0.1% coordinate sparsity and Swarm-SGD).

Insights from Proof of Theorem 1

- Denote $\mathbf{v}^t := \widehat{\lambda}^t + \frac{\alpha}{\eta}\nabla\mathbf{f}((\mathbf{1}\otimes\mathbf{I})\bar{\mathbf{x}}^t) \leftarrow$ **'gradient tracking' variable**.
- Using potential function:

$$F_t = \mathbb{E}\left[F(\bar{\mathbf{x}}^t) + \mathsf{a}\|\mathbf{x}^t\|^2_{\mathbf{K}} + \mathsf{b}\|\mathbf{v}^t\|^2_{\mathbf{Q}+\mathsf{c}\mathbf{K}} + \mathsf{d}\left\langle \mathbf{x}^t \mid \mathbf{v}^t\right\rangle_{\mathbf{K}}\right]. \tag{6}$$

where $\mathbf{Q} = \mathbf{U}\mathbf{\Lambda}^{-1}\mathbf{U}^\top$ for orthogonal matrix $\mathbf{U}$ such that $\mathbf{A}^\top\mathbf{R}\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$.

- We can show

$$F_{t+1} \leq F_t - \frac{\alpha}{8}\mathbb{E}\left[\|\nabla F(\bar{\mathbf{x}}^t)\|^2\right] - \frac{\mathsf{a}\gamma\rho_{\min}}{8}\mathbb{E}\left[\|\mathbf{x}^t\|^2_{\mathbf{K}}\right] - \mathsf{a}\eta^2\mathbb{E}\left[\|\mathbf{v}^t\|^2_{\mathbf{K}}\right] + \mathbb{C}_\sigma\alpha^2\bar{\sigma}^2.$$

- We guarantee the convergence of $1/T\sum_{t=0}^{T-1}\mathbb{E}\left[\|\mathbf{v}^t\|^2_{\mathbf{K}}\right] = \mathcal{O}(1/T)$ where

$$\|\mathbf{v}^t\|^2_{\mathbf{K}} = 0 \Leftrightarrow \widehat{\lambda}^t + \frac{\alpha}{\eta}\nabla\mathbf{f}((\mathbf{1}\otimes\mathbf{I})\bar{\mathbf{x}}^t) - \frac{\alpha}{\eta n}\mathbf{1}\mathbf{1}^\top\nabla\mathbf{f}((\mathbf{1}\otimes\mathbf{I})\bar{\mathbf{x}}^t) = \mathbf{0} \tag{7}$$

$$\Leftrightarrow \widehat{\lambda}^t = \frac{\alpha}{\eta n}\mathbf{1}\mathbf{1}^\top\nabla\mathbf{f}((\mathbf{1}\otimes\mathbf{I})\bar{\mathbf{x}}^t) - \frac{\alpha}{\eta}\nabla\mathbf{f}((\mathbf{1}\otimes\mathbf{I})\bar{\mathbf{x}}^t) \tag{8}$$

- The dual variable $\widehat{\lambda}$ **corrects the local gradient into global gradient**.
- (Recall the primal update) $\mathbf{x}^{t+1} = \mathbf{x}^t - \alpha\nabla\mathbf{f}(\mathbf{x}^t;\xi^t) - \eta\widehat{\lambda}^t - \gamma\mathbf{A}^\top\mathbf{A}(\xi^t)\mathbf{x}^t$

Conclusion

- The first primal-dual (PD) decentralized algorithm in **a fully stochastic** setting, which supports:
  1. stochastic gradient
  2. random graph
  3. random sparsification
  4. asynchronous updates
- With **deterministic gradient** $+$ PL condition, FSPDA converges **linearly**.
- The FSPDA framework includes *EXTRA*, *Gradient Tracking* on static graph as special cases $\Rightarrow$ suggests a random graph extension for the latter.
- We have also extended the FSPDA framework to handle nonlinear compression (e.g., quantization) $\leftarrow$ requires a **two-timescale** updates.

Thank you. Comments are welcomed!

More details in `https://arxiv.org/abs/2410.18774`

Extension: TiCoPD - Primal-Dual with Error Feedback

- We extended the FSPDA algorithm to support nonlinear compression[2] (e.g., **quantization**) $\Rightarrow$ **T**wo-**ti**mescale **Co**mpressed **P**rimal-**D**ual algorithm (TiCoPD).

TiCoPD Algorithm. (the case of determ. gradient + static graph) [Liu et al., 2024].

$$\mathbf{x}^{t+1} = \mathbf{x}^t - \alpha \left(\nabla \mathbf{f}(\mathbf{x}^t) + \mathbf{A}^\top \lambda^t + \theta \mathbf{A}^\top \mathbf{A} \hat{\mathbf{x}}^t \right) \tag{9a}$$

$$\lambda^{t+1} = \lambda^t + \eta \mathbf{A} \mathbf{x}^t \tag{9b}$$

$$\hat{\mathbf{x}}_i^{t+1} = \hat{\mathbf{x}}_i^t + \gamma Q(\mathbf{x}^t - \hat{\mathbf{x}}^t; \xi^t) \tag{9c}$$

- Suppose there is a surrogate sequence $\hat{\mathbf{x}} \approx \mathbf{x}$. The augmented Lagrangian function can be majorized by

$$\|\mathbf{A}\mathbf{x}\|^2 \leq \|\mathbf{A}\hat{\mathbf{x}}^t\|^2 + 2(\mathbf{x} - \hat{\mathbf{x}}^t)^\top \mathbf{A}^\top \mathbf{A} \hat{\mathbf{x}}^t + M\|\mathbf{x} - \hat{\mathbf{x}}^t\|^2.$$

- Eq. (9a), (9b) is the resulting 'FSPDA' algorithm based on $\hat{\mathbf{x}}^t$.
- Eq. (9c) is a relaxed **fixed point iteration** step to achieve $\hat{\mathbf{x}} \approx \mathbf{x}$ — can be achieved using *contractive compression* such as randomized quantization.
- The convergence of TiCoPD requires **two-timescale** update such that $\alpha/\gamma \ll 1$.

[2][Liu et al., 2024] H. Liu, C.-Y. Yau, **H.-T.**, A Two-timescale Primal-dual Algorithm for Decentralized Optimization with Compression, in submission.

Convergence of TiCoPD

[Theorem] There exists a constant step size & parameter choices such that for any $T \geq 1$, it holds

$$\begin{aligned} \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla f(\bar{\mathbf{x}}^t)\right\|^2\right] &= \mathcal{O}\left(\frac{1}{T}\right), \\ \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\mathbf{x}^t\|_{\mathbf{K}}^2\right] &= \mathcal{O}\left(\frac{1}{T}\right). \end{aligned} \tag{10}$$

- The proof can be viewed as an extension of FSPDA's analysis.
- Same (asymptotic) convergence rate as state-of-the-art centralized algorithm.

Experiments with SOTA

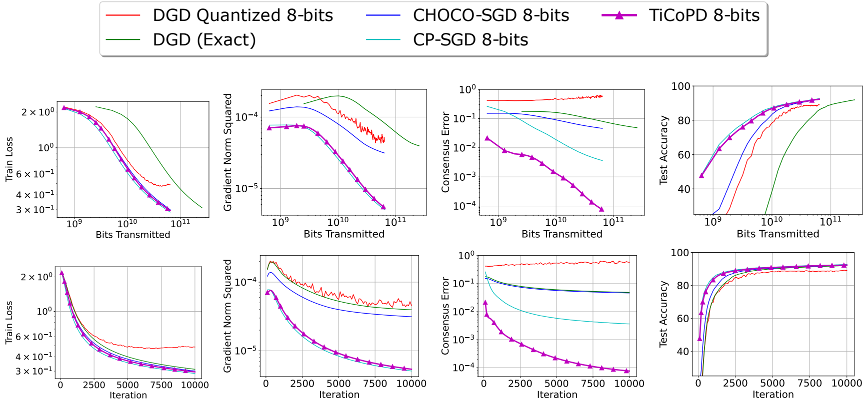


Figure: Training a 2-layer feedforward network using the MNIST data. The bit-rates for communication quantization are denoted in the legend.

Reference I

[Boyd et al., 2011] Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., et al. (2011).
Distributed optimization and statistical learning via the alternating direction method of multipliers.
Foundations and Trends® in Machine learning, 3(1):1–122.

[He et al., 2019] He, C., Tan, C., Tang, H., Qiu, S., and Liu, J. (2019).
Central server free federated learning over single-sided trust social networks.
arXiv preprint arXiv:1910.04956.

[Hong et al., 2017] Hong, M., Hajinezhad, D., and Zhao, M.-M. (2017).
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.

[Koloskova et al., 2019] Koloskova, A., Stich, S. U., and Jaggi, M. (2019).
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.

Reference II

[Liu et al., 2024] Liu, H., Yau, C.-Y., and Wai, H.-T. (2024).
A two-timescale primal-dual algorithm for decentralized optimization with compression.
Coming Soon.

[Mansoori and Wei, 2021] Mansoori, F. and Wei, E. (2021).
Flexpd: A flexible framework of first-order primal-dual algorithms for distributed optimization.
IEEE Transactions on Signal Processing, 69:3500–3512.

[Nadiradze et al., 2021] Nadiradze, G., Sabour, A., Davies, P., Li, S., and Alistarh, D. (2021).
Asynchronous decentralized sgd with quantized and local updates.
Advances in Neural Information Processing Systems, 34:6829–6842.

[Nedic et al., 2017] Nedic, A., Olshevsky, A., and Shi, W. (2017).
Achieving geometric convergence for distributed optimization over time-varying graphs.
SIAM Journal on Optimization, 27(4):2597–2633.

[Nedic and Ozdaglar, 2009] Nedic, A. and Ozdaglar, A. (2009).
Distributed subgradient methods for multi-agent optimization.
IEEE Trans. Automat. Control.

Reference III

[Qu and Li, 2017] Qu, G. and Li, N. (2017).
Harnessing smoothness to accelerate distributed optimization.
IEEE Transactions on Control of Network Systems, 5(3):1245–1260.

[Rabbat and Nowak, 2004] Rabbat, M. and Nowak, R. (2004).
Distributed optimization in sensor networks.
In *Proceedings of the 3rd international symposium on Information processing in sensor networks*, pages 20–27.

[Ryabinin and Gusev, 2020] Ryabinin, M. and Gusev, A. (2020).
Towards crowdsourced training of large neural networks using decentralized mixture-of-experts.
Advances in Neural Information Processing Systems, 33:3659–3672.

[Shi et al., 2015] Shi, W., Ling, Q., Wu, G., and Yin, W. (2015).
Extra: An exact first-order algorithm for decentralized consensus optimization.
SIAM Journal on Optimization, 25(2):944–966.

[Vlaski and Sayed, 2021] Vlaski, S. and Sayed, A. H. (2021).
Distributed learning in non-convex environments - part i: Agreement at a linear rate.
IEEE Transactions on Signal Processing, 69:1242–1256.

Reference IV

[Warnat-Herresthal et al., 2021] Warnat-Herresthal, S., Schultze, H., Shastry, K. L., Manamohan, S., Mukherjee, S., Garg, V., Sarveswara, R., Händler, K., Pickkers, P., Aziz, N. A., et al. (2021).
Swarm learning for decentralized and confidential clinical machine learning.
Nature, 594(7862):265–270.

[Xi and Khan, 2017] Xi, C. and Khan, U. A. (2017).
Dextra: A fast algorithm for optimization over directed graphs.
IEEE Transactions on Automatic Control, 62(10):4980–4993.

[Yau et al., 2024] Yau, C.-Y., Liu, H., and Wai, H.-T. (2024).
Fully stochastic primal-dual gradient algorithm for non-convex optimization on random graphs.
Coming Soon.

[Yi et al., 2021] Yi, X., Zhang, S., Yang, T., Chai, T., and Johansson, K. H. (2021).
Linear convergence of first-and zeroth-order primal–dual algorithms for distributed nonconvex optimization.
IEEE Transactions on Automatic Control, 67(8):4194–4201.

Reference V

[Yuan et al., 2022] Yuan, B., He, Y., Davis, J., Zhang, T., Dao, T., Chen, B., Liang, P. S., Re, C., and Zhang, C. (2022).
Decentralized training of foundation models in heterogeneous environments.
Advances in Neural Information Processing Systems, 35:25464–25477.