The 38th Annual Conference on Neural Information Processing Systems

NEURAL INFORMATION PROCESSING SYSTEMS

Achieving Near-Optimal Convergence for Distributed Minimax Optimization with Adaptive Stepsizes

Yan Huang [1], Xiang Li [2], Yipeng Shen [1], Niao He [2], Jinming Xu [1]

[1] Zhejiang University, China
[2] ETH Zurich, Switzerland

Vancouver, 13 Dec., 2024

浙江大学 Zhejiang University

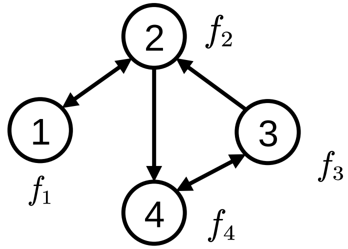
ETH zürich

OUTLINE

Motivation

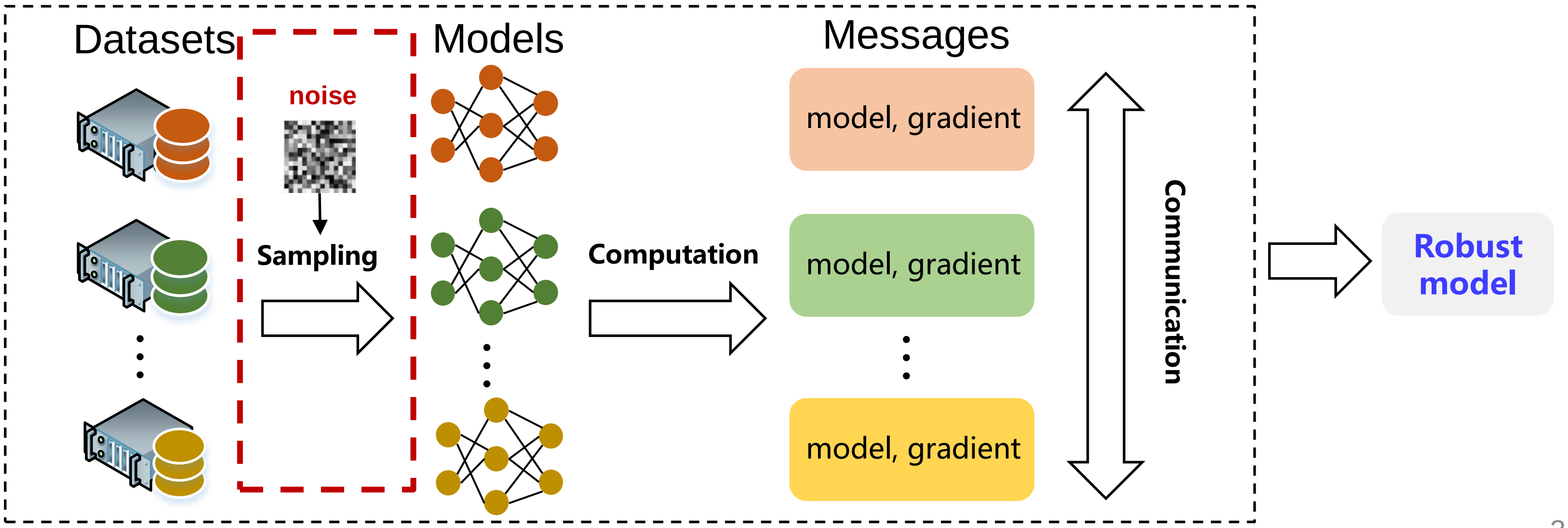
Distributed Minimax Optimization

$$\min_{x \in \mathbb{R}^p} \max_{y \in \mathcal{Y}} f(x,y) = \frac{1}{n} \sum_{i=1}^{n} f_i(x,y)$$



- $x \in \mathbb{R}^p$, $y \in \mathcal{Y}$ denote the min and max variables
- $\mathcal{Y} \subset \mathbb{R}^d$ is a compact set

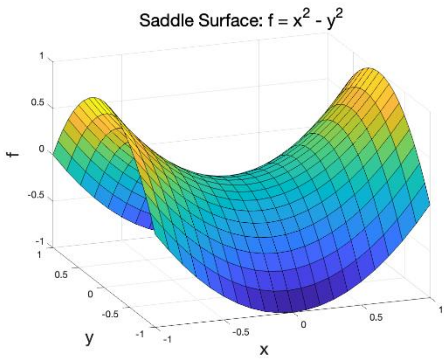
Example: robust distributed training



Motivation

Distributed Gradient Decent Ascent (DGDA)

$$\mathbf{x}_{k+1} = W(\mathbf{x}_k - \gamma_x \nabla_x F(\mathbf{x}_k, \mathbf{y}_k; \xi_k)),$$
$$\mathbf{y}_{k+1} = W(\mathbf{y}_k + \gamma_y \nabla_y F(\mathbf{x}_k, \mathbf{y}_k; \xi_k)),$$



- $W\mathbf{1} = \mathbf{1}, \quad \mathbf{1}^T W = \mathbf{1}^T, \quad \mathbf{x}_k = [x_{1,k}, \cdots, x_{n,k}]^T$
- $\nabla_x F(\mathbf{x}_k, \mathbf{y}_k; \xi_k) = [\cdots, \nabla_x f(x_{i,k}, x_{i,k}; \xi_{i,k}), \cdots]^T$

➢ For **non-convex and smooth** objectives, if $\gamma_x, \ \gamma_y \sim (L, \ \kappa, \ \rho_W)$

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}[\|\nabla \Phi(\bar{x}_k)\|^2] = \mathcal{O}\left(\frac{1}{K}\right) + \mathcal{O}\left(\frac{\gamma \kappa L}{n}\sigma^2 + \frac{\gamma^2 L^2 \kappa^2 \rho_W}{(1-\rho_W)^2}(\zeta^2 + \sigma^2)\right) + \mathcal{O}\left(\frac{\kappa^4 \gamma_x^2}{n \gamma_y^2}\sigma^2\right)$$

- $\Phi(x_k) := \max_{y_k} \{f(x_k, y_k)\}$ is the envelope function

➢ Depend on the prior knowledge of the objective and network

➢ Require two time-scale separation to achieve exact convergence

Motivation

Related works

- (Centralized) nonconvex minimax methods
  - Sharma et al. (2022) provide improved sample complexity of $\tilde{\mathcal{O}}(\epsilon^{-4})$ matching that of the lower bound of first-order algorithms for NC-SC problem (Li et al., 2021; Zhang et al., 2021a)
  - Requiring the prior knowledge about problem-dependent parameters

- (Federated) adaptive minimax methods.
  - Centralized parameter-agnostic methods such as NeAda (Yang et al., 2022b) and TiAda (Li et al., 2023)
  - Ju et al. (2023) and Huang et al. (2024) introduce Adam-based federated adaptive minimax algorithms with full-client participation

Question: Can we design a parameter-agnostic adaptive minimax method that ensures exact convergence in fully decentralized settings?

Algorithm Design

A direct extension： D-TiAda

- Extending TiAda (Li et al., 2023) to decentralized setting

$$\mathbf{x}_{k+1} = W\left(\mathbf{x}_k - \gamma_x V_{k+1}^{-\alpha} \nabla_x F(\mathbf{x}_k, \mathbf{y}_k; \xi_k)\right),$$

$$\mathbf{y}_{k+1} = W\left(\mathbf{y}_k + \gamma_y U_{k+1}^{-\beta} \nabla_y F(\mathbf{x}_k, \mathbf{y}_k; \xi_k)\right),$$

- $V_{k+1} = \mathrm{diag}\{v_{i,k+1}\}, \quad v_{i,k+1} = v_{i,k} + \|\nabla_x f_i(x_{i,k}, y_{i,k}; \xi_{i,k})\|^2, \ i \in [n]$
- $U_{k+1} = \mathrm{diag}\{u_{i,k+1}\}, \quad u_{i,k+1} = u_{i,k} + \|\nabla_y f_i(x_{i,k}, y_{i,k}; \xi_{i,k})\|^2, \ i \in [n]$

AdaGrad

- $0 < \beta < \alpha < 1$

- Achieve two time-scale separation automatically
- There is a bias on the gradient due to the inconsistent adaptive stepsizes

Algorithm Design

Bias caused by inconsistent scalars

$$\mathbf{x}_{k+1} = W\left(\mathbf{x}_k - \gamma_x V_{k+1}^{-\alpha} \nabla_x F(\mathbf{x}_k, \mathbf{y}_k; \xi_k)\right),$$

averaged system

$$\bar{x}_{k+1} = \underbrace{\bar{x}_k - \gamma_x \bar{v}_{k+1}^{-\alpha} \frac{\mathbf{1}^T}{n} \nabla_x F(\mathbf{x}_k, \mathbf{y}_k; \xi_k)}_{\textbf{SGD}} + \underbrace{\gamma_x \frac{\left(\tilde{\boldsymbol{v}}_{k+1}\right)^T}{n} \nabla_x F(\mathbf{x}_k, \mathbf{y}_k; \xi_k)}_{\textbf{inconsistency}}$$

- $\bar{x}_k := \frac{1}{n}\sum_{i=1}^{n} x_{i,k}, \quad \bar{v}_k = \frac{1}{n}\sum_{i=1}^{n} v_{i,k}, \quad \left(\tilde{\boldsymbol{v}}_k^{-\alpha}\right)^T = \left[\cdots \quad v_{i,k}^{-\alpha} - \bar{v}_k^{-\alpha} \quad \cdots\right]$

➢ Bounded inconsistency of the adaptive step-sizes

$$\zeta_v^2 := \sup_{i\in[n],k>0} \left\{ \left(v_{i,k}^{-\alpha} - \bar{v}_k^{-\alpha}\right)^2 / \left(\bar{v}_k^{-\alpha}\right)^2 \right\},$$

$$\zeta_u^2 := \sup_{i\in[n],k>0} \left\{ \left(u_{i,k}^{-\beta} - \bar{u}_k^{-\beta}\right)^2 / \left(\bar{u}_k^{-\beta}\right)^2 \right\}.$$

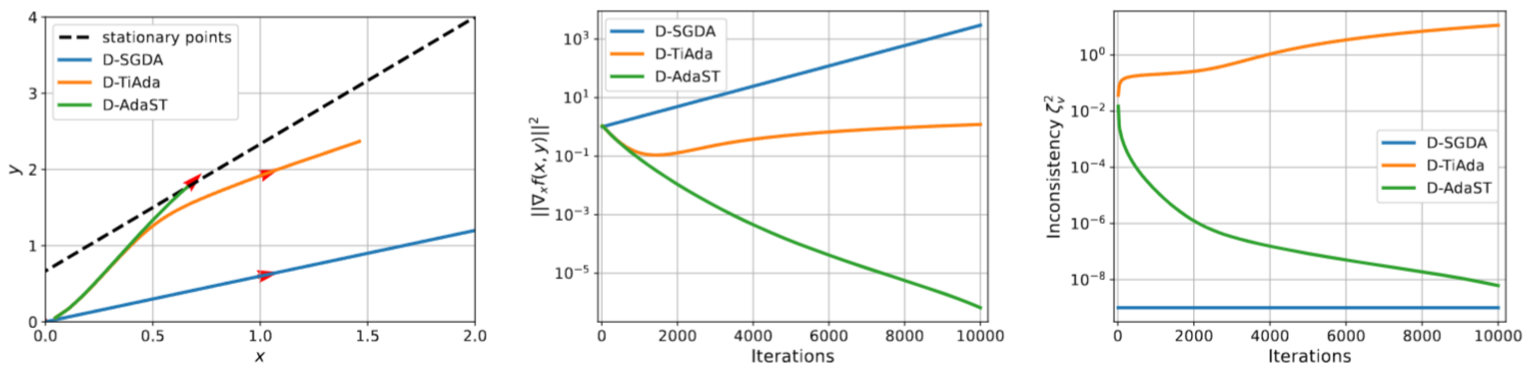
Algorithm Design

Counterexample

Theorem 1 (impact of the inconsistency)

There exists a distributed minimax problem and certain initialization such that after running an adaptive method, it holds that for $t \geqslant 0$

$$\|\nabla_x f(x_t, y_t)\| = \|\nabla_x f(x_0, y_0)\| \quad \text{and} \quad \|\nabla_y f(x_t, y_t)\| = \|\nabla_y f(x_0, y_0)\|$$



(a) trajectory (b) convergence of $\|\nabla_x f(x, y)\|^2$ (c) convergence of ζ_v^2

- Directly applying adaptive methods might lead to non-convergence in distributed settings

Algorithm Design

D-AdaST: compact form

Stepsize tracking

$$\begin{cases} \mathbf{m}_{k+1}^x = W(\mathbf{m}_k^x + \mathbf{h}_k^x) \\ \mathbf{m}_{k+1}^y = W(\mathbf{m}_k^y + \mathbf{h}_k^y) \end{cases}$$

Adaptive update

$$\begin{cases} \mathbf{x}_{k+1} = W\left(\mathbf{x}_k - \gamma_x V_{k+1}^{-\alpha} \nabla_x F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x)\right) \\ \mathbf{y}_{k+1} = \mathcal{P}_{\mathcal{Y}}\left(W\left(\mathbf{y}_k + \gamma_y U_{k+1}^{-\beta} \nabla_y F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y)\right)\right) \end{cases}$$

- $\boldsymbol{h}_k^x = [\cdots, \| g_{i,k}^x \|^2, \cdots]^T \in \mathbb{R}^n, \ \boldsymbol{h}_k^y = [\cdots, \| g_{i,k}^y \|^2, \cdots]^T \in \mathbb{R}^n,$
- $V_{k+1}^{-\alpha} = \mathrm{diag}\{v_{i,k+1}^{-\alpha}\}_{i=1}^n, \quad v_{i,k+1} = \max\{m_{i,k+1}^x, m_{i,k+1}^y\},$
- $U_{k+1}^{-\beta} = \mathrm{diag}\{u_{i,k+1}^{-\beta}\}_{i=1}^n, \quad u_{i,k+1} = m_{i,k+1}^y$

➢ Achieving consistency in local adaptive stepsizes asymptoticly

Main Results

Assumptions

- (**NCSC**) Each f_i is μ -strongly concave in y
- (**Joint smoothness**) Each f_i is L -smooth and second-order Lipschitz continuous in y
- (**Stochastic gradient**) The stochastic gradient of each node is unbiased and there exists a constant $C > 0$ such that $\|\nabla_z F_i(x, y; \xi_i)\|^2 \leqslant C,\ z \in \{x, y\}$
- (**Graph connectivity**) The spectral norm of the doubly stochastic matrix satisfies $\rho_W := \|W - \mathbf{J}_n\|_2^2 < 1$

Main Results

Convergence results

Theorem 2 (near-optimal convergence)

Suppose assumptions hold. Let $0 < \alpha < \beta < 1$ and the total iteration satisfy

$$K = \Omega\Big(\max\Big\{(\gamma_x^2 \kappa^4 / \gamma_y^2)^{1/(\alpha-\beta)}, \quad (1/(1-\rho_W)^2)^{\max\{1/\alpha,\ 1/\beta\}}\Big\}\Big),$$

to ensure time-scale separation and quasi-independence of network. Then,

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\left[\|\nabla \Phi(\bar{x}_k)\|^2\right] = \tilde{\mathcal{O}}\left(\frac{1}{K^{1-\alpha}} + \frac{1}{(1-\rho_W)^{\alpha} K^{\alpha}} + \frac{1}{K^{1-\beta}} + \frac{1}{(1-\rho_W) K^{\beta}}\right).$$

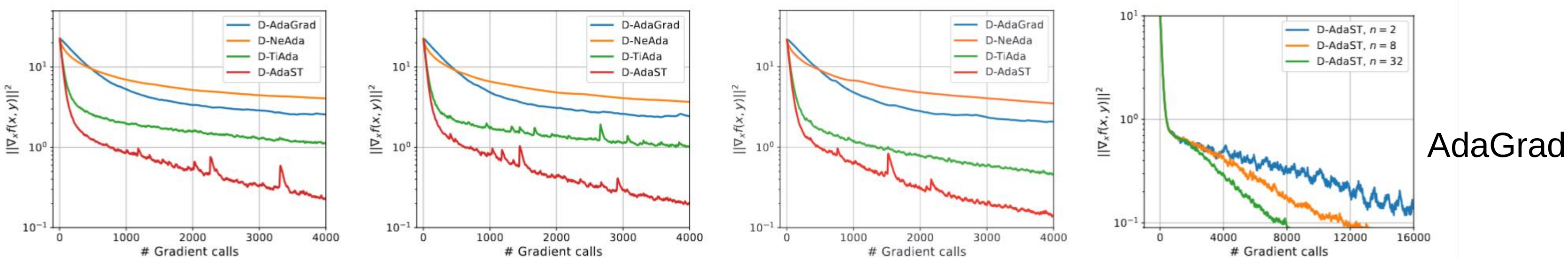
- $\Phi(x_k) := \max_{y_k} \{f(x_k, y_k)\}$ is the envelope function

- Near-optimal convergence rate $\tilde{\mathcal{O}}(\epsilon^{-(4+\delta)})$ with arbitrary small $\delta > 0$
- Parameter-agnostic property without requiring to know prior knowledge

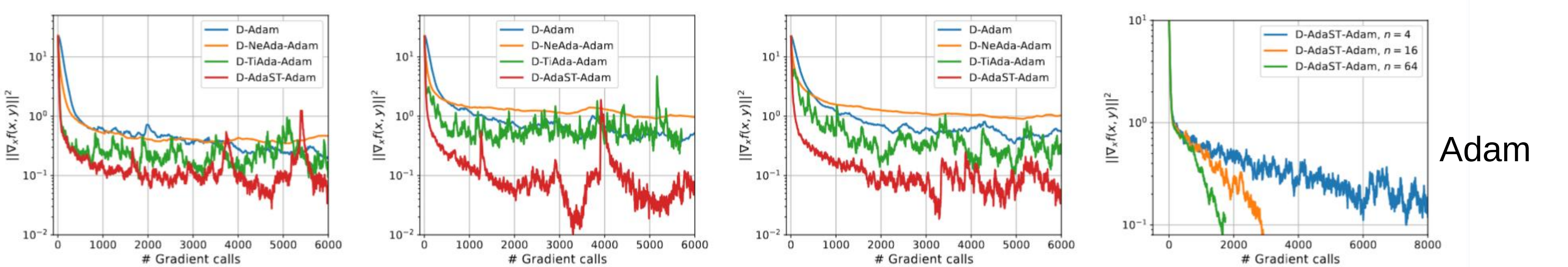
Experiment

Training robust CNN on MNIST

$$\min_{x} \max_{y} 1/n \sum_{j=1}^{n} f_i(x;\xi_i + y) - \eta \|y\|^2$$



(a) ring, $\rho_W = 0.97$ (b) exp., $\rho_W = 0.67$ (c) dense, $\rho_W = 0.55$ (d) scalability

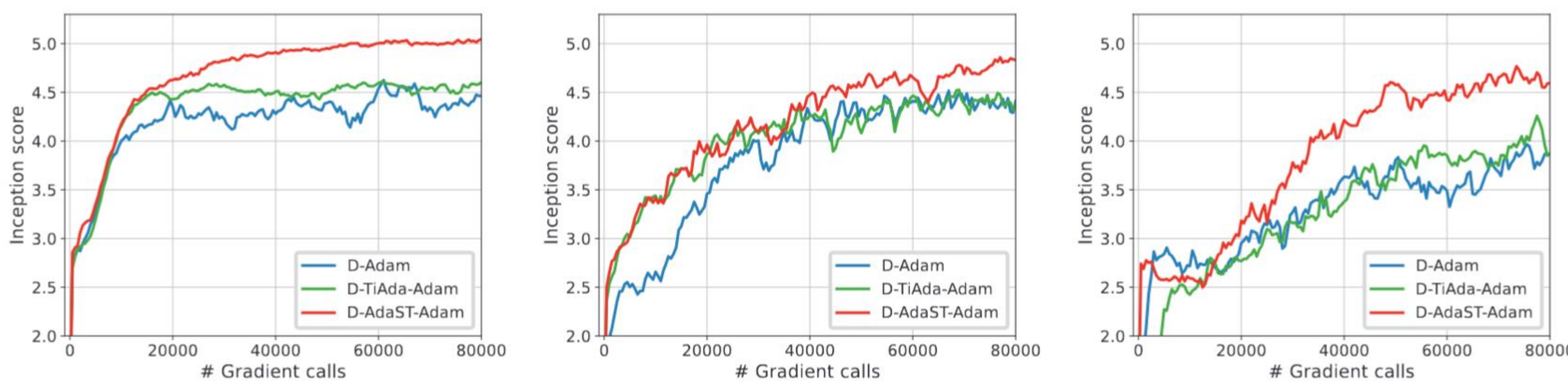


(e) ring, $\rho_W = 0.97$ (f) exp., $\rho_W = 0.67$ (g) dense, $\rho_W = 0.55$ (h) scalability

Experiment

Training GANs on CIFAR-10

- Generator (min player): four-layer transposed CNN
- Discriminator (max player): four-layer CNN



(a) $\gamma_x = \gamma_y = 0.001$ (b) $\gamma_x = \gamma_y = 0.01$ (c) $\gamma_x = \gamma_y = 0.05$

- D-AdaST exhibits the best performance under different settings

Conclusion

Takeaways

- Directly extending centralized adaptive method to decentralized setting, e.g., D-TiAda, might lead to non-convergence
- The proposed D-AdaST achieves a near-optimal convergence rate by stepsize tracking and is parameter-agnostic

Future works

- Incorporate gradient tracking to remove assumptions about the bounded gradient norm
- Consider non-monotonic adaptive stepsizes, such as Adam, and provide theoretical guarentee

Reference

- Sharma, P., Panda, R., Joshi, G., and Varshney, P. (2022). Federated minimax optimization: Improved convergence analyses and algorithms. ICML.
- Li, H., Tian, Y., Zhang, J., and Jadbabaie, A. (2021). Complexity lower bounds for nonconvex-strongly-concave min-max optimization. NeurIPS.
- Zhang, S., Yang, J., Guzmán, C., Kiyavash, N., and He, N. (2021a). The complexity of nonconvex-strongly-concave minimax optimization. In Uncertainty in Artificial Intelligence.
- Yang, J., Li, X., and He, N. (2022b). Nest your adaptive algorithm for parameter-agnostic nonconvex minimax optimization. NeurIPS.
- Li, X., YANG, J., and He, N. (2023). Tiada: A time-scale adaptive algorithm for nonconvex minimax optimization. ICLR.
- Ju, L., Zhang, T., Toor, S., and Hellander, A. (2023). Accelerating fair federated learning: Adaptive federated adam. arXiv:2301.09357.
- Huang, F., Wang, X., Li, J., and Chen, S. (2024). Adaptive federated minimax optimization with lower complexities. AISTATS.

Thank you!

Yan Huang

huangyan5616@zju.edu.cn

(Full paper)