# Towards Stability and Generalization Bounds in Decentralized Minibatch Stochastic Gradient Descent

**Jiahuan Wang**[1], **Hong Chen**[1, 2, 3*]

[1]College of Informatics, Huazhong Agricultural University, Wuhan, China
[2]Engineering Research Center of Intelligent Technology for Agriculture, Ministry of Education, Wuhan, China
[3]Hubei Engineering Technology Research Center of Agricultural Big Data, Wuhan, China
chenh@mail.hzau.edu.cn

## Abstract

Decentralized Stochastic Gradient Descent (D-SGD) represents an efficient communication approach tailored for mastering insights from vast, distributed datasets. Inspired by parallel optimization paradigms, the incorporation of minibatch serves to diminish variance, consequently expediting the optimization process. Nevertheless, as per our current understanding, the existing literature has not thoroughly explored the learning theory foundation of Decentralized Minibatch Stochastic Gradient Descent (DM-SGD). In this paper, we try to address this theoretical gap by investigating the generalization properties of DM-SGD. We establish the sharper generalization bounds for the DM-SGD algorithm with replacement (without replacement) on (non)convex and (non)smooth cases. Moreover, our results consistently recover to the results of Centralized Stochastic Gradient Descent (C-SGD). In addition, we derive generalization analysis for Zero-Order (ZO) version of DM-SGD.

## Introduction

Decentralized Stochastic Gradient Descent (D-SGD) is a distributed optimization algorithm used to train machine learning models across multiple devices or nodes while minimizing communication overhead (Nedic and Ozdaglar 2009; Sundhar Ram, Nedić, and Veeravalli 2010; Lian et al. 2017). It is particularly useful when dealing with large datasets or complex models that cannot be trained on a single machine due to memory or computational limitations.

In traditional (centralized) SGD, each iteration of training involves computing gradients using a random subset (minibatch) of the training data and updating the model parameters based on these gradients. Decentralized SGD takes this concept and extends it to a distributed setting (Predd, Kulkarni, and Poor 2006; Agarwal and Duchi 2011), where each node has access to its local subset of data and computes gradients independently.

As far as we know, utilizing a minibatch aids in decreasing variance and expediting the optimization process. Therefore, it is natural to consider the Decentralized Minibatch Stochastic Gradient Descent (DM-SGD) which combines the principles of decentralized optimization and Minibatch stochastic gradient descent (Gower et al. 2019; Cotter et al. 2011; Dekel et al. 2012; Shamir and Srebro 2014; Li et al. 2014; Ghadimi, Lan, and Zhang 2016; Yin et al. 2018) to enable efficient distributed training of machine learning models (Lian et al. 2017; Zinkevich et al. 2010).

While the studies on the convergence analysis of CM-SGD/DM-SGD are increasing (Cotter et al. 2011; Dekel et al. 2012; Shamir and Srebro 2014; Woodworth, Patel, and Srebro 2020; Lian et al. 2017), there are far fewer results to investigate the generalization ability of DM-SGD in learning theory (also see Table 1).

Within the realm of learning theory analysis, algorithmic stability tools stand out as a crucial factor. Notably, they bring forth advantages such as independence from dimensionality and adaptability to a wide array of learning paradigms (Bousquet and Elisseeff 2002; Shalev-Shwartz et al. 2010; Hardt, Recht, and Singer 2016; Feldman and Vondrak 2018, 2019). Specially, the stability and generalization of D-SGD has recently studied by (Sun, Li, and Wang 2021; Zhu et al. 2022; Taheri and Thrampoulidis 2023; Bars, Bellet, and Tommasi 2023). However, they did not take into account the stability analysis work and outcomes of DM-SGD. Thus, the following questions are raised:

**Questions**

*What are the results of the stability and generalization analysis of DM-SGD? Is the generalization bound of DM-SGD consistent with D-SGD?*

This paper focuses on answering the above questions, and gets the satisfactory generalization bounds compared with D-SGD (Sun, Li, and Wang 2021; Bars, Bellet, and Tommasi 2023).

### Contributions

As we all know, this paper is the first work to study the generalization behavior of DM-SGD, where both the generalization error and optimization error are considered. We state comprehensive theoretical results of DM-SGD under the convex, strongly convex and non-convex settings. We elaborate on our contributions as below.

- **Stability and generalization bounds of DM-SGD.** We conduct a comprehensive analysis of the algorithmic sta-

*Corresponding author.

bility inherent in DM-SGD and proceed to deduce its associated generalization error bounds accordingly. Our stability bounds are constructed to account for the phenomenon of low training errors, a common characteristic found in over-parameterized models. We also derive the excess generalization error under $\ell_2$ on-average stability, which extends the related work for C-SGD (Lei and Ying 2020) to the DM-SGD setting. As far as our current knowledge goes, this marks the initial stability analysis for DM-SGD.

- **Stability bounds of Zeroth-Order DM-SGD (ZO DM-SGD).** Under the assumption of Lipschitz and smooth loss functions, we further establish the stability bounds for DM-SGD in scenarios involving the intricate gradient information.

## Related Work

In this section, we review the related works: Decentralized and distributed learning, and Stability and Generalization.

### Decentralized and Distributed Learning

Decentralized algorithms are designed to operate in distributed systems without a central coordinating entity (Tsitsiklis 1984; Tsitsiklis, Bertsekas, and Athans 1986; Nedic and Ozdaglar 2009).

Emerging as a quintessential decentralized optimization technique, D-SGD has found its way into various dimensions of deep learning, undergoing extensions that encompass an array of contexts, such as: minibatch setting (Lian et al. 2017), local D-SGD (Li et al. 2019; Nadiradze et al. 2020), gradient tracking (Zhang and You 2019; Xin, Khan, and Kar 2021), asynchronous settings (Sirb and Ye 2016; Lian et al. 2018; Xu, Zhang, and Wang 2021; Nadiradze et al. 2021) , data-heterogeneous (Tang et al. 2018; Vogels et al. 2021), and markov chain sampling scenarios (Sun, Li, and Wang 2023).

### Stability and Generalization

In the realm of statistical learning theory (SLT), algorithmic stability stands as a cornerstone (Bousquet and Elisseeff 2002; Elisseeff et al. 2005; Shalev-Shwartz et al. 2010). It serves as a vital gauge for measuring the sensitivity of an algorithm to fluctuations caused by perturbations in the training data (Hardt, Recht, and Singer 2016; Bousquet, Klochkov, and Zhivotovskiy 2020).

Algorithmic stability has demonstrated remarkable efficacy in establishing dimension-independent generalization bounds for a broad range of learning frameworks. A foundational framework for analyzing stability was formulated by Bousquet and Elisseeff (2002), introducing the concepts of uniform stability and hypothesis stability. This framework paved the way for subsequent advancements, including the extension of uniform stability measurements to analyze stochastic algorithms (Elisseeff et al. 2005; Hardt, Recht, and Singer 2016). It also served as inspiration for various stability concepts such as uniform argument stability (Liu et al. 2017), locally elastic stability (Deng, He, and Su 2021), on-average loss stability (Lei, Ledent, and Kloft 2020; Lei, Liu, and Ying 2021), and on-average argument stability (Shalev-Shwartz et al. 2010; Lei and Ying 2020; Lei, Liu, and Ying 2021).

While the above studies do not consider the decentralized and distributed settings, Sun, Li, and Wang (2021) proposed stability and generalization bounds for D-SGD (AWC version (5)) based on uniform stability. Zhu et al. (2022) also studied the impact of communication topology on AWC version with the on-average stability tool under the Hölder smooth (Lei and Ying 2020) and convex condition. On the other hand, Richards et al. (2020) introduces a generalization bound of D-SGD (CAU version (4)) using the concepts of algorithmic stability and Rademacher complexity in both smooth and nonsmooth scenarios. Bars, Bellet, and Tommasi (2023) considers that the CAU version establishes the same generalization bounds as the classical C-SGD (Hardt, Recht, and Singer 2016).

It is worth noting that the difficulty in solving the stability analysis of the decentralization problem lies in how to deal with the mixing matrix. Sun, Li, and Wang (2021) used the convergence relationship between the mixing matrix and the identity matrix to solve the difficulty, Zhu et al. (2022) trained the ResNet-18 model in the MINIST dataset, and then obtained the assumption that the weight update difference satisfies the Gaussian distribution, and finally obtained the result of $\ell_2$ stability bounds. Recently, inspired by the work of Bars, Bellet, and Tommasi (2023), due to the different update methods of D-SGD (CAU version), the problem is relatively simple and there are more angles for analysis. To better understand the difference of the above related work, we summarize the main results in Table 2. In the final stage of the current paper, we noticed relevant work (Lei, Sun, and Liu 2023) on the generalization analysis of M-SGD and Local-SGD, achieving linear acceleration to both bath size and machine quantity.

## Preliminaries

This section introduces the problem formulation of DM-SGD and the definitions of algorithmic stability.

### Distributed Learning

Consider a distributed system with $m$ computing workers to train a ML model under data parallelism. Denote the training dataset as $S := \{S_1, \cdots, S_m\}$, where $S_j = \{Z_{j1}, \cdots, Z_{jn}\}$ is the agent $j$'s training data that each sample is independently drawn from a probability measure $\mathcal{D}$ defined over a sample space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Suppose that $\mathcal{X} \subset \mathbb{R}^d$ represents an input space of dimension $d$ and $\mathcal{Y} \subset \mathbb{R}$ is an output space.

Consider $\mathcal{W}$ as the designated parameter space for learning models. The objective of distributed learning is to seek a model parameterized by $\mathbf{w}$ in such a way that the population risk (or expected risk), denoted as

$$R(\mathbf{w}) = \mathbb{E}_{z\sim\mathcal{D}}[\ell(\mathbf{w}; z)], \tag{1}$$

is minimized to the greatest extent possible. Here, $\ell : \mathcal{W} \times \mathcal{Z} \to [0, \infty)$ represents a loss function.

Due to the unknown intrinsic distribution $\mathcal{D}$ , we can't get the minimizer of $R(\mathbf{w})$ directly. As a practical substitute for

| Type | Reference | Algorithm | CA | GA |
|---|---|---|---|---|
| Centralized | Cotter et al. (2011) | M-SGD | ✓ | × |
| | Dekel et al. (2012) | M-SGD | ✓ | × |
| | Shamir and Srebro (2014) | M-SGD | ✓ | × |
| | Woodworth, Patel, and Srebro (2020) | M-SGD/Local-SGD | ✓ | × |
| Decentralized | Lian et al. (2017) | SGD/M-SGD | ✓ | × |

Table 1: Summary of anlysis for Minibacth SGD and DM-SGD.(CA:Convergence Analysis; GA: Generalization Analysis; M-SGD: Minibatch SGD; "✓": YES; "×": NO)

| Type | Reference | Analysis Tool | Algorithm | L | S | C | SC | NC |
|---|---|---|---|---|---|---|---|---|
| AWC (5) | Sun, Li, and Wang (2021) | Consensus Distance | D-SGD | ✓ | ✓ | ✓ | ✓ | ✓ |
| | Zhu et al. (2022) | Gaussian Weight | D-SGD | × | ✓[H] | ✓ | × | × |
| | Deng et al. (2023) | Consensus Distance | AD-SGD | ✓ | ✓ | ✓ | ✓ | ✓ |
| | Taheri and Thrampoulidis (2023) | Consensus Distance | D-GD | × | ✓ | ✓ | × | ✓ |
| CAU (4) | Bars, Bellet, and Tommasi (2023) | Stripped $W_{ij}$ | D-SGD | ✓ | ✓ | ✓ | ✓ | ✓ |
| | Ours | Stripped $W_{ij}$ | DM-SGD | ✓ | ✓/× | ✓ | ✓ | ✓ |

Table 2: Summary of related work.("L" and "S" denote Lipschitz and smoothness properties respectively. "H" denotes to the Hölder continuous smooth. "C," "SC," and "NC" correspond to convex, strongly convex, and nonconvex, respectively.)

algorithmic development, we frequently examine the associated empirical risk, which is defined as

$$R_S(\mathbf{w}) := \frac{1}{m}\sum_{j=1}^{m} R_{S_j}(\mathbf{w}) = \frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}\ell(\mathbf{w}; z_{ji}) \quad (2)$$

For conciseness, let $\mathcal{A}(S)$ represent the model derived from applying algorithm $\mathcal{A}$ (e.g., SGD or D-SGD) to dataset $S$. While $\mathcal{A}(S)$ may show a small empirical risk during training by fitting examples perfectly, its empirical effectiveness does not assure a similarly small population risk. So it is natural to study the difference between the population risk and empirical risk

$$R(\mathcal{A}(S)) - R_S(\mathcal{A}(S)). \quad (3)$$

We are also interested in studying the excess generalization error $R(\mathcal{A}(S)) - R(\mathbf{w}^\star)$, where $\mathbf{w}^\star \in \arg\min R(\mathbf{w})$ is the theoretically optimal solution. It can be decomposed as

$$\begin{aligned}&\mathbb{E}\left[R(\mathcal{A}(S)) - R(\mathbf{w}^\star)\right]\\ =&\underbrace{\mathbb{E}\left[R(A(S)) - R_S(A(S))\right]}_{Generalization\quad Error} + \underbrace{\mathbb{E}\left[R_S(A(S)) - R_S(\mathbf{w}^\star)\right]}_{Optimization\quad Error}.\end{aligned}$$

DM-SGD was proposed in Lian et al. (2017), which use $m$ machines to optimize problem (2) in batches according to a decentralized communication graph $\mathcal{G} = (\mathcal{V}, W)$. The vertex set $\mathcal{V} = \{1, \cdots, m\}$ denotes the set of $m$ workers and $W = [w_{ij}] \in \mathcal{R}^{m\times m}$ represents the communication link between each nodes. It is important to note that matrix $W$ is a doubly stochastic matrix (Sinkhorn 1964; Sinkhorn and Knopp 1967), and for a given graph, the mixing matrix lacks uniqueness. We will also introduce the four conditions of mixing matrix below.

**Definition 1.** *(Mixing matrix/Gossip matrix) For any given graph $\mathcal{G} = (\mathcal{V}, W)$, the mixing matrix $W$ is defined on the edge set $\mathcal{E}$ that satisfies: (1) If $i \neq j$ and $(i,j) \notin \mathcal{E}$, then $W_{ij} = 0$ (disconnected); otherwise, $W_{ij} > 0$ (connected); (2) $W = W^T$ (symmetry); (3) $W_{ij} \in [0,1] \quad \forall i,j \in m$; (4) $\mathbf{1}_m^T W = W\mathbf{1}_m$ (Standard additivity);*

Let $\lambda_i$ represent the $i$-th largest eigenvalue of $\mathcal{W}$, and introduce a significant constant $\lambda := \max\{|\lambda_2|, |\lambda_m(W)|\}$. The significance of the mixing matrix's definition lies in the fact that it results in $0 \leq \lambda < 1$. In the context of a connected graph, if $\lambda = 0$, this indicates a fully-connected communication topology, where all elements of $W$ are $\frac{1}{m}$.

## Decentralized Minibatch SGD

In this section, we first introduce the Decetralized Stochastic Gradient Descent (D-SGD) algorithm (Lian et al. 2017). It is broken down into the following steps:

1. **Initialization**: Each node initializes its model parameters randomly or based on some initial configuration.
2. **Gradient Computation**: Nodes compute gradients using their local data (mini-batches) and the current model parameters. These gradients represent the direction in which the model should be adjusted to minimize the loss function.

3. **Communication and Aggregation**: Nodes communicate their gradients directly with each other.
4. **Parameter Update**: After aggregation, each node updates its local model parameters using the aggregated gradient information. This step is similar to the update step in traditional SGD.
5. **Output**: The average of the last updated parameters of each machine.

**Remark 1.** *In Lian et al. (2017) mentions that, the Step 3 and Step 4 can be exchanged. The above updated strategy is shown to compute local stochastic gradients and machine communication in parallel (communication time can be completely hidden if communication time is less than computation time). The replacement version can ensure the interaction of more gradient information. We will show the specific two update formulas as follows.*

- *Type I (Communicate After Update [CAU])*

$$\mathbf{w}_j^{t+1} = \sum_{k=1}^{m} W_{jk} \left( \mathbf{w}_k^t - \eta_t \nabla \ell(\mathbf{w}_k^t; Z_{k,i_{t,k}}) \right) \quad (4)$$

- *Type II (Adapt while Communicate [AWC])*

$$\mathbf{w}_j^{t+1} = \sum_{k=1}^{m} W_{jk} \mathbf{w}_k^t - \eta_t \nabla \ell(\mathbf{w}_j^t; Z_{j,i_{t,j}}) \quad (5)$$

Inspiring work by Lian et al. (2017) points out that the computed stochastic gradients can be replaced by mini-batch stochastic gradients without compromising their theoretical results.

At the $t$-th iteration, minibatch SGD randomly draws (with replacement) $b$ numbers $i_{t,1}, \ldots, i_{t,b}$ independently from the uniform distribution over $[n]$, where $b \in [n]$ is the batch size. Then it updates $\{\mathbf{w}_t\}$ by

$$\mathbf{w}_j^{t+\frac{1}{2}} = \mathbf{w}_j^t - \frac{\eta_t}{b} \sum_{r=1}^{b} \nabla \ell(\mathbf{w}_j^t; Z_{j,i_{t,r}}), \quad (6)$$

where $\{\eta_t\}$ is a positive step size sequence. If $b = 1$, equation (6) returns to the SGD. However, when $b = n$, the aforementioned approach remains distinct from Gradient Descent due to the incorporation of selection with replacement. We will show the DM-SGD with replacement (WR) algorithm flow chart in Algorithm 1.

Algorithm 1: Decentralized Minibatch Stochastic Gradient Descent (DM-SGD)

**Require:** Initialize $\mathbf{w}_j^1 = \mathbf{w}^1$, stepsizes $\{\eta_t\}_{t=1}^T$,weight matrix $W$, batch size $b$ and iterations $T$.

1: **for** $t = 1, 2, \cdots, T$ **do**
2:  **for** $j = 1, 2, \cdots, m$ **do**
3:  draw(with replacement) b numbers $i_{t,1}, \cdots, i_{t,b}$ uniformly over local data of the $j$-th worker
4:  $\mathbf{w}_j^{t+\frac{1}{2}} = \mathbf{w}_j^t - \frac{\eta_t}{b} \sum_{r=1}^{b} \nabla \ell(\mathbf{w}_j^t; Z_{j,i_{t,r}})$
5:  $\mathbf{w}_j^{t+1} = \sum_{k=1}^{m} W_{jk} \mathbf{w}_j^{t+\frac{1}{2}}$
6:  **end for**
7: **end for**

**Output:** $\sum_{j=1}^{m} \mathbf{w}_j^{T+1}$

## Algorithmic Stability

Algorithmic stability is a pivotal concept in statistical learning, quantifying how algorithms respond to changes in training data. This paper focuses on analyzing on-average argument stability techniques (Shalev-Shwartz et al. 2010; Kuzborskij and Lampert 2018; Lei and Ying 2020; Lei, Liu, and Ying 2021).

Unlike uniform stability (Bousquet and Elisseeff 2002; Hardt, Recht, and Singer 2016; Bousquet, Klochkov, and Zhivotovskiy 2020), which hinges on loss function shifts, on-average argument evaluates stability by observing model $\mathcal{A}(S)$ changes.

**Definition 2.** *(On-average argument/model stability) Let $S = (S_1, \cdots, S_m)$ with $S_j = \{Z_{j1}, \cdots, Z_{jn}\}$ and $\tilde{S} = (\tilde{S}_1, \cdots, \tilde{S}_m)$ with $\tilde{S}_j = \{\tilde{Z}_{j1}, \cdots, \tilde{Z}_{jn}\}$ be two independent copies drawn from the same distribution $\mathcal{D}$. Assume that $S_{ki}$ is the $i$-th sample in the $k$-th worker's training set becomes $\tilde{Z}_{ki}$ for any $i \in [n], k \in [m]$. Let $\mathbf{w}_k$ and $\tilde{\mathbf{w}}_k$ represent the weights assigned to the $k$-th worker through the stochastic algorithm $\mathcal{A}$ based on $S$ and $S_{ki}$ respectively.*

*$\mathcal{A}$ is $\ell_1$ on-average argument $\epsilon$-stability for all training sets $S$ and $S_{ki}$ if*

$$\frac{1}{mn} \sum_{k=1}^{m} \sum_{i=1}^{n} \mathbb{E} \left[ \|\mathbf{w}_k - \tilde{\mathbf{w}}_k\|_2 \right] \leq \epsilon, \quad (7)$$

*and $\ell_2$ on-average argument $\epsilon$-stability if*

$$\frac{1}{mn} \sum_{k=1}^{m} \sum_{i=1}^{n} \mathbb{E} \left[ \|\mathbf{w}_k - \tilde{\mathbf{w}}_k\|_2^2 \right] \leq \epsilon^2. \quad (8)$$

**Lemma 1.** *(Generalization error via on-average stability. (Lei and Ying 2020) [Theorem 2]). Let $S$ and $S_{ki}$ be constructed as the definition of on-average argument stability. Let $\gamma > 0$.*

- *Let $\mathcal{A}$ be $\ell_1$ on-average argument $\epsilon$-stable and Assumption 1 hold. Then*

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

- *Let $\mathcal{A}$ be $\ell_2$ on-average argument $\epsilon$-stable and suppose that $\ell(\mathbf{w}; z)$ is non-negative and $\beta$-smooth. Then*

$$\mathbb{E}_{S,\mathcal{A}} \left[ R(\mathcal{A}(S)) - R_S(\mathcal{A}(S)) \right] \leq \frac{\beta}{\gamma} \mathbb{E}_{S,\mathcal{A}} \left[ R_S(\mathcal{A}(S)) \right] + \frac{\beta + \gamma}{2mn} \sum_{k=1}^{m} \sum_{i=1}^{n} \mathbb{E}_{S,\tilde{S},\mathcal{A}} \left[ \|\mathcal{A}(S_{ki}) - \mathcal{A}(S)\|_2^2 \right]. \quad (9)$$

# Main Results

This section presents our main results on the generalization bounds of DM-SGD algorithms based on on-average stability. In order to constrain the gradient update process of DM-SGD, it becomes essential to make some assumptions about the characteristics of the loss functions.

**Definition 3.** *$\ell$ is $\mu$-strongly convex if for any $z$ and $u, v \in \mathcal{W}$,*

$$\ell(u; z) \geq \ell(v; z) + \langle \nabla \ell(v; z), u - v \rangle + \frac{\mu}{2} \|u - v\|_2^2.$$

*Specially, $\ell(\cdot; z)$ is convex if $\mu = 0$.*

Indeed, it's evident that a strongly convex loss function must possess convexity, but the reverse might not hold true. And convexity plays a pivotal role in various optimization analyses of learning algorithms (Hardt, Recht, and Singer 2016; Harvey et al. 2019).

**Assumption 1.** *For any $z$ and $u, v \in \mathcal{W}$, $\ell$ is $L$-Lipschitz if*

$$\|\ell(u; z) - \ell(v; z)\|_2 \leq L\|u - v\|_2.$$

The mentioned inequality is synonymous with the constraint on the gradient of $\ell$, denoted as $|\nabla\ell(\cdot; z)|_2 \leq L$.

**Assumption 2.** *For any $z$ and $u, v \in \mathcal{W}$, $\ell$ is $\beta$-smooth if*

$$\|\nabla\ell(u; z) - \nabla\ell(v; z)\|_2 \leq \beta\|u - v\|_2.$$

**Remark 2.** *Assumptions 1 and 2 are common stability analysis assumptions, where Assumption 1 can be relaxed to avoid (Lei and Ying 2020; Nikolakakis et al. 2022b), and more methods are to use the self-bounding property of the "non-negative+smooth" function to scale to the function value itself (Srebro, Sridharan, and Tewari 2010). Following the steps in (Hardt, Recht, and Singer 2016; Lei, Ledent, and Kloft 2020), we can verify that the gradient update is non-expansive when $\ell$ is convex and $\beta$-smooth.*

Now we are ready to present the generalization bounds of DM-SGD. Due to the space limitations, please refer to the *supplementary material B* for detailed theoretical proofs of the convex case.

## Convex Case

**Theorem 1.** *(Stability bounds) Suppose for any $z \in \mathcal{Z}$, $\ell(\mathbf{w}; z)$ is non-negative, convex and $\beta$-smooth with respect to (w.r.t) $\mathbf{w} \in \mathcal{W}$. Without loss of generality, let $S$ and $S_{ki}$ be constructed in Definition 2. Let $\mathbf{w}_k^{T+1}$ and $\tilde{\mathbf{w}}_k^{T+1}$ be the $T$-th iteration on the $k$-th worker produced by DM-SGD (WR) based $S$ and $S_{ki}$, respectively. If $\eta_t \leq 2/\beta$ and $p > 0$, then with $T$ iterations we have*

$$\frac{1}{mn}\sum_{k=1}^{m}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\mathbf{w}_k^{T+1} - \tilde{\mathbf{w}}_k^{T+1}\right\|_2\right] \leq \frac{2\sqrt{2\beta}}{mn}\sum_{t=1}^{T}\eta_t\sqrt{\mathbb{E}\left[R_S(\mathbf{w}^t)\right]}$$

*and*

$$\frac{1}{mn}\sum_{k=1}^{m}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\mathbf{w}_k^{T+1} - \tilde{\mathbf{w}}_k^{T+1}\right\|_2^2\right] \leq (1+\frac{1}{p})\frac{8\beta\lambda^2(n+b-1)}{n^2b}\sum_{t=1}^{T}\eta_t^2(1+p)^{T-t}\mathbb{E}\left[R_S(\mathbf{w}^t)\right].$$

**Remark 3.** *The key to solving the decentralized SGD problem is how to deal with the mixing matrix, and then transform it into a common update iterative analysis. Motivate by Bars, Bellet, and Tommasi (2023), we diversify its decomposition technique to isolate the mixing matrix and further transform it into the regular Minibacth SGD case.*

*The conventional SGD stability analysis process (Hardt, Recht, and Singer 2016; Lei and Ying 2020; Bassily et al. 2020) considers whether to select abnormal samples or not, and then considers the results according to the situation. This argument does not apply to mini-batch SGD with replacement, since we can draw a specific example multiple times. We solve this difficulty by introducing the concept of binomial distribution, turning the original $b$ samples of a batch into counts of all sample traversals (Lei, Sun, and Liu 2023). Our $\ell_1$ results yield a similar generalization bound for DM-SGD as the one derived by Lei and Ying (2020) for C-SGD. We also elaborate the various stability bounds under convex case in Table 3 for comparison.*

**Remark 4.** *Compared with Bars, Bellet, and Tommasi (2023), we relax the Lipschitz assumptions instead of the non-negative condition. We obtain that this stability bound includes the empirical risk of $\mathbf{w}^t$, and then the empirical risk (training error) can be minimized by the optimization algorithm. It is inherent that $\mathbb{E}\left[R_S(\mathbf{w}^t)\right]$ can be guaranteed to be much smaller than the Lipschitz constant (Assumption 1 holds) under certain instances (Lei and Ying 2020).*

We also derive generalization results for DM-SGD without replacement (WOR). Unlike replacement case, the non-replacement framework ensures that anomalous samples are selected only once rather than multiple times.

**Proposition 1.** *Suppose $\ell(\mathbf{w}; z)$ is convex and Assumption 1 holds.*

- *(Smooth case) If $\ell(\mathbf{w}; z)$ is $\beta$-smooth and $\eta_t \leq 2/\beta$, then, for DM-SGD (WOR) with $T$ iterations, we have*

$$\frac{1}{mn}\sum_{k=1}^{m}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\mathbf{w}_k^{T+1} - \tilde{\mathbf{w}}_k^{T+1}\right\|_2\right] \leq \frac{2L}{mn}\sum_{t=1}^{T}\eta_t.$$

- *(Non-smooth case) $\mathcal{A}$ is on-average $\epsilon$-argument stable with*

$$\frac{1}{mn}\sum_{k=1}^{m}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\mathbf{w}_k^{T+1} - \tilde{\mathbf{w}}_k^{T+1}\right\|_2\right] \leq \frac{2L}{m}\sqrt{\sum_{t=1}^{T}\eta_t^2} + \frac{4L}{mn}\sum_{t=1}^{T-1}\eta_{t+1}.$$

**Remark 5.** *Extending the analysis framework of the work (Wang et al. 2022), we divide the analysis process into two cases $i \in \{i_{t,r}, r = 1, \cdots, b\}$ and $i \notin \{i_{t,r}, r = 1, \cdots, b\}$, and then get a similar boundary.*

**Remark 6.** *We briefly describe the development of stability analysis for nonsmooth problems in SGD. Lei and Ying (2020) put forward the assumption of Hölder continuity and obtained the first non-smooth stability analysis. Then Bassily et al. (2020) proposed sharper upper and lower bounds for SGD under Lipschitz and nonsmooth convex case. Recently, based on Bassily's analysis framework, Wang et al. (2022) added the recursive sequence lemma (Schmidt, Roux, and Bach 2011) to get better results. Their works paved the way to achieve more intricate and refined stability analysis outcomes.*

| Type | Reference | Algorithmic Stability | Bounds |
|---|---|---|---|
| C-SGD | Hardt, Recht, and Singer (2016) | Uniform Stability | $\mathcal{O}\left(\sum_{t=1}^{T}\eta_t/n\right)$ |
| | Lei and Ying (2020) | $\ell_1$ On-average Stability | $\mathcal{O}\left(\frac{1}{n}\sum_{t=1}^{T}\eta_t\mathbb{E}\left[\sqrt{R_S(\mathbf{w}^t)}\right]\right)$ |
| | | $\ell_2$ On-average Stability | $\mathcal{O}\left(\frac{1}{n}\sum_{t=1}^{T}\eta_t^2\mathbb{E}\left[R_S(\mathbf{w}^t)\right]\right)$ |
| DM-SGD (WR) | Ours (Theorem 1) | $\ell_1$ On-average Stability | $\mathcal{O}\left(\frac{1}{mn}\sum_{t=1}^{T}\eta_t\mathbb{E}\left[\sqrt{R_S(\mathbf{w}^t)}\right]\right)$ |
| | | $\ell_2$ On-average Stability | $\mathcal{O}\left(\frac{1}{nb}\sum_{t=1}^{T}\eta_t^2\mathbb{E}\left[R_S(\mathbf{w}^t)\right]\right)$ |
| D-SGD | Sun, Li, and Wang (2021) | Uniform Stability | $\mathcal{O}\left(\frac{L^2\sum_{t=1}^{T}\eta_t}{mn}+\frac{\eta T}{1-\lambda}\right)$ |
| | Bars, Bellet, and Tommasi (2023) | $\ell_1$ On-average Stability | $\mathcal{O}\left(\sum_{t=1}^{T}\eta_t/mn\right)$ |

Table 3: Summary of stability-based generalization bounds under convex case.

**Theorem 2.** *(Excess generalization bounds) Assume that for any $z \in \mathcal{Z}$, $\ell(\mathbf{w};z)$ is non-negative, convex and $\beta$-smooth w.r.t $\mathbf{w}\in\mathcal{W}$. Assuming $\mathbf{w}^t=\frac{1}{m}\sum_{k=1}^{m}\mathbf{w}_k^t$ is generated through DM-SGD (WR) with step sizes that decrease over time, and given $\eta_t\le\frac{1}{2\beta}$ and $\gamma>1$, there holds*

$$\begin{aligned}\sum_{t=1}^{T}\eta_t\mathbb{E}_{S,\mathcal{A}}[R(\mathbf{w}^t)-R(\mathbf{w}^\star)] &\le \frac{\beta}{\gamma}\sum_{t=1}^{T}\eta_t[R(\mathbf{w}^\star)]\\ &+(1+\frac{\beta}{\gamma})\left[(1/2+\beta\eta_1)\|\mathbf{w}\|_2^2+2\beta\sum_{t=1}^{T}\eta_t^2R(\mathbf{w}^\star)\right]\\ &+\left\{\frac{4(\beta+\gamma)(1+T)\beta\lambda^2e}{nb}(1+\frac{b-1}{n})\right\}\times\\ &\left\{\sum_{t=1}^{T}\eta_t(\eta_1\|\mathbf{w}^\star\|_2^2+2\sum_{t=1}^{T}\eta_t^2R(\mathbf{w}^\star))\right\}.\end{aligned}$$

*Furthermore, if step size $\eta_t=\eta=c/\sqrt{T}\le\frac{1}{2\beta}$ and $T\asymp n$, then*

$$\mathbb{E}_{S,\mathcal{A}}[R(\bar{\mathbf{w}}^T)-R(\mathbf{w}^\star)]=\mathcal{O}\left(\frac{n+b}{n^2b}+\frac{R(\mathbf{w}^\star)}{\sqrt{n}}\right).$$

**Remark 7.** *When $R(\mathbf{w}^\star)=0$ and $T\asymp n$, we can get*

$$\mathbb{E}_{S,\mathcal{A}}[R(\bar{\mathbf{w}}^T)-R(\mathbf{w}^\star)]=\mathcal{O}\left(\frac{n+b}{n^2b}\right).$$

*The obtained rate of $\mathcal{O}(n^{-1})$ is commonly deemed sufficiently tight in statistical learning theory (Shalev-Shwartz et al. 2010; Lei and Ying 2020).*

## Strongly Convex Case

We now consider strongly convex functions. *Supplementary Material C* provides the detailed proof of the following theorem.

**Theorem 3.** *Assume that $\ell(\mathbf{w};z)$ is non-negative, $\mu$-strongly convex and $\beta$-smooth w.r.t $\mathbf{w}\in\mathcal{W}$. Then for DM-SGD (WR) with $T$ iterations, we have*

$$\begin{aligned}&\frac{1}{mn}\sum_{k=1}^{m}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\mathbf{w}_k^{T+1}-\tilde{\mathbf{w}}_k^{T+1}\right\|_2\right]\\ &\quad\le\frac{2\sqrt{2\beta}}{mn}\sum_{t=1}^{T}\eta_t\sqrt{\mathbb{E}\left[R_S(w^t)\right]}\prod_{\tilde{t}=t+1}^{T}\left(1-\frac{\eta_{\tilde{t}}\mu}{2}\right).\end{aligned}$$

*and*

$$\begin{aligned}&\frac{1}{mn}\sum_{k=1}^{m}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\mathbf{w}_k^{T+1}-\tilde{\mathbf{w}}_k^{T+1}\right\|_2^2\right]\\ &\le\left\{(1+1/p)\frac{8\beta\lambda^2}{nb}\left(1+\frac{b-1}{n}\right)\right\}\times\\ &\quad\left\{\sum_{t=1}^{T}\eta_t^2[(1+p)]^{T-t}\mathbb{E}\left[R_S(\mathbf{w}^t)\right]\prod_{\tilde{t}=t+1}^{T}\left(1-\frac{\eta_{\tilde{t}}\mu}{2}\right)\right\}.\end{aligned}$$

**Remark 8.** *The difference between the update analysis process under strongly convex and convex condition lies in the expansion operator. Under $\eta<1/\beta$, the expansion operator under strongly convexity is $1-\eta\mu/2$.*

**Proposition 2.** *Suppose that $\ell$ is is non-negative, $\mu$-strongly convex $L$-Lipschitz and $\beta$-smooth w.r.t $\mathbf{w}\in\mathcal{W}$. Then for DM-SGD (WOR) with $T$ iterations, we have*

$$\begin{aligned}&\frac{1}{mn}\sum_{k=1}^{m}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\mathbf{w}_k^{T+1}-\tilde{\mathbf{w}}_k^{T+1}\right\|_2\right]\\ &\quad\le\frac{2L}{mn}\sum_{t=1}^{T}\eta_t\prod_{\tilde{t}=t+1}^{T}\left(1-\frac{\eta_{\tilde{t}}\mu}{2}\right).\end{aligned}$$

*Furthermore, with a fixed step size $\eta_t=\eta\le1/\beta$, we can obtain*

$$\frac{1}{mn}\sum_{k=1}^{m}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\mathbf{w}_k^{T+1}-\tilde{\mathbf{w}}_k^{T+1}\right\|_2\right]\le\frac{4L}{\mu mn}.$$

## Nonconvex Case

In this subsection, we provide error bound for non-convex case which is more challenging and important. The proof is deferred to *Supplementary Material D*.

**Theorem 4.** *(Stability bounds) Assume that $\ell(\mathbf{w}; z)$ is non-negative and $\beta$-smooth w.r.t $\mathbf{w} \in \mathcal{W}$. Then for DM-SGD (WR) with $T$ iterations, we have*

$$\frac{1}{mn}\sum_{k=1}^{m}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\mathbf{w}_k^{T+1}-\tilde{\mathbf{w}}_k^{T+1}\right\|_2\right] \leq \frac{2\sqrt{2\beta}}{mn}\sum_{t=1}^{T}\eta_t\sqrt{\mathbb{E}\left[R_S(w^t)\right]}\prod_{\tilde{t}=t+1}^{T}\left(1+\frac{\eta_{\tilde{t}}(n-1)}{n}\beta\right)$$

*and*

$$\frac{1}{mn}\sum_{k=1}^{m}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\mathbf{w}_k^{T+1}-\tilde{\mathbf{w}}_k^{T+1}\right\|_2^2\right] \leq \left\{(1+1/p)\frac{8\beta\lambda^2}{nb}\left(1+\frac{b-1}{n}\right)\right\}\times \left\{\sum_{t=1}^{T}\eta_t^2[(1+p)]^{T-t}\mathbb{E}\left[R_S(\mathbf{w}^t)\right]\prod_{\tilde{t}=t+1}^{T}(1+\eta_{\tilde{t}}(\beta)^2\right\}.$$

**Remark 9.** *Without the convexity assumption, we can't use the lemma of the expansion factor, need to divide the gradient part into those with and without abnormal points, and use the smooth condition for the part without. Notably, our $\ell_1$ results match the stability error derived from Nikolakakis et al. (2022b) for full-batch GD in the nonconvex case, as we transform Minibatch SGD into a full-sample update that incorporates binomial distributed variables.*

**Proposition 3.** *Suppose that $\ell(\mathbf{w}; z)$ is an L-Lipschitz and $\beta$-smooth loss function for any $z$. Then for DM-SGD (WOR) with $T$ iterations, we have*

$$\frac{1}{mn}\sum_{k=1}^{m}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\mathbf{w}_k^{T+1}-\tilde{\mathbf{w}}_k^{T+1}\right\|_2\right] \leq \frac{2L}{mn}\sum_{t=1}^{T}\eta_t\prod_{\tilde{t}=t+1}^{T}\left(1+\eta_{\tilde{t}}\beta\right).$$

## Extension to Zeroth-Order Oracle

In this section, we consider the Zeroth-Order (ZO) version of DM-SGD (the gradient may not be available in this case). The first-order information is usually obtained by using one point (Bach and Perchet 2016)/two or coordinate points (Nesterov and Spokoiny 2017; Duchi et al. 2015; Shamir 2017) feedback strategy. Zero-order optimization does not require the gradient we usually seek, but it defines a "substitute" based on sampling and difference, which we call "zero-order gradient" for the time being.

We consider the two-points approximation method and use the following formula to estimate $\nabla\ell(\mathbf{w}; z)$,

$$\Delta\ell\left(\mathbf{w}; z\right) := \frac{1}{Q}\sum_{q=1}^{Q}\frac{\ell(\mathbf{w}+\delta U_{qt}; z)-\ell(\mathbf{w}; z)}{\delta}U_{qt}, \quad (10)$$

where $U_{qt} \sim \mathcal{N}(0, I_d)$ is a standard normal random vector, $Q$ is the number of evaluations, and $\delta > 0$ is a small parameter. Clearly, when $Q \to \infty$ and $\delta \to 0$, the "substitute" $\Delta\ell(\mathbf{w}; z)$ becomes the $\nabla\ell(\mathbf{w}; z)$. We replace the fourth step of Algorithm 1 with the following iterative method,

$$\mathbf{w}_j^{t+\frac{1}{2}} = \mathbf{w}_j^t - \frac{\eta_t}{b}\sum_{r=1}^{b}\Delta\ell(\mathbf{w}_j^t; Z_{j,i_{t,r}}), \quad (11)$$

and then get ZO DM-SGD algorithm.

The stability bounds of ZO DM-SGD with convex and non-convex setting are presented in the following propositions, which driven by the recent analysis in the black-box learning (Nikolakakis et al. 2022a). Detailed proofs can be found in *Supplementary Material E*.

**Corollary 1.** *Assume that the loss function $\ell(\mathbf{w}; z)$ is convex, L-Lipschitz and $\beta$-smooth for any $z \in \mathcal{Z}$. Consider the ZO DM-SGD (WR) algorithm with $T$ iterations and $\eta_t \leq 2/\beta$, we have*

$$\frac{1}{mn}\sum_{k=1}^{m}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\mathbf{w}_k^{T+1}-\tilde{\mathbf{w}}_k^{T+1}\right\|_2\right] \leq \frac{2L}{mn}\left(1+\zeta_Q^d\right)\sum_{t=1}^{T}\eta_t\prod_{\tilde{t}=t+1}^{T}\left(1+\eta_{\tilde{t}}\beta\zeta_Q^d\right),$$

*where $\zeta_Q^d = \sqrt{\frac{3d-1}{Q}}$.*

**Corollary 2.** *Suppose that the loss function $\ell(\mathbf{w}; z)$ is L-Lipschitz and $\beta$-smooth for any $z \in \mathcal{Z}$. Consider the ZO DM-SGD (WR) with $T$ iterations, then*

$$\frac{1}{mn}\sum_{k=1}^{m}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\mathbf{w}_k^{T+1}-\tilde{\mathbf{w}}_k^{T+1}\right\|_2\right] \leq \frac{2L}{mn}\left(1+\zeta_Q^d\right)\sum_{t=1}^{T}\eta_t\prod_{\tilde{t}=t+1}^{T}\left(1+\frac{\eta_{\tilde{t}}\beta(n-1)}{n}\left(1+\zeta_Q^d\right)\right).$$

**Remark 10.** *Here only the first-order Taylor expansion is considered for simplicity. Following the strategies in Nikolakakis et al. (2022a), similar results can also be obtained for the second-order approximation.*

## Conclusion

This paper presented the analysis on the stability and generalization of the DM-SGD. By developing analysis techniques associated with algorithmic stability, we established on-average stability bounds for DM-SGD with convex, strongly convex and nonconvex optimization objectives, respectively. The derived convergence rates are comparable with the existing results for D-SGD (Sun, Li, and Wang 2021; Bars, Bellet, and Tommasi 2023). Additionally, the excess generalization error is bounded in terms of the $\ell_2$ on-average stability.

In the future, delving deeper into stability-based generalization bounds for DM-SGD in the realm of non-i.i.d. sampling, such as Markov chain sampling (Sun, Sun, and Yin 2018; Wang et al. 2022; Sun, Li, and Wang 2023), would be an intriguing avenue for exploration.

## Acknowledgments

This work was supported in part by National Natural Science Foundation of China (Nos. 12071166 and 62376104) and the Fundamental Research Funds for the Central Universities of China (Nos. 2662020LXQD002 and 2662023LXPY005).

## References

Agarwal, A.; and Duchi, J. C. 2011. Distributed delayed stochastic optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Bach, F.; and Perchet, V. 2016. Highly-smooth zero-th order online optimization. In *Conference on Learning Theory (COLT)*, 257–283.

Bars, B. L.; Bellet, A.; and Tommasi, M. 2023. Improved Stability and Generalization Analysis of the Decentralized SGD Algorithm. arXiv:2306.02939.

Bassily, R.; Feldman, V.; Guzmán, C.; and Talwar, K. 2020. Stability of stochastic gradient descent on nonsmooth convex losses. In *Advances in Neural Information Processing Systems (NeurIPS)*, 4381–4391.

Bousquet, O.; and Elisseeff, A. 2002. Stability and generalization. *Journal of Machine Learning Research*, 2: 499–526.

Bousquet, O.; Klochkov, Y.; and Zhivotovskiy, N. 2020. Sharper bounds for uniformly stable algorithms. In *Conference on Learning Theory (COLT)*, 610–626.

Cotter, A.; Shamir, O.; Srebro, N.; and Sridharan, K. 2011. Better mini-batch algorithms via accelerated gradient methods. In *Advances in Neural Information Processing Systems(NeurIPS)*, 1647–1655.

Dekel, O.; Gilad-Bachrach, R.; Shamir, O.; and Xiao, L. 2012. Optimal Distributed Online Prediction Using Mini-Batches. *Journal of Machine Learning Research*, 13.

Deng, X.; Sun, T.; Li, S.; and Li, D. 2023. Stability-Based Generalization Analysis of the Asynchronous Decentralized SGD. In *AAAI Conference on Artificial Intelligence*, 7340–7348.

Deng, Z.; He, H.; and Su, W. 2021. Toward better generalization bounds with locally elastic stability. In *International Conference on Machine Learning (ICML)*, 2590–2600.

Duchi, J. C.; Jordan, M. I.; Wainwright, M. J.; and Wibisono, A. 2015. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61: 2788–2806.

Elisseeff, A.; Evgeniou, T.; Pontil, M.; and Kaelbing, L. P. 2005. Stability of randomized learning algorithms. *Journal of Machine Learning Research*, 6: 55–79.

Feldman, V.; and Vondrak, J. 2018. Generalization bounds for uniformly stable algorithms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 9770–9780.

Feldman, V.; and Vondrak, J. 2019. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Conference on Learning Theory (COLT)*, 1270–1279.

Ghadimi, S.; Lan, G.; and Zhang, H. 2016. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155: 267–305.

Gower, R. M.; Loizou, N.; Qian, X.; Sailanbayev, A.; Shulgin, E.; and Richtárik, P. 2019. SGD: General analysis and improved rates. In *International Conference on Machine Learning(ICML)*, 5200–5209.

Hardt, M.; Recht, B.; and Singer, Y. 2016. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning (ICML)*, 1225–1234.

Harvey, N. J.; Liaw, C.; Plan, Y.; and Randhawa, S. 2019. Tight analyses for non-smooth stochastic gradient descent. In *Conference on Learning Theory (COLT)*, 1579–1613.

Kuzborskij, I.; and Lampert, C. 2018. Data-dependent stability of stochastic gradient descent. In *International Conference on Machine Learning (ICML)*, 2815–2824.

Lei, Y.; Ledent, A.; and Kloft, M. 2020. Sharper generalization bounds for pairwise learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 21236–21246.

Lei, Y.; Liu, M.; and Ying, Y. 2021. Generalization guarantee of SGD for pairwise learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 21216–21228.

Lei, Y.; Sun, T.; and Liu, M. 2023. Stability and Generalization for Minibatch SGD and Local SGD. arXiv:2310.01139.

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

Li, M.; Zhang, T.; Chen, Y.; and Smola, A. J. 2014. Efficient mini-batch training for stochastic optimization. In *International Conference on Knowledge Discovery and Data Mining*, 661–670.

Li, X.; Yang, W.; Wang, S.; and Zhang, Z. 2019. Communication-efficient local decentralized SGD methods. arXiv:1910.09126.

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

Lian, X.; Zhang, W.; Zhang, C.; and Liu, J. 2018. Asynchronous decentralized parallel stochastic gradient descent. In *International Conference on Machine Learning (ICML)*, 3043–3052.

Liu, T.; Lugosi, G.; Neu, G.; and Tao, D. 2017. Algorithmic stability and hypothesis complexity. In *International Conference on Machine Learning (ICML)*, 2159–2167.

Nadiradze, G.; Sabour, A.; Davies, P.; Li, S.; and Alistarh, D. 2021. Asynchronous decentralized SGD with quantized and local updates. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 6829–6842.

Nadiradze, G.; Sabour, A.; Davies, P.; Markov, I.; Li, S.; and Alistarh, D. 2020. Decentralized SGD with asynchronous, local and quantized updates. https://openreview.net/pdf?id=x6x7FWFNZpg. Accessed: 2020-09-28.

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

Nesterov, Y.; and Spokoiny, V. 2017. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17: 527–566.

Nikolakakis, K.; Haddadpour, F.; Kalogerias, D.; and Karbasi, A. 2022a. Black-box generalization: Stability of zeroth-order learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 31525–31541.

Nikolakakis, K. E.; Haddadpour, F.; Karbasi, A.; and Kalogerias, D. S. 2022b. Beyond lipschitz: Sharp generalization and excess risk bounds for full-batch gd. arXiv:2204.12446.

Predd, J. B.; Kulkarni, S. B.; and Poor, H. V. 2006. Distributed learning in wireless sensor networks. *IEEE Signal Processing Magazine*, 23: 56–69.

Richards, D.; et al. 2020. Graph-dependent implicit regularisation for distributed stochastic subgradient descent. *Journal of Machine Learning Research*, 21: 1–44.

Schmidt, M.; Roux, N.; and Bach, F. 2011. Convergence rates of inexact proximal-gradient methods for convex optimization. In *Advances in Neural Information Processing Systems(NeurIPS)*.

Shalev-Shwartz, S.; Shamir, O.; Srebro, N.; and Sridharan, K. 2010. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11: 2635–2670.

Shamir, O. 2017. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *Journal of Machine Learning Research*, 18: 1703–1713.

Shamir, O.; and Srebro, N. 2014. Distributed stochastic optimization and learning. In *Annual Allerton Conference on Communication, Control, and Computing*, 850–857.

Sinkhorn, R. 1964. A relationship between arbitrary positive matrices and doubly stochastic matrices. *The Annals of Mathematical Statistics*, 35: 876–879.

Sinkhorn, R.; and Knopp, P. 1967. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21: 343–348.

Sirb, B.; and Ye, X. 2016. Consensus optimization with delayed and stochastic gradients on decentralized networks. In *IEEE International Conference on Big Data*, 76–85.

Srebro, N.; Sridharan, K.; and Tewari, A. 2010. Smoothness, low noise and fast rates. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2199–2207.

Sun, T.; Li, D.; and Wang, B. 2021. Stability and generalization of decentralized stochastic gradient descent. In *AAAI Conference on Artificial Intelligence*, 9756–9764.

Sun, T.; Li, D.; and Wang, B. 2023. On the decentralized stochastic gradient descent with markov chain sampling. *IEEE Transactions on Signal Processing*, 1–14.

Sun, T.; Sun, Y.; and Yin, W. 2018. On Markov chain gradient descent. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Sundhar Ram, S.; Nedić, A.; and Veeravalli, V. V. 2010. Distributed stochastic subgradient projection algorithms for convex optimization. *Journal of Optimization Theory and Applications*, 147: 516–545.

Taheri, H.; and Thrampoulidis, C. 2023. On generalization of decentralized learning with separable data. In *International Conference on Artificial Intelligence and Statistics*, 4917–4945.

Tang, H.; Lian, X.; Yan, M.; Zhang, C.; and Liu, J. 2018. Decentralized training over decentralized data. In *International Conference on Machine Learning (ICML)*, 4848–4856.

Tsitsiklis, J.; Bertsekas, D.; and Athans, M. 1986. Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE Transactions on Automatic Control*, 31: 803–812.

Tsitsiklis, J. N. 1984. *Problems in decentralized decision making and computation*. Ph.D. thesis, Massachusetts Institute of Technology.

Vogels, T.; He, L.; Koloskova, A.; Karimireddy, S. P.; Lin, T.; Stich, S. U.; and Jaggi, M. 2021. Relaysum for decentralized deep learning on heterogeneous data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 28004–28015.

Wang, P.; Lei, Y.; Ying, Y.; and Zhou, D.-X. 2022. Stability and generalization for markov chain stochastic gradient methods. In *Advances in Neural Information Processing Systems (NeurIPS)*, 37735–37748.

Woodworth, B. E.; Patel, K. K.; and Srebro, N. 2020. Minibatch vs local sgd for heterogeneous distributed learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 6281–6292.

Xin, R.; Khan, U. A.; and Kar, S. 2021. An improved convergence analysis for decentralized online stochastic nonconvex optimization. *IEEE Transactions on Signal Processing*, 69: 1842–1858.

Xu, J.; Zhang, W.; and Wang, F. 2021. A (DP)2 SGD: Asynchronous decentralized parallel stochastic gradient descent With differential privacy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44: 8036–8047.

Yin, D.; Pananjady, A.; Lam, M.; Papailiopoulos, D.; Ramchandran, K.; and Bartlett, P. 2018. Gradient diversity: a key ingredient for scalable distributed learning. In *International Conference on Artificial Intelligence and Statistics*, 1998–2007.

Zhang, J.; and You, K. 2019. Decentralized stochastic gradient tracking for non-convex empirical risk minimization. arXiv:1909.02712.

Zhu, T.; He, F.; Zhang, L.; Niu, Z.; Song, M.; and Tao, D. 2022. Topology-aware generalization of decentralized sgd. In *International Conference on Machine Learning (ICML)*, 27479–27503.

Zinkevich, M.; Weimer, M.; Li, L.; and Smola, A. 2010. Parallelized stochastic gradient descent. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 23.