Stability and Generalization for Decentralized Markov SGD

Jiahuan Wang[1] , **Ziqing Wen**[1] , **Ping Luo**[1] , **Dongsheng Li**[1] , **Tao Sun**[1*]
[1]National Key Laboratory of Parallel and Distributed Computing,
National University of Defense Technology, Changsha, 410073, China
{wangjiahuan, zqwen, luoping, dsli}@ nudt.edu.cn, suntao.saltfish@outlook.com

Abstract

Stochastic gradient methods are central to large-scale learning, yet their generalization theory typically relies on independent sampling assumptions. In many practical applications, data are generated by Markov chains and learning is performed in a decentralized manner, which introduces significant analytical challenges. In this work, we investigate the stability and generalization of decentralized stochastic gradient descent (SGD) and stochastic gradient descent ascent (SGDA) under Markov chain sampling. Leveraging a stability-based framework, we characterize how Markovian dependence and decentralized communication jointly influence generalization behavior. Our analysis captures the effects of network topology, Markov chain mixing properties, and primal–dual dynamics. We establish non-asymptotic generalization bounds for both algorithms, extending existing results on Markov stochastic gradient methods to decentralized and minimax settings.

1 Introduction

Stochastic gradient methods form the backbone of modern large-scale machine learning, owing to their simplicity and scalability [Li *et al.*, 2014; Lin *et al.*, 2018]. In many practical applications, however, data samples are not independently drawn at each iteration. Instead, they are generated sequentially by a stochastic process, such as a Markov chain, which introduces temporal dependence into the training procedure. This phenomenon naturally arises in reinforcement learning [Wai *et al.*, 2018; Doan *et al.*, 2020], recommendation systems [Li *et al.*, 2010], and distributed data collection [Lopes and Sayed, 2007; Duchi *et al.*, 2011; Chen *et al.*, 2014; Madry *et al.*, 2017; Shah and Avrachenkov, 2018], where i.i.d. sampling is either infeasible or prohibitively expensive.

In parallel, the increasing scale of data has driven the adoption of decentralized optimization architectures, where multiple workers collaboratively train a shared model using local computations and limited communication [Nedic and Ozdaglar, 2009; Sundhar Ram *et al.*, 2010]. Decentralized stochastic gradient descent (D-SGD) and its variants [Lian *et al.*, 2017; Lian *et al.*, 2018] have attracted significant attention due to their robustness, communication efficiency, and suitability for large networks. While the optimization properties of decentralized algorithms are now relatively well understood [Shi *et al.*, 2015; Yuan *et al.*, 2016; Lian *et al.*, 2017; Koloskova *et al.*, 2020; Sun *et al.*, 2021; Yuan *et al.*, 2023], their statistical behavior—particularly generalization—remains less explored under realistic sampling assumptions.

Recent progress has begun to address these challenges from different perspectives. Stability-based analyses have established rigorous generalization guarantees for stochastic gradient methods under Markovian sampling, highlighting the role of algorithmic stability in controlling the discrepancy between empirical and population risks [Wang *et al.*, 2022]. Separately, decentralized SGD with Markov chain sampling has been studied primarily from an optimization viewpoint [Sun *et al.*, 2022; Sun *et al.*, 2023; Koloskova *et al.*, 2020], focusing on convergence rates and consensus errors. More recently, stability and generalization properties of decentralized stochastic gradient descent (ascent) (SGDA) algorithms have been investigated under standard i.i.d. sampling assumptions [Sun *et al.*, 2021; Zhu *et al.*, 2022; Bars *et al.*, 2024; Zhu *et al.*, 2023; Zeng and Lei, 2025]. Despite these advances, a unified theoretical understanding of decentralized stochastic gradient methods under Markovian data dependence is still lacking. In particular, existing works either (i) analyze Markov chain sampling in centralized settings, (ii) study decentralized algorithms assuming independent samples, or (iii) focus on optimization performance without addressing generalization. As a result, fundamental questions remain open:

> *Can one obtain sharp generalization guarantees for decentralized SGD and SGDA when data are generated by Markov chains?*

Motivated by these gaps, this paper investigates the theoretical foundations of decentralized stochastic gradient descent and descent–ascent algorithms with Markov chain sampling, and aims to provide a rigorous characterization of their gen-

*Corresponding author.

eralization behavior. Building on the stability framework, we develop novel stability bounds for both decentralized SGD and SGDA under Markovian data dependence. Our analysis explicitly captures the interplay between algorithmic stability, network consensus, and the mixing properties of the underlying Markov chains. We generalize stability-based generalization theory for Markov chain stochastic gradient methods to decentralized settings, and complement prior studies on decentralized optimization by providing statistically meaningful guarantees. To the best of our knowledge, this is the first work to systematically study the stability and generalization of decentralized SGD and SGDA under Markovian sampling.

1.1 Differences and Technical Challenges

Unlike existing analyses of decentralized SGD under i.i.d. sampling [Richards and Rebeschini, 2020; Sun *et al.*, 2021; Zhu *et al.*, 2022; Bars *et al.*, 2024; Zeng and Lei, 2025], this work allows each worker to access data through a Markov chain, which introduces temporal dependence and breaks the sample-wise independence typically used in stability arguments. Compared with centralized Mc-SGD [Wang *et al.*, 2022], decentralization further introduces consensus dynamics, leading to an additional error accumulation through network disagreement. A key technical challenge is to control the interaction between these two effects without imposing mixing-time or spectral assumptions on the Markov chain. This is resolved by exploiting aggregation identities that depend only on the update structure, rather than on the specific sampling mechanism. As a result, i.i.d.-type stability bounds can still be recovered in the decentralized Markovian setting.

1.2 Contributions

Our main contributions can be summarized as follows.

- **Stability analysis of decentralized SGD with Markovian sampling.** We establish on-average stability bounds for decentralized SGD under Markov chain sampling (See Theorem 1). The resulting bounds match those of decentralized SGD under i.i.d. sampling up to the same consensus-dependent terms, showing that Markovian sampling does not incur additional stability degradation in the decentralized setting.
- **Excess risk and generalization guarantees.** Based on the stability analysis, we derive excess risk and generalization bounds for Decentralized Markov SGD (DMc-SGD) under both smooth and non-smooth losses (See Theorem 3-4). Our results explicitly characterize the roles of network connectivity, Markov chain mixing, and stepsize selection, and recover known centralized and i.i.d.-based decentralized rates as special cases.
- **Stability and generalization of decentralized SGDA with Markovian sampling.** We extend the stability-based framework to decentralized stochastic gradient descent ascent (DMc-SGDA) for convex–concave minimax problems. We obtain on-average argument stability bounds and corresponding generalization guarantees for both weak primal–dual risk and primal population risk (See Theorem 6-7), covering smooth setting.

2 Related Work

Generalization Analysis of D-SGD. Recent studies have investigated the generalization behavior of decentralized learning primarily through stability-based analyses, which relate the generalization gap to algorithmic sensitivity and network-induced perturbations. These works reveal how communication topology and consensus dynamics influence generalization, though the resulting bounds often grow with the number of iterations unless additional structural assumptions, such as strong convexity, are imposed. Representative results include stability- and complexity-based bounds for D-SGD in both smooth and non-smooth settings [Richards and Rebeschini, 2020; Sun *et al.*, 2021], analyses of topology-dependent effects under on-average stability frameworks [Zhu *et al.*, 2022], and improvements that recover rates comparable to centralized SGD [Bars *et al.*, 2024]. Extensions to heterogeneous data distributions and weaker regularity conditions have also been considered [Ye *et al.*, 2025; Zeng and Lei, 2025]. A detailed comparison of these results is provided in Table 1. Related generalization analyses have further been developed for several D-SGD variants, including asynchronous implementations [Deng *et al.*, 2023], minibatch methods [Wang and Chen, 2024], decentralized SGDA [Zhu *et al.*, 2023], and zeroth-order optimization schemes [Wang and Chen, 2024; Hu *et al.*, 2025].

Markov Chain Gradient Descent. Beyond stability and generalization, Markov chain stochastic gradient methods have been extensively investigated from an optimization perspective, motivated by scenarios where independent sampling is infeasible and data are generated sequentially [Ram *et al.*, 2009; Tadić and Doucet, 2011; Duchi *et al.*, 2012]. A central theme in this line of work is to understand how the temporal dependence induced by Markovian sampling affects convergence rates. Convergence properties of SGD with Markovian sampling for convex and non-convex problems were investigated in Sun *et al.* [2021]. The same work also developed convergence guarantees for non-convex problems under Markovian sampling. Decentralized SGD with gradients sampled from non-reversible Markov chains was further studied in Sun *et al.* [2023], which characterized the impact of non-reversibility on convergence behavior. Acceleration techniques have also been considered in this setting. Doan *et al.* [2020] analyzed accelerated ergodic Markov chain SGD for both convex and non-convex objectives, while Doan [2022] further relaxed standard assumptions by deriving convergence rates without requiring bounded gradients.

3 Preliminaries

We study a decentralized learning setting with m computing nodes, where each node stores n training samples. The full dataset is denoted by $S = \{S_1, S_2, \cdots, S_m\}$, and each local dataset $S_r = \{Z_{1(1)}, \cdots, Z_{k(r)}, \cdots, Z_{n(m)}\}$ consists of n samples drawn from an unknown distribution $\mathcal{D}$. A learning algorithm $\mathcal{A}$, such as (decentralized) SGD, maps the dataset S to a model parameter in a hypothesis space $\mathcal{W}$. The performance of a model $\mathbf{w} \in \mathcal{W}$ is evaluated through a loss function $f(\mathbf{w}; Z)$, which induces the population risk (expected risk) $\min_{w\in\mathcal{W}} R(\mathbf{w}) := \mathbb{E}_{Z\sim\mathcal{D}}[f(\mathbf{w}; Z)]$. Since the data-

Update Type	Data sampling	Reference	Situation	Bounds
GtC (5)	i.i.d	[Richards and Rebeschini, 2020]	$\eta \leq 2/\beta$	$\mathcal{O}\left(1/\sqrt{mn}\right)$
	Markov chain	Ours (Theorem 8)	$\eta \leq 2/\beta$	$\mathcal{O}\left(1/\sqrt{mn}\right)$
CtG (6)	i.i.d	[Sun *et al.*, 2021]	$\eta \leq 2/\beta$	$\mathcal{O}\left(1/\sqrt{mn}\right)$
		[Bars *et al.*, 2024]	$\eta \leq 2\min_k P_{kk}/\beta$	$\mathcal{O}\left(1/\sqrt{mn}\right)$
	Markov chain	Ours (Theorem 2)	$\eta \leq 2/\beta$	$\mathcal{O}\left(1/\sqrt{mn}\right)$

Table 1: Generalization Bounds for D-SGD Problem. (GtC: Gradient-then-Consensus, CtG: Consensus-then-Gradient. P_{kk}: the k-th diagonal element of the gossip matrix; η: stepsize; β: smoothness property.)

generating distribution is inaccessible, the algorithm instead minimizes its empirical analogue computed over the observed samples. In the decentralized setting, this empirical objective can be written as

$$R_S(\mathbf{w}) = \frac{1}{m}\sum_{r=1}^{m} R_{S_r} = \frac{1}{mn}\sum_{r=1}^{m}\sum_{k=1}^{n} f\left(\mathbf{w}; Z_{k(r)}\right), \quad (1)$$

with R_{S_r} denoting the corresponding local empirical risk at node/machine r.

Although $\mathcal{A}(S)$ may fit the training data well, good empirical performance does not necessarily translate to good population performance. This motivates the study of the expected gap between the population and empirical risks evaluated at the algorithm output, namely

$$\epsilon_{\text{gen}} := \mathbb{E}_{S,\mathcal{A}}\left[R(\mathcal{A}(S)) - R_S(\mathcal{A}(S))\right]. \quad (2)$$

which we refer to as the generalization error. Beyond generalization, we are also interested in how far the learned model is from the optimal population solution.

Definition 1. *The excess generalization error of the learned model, $\epsilon_{excess} := \mathbb{E}_{S,\mathcal{A}}\left[R(\mathcal{A}(S)) - R(\mathbf{w}^*)\right]$, where $\mathbf{w}^*$ is denote the minimizer of $R(\cdot)$. This term can then be expressed as*

$$\mathbb{E}_{S,\mathcal{A}}\left[R(\mathcal{A}(S)) - R(\mathbf{w}^*)\right] = \underbrace{\mathbb{E}_{S,\mathcal{A}}\left[R(\mathcal{A}(S)) - R_S(\mathcal{A}(S))\right]}_{\text{Generalization Error}}$$
$$+\underbrace{\mathbb{E}_{S,\mathcal{A}}\left[R_S(\mathcal{A}(S)) - R_S(\mathbf{w}_S^*)\right]}_{\epsilon_{opt}\text{: Optimization Error}} + \underbrace{\mathbb{E}_{S,\mathcal{A}}\left[R_S(\mathbf{w}_S^*) - R(\mathbf{w}^*)\right]}_{\text{Test Error}},$$

where $\mathbf{w}^\star$ be the minimizer of $R_S(\cdot)$.

Remark 1. *Since $\mathbb{E}\left[R_S\left(\mathbf{w}^*\right)\right] = \mathbb{E}\left[R\left(\mathbf{w}^*\right)\right]$ and the empirical risk minimizer satisfies $R_S(\mathbf{w}_S^*) \leq R_S(\mathbf{w}^*)$, the last item is guaranteed to be non-positive. Consequently, most theoretical analyses can be reduced to concentrating on the first two terms, the generalization error and the optimization error.*

We now turn to decentralized minimax learning problems, which naturally arise in adversarial learning and robust optimization [Goodfellow *et al.*, 2014; Madry *et al.*, 2017]. In this setting, the objective depends on a pair of variables: a primal variable $\mathbf{w} \in \mathcal{W} \subset \mathbb{R}^d$ and a dual variable $\mathbf{v} \in \mathcal{V} \subset \mathbb{R}^d$. Given a loss function $f : \mathcal{W} \times \mathcal{V} \times \mathcal{Z} \to [0, \infty)$, the population-level minimax objective is defined as

$$\min_{\mathbf{w}\in\mathcal{W}} \max_{\mathbf{v}\in\mathcal{V}} R(\mathbf{w}, \mathbf{v}),\ R(\mathbf{w}, \mathbf{v}) := \mathbb{E}_{z\sim\mathcal{D}}\left[f\left(\mathbf{w}, \mathbf{v}; Z\right)\right]. \quad (3)$$

As before, the underlying distribution $\mathcal{D}$ is unknown, and learning algorithms operate on a finite dataset distributed across m workers. The corresponding empirical objective is therefore given by

$$R_S(\mathbf{w}, \mathbf{v}) = \frac{1}{mn}\sum_{r=1}^{m}\sum_{k=1}^{n} f\left(\mathbf{w}, \mathbf{v}; Z_{k(r)}\right). \quad (4)$$

In contrast to standard minimization, minimax learning involves two coupled variables, which makes generalization more nuanced. To streamline notation, we write the (possibly randomized) output of algorithm $\mathcal{A}$ as $\mathcal{A}(S) = (\mathcal{A}_{\mathbf{w}}(S), \mathcal{A}_{\mathbf{v}}(S))$. We focus on two notions of performance that will be used throughout the paper [Farnia and Ozdaglar, 2021; Lei *et al.*, 2021b; Zhu *et al.*, 2023].

Definition 2 (Weak Primal-Dual (PD) Risk)**.** *The weak Primal-Dual population risk $\Delta^{\mathbf{w}}\left(\mathcal{A}_{\mathbf{w}}, \mathcal{A}_{\mathbf{v}}\right)$:*

$$\max_{\mathbf{v}\in\mathcal{V}} \mathbb{E}\left[R\left(\mathcal{A}_{\mathbf{w}}(S), \mathbf{v}\right)\right] - \min_{\mathbf{w}\in\mathcal{W}} \mathbb{E}\left[R\left(\mathbf{w}, \mathcal{A}_{\mathbf{v}}(S)\right)\right].$$

The weak PD empirical risk $\Delta^{\mathbf{w}}_{\text{emp}}\left(\mathcal{A}_{\mathbf{w}}, \mathcal{A}_{\mathbf{v}}\right)$:

$$\max_{\mathbf{v}\in\mathcal{V}} \mathbb{E}\left[R_S\left(\mathcal{A}_{\mathbf{w}}(\mathcal{S}), \mathbf{v}\right)\right] - \min_{\mathbf{w}\in\mathcal{W}} \mathbb{E}\left[R_S\left(\mathbf{w}, \mathcal{A}_{\mathbf{v}}(S)\right)\right].$$

The weak PD generalization error of the model:

$$\epsilon^{\mathbf{w}}_{\text{gen}} = \Delta^{\mathbf{w}}\left(\mathcal{A}_{\mathbf{w}}, \mathcal{A}_{\mathbf{v}}\right) - \Delta^{\mathbf{w}}_{\text{emp}}\left(\mathcal{A}_{\mathbf{w}}, \mathcal{A}_{\mathbf{v}}\right).$$

Definition 3 (Primal Risk)**.** *The primal population and empirical risks of $\mathcal{A}(\mathcal{S})$:*

$$F\left(\mathcal{A}_{\mathbf{w}}(S)\right) = \max_{\mathbf{v}\in\mathcal{V}} R\left(\mathcal{A}_{\mathbf{w}}(\mathcal{S}), \mathbf{v}\right),$$
$$F_{\mathcal{S}}\left(\mathcal{A}_{\mathbf{w}}(S)\right) = \max_{\mathbf{v}\in\mathcal{V}} R_S\left(\mathcal{A}_{\mathbf{w}}(\mathcal{S}), \mathbf{v}\right).$$

The primal generalization error is defined as

$$\epsilon^{\text{P}}_{\text{gen}} = \mathbb{E}_{S,\mathcal{A}}\left[F\left(\mathcal{A}_{\mathbf{w}}(S)\right) - F_S\left(\mathcal{A}_{\mathbf{w}}(S)\right)\right].$$

The excess primal population risk of the model:

$$\epsilon^{\text{P}}_{\text{excess}} = \mathbb{E}_{S,\mathcal{A}}\left[F\left(\mathcal{A}_{\mathbf{w}}(S)\right) - \min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w})\right].$$

3.1 Decentralized SGD with Markov Sampling

We consider a decentralized optimization framework based on stochastic gradient descent, originally proposed in Lian *et al.* [2017], and adopt its projected variant to handle constrained parameter domains. In this setting, a network of m

nodes cooperatively solves a learning problem while maintaining local model copies. Communication between nodes is governed by a weighted graph, through which neighboring nodes exchange information at each iteration. At a high level, each iteration of the algorithm consists of two stages. First, nodes perform a consensus operation by aggregating model information from their neighbors. Subsequently, each node updates its local model using a stochastic gradient computed from a data sample generated by a local Markov chain, followed by a projection step to ensure feasibility. This procedure naturally extends decentralized SGD to scenarios where data are not independently sampled.

The complete procedure, referred to as Decentralized Markov Stochastic Gradient Descent (Ascent), is summarized in Algorithm 1. For each node i, the algorithm initializes local primal (and dual) variables and iteratively updates them using a combination of communication and stochastic gradient steps. After T iterations, the network outputs the averaged primal (and dual) iterates. The communication structure of

Algorithm 1 Decentralized Markov Stochastic Gradient Descent (Ascent)

Input: Initialize $\forall i, \mathbf{w}^0(i) = \mathbf{w}^0, (\mathbf{v}^0(i) = \mathbf{v}^0)$, stepsizes $\{\eta_t\}_{t=1}^T$,weight matrix P and the iteration number T.

for $t = 1, 2, \cdots, T$ **do**
  for $i = 1, 2, \cdots, m$ **do**
    Sample $Z_{j_t(i)}$ via a Markov chain
    $\mathbf{w}_{t+\frac{1}{2}}^{(i)} = \sum_{l=1}^m P_{il}\mathbf{w}_t^{(l)}$, $(\mathbf{v}_{t+\frac{1}{2}}^{(i)} = \sum_{l=1}^m P_{il}\mathbf{v}_t^{(l)})$
    $\mathbf{w}_{t+1}^{(i)} = \mathbf{P}_{\mathcal{W}}\left(\mathbf{w}_{t+\frac{1}{2}}^{(i)} - \eta_t \nabla_{\mathbf{w}} f(\mathbf{w}_t^{(i)}; Z_{j_t(i)})\right)$
    $(\mathbf{v}_{t+1}^{(i)} = \mathbf{P}_{\mathcal{V}}\left(\mathbf{v}_{t+\frac{1}{2}}^{(i)} + \eta_t \nabla_{\mathbf{v}} f(\mathbf{v}_t^{(i)}; Z_{j_t(i)})\right))$
  end for
end for

Output: $\bar{\mathbf{w}}_{T+1} = \frac{1}{m}\sum_{i=1}^m \mathbf{w}_{T+1}^{(i)}$, $\bar{\mathbf{v}}_{T+1} = \frac{1}{m}\sum_{i=1}^m \mathbf{v}_{T+1}^{(i)}$

the decentralized network is encoded by a matrix $P \in R^{m\times m}$, which specifies how information is mixed across nodes. The matrix P is assumed to satisfy the following properties: (1) P is symmetric; (2) all entries satisfy $P_{ij} \in [0, 1]$; (3) P is doubly stochastic, i.e., $\mathbf{1}_m^T P = \mathbf{1}_m^T$ and $P\mathbf{1}_m = \mathbf{1}_m$.

Gossip Matrix. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$ denote the eigenvalues of P, and define $\gamma := 1 - \lambda = 1 - \max\{|\lambda_2(P)|, |\lambda_m(P)|\}$ [Sun *et al.*, 2021; Zhu *et al.*, 2022; Deng *et al.*, 2023; Bars *et al.*, 2024; Wang and Chen, 2024; Zeng and Lei, 2025]. The quantity $\lambda \in [0, 1)$ characterizes the speed at which consensus is achieved across the network. In particular, smaller values of λ correspond to faster information mixing. A fully connected network yields $\lambda = 0$, in which case P reduces to the uniform averaging matrix.

Update Order. Following existing decentralized SGD schemes [Lian *et al.*, 2017], the update at each node can be organized according to two distinct patterns. In the Gradient-then-Consensus (GtC) variant, each node first performs a local stochastic gradient step and subsequently mixes the updated parameters with its neighbors:

$$\mathbf{w}^{t+1}(i) = \sum_{l=1}^m P_{il}\left[\mathbf{w}^t(l) - \eta_t \nabla f(\mathbf{w}^t(l); Z_{j_t(l)})\right]. \quad (5)$$

In contrast, the Consensus-then-Gradient (CtG) variant reverses this order. Nodes first aggregate information through the gossip matrix and then apply a stochastic gradient update using locally sampled data:

$$\mathbf{w}^{t+1}(i) = \sum_{l=1}^m P_{il}\mathbf{w}^t(l) - \eta_t \nabla f(\mathbf{w}^t(i); Z_{j_t(i)}). \quad (6)$$

From an optimization viewpoint, GtC and CtG exhibit similar convergence behavior, enabling communication and computation to be overlapped in practice. Their difference becomes pronounced in stability and generalization analysis: the CtG update relies solely on local gradients, which magnifies the effects of data heterogeneity and network disagreement. Consequently, establishing stability guarantees for CtG typically requires more refined analysis or stronger structural assumptions [Sun *et al.*, 2021; Deng *et al.*, 2023; Zeng and Lei, 2025; Le Bars *et al.*, 2023].

Markov Chain Sampling. We now summarize the basic concepts of finite-state, time-homogeneous Markov chains that are relevant to our analysis. These notions formalize the dependence structure induced by Markovian data sampling.

Definition 4. *A stochastic process $\{X_k\}_{k\geq 1}$ taking values in a finite state space $\{1, 2, \ldots, n\}$ is called a time-homogeneous Markov chain with transition matrix $H \in \mathbb{R}^{n\times n}$ if $\Pr(X_{k+1} = j \mid X_0, \ldots, X_k = i) = \Pr(X_{k+1} = j \mid X_k = i) = H_{i,j}$, for all states $i, j \in \{1, 2, \ldots, n\}$ and $k \geq 0$.*

Let π^k denote the distribution of X_k, viewed as a row vector. Then $\pi^k = \pi^{k-1}H = \pi^0 H^k$, where H^k denotes the k-step transition matrix. A Markov chain is said to be irreducible if every state can be reached from any other state in a finite number of steps. A state is aperiodic if it does not exhibit deterministic cyclic behavior, and the chain is aperiodic if all states are aperiodic. Under irreducibility and aperiodicity, the chain admits a unique stationary distribution π^* satisfying $\pi^* = \pi^* H$ and $\min_i \pi_i^* > 0$. Moreover, the transition matrix converges as $\lim_{k\to\infty} H^k = \left[(\pi^*)^\top, \cdots, (\pi^*)^\top\right]^\top =: \Pi^*$.

Mixing Behavior. An important quantity governing the statistical behavior of Markov chain sampling is the mixing time, which measures how quickly the distribution of the chain approaches its stationary distribution. Mixing properties provide explicit bounds on the deviation between π^k and π^*a s a function of k, and play a central role in controlling the bias introduced by Markovian dependence. Following [Sun *et al.*, 2018; Sun *et al.*, 2023], we will characterize this deviation through matrix-based bounds that are well suited for stability analysis.

4 Main Results

This section presents our main theoretical results and the assumptions required for our analysis.

Type	Algorithm	Reference	Case	Generalization Bounds
Centralized	SGD	[Hardt *et al.*, 2016]	Smooth	$\eta T/n$
		[Bassily *et al.*, 2020]	Non-smooth	$\sqrt{T}\eta + \eta T/n$
	Mc-SGD	[Wang *et al.*, 2022]	Smooth	$\eta T/n$
			Non-smooth	$\sqrt{T}\eta + \eta T/n$
Decentralized	D-SGD	[Sun *et al.*, 2021]	Smooth	$\eta T/mn + \eta T/(1-\lambda)$
		[Zeng and Lei, 2025]	Non-smooth	$\sqrt{T}\eta/(1-\lambda^2) + \eta T/mn$
	DMc-SGD	Ours (Theorem 2)	Smooth	$\eta T/mn + \eta T/(1-\lambda)$
			Non-smooth	$\sqrt{T}\eta/(\sqrt{1-\lambda}) + \eta T/mn$

Table 2: Generalization Bounds for Convex Problem. (T: Iterate numbers, η: Stepsize, n: Number of (local) dataset, m: Number of worker, λ: Consensus rate)

Assumption 1 (L-Lipschitz)**.** *For any $Z \sim \mathcal{D}$ and $\mathbf{w}, \tilde{\mathbf{w}} \in \mathcal{W}$, the loss function $f(\mathbf{w}; Z)$ satisifies*

$$|f(\mathbf{w}; Z) - f(\tilde{\mathbf{w}}; Z)| \leq L\|\mathbf{w} - \tilde{\mathbf{w}}\|_2. \tag{7}$$

Remark 2. *This condition ensures that perturbations in the model parameters induce proportionally bounded changes in the loss. In particular, whenever the gradient exists, its norm is uniformly bounded by L.*

Assumption 2 (β-Smoothness)**.** *For any $Z \sim \mathcal{D}$ and $\mathbf{w}, \tilde{\mathbf{w}} \in \mathcal{W}$, the loss function has Lipschitz-continuous gradient:*

$$\|\nabla f(\mathbf{w}; Z) - \nabla f(\tilde{\mathbf{w}}; Z)\|_2 \leq \beta\|\mathbf{w} - \tilde{\mathbf{w}}\|_2. \tag{8}$$

Remark 3. *Smoothness guarantees regular variation of the gradient field and plays a key role in controlling the propagation of perturbations along the optimization trajectory. We next formalize the assumptions on the sampling mechanism employed at each worker.*

Assumption 3. *Each worker accesses data through a finite-state, time-homogeneous Markov chain that is irreducible and aperiodic. All workers are assumed to employ Markov chains with a common transition matrix H and the same sationary distrubtion.*[1]

Remark 4. *Such Markovian sampling schemes have been widely studied in the literature [Sun* et al.*, 2018; Mao* et al.*, 2020; Doan* et al.*, 2020; Sun* et al.*, 2023]. For example, Markov chain–based SGD has been proposed for pairwise learning problems, including AUC maximization, bipartite ranking, and metric learning [Lei* et al.*, 2020; Lei* et al.*, 2021a; Yang* et al.*, 2021]. In these settings, model updates are driven by data pairs generated via an auxiliary Markov chain, which satisfies irreducibility and aperiodicity under mild conditions. Similar assumptions also arise in decentralized consensus optimization over multi-agent networks, where the Markov chain state space coincides with the finite set of agents and a common transition matrix is shared across nodes.*

[1]This assumption is introduced primarily to simplify exposition. The analysis can be extended to heterogeneous Markov chains across workers with additional technical effort.

Assumption 4 (Reversibility)**.** *The Markov transition matrix satisfies $H = H^T$.*

Reversibility enables a clean characterization of mixing behavior and will be used to derive explicit stability bounds.

4.1 Generalization Analysis for DMc-SGD

We adopt an on-average notion of argument stability [Lei and Ying, 2020; Zhu *et al.*, 2022; Wang and Chen, 2024; Zeng and Lei, 2025], which measures the expected deviation between algorithm outputs when a single training example is replaced.

Definition 5. *(On-Average Argument Stability) Let $S = (S_1, \cdots, S_m)$ and $\tilde{S} = (\tilde{S}_1, \cdots, \tilde{S}_m)$ be two independent datasets drawn from $\mathcal{D}$. Let S_{rk} denotes denote the dataset obtained by replacing the k-th sample at the r-th worker in S with the corresponding sample from $\tilde{S}$. A (possibly randomized) algorithm $\mathcal{A}$ is said to be ℓ_1 on-average argument ϵ-stable if*

$$\frac{1}{mn}\sum_{r=1}^{m}\sum_{k=1}^{n}\mathbb{E}\left[\|\mathcal{A}(S) - \mathcal{A}(S_{rk})\|_2\right] \leq \epsilon. \tag{9}$$

Lemma 1. *(Generalization via On-Average Stability. [Lei and Ying, 2020]). Let $\mathcal{A}$ be ℓ_1 on-average argument ϵ-stable and suppose Assumption 1 holds. Then,*

$$\left|\mathbb{E}_{S,\mathcal{A}}\left[R(\mathcal{A}(S)) - R_S(\mathcal{A}(S))\right]\right| \leq L\epsilon.$$

This result reduces the analysis of generalization error to that of stability. We now present explicit stability bounds for DMc-SGD under convexity.

Theorem 1 (Stability Bounds for DMc-SGD)**.** *Assume that $f(\mathbf{w}; Z)$ is convex and L-Lipshitz.*

- *(Smooth Case) If $f(\mathbf{w}; Z)$ is β-smooth and $\eta \leq 2/\beta$, then after T iterations of DMc-SGD,*

$$\epsilon \leq 4\beta L\sum_{t=1}^{T}\eta_t\sum_{q=1}^{t}\eta_q\lambda^{t-q} + \frac{2L}{mn}\sum_{t=1}^{T}\eta_t.$$

- *(Non-smooth Case) Without smoothness, DMc-SGD remains on-average ϵ-stable with*

$$\epsilon \leq 2L\sqrt{\sum_{t=1}^{T}\eta_t^2} + 4L\sqrt{\sum_{t=1}^{T}\eta_t\sum_{q=1}^{t}\eta_q\lambda^{t-q}} + \frac{4L}{mn}\sum_{t=1}^{T}\eta_t.$$

Remark 5. *Compared with stability bounds for centralized SGD with Markovian sampling [Wang* et al.*, 2022], Theorem 1 contains an additional consensus error term that captures the effect of decentralized communication. This term naturally arises in decentralized optimization and is absent in the centralized setting. More importantly, when compared with decentralized SGD under i.i.d. sampling [Sun* et al.*, 2021; Zeng and Lei, 2025], the stability bound obtained here has exactly the* ***same form and order****. In particular, the result is derived without imposing any assumptions on the underlying Markov chain, such as mixing-time or spectral conditions. This shows that Markovian sampling* ***does not lead to worse stability*** *guarantees than i.i.d. sampling for decentralized SGD.*

The key technical ingredient behind this result is the identity $\sum_{k=1}^{n} \mathbb{I}_{\{j_t=k\}} = 1$*, where* $\mathbb{I}_{\{\cdot\}}$ *denotes the indicator function. Since this property is independent of the specific sampling mechanism, the argument is not tied to Markov chain sampling. As a consequence, the same stability analysis* ***applies to other non-i.i.d. sampling schemes****, such as particle filtering or SGLD-based sampling. A detailed comparison with existing stability results is summarized in Table 2.*

The stability bounds in Theorem 2 explicitly depend on the stepsize sequence, revealing how the choice of learning rate mediates the effect of decentralization. In particular, smaller stepsizes mitigate the amplification of instability caused by imperfect consensus across the network. To make this dependence more transparent, we specialize the general bounds to two commonly adopted stepsize schedules.

Corollary 1 (Stability under Smooth Losses)**.** *Suppose* $f(\mathbf{w}; Z)$ *is convex and satisfies both the Lipschitz and smoothness conditions.*

- *(Constant Stepsize) If the stepsize* $\eta_t = \eta \leq 2/\beta$*, the average stability parameter satisfies* $\epsilon \leq \frac{4\eta^2\beta LT}{1-\lambda} + \frac{2\eta LT}{mn}$.
- *(Decreasing Stepsize) If the stepsize* $\eta_t = \frac{1}{t+1} \leq 2/\beta$*, then* $\epsilon \leq \frac{4\beta LC_\lambda T}{T+1} + \frac{2L\ln(T+1)}{mn}$.

The next result addresses the non-smooth setting, where stability exhibits a qualitatively different dependence on the iteration horizon.

Corollary 2 (Stability under Non-smooth Losses)**.** *Suppose* $f(\mathbf{w}; Z)$ *is convex and L-Lipschitz.*

- *(Constant Stepsize) If the stepsize* $\eta_t = \eta \leq 2/\beta$*, the average stability we get* $\epsilon \leq 2L\eta\sqrt{T} + \frac{4\eta L\sqrt{T}}{\sqrt{1-\lambda}} + \frac{4\eta LT}{mn}$.
- *(Decreasing Stepsize) If the stepsize* $\eta_t = \frac{1}{t+1} \leq 2/\beta$*, the average stability we have* $\epsilon \leq 2L + 4L\sqrt{C_\lambda} + \frac{2L\ln T}{mn}$.

The above corollaries characterize the stability of the final iterate produced by DMc-SGD. However, in general convex optimization, performance guarantees are often stated for a weighted average of iterates rather than the last iterate. Following standard practice in decentralized optimization [Sun *et al.*, 2023], we therefore consider the stepsize-weighted average $\bar{\mathbf{w}}^T = \sum_{t=1}^{T} \eta_t \mathbf{w}^t / \sum_{t=1}^{T} \eta_t$. The next theorem establishes generalization guarantees for this averaged solution.

Theorem 2 (Generalization of Averaged Iterates)**.** *Suppose* $f(\mathbf{w}; Z)$ *is convex, satisfying Lipschitz property.*
(Smooth Case) If $f(\mathbf{w}; Z)$ *is* β*-smooth and the stepsize* $\eta_t = \eta \leq 2/\beta$*, the generalization error we get*

$$\left|\mathbb{E}_{S,\mathcal{A}}\left[R(\bar{\mathbf{w}}^T) - R_S(\bar{\mathbf{w}}^T)\right]\right| \leq \frac{2\eta^2\beta LT}{1-\lambda} + \frac{\eta LT}{mn}.$$

(Non-smooth Case) If the stepsize $\eta_t = \eta$*, then*

$$\left|\mathbb{E}_{S,\mathcal{A}}\left[R(\bar{\mathbf{w}}^T) - R_S(\bar{\mathbf{w}}^T)\right]\right| \leq 2L\eta\sqrt{T} + \frac{4\eta L\sqrt{T}}{\sqrt{1-\lambda}} + \frac{4\eta LT}{mn}.$$

We are now ready to combine the optimization and generalization results to obtain explicit excess risk bounds for DMc-SGD. The following results characterize how decentralization and Markovian sampling jointly influence the statistical performance of the averaged iterate.

Theorem 3 (Excess Risk under Smooth Case)**.** *Assume that the loss function* $f(\mathbf{w}; Z)$ *is convex, L-Lipschitz and* β*-Smooth, and that Assumptions 4 holds. Consider DMc-SGD initialized at* $\mathbf{w}^0 = 0$*, and let* $\{w^t\}_{t=1}^{T}$ *denote the iterates generated with the constant stepsize* $\eta_t = \eta \leq 2/\beta$*. If we select* $T = \mathcal{O}(mn)$ *and* $\eta = \frac{1}{\sqrt{T\log T}}$*, then*

$$\mathbb{E}\left[R(\bar{\mathbf{w}}^T) - R(\mathbf{w}^*)\right] = \mathcal{O}\left(\frac{1}{\gamma\log T} + \frac{\sqrt{\log T}}{\sqrt{T}\log(1/\lambda(H))}\right).$$

The next theorem addresses the non-smooth regime, where the interaction between stepsize selection and network effects leads to a different scaling behavior.

Theorem 4 (Excess Risk under Nonsmooth Case)**.** *Suppose that the loss function* $f(\mathbf{w}; Z)$ *is convex, L-Lipschitz , and that Assumptions 4 holds. Let DMc-SGD run for* T *iterations with a constant stepsize* $\eta_t = \eta$*. If we choose* $T = \mathcal{O}\left(\frac{m^2n^2}{1-\lambda}\right)$ *and* $\eta = \sqrt{1-\lambda}T^{-3/4} = \sqrt{\gamma}T^{-3/4}$*, then*

$$\mathbb{E}\left[R(\bar{\mathbf{w}}^T) - R(\mathbf{w}^*)\right] = \mathcal{O}\left(\frac{\log(mn)}{(mn)^{3/2}(\gamma)^{1/4}\log(1/\lambda(H))}\right).$$

Remark 6. *In the smooth case, the excess risk consists of two distinct components: a network-dependent term scaling as* $1/(\gamma\log T)$*, which reflects* ***the effect of imperfect consensus****, and a Markovian sampling term governed by the mixing property of the underlying chain through* $\log(1/\lambda(H))$*. When* $\gamma = 1$*, the bound reduces to the known excess risk rate of centralized Mc-SGD [Wang* et al.*, 2022].*

In the non-smooth regime, the slower rate arises from the combined effect of non-smoothness and decentralized communication, leading to a stronger dependence on both the network ***spectral gap*** γ *and the* ***sample size*** mn*. Nevertheless, the resulting bound remains consistent with existing excess risk guarantees for Mc-SGD. We note, however, that the existing analysis [Wang* et al.*, 2022] reports a* $\mathcal{O}(1/\sqrt{n})$ *rate in the non-smooth case, whereas a careful derivation* ***yields a sharper*** $\mathcal{O}(\log n/n^{3/2})$ *dependence (up to logarithmic factors involving the Markov chain mixing). A detailed discussion of this discrepancy is provided in the appendix F. Overall, these results indicate that Markovian sampling does not introduce an additional statistical penalty beyond what is already induced by decentralization and non-smoothness, and that its effect is explicitly captured through the mixing characteristics of the Markov chain.*

4.2 Generalization Analysis for DMc-SGDA

We now extend the stability-based generalization analysis to decentralized stochastic gradient descent ascent applied to minimax optimization problems. Let $(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T)$ denote the output of DMc-SGDA after T iterations, where the averaged primal and dual variables are defined a

$$\bar{\mathbf{w}}^T = \frac{\sum_{t=1}^T \eta_t \mathbf{w}^t}{\sum_{t=1}^T \eta_t},\ \bar{\mathbf{v}}^T = \frac{\sum_{t=1}^T \eta_t \mathbf{v}^t}{\sum_{t=1}^T \eta_t}.$$

Our goal is to characterize how decentralized communication and Markovian sampling affect the generalization behavior of DMc-SGDA in minimax settings. We first impose some tandard definitions and assumptions [Lei *et al.*, 2021b; Farnia and Ozdaglar, 2021; Ozdaglar *et al.*, 2022; Zhu *et al.*, 2023].

Definition 6. *Let $\rho \geq 0$. A function $f(\mathbf{w}, \mathbf{v}) : \mathcal{W} \times \mathcal{V} \mapsto \mathbb{R}$ is said to be ρ-strongly-convex-strongly-concave (ρ-SC-SC) if, for any $\mathbf{v} \in \mathcal{V}$, the mapping $\mathbf{w} \mapsto f(\mathbf{w}, \mathbf{v})$ is ρ-strongly-convex and, for any $\mathbf{w} \in \mathcal{W}$, the mapping $\mathbf{v} \mapsto f(\mathbf{w}, \mathbf{v})$ is ρ-strongly-concave. The special case $\rho = 0$ corresponds to convex–concave objectives.*

Assumption 5. *For all $\mathbf{w} \in \mathcal{W}, \mathbf{v} \in \mathcal{V}$ and $Z \in \mathcal{Z}$, $\|\nabla_{\mathbf{w}} f(\mathbf{w}, \mathbf{v}; Z)\|_2 \leq L,\ \|\nabla_{\mathbf{v}} f(\mathbf{w}, \mathbf{v}; Z)\|_2 \leq L$.*

Assumption 6. *For any Z, the mapping $(\mathbf{w}, \mathbf{v}) \mapsto f(\mathbf{w}, \mathbf{v}; Z)$ is β-smooth in the sense that, for all $(\mathbf{w}, \mathbf{v})$ and $(\tilde{\mathbf{w}}, \tilde{\mathbf{v}})$,*

$$\left\| \begin{pmatrix} \nabla_{\mathbf{w}} f(\mathbf{w}, \mathbf{v}; Z) - \nabla \mathbf{w} f(\tilde{\mathbf{w}}, \tilde{\mathbf{v}}; Z) \\ \nabla_{\mathbf{v}} f(\mathbf{w}, \mathbf{v}; Z) - \nabla_{\mathbf{v}} f(\tilde{\mathbf{w}}, \tilde{\mathbf{v}}; Z) \end{pmatrix} \right\| \leq \beta \left\| \begin{pmatrix} \mathbf{w} - \tilde{\mathbf{w}} \\ \mathbf{v} - \tilde{\mathbf{v}} \end{pmatrix} \right\|.$$

To quantify the sensitivity of DMc-SGDA, we adopt an on-average notion of argument stability that simultaneously accounts for perturbations in both the primal and dual variables.

Definition 7 (Dencentralized Argument Stability for Minmax Problems)**.** *Let S and $S^{(rk)}$ be constructed as in Definition 5. A randomized algorithm $\mathcal{A}$ is said to be on-average ϵ-argument-stable for minimax problems if $\frac{1}{mn} \sum_{r=1}^m \sum_{k=1}^n \mathbb{E}\Big[\|\mathcal{A}_{\mathrm{w}}(S) - \mathcal{A}_{\mathrm{w}}(S_{rk})\|_2 + \|\mathcal{A}_{\mathrm{v}}(S) - \mathcal{A}_{\mathrm{v}}(S_{rk})\|_2 \Big] \leq \epsilon$.*

We will establish explicit stability bounds for DMc-SGDA applied to convex–concave objectives.

Theorem 5. *(Stability of DMc-SGDA) Consider DMc-SGDA run for T iterations on a convex–concave objective, i.e., for every $Z \in \mathcal{Z}$ the mapping $(\mathbf{w}, \mathbf{v}) \mapsto f(\mathbf{w}, \mathbf{v}; Z)$ is convex in $\mathbf{w}$ and concave in $\mathbf{v}$. Assume $\mathcal{W} = \mathbb{R}^d$ and Assumption 5 holds. Then the algorithm output is on-average ϵ-argument stable, where ϵ can be bounded as follows.*

Smooth regime. *If Assumption 6 holds and the stepsizes satisfy $\sum_{t=1}^T \eta_t \leq 1/(2\beta)$, then*

$$\epsilon \leq 8\sqrt{2}\beta L \sum_{t=1}^T \eta_t \sum_{q=1}^{t-1} \eta_q \lambda^{t-q-1} + \frac{4\sqrt{2}L}{mn} \sum_{t=1}^T \eta_t.$$

Non-smooth regime. *Without Assumption 6, a valid bound is*

$$\epsilon = \mathcal{O}\left(\sqrt{\sum_{t=1}^T \eta_t^2} + \sqrt{\sum_{t=1}^T \eta_t \sum_{q=1}^{t-1} \eta_q \lambda^{t-q-1}} + \frac{1}{mn} \sum_{t=1}^T \eta_t \right).$$

Remark 7. *This theorem exhibits the same stability structure as decentralized SGDA under i.i.d. sampling [Zhu* et al.*, 2023] in the smooth regime, with comparable dependence on the stepsize sequence and the network connectivity. Moreover, we provide stability guarantees for the **non-smooth setting**, which has not been explicitly addressed in prior work. Compared with centralized SGDA or Mc-SGDA [Lei* et al.*, 2021b; Wang* et al.*, 2022], the above bounds contain an additional consensus-related term that arises from decentralized communication.*

The following results characterize the generalization behavior of DMc-SGDA under different performance metrics.

Theorem 6 (Generalization Error of DMc-SGDA)**.** *Consider DMc-SGDA with a constant stepsize η run for T iterations, and define $\mathcal{A}_{\mathbf{w}}(S) = \bar{\mathbf{w}}_T$ and $\mathcal{A}_{\mathbf{v}}(S) = \bar{\mathbf{v}}_T$. Suppose Assumption 5 holds, $\mathcal{W} = \mathbb{R}^d$, and for every Z the function $(\mathbf{w}, \mathbf{v}) \mapsto f(\mathbf{w}, \mathbf{v}; Z)$ is convex–concave. (In addition, suppose the mapping $\mathbf{v} \mapsto R(\mathbf{w}, \mathbf{v})$ is ρ-strongly concave.)*

- *(Smooth case) If Assumption 6 holds and the stepsizes satisfy $\sum_{t=1}^T \eta_t \leq 1/(2\beta)$, then*

$$\epsilon_{\mathrm{gen}}^{\mathbf{w}} \leq \frac{4\sqrt{2}\eta^2 \beta L^2 T}{1-\lambda} + \frac{2\sqrt{2}\eta L^2 T}{mn},$$
$$\epsilon_{\mathrm{gen}}^{\mathrm{P}} \leq 2\sqrt{2}L^2(1+\beta/\rho)\left(\frac{2\eta^2 \beta T}{1-\lambda} + \frac{\eta T}{mn} \right).$$

- *(Non-smooth case) In the absence of smoothness, the same quantity can be bounded as*

$$\epsilon_{\mathrm{gen}}^{\mathbf{w}} \leq 2\sqrt{2}L^2 \eta \sqrt{T} + \frac{4\eta L^2 \sqrt{T}}{\sqrt{1-\lambda}} + \frac{4\sqrt{2}\eta L^2 T}{mn},$$
$$\epsilon_{\mathrm{gen}}^{\mathrm{P}} \leq 2\sqrt{2}L^2(1+\beta/\rho)\left(\eta\sqrt{T} + \frac{\eta\sqrt{2T}}{\sqrt{1-\lambda}} + \frac{2\eta T}{mn} \right).$$

We next characterize the population-level performance of DMc-SGDA in terms of the weak primal–dual (PD) risk. Throughout, let $D_{\mathbf{w}}$ and $D_{\mathbf{v}}$ denote the diameters of the feasible sets $\mathcal{W}$ and $\mathcal{V}$, respectively.

Theorem 7 (Weak PD Population Risk)**.** *Consider DMc-SGDA run with a constant stepsize $\eta_t = \eta$ for T iterations, and let $(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T)$ denote the averaged iterates. Suppose that Assumptions 3-6 hold, and that for every Z the function $(\mathbf{w}, \mathbf{v}) \mapsto f(\mathbf{w}, \mathbf{v}; Z)$ is convex–concave. If $T = mn$ and $\eta = (T \log(T))^{-1/2}$, then*

$$\Delta^{\mathbf{w}}(\overline{\mathbf{w}}_T, \overline{\mathbf{v}}_T) = \mathcal{O}\left(\frac{1}{(1-\lambda)\log T} + \frac{\sqrt{\log T}}{\sqrt{T}\log(1/\lambda(H))} \right).$$

5 Conclusion

We investigated the stability and generalization of decentralized stochastic gradient methods with Markovian sampling. Our analysis shows that, despite temporal dependence and decentralized communication, i.i.d.-type stability and generalization guarantees can still be obtained under appropriate conditions. The results extend naturally to decentralized minimax optimization and highlight the robustness of stability-based analyses beyond independent sampling.

Acknowledgements

This work is supported in part by the National Natural Science Foundation of China (Grant Nos. 62522610, 62376278), and NUDT Foundational Research Funding (JS25-02).

References

[Bars *et al.*, 2024] Batiste Le Bars, Aurélien Bellet, and Marc Tommasi. Improved stability and generalization analysis of the decentralized sgd algorithm. In *International Conference on Machine Learning (ICML)*, 2024.

[Bassily *et al.*, 2020] Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4381–4391, 2020.

[Chen *et al.*, 2014] Jianshu Chen, Zaid J Towfic, and Ali H Sayed. Dictionary learning over distributed models. *IEEE Transactions on Signal Processing*, 63(4):1001–1016, 2014.

[Deng *et al.*, 2023] Xiaoge Deng, Tao Sun, Shengwei Li, and Dongsheng Li. Stability-based generalization analysis of the asynchronous decentralized sgd. In *AAAI Conference on Artificial Intelligence*, pages 7340–7348, 2023.

[Doan *et al.*, 2020] Thinh T Doan, Lam M Nguyen, Nhan H Pham, and Justin Romberg. Convergence rates of accelerated markov gradient descent with applications in reinforcement learning, 2020.

[Doan, 2022] Thinh T Doan. Finite-time analysis of markov gradient descent. *IEEE Transactions on Automatic Control*, 68(4):2140–2153, 2022.

[Duchi *et al.*, 2011] John C Duchi, Alekh Agarwal, and Martin J Wainwright. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic control*, 57(3):592–606, 2011.

[Duchi *et al.*, 2012] John C Duchi, Alekh Agarwal, Mikael Johansson, and Michael I Jordan. Ergodic mirror descent. *SIAM Journal on Optimization*, 22(4):1549–1578, 2012.

[Farnia and Ozdaglar, 2021] Farzan Farnia and Asuman Ozdaglar. Train simultaneously, generalize better: Stability of gradient-based minimax learners. In *International Conference on Machine Learning (ICML)*, pages 3174–3185. PMLR, 2021.

[Goodfellow *et al.*, 2014] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

[Hardt *et al.*, 2016] Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning (ICML)*, pages 1225–1234, 2016.

[Hu *et al.*, 2025] Xiaolin Hu, Zixuan Gong, Gengze Xu, Wei Liu, Jian Luan, Bin Wang, and Yong Liu. Stability and generalization of zeroth-order decentralized stochastic gradient descent with changing topology. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 17342–17350, 2025.

[Koloskova *et al.*, 2020] Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian Stich. A unified theory of decentralized sgd with changing topology and local updates. In *International Conference on Machine Learning (ICML)*, pages 5381–5393. PMLR, 2020.

[Le Bars *et al.*, 2023] Batiste Le Bars, Aurélien Bellet, Marc Tommasi, Erick Lavoie, and Anne-Marie Kermarrec. Refined convergence and topology learning for decentralized sgd with heterogeneous data. In *International Conference on Artificial Intelligence and Statistics*, pages 1672–1702. PMLR, 2023.

[Lei and Ying, 2020] Yunwen Lei and Yiming Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *International Conference on Machine Learning (ICML)*, pages 5809–5819, 2020.

[Lei *et al.*, 2020] Yunwen Lei, Antoine Ledent, and Marius Kloft. Sharper generalization bounds for pairwise learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 21236–21246, 2020.

[Lei *et al.*, 2021a] Yunwen Lei, Mingrui Liu, and Yiming Ying. Generalization guarantee of sgd for pairwise learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 21216–21228, 2021.

[Lei *et al.*, 2021b] Yunwen Lei, Zhenhuan Yang, Tianbao Yang, and Yiming Ying. Stability and generalization of stochastic gradient methods for minimax problems. In *International Conference on Machine Learning (ICML)*, pages 6175–6186. PMLR, 2021.

[Li *et al.*, 2010] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *International Conference on World Wide Web (WWW)*, pages 661–670, 2010.

[Li *et al.*, 2014] Mu Li, Tong Zhang, Yuqiang Chen, and Alexander J Smola. Efficient mini-batch training for stochastic optimization. In *International Conference on Knowledge Discovery and Data Mining*, pages 661–670, 2014.

[Lian *et al.*, 2017] Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[Lian *et al.*, 2018] Xiangru Lian, Wei Zhang, Ce Zhang, and Ji Liu. Asynchronous decentralized parallel stochastic gradient descent. In *International Conference on Machine Learning (ICML)*, pages 3043–3052, 2018.

[Lin *et al.*, 2018] Tao Lin, Sebastian U Stich, Kumar Kshitij Patel, and Martin Jaggi. Don't use large mini-batches, use local sgd. *arXiv preprint arXiv:1808.07217*, 2018.

[Lopes and Sayed, 2007] Cassio G Lopes and Ali H Sayed. Incremental adaptive strategies over distributed networks. *IEEE Transactions on Signal Processing*, 55(8):4064–4077, 2007.

[Madry *et al.*, 2017] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks, 2017.

[Mao *et al.*, 2020] Xianghui Mao, Kun Yuan, Yubin Hu, Yuantao Gu, Ali H Sayed, and Wotao Yin. Walkman: A communication-efficient random-walk algorithm for decentralized optimization. *IEEE Transactions on Signal Processing*, 68:2513–2528, 2020.

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

[Ozdaglar *et al.*, 2022] Asuman Ozdaglar, Sarath Pattathil, Jiawei Zhang, and Kaiqing Zhang. What is a good metric to study generalization of minimax learners? In *Advances in Neural Information Processing Systems (NeuIPS)*, volume 35, pages 38190–38203, 2022.

[Ram *et al.*, 2009] S Sundhar Ram, A Nedić, and Venugopal V Veeravalli. Incremental stochastic subgradient algorithms for convex optimization. *SIAM Journal on Optimization*, 20(2):691–717, 2009.

[Richards and Rebeschini, 2020] Dominic Richards and Patrick Rebeschini. Graph-dependent implicit regularisation for distributed stochastic subgradient descent. *Journal of Machine Learning Research*, 21:1–44, 2020.

[Shah and Avrachenkov, 2018] Suhail M Shah and Konstantin E Avrachenkov. Linearly convergent asynchronous distributed admm via markov sampling, 2018.

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

[Sun *et al.*, 2018] Tao Sun, Yuejiao Sun, and Wotao Yin. On markov chain gradient descent. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[Sun *et al.*, 2021] Tao Sun, Dongsheng Li, and Bao Wang. Stability and generalization of decentralized stochastic gradient descent. In *AAAI Conference on Artificial Intelligence*, pages 9756–9764, 2021.

[Sun *et al.*, 2022] Tao Sun, Dongsheng Li, and Bao Wang. Adaptive random walk gradient descent for decentralized optimization. In *International Conference on Machine Learning*, pages 20790–20809. PMLR, 2022.

[Sun *et al.*, 2023] Tao Sun, Dongsheng Li, and Bao Wang. On the decentralized stochastic gradient descent with markov chain sampling. *IEEE Transactions on Signal Processing*, pages 1–14, 2023.

[Sundhar Ram *et al.*, 2010] S Sundhar Ram, Angelia Nedić, and Venugopal V Veeravalli. Distributed stochastic subgradient projection algorithms for convex optimization. *Journal of Optimization Theory and Applications*, 147:516–545, 2010.

[Tadić and Doucet, 2011] Vladislav B Tadić and Arnaud Doucet. Asymptotic bias of stochastic gradient search. In *IEEE Conference on Decision and Control and European Control Conference*, pages 722–727. IEEE, 2011.

[Wai *et al.*, 2018] Hoi-To Wai, Zhuoran Yang, Zhaoran Wang, and Mingyi Hong. Multi-agent reinforcement learning via double averaging primal-dual optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

[Wang and Chen, 2024] Jiahuan Wang and Hong Chen. Towards stability and generalization bounds in decentralized minibatch stochastic gradient descent. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 15511–15519, 2024.

[Wang *et al.*, 2022] Puyu Wang, Yuewen Lei, Yiming Ying, and Ding-Xuan Zhou. Stability and generalization for markov chain stochastic gradient methods. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 37735–37748, 2022.

[Yang *et al.*, 2021] Zhenhuan Yang, Yunwen Lei, Puyu Wang, Tianbao Yang, and Yiming Ying. Simple stochastic and online gradient descent algorithms for pairwise learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 20160–20171, 2021.

[Ye *et al.*, 2025] Haoxiang Ye, Tao Sun, and Qing Ling. Generalization error analysis for attack-free and byzantine-resilient decentralized learning with data heterogeneity, 2025.

[Yuan *et al.*, 2016] Kun Yuan, Qing Ling, and Wotao Yin. On the convergence of decentralized gradient descent. *SIAM Journal on Optimization*, 26(3):1835–1854, 2016.

[Yuan *et al.*, 2023] Kun Yuan, Sulaiman A Alghunaim, and Xinmeng Huang. Removing data heterogeneity influence enhances network topology dependence of decentralized sgd. *Journal of Machine Learning Research*, 24(280):1–53, 2023.

[Zeng and Lei, 2025] Shuang Zeng and Yunwen Lei. Stability and generalization analysis of decentralized sgd: sharper bounds beyond lipschitzness and smoothness. In *International Conference on Machine Learning (ICML)*. PMLR, 2025.

[Zhu *et al.*, 2022] Tongtian Zhu, Fengxiang He, Lan Zhang, Zhengyang Niu, Mingli Song, and Dacheng Tao. Topology-aware generalization of decentralized sgd. In *International Conference on Machine Learning (ICML)*, pages 27479–27503, 2022.

[Zhu *et al.*, 2023] Miaoxi Zhu, Li Shen, Bo Du, and Dacheng Tao. Stability and generalization of the decentralized stochastic gradient descent ascent algorithm. In *Advances in Neural Information Processing Systems (NeuIPS)*, volume 36, 2023.