

Topology-aware Generalization of Decentralized SGD

Tongtian Zhu^{1 2}, Fengxiang He², Lan Zhang³, Zhengyang Niu^{2 4}, Mingli Song¹, Dacheng Tao²

1 Zhejiang University, 2 JD Explore Academy, 3 University of Science and Technology of China, 4 Wuhan University



Content

- 1 Background
- 2 Theoretical results
- 3 Empirical results
- 4 Implications
- 5 Reference
- 6 Acknowledgement

Decentralized Learning



Compared with centralized training, training in a decentralized fashion

- improves computation efficiency and privacy¹;
- suffers from a severe drop in generalizability, especially when the communication topology is large or sparse.

To date, research on decentralized optimization algorithms has mainly focused on their convergence. Understanding their generalizability is still premature.

¹Swarm Learning for decentralized and confidential clinical machine learning.

Decentralized Learning

Decentralized SGD (D-SGD)

- Distributed learning jointly trains a global model \mathbf{w} by optimizing the empirical risk $\frac{1}{mn} \sum_{k=1}^m \sum_{\zeta=1}^n f(\mathbf{w}; z_{k,\zeta})$, where $z_{k,\zeta} \in \mathcal{S}_k$ ($\zeta = 1, \dots, n$) is the local training data on k -th worker.
- D-SGD aims to learn a consensus model $\bar{\mathbf{w}} = \frac{1}{m} \sum_{k=1}^m \mathbf{w}_k$ on m workers, where \mathbf{w}_k denotes the local model on the k -th worker. D-SGD updates the model on k -th worker by

$$\mathbf{w}_k^{(t+1)} = \overbrace{\sum_{l=1}^m \mathbf{P}_{k,l} \mathbf{w}_l^{(t)}}^{\text{Communication}} - \overbrace{\eta_t \nabla f(\mathbf{w}_k^{(t)}; z_{k,\zeta_t})}_{\text{Computation}},$$

where \mathbf{P} is a doubly stochastic gossip matrix characterizing the topology \mathcal{G} .

Decentralized Learning

Optimization


- D-SGD can achieve the same asymptotic linear speedup in convergence rate as centralized SGD (Lian et al., 2017).
- The convergence of D-SGD highly relies on the communication topology (Bars et al., 2022; Bianchi and Jakubowicz, 2012).
- D-SGD has been extended to various settings, including asynchronous settings (Lian et al., 2018; Xu et al., 2021), time-varying topologies (Koloskova et al., 2020; Lu and Wu, 2020), and data heterogeneous scenarios (Tang et al., 2018; Vogels et al., 2021).

Decentralized Learning

Generalization

- Sun et al. (2021) derives generalization bounds² of projected variant of D-SGD with order $\mathcal{O}(N^{-1}+(1-\lambda)^{-1})$, where $(1-\lambda)^{-1}$ is very large for sparse or large topology.
- Another work by Richards et al. (2020) proves generalization bounds³ of the Adaptation-Then-Combination (ATC) version of D-SGD through algorithmic stability and Rademacher complexity. However, their generalization bounds are invariant to the communication topology, which contradicts the empirical results.

²Stability and Generalization of Decentralized Stochastic Gradient Descent.

³Graph-dependent implicit regularisation for distributed stochastic subgradient descent. 

Motivation

Our Empirical results⁴ show that the generalizability of decentralized SGD (D-SGD) depends on the network topology and the total number of workers⁵:

worker number	models & topology	test accuracy of training ResNet-18 on CIFAR-10			
		Ring	Grid	Exponential	Fully-connected
m=16		82.00±0.18	81.85±0.17	81.89±0.02	82.02±0.11
m=32		80.51±0.09	81.35±0.26	81.64±0.12	81.40±0.06
m=64		78.75±0.25	80.72±0.23	81.15±0.08	81.37±0.13

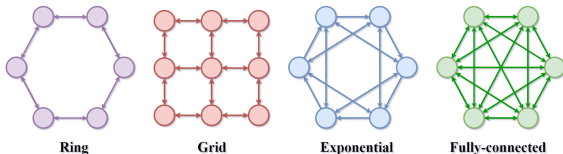


Figure: Illustration of commonly used communication topologies.

⁴Techniques including batch normalization, dropout, momentum, weight decay and data augmentation are disabled.

⁵See similar results in Kong et al. (2021) and Ying et al. (2021).

Contribution

Thus, a question is raised:

Our question

How does the communication topology and the number of workers impact the generalizability of the consensus model learned by D-SGD?

Our contribution

- We establish topology-aware algorithmic stability and generalization bounds of the consensus model learned by D-SGD in non-convex non-smooth setting. Our theory suggests that the generalizability of D-SGD has a positive relationship with the spectral gap of the communication topology.
- We calculate the difference between validation loss and training loss of D-SGD on deep learning models to collaborate the theory.

Assumptions

- **Non-smoothness:**

$\nabla f(\cdot, z)$ is (α, L) -Hölder continuous if for all $\mathbf{w}, \tilde{\mathbf{w}} \in \mathbb{R}^d$ and $z \in \mathcal{Z}$,

$$\|\nabla f(\mathbf{w}; z) - \nabla f(\tilde{\mathbf{w}}; z)\|_2 \leq L \|\mathbf{w} - \tilde{\mathbf{w}}\|_2^\alpha.$$

- **Gaussian weight difference:**

Let $\mathbf{w}_k^{(t)}$ and $\tilde{\mathbf{w}}_k^{(t)}$ denote the t -th iteration on the k -th worker produced by D-SGD based on \mathcal{S}_k and $\mathcal{S}_k^{(i)}$ that differ by only one data point, respectively. We assume $\mathbf{w}_k^{(t)} - \tilde{\mathbf{w}}_k^{(t)} \sim \mathcal{N}(\mu_{t,k}, \sigma_{t,k}^2 I_d)$ for all k with unknown parameters $\mu_{t,k}$ and $\sigma_{t,k}$.

Distributed on-average stability

We define a new algorithmic stability specifically for distributed optimization algorithms:

Definition (Distributed on-average stability)

A distributed optimization algorithm A is ℓ_2 distributed on-average ϵ -stable for all training data sets \mathcal{S}_k and $\mathcal{S}_k^{(i)}$ ($k = 1 \dots m$) if

$$\frac{1}{mn} \sum_{i=1}^n \sum_{k=1}^m \mathbb{E}_{\mathcal{S}_k, \mathcal{S}_k^{(i)}, A} [\|\mathbf{w}_k - \tilde{\mathbf{w}}_k\|_2^2] \leq \epsilon^2,$$

where $\mathbb{E}_A[\cdot]$ denotes the expectation w.r.t. the randomness of the algorithm A .

- The stability characterizes the on-average sensitivity of models across multiple workers.

Distributed on-average stability of D-SGD

Lemma (Distributed on-average stability bound)

The distributed on-average stability of D-SGD can be bounded as⁶

$$\frac{1}{mn} \sum_{i=1}^n \sum_{k=1}^m \mathbb{E}_{\mathcal{S}_k, \mathcal{S}_k^{(i)}, \mathcal{A}} [\|\mathbf{w}_k^{(t+1)} - \tilde{\mathbf{w}}_k^{(t+1)}\|_2^2] \leq \frac{1}{1 - 2\eta L(1 - \frac{1}{n})} \left\{ \underbrace{\mathcal{O}\left(\frac{\epsilon_S \eta^2}{n}\right)}_{\mathcal{O}\left(\frac{m}{N}\right)} + \underbrace{\mathcal{O}\left(\left(1 - \frac{1}{m}\right)\lambda^2 + \frac{1}{m}\right)}_{\text{Error from decentralization}} \right\}.$$

- The algorithmic stability bound of D-SGD increases monotonically with λ , which measures the connectivity of the communication topology.

⁶For simplicity, we fix the learning rate $\eta_t \equiv \eta$.

Topology-aware Generalization bound of D-SGD

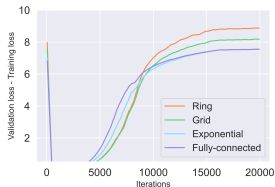
Main theorem (Generalization bound)

If we denote the global model $\bar{\mathbf{w}}^{(t)} = \frac{1}{m} \sum_{k=1}^m \mathbf{w}_k^{(t)}$, the generalization error of $\bar{\mathbf{w}}^{(t)}$ learned by D-SGD can be controlled as

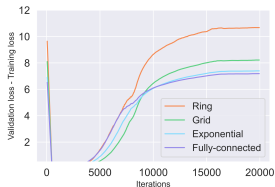
$$\mathbb{E}_{S, \mathcal{A}} [F(\bar{\mathbf{w}}^{(t)}) - F_S(\bar{\mathbf{w}}^{(t)})] \leq \frac{L}{[1 - 2\eta L(1 - \frac{1}{n})]^{\frac{\alpha}{2}}} \left\{ \underbrace{\mathcal{O}\left(\frac{\epsilon_S^{\frac{\alpha}{2}}}{N}\right)}_{\mathcal{O}(\frac{1}{N})} + \underbrace{\frac{n^{\frac{\alpha}{2}}}{N} \left((1 - \frac{1}{m})\lambda^2 + \frac{1}{m} \right)^{\frac{\alpha}{2}}}_{\text{Error from decentralization}} \right\}.$$

- The generalization bounds are non-vacuous, even when the worker number is sufficiently large or the communication graph is sufficiently sparse ($\lambda \rightarrow 1$).
- The generalizability of D-SGD positively correlates with the spectral gap $1 - \lambda$.

Empirical results



(a) Tiny ImageNet, 32 workers



(b) Tiny ImageNet, 64 workers

Graph topology	Spectral gap $1 - \lambda$
Disconnected	0
Ring	$\mathcal{O}(1/m^2)$
Grid	$\mathcal{O}(1/(m \log_2(m)))$
Exponential	$\mathcal{O}(1/\log_2(m))$
Fully-connected	1

Figure: Loss differences of training VGG-11 with D-SGD on different topologies⁷.

- D-SGD generalizes better on well-connected topology with a larger spectral gap;
- As the number of workers increases, the generalization gap of D-SGD on different topologies increase.

⁷ Similar results are obtained on CIFAR-10 and CIFAR-100.

Implications

Corollary

Suppose that the consensus distance satisfies $\Gamma^2 \leq \frac{1}{m} \sum_{k=1}^m \|\mathbf{w}_k^{(\tau)} - \bar{\mathbf{w}}^{(\tau)}\|_2^2 \leq K^2$ for $\tau \leq t_\Gamma$, and is controlled below Γ^2 for $\tau > t_\Gamma$. We can conclude that the distributed on-average stability bound of D-SGD increases monotonically with t_Γ , if the total number of iterations $t \geq \frac{-C}{2 \ln C}$.

- Consensus distance control is beneficial for the distributed on-average stability and thus for the generalizability of D-SGD;
- Controlling the consensus distance at initial stage of training is more effective than at the end.

Summary

Objective: We theoretically analyze the impact of the communication topology of Decentralized SGD (D-SGD) on its generalizability in non-convex non-smooth setting.

Results: We establish topology-aware algorithmic stability and generalization bounds of the consensus model learned by D-SGD. Our theory indicates that the generalizability of D-SGD has a positive correlation with the spectral gap of the underlying communication topology, which is further justified by comprehensive deep learning experiments.

Implications: The theory we propose can explain (1) why D-SGD generalizes poorly on large and sparse topologies; (2) why consensus distance control in the initial training phase can ensure better generalization. The findings also provide guidelines for designing better communication topologies for decentralized deep learning problems.

Reference

- Bars, B. L., Bellet, A., Tommasi, M., and Kermarrec, A.-M. (2022). Yes, topology matters in decentralized optimization: Refined convergence and topology learning under heterogeneous data. *arXiv preprint arXiv:2204.04452*.
- Bianchi, P. and Jakubowicz, J. (2012). Convergence of a multi-agent projected stochastic gradient algorithm for non-convex optimization. *IEEE transactions on automatic control*.
- Koloskova, A., Loizou, N., Boreiri, S., Jaggi, M., and Stich, S. (2020). A unified theory of decentralized SGD with changing topology and local updates. In *Proceedings of the 37th International Conference on Machine Learning*.
- Kong, L., Lin, T., Koloskova, A., Jaggi, M., and Stich, S. (2021). Consensus control for decentralized deep learning. In *Proceedings of the 38th International Conference on Machine Learning*.
- Lian, X., Zhang, C., Zhang, H., Hsieh, C.-J., Zhang, W., and Liu, J. (2017). Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. In *Advances in Neural Information Processing Systems*.

Reference

- Lian, X., Zhang, W., Zhang, C., and Liu, J. (2018). Asynchronous decentralized parallel stochastic gradient descent. In *International Conference on Machine Learning*.
- Lu, S. and Wu, C. W. (2020). Decentralized stochastic non-convex optimization over weakly connected time-varying digraphs. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
- Richards, D. et al. (2020). Graph-dependent implicit regularisation for distributed stochastic subgradient descent. *Journal of Machine Learning Research*.
- Sun, T., Li, D., and Wang, B. (2021). Stability and generalization of decentralized stochastic gradient descent. *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Tang, H., Lian, X., Yan, M., Zhang, C., and Liu, J. (2018). D2: Decentralized training over decentralized data. In *International Conference on Machine Learning*.
- Vogels, T., He, L., Koloskova, A., Karimireddy, S. P., Lin, T., Stich, S. U., and Jaggi, M. (2021). Relaysun for decentralized deep learning on heterogeneous data. *Advances in Neural Information Processing Systems*.

Reference

- Xu, J., Zhang, W., and Wang, F. (2021). A (dp)2 2sgd: Asynchronous decentralized parallel stochastic gradient descent with differential privacy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Ying, B., Yuan, K., Chen, Y., Hu, H., PAN, P., and Yin, W. (2021). Exponential graph is provably efficient for decentralized deep training. In *Advances in Neural Information Processing Systems*.

Acknowledgement

- The authors would like to thank Chun Li, Yingjie Wang, Haowen Chen, and Shaopeng Fu for their insightful comments on the revision of this manuscript and appreciate Li Shen, Rong Dai, and Luofeng Liao for their helpful discussions. We also sincerely thank the anonymous ICML reviewers and chairs for their constructive comments.

Thank You!

Contact: raiden@zju.edu.cn (Tongtian Zhu)