Topology-aware Generalization of Decentralized SGD

Tongtian Zhu [1 2 3] **Fengxiang He** [3] **Lan Zhang** [4 5] **Zhengyang Niu** [6 3] **Mingli Song** [7] **Dacheng Tao** [3]

Abstract

This paper studies the algorithmic stability and generalizability of decentralized stochastic gradient descent (D-SGD). We prove that the consensus model learned by D-SGD is $\mathcal{O}(m/N+1/m+\lambda^2)$-stable in expectation in the non-convex non-smooth setting, where N is the total sample size of the whole system, m is the worker number, and $1-\lambda$ is the spectral gap that measures the connectivity of the communication topology. These results then deliver an $\mathcal{O}(1/N+((m^{-1}\lambda^2)^{\frac{\alpha}{2}}+m^{-\alpha})/N^{1-\frac{\alpha}{2}})$ in-average generalization bound, which is non-vacuous even when λ is closed to 1, in contrast to vacuous as suggested by existing literature on the projected version of D-SGD. Our theory indicates that the generalizability of D-SGD has a positive correlation with the spectral gap, and can explain why consensus control in initial training phase can ensure better generalization. Experiments of VGG-11 and ResNet-18 on CIFAR-10, CIFAR-100 and Tiny-ImageNet justify our theory. To our best knowledge, this is the first work on the topology-aware generalization of vanilla D-SGD. Code is available at `https://github.com/Raiden-Zhu/Generalization-of-DSGD`.

1. Introduction

Decentralized stochastic gradient descent (D-SGD) facilitates simultaneous model training on a massive number of workers without a central server (Lopes & Sayed, 2008; Nedic & Ozdaglar, 2009b). In D-SGD, every worker only communicates with the directly connected neighbors through "gossip communication" (Xiao & Boyd, 2004; Lian et al., 2017; Koloskova et al., 2020). The communication intensity is controlled by the communication topology. This decentralized nature eliminates the requirement for an expensive central server dedicated to heavy communication and computation. Surprisingly, existing theoretical results demonstrate that the massive models on the edge converge to a unique steady model, the consensus model, even without the control of a central server (Lu et al., 2011; Shi et al., 2015; Lian et al., 2017). Compared with the centralized synchronized SGD (C-SGD) (Dean et al., 2012; Li et al., 2014), D-SGD can achieve the same asymptotic linear speedup in convergence rate (Zhang et al., 2019; 2021). In this way, D-SGD provides a promising distributed machine learning paradigm with improved privacy (Nedic, 2020), scalability (Lian et al., 2017; Kairouz et al., 2021), and communication efficiency (Ying et al., 2021b).

To date, the theoretical research on D-SGD has mainly focused on its convergence (Nedic & Ozdaglar, 2009b; Lian et al., 2017; Koloskova et al., 2020; Alghunaim & Yuan, 2021), while the understanding on the generalizability (Mohri et al., 2018; He & Tao, 2020) of D-SGD is still premature. A large amount of empirical evidence have shown that D-SGD generalizes well on well-connected topologies (Assran et al., 2019a; Ying et al., 2021a). Meanwhile, empirical results by Assran et al. (2019b), Kong et al. (2021) and Ying et al. (2021a) demonstrate that for ring topologies, the validation accuracy of the consensus model learned by D-SGD decreases as the number of workers increases. Thus, a question is raised:

> Our question
>
> *How does the communication topology of D-SGD impact its generalizability?*

This paper answers this question. We prove a topology-aware generalization error bound for the consensus model learned by D-SGD, which characterizes the impact of the communication topology on the generalizability of D-SGD. Our contributions are summarized as follows:

- **Stability and generalization bounds of D-SGD.** This work proves the algorithmic stability (Bousquet & Elis-

[1]College of Computer Science and Technology, Zhejiang University [2]Shanghai Institute for Advanced Study of Zhejiang University [3]JD Explore Academy, JD.com Inc. [4]School of Computer Science and Technology, University of Science and Technology of China [5]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center [6]School of Computer Science, Wuhan University [7]Zhejiang University City College. Correspondence to: Fengxiang He <fengxiang.f.he@gmail.com>.

Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

seeff, 2002) and generalization bounds of vanilla D-SGD in the non-convex non-smooth setting. In Section 4, we present an $\mathcal{O}(m/N+1/m+\lambda^2)$ distributed on-average stability (see Corollary 2), where $1-\lambda$ denotes the spectral gap of the network, a measure of the connectivity of the communication topology $\mathcal{G}$. These results would suffice to derive a $\mathcal{O}(1/N + ((m^{-1}\lambda^2)^{\frac{\alpha}{2}} + m^{-\alpha})/N^{1-\frac{\alpha}{2}})$ generalization bound in expectation of D-SGD (see Theorem 4). Our error bounds are non-vacuous, even when the worker[1] number is sufficiently large, or the communication graph is sufficiently sparse. The theory can be directly applied to explain why consensus distance control in the initial phase of training can ensure better generalization.

- **Communication topology and generalization of D-SGD.** Our theory shows that the generalizability of D-SGD has a positive relationship with the spectral gap $1-\lambda$ of the communication topology $\mathcal{G}$. Besides, we prove that the generalizability of D-SGD decreases when the worker number increases for the ring, grid, and exponential graphs. We conduct comprehensive experiments of VGG-11 (Simonyan & Zisserman, 2014) and ResNet-18 (He et al., 2016b) on CIFAR-10, CIFAR-100 (Krizhevsky et al., 2009) and Tiny-ImageNet (Le & Yang, 2015) to verify our theory.

To our best knowledge, this work offers the first investigation into the topology-aware generalizability of vanilla D-SGD. The closest work in the existing literature is by Sun et al. (2021), which derives $\mathcal{O}(N^{-1} + (1-\lambda)^{-1})$ generalization bounds for projected D-SGD based on uniform stability (Bousquet & Elisseeff, 2002). They show that the decentralized nature hurts the stability, and thus undermines generalizability. Compared with the results by Sun et al. (2021), our work makes two contributions: (1) we analyze the vanilla D-SGD, which is capable of solving optimization problems on unbounded domains, rather than the projected D-SGD; and (2) our stability and generalization bounds are non-vacuous, even in the cases where the spectral gap $1-\lambda$ is sufficiently close to 0, which characterizes the cases where the worker number is sufficiently large or the communication graph is sufficiently sparse.

2. Related Work

The earliest work of classical decentralized optimization can be traced back to Tsitsiklis (1984), Tsitsiklis et al. (1986) and Nedic & Ozdaglar (2009a). D-SGD has been extended to various settings in deep learning, including time-varying topologies (Lu & Wu, 2020; Koloskova et al., 2020), asynchronous settings (Lian et al., 2018; Xu et al., 2021; Nadiradze et al., 2021), directed topologies (Assran et al., 2019a; Taheri et al., 2020), and data-heterogeneous scenarios (Tang et al., 2018; Vogels et al., 2021). It has been proved that the convergence of D-SGD heavily relies on the communication topology (Hambrick et al., 1996; Bianchi & Jakubowicz, 2012; Lian et al., 2017; Assran et al., 2019b; Wang et al., 2019; Guo et al., 2020), especially in the scenarios where the local data is heterogeneous across workers (Yuan et al., 2020; Koloskova et al., 2020; Dai et al., 2022; Bars et al., 2022; Huang et al., 2022). However, the impact of the communication topology on the generalizability of D-SGD is still in its infancy.

Recently, inspiring work by Zhang et al. (2021) gives insights to how gossip communication in D-SGD promotes generalization in large batch settings. They prove that a self-adjusting noise exists in D-SGD, which may help D-SGD find flatter minima with better generalization. Another work by Richards et al. (2020) presents a generalization bound of the Adaptation-Then-Combination (ATC) version of D-SGD through algorithmic stability and Rademacher complexity (Mohri et al., 2018) in both smooth and non-smooth settings. However, their generalization bounds are invariant to the communication topology, which contradicts the experimental results (see Figure 3). In contrast, our generalization bounds are topology-aware and characterize the effects of decentralization on generalization.

3. Preliminaries

Supervised learning. Supposed $\mathcal{X} \subseteq \mathbb{R}^{d_x}$ and $\mathcal{Y} \subseteq \mathbb{R}$ are the input and output spaces, respectively. We denote the training set as $\mathcal{S} = \{z_1, \ldots, z_N\}$, where $z_\zeta = (x_\zeta, y_\zeta), \zeta = 1, \ldots, N$ are sampled independent and identically distributed (i.i.d.) from an unknown data distribution $\mathcal{D}$ defined on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$.

The goal of supervised learning is to learn a predictor (hypothesis) $g(\cdot; \mathbf{w})$, parameterized by $\mathbf{w} = \mathbf{w}(z_1, z_2, \ldots, z_N) \in \mathbb{R}^d$, to approximate the mapping between the input variable $x \in \mathcal{X}$ and the output variable $y \in \mathcal{Y}$, based on the training set $\mathcal{S}$. Let $c : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}^+$ be a loss function that evaluates the prediction performance of the hypothesis g. The loss of a hypothesis g with respect to (w.r.t.) the example $z_\zeta = (x_\zeta, y_\zeta)$ is denoted by $f(\mathbf{w}; z_\zeta) = c(g(x_\zeta; \mathbf{w}), y_\zeta)$, in order to measure the effectiveness of the learned model. Then, the empirical and population risks of $\mathbf{w}$ are defined as follows:

$$F_\mathcal{S}(\mathbf{w}) = \frac{1}{N}\sum_{\zeta=1}^{N} f(\mathbf{w}; z_\zeta), \quad F(\mathbf{w}) = \mathbb{E}_{z\sim D}[f(\mathbf{w}; z)].$$

Distributed learning. Distributed learning jointly trains a learning model $\mathbf{w}$ on multiple workers (Shamir & Srebro, 2014). In this framework, the k-th worker $(k = 1, \ldots, m)$ can access n_k independent and identically distributed (i.i.d.)

[1]Throughout this work, we use the term *worker* to represent the local model.

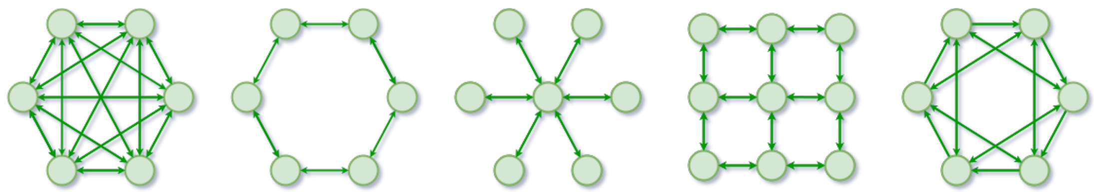


Figure 1. Illustration of some commonly-used communication topologies.

training examples $\mathcal{S}_k = \{z_{k,1}, \ldots, z_{k,n_k}\}$, drawn from the data distribution $\mathcal{D}$. If set $n_k = n$, the total sample size will be $N = nm$. In this case, the global empirical risk of $\mathbf{w}$ is defined as:

$$*\hat{F}(\mathbf{w}) = \frac{1}{m}\sum_{k=1}^{m} F_{\mathcal{S}_k}(\mathbf{w}),$$

where $F_{\mathcal{S}_k}(\mathbf{w}) = \frac{1}{n}\sum_{\zeta=1}^{n} f(\mathbf{w}; z_{k,\zeta})$ denotes the local empirical risk on the k-th worker and $z_{k,\zeta} \in \mathcal{S}_k$ $(\zeta = 1, \ldots, n)$ is the local sample set.

Decentralized Stochastic Gradient Descent (D-SGD). The goal of D-SGD is to learn a consensus model $\overline{\mathbf{w}} = \frac{1}{m}\sum_{k=1}^{m} \mathbf{w}_k$, on m workers, where $\mathbf{w}_k$ denotes the local model on the k-th worker. For any k, let $\mathbf{w}_k^{(t)} \in \mathbb{R}^d$ be the d-dimensional local model on the k-th worker in the t-th iteration, while $\mathbf{w}_k^{(1)} = 0$ is the initial point. We denote $\mathbf{P}$ as a doubly stochastic gossip matrix that characterizes the underlying topology $\mathcal{G}$ (see Definition A.5 and Figure 1). The intensity of gossip communications is measured by the spectral gap (Seneta, 2006) of $\mathbf{P}$ (i.e., $1 - \max\{|\lambda_2|, |\lambda_n|\}$, where λ_i $(i = 2, \ldots, m)$ denotes the i-th largest eigenvalue of $\mathbf{P}$ (see Definition A.6). The vanilla Adapt-While-Communicate (AWC) version of D-SGD without projecting operations updates the model on the k-th worker by

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

where $\{\eta_t\}$ is a sequence of positive learning rates, and $\nabla f(\mathbf{w}_k^{(t)}; z_{k,\zeta_t}^{(t)})$ is the gradient of f w.r.t. the first argument on the k-th worker, and ζ_t is i.i.d. variable drawn from the uniform distribution over $\{1, \ldots, n\}$ at the t-th iteration (Lian et al., 2017). In this paper, matrix $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_m]^T \in \mathbb{R}^{m \times d}$ stands for all local models across the network, while matrix $\nabla f(\mathbf{W}; \mathbf{Z}) = [\nabla f(\mathbf{w}_1; z_1), \cdots, \nabla f(\mathbf{w}_m; z_m)]^T \in \mathbb{R}^{m \times d}$ stacks all local gradients w.r.t. the first argument. In this way, the matrix form of Equation (1) is as follows:

$$\mathbf{W}^{(t+1)} = \mathbf{P}\mathbf{W}^{(t)} - \eta_t \nabla f(\mathbf{W}^{(t)}; \mathbf{Z}_{\zeta_t}^{(t)}).$$

4. Topology-aware Generalization Bounds of D-SGD

This section proves stability and generalization bounds for D-SGD. We start with the definition of a new parameter-level stability for distributed settings. Then, the stability of D-SGD under a non-smooth condition is obtained (see Theorem 1 and Corollary 2). This implies a connection between stability and generalization in expectation (see Lemma 3), which suffices to prove the expected generalization bound of D-SGD, of order $\mathcal{O}(1/N + ((m^{-1}\lambda^2)^{\frac{\alpha}{2}} + m^{-\alpha})/N^{1-\frac{\alpha}{2}})$.

4.1. Algorithmic Stability of D-SGD

Understanding generalization using algorithmic stability can be traced back to Bousquet & Elisseeff (2002) and Shalev-Shwartz et al. (2010), and has been applied to stochastic gradient methods (Hardt et al., 2016; Lei & Ying, 2020). For more details, please see Appendix B.

We define a new algorithmic stability of distributed optimization algorithms below, which better characterizes the on-average sensitivity of models across multiple workers.

Definition 1 (Distributed On-average Stability)**.** *Let $\mathcal{S}_k = \{z_{k,1}, \ldots, z_{k,n}\}$ denote the i.i.d. local samples on the k-th worker drawn from the distribution $\mathcal{D}$ $(k = 1, \ldots, m)$. $\mathcal{S}_k^{(i)} = z_{k,1}, \ldots, \tilde{z}_{k,i}, \ldots, z_{k,n}\} \in \mathcal{Z}^n$ is formed by replacing the i-th element of $\mathcal{S}_k$ with a local sample $\tilde{z}_{k,i}$ drawn from the distribution $\mathcal{D}$. We denote $\mathbf{w}_k$ and $\widetilde{\mathbf{w}}_k$ as the weight vectors on the k-th worker produced by the stochastic algorithm A based on $\mathcal{S}_k$ and $\mathcal{S}_k^{(i)}$, respectively. A is ℓ_2 distributed on-average ϵ-stable for all training data sets $\mathcal{S}_k$ and $\mathcal{S}_k^{(i)}$ $(k = 1 \ldots m)$ if*

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

where $\mathbb{E}_A[\cdot]$ stands for the expectation w.r.t. the randomness of the algorithm A (see more details in Appendix A).

We then prove that D-SGD is distributed on-average stable.

Theorem 1. *Let $\mathcal{S}_k$ and $\mathcal{S}_k^{(i)}$ be constructed in Definition 1. Let $\mathbf{w}_k^{(t)}$ and $\widetilde{\mathbf{w}}_k^{(t)}$ be the t-th iteration on the k-th worker* *produced by Equation (1) based on $\mathcal{S}_k$ and $\mathcal{S}_k^{(i)}$ respectively, and $\{\eta_t\}$ be a non-increasing sequence of positive learning rates. We assume that for all $z \in \mathcal{Z}$, the function $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is non-negative and convex with its gradient $\nabla f(\mathbf{w}; z)$ being (α, L)-Hölder continuous (see Assumption A.3). We further assume that the weight differences at the t-th iteration are multivariate normally distributed: $\mathbf{w}_k^{(t)} - \widetilde{\mathbf{w}}_k^{(t)} \overset{i.i.d.}{\sim} \mathcal{N}(\mu_{t,k}, \sigma_{t,k}^2 I_d)$ for all k where d denotes the dimension of weights, with **unknown** parameters $\mu_{t,k}$ and $\sigma_{t,k}$ satisfying some technical conditions (see Assumption A.4). Then we have the following:*

$$\frac{1}{mn}\sum_{i=1}^{n}\sum_{k=1}^{m}\mathbb{E}_{\mathcal{S}_k,\mathcal{S}_k^{(i)},A}\left[\left\|\mathbf{w}_k^{(t+1)} - \widetilde{\mathbf{w}}_k^{(t+1)}\right\|_2^2\right]$$
$$\leq \sum_{\tau=0}^{t} C^{t-\tau}\{\underbrace{\mathcal{O}\big((1-\frac{1}{m})\boldsymbol{\lambda}^{\mathbf{2}} + \frac{1}{m}\big)}_{\text{Error from decentralization}}$$
$$\underbrace{+\,\mathcal{O}(\frac{{\eta_\tau}^2}{n}\cdot\frac{1}{m}\sum_{k=1}^{m}\mathbb{E}_{\mathcal{S}_k,A}\big[F_{\mathcal{S}_k}^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_k^{(\tau)})\big])\},}_{\mathcal{O}(\frac{m}{N})\,\cdot\,\text{Averaged empirical risk}}$$

where $C = 2\eta_0 L(1-\frac{1}{n})$ and $F_{\mathcal{S}_k}(\mathbf{w}_k^{(\tau)})$ is the local empirical risk of the k-th worker at iteration τ.

Theorem 1 suggests that the distributed on-average stability of D-SGD is positively related to the spectral gap of the underlying topology and the accumulation of the averaged empirical risk. The proof is provided in Appendix D.2.

We can obtain a simplified result with fixed learning rates.

Corollary 2 (Stability in Expectation with $\eta_t \equiv \eta$). *Suppose all the assumptions of Theorem 1 hold. With a fixed learning rate $\eta_t \equiv \eta \leq \frac{1}{2L}(1-\frac{2}{m})$, the distributed on-average stability of D-SGD can be bounded as*

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

where $\epsilon_{\mathcal{S}}$ denotes the upper bound of averaged empirical risk $\frac{1}{m}\sum_{k=1}^{m}\mathbb{E}_{\mathcal{S}_k,A}\big[F_{\mathcal{S}_k}^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_k^{(t)})\big]$.

Corollary 2 shows that the distributed on-averge stability of D-SGD is of the order $\mathcal{O}(m/N{+}1/m{+}\lambda^2)$. We defer the proof to Appendix D.2.

Comparison with existing results. Compared with Sun et al. (2021), we relax the restrictive bounded gradient and the smoothness assumptions. Instead, a much weaker

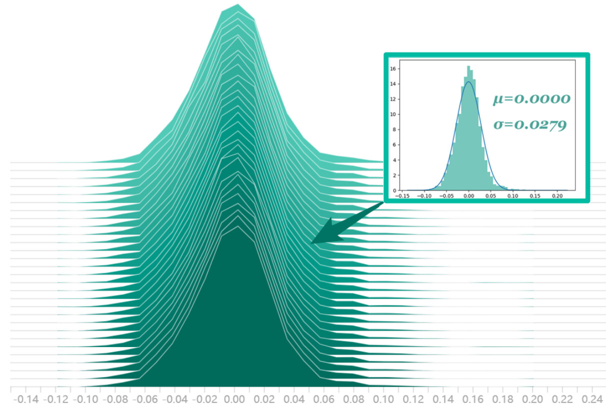


Figure 2. Histograms of the weight differences of the last layers of the ResNet-18 models (1024 dimensions $\times$ 10 classes=10240 parameters) trained by AWC D-SGD on $\mathcal{S}$ and $\mathcal{S}^{(i)}$ that differ by only one data point. Thirty ResNet-18 models are trained on data sampled from the MNIST dataset (LeCun et al., 1998), each model is trained with 16 workers for 3000 iterations.

Hölder condition (see Assumption A.3) is adopted. We make a mild assumption that the weight difference $\big(\mathbf{w}_k^{(t)} - \widetilde{\mathbf{w}}_k^{(t)}\big)$ is multivariate normally distributed (see Assumption A.4), which stems from our empirical observations: Figure 2 illustrates that the distribution of the weight differences in ResNet-18 models trained by D-SGD is close to a centered Gaussian. Intuitively, the assumption is based upon the fact that the weights of the consensus model are very insensitive to the change of a single data point.

We also compare the order of the derived bound with the existing literature. Hardt et al. (2016) proves that SGD is $\mathcal{O}(\sum_{\tau=1}^{t-1}\eta_\tau/n)$-stable in convex and smooth settings, which corresponds to the $\mathcal{O}(\frac{m}{N})$ term in Corollary 2. Under Hölder continuous condition, Lei & Ying (2020) proposes a parameter-level stability bound of SGD of the order $\mathcal{O}(\frac{\epsilon_{\mathcal{S}}\eta^2}{n}{+}\eta^{\frac{2}{1-\alpha}})$. In contrast, Corollary 2 shows that D-SGD suffers from additional terms $\mathcal{O}\big((1{-}\frac{1}{m})\lambda^2{+}\frac{1}{m}\big)$, where the first term $\mathcal{O}\big((1{-}\frac{1}{m})\lambda^2\big)$ can characterize the degree of disconnection of the underlying communication topology. Close work by Sun et al. (2021) proved that the stability of the projected variant of D-SGD is bounded by $\mathcal{O}\left(\frac{\eta t B^2}{N}{+}\frac{\eta t B^2}{1-\lambda}\right)$ in the convex smooth setting, where B is the upper bound of the gradient norm. The term $\mathcal{O}(\frac{\eta t B^2}{1-\lambda})$ brought by decentralization is of the order $\mathcal{O}(m^2B^2)$ for ring topologies and $\mathcal{O}(mB^2)$ for grids, respectively. Our error bound in Corollary 2 is tighter than their results, since

$$\frac{1}{1-2\eta L(1-\frac{1}{n})}\cdot\mathcal{O}\big((1-\frac{1}{m})\lambda^2 + \frac{1}{m}\big)$$
$$\leq \mathcal{O}\big(m\big) \ll \mathcal{O}(mB^2) \leq \mathcal{O}(m^2B^2).$$

4.2. Generalization bounds of D-SGD

The following lemma bridges the gap between generalization and the newly proposed distributed on-average stability.

Lemma 3 (Generalization via Distributed On-average Stability)**.** *Let $\mathcal{S}_k$ and $\mathcal{S}_k^{(i)}$ be constructed in Definition 1. If for any z, the pre-specified function $f(\mathbf{w}; z)$ is non-negative and convex, with its gradient $\nabla f(\mathbf{w}; z)$ being (α, L)-Hölder continuous, then*

$$\mathbb{E}_{\mathcal{S},A}\big[F(\overline{\mathbf{w}}^{(t)}) - F_{\mathcal{S}}(\overline{\mathbf{w}}^{(t)})\big] \le \frac{L}{mn^{1-\frac{\alpha}{2}}}\Big\{\frac{1}{mn}\sum_{k=1}^{m}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{S}_k,\mathcal{S}_k^{(i)},A}\big[\|\mathbf{w}_k^{(t)} - \widetilde{\mathbf{w}}_k^{(t)}\|_2^2\big]\Big\}^{\frac{\alpha}{2}},$$

where $\overline{\mathbf{w}}^{(t)} = \frac{1}{m}\sum_{k=1}^{m}\mathbf{w}_k^{(t)}$.

We give the proof in Appendix D.3.

Lemma 3 suggests that if the consensus model learned by the distributed SGD A is ϵ-stable in the sense of Definition 1, the generalization error of the consensus model is bounded by $L\epsilon^{\frac{\alpha}{2}}/(mn^{1-\frac{\alpha}{2}})$. Lemma 3 improves Theorem 2 (c) of Lei & Ying (2020) by removing the $\mathcal{O}(\mathbb{E}_{S,A}\big[F^{\frac{2\alpha}{1+\alpha}}(A(S))\big])$ term, where $\mathbb{E}_{S,A}\big[F(A(S))\big]$ denotes the population risk of the learned model $A(\mathcal{S})$ (see Equation (D.26)). This improvement is significant, because $\mathbb{E}_{S,A}\big[F(A(S))\big]$ usually does not converge to zero in practice.

We now prove the generalization bound of D-SGD.

Theorem 4 (Generalization Bound in Expectation with $\eta_t \equiv \eta$)**.** *Let all the assumptions of Theorem 1 hold. With a fixed step sizes of $\eta_t \equiv \eta \le 2L(1-\frac{2}{m})$, the generalization error of the consensus model learned by D-SGD can be controlled as*

$$\mathbb{E}_{\mathcal{S},A}\big[F(\overline{\mathbf{w}}^{(t)}) - F_{\mathcal{S}}(\overline{\mathbf{w}}^{(t)})\big] \le \frac{L}{[1-2\eta L(1-\frac{1}{n})]^{\frac{\alpha}{2}}}\Big\{\mathcal{O}\big(\underbrace{\frac{\epsilon_{\mathcal{S}}^{\frac{\alpha}{2}}}{N}}_{\mathcal{O}(\frac{1}{N})} + \underbrace{\frac{n^{\frac{\alpha}{2}}}{N}\big((1-\frac{1}{m})\boldsymbol{\lambda}^{\mathbf{2}}+\frac{1}{m}\big)^{\frac{\alpha}{2}}}_{\textit{Error from decentralization}}\big)\Big\},$$

where $\overline{\mathbf{w}}^{(t)} = \frac{1}{m}\sum_{k=1}^{m}\mathbf{w}_k^{(t)}$ and $\epsilon_{\mathcal{S}}$ is the upper bounder of $\frac{1}{m}\sum_{k=1}^{m}\mathbb{E}_{\mathcal{S}_k,A}\big[F_{\mathcal{S}_k}^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_k^{(t)})\big]$.

The order of the generalization bound in Theorem 4 is $\mathcal{O}(1/N + ((m^{-1}\lambda^2)^{\frac{\alpha}{2}} + m^{-\alpha})/N^{1-\frac{\alpha}{2}})$ and will become $\mathcal{O}(1/N + (m^{-\frac{1}{2}}\lambda^2 + m^{-1})/N^{-\frac{1}{2}})$ in the smooth settings where $\alpha = 1$. The proof is provided in Appendix D.3.

Remark 1. *Corollary 2 and Theorem 4 indicate that the stability and generalization of D-SGD are positively related to the spectral gap $1-\lambda$. The intuition of the results is that D-SGD with a denser connection topology (i.e., larger λ) can aggregate more information from its neighbors, thus "indirectly" accessing more data at each iteration, leading to better generalization.*

Table 1. Spectral gap of gossip matrices with different topology.

Graph topoloy	**Spectral gap** $1-\lambda$
Fully-connected	1
Disconnected	0
Ring	$\approx 16\pi^2/3m^2$
Grid	$\mathcal{O}(1/(m\log_2(m)))$
Exponential graph (even)	$2/(1+\log_2(m))$

4.3. Practical Implications

Our theory delivers significant practical implications.

Communication topology and generalization. The intensity of communication is controlled by the spectral gap $1-\lambda$ of the underlying communication topologies (see Table 1). Detailed analyses of the spectral gaps of some commonly-used topologies can be found in Proposition 5 of Nedić et al. (2018) and Ying et al. (2021a). Substituting the spectral gap of different topologies in Table 1 into Theorem 4, we can conclude that the generalization error of different topologies can be ranked as follows: fully-connected < exponential < grid < ring, since

$$0 < 1 - \mathcal{O}((\log_2(m))^{-1}) < 1 - \mathcal{O}((m\log_2(m))^{-1}) < 1 - \mathcal{O}(m^{-2}).$$

On the one hand, our theory provides theoretical evidence that D-SGD will generalize better on well-connected topologies (i.e., topologies with larger spectral gap). On the other hand, we prove that for a specific topology, the worker number impacts the generalization of D-SGD through affecting the spectral gap of the topology.

Consensus distance control. Recently, a line of studies have been devoted to understanding the connection between optimization and generalization through studying the effect of early phase training (Keskar et al., 2017; Achille et al., 2018; Frankle et al., 2020). In the decentralized settings, Kong et al. (2021) claims that there exists a "critical consensus distance" in the initial training phase—consensus distance (i.e. $\frac{1}{m}\sum_{i=1}^{m}\|\mathbf{w}_k^{(t)} - \frac{1}{m}\sum_{k=1}^{m}\mathbf{w}_k^{(t)}\|_F^2$) below the critical threshold will ensure good generalization. However, the reason why consensus control can promote generalization remains an open problem. Fortunately, the following theorem can explain this phenomenon by connecting the consensus distance notion in Kong et al. (2021) with the algorithmic stability and the generalizability of D-SGD.

Corollary 5. *Let all the assumptions of Theorem 1 plus Assumption A.1 and Assumption A.2 hold. Suppose that the consensus distance satisfies the condition $\Gamma^2 \le \frac{1}{m}\sum_{k=1}^{m}\|\mathbf{w}_k^{(t)} - \overline{\mathbf{w}}^{(t)}\|_2^2 \le K^2$ if $t \le t_\Gamma$, and is controlled below Γ^2 from t_Γ-th iterate to the end of training. Then one can conclude that the upper bound of the distributed* *on-average stability of D-SGD increase monotonically with* t_Γ *if the total number of iterations* $t \geq \frac{-C}{2\ln C}$.

We give the proof in Appendix D.4.

Corollary 5 provides theoretical evidence for the following empirical findings: (1) consensus control is beneficial to the algorithmic stability and the generalizability of D-SGD; and (2) it is more effective to control the consensus distance in the initial stage of training than at the end of training.

5. Empirical Results

This section empirically validates our theoretical results. We first introduce the experimental setup and then study how the communication topology and the worker number affect the generalization of D-SGD. The code is available at https://github.com/Raiden-Zhu/Generalization-of-DSGD.

5.1. Experimental Setup

Networks and datasets. Network architectures VGG-11 (Simonyan & Zisserman, 2014) and ResNet-18 (He et al., 2016b) are employed in our experiments. The models are trained on CIFAR-10, CIFAR-100 (Krizhevsky et al., 2009) and Tiny ImageNet (Le & Yang, 2015), three popular benchmark image classification datasets. The CIFAR-10 dataset consists of 60,000 32×32 color images across 10 classes, with each class containing 5,000 training and 1,000 testing images. The CIFAR-100 dataset also consists of 60,000 32×32 color images, except that it has 100 classes, each class containing 5,00 training and 1,00 testing images. Tiny ImageNet contains 120,000 64× 64 color images in 200 classes, each class containing 500 training images, 50 validation images, and 50 test images. No other pre-processing methods are employed.

Training setting. Vanilla D-SGD is employed to train image classifiers based on VGG-11 and ResNet-18 on fully-connected, ring, grid, and static exponential topologies. The number of workers is set as 32 and 64. Batch normalization (Ioffe & Szegedy, 2015) and dropout (Srivastava et al., 2014) are employed in training VGG-11. The local batch size is set as 64. To control the impact of different total batch size (local batch size × worker number) caused by the different number of workers, we apply the linear scaling law (i.e., linearly increase learning rate w.r.t. total batch size) (He et al., 2016a; Goyal et al., 2017). The initial learning rate is set as 0.1 and will be divided by 10 when the model has accessed $2/5$ and $4/5$ of the total number of iterations (He et al., 2016a). All other techniques, including momentum (Qian, 1999), weight decay (Tihonov, 1963), and data augmentation (LeCun et al., 1998) are disabled.

Implementations. All our experiments are conducted on a computing cluster with GPUs of NVIDIA® Tesla™ V100 16GB and CPUs of Intel® Xeon® Gold 6140 CPU @ 2.30GHz. Our code is implemented based on PyTorch (Paszke et al., 2019).

5.2. Communication topology and generalization

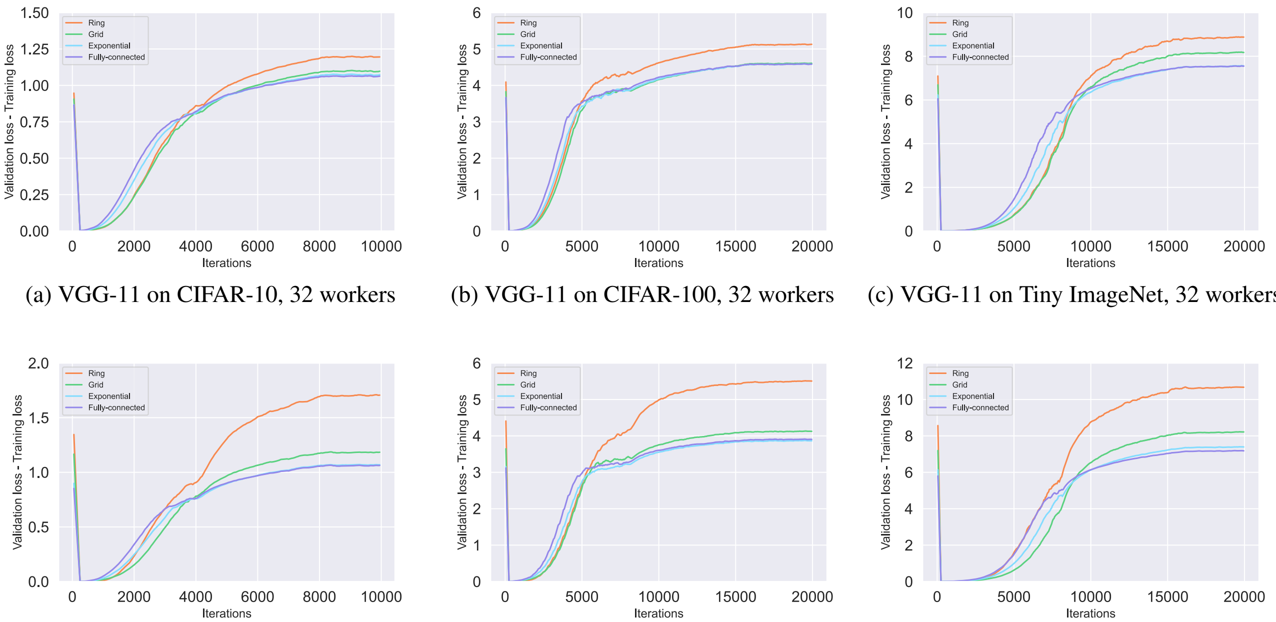
We calculate the difference between the validation loss and the training loss on different topologies separately, as shown in Figure 3 and Figure C.1. Two observations are obtained from those figures: (1) for large topologies, the loss differences can be sorted as follows: fully-connected $\approx$ exponential $<$ grid $<$ ring; (2) as the worker number increases, the loss differences of D-SGD on different topologies increase. These observations suggest that (1) D-SGD generalizes better on well-connected topologies with a larger spectral gap; (2) the generalizability gap of D-SGD on different topologies grows as the worker number increases.

In comparison to ResNet-18, we observe when VGG-11 is chosen as the backbone, the generalization gaps between different topologies are larger (see Figure 3 and Figure C.1). We conjecture that the generalization gaps are amplified because the Hölder constant L of VGG-11 is larger than that of ResNet-18, according to our theory suggesting that the generalization error of D-SGD linearly increases with L (see Theorem 4).

One may observe that the loss difference of the fully-connected topology is larger than that of other topologies in the initial training phase, which seems to be inconsistent with our theory. Theoretically, explaining this phenomenon is an optimization problem, which is beyond the scope of this work. One possible explanation is motivated by the stability-convergence trade-off in iterative optimization algorithms (Chen et al., 2018). Since the fully-connected system converges faster, the corresponding optimization error is smaller than the other three kinds of systems at the beginning of training. Therefore, the fully-connected system is less stable in the initial phase of training.

6. Future Works

Generalization of D-SGD with non-i.i.d. data. In real-world settings, a fundamental challenge is that data may not be i.i.d. across the workers. In this case, different workers may collect very distinct or even contradictory samples (i.e., data-heterogeneity) (Criado et al., 2021). However, generalization analyses of distributed learning algorithms are mostly based on the key assumption that the data is i.i.d. over workers. Therefore, more sophisticated techniques are needed to broaden our knowledge on the following questions: Would the consensus model learned by D-SGD generalize well in the non-i.i.d. settings? Would the effect of the communication topology on generalization be reduced



(a) VGG-11 on CIFAR-10, 32 workers (b) VGG-11 on CIFAR-100, 32 workers (c) VGG-11 on Tiny ImageNet, 32 workers

(d) VGG-11 on CIFAR-10, 64 workers (e) VGG-11 on CIFAR-100, 64 workers (f) VGG-11 on Tiny ImageNet, 64 workers

Figure 3. Loss differences in training VGG-11 using D-SGD with different topologies.

or amplified in these scenarios?

Implicit bias of D-SGD. As pointed out in Zhang et al. (2021), the additional gradient noise in D-SGD helps it converge to a flatter minima compared to centralized distributed SGD. Therefore, a direct question is whether there is a superior implicit bias effect (Soudry et al., 2018; Ji & Telgarsky, 2019; Arora et al., 2019; Wang et al., 2021) in D-SGD compared to centralized distributed SGD, which involves the convergence direction? Can we theoretically analyze the implicit bias of D-SGD and derive fine-grained generalization bounds of D-SGD?

7. Conclusion

In this paper, we analyze the algorithmic stability and generalizability of decentralized stochastic gradient descent (D-SGD). We prove that the consensus model learned by D-SGD is $\mathcal{O}(m/N+1/m+\lambda^2)$-stable, where N is the total sample size of the whole system, m is the worker number, and $1-\lambda$ is the spectral gap of the communication topology. Based on this stability result, we obtain an $\mathcal{O}(1/N + ((m^{-1}\lambda^2)^{\frac{\alpha}{2}} + m^{-\alpha})/N^{1-\frac{\alpha}{2}})$ in-average generalization bound, characterizing the gap between the training performance and the test performance. Our error bounds are non-vacuous, even when the worker number is sufficiently large, or the communication graph is sufficiently sparse. According to our theory, we can conclude: (1) the generalizability of D-SGD is positively correlated with the spectral gap of the underlying topology; (2) the generalizability of D-SGD decreases when the worker number increases. These theoretical findings are empirically justified by the experiments of VGG-11 and ResNet-18 on CIFAR-10, CIFAR-100, and Tiny ImageNet. The theory can also explain why consensus control at the beginning of training can promote the generalizability of D-SGD. To our best knowledge, this is the first study on the topology-aware generalizations of vanilla D-SGD.

Acknowledgement

This work is supported by the Major Science and Technology Innovation 2030 key projects "New Generation Artificial Intelligence" (No. 2021ZD0111700), the National Natural Science Foundation of China (No. 61976186, U20B2066, and 61932016), the Key Research and Development Program of Zhejiang Province (No. 2021C01164), the Fundamental Research Funds for the Central Universities (No. 226-2022-00064 and WK2150110024), and the Starry Night Science Fund of Zhejiang University Shanghai Institute for Advanced Study (No. SN-ZJU-SIAS-001). This work is completed when Tongtian Zhu and Zhengyang Niu were interns at JD Explore Academy.

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.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} symposium on operating systems design and implementation*, 2016.

Achille, A., Rovere, M., and Soatto, S. Critical learning periods in deep networks. In *International Conference on Learning Representations*, 2018.

Alghunaim, S. A. and Yuan, K. A unified and refined convergence analysis for non-convex decentralized learning. *arXiv preprint arXiv:2110.09993*, 2021.

Arora, S., Du, S. S., Hu, W., Li, Z., Salakhutdinov, R. R., and Wang, R. On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems*, 2019.

Assran, M., Loizou, N., Ballas, N., and Rabbat, M. Stochastic gradient push for distributed deep learning. In *Proceedings of the 36th International Conference on Machine Learning*, 2019a.

Assran, M., Loizou, N., Ballas, N., and Rabbat, M. Stochastic gradient push for distributed deep learning. In *Proceedings of the 36th International Conference on Machine Learning*, 2019b.

Bars, B. L., Bellet, A., Tommasi, M., and Kermarrec, A.-M. Yes, topology matters in decentralized optimization: Refined convergence and topology learning under heterogeneous data. *arXiv preprint arXiv:2204.04452*, 2022.

Bassily, R., Feldman, V., Guzmán, C., and Talwar, K. Stability of stochastic gradient descent on nonsmooth convex losses. In *Advances in Neural Information Processing Systems*, 2020.

Bianchi, P. and Jakubowicz, J. Convergence of a multi-agent projected stochastic gradient algorithm for non-convex optimization. *IEEE transactions on automatic control*, 2012.

Bottou, L. and Bousquet, O. The tradeoffs of large scale learning. In Platt, J., Koller, D., Singer, Y., and Roweis, S. (eds.), *Advances in Neural Information Processing Systems*, 2008.

Bousquet, O. and Elisseeff, A. Stability and generalization. *Journal of Machine Learning Research*, 2002.

Chen, Y., Jin, C., and Yu, B. Stability and convergence tradeoff of iterative optimization algorithms. *arXiv preprint arXiv:1804.01619*, 2018.

Criado, M. F., Casado, F. E., Iglesias, R., Regueiro, C. V., and Barro, S. Non-iid data and continual learning processes in federated learning: A long road ahead. *arXiv preprint arXiv:2111.13394*, 2021.

Dai, R., Shen, L., He, F., Tian, X., and Tao, D. Dispfl: Towards communication-efficient personalized federated learning via decentralized sparse training. In *International Conference on Machine Learning*, 2022.

Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Mao, M., Ranzato, M., Senior, A., Tucker, P., Yang, K., et al. Large scale distributed deep networks. *Advances in neural information processing systems*, 2012.

Deng, Z., He, H., and Su, W. Toward better generalization bounds with locally elastic stability. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Devroye, L. and Wagner, T. Distribution-free performance bounds with the resubstitution error estimate. *IEEE Transactions on Information Theory*, 1979.

Fazlyab, M., Robey, A., Hassani, H., Morari, M., and Pappas, G. Efficient and accurate estimation of lipschitz constants for deep neural networks. In *Advances in Neural Information Processing Systems*, 2019.

Foster, D. J., Greenberg, S., Kale, S., Luo, H., Mohri, M., and Sridharan, K. Hypothesis set stability and generalization. In *Advances in Neural Information Processing Systems*, 2019.

Frankle, J., Schwab, D. J., and Morcos, A. S. The early phase of neural network training. In *International Conference on Learning Representations*, 2020.

Geiping, J., Bauermeister, H., Dröge, H., and Moeller, M. Inverting gradients - how easy is it to break privacy in federated learning? In *Advances in Neural Information Processing Systems*, 2020.

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., and He, K. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

Guo, Z., Liu, M., Yuan, Z., Shen, L., Liu, W., and Yang, T. Communication-efficient distributed stochastic auc maximization with deep neural networks. In *International Conference on Machine Learning*, pp. 3864–3874. PMLR, 2020.

Hambrick, D. C., Cho, T. S., and Chen, M.-J. The influence of top management team heterogeneity on firms' competitive moves. *Administrative science quarterly*, 1996.

Hardt, M., Recht, B., and Singer, Y. Train faster, generalize better: Stability of stochastic gradient descent. In *Proceedings of The 33rd International Conference on Machine Learning*, 2016.

He, F. and Tao, D. Recent advances in deep learning theory. *arXiv preprint arXiv:2012.10931*, 2020.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016a.

He, K., Zhang, X., Ren, S., and Sun, J. Identity mappings in deep residual networks. In *European conference on computer vision*. Springer, 2016b.

Huang, T., Liu, S., Shen, L., He, F., Lin, W., and Tao, D. Achieving personalized federated learning with sparse local models. *arXiv preprint arXiv:2201.11380*, 2022.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, 2015.

Ji, Z. and Telgarsky, M. Gradient descent aligns the layers of deep linear networks. In *International Conference on Learning Representations*, 2019.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawit, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Eichner, H., El Rouayheb, S., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., Gruteser, M., Harchaoui, Z., He, C., He, L., Huo, Z., Hutchinson, B., Hsu, J., Jaggi, M., Javidi, T., Joshi, G., Khodak, M., Konecný, J., Korolova, A., Koushanfar, F., Koyejo, S., Lepoint, T., Liu, Y., Mittal, P., Mohri, M., Nock, R., Özgür, A., Pagh, R., Qi, H., Ramage, D., Raskar, R., Raykova, M., Song, D., Song, W., Stich, S. U., Sun, Z., Theertha Suresh, A., Tramèr, F., Vepakomma, P., Wang, J., Xiong, L., Xu, Z., Yang, Q., Yu, F. X., Yu, H., and Zhao, S. *Advances and Open Problems in Federated Learning*. IEEE, 2021.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. The early phase of neural network training. In *International Conference on Learning Representations*, 2017.

Koloskova, A., Loizou, N., Boreiri, S., Jaggi, M., and Stich, S. A unified theory of decentralized SGD with changing topology and local updates. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Kong, L., Lin, T., Koloskova, A., Jaggi, M., and Stich, S. Consensus control for decentralized deep learning. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images (tech. rep.). *University of Toronto*, 2009.

Kuzborskij, I. and Lampert, C. Data-dependent stability of stochastic gradient descent. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Le, Y. and Yang, X. Tiny imagenet visual recognition challenge. *CS 231N*, 2015.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.

Lei, Y. and Ying, Y. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Li, M., Andersen, D. G., Smola, A. J., and Yu, K. Communication efficient distributed machine learning with the parameter server. *Advances in Neural Information Processing Systems*, 2014.

Lian, X., Zhang, C., Zhang, H., Hsieh, C.-J., Zhang, W., and Liu, J. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. In *Advances in Neural Information Processing Systems*, 2017.

Lian, X., Zhang, W., Zhang, C., and Liu, J. Asynchronous decentralized parallel stochastic gradient descent. In *International Conference on Machine Learning*, 2018.

Liu, T., Lugosi, G., Neu, G., and Tao, D. Algorithmic stability and hypothesis complexity. In *International Conference on Machine Learning*, 2017.

Lopes, C. G. and Sayed, A. H. Diffusion least-mean squares over adaptive networks: Formulation and performance analysis. *IEEE Transactions on Signal Processing*, 2008.

Lu, J., Tang, C. Y., Regier, P. R., and Bow, T. D. Gossip algorithms for convex consensus optimization over networks. *IEEE Transactions on Automatic Control*, 2011.

Lu, S. and Wu, C. W. 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)*, 2020.

Lu, Y. and De Sa, C. Optimal complexity in decentralized training. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of machine learning*. MIT press, 2018.

Montenegro, R. and Tetali, P. Mathematical aspects of mixing times in markov chains. *Foundations and Trends® in Theoretical Computer Science*, 2006.

Nadiradze, G., Sabour, A., Davies, P., Li, S., and Alistarh, D. Asynchronous decentralized sgd with quantized and local updates. *Advances in Neural Information Processing Systems*, 2021.

Nedic, A. Distributed gradient methods for convex machine learning problems in networks: Distributed optimization. *IEEE Signal Processing Magazine*, 2020.

Nedic, A. and Ozdaglar, A. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 2009a.

Nedic, A. and Ozdaglar, A. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 2009b.

Nedić, A., Olshevsky, A., and Rabbat, M. G. Network topology and communication-computation tradeoffs in decentralized optimization. *Proceedings of the IEEE*, 2018.

Nesterov, Y. *Introductory lectures on convex optimization: A basic course*. Springer Science & Business Media, 2003.

Nesterov, Y. Universal gradient methods for convex optimization problems. *Mathematical Programming*, 2015.

Neu, G., Dziugaite, G. K., Haghifam, M., and Roy, D. M. Information-theoretic generalization bounds for stochastic gradient descent. In *Proceedings of Thirty Fourth Conference on Learning Theory*, 2021.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 2019.

Qian, N. On the momentum term in gradient descent learning algorithms. *Neural networks*, pp. 145–151, 1999.

Richards, D. et al. Graph-dependent implicit regularisation for distributed stochastic subgradient descent. *Journal of Machine Learning Research*, 2020.

Seneta, E. *Non-negative matrices and Markov chains*. Springer Science & Business Media, 2006.

Shalev-Shwartz, S., Shamir, O., Srebro, N., and Sridharan, K. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 2010.

Shamir, O. and Srebro, N. Distributed stochastic optimization and learning. In *2014 52nd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2014.

Shi, W., Ling, Q., Wu, G., and Yin, W. Extra: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 2015.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., and Srebro, N. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 2018.

Srebro, N., Sridharan, K., and Tewari, A. Smoothness, low noise and fast rates. In *Advances in Neural Information Processing Systems*, 2010.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 2014.

Sun, T., Li, D., and Wang, B. Stability and generalization of decentralized stochastic gradient descent. *Proceedings of the AAAI Conference on Artificial Intelligence*, May 2021.

Taheri, H., Mokhtari, A., Hassani, H., and Pedarsani, R. Quantized decentralized stochastic learning over directed graphs. In *International Conference on Machine Learning*, 2020.

Tang, H., Lian, X., Yan, M., Zhang, C., and Liu, J. D2: Decentralized training over decentralized data. In *International Conference on Machine Learning*, 2018.

Tihonov, A. N. Solution of incorrectly formulated problems and the regularization method. *Soviet Math.*, 1963.

Tsitsiklis, J., Bertsekas, D., and Athans, M. Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IRE Transactions on Automatic Control*, 1986.

Tsitsiklis, J. N. Problems in decentralized decision making and computation. Technical report, Massachusetts Inst of Tech Cambridge Lab for Information and Decision Systems, 1984.

Vapnik, V. and Chervonenkis, A. Theory of pattern recognition, 1974.

Vogels, T., He, L., Koloskova, A., Karimireddy, S. P., Lin, T., Stich, S. U., and Jaggi, M. Relaysum for decentralized deep learning on heterogeneous data. *Advances in Neural Information Processing Systems*, 2021.

Wan, X., Zhang, H., Wang, H., Hu, S., Zhang, J., and Chen, K. Rat-resilient allreduce tree for distributed machine learning. In *4th Asia-Pacific Workshop on Networking*, 2020.

Wang, B., Meng, Q., Chen, W., and Liu, T.-Y. The implicit bias for adaptive optimization algorithms on homogeneous neural networks. In *International Conference on Machine Learning*, 2021.

Wang, J., Sahu, A. K., Yang, Z., Joshi, G., and Kar, S. Matcha: Speeding up decentralized sgd via matching decomposition sampling. In *2019 Sixth Indian Control Conference (ICC)*, 2019.

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. Swarm learning for decentralized and confidential clinical machine learning. *Nature*, 2021.

Xiao, L. and Boyd, S. Fast linear iterations for distributed averaging. *Systems & Control Letters*, 2004.

Xu, J., Zhang, W., and Wang, F. A (dp)2 2sgd: Asynchronous decentralized parallel stochastic gradient descent with differential privacy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Yin, H., Mallya, A., Vahdat, A., Alvarez, J. M., Kautz, J., and Molchanov, P. See through gradients: Image batch recovery via gradinversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Ying, B., Yuan, K., Chen, Y., Hu, H., Pan, P., and Yin, W. Exponential graph is provably efficient for decentralized deep training. In *Advances in Neural Information Processing Systems*, 2021a.

Ying, B., Yuan, K., Hu, H., Chen, Y., and Yin, W. Bluefog: Make decentralized algorithms practical for optimization and deep learning. *arXiv preprint arXiv:2111.04287*, 2021b.

Ying, Y. and Zhou, D.-X. Unregularized online learning algorithms with general loss functions. *Applied and Computational Harmonic Analysis*, 2017.

Yuan, K., Alghunaim, S. A., Ying, B., and Sayed, A. H. On the influence of bias-correction on distributed stochastic optimization. *IEEE Transactions on Signal Processing*, 2020.

Zhang, W., Cui, X., Finkler, U., Saon, G., Kayi, A., Buyuktosunoglu, A., Kingsbury, B., Kung, D., and Picheny, M. A highly efficient distributed deep learning system for automatic speech recognition. *arXiv preprint arXiv:1907.05701*, 2019.

Zhang, W., Liu, M., Feng, Y., Cui, X., Kingsbury, B., and Tu, Y. Loss landscape dependent self-adjusting learning rates in decentralized stochastic gradient descent. *arXiv preprint arXiv:2112.01433*, 2021.

Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., and Chandra, V. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.

Zhu, L., Liu, Z., and Han, S. Deep leakage from gradients. In *Advances in Neural Information Processing Systems*, 2019.

A. Additional Background

The following remarks clarify some notations in the literature.

Remark A.1. *Stochastic learning algorithms $A : \cup_n \mathcal{Z}^n \mapsto \mathcal{W}$ are often applied to produce an output model $A(\mathcal{S}) \in \mathbb{R}^d$ based on the training set $\mathcal{S}$. To avoid ambiguity, let $A(\mathcal{S})$ denote the model generated by a general learning algorithm, and $\mathbf{w}$ denote the models generated by a specific stochastic learning algorithm.*

Remark A.2. *Stochastic learning introduces two kinds of randomness: one from the sampling of training examples and another from the adopted randomized algorithm. In the following analysis, $\mathbb{E}_A[\cdot]$ stands for the expectation w.r.t. the randomness of the algorithm A, and $\mathbb{E}_S[\cdot]$ denotes the expectation w.r.t. the randomness originating from sampling the data set $\mathcal{S}$. Notice that $\mathcal{S} \sim \mathcal{D}^n$ and $z \sim \mathcal{D}$, therefore $\mathbb{E}_{\mathcal{S}}[\cdot]$ differs from $\mathbb{E}_z[\cdot]$ defined in Section 3.*

Based on the work by Hardt et al. (2016) and Bottou & Bousquet (2008), we give a formulation of excess error decomposition and demonstrate how to understand generalization through error decomposition.

Definition A.1 (Excess Error Decomposition). *We denote the empirical risk minimization (ERM) solution by $\mathbf{w}_{\mathcal{S}}^* = \arg\min_{\mathbf{w}} F_{\mathcal{S}}(\mathbf{w})$ and $\mathbf{w}^* = \arg\min_{\mathbf{w}} F(\mathbf{w})$. The excess error $F(A(\mathcal{S})) - F(\mathbf{w}^*)$ can be decomposed as*

$$
\begin{aligned}
&\underbrace{\mathbb{E}_{\mathcal{S},A}\left[F(A(\mathcal{S})) - F\left(\mathbf{w}^*\right)\right]}_{\text{Excess error}} \\
&\quad= \underbrace{\mathbb{E}_{\mathcal{S},A}\left[F(A(\mathcal{S})) - F_{\mathcal{S}}(A(\mathcal{S}))\right]}_{\text{Generalization error}} + \underbrace{\mathbb{E}_{\mathcal{S},A}\left[F_{\mathcal{S}}(A(\mathcal{S})) - F_{\mathcal{S}}\left(\mathbf{w}_{\mathcal{S}}^*\right)\right]}_{\text{Optimization error}} + \mathbb{E}_{\mathcal{S},A}\left[F_{\mathcal{S}}(\mathbf{w}_{\mathcal{S}}^*) - F\left(\mathbf{w}^*\right)\right] \\
&\quad\leq \underbrace{\mathbb{E}_{\mathcal{S},A}\left[F(A(\mathcal{S})) - F_{\mathcal{S}}(A(\mathcal{S}))\right]}_{\text{Generalization error}} + \underbrace{\mathbb{E}_{\mathcal{S},A}\left[F_{\mathcal{S}}(A(\mathcal{S})) - F_{\mathcal{S}}\left(\mathbf{w}_{\mathcal{S}}^*\right)\right]}_{\text{Optimization error}}
\end{aligned} \tag{A.1}
$$

The last inequality holds since and $\mathbb{E}_{\mathcal{S},A}\left[F_{\mathcal{S}}(\mathbf{w}_{\mathcal{S}}^*)\right] \leq \mathbb{E}_{\mathcal{S},A}\left[F_{\mathcal{S}}\left(\mathbf{w}^*\right)\right] = \mathbb{E}_{\mathcal{S}}\left[F(\mathbf{w}^*)\right.$. The empirical risk and the population risk above are defined in Section 3. This paper considers upper bounding the first term called the generalization error.

Lipschitzness and smoothness are two commonly adopted assumptions to establish the uniform stability guarantees of SGD.

Assumption A.1 (Lipschitzness). $\left\|\nabla f(\mathbf{w}; z)\right\|_2 \leq G$ *for all* $\mathbf{w} \in \mathbb{R}^d$ *and* $z \in \mathcal{Z}$.

Assumption A.2 (Smoothness). *f is β-smooth if for any z and $\mathbf{w}, \widetilde{\mathbf{w}} \in \mathbb{R}^d$,*

$$\left\|\nabla f(\mathbf{w}; z) - \nabla f(\widetilde{\mathbf{w}}; z)\right\|_2 \leq \beta \|\mathbf{w} - \widetilde{\mathbf{w}}\|_2. \tag{A.2}$$

These two restrictive assumptions are not satisfied in many real contexts. For example, the Lipschitz constant G can be very large for some learning problems (Fazlyab et al., 2019; Lei & Ying, 2020). In addition, neural nets with piecewise linear activation functions like ReLU are not smooth. Smoothness is generally difficult to ensure at the beginning and intermediate phases of deep neural network training (Bassily et al., 2020).

Assumption A.3 (Hölder Continuity). *Let $L > 0, \alpha \in [0, 1]$. $\nabla f(\cdot, z)$ is (α, L)-Hölder continuous if for all $\mathbf{w}, \widetilde{\mathbf{w}} \in \mathbb{R}^d$ and $z \in \mathcal{Z}$,*

$$\left\|\nabla f(\mathbf{w}; z) - \nabla f(\widetilde{\mathbf{w}}; z)\right\|_2 \leq L \|\mathbf{w} - \widetilde{\mathbf{w}}\|_2^{\alpha}. \tag{A.3}$$

Hölder continuous gradient assumption is much weaker than smoothness by definition. Serving as an intermediate class of functions ($C^{1,\alpha}\left(\mathbb{R}^n\right)$) between smooth functions ($C^{1,1}\left(\mathbb{R}^n\right)$) and functions with Lipschitz continuous gradients ($C^{1,0}\left(\mathbb{R}^n\right)$), the main advantage of functions with Hölder continuous gradients lies in the ability to automatically adjust the smoothness parameter to a proper level (Nesterov, 2015): Inequality (A.3) with $\alpha = 1$ corresponds to smoothness and Inequality (A.3) with $\alpha = 0$ is equivalent to Lipschitzness (see Assumption A.1).

Assumption A.4 (Gaussian Weight Difference). *We assume that the difference between $\mathbf{w}_k^{(t)}$ and $\widetilde{\mathbf{w}}_k^{(t)}$ (the t-th iterate on k-th worker produced by Equation (1) based on $\mathcal{S}_k$ and $\mathcal{S}_k^{(i)}$ respectively) is independent and normally distributed:*

$$\left(\mathbf{w}_k^{(t)} - \widetilde{\mathbf{w}}_k^{(t)}\right) \sim \mathcal{N}(\mu_{t,k}, \sigma_{t,k}^2 I_d), \quad k = 1, \ldots, m$$

where I_d denotes an identity matrix with size d, and $\mu_{t,k}, \sigma_{t,k}^2$ are unknown parameters. We also give a mild constraint that the d-dimensional parameter $\mu_{t,k}$ satisfies $d \cdot \mu_0^2 \leq \|\mu_{t,k}\|_2^2 \leq d \cdot \mu^2$ and the parameter $\sigma_{t,k}^2 \in \mathbb{R}$ is bounded by σ^2. Assumption A.4 is mild since $\mathcal{S}_k$ and $\mathcal{S}_k^{(i)}$ only vary at one point.

Commonly used stability notions are listed below.

Definition A.2 (Hypothesis Stability). *A stochastic algorithm A is hypothesis ϵ-stable w.r.t. the loss function f if for all training data sets $\mathcal{S}, \mathcal{S}^{(i)} \in \mathcal{Z}^n$ that differ by at most one example, we have*

$$\mathbb{E}_z \mathbb{E}_A \big[f(A(\mathcal{S}); z) - f(A(\mathcal{S}^{(i)}); z) \big] \leq \epsilon. \tag{A.4}$$

Definition A.3 (Uniform Stability). *A stochastic algorithm A is ϵ-uniformly stable w.r.t. the loss function f if for all training data sets $\mathcal{S}, \mathcal{S}^{(i)} \in \mathcal{Z}^n$ that differ by at most one example, we have*

$$\sup_z \mathbb{E}_A \big[f(A(\mathcal{S}); z) - f(A(\mathcal{S}^{(i)}); z) \big] \leq \epsilon. \tag{A.5}$$

Definition A.4 (On-average Model Stability, (Lei & Ying, 2020)). *A stochastic algorithm A is ℓ_2 on-average model ϵ-stable for all training data sets $\mathcal{S}, \mathcal{S}^{(i)} \in \mathcal{Z}^n$ that differ by at most one example, we have*

$$\frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{\mathcal{S}, \mathcal{S}^{(i)}, A} \big[\| A(\mathcal{S}) - A(\mathcal{S}^{(i)}) \|_2^2 \big] \leq \epsilon^2.$$

Some widely used notions regarding decentralized training are listed as follows.

Definition A.5 (Doubly Stochastic Matrix). *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ stand for the decentralized communication topology where $\mathcal{V}$ denotes the set of m computational nodes and $\mathcal{E}$ represents the edge set. For any given graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, the doubly stochastic gossip matrix $\mathbf{P} = [\mathbf{P}_{k,l}] \in \mathbb{R}^{m \times m}$ is defined on the edge set $\mathcal{E}$ that satisfies: (1) If $k \neq l$ and $(k,l) \notin \mathcal{E}$, then $\mathbf{P}_{k,l} = 0$ (disconnected); otherwise, $\mathbf{P}_{k,l} > 0$ (connected); (2) $\mathbf{P}_{k,l} \in [0,1]\ \forall k, l$; (3) $\mathbf{P} = \mathbf{P}^\top$; and (4) $\sum_k \mathbf{P}_{k,l} = \sum_l \mathbf{P}_{k,l} = 1$ (standard weight matrix for undirected graph).*

Definition A.6 (Spectral Gap). *Denote $\lambda = \max\{|\lambda_2|, |\lambda_m|\}$ where λ_i $(i = 2, \ldots, m)$ is the i-th largest eigenvalue of gossip matrix $\mathbf{P} \in \mathbb{R}^{m \times m}$. The spectral gap of a gossip matrix $\mathbf{P}$ can be defined as follows:*

$$\textit{spectral gap} := 1 - \lambda.$$

According to the definition of doubly stochastic matrix (Definition A.5), we have $0 \leq \lambda < 1$. The spectral gap measures the connectivity of the communication topology, which is close to 0 for sparse topologies and will approach 1 for well-connected topologies.

To facilitate our subsequent analysis, we provide some preliminaries of matrix algebra here.

Definition A.7 (Frobenius Norm). *The Frobenius norm (Euclidean norm, or Hilbert–Schmidt norm) is the matrix norm of a matrix $\mathbf{A} \in \mathbb{R}^{p \times q}$ defined as the square root of the sum of the squares of its elements:*

$$\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^{q} \sum_{j=1}^{q} |a_{ij}|^2} = \sqrt{\operatorname{Tr}(\mathbf{A}^\intercal \mathbf{A})}.$$

For any $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{p \times q}$, the following identity holds:

$$\|\mathbf{A} + \mathbf{B}\|_F^2 = \|\mathbf{A}\|_F^2 + \|\mathbf{B}\|_F^2 + 2\langle \mathbf{A}, \mathbf{B} \rangle_F,$$

where $\langle \cdot, \cdot \rangle_F$ is the Frobenius inner product.

B. Additional Related Work

B.1. Non-centralized learning.

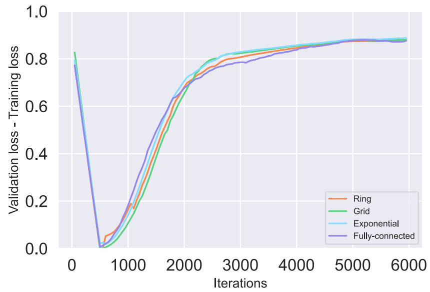
To handle an increasing amount of data and model parameters, distributed learning across multiple computing nodes (workers) emerges. A traditional distributed learning system usually follows a centralized setup (Abadi et al., 2016). However, such a central server-based learning scheme suffers from two main issues: (1) A centralized communication protocol significantly slows down the training since central servers are easily overloaded, especially in low-bandwidth or high-latency cases (Lian et al., 2017); (2) There exists potential information leakage through privacy attacks on model parameters despite decentralizing data using Federated Learning (Zhu et al., 2019; Geiping et al., 2020; Yin et al., 2021). As an alternative, training in a non-centralized fashion allows workers to balance the load on the central server through the gossip technique, as well as maintain confidentiality (Warnat-Herresthal et al., 2021). Model decentralization can be divided into three kinds of categories by layers (Lu & De Sa, 2021): (1) On the application layer, decentralized training usually refers to federated learning (Zhao et al., 2018; Dai et al., 2022) ; (2) On the protocol layer, decentralization denotes average gossip where local workers communicate by averaging their parameters with their neighbors on a graph (Lian et al., 2017) and (3) on the topology layer, it means a sparse topology graph (Wan et al., 2020).

B.2. Generalization via algorithmic stability.

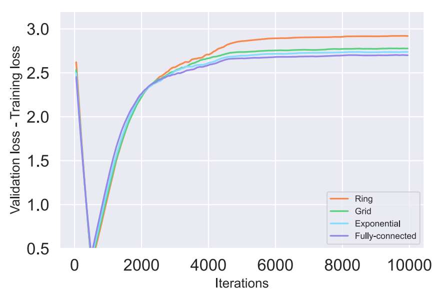
Algorithmic stability theory, PAC-Bayes theory, and information theory are major tools for constructing algorithm-dependent generalization bounds (Neu et al., 2021). A direct intuition behind algorithmic stability is that if an algorithm does not rely excessively on any single data point, it can generalize well. Proving generalization bounds based on the sensitivity of the algorithm to changes in the learning sample can be traced back to Vapnik & Chervonenkis (1974) and Devroye & Wagner (1979). After that, the celebrated work by Bousquet & Elisseeff (2002) establishes the relationship between uniform stability and generalization in high probability. Follow-up work by Shalev-Shwartz et al. (2010) identifies stability as the major necessary and sufficient condition for learnability. Then, Hardt et al. (2016) provide uniform stability bounds for stochastic gradient methods (SGM) and show the strong stability properties of SGD with convex and smooth losses. Recent work by Lei & Ying (2020) defines a new on-average stability notion and conducts generalization analyses on SGD with the Hölder continuous assumption. In addition to uniform stability, there are other stability notions including on-average stability (Shalev-Shwartz et al., 2010), uniform argument stability (Liu et al., 2017), data-dependent stability (Kuzborskij & Lampert, 2018), hypothesis set stability (Foster et al., 2019) and locally elastic stability (Deng et al., 2021).

C. Additional Experimental Results

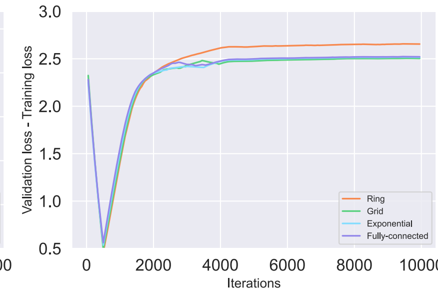
We also calculate the difference between the validation loss and the training loss in training ResNet-18 using D-SGD.



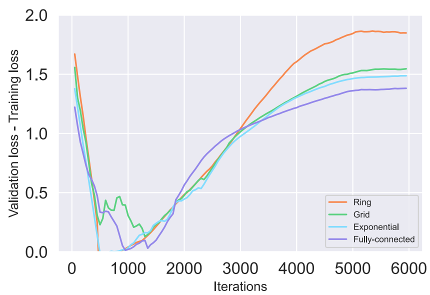
(a) ResNet-18 on CIFAR-10, 32 workers



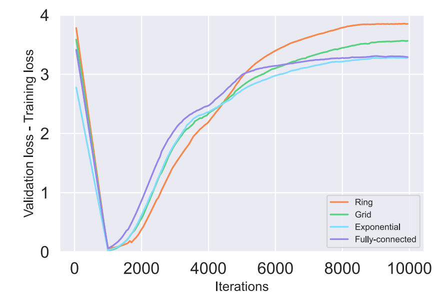
(b) ResNet-18 on CIFAR-100, 32 workers



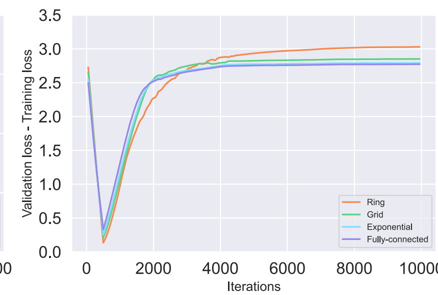
(c) ResNet-18 on Tiny ImageNet, 32 workers



(d) ResNet-18 on CIFAR-10, 64 workers



(e) ResNet-18 on CIFAR-100, 64 workers



(f) ResNet-18 on Tiny ImageNet, 64 workers

Figure C.1. Loss differences in training ResNet-18 using D-SGD with different topologies.

D. Proof

D.1. Technical lemmas

To complete our proof, we first introduce some technical lemmas.

Lemma D.1 (Corollary 1.14., (Montenegro & Tetali, 2006))**.** *Let $\mathbf{M}$ stand for the matrix with all the elements be $1/m$ and $\mathbf{P}$ is defined in Definition A.5. For any $k \in \mathbb{Z}^+$, the following inequality holds:*

$$\left\|\mathbf{P}^k - \mathbf{M}\right\|_{2,2} \leq \|\mathbf{P}\|_\lambda^k. \tag{D.1}$$

Lemma D.2. *For any $a, b \in \mathbb{R}$ and $p \in \mathbb{R}^+$, the following inequality holds:*

$$(a+b)^2 \leq (1+p)a^2 + (1+p^{-1})b^2. \tag{D.2}$$

Lemma D.3 (Self-bounding Property, (Lei & Ying, 2020))**.** *Assume that for all $z \in \mathcal{Z}$, the map $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is nonnegative with its gradient $\nabla f(\mathbf{w}; z)$ being (α, L)-Hölder continuous (Assumption A.3), then $\mathbf{w} \mapsto f(\mathbf{w}; z)$ can be bounded as*

$$\|\nabla f(\mathbf{w}, z)\|_2 \leq c_{\alpha,1} f^{\frac{\alpha}{1+\alpha}}(\mathbf{w}, z), \quad \forall \mathbf{w} \in \mathbb{R}^d, z \in \mathcal{Z}.$$

where

$$c_{\alpha,1} = \begin{cases} (1+1/\alpha)^{\frac{\alpha}{1+\alpha}} L^{\frac{1}{1+\alpha}}, & \text{if } \alpha > 0 \\ \sup_z \|\nabla f(0; z)\|_2 + L, & \text{if } \alpha = 0. \end{cases} \tag{D.3}$$

Remark D.1. *The self-binding property implies that H"older continuous gradients can be controlled by function values. The $\alpha = 1$ and $\alpha \in (0, 1)$ case are established by Srebro et al. (2010) and Ying & Zhou (2017), respectively. The case where $\alpha = 0$ follows directly from Assumption A.3.*

Lemma D.4 (Co-coercivity)**.** *Assume that for all $z \in \mathcal{Z}$, the map $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is nonnegative and convex, with its gradient $\nabla f(\mathbf{w}; z)$ being (α, L)-Hölder continuous (see Assumption A.3). Then for all $\mathbf{w}, \widetilde{\mathbf{w}}$, the following inequality holds:*

$$\|\nabla f(\mathbf{w}; z) - \nabla f(\widetilde{\mathbf{w}}; z)\|_2 \leq \frac{(1+\alpha)L^{\frac{1}{\alpha}}}{2\alpha} \left\langle \mathbf{w} - \widetilde{\mathbf{w}}, \nabla f(\mathbf{w}; z) - \nabla f(\widetilde{\mathbf{w}}; z)\right\rangle^{\frac{\alpha}{1+\alpha}}. \tag{D.4}$$

Remark D.2. *Lemma D.4 establishes the co-coercivity of the gradients for nonnegative convex functions with Hölder continuous gradients. The $\alpha = 1$ and $\alpha \in (0, 1)$ cases can be found in Nesterov (2003) and Ying & Zhou (2017) respectively. The proof of the $\alpha = 0$ case can be obtained directly from the convexity of f.*

Remark D.3. *Using Young's inequality*

$$ab \leq p^{-1}|a|^p + q^{-1}|b|^q \quad (a, b \in \mathbb{R}, p, q > 0 \text{ with } p^{-1} + q^{-1} = 1),$$

for any $\eta > 0$, the right-hand side of Inequality (D.4) can be further controlled by

$$2\eta^{-1}\left\langle \mathbf{w} - \widetilde{\mathbf{w}}, \nabla f(\mathbf{w}; z) - \nabla f(\widetilde{\mathbf{w}}; z)\right\rangle + c_{\alpha,3}^2 \eta^{\frac{2\alpha}{1-\alpha}}, \tag{D.5}$$

where $c_{\alpha,3} = \frac{\sqrt{1-\alpha}}{\sqrt{1+\alpha}}(2^{-\alpha}L)^{\frac{1}{1-\alpha}}$. For more details, see Appendix D of Lei & Ying (2020).

Lemma D.5. *For any $a, b \in \mathbb{R}^d$ with a_i, b_i being their i-th components, respectively, the following inequality holds:*

$$a^\mathsf{T} b = \sum_i a_i b_i \leq \sqrt{\sum_i a_i^2 \sum_i b_i^2} \leq \frac{\sum_i a_i^2 + \sum_i b_i^2}{2}. \tag{D.6}$$

D.2. Algorithmic stability of D-SGD

Proof of Theorem 1.

To begin with, we decompose $\sum_{k=1}^m \left\|\mathbf{w}_k^{(t+1)} - \widetilde{\mathbf{w}}_k^{(t+1)}\right\|_2^2$, the on-average stability of D-SGD at the t-th iteration, into three parts by the definition of the vector 2-norm. In the following, we will let $z_{k,\zeta_t}^{(t)}$ and $\tilde{z}_{k,\zeta_t}^{(t)}$ denote two random samples drawn from $\mathcal{S}$ and $\mathcal{S}^{(i)}$ respectively on the k-th worker at the i-th iteration, respectively.

$$
\begin{aligned}
&\sum_{k=1}^m \left\|\mathbf{w}_k^{(t+1)} - \widetilde{\mathbf{w}}_k^{(t+1)}\right\|_2^2 \\
&\quad= \sum_{k=1}^m \Big\| \sum_{l=1}^m \mathbf{P}_{k,l}\mathbf{w}_t(l) - \eta_t \nabla f\left(\mathbf{w}_k^{(t)}; z_{k,\zeta_t}^{(t)}\right) - \sum_{l=1}^m \mathbf{P}_{k,l}\widetilde{\mathbf{w}}_l^{(t)} + \eta_t \nabla f\left(\widetilde{\mathbf{w}}_k^{(t)}, \tilde{z}_{k,\zeta_t}^{(t)}\right) \Big\|_2^2 \\
&\quad= \sum_{k=1}^m \Big\| \sum_{l=1}^m \mathbf{P}_{k,l}(\mathbf{w}_l^{(t)} - \widetilde{\mathbf{w}}_l^{(t)})\Big\|_2^2 + \sum_{k=1}^m {\eta_t}^2 \Big\|\nabla f\left(\mathbf{w}_k^{(t)}; z_{k,\zeta_t}^{(t)}\right) - \nabla f\left(\widetilde{\mathbf{w}}_k^{(t)}, \tilde{z}_{k,\zeta_t}^{(t)}\right)\Big\|_2^2 \\
&\quad - 2\sum_{k=1}^m \eta_t \Big\langle \sum_{l=1}^m \mathbf{P}_{k,l}(\mathbf{w}_l^{(t)} - \widetilde{\mathbf{w}}_l^{(t)}), \nabla f\left(\mathbf{w}_k^{(t)}; z_{k,\zeta_t}^{(t)}\right) - \nabla f\left(\widetilde{\mathbf{w}}_k^{(t)}, \tilde{z}_{k,\zeta_t}^{(t)}\right) \Big\rangle \\
&\quad= \underbrace{\left\|\mathbf{P}(\mathbf{W}^{(t)} - \widetilde{\mathbf{W}}^{(t)})\right\|_F^2}_{T_1} + \underbrace{\sum_{k=1}^m \mathbb{1}_{z_{k,\zeta_t}^{(t)} \neq \tilde{z}_{k,\zeta_t}^{(t)}} \Big[\left\|\mathbf{w}_k^{(t+1)} - \widetilde{\mathbf{w}}_k^{(t+1)}\right\|_2^2 - \Big\| \sum_{l=1}^m \mathbf{P}_{k,l}(\mathbf{w}_l^{(t)} - \widetilde{\mathbf{w}}_l^{(t)})\Big\|_2^2\Big]}_{T_2} \\
&\quad + \underbrace{\begin{aligned}\sum_{k=1}^m \mathbb{1}_{z_{k,\zeta_t}^{(t)} = \tilde{z}_{k,\zeta_t}^{(t)}} \Big[& {\eta_t}^2 \Big\|\nabla f\left(\mathbf{w}_k^{(t)}; z_{k,\zeta_t}^{(t)}\right) - \nabla f\left(\widetilde{\mathbf{w}}_k^{(t)}, \tilde{z}_{k,\zeta_t}^{(t)}\right)\Big\|_2^2 \\ & -2\eta_t \Big\langle \sum_{l=1}^m \mathbf{P}_{k,l}(\mathbf{w}_l^{(t)} - \widetilde{\mathbf{w}}_l^{(t)}), \nabla f\left(\mathbf{w}_k^{(t)}; z_{k,\zeta_t}^{(t)}\right) - \nabla f\left(\widetilde{\mathbf{w}}_k^{(t)}, \tilde{z}_{k,\zeta_t}^{(t)}\right) \Big\rangle\Big]\end{aligned}}_{T_3}
\end{aligned}
$$

where $\|\cdot\|_F$ denotes the Frobenius norm (see Definition A.7).

(1) To construct our proof, we start by constructing an upper bound for the expectation of T_1:

$$
\mathbb{E}_A(T_1) = \mathbb{E}_A\left[\left\|\mathbf{P}(\mathbf{W}^{(t)} - \widetilde{\mathbf{W}}^{(t)})\right\|_F^2 \leq d(\sigma^2 + \mu^2)\left[(m-1)\lambda^2 + 1\right]. \tag{D.7}
$$

Proof.

The on-averaged stability after a single gossip communication can be written as

$$
\begin{aligned}
\left\|\mathbf{P}(\mathbf{W}^{(t)} - \widetilde{\mathbf{W}}^{(t)})\right\|_F^2 &= \sum_{k=1}^m \Big\| \sum_{l=1}^m \mathbf{P}_{k,l}(\mathbf{w}_l^{(t)} - \widetilde{\mathbf{w}}_l^{(t)})\Big\|_2^2 = \sum_{v=1}^d \sum_{k=1}^m \Big\| \sum_{l=1}^m \mathbf{P}_{k,l}(\mathbf{w}_l^{v,(t)} - \widetilde{\mathbf{w}}_l^{v,(t)})\Big\|_2^2 \\
&= \sum_{v=1}^d \sum_{k=1}^m \sigma_{t,k}^2 (\sum_{l=1}^m \mathbf{P}_{k,l}^2) \{ \frac{\sum_{l=1}^m \mathbf{P}_{k,l}[(\mathbf{w}_l^{v,(t)} - \widetilde{\mathbf{w}}_l^{v,(t)}) - \mu_{t,k}^v] + \sum_{l=1}^m \mathbf{P}_{k,l}\mu_{t,l}^v]}{\sigma_{t,k}\sqrt{\sum_{l=1}^m \mathbf{P}_{k,l}^2}} \}^2 \\
&\leq \sum_{v=1}^d \sum_{k=1}^m \sigma_{t,k}^2 (\sum_{l=1}^m \mathbf{P}_{k,l}^2) \{ \frac{\sum_{l=1}^m \mathbf{P}_{k,l}[(\mathbf{w}_l^{v,(t)} - \widetilde{\mathbf{w}}_l^{v,(t)}) - \mu_{t,k}^v]}{\sigma_{t,k}\sqrt{\sum_{l=1}^m \mathbf{P}_{k,l}^2}} \}^2 + \mu^2 dm \cdot \sum_{k=1}^m \sum_{l=1}^m \mathbf{P}_{k,l}^2 \\
&+ \sum_{v=1}^d \sum_{k=1}^m (\sum_{l=1}^m \mathbf{P}_{k,l}^2) \frac{\sum_{l_1=1}^m \sum_{l_2=1}^m \mathbf{P}_{k,l_1}\mathbf{P}_{k,l_2}\mu_{t,l_1}^v [(\mathbf{w}_{l_1}^{v,(t)} - \widetilde{\mathbf{w}}_{l_1}^{v,(t)}) - \mu_{t,l_1}^v]}{\sqrt{\sum_{l_1=1}^m \mathbf{P}_{k,l}^2}\sqrt{\sum_{l_2=1}^m \mathbf{P}_{k,l}^2}},
\end{aligned} \tag{D.8}
$$

where $\mathbf{w}_l^{v,(t)}$ and $\widetilde{\mathbf{w}}_l^{v,(t)}$ stacks the v-th entry of the d-dimensional vector $\mathbf{w}_l^{(t)}$ and $\widetilde{\mathbf{w}}_l^{(t)}$, respectively. The last inequality holds since $\sum_{v=1}^d (\mu_{t,l}^v)^2 \leq d\mu^2$.

Since the weight difference is normally distributed:

$$\big(\mathbf{w}_k^{(t)} - \widetilde{\mathbf{w}}_k^{(t)}\big) \sim \mathcal{N}(\mu_{t,k}, \sigma_{t,k}^2 I_d), \quad k = 1, \dots, m,$$

with $\mu_{t,k}$ satisfying $\|\mu_{t,k}\|_2^2 \le d \cdot \mu^2$ and $\sigma_{t,k}^2 \in \mathbb{R}$ being bounded by σ^2, we obtain

$$\frac{\sum_{l=1}^m \mathbf{P}_{k,l}[(\mathbf{w}_l^{v,(t)} - \widetilde{\mathbf{w}}_l^{v,(t)}) - \mu_{t,k}^v]}{\sigma_{t,k}\sqrt{\sum_{l=1}^m \mathbf{P}_{k,l}^2}} \overset{i.i.d.}{\sim} \mathcal{N}(0,1), \quad l = 1, \dots, m.$$

Since the sum of squared i.i.d. standard normal variables follows a Chi-Square distribution with 1 degree of freedom, we arrive at

$$\{\frac{\sum_{l=1}^m \mathbf{P}_{k,l}[(\mathbf{w}_l^{v,(t)} - \widetilde{\mathbf{w}}_l^{v,(t)}) - \mu_{t,k}^v]}{\sigma_{t,k}\sqrt{\sum_{l=1}^m \mathbf{P}_{k,l}^2}}\}^2 \sim \mathcal{X}^2(1), \quad l = 1, \dots, m.$$

Furthermore, since $\forall l \ \mathbb{E}_A[(\mathbf{w}_l^{v,(t)} - \widetilde{\mathbf{w}}_l^{v,(t)}) - \mu_{t,l}^v] = 0$, we have

$$\mathbb{E}_A\left[\sum_{v=1}^d \sum_{k=1}^m (\sum_{l=1}^m \mathbf{P}_{k,l}^2) \frac{\sum_{l_1=1}^m \sum_{l_2=1}^m \mathbf{P}_{k,l_1}\mathbf{P}_{k,l_2}\mu_{t,l_1}^v[(\mathbf{w}_{l_1}^{v,(t)} - \widetilde{\mathbf{w}}_{l_1}^{v,(t)}) - \mu_{t,l_1}^v]}{\sqrt{\sum_{l_1=1}^m \mathbf{P}_{k,l}^2}\sqrt{\sum_{l_2=1}^m \mathbf{P}_{k,l}^2}}\right] = 0.$$

As a consequence,

$$\mathbb{E}_A\big[\big\|\mathbf{P}(\mathbf{W}^{(t)} - \widetilde{\mathbf{W}}^{(t)})\big\|_F^2 \le d\sigma^2 \sum_{k=1}^m \lambda_k^2 + d\mu^2 \sum_{k=1}^m \lambda_k^2 \le d(\sigma^2 + \mu^2)[1 + (m-1)\lambda^2],$$

where $1 - \lambda$ is the spectral gap of the communication topology.

(2) For the second part T_2, we have

$$T_2 \le \sum_{k=1}^m \mathbb{1}_{z_{k,\zeta_t}^{(t)} \neq \tilde{z}_{k,\zeta_t}^{(t)}} p\big\|\sum_{l=1}^m \mathbf{P}_{k,l}(\mathbf{w}_l^{(t)} - \widetilde{\mathbf{w}}_l^{(t)})\big\|_2^2 + \sum_{k=1}^m \mathbb{1}_{z_{k,\zeta_t}^{(t)} \neq \tilde{z}_{k,\zeta_t}^{(t)}} (1 + p^{-1})c_{\alpha,1}^2 {\eta_t}^2 \big(f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_k^{(t)}; z_{k,\zeta_t}^{(t)}) + f^{\frac{2\alpha}{1+\alpha}}(\widetilde{\mathbf{w}}_k^{(t)}, \tilde{z}_{k,\zeta_t}^{(t)})\big). \tag{D.9}$$

Proof.

Inequality (D.9) are mainly based on Lemma D.2 and Lemma D.3:

$$\begin{aligned}
T_2 &= \sum_{k=1}^m \mathbb{1}_{z_{k,\zeta_t}^{(t)} \neq \tilde{z}_{k,\zeta_t}^{(t)}} \big[\big\|\mathbf{w}_k^{(t+1)} - \widetilde{\mathbf{w}}_k^{(t+1)}\big\|_2^2 - \big\|\sum_{l=1}^m \mathbf{P}_{k,l}(\mathbf{w}_l^{(t)} - \widetilde{\mathbf{w}}_l^{(t)})\big\|_2^2\big] \\
&= \sum_{k=1}^m \mathbb{1}_{z_{k,\zeta_t}^{(t)} \neq \tilde{z}_{k,\zeta_t}^{(t)}} \big[\big\|\sum_{l=1}^m \mathbf{P}_{k,l}\mathbf{w}_t(l) - \eta_t \nabla f\left(\mathbf{w}_k^{(t)}; z_{k,\zeta_t}^{(t)}\right) - \sum_{l=1}^m \mathbf{P}_{k,l}\widetilde{\mathbf{w}}_l^{(t)} + \eta_t \nabla f\big(\widetilde{\mathbf{w}}_k^{(t)}, \tilde{z}_{k,\zeta_t}^{(t)}\big)\big\|_2^2 - \big\|\sum_{l=1}^m \mathbf{P}_{k,l}(\mathbf{w}_l^{(t)} - \widetilde{\mathbf{w}}_l^{(t)})\big\|_2^2\big] \\
&\underset{\text{Lemma D.2}}{\le} \sum_{k=1}^m \mathbb{1}_{z_{k,\zeta_t}^{(t)} \neq \tilde{z}_{k,\zeta_t}^{(t)}} \big[p\big\|\sum_{l=1}^m \mathbf{P}_{k,l}(\mathbf{w}_l^{(t)} - \widetilde{\mathbf{w}}_l^{(t)})\big\|_2^2 + (1 + p^{-1}){\eta_t}^2 \big\|\nabla f\left(\mathbf{w}_k^{(t)}; z_{k,\zeta_t}^{(t)}\right) - \nabla f\left(\widetilde{\mathbf{w}}_k^{(t)}, \tilde{z}_{k,\zeta_t}^{(t)}\right)\big\|_2^2\big] \\
&\underset{\text{Lemma D.3}}{\le} \underbrace{\sum_{k=1}^m \mathbb{1}_{z_{k,\zeta_t}^{(t)} \neq \tilde{z}_{k,\zeta_t}^{(t)}} \big[p\big\|\sum_{l=1}^m \mathbf{P}_{k,l}(\mathbf{w}_l^{(t)} - \widetilde{\mathbf{w}}_l^{(t)})\big\|_2^2}_{T_{2.1}} + (1 + p^{-1})c_{\alpha,1}^2 {\eta_t}^2 \big(f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_k^{(t)}; z_{k,\zeta_t}^{(t)}) + f^{\frac{2\alpha}{1+\alpha}}(\widetilde{\mathbf{w}}_k^{(t)}, \tilde{z}_{k,\zeta_t}^{(t)})\big)\big].
\end{aligned}$$

(3) T_3 can be controlled as follows:

$$T_3 \le \sum_{k=1}^m \mathbb{1}_{z_{k,\zeta_t}^{(t)} = \tilde{z}_{k,\zeta_t}^{(t)}} \eta_t \big[2L\big\|\mathbf{w}_k^{(t)} - \widetilde{\mathbf{w}}_k^{(t)}\big\|_2^{2\alpha} + \big\|\sum_{l=1}^m \mathbf{P}_{k,l}(\mathbf{w}_l^{(t)} - \widetilde{\mathbf{w}}_l^{(t)})\big\|_2^2\big]. \tag{D.10}$$

Proof.

According to the Hölder continuous assumption, we have

$$\sum_{k=1}^{m} \mathbb{1}_{z_{k,\zeta_t}^{(t)}=\tilde{z}_{k,\zeta_t}^{(t)}} \big\|\nabla f(\mathbf{w}_k^{(t)}; z_{k,\zeta_t}^{(t)}) - \nabla f(\widetilde{\mathbf{w}}_k^{(t)}; \tilde{z}_{k,\zeta_t}^{(t)})\big\|_2^2 \leq L \sum_{k=1}^{m} \mathbb{1}_{z_{k,\zeta_t}^{(t)}=\tilde{z}_{k,\zeta_t}^{(t)}} \big\|\mathbf{w}_k^{(t)} - \widetilde{\mathbf{w}}_k^{(t)}\big\|_2^{2\alpha}. \tag{D.11}$$

Consequently,

$$\begin{aligned}
T_3 &\leq \sum_{k=1}^{m} \mathbb{1}_{z_{k,\zeta_t}^{(t)}=\tilde{z}_{k,\zeta_t}^{(t)}} \eta_t \big[L\big\|\mathbf{w}_k^{(t)} - \widetilde{\mathbf{w}}_k^{(t)}\big\|_2^{2\alpha} - 2\big\langle \sum_{l=1}^{m} \mathbf{P}_{k,l}(\mathbf{w}_l^{(t)} - \widetilde{\mathbf{w}}_l^{(t)}), \nabla f\left(\mathbf{w}_k^{(t)}; z_{k,\zeta_t}^{(t)}\right) - \nabla f\left(\widetilde{\mathbf{w}}_k^{(t)}, \tilde{z}_{k,\zeta_t}^{(t)}\right)\big\rangle\big] \\
&\leq \sum_{k=1}^{m} \mathbb{1}_{z_{k,\zeta_t}^{(t)}=\tilde{z}_{k,\zeta_t}^{(t)}} \eta_t \big[L\big\|\mathbf{w}_k^{(t)} - \widetilde{\mathbf{w}}_k^{(t)}\big\|_2^{2\alpha} + 2\big\|\sum_{l=1}^{m} \mathbf{P}_{k,l}(\mathbf{w}_l^{(t)} - \widetilde{\mathbf{w}}_l^{(t)})\big\|_2 \cdot \big\|\nabla f(\mathbf{w}_k^{(t)}; z_{k,\zeta_t}^{(t)}) - \nabla f(\widetilde{\mathbf{w}}_k^{(t)}; \tilde{z}_{k,\zeta_t}^{(t)})\big\|_2\big] \\
&\underset{\text{Lemma D.5}}{\leq} \sum_{k=1}^{m} \mathbb{1}_{z_{k,\zeta_t}^{(t)}=\tilde{z}_{k,\zeta_t}^{(t)}} \eta_t \big[L\big\|\mathbf{w}_k^{(t)} - \widetilde{\mathbf{w}}_k^{(t)}\big\|_2^{2\alpha} + \big\|\sum_{l=1}^{m} \mathbf{P}_{k,l}(\mathbf{w}_l^{(t)} - \widetilde{\mathbf{w}}_l^{(t)})\big\|_2^2 + \big\|\nabla f(\mathbf{w}_k^{(t)}; z_{k,\zeta_t}^{(t)}) - \nabla f(\widetilde{\mathbf{w}}_k^{(t)}; \tilde{z}_{k,\zeta_t}^{(t)})\big\|_2^2\big] \\
&\leq \underbrace{\sum_{k=1}^{m} \mathbb{1}_{z_{k,\zeta_t}^{(t)}=\tilde{z}_{k,\zeta_t}^{(t)}} \eta_t \big[2L\big\|\mathbf{w}_k^{(t)} - \widetilde{\mathbf{w}}_k^{(t)}\big\|_2^{2\alpha}}_{T_{3.1}} + \big\|\sum_{l=1}^{m} \mathbf{P}_{k,l}(\mathbf{w}_l^{(t)} - \widetilde{\mathbf{w}}_l^{(t)})\big\|_2^2\big].
\end{aligned} \tag{D.12}$$

(4) A simple combination of Inequality (D.7), Inequality (D.9) and Inequality (D.10) provides the following:

$$\begin{aligned}
&\sum_{k=1}^{m} \big\|\mathbf{w}_k^{(t+1)} - \widetilde{\mathbf{w}}_k^{(t+1)}\big\|_2^2 = T_1 + T_2 + T_3 \\
&\quad\leq T_1 + \sum_{k=1}^{m} \mathbb{1}_{z_{k,\zeta_t}^{(t)}\neq\tilde{z}_{k,\zeta_t}^{(t)}} \big[p\big\|\sum_{l=1}^{m} \mathbf{P}_{k,l}(\mathbf{w}_l^{(t)} - \widetilde{\mathbf{w}}_l^{(t)})\big\|_2^2 + (1+p^{-1})c_{\alpha,1}^2 \eta_t^2 \big(f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_k^{(t)}; z_{k,\zeta_t}^{(t)}) + f^{\frac{2\alpha}{1+\alpha}}(\widetilde{\mathbf{w}}_k^{(t)}, \tilde{z}_{k,\zeta_t}^{(t)})\big)\big] \\
&\quad+ \sum_{k=1}^{m} \mathbb{1}_{z_{k,\zeta_t}^{(t)}=\tilde{z}_{k,\zeta_t}^{(t)}} \eta_t \big[2L\big\|\mathbf{w}_k^{(t)} - \widetilde{\mathbf{w}}_k^{(t)}\big\|_2^{2\alpha} + \big\|\sum_{l=1}^{m} \mathbf{P}_{k,l}(\mathbf{w}_l^{(t)} - \widetilde{\mathbf{w}}_l^{(t)})\big\|_2^2\big].
\end{aligned} \tag{D.13}$$

$\mathcal{S}_k^{(i)} = \{z_{k,1}, \ldots, \tilde{z}_{i,k}, \ldots, z_{k,n}\}$ differs from $\mathcal{S}_k$ by only the i-th element. Consequently, at the t-th iterate, with a probability of $1 - \frac{1}{n}$, the example $z_{k,\zeta_t}^{(t)}$ selected by D-SGD on worker k in both $\mathcal{S}_k$ and $\mathcal{S}_k^{(i)}$ is the same (i.e. $z_{k,\zeta_t}^{(t)} = \tilde{z}_{k,\zeta_t}^{(t)}$), and with a probability of $\frac{1}{n}$, the selected example is different (i.e. $z_{k,\zeta_t}^{(t)} \neq \tilde{z}_{k,\zeta_t}^{(t)}$).

Since A is independent of k, $\mathbb{E}_A(T_{2.1})$ and $\mathbb{E}_A(T_{3.1})$ can be controlled accordingly as follows:

$$\begin{aligned}
\mathbb{E}_A(T_{2.1}) &= \mathbb{E}_A\big[p\sum_{k=1}^{m} \mathbb{1}_{z_{k,\zeta_t}^{(t)}\neq\tilde{z}_{k,\zeta_t}^{(t)}} \big\|\sum_{l=1}^{m} \mathbf{P}_{k,l}(\mathbf{w}_l^{(t)} - \widetilde{\mathbf{w}}_l^{(t)})\big\|_2^2\big] \\
&= \frac{p}{n}\sum_{k=1}^{m} \mathbb{E}_A\big[\big\|\sum_{l=1}^{m} \mathbf{P}_{k,l}(\mathbf{w}_l^{(t)} - \widetilde{\mathbf{w}}_l^{(t)})\big\|_2^2\big] \underset{\text{Inequality (D.7)}}{\leq} \frac{p}{n} d(\sigma^2 + \mu^2)\big[(m-1)\lambda^2 + 1\big]
\end{aligned} \tag{D.14}$$

where the proof of the last inequality is analogous to part (1).

By the concavity of the mapping $x \mapsto x^\alpha$ ($\alpha \in [0, 1]$), we have

$$\begin{aligned}
\mathbb{E}_A(T_{3.1}) &\leq 2\eta_t L(1 - \frac{1}{n})\mathbb{E}_A\big[\sum_{k=1}^{m} \big\|\mathbf{w}_k^{(t)} - \widetilde{\mathbf{w}}_k^{(t)}\big\|_2^{2\alpha}\big] \\
&\leq 2\eta_t L(1 - \frac{1}{n})\{\mathbb{E}_A\big[\sum_{k=1}^{m} \big\|\mathbf{w}_k^{(t)} - \widetilde{\mathbf{w}}_k^{(t)}\big\|_2^2\big]\}^\alpha \leq 2\eta_t L(1 - \frac{1}{n})\mathbb{E}_A\big[\sum_{k=1}^{m} \big\|\mathbf{w}_k^{(t)} - \widetilde{\mathbf{w}}_k^{(t)}\big\|_2^2\big].
\end{aligned} \tag{D.15}$$

The last inequality holds if $m \geq \frac{1}{d\mu_0^2}$, where $d\mu_0^2$ is the lower bound of $\|\mu_{t,k}\|_2^2$ ($k = 1 \dots m$), which leads to $\mathbb{E}_A\big[\sum_{k=1}^m \|\mathbf{w}_k^{(t)} - \widetilde{\mathbf{w}}_k^{(t)}\|_2^2\big] \geq 1$. The condition $m \geq \frac{1}{d\mu_0^2}$ can be easily satisfied in training overparameterized models in a decentralized manner, since both m and d are large in these cases.

Therefore, taking the expectation on both sides of Inequality (D.13) provides

$$\mathbb{E}_A\Big[\sum_{k=1}^m \big\|\mathbf{w}_k^{(t+1)} - \widetilde{\mathbf{w}}_k^{(t+1)}\big\|_2^2\Big] \leq 2\eta_t L(1-\frac{1}{n})\mathbb{E}_A\Big[\sum_{k=1}^m \big\|\mathbf{w}_k^{(t)} - \widetilde{\mathbf{w}}_k^{(t)}\big\|_2^2\Big]$$
$$+[1+\frac{p}{n}+(1-\frac{1}{n})\eta_t]d(\sigma^2+\mu^2)\big[(m-1)\lambda^2+1\big] + \frac{1}{n}\sum_{k=1}^m \big[(1+p^{-1})c_{\alpha,1}^2{\eta_t}^2\big(f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_k^{(t)};z_{k,\zeta_t}^{(t)})+f^{\frac{2\alpha}{1+\alpha}}(\widetilde{\mathbf{w}}_k^{(t)},\tilde{z}_{k,\zeta_t}^{(t)})\big)\big]. \tag{D.16}$$

Knowing that $z_{k,\zeta_t}^{(t)}$ and $\tilde{z}_{k,\zeta_t}^{(t)}$ follow the same distribution, we have

$$\mathbb{E}_{\mathcal{S}_k,\mathcal{S}_k^{(i)},A}\big[f^{\frac{2\alpha}{1+\alpha}}(\widetilde{\mathbf{w}}_k^{(t)};\tilde{z}_{k,\zeta_t}^{(t)})\big] = \mathbb{E}_{\mathcal{S}_k,A}\big[f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_k^{(t)};z_{k,\zeta_t}^{(t)})\big].$$

Note that $\{\eta_t\}$ is an non-increasing sequence. As a consequence,

$$\sum_{k=1}^m \mathbb{E}_{\mathcal{S}_k,\mathcal{S}_k^{(i)},A}\big[\big\|\mathbf{w}_k^{(t+1)} - \widetilde{\mathbf{w}}_k^{(t+1)}\big\|_2^2\big] \leq \underbrace{2\eta_0 L(1-\frac{1}{n})}_{\triangleq C \text{ (scaling coefficient)}} \mathbb{E}_{\mathcal{S}_k,\mathcal{S}_k^{(i)},A}\Big[\sum_{k=1}^m \big\|\mathbf{w}_k^{(t)} - \widetilde{\mathbf{w}}_k^{(t)}\big\|_2^2\Big]$$
$$+\underbrace{[1+\frac{p}{n}+(1-\frac{1}{n})\eta_t]d(\sigma^2+\mu^2)\big[(m-1)\lambda^2+1\big]}_{\text{topology-dependent}} + \underbrace{\frac{2}{n}\big[(1+p^{-1})c_{\alpha,1}^2{\eta_t}^2\sum_{k=1}^m \mathbb{E}_{\mathcal{S}_k,A}\big[f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_k^{(t)};z_{k,\zeta_t}^{(t)})\big]\big]}_{\text{topology-independent}}. \tag{D.17}$$

Multiplying both sides of Inequality (D.17) with $C^{-(t+1)}$ provides

$$C^{-(t+1)}\sum_{k=1}^m \mathbb{E}_{\mathcal{S}_k,\mathcal{S}_k^{(i)},A}\big[\big\|\mathbf{w}_k^{(t+1)} - \widetilde{\mathbf{w}}_k^{(t+1)}\big\|_2^2\big] \leq C^{-t}\mathbb{E}_{\mathcal{S}_k,\mathcal{S}_k^{(i)},A}\Big[\sum_{k=1}^m \big\|\mathbf{w}_k^{(t)} - \widetilde{\mathbf{w}}_k^{(t)}\big\|_2^2\Big]$$
$$+C^{-(t+1)}\big\{[1+\frac{p}{n}+(1-\frac{1}{n})\eta_t]d(\sigma^2+\mu^2)\big[(m-1)\lambda^2+1\big] + \frac{2}{n}\big[(1+p^{-1})c_{\alpha,1}^2{\eta_t}^2\sum_{k=1}^m \mathbb{E}_{\mathcal{S}_k,A}\big[f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_k^{(t)};z_{k,\zeta_t}^{(t)})\big]\big]\big\}. \tag{D.18}$$

Taking the summation over the iteration τ, we can write

$$\sum_{\tau=0}^t C^{-(\tau+1)}\sum_{k=1}^m \mathbb{E}_{\mathcal{S}_k,\mathcal{S}_k^{(i)},A}\big[\big\|\mathbf{w}_k^{(\tau+1)} - \widetilde{\mathbf{w}}_k^{(\tau+1)}\big\|_2^2\big] \leq \sum_{\tau=0}^t C^{-\tau}\mathbb{E}_{\mathcal{S}_k,\mathcal{S}_k^{(i)},A}\Big[\sum_{k=1}^m \big\|\mathbf{w}_k^{(\tau)} - \widetilde{\mathbf{w}}_k^{(\tau)}\big\|_2^2\Big]$$
$$+\sum_{\tau=0}^t C^{-(\tau+1)}\big\{[1+\frac{p}{n}+(1-\frac{1}{n})\eta_\tau]d(\sigma^2+\mu^2)\big[(m-1)\lambda^2+1\big] + \frac{2}{n}\big[(1+p^{-1})c_{\alpha,1}^2{\eta_\tau}^2\sum_{k=1}^m \mathbb{E}_{\mathcal{S}_k,A}\big[f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_k^{(\tau)};z_{k,\zeta_t}^{(\tau)})\big]\big]\big\}. \tag{D.19}$$

Since for all k, $\mathbf{w}_1(k) = \widetilde{\mathbf{w}}_1(k) = 0$ (see Definition Page 3), we have

$$\sum_{k=1}^m \mathbb{E}_{\mathcal{S}_k,\mathcal{S}_k^{(i)},A}\big[\big\|\mathbf{w}_k^{(t+1)} - \widetilde{\mathbf{w}}_k^{(t+1)}\big\|_2^2\big]$$
$$\leq \sum_{\tau=0}^t C^{t-\tau}\big\{[1+\frac{p}{n}+(1-\frac{1}{n})\eta_\tau]d(\sigma^2+\mu^2)\big[(m-1)\lambda^2+1\big] + \frac{2}{n}\big[(1+p^{-1})c_{\alpha,1}^2{\eta_\tau}^2\sum_{k=1}^m \mathbb{E}_{\mathcal{S}_k,A}\big[f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_k^{(\tau)};z_{k,\zeta_t}^{(\tau)})\big]\big]\big\}. \tag{D.20}$$

Since the mapping $x \mapsto x^{\frac{2\alpha}{1+\alpha}}$ is concave, we have $\frac{1}{n}\sum_{i=1}^{n} f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_k^{(\tau)}, z_{k,\zeta_\tau}^{(\tau)}) \leq F_{\mathcal{S}_k}^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_k^{(\tau)})$. Then the distributed on-average stability of D-SGD can be controlled as follows:

$$
\begin{aligned}
&\frac{1}{mn}\sum_{i=1}^{n}\sum_{k=1}^{m}\mathbb{E}_{\mathcal{S}_k,\mathcal{S}_k^{(i)},A}\big[\big\|\mathbf{w}_k^{(t+1)}-\widetilde{\mathbf{w}}_k^{(t+1)}\big\|_2^2\big]\\
&\leq \sum_{\tau=0}^{t} C^{t-\tau}\big\{[1+\frac{p}{n}+(1-\frac{1}{n})\eta_\tau]d(\sigma^2+\mu^2)\big[(1-\frac{1}{m})\lambda^2+\frac{1}{m}\big]+\frac{2}{n}\big[(1+p^{-1})c_{\alpha,1}^2{\eta_\tau}^2\frac{1}{m}\sum_{k=1}^{m}\mathbb{E}_{\mathcal{S}_k,A}\big[F_{\mathcal{S}_k}^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_k^{(\tau)})\big]\big]\big\}. \quad \text{(D.21)}
\end{aligned}
$$

The proof is complete.

□

Proof of Corollary 2.

With constant step size $\eta_t \equiv \eta \leq \frac{1}{2L}(1-\frac{2}{m})$, $\sum_{\tau=0}^{t} C^{t-\tau}$ can be written as

$$
\sum_{\tau=0}^{t} C^{t-\tau} = \sum_{\tau=0}^{t}[2\eta L(1-\frac{1}{n})]^{t-\tau} = \sum_{\tau=0}^{t-1}[2\eta L(1-\frac{1}{n})]^{\tau} = \frac{1-[2\eta L(1-\frac{1}{n})]^t}{1-2\eta L(1-\frac{1}{n})}.
$$

Consequently, the distributed on-average stability of D-SGD becomes:

$$
\begin{aligned}
&\frac{1}{mn}\sum_{i=1}^{n}\sum_{k=1}^{m}\mathbb{E}_{\mathcal{S}_k,\mathcal{S}_k^{(i)},A}\big[\big\|\mathbf{w}_k^{(t+1)}-\widetilde{\mathbf{w}}_k^{(t+1)}\big\|_2^2\big]\\
&\qquad\leq \frac{1-[2\eta L(1-\frac{1}{n})]^t}{1-2\eta L(1-\frac{1}{n})}\big\{[1+\frac{p}{n}+(1-\frac{1}{n})\eta]d(\sigma^2+\mu^2)\big[(1-\frac{1}{m})\lambda^2+\frac{1}{m}\big]+\frac{2}{n}\big[(1+p^{-1})c_{\alpha,1}^2\eta^2\epsilon_{\mathcal{S}}\big]\big\}, \quad \text{(D.22)}
\end{aligned}
$$

where $\epsilon_{\mathcal{S}}$ denotes the upper bound of $\frac{1}{m}\sum_{k=1}^{m}\mathbb{E}_{\mathcal{S}_k,A}\big[F_{\mathcal{S}_k}^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_k^{(t)})\big]\ \forall t$.

When $t \to \infty$, we further get

$$
\begin{aligned}
&\frac{1}{mn}\sum_{i=1}^{n}\sum_{k=1}^{m}\mathbb{E}_{\mathcal{S}_k,\mathcal{S}_k^{(i)},A}\big[\big\|\mathbf{w}_k^{(t+1)}-\widetilde{\mathbf{w}}_k^{(t+1)}\big\|_2^2\big]\\
&\qquad\leq \frac{1}{1-2\eta L(1-\frac{1}{n})}\big\{[1+\frac{p}{n}+(1-\frac{1}{n})\eta_\tau]d(\sigma^2+\mu^2)\big[(1-\frac{1}{m})\lambda^2+\frac{1}{m}\big]+\frac{2}{n}\big[(1+p^{-1})c_{\alpha,1}^2{\eta_\tau}^2\epsilon_{\mathcal{S}}\big]\big\}. \quad \text{(D.23)}
\end{aligned}
$$

□

D.3. Generalization of D-SGD

Proof of Lemma 3.

We denote $A(\mathcal{S})$ as the model produced by algorithm A based on the training dataset $\mathcal{S}$.

To begin with, we can write

$$
\begin{aligned}
\mathbb{E}_{\mathcal{S},A}\big[F(A(\mathcal{S}))-F_{\mathcal{S}}(A(\mathcal{S}))\big] &= \mathbb{E}_{\mathcal{S},\mathcal{S}^{(i)},A}\big[\frac{1}{N}\sum_{i=1}^{N}\big(F(A(\mathcal{S}^{(i)}))-F_{\mathcal{S}}(A(\mathcal{S}))\big)\big]\\
&= \mathbb{E}_{\mathcal{S},\mathcal{S}^{(i)},A}\big[\frac{1}{N}\sum_{i=1}^{N}\big(\mathbb{E}_{\tilde{z}\sim D}(f(A(\mathcal{S}^{(i)});\tilde{z}))-f(A(\mathcal{S});z_i)\big)\big] \quad \text{(D.24)}\\
&= \mathbb{E}_{\mathcal{S},\mathcal{S}^{(i)},A}\big[\frac{1}{N}\sum_{i=1}^{N}\big(f(A(\mathcal{S}^{(i)});z_i)-f(A(\mathcal{S});z_i)\big)\big], \quad \text{(D.25)}
\end{aligned}
$$

where the first line follows from noticing that $\mathbb{E}_{\mathcal{S},A}\big[F(A(\mathcal{S}))\big] = \mathbb{E}_{\mathcal{S}^{(i)},A}\big[F(A(\mathcal{S}^{(i)}))\big]$ and the last identity holds since $A(\mathcal{S}^{(i)})$ is independent of z_i and thus $\mathbb{E}_{\mathcal{S},\mathcal{S}^{(i)},A}\big[\mathbb{E}_{\tilde{z}\sim\mathcal{S}^{(i)}}(f(A(\mathcal{S}^{(i)});\tilde{z})\big] = \mathbb{E}_{\mathcal{S},\mathcal{S}^{(i)},A}\big[f(A(\mathcal{S}^{(i)});z_i)\big]$.

The Hölder continuity of f and the concavity of the $x \mapsto x^{\frac{\alpha}{2}}$ further guarantees

$$\begin{aligned}\mathbb{E}_{\mathcal{S},A}\big[F(A(\mathcal{S})) - F_{\mathcal{S}}(A(\mathcal{S}))\big] &\le \mathbb{E}_{\mathcal{S},\mathcal{S}^{(i)},A}\big[\frac{1}{N}\sum_{i=1}^{N} L\|A(\mathcal{S}) - A(\mathcal{S}^{(i)})\|_2^{\alpha}\big] \\ &= \mathbb{E}_{\mathcal{S},\mathcal{S}^{(i)},A}\big[\frac{L}{N^{1-\frac{\alpha}{2}}}\big(\frac{1}{N}\sum_{i=1}^{N}\|A(\mathcal{S}) - A(\mathcal{S}^{(i)})\|_2^2\big)^{\frac{\alpha}{2}}\big].\end{aligned} \tag{D.26}$$

Finally, consider $\frac{1}{m}\sum_{k=1}^{m}\mathbf{w}_k^{(t)}$ as an output of algorithm A on dataset $\mathcal{S}$, we can complete the proposition by the convexity of vector 2-norm and square function:

$$\begin{aligned}\mathbb{E}_{\mathcal{S},A}\big[F(\frac{1}{m}\sum_{k=1}^{m}\mathbf{w}_k^{(t)}) - F_{\mathcal{S}}(\frac{1}{m}\sum_{k=1}^{m}\widetilde{\mathbf{w}}_k^{(t)})\big] &\le \mathbb{E}_{\mathcal{S},\mathcal{S}^{(i)},A}\big[\frac{L}{(mn)^{1-\frac{\alpha}{2}}}\big(\frac{1}{n}\sum_{i=1}^{n}\|\frac{1}{m}\sum_{k=1}^{m}\mathbf{w}_k^{(t)} - \frac{1}{m}\sum_{k=1}^{m}\widetilde{\mathbf{w}}_k^{(t)}\|_2^2\big)^{\frac{\alpha}{2}}\big] \\ &\le \frac{L}{mn^{1-\frac{\alpha}{2}}}\big\{\frac{1}{mn}\sum_{k=1}^{m}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{S}_k,\mathcal{S}_k^{(i)},A}\big[\|\mathbf{w}_k^{(t)} - \widetilde{\mathbf{w}}_k^{(t)}\|_2^2\big]\big\}^{\alpha/2}.\end{aligned} \tag{D.27}$$

□

Proof of Theorem 4.

We start by rewriting Inequality (D.21) as

$$\begin{aligned}&\frac{1}{mn}\sum_{i=1}^{n}\sum_{k=1}^{m}\mathbb{E}_{\mathcal{S}_k,\mathcal{S}_k^{(i)},A}\big[\|\mathbf{w}_k^{(t+1)} - \widetilde{\mathbf{w}}_k^{(t+1)}\|_2^2\big] \\ &\le \sum_{\tau=0}^{t}C^{t-\tau}\big\{[1+\frac{p}{n}+(1-\frac{1}{n})\eta_\tau]d(\sigma^2+\mu^2)\big[(m-1)\lambda^2+1\big] + \frac{1}{n}\big[2(1+p^{-1})c_{\alpha,1}^2{\eta_\tau}^2\sum_{k=1}^{m}\mathbb{E}_{\mathcal{S}_k,A}\big[F_{\mathcal{S}_k}^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_k^{(\tau)})\big]\big]\big\}.\end{aligned} \tag{D.28}$$

To facilitate subsequent analysis, we denote $T_{\text{dec}} = \sum_{\tau=0}^{t}C^{t-\tau}[1+\frac{p}{n}+(1-\frac{1}{n})\eta_\tau]d(\sigma^2+\mu^2)\big[(m-1)\lambda^2+1\big]$ and $T_{\text{avg}} = \sum_{\tau=0}^{t}C^{t-\tau}\big[2(1+p^{-1})c_{\alpha,1}^2{\eta_\tau}^2\sum_{k=1}^{m}\mathbb{E}_{\mathcal{S}_k,A}\big[F_{\mathcal{S}_k}^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_k^{(\tau)})\big]\big]$.

A combination of Inequality (D.21) and Lemma 3 yields

$$\mathbb{E}_{\mathcal{S},A}\big[F(\frac{1}{m}\sum_{k=1}^{m}\mathbf{w}_k^{(t+1)}) - F_{\mathcal{S}}(\frac{1}{m}\sum_{k=1}^{m}\mathbf{w}_k^{(t+1)})\big] \le \frac{L}{mn^{1-\frac{\alpha}{2}}}\big(\frac{1}{n}T_{\text{avg}} + T_{\text{dec}}\big)^{\alpha/2}. \tag{D.29}$$

Since the inequality $(1+x)^{\frac{\alpha}{2}} \le 2^{\frac{\alpha}{2}} - 1 + x^{\frac{\alpha}{2}} \le 1 + x^{\frac{\alpha}{2}}$ holds for all $x \ge 1$ and $\alpha \in [0,1]$, the right hand side of Inequality (D.28) can be bounded as

$$\begin{aligned}\frac{L}{mn^{1-\frac{\alpha}{2}}}\big(\frac{1}{n}T_{\text{avg}} + T_{\text{dec}}\big)^{\frac{\alpha}{2}} &= \frac{L}{mn^{1-\frac{\alpha}{2}}}\big(\frac{T_{\text{avg}}}{n}\big)^{\frac{\alpha}{2}}\big(1 + T_{\text{dec}}n\big)^{\frac{\alpha}{2}} \le \frac{L}{mn^{1-\frac{\alpha}{2}}}\big(\frac{T_n}{n}\big)^{\frac{\alpha}{2}}[1 + (T_{\text{dec}}n)^{\frac{\alpha}{2}}] \\ &= \frac{LT_{\text{avg}}^{\frac{\alpha}{2}}}{mn}[1 + (T_{\text{dec}}n)^{\frac{\alpha}{2}}] = \frac{L}{m}[\frac{T_{\text{avg}}^{\frac{\alpha}{2}}}{n} + T_{\text{avg}}T_{\text{dec}}^{\frac{\alpha}{2}}n^{\frac{\alpha}{2}-1}].\end{aligned} \tag{D.30}$$

Consequently, the generalization bound of D-SGD can be controlled as

$$\begin{aligned}\mathbb{E}_{\mathcal{S},A}\big[F(\frac{1}{m}\sum_{k=1}^{m}\mathbf{w}_k^{(t+1)}) - F_{\mathcal{S}}(\frac{1}{m}\sum_{k=1}^{m}\mathbf{w}_k^{(t+1)})\big] &\le \frac{L}{m}[\frac{T_{\text{avg}}^{\frac{\alpha}{2}}}{n} + T_{\text{dec}}^{\frac{\alpha}{2}}n^{\frac{\alpha}{2}-1}] \\ &= \frac{L}{N}\big[\sum_{\tau=0}^{t}C^{t-\tau}2(1+p^{-1})c_{\alpha,1}^2{\eta_\tau}^2\epsilon_{\mathcal{S}}\big]^{\frac{\alpha}{2}} + \frac{L\cdot n^{\frac{\alpha}{2}}}{N}\big\{\sum_{\tau=0}^{t}C^{t-\tau}[(1-\frac{1}{m})\lambda^2+\frac{1}{m}]\big\}^{\frac{\alpha}{2}}.\end{aligned} \tag{D.31}$$

where $\epsilon_{\mathcal{S}}$ denotes the upper bound of $\frac{1}{m}\sum_{k=1}^{m}\mathbb{E}_{\mathcal{S}_k,A}\big[F_{\mathcal{S}_k}^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_k^{(t)})\big]\ \forall t$.

If we set the $\eta_t \equiv \eta \leq \frac{1}{2L}$, Inequality (D.31) can be written as

$$\begin{aligned}\mathbb{E}_{\mathcal{S},A}\big[F(\frac{1}{m}\sum_{k=1}^{m}\mathbf{w}_k^{(t+1)}) - F_{\mathcal{S}}(\frac{1}{m}\sum_{k=1}^{m}\mathbf{w}_k^{(t+1)})\big] &\leq \frac{1}{[1-2\eta L(1-\frac{1}{n})]^{\frac{\alpha}{2}}}\{\mathcal{O}(\frac{L\epsilon_{\mathcal{S}}^{\frac{\alpha}{2}}}{N}) + \mathcal{O}(\frac{Ln^{\frac{\alpha}{2}}}{N}[(1-\frac{1}{m})\lambda^2+\frac{1}{m}]^{\frac{\alpha}{2}})\} \\ &\leq \frac{1}{[1-2\eta L(1-\frac{1}{n})]^{\frac{\alpha}{2}}}\{\mathcal{O}(\frac{L\epsilon_{\mathcal{S}}^{\frac{\alpha}{2}}}{N}) + \mathcal{O}(\frac{Ln^{\frac{\alpha}{2}}}{N}(\lambda^{\alpha}+m^{-\frac{\alpha}{2}}))\},\end{aligned} \tag{D.32}$$

which completes the proof.

□

D.4. Implications

Proof of Corollary 5.

Inequality (D.21) shows that in the smooth settings ($\alpha = 1$), the distributed on-average stability of D-SGD is bounded as

$$\begin{aligned}&\frac{1}{mn}\sum_{i=1}^{n}\sum_{k=1}^{m}\mathbb{E}_{\mathcal{S}_k,\mathcal{S}_k^{(i)},A}\big[\|\mathbf{w}_k^{(t+1)} - \widetilde{\mathbf{w}}_k^{(t+1)}\|_2^2\big] \\ &\leq \sum_{\tau=0}^{t}C^{t-\tau}\big\{[1+\frac{p}{n}+(1-\frac{1}{n})\eta_\tau]d(\sigma^2+\mu^2)\big[(1-\frac{1}{m})\lambda_\tau^2+\frac{1}{m}\big] + \frac{2}{n}\big[(1+p^{-1})c_{1,1}^2\eta_\tau{}^2\frac{1}{m}\sum_{k=1}^{m}\mathbb{E}_{\mathcal{S}_k,A}\big[F_{\mathcal{S}_k}(\mathbf{w}_k^{(\tau)})\big]\big]\big\},\end{aligned} \tag{D.33}$$

where $C = 2\eta L(1-\frac{1}{n})$.

Our goal is to prove that the upper bound of the stability increase with the number of iterations that we start to control the "consensus distance".

According to the descent lemma in Koloskova et al. (2020), the empirical risk of the consensus model can be bounded by the consensus distance as follows:

$$\mathbb{E}_A f(\overline{\mathbf{w}}^{(\tau+1)}; z_{\zeta_t}^{(\tau+1)}) \leq \mathbb{E}_A f(\overline{\mathbf{w}}^{(\tau)}; z_{\zeta_t}^{(\tau+1)}) + \eta L^2\underbrace{\frac{1}{m}\sum_{k=1}^{m}\|\mathbf{w}_k^{(\tau)} - \overline{\mathbf{w}}^{(\tau)}\|_2^2}_{\text{consensus distance}} + \mathbb{E}_A\frac{L}{n}\eta^2\underbrace{\|\frac{1}{m}\sum_{k=1}^{m}\nabla f(\mathbf{w}_k^{(\tau)}; z_{\zeta_\tau}^{(\tau)})\|_2^2}_{\text{norm of the averaged gradient}}, \tag{D.34}$$

where $\overline{\mathbf{w}}^{(\tau)} = \frac{1}{m}\sum_{k=1}^{m}\mathbf{w}_k^{(\tau)}$.

Due to the fact that the gradient of f w.r.t. the first parameter is bounded by B and the square of the vector 2-norm $\|\cdot\|_2^2$ is convex, we have

$$\frac{1}{m}\sum_{k=1}^{m}\mathbb{E}_{\mathcal{S}_k,A}\big[F_{\mathcal{S}_k}(\overline{\mathbf{w}}^{(\tau)})\big] \leq \eta L^2\sum_{\nu=0}^{\tau}\frac{1}{m}\sum_{k=1}^{m}\|\overline{\mathbf{w}}^{(\nu)} - \mathbf{w}_k^{(\nu)}\|_2^2 + \frac{L}{m}\eta^2\tau B^2. \tag{D.35}$$

To connect the stability upper bound in Inequality (D.33), we perform the Taylor expansion of $\frac{1}{m}\sum_{k=1}^{m}\mathbb{E}_{\mathcal{S}_k,A}\big[F_{\mathcal{S}_k}(\cdot)\big]$ around $\overline{\mathbf{w}}^{(\tau)}$:

$$\begin{aligned}\frac{1}{m}\sum_{k=1}^{m}\mathbb{E}_{\mathcal{S}_k,A}\big[F_{\mathcal{S}_k}(\mathbf{w}_k^{(\tau)})\big] =& \frac{1}{m}\sum_{k=1}^{m}\mathbb{E}_{\mathcal{S}_k,A}\big[F_{\mathcal{S}_k}(\overline{\mathbf{w}}^{(\tau)})\big] + \frac{1}{m}\sum_{k=1}^{m}\big[\mathbb{E}_{\mathcal{S}_k,A}\big[\nabla F_{\mathcal{S}_k}(\overline{\mathbf{w}}^{(\tau)})\big]\big]^{\mathsf{T}}(\mathbf{w}_k^{(\tau)} - \overline{\mathbf{w}}^{(\tau)}) \\ &+ (\mathbf{w}_k^{(\tau)} - \overline{\mathbf{w}}^{(\tau)})^{\mathsf{T}}\frac{1}{m}\sum_{k=1}^{m}\mathbb{E}_{\mathcal{S}_k,A}\big[\nabla^2F_{\mathcal{S}_k}(\overline{\mathbf{w}}^{(\tau)})\big](\mathbf{w}_k^{(\tau)} - \overline{\mathbf{w}}^{(\tau)}) + \mathcal{O}(\|\mathbf{w}_k^{(\tau)} - \overline{\mathbf{w}}^{(\tau)}\|_2^3).\end{aligned} \tag{D.36}$$

According to Assumption A.1, the gradient of $F_{\mathcal{S}_k}$ w.r.t. the first parameter is bounded by B. Consequently, the averaged empirical loss $\frac{1}{m}\sum_{k=1}^m \mathbb{E}_{\mathcal{S}_k,A}\big[F_{\mathcal{S}_k}(\mathbf{w}_k^{(\tau)})\big]$ can be bounded as

$$\begin{aligned}
\frac{1}{m}\sum_{k=1}^m \mathbb{E}_{\mathcal{S}_k,A}\big[F_{\mathcal{S}_k}(\mathbf{w}_k^{(\tau)})\big] &\le \frac{1}{m}\sum_{k=1}^m \mathbb{E}_{\mathcal{S}_k,A}\big[F_{\mathcal{S}_k}(\overline{\mathbf{w}}^{(\tau)})\big] + \underbrace{\|\frac{1}{m}\sum_{k=1}^m(\mathbf{w}_k^{(\tau)}-\overline{\mathbf{w}}^{(\tau)})\|_2}_{=0}\underbrace{\|\mathbb{E}_{\mathcal{S}_k,A}\Big[\nabla F_{\mathcal{S}_k}(\overline{\mathbf{w}}^{(\tau)})\Big]\|_2}_{\le B} \\
&+ (\mathbf{w}_k^{(\tau)}-\overline{\mathbf{w}}^{(\tau)})^\mathsf{T}\frac{1}{m}\sum_{k=1}^m \mathbb{E}_{\mathcal{S}_k,A}\big[\nabla^2 F_{\mathcal{S}_k}(\overline{\mathbf{w}}^{(\tau)})\big](\mathbf{w}_k^{(\tau)}-\overline{\mathbf{w}}^{(\tau)}) + \mathcal{O}(\|\mathbf{w}_k^{(\tau)}-\overline{\mathbf{w}}^{(\tau)}\|_2^3) \\
&\le \frac{1}{m}\sum_{k=1}^m \mathbb{E}_{\mathcal{S}_k,A}\big[F_{\mathcal{S}_k}(\overline{\mathbf{w}}^{(\tau)})\big] + \underbrace{L\frac{1}{m}\sum_{k=1}^m\|\mathbf{w}_k^{(\tau)}-\overline{\mathbf{w}}^{(\tau)}\|_2^2}_{\text{consensus distance}} + \mathcal{O}(\|\mathbf{w}_k^{(\tau)}-\overline{\mathbf{w}}^{(\tau)}\|_2^3).
\end{aligned} \tag{D.37}$$

The last inequality holds since the smooth condition

$$\|\nabla f(x)-\nabla f(y)\|_2 \le L\|x-y\|_2 \Longleftrightarrow \nabla^2 f \preceq LI, \tag{D.38}$$

and thus we have

$$\begin{aligned}
&(\mathbf{w}_k^{(\tau)}-\overline{\mathbf{w}}^{(\tau)})^\mathsf{T}\frac{1}{m}\sum_{k=1}^m \mathbb{E}_{\mathcal{S}_k,A}\big[\nabla^2 F_{\mathcal{S}_k}(\overline{\mathbf{w}}^{(\tau)})\big](\mathbf{w}_k^{(\tau)}-\overline{\mathbf{w}}^{(\tau)}) \\
&\qquad \le L\frac{1}{m}\sum_{k=1}^m(\mathbf{w}_k^{(\tau)}-\overline{\mathbf{w}}^{(\tau)})^\mathsf{T} I(\mathbf{w}_k^{(\tau)}-\overline{\mathbf{w}}^{(\tau)}) = L\frac{1}{m}\sum_{k=1}^m\|\mathbf{w}_k^{(\tau)}-\overline{\mathbf{w}}^{(\tau)}\|_2^2.
\end{aligned} \tag{D.39}$$

If we omit the third-order difference, a combination of Inequality (D.35) and Inequality (D.37) provides

$$\frac{1}{m}\sum_{k=1}^m \mathbb{E}_{\mathcal{S}_k,A}\big[F_{\mathcal{S}_k}(\mathbf{w}_k^{(\tau)})\big] \le L\frac{1}{m}\sum_{k=1}^m\|\mathbf{w}_k^{(\tau)}-\overline{\mathbf{w}}^{(\tau)}\|_2^2 + \frac{\eta L^2}{m}\sum_{\nu=0}^{\tau}\sum_{k=1}^m\|\overline{\mathbf{w}}^{(\nu)}-\mathbf{w}_k^{(\nu)}\|_2^2 + \frac{L}{m}\eta^2\tau B^2. \tag{D.40}$$

This inequality would suffice to prove that the distributed on-average stability increase with the accumulation of the "consensus distance".

Suppose that the consensus distance is controlled below the critical consensus distance Γ^2 from t_Γ-th iterate to the end of the training. For simplicity, we make a mild assumption that the consensus distance $\Gamma^2 \le \frac{1}{m}\sum_{k=1}^m\|\mathbf{w}_k^{(t)}-\overline{\mathbf{w}}^{(t)}\|_2^2 \le K^2$ if $t \le t_\Gamma$. Therefore, the averged empirical risk at τ-th iterate can be bounded as

$$\begin{aligned}
\frac{1}{m}\sum_{k=1}^m \mathbb{E}_{\mathcal{S}_k,A}\big[F_{\mathcal{S}_k}(\mathbf{w}_k^{(\tau)})\big] &\le L\Gamma^2 + \frac{\eta L^2}{m}\sum_{\nu=0}^{t_\Gamma}\sum_{k=1}^m\|\overline{\mathbf{w}}^{(\nu)}-\mathbf{w}_k^{(\nu)}\|_2^2 + \eta L^2(\tau-t_\Gamma)\Gamma^2 + \frac{L}{m}\eta^2\tau B^2 \\
&\le L\Gamma^2 + \eta L^2\sum_{\nu=0}^{t_\Gamma}\big(\frac{1}{m}\sum_{k=1}^m\|\overline{\mathbf{w}}^{(\nu)}-\mathbf{w}_k^{(\nu)}\|_2^2-\Gamma^2\big) + (\eta L^2\Gamma^2+\frac{L}{m}\eta^2B^2)\cdot\boldsymbol{\tau} \\
&\le L\Gamma^2 + \eta L^2(t_\Gamma+1)\big(K^2-\Gamma^2\big) + (\eta L^2\Gamma^2+\frac{L}{m}\eta^2B^2)\cdot\boldsymbol{\tau},
\end{aligned} \tag{D.41}$$

if τ is greater than t_Γ; and

$$\begin{aligned}
\frac{1}{m}\sum_{k=1}^m \mathbb{E}_{\mathcal{S}_k,A}\big[F_{\mathcal{S}_k}(\mathbf{w}_k^{(\tau)})\big] &\le L\frac{1}{m}\sum_{k=1}^m\|\mathbf{w}_k^{(\tau)}-\overline{\mathbf{w}}^{(\tau)}\|_2^2 + \frac{\eta L^2}{m}\sum_{\nu=0}^{\tau}\sum_{k=1}^m\|\overline{\mathbf{w}}^{(\nu)}-\mathbf{w}_k^{(\nu)}\|_2^2 + \frac{L}{m}\eta^2B^2\tau \\
&\le LK^2 + (\eta L^2K^2+\frac{L}{m}\eta^2B^2)\cdot\boldsymbol{\tau},
\end{aligned} \tag{D.42}$$

if if τ is smaller than t_Γ.

Consequently,

$$
\begin{aligned}
G(t_\Gamma) = \sum_{\tau=0}^{t} C^{t-\tau}\frac{1}{m}\sum_{k=1}^{m}\mathbb{E}_{\mathcal{S}_k,A}\big[F_{\mathcal{S}_k}(\mathbf{w}_k^{(\tau)})\big] &= (\sum_{\tau=0}^{t_\Gamma} + \sum_{\tau=t_\Gamma+1}^{t})C^{t-\tau}\frac{1}{m}\sum_{k=1}^{m}\mathbb{E}_{\mathcal{S}_k,A}\big[F_{\mathcal{S}_k}(\mathbf{w}_k^{(\tau)})\big] \\
&\leq \frac{L}{m}\eta^2 B^2 \sum_{\tau=0}^{t}\tau C^{t-\tau} + \eta L^2 K^2 \sum_{\tau=0}^{t_\Gamma}\tau C^{t-\tau} + \eta L^2\Gamma^2 \sum_{\tau=t_\Gamma+1}^{t}\tau C^{t-\tau} \\
&\quad + LK^2(\sum_{\tau=0}^{t_\Gamma} C^{t-\tau}) + [L\Gamma^2 + \eta L^2(t_\Gamma+1)\big(K^2-\Gamma^2\big)](\sum_{\tau=t_\Gamma+1}^{t} C^{t-\tau}).
\end{aligned} \tag{D.43}
$$

Recall that **our goal** is to prove that $G(t_\Gamma)$ increase with t_Γ.

Due to the fact that $\sum_{\tau=0}^{t}\tau C^{t-\tau} = \sum_{\tau=0}^{t}(t-\tau)C^{\tau} \leq t\sum_{\tau=0}^{t} C^{\tau}$, we can obtain

$$
\begin{aligned}
G(t_\Gamma) \leq & \frac{t\frac{L}{m}\eta^2 B^2}{1-C} + \eta L^2 K^2 \sum_{\tau=0}^{t_\Gamma}\tau C^{t-\tau} + \eta L^2\Gamma^2 \sum_{\tau=t_\Gamma+1}^{t}\tau C^{t-\tau} \\
& + \frac{C^t L K^2}{1-C} + [L\Gamma^2 + \eta L^2(t_\Gamma+1)\big(K^2-\Gamma^2\big)]\frac{C^t C^{t_\Gamma+1}}{1-C}.
\end{aligned} \tag{D.44}
$$

Since the finite sum of the arithmetico-geometric sequence can be written as

$$
\sum_{\tau=0}^{t}\tau C^{-\tau} = [\frac{t}{C^{-1}-1} - \frac{1}{(C^{-1}-1)^2}]C^{-(t+1)} + [\frac{1}{(C^{-1}-1)^2} + \frac{1}{C^{-1}-1}], \tag{D.45}
$$

we can upper bound $G(t_\Gamma)$ as follows:

$$
\begin{aligned}
G(t_\Gamma) \leq & \frac{t\frac{L}{m}\eta^2 B^2}{1-C} + \eta L^2\Gamma^2 C^t\{[\frac{t}{C^{-1}-1} - \frac{1}{(C^{-1}-1)^2}]C^{-(t+1)} + [\frac{1}{(C^{-1}-1)^2} + \frac{1}{C^{-1}-1}]\} \\
& + \eta L^2(K^2-\Gamma^2)C^t\{[\frac{t}{C^{-1}-1} - \frac{1}{(C^{-1}-1)^2}]C^{-(\boldsymbol{t}_{\boldsymbol{\Gamma}}+1)} + [\frac{1}{(C^{-1}-1)^2} + \frac{1}{C^{-1}-1}]\} \\
& + \frac{C^t L K^2}{1-C} + [L\Gamma^2 + \eta L^2(\boldsymbol{t}_{\boldsymbol{\Gamma}}+1)\big(K^2-\Gamma^2\big)]\frac{C^t C^{\boldsymbol{t}_{\boldsymbol{\Gamma}}+1}}{1-C}.
\end{aligned} \tag{D.46}
$$

Rewrite the inequality above, then we arrive at

$$
\begin{aligned}
G(t_\Gamma) \leq & \frac{t\frac{L}{m}\eta^2 B^2}{1-C} + \eta L^2\Gamma^2 C^t[\frac{t}{C^{-1}-1} - \frac{1}{(C^{-1}-1)^2}]C^{-(t+1)} + \frac{C^t L K^2}{1-C} \\
& + \eta L^2(K^2-\Gamma^2)C^t\{[\frac{t}{C^{-1}-1} - \frac{1}{(C^{-1}-1)^2}]C^{-(\boldsymbol{t}_{\boldsymbol{\Gamma}}}+1) + \boldsymbol{t}_{\boldsymbol{\Gamma}}\frac{C^{\boldsymbol{t}_{\boldsymbol{\Gamma}}+1}}{1-C}\} + [L\Gamma^2 + \eta L^2\big(K^2-\Gamma^2\big)]C^t\frac{C^{\boldsymbol{t}_{\boldsymbol{\Gamma}}+1}}{1-C}.
\end{aligned} \tag{D.47}
$$

One can prove that if $t \geq \frac{-C}{2\ln C}$, the upper bound of $G(t_\Gamma)$ will be a monotonically increasing function of t_Γ. Consequently, we can conclude that the distributed on-average stability bound and the generalization bound of D-SGD increase monotonically with t_Γ if the total number of iterations satisfies $t \geq \frac{-C}{2\ln C}$.

$\square$