# Stability and Generalization of Zeroth-Order Decentralized Stochastic Gradient Descent with Changing Topology

**Xiaolin Hu**[1*†] **Zixuan Gong**[1*], **Gengze Xu**[1], **Wei Liu**[2], **Jian Luan**[2], **Bin Wang**[2], **Yong Liu**[1‡]

[1]Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China
[2]XiaoMi AI Lab, Beijing, China
{xiaolinhu, liuyonggsai}@ruc.edu.cn

## Abstract

Zeroth-order (ZO) optimization as the gradient-free method has become a powerful tool when the first-order gradient is unavailable or expensive to obtain, especially in decentralized learning scenarios where data and computational resources are distributed across multiple clients. There have been many efforts to analyze the optimization convergence rate of zeroth-order decentralized stochastic gradient descent (ZO-DSGD) algorithms. However, the generalization of these methods has not been well studied. In this paper, we provide a generalization analysis of ZO-DSGD with changing topology, where the clients run zeroth-order SGD with local data and communicate with each other according to time-varying topology. We systematically analyze the generalization error in convex, strongly convex, and non-convex cases. The obtained results in the convex and strongly convex cases with zeroth-order oracles recover the results of SGD. Moreover, the generalization bounds derived in non-convex cases align with that of DSGD. To capture the influence of communication topology on the generalization performance, we analyze local generalization bounds concerning local models held at different clients. The obtained results reflect the influence of the number of clients, local sample size, and topology on the generalization error. To the best of our knowledge, this is the first work that provides a generalization analysis of zeroth-order decentralized stochastic gradient descent methods and recovers the results of SGD.

## Introduction

Zeroth-order (ZO) optimization algorithms play an important role in scenarios where gradients are either unavailable or too costly to compute, such as black-box optimization (Chen et al. 2024; Fang et al. 2022), reinforcement learning (Zhong et al. 2024), and prompt engineering (Liu et al. 2023). In addition to the constraints on gradient computation, practical applications often involve decentralized clients with limited communication capabilities (Tang et al. 2018; Koloskova et al. 2020; Martínez Beltrán et al. 2023). Zeroth-order decentralized stochastic gradient descent(ZO-DSGD) and its variants offer a promising solution to address constraints on data privacy (Chen et al. 2023), network communication (Qiu, Shanbhag, and Yousefian 2023), and gradient computation (Chen et al. 2024). In this approach, clients perform local updates using zeroth-order estimated gradients and communicate with each other according to the network topology.

The theoretical analysis of zeroth-order stochastic algorithms can be divided into two primary categories: optimization error (Nesterov and Spokoiny 2017) and generalization error (Nikolakakis et al. 2022). Specifically, optimization error reflects the convergence speed of the algorithm on seen data during training (Duchi et al. 2015). Generalization error measures the algorithm's performance on unseen data, which is fundamental for evaluating the machine learning models (Liu et al. 2024). The optimization analysis of zeroth-order algorithms in the decentralized learning setting concern the communication complexity (Li et al. 2019; Fang et al. 2022), dimension dependence (Gu et al. 2024; Li et al. 2024) and convergence rate (Shamir 2017; Ling et al. 2024). Existing generalization analysis of zeroth-order algorithms primarily focuses on centralized settings (Nikolakakis et al. 2022; Chen et al. 2023; Liu et al. 2024). Although these works indicate that the generalization bounds for centralized ZO-SGD are comparable to those of SGD in non-convex case (Nikolakakis et al. 2022), it remains an open question whether these bounds align in convex case (Liu et al. 2024). Furthermore, the generalization performance of zeroth-order stochastic algorithms in decentralized settings (Wang and Chen 2024) has not been carefully examined.

Recent works have developed generalization error bounds for decentralized stochastic gradient descent (DSGD) using algorithm stability tools (Richards and Rebeschini 2020; Sun, Li, and Wang 2021; Deng et al. 2023; Zhu et al. 2022; Le Bars et al. 2024). Algorithm stability assesses the sensitivity of the random algorithm to perturbations in the training dataset, providing insights into the generalization error of the algorithm (Bousquet and Elisseeff 2002; Lei and Ying 2020). The stability analysis of DSGD involves the randomness of the data samples and the communication topology among clients at each iteration (Wang and Chen 2024). However, the influence of the communication topology on the generalization performance has not been well understood. Some of the existing generalization bounds of DSGD are topology-dependent and suggest that sparse communica-

*These authors contributed equally.
†Work done during the internship at XiaoMi.
‡Corresponding author.

tion topology has a negative impact on generalization (Sun, Li, and Wang 2021; Deng et al. 2023; Zhu et al. 2022). However, another line of work derives topology-independent generalization bounds for DSGD, which indicates that the communication topology has no effect on the generalization error (Richards and Rebeschini 2020; Le Bars et al. 2024).

The inconsistency of generalization bounds on centralized zeroth-order SGD and first-order DSGD algorithms motivates us to ask the following question: **How does the communication topology of ZO-DSGD impact its generalization error? Is the generalization bound of ZO-DSGD consistent with DSGD and SGD?**

To answer the above questions, we investigate the generalization error of zeroth-order decentralized SGD (ZO-DSGD) with changing topology. Our analysis framework covers zeroth-order federated learning (Local SGD) as a special case. We provide systematic generalization error bounds for ZO-DSGD in the convex, strongly convex, and non-convex cases. The obtained results for the consensus model match those of SGD in the convex and strongly convex cases, and they recover the results of DSGD in the non-convex case. Moreover, we derive generalization error bounds for local models, and the obtained results reflect the influence of the communication topology. Our contributions are summarized as follows:

- We conduct a comprehensive analysis of the generalization error of ZO-DSGD. Instead of analyzing the decentralized systems with fixed communication topology, we consider the case where the communication topology may change over time. Our unifying framework covers a class of methods, such as zeroth-order federated learning (Local SGD) and its decentralized variants.
- We provide generalization error bounds for the consensus model of ZO-DSGD using algorithm stability tools. The obtained results recover those of SGD in the convex, strongly convex, and non-convex cases. To the best of our knowledge, this is the first work that provides generalization bounds matching those of SGD in the convex and strongly convex cases.
- We derive generalization error bounds for the local models of ZO-DSGD, which is more practical when the distributed data is heterogeneous and communication is limited or unstable. The derived generalization bounds of local models reflect the influence of the communication topology. Our results match the generalization bounds of DSGD in the convex, strongly convex, and non-convex cases with proper learning rates.

**Notations.** Throughout this paper, we denote the unit sphere by $\mathbb{S}^{d-1}$. For any $\mathbf{x} \in \mathbb{R}^d$, we denote by $\|\mathbf{x}\| =$ the standard $\ell_2$-norm. $\nabla f(\cdot)$ denotes the gradient of a function $f$, and $\widetilde{\nabla} f(\cdot)$ denotes the zeroth-order gradient estimator. $[m]$ denotes the set $\{1, \cdots, m\}$. $I_m \in \mathbb{R}^{m\times m}$ denotes the identity matrix. The parameter $\Gamma_K^d = \sqrt{\frac{K+d-1}{K}}$ is frequently used in our analysis.

Algorithm 1: ZEROTH-ORDER DECENTRALIZED SGD

**Require:** for each client $i \in [m]$ initialize $\mathbf{x}_i^0 \in \mathbb{R}^d$, learning rate $\{\eta_t\}_{t=0}^{T-1}$, iteration number $T$, mixing matrix distributions $\mathcal{W}^t$.

**for** $t$ **in** $0 \dots T-1$ **do**
  Sample $\mathbf{W}^t \sim \mathcal{W}^t$
  **for** client $i \in [m]$ in parallel **do**
    Generate $Z_i^{j_t}$ and $\{\mathbf{v}_{i,k}^t\}_{k=1}^K$ from $S_i$ and $\sqrt{d}\mathbb{S}^{d-1}$ uniformly
    Compute $\mathbf{g}_i^t \leftarrow \widetilde{\nabla} f_i(\mathbf{x}_i^t; Z_i^{j_t}, \{\mathbf{v}_{i,k}^t\}_{k=1}^K, \epsilon)$
    $\mathbf{x}_i^{t+\frac{1}{2}} = \mathbf{x}_i^t - \eta_t \mathbf{g}_i^t$ ▷ stochastic updates with zeroth-order oracles
    $\mathbf{x}_i^{t+1} \leftarrow \sum_{j=1}^m \mathbf{W}_{ij}^t \mathbf{x}_j^{t+\frac{1}{2}}$ ▷ gossip averaging
  **end for**
**end for**

# Preliminaries

## Distributed Learning with Heterogeneous Data

Consider a distributed system with $m$ clients (Shamir and Srebro 2014). Let $D_i$ be the unknown data distribution associated with the $i$-th client $(i = 1, \cdots, m)$, which is defined on a sample space $\mathcal{Z}$. If $D_i \equiv D$ for all clients, the system is called homogeneous. In this paper, we consider the heterogeneous distributed setting where the $D_i$ varies across clients (Zhao et al. 2018; Reddi et al. 2020). The goal of distributed learning is to learn a hypothesis parameterized by $\mathbf{x} \in \Omega, \Omega \subseteq \mathbb{R}^d$, to minimize the global population risk $F(\mathbf{x}) : \Omega \to \mathbb{R}$. That is,

$$F(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^m F_i(\mathbf{x}), \quad F_i(\mathbf{x}) = \mathbb{E}_{Z_i\sim D_i}[f(\mathbf{x}; Z_i)],$$

where each client $i \in [m]$ holds the local population risk $F_i(\mathbf{x})$ realized by its distribution $D_i$, and $f(\mathbf{x}; \cdot)$ is the objective function.

In practice, it is impossible to access the population risk $F(\mathbf{x})$ since the distributions $\{D_i\}_{i=1}^m$ are unknown. Therefore, we turn to optimize the global empirical risk realized by the data set sampled from the unknown distributions $\{D_i\}_{i=1}^m$. Let $S_i = \{Z_i^j\}_{j=1}^n$ denote the local training set located on the $i$-th client, which consists of $n$ i.i.d. realizations of $Z$ following $D_i$. The global empirical risk $F_S(\mathbf{x})$ is defined as

$$F_S(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^m F_{S_i}(\mathbf{x}), \quad F_{S_i}(\mathbf{x}) = \frac{1}{n}\sum_{j=1}^n f(\mathbf{x}; Z_i^j),$$

where each $F_{S_i}(\mathbf{x})$ denotes the local empirical risk at the $i$-th client, and $S = S_1 \cup \cdots \cup S_m$ represents the global training set across all clients.

## Decentralized Learning with Changing Topology

In a decentralized setting, the topology of the communication network can be represented as an undirected graph: $(V, E)$, where $V = \{1, \cdots, m\}$ denotes the client set and

$E \subseteq V \times V$ represents the edge set. The communication topology is represented by a mixing matrix $\mathbf{W} \in \mathbb{R}^{m\times m}$, where $\mathbf{W}_{ij} \geq 0$ and $\mathbf{W}_{ij}$ denotes the communication weight from client $j$ to client $i$. The goal of decentralized learning is to learn local models $\{\mathbf{x}_i^T\}_{i=1}^m$ or a consensus model $\mathbf{x}^T = \frac{1}{m}\sum_{i=1}^m \mathbf{x}_i^T$, where $\mathbf{x}_i$ denotes the local model on the $i$-th client and $T$ is the total iteration number.

Let $\mathbf{x}_i^t$ represent the local model on the $i$-th client at the $t$-th step, and $\mathbf{x}_i^0 \equiv \mathbf{x}^0$ be the initial point. At iteration $t$, each client first updates its local model by the stochastic gradient $\nabla f(\mathbf{x}_i^t; Z_i^{j_t})$, where $Z_i^{j_t}$ is the data uniformly sampled from $S_i$. The local update rule can be written as $\mathbf{x}_i^{t+\frac{1}{2}} = \mathbf{x}_i^t - \eta_t \nabla f(\mathbf{x}_i^t; Z_i^{j_t})$, where $\eta_t$ is the learning rate at step $t$. Then, each client aggregates its updated models with its neighbors according to the communication matrix $\mathbf{W}^t$ at iteration $t$. The aggregating step can be written as $\mathbf{x}_i^{t+1} = \sum_{j=1}^m \mathbf{W}_{ij}^t \mathbf{x}_j^{t+\frac{1}{2}}$.

## Decentralized Learning with Zeroth-order Oracles

ZO optimization methods are widely used in scenarios where the gradients are not available or too expensive to compute. Instead of computing the first-order gradient, ZO methods perturb the model and use the corresponding function values to estimate the gradient. In this paper, we consider the forward difference estimator, which is defined as

$$\begin{aligned}&\widetilde{\nabla} f(\mathbf{x}_i^t; Z_i^{j_t}, \{\mathbf{v}_{i,k}^t\}_{k=1}^K, \epsilon)\\ &= \frac{1}{K}\sum_{k=1}^K \frac{f(\mathbf{x}_i^t + \epsilon \mathbf{v}_{i,k}^t; Z_i^{j_t}) - f(\mathbf{x}_i^t; Z_i^{j_t})}{\epsilon}\mathbf{v}_{i,k},\end{aligned}$$

where $\mathbf{x}_i^t$ is the local model at time $t$, $Z_i^{j_t}$ is the data uniformly sampled from $S_i$, $\epsilon > 0$ is the perturbation parameter and $\{\mathbf{v}_{i,k}^t\}_{k=1}^K$ are i.i.d. random vectors uniformly sampled from the sphere $\sqrt{d}\mathbb{S}^{d-1}$. Although our theoretical results are derived for random vectors sampled from $\sqrt{d}\mathbb{S}^{d-1}$, they can be extended to other distributions, such as standard Gaussian distribution $\mathcal{N}(0, I_d)$ or uniformly $l_2$-ball distribution $\mathcal{U}(\sqrt{d+2}\mathbb{B}^{d-1})$ (Liu et al. 2024).

In this paper, we consider zeroth-order decentralized SGD defined in Algorithm 1. At iteration $t$, each client $i$ first samples a data point $Z_i^{j_t}$ and $K$ random vectors $\{\mathbf{v}_{i,k}^t\}_{k=1}^K$ from the corresponding uniform distribution, and then estimate the gradient $\widetilde{\nabla} f(\mathbf{x}_i^t; Z_i^{j_t}, \{\mathbf{v}_{i,k}^t\}_{k=1}^K, \epsilon)$ using the forward difference estimator. Then, each client updates its local model $\mathbf{x}_i^{t+\frac{1}{2}}$ using the estimated gradient. Finally, each client averages its updated models with its neighbors according to the communication matrix $\mathbf{W}^t$ generated from a distribution $\mathcal{W}^t$ at iteration $t$.

Note that zeroth-order federated learning (Local SGD) can be seen as a special case of decentralized learning with a time-varying topology. That is, the mixing matrix $\mathbf{W}^t$ is set as the identity matrix $I_m$ at local update steps and is set as the original mixing matrix at gossip averaging steps.

## Assumptions and Definitions

**Assumption 1 (Lipschitz continuous)** *Assume the objective function $f$ is $L$-Lipschitz. That is, for any $z \in \mathcal{Z}$ and any $\mathbf{x}, \mathbf{x}' \in \Omega$, there exists constant $L \geq 0$ such that*

$$\|f(\mathbf{x}, z) - f(\mathbf{x}', z)\| \leq L\|\mathbf{x} - \mathbf{x}'\|.$$

**Assumption 2 (Smoothness)** *Let $\beta \geq 0$. For any sample $z \in \mathcal{Z}$ and $\mathbf{x}, \mathbf{x}' \in \Omega$, there has*

$$\|\nabla f(\mathbf{x}, z) - \nabla f(\mathbf{x}', z)\| \leq \beta\|\mathbf{x} - \mathbf{x}'\|.$$

**Assumption 3 (Strongly convex)** *Let $\Omega$ be a convex compact set, the objective function $f : \Omega \to \mathbb{R}$ is $\mu$-strongly convex, i.e., for all $\mathbf{x}, \mathbf{x}' \in \Omega$, we have*

$$f(\mathbf{x}) \geq f(\mathbf{x}') + \nabla f(\mathbf{x}')^\top(\mathbf{x} - \mathbf{x}') + \frac{\mu}{2}\|\mathbf{x} - \mathbf{x}'\|^2.$$

**Remark 1** *Lipschitz continuity and smoothness are widely used in the analysis of optimization and generalization (Reddi et al. 2016; Allen-Zhu and Hazan 2016; Bousquet and Elisseeff 2002; Hardt, Recht, and Singer 2016). Although standard, these properties can be relaxed using the techniques for generalization analysis in (Lei and Ying 2020; Chen et al. 2023; Lei, Sun, and Liu 2023).*

**Assumption 4** *The mixing matrix $\mathbf{W}^t$ is doubly stochastic, i.e., $\mathbf{W}_{ij}^t \geq 0$ for all $i, j \in [m]$ and $\sum_{j=1}^m \mathbf{W}_{ij}^t = \sum_{i=1}^m \mathbf{W}_{ij}^t = 1$ for all $i, j \in [m]$.*

**Remark 2** *In this paper, we do not require the communication network to be connected. The mixing matrix $\mathbf{W}^t$ can be set as identify matrix $I_m$, which means that each client runs zeroth-order local update without communication at step $t$.*

Excess error measures the distance between the learned consensus model $A(S)$ and the optimal model $\mathbf{x}^*$ with respect to the global population risk. It can be decomposed into generalization error and optimization error.

**Definition 1 (Excess Error Decomposition)** *Denote $\mathbf{x}_S^* = \arg\min_{\mathbf{x}} F_S(\mathbf{x})$ and $\mathbf{x}^* = \arg\min_{\mathbf{x}} F(\mathbf{x})$ as the optimal model of global empirical risk and population risk minimization respectively. Let $A(S)$ represent the consensus model generating by algorithm $A$ on dataset $S$, then the excess error $F(A(S)) - F(\mathbf{x}^*)$ can be decomposed as*

$$\begin{aligned}\underbrace{\mathbb{E}_{A,S}\left[F(A(S)) - F\left(\mathbf{x}^*\right)\right]}_{\text{Excess error}} \leq &\underbrace{\mathbb{E}_{A,S}\left[F(A(S)) - F_S(A(S))\right]}_{\text{Generalization error}}\\ &+ \underbrace{\mathbb{E}_{A,S}\left[F_S(A(S)) - F_S\left(\mathbf{x}_S^*\right)\right]}_{\text{Optimization error}},\end{aligned}$$

*where the inequality holds since $\mathbb{E}_{A,S}\left[F_S\left(\mathbf{x}_S^*\right)\right] \leq \mathbb{E}_{A,S}\left[F_S\left(\mathbf{x}^*\right)\right] = \mathbb{E}_S\left[F\left(\mathbf{x}^*\right)\right]$.*

The second term in the decomposition is the optimization error, which measures the performance of consensus model $A(S)$ with respect to the global empirical risk. It can be addressed by optimization tools in decentralized learning. In this paper, we focus on the first term, generalization error, which measures the performance gap between the global population risk and the global empirical risk.

Algorithm stability is a popular tool for bounding the generalization error of randomized algorithms (Bousquet and Elisseeff 2002; Elisseeff, Evgeniou, and Pontil 2005). The stability measures the sensitivity of a random algorithm to perturbations in the training samples. To bound the generalization error of the consensus model, we first introduce the definition of uniform stability (Hardt, Recht, and Singer 2016; Liu et al. 2024) for heterogeneous data.

**Definition 2 (Neighboring Datasets)** *Two datasets are said to be neighboring if they only differ in one data point. Let $Z_i^j \in S$ be the $j$-th data point on the $i$-th client and $\hat{Z}_i^j \in S^{(i,j)}$ be the independent copy of $Z_i^j$, we use $S$ and $S^{(i,j)}$ to denote the neighboring datasets with different data points $Z_i^j$ and $\hat{Z}_i^j$.*

**Definition 3 (Uniform Stability (Lei and Ying 2020))** *Let $A$ be a randomized algorithm, we say $A$ is $\epsilon$-uniformly stable if for all neighboring datasets $S, S^{(i,j)}$, we have*

$$\sup_Z \mathbb{E}_A \left[ \left| f(A(S), Z) - f\left(A\left(S^{(i,j)}\right), Z\right) \right| \right] \leq \epsilon.$$

For generalization analysis of centralized or homogeneous distributed algorithms, the definition of uniform stability only involves one unknown data distribution. In heterogeneous decentralized systems with $m$ clients, its definition involves $m$ unknown data distributions. The relationship between uniform stability and generalization error is given in the following lemma (Lei and Ying 2020).

**Lemma 1 (Generalization via Uniform Stability)** *If the consensus model learned by ZO-DSGD is $\epsilon$-stable under uniform stability in function values, we have*

$$|\mathbb{E}_{A,S} \left[ F(A(S)) - F_S(A(S)) \right]| \leq \epsilon.$$

**Remark 3** *Uniform stability measures the distance between learned models conducted on neighboring datasets $S$ and $S^{(i,j)}$. With the assumption that the loss function is L-Lipschitz, the uniform stability $\epsilon$ can be chosen as $L\mathbb{E}_A \left[ \left\| A(S) - A(S^{(i,j)}) \right\| \right]$. For stochastic gradient algorithms, the expectation is taken over the randomness of the sampled data at each iteration. For the zeroth-order stochastic algorithms, the expectation is taken over the randomness of the sampled data and the random vectors used to perturb the model at each step.*

Our generalization error bounds depend on the communication matrix product defined as follows.

**Definition 4 (Communication Matrix Product $\mathbf{W}^{T-1:t}$ )** *For any $t \leq T-1$, let $\mathbf{W}^{T-1:t} = \mathbf{W}^{T-1}\mathbf{W}^{T-2}\cdots\mathbf{W}^{t}$ be the product of matrices from iteration $T-1$ to $t$, where $\mathbf{W}^t$ is the communication matrix at iteration $t$.*

## Generalization Error for Consensus Model

In this section, we present the generalization bounds of ZO-DSGD for the consensus model $A(S) = \frac{1}{m}\sum_{i=1}^m \mathbf{x}_i^T$. Our results recover the generalization bounds of DSGD in the non-convex cases and match the generalization bounds of SGD in the convex and strongly convex cases. As far as we know, this is the first work that recovers the results of SGD in the line of zeroth-order stochastic algorithms.

### Stability Analysis

Stability analysis involves the upper bound of difference between the consensus model $A(S)$ and $A(S^{(ij)})$, where $S^{(ij)}$ is the neighboring dataset of $S$. Let $\mathbf{x}^t = \frac{1}{m}\sum_{i=1}^m \mathbf{x}_i^t$ and $\tilde{\mathbf{x}}^t = \frac{1}{m}\sum_{i=1}^m \tilde{\mathbf{x}}_i^t$ be the consensus models induced by neighboring dataset $S$ and $S^{(ij)}$, respectively. The update rules of $i$-th client at step $t$ can be written as:

$$\mathbf{x}_i^{t+1} = \mathbf{x}_i^t - \eta_t \widetilde{\nabla} f(\mathbf{x}_i^t; Z_i^{j_t}, \{\mathbf{v}_{i,k}^t\}_{k=1}^k, \epsilon),$$
$$\tilde{\mathbf{x}}_i^{t+1} = \tilde{\mathbf{x}}_i^t - \eta_t \widetilde{\nabla} f(\tilde{\mathbf{x}}_i^t; \tilde{Z}_i^{j_t}, \{\mathbf{v}_{i,k}^t\}_{k=1}^k, \epsilon),$$

where $j_t$ is the index of the data point uniformly sampled from $[n]$ at iteration $t$. The key step of stability-based generalization analysis is to bound the iterate stability error $\mathbb{E}_A[\|\mathbf{x}^T - \tilde{\mathbf{x}}^T\|]$. Previous generalization analysis of zeroth-order algorithms rely on the decomposition technique developed in (Nikolakakis et al. 2022), which directly breaks the zeroth-order stability error into gradient approximation error and first-order stability term. However, previous results based on this decomposition technique fail to recover the generalization bounds of SGD in the convex and strongly convex cases. Instead of decomposing the zeroth-order stability error into the gradient stability term and gradient approximation term, we carefully decompose the zeroth-order stability error by considering the convexity of the objective function. With this new decomposition technique, we derive the following lemma on the zeroth-order growth recursion.

**Lemma 2 (Zeroth-order Growth Recursion)** *Consider two update sequences $\{\mathbf{x}_i^t\}_{t=0}^T$ and $\{\tilde{\mathbf{x}}_i^t\}_{t=0}^T$ started from the same initial point $\mathbf{x}_i^0 = \tilde{\mathbf{x}}_i^0$. Let $\delta_i^t = \|\mathbf{x}_i^t - \tilde{\mathbf{x}}_i^t\|$ be the stability error at step $t$. If $f$ is $\beta$-smooth, there exist a sequence of weights $\alpha_t$ such that the following properties hold for any $i \in [m]$, and $0 \leq t < T$.*

- *If $Z_t^{j_t} = \tilde{Z}_t^{j_t}$, we have $\mathbb{E}\|\delta_i^{t+1}\| \leq \alpha_t \mathbb{E}\|\delta_i^t\| + \epsilon\eta_t\beta d^{3/2}$.*
- *If $Z_t^{j_t} \neq \tilde{Z}_t^{j_t}$, we have $\mathbb{E}\|\delta_i^{t+1}\| \leq \min(\alpha_t, 1)\mathbb{E}\|\delta_i^t\| + 2\eta_t L\Gamma_K^d + \epsilon\eta_t\beta d^{\frac{3}{2}}$.*

*Specifically, $\alpha_t$ can be chosen as follows.*

- *If $f$ is $\beta$-smooth, $\alpha_t = 1 + \eta_t\beta\Gamma_K^d$.*
- *If $f$ is convex and $\beta$-smooth, then for $\eta_t \leq \frac{2K}{\beta(K+d-1)}$, $\alpha_t = 1$.*
- *If $f$ is $\mu$-strongly convex and $\beta$-smooth, then for any $\eta_t \leq \frac{2K}{(\beta+\mu)(K+d-1)}$, $\alpha_t = 1 - \frac{\eta_t\beta\mu}{\beta+\mu}$.*

**Remark 4** *Lemma 2 provides the growth recursion of the zeroth-order update rule in the convex, strongly convex, and non-convex cases. The last term in the growth recursion is the second-order approximation error of the ZO estimator. Compared with the growth recursion of SGD developed in (Hardt, Recht, and Singer 2016), the choice of learning rate $\eta_t$ in Lemma 2 depends on the model parameter dimension $d$ and the number of random perturbations $K$. Compared with the growth recursion of ZO-SGD developed in (Nikolakakis et al. 2022; Liu et al. 2024), our results are tighter.*

## Generalization Error Bounds for Consensus Model

With the zeroth-order growth recursion in Lemma 2, we derive generalization error bounds for the consensus model $A(S) = \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i^T$ learned by ZO-DSGD. Recent works have shown that the generalization bounds of DSGD match those of SGD in the strongly convex case. However, there is still a gap between the generalization bounds of ZO-SGD and SGD in the convex and strongly convex cases (Liu et al. 2024). In this paper, we aim to bridge this gap by providing generalization bounds for ZO-DSGD.

**Theorem 3 (Convex Case)** *Assume that the loss function $f(\cdot, z)$ is convex, L-Lipschitz (Assumption 1) and $\beta$-smooth (Assumption 2). Let $A(S) = \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i^T$ be the final consensus model produced by algorithm ZO-DSGD with mixing matrix $\mathbf{W}^t$ satisfying Assumption 4. When $\eta_t \leq \frac{2K}{\beta(K+d-1)}$ and $\epsilon \leq \frac{L\Gamma_k^d}{\beta m n d^{3/2}}$, ZO-DSGD has a bounded expected generalization error:*

$$|\mathbb{E}_{A,S}\left[F(A(S)) - F_S(A(S))\right]| \leq \frac{3L^2\Gamma_k^d\sum_{t=0}^{T-1}\eta_t}{mn}.$$

*Moreover, with constant learning rate $\eta_t \equiv \eta_0$, the upper bound is simplified as $\frac{3L^2\Gamma_k^d\eta_0 T}{mn}$.*

**Remark 5** *Theorem 3 provides global generalization bounds on average final model $\mathbf{x}^T = \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i^T$. When applying constant learning rate, Theorem 3 matches the optimal rate of order $\mathcal{O}(\frac{T}{mn})$ in centralized SGD (Hardt, Recht, and Singer 2016; Zhang et al. 2022). Previous generalization bounds of order $\mathcal{O}(\frac{T}{mn})$ for ZO-SGD require applying monotonically non-increasing learning rate $\frac{C}{t+1}$ (Liu et al. 2024), which is a stronger assumption.*

Next, we consider the strongly convex case. Let $\Pi_\Omega(\mathbf{x}) = \arg\min_{\mathbf{x}'\in\Omega}\|\mathbf{x}-\mathbf{x}'\|$ be the Euclidean projection onto the convex compact set $\Omega$. In the strongly convex case, we replace the stochastic update step in Algorithm 1 with its projected version as follows:

$$\mathbf{x}_i^{t+\frac{1}{2}} = \Pi_\Omega\left(\mathbf{x}_i^t - \eta_t\nabla f(\mathbf{x}_i^t; Z_i^{j_t})\right).$$

**Theorem 4 (Strongly Convex Case)** *Assume that the loss function $f(\cdot, z)$ is strongly convex, L-Lipschitz (Assumption 1) and $\beta$-smooth (Assumption 2). Let $A(S) = \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i^T$ be the final consensus model produced by algorithm ZO-DSGD with mixing matrix $\mathbf{W}^t$ satisfying Assumption 4. When $\eta_t \leq \frac{K}{\beta(K+d-1)}, \epsilon \leq \frac{L\Gamma_k^d}{\beta m n d^{3/2}}$, ZO-DSGD has a bounded expected generalization error:*

$$|\mathbb{E}_{A,S}\left[F(A(S)) - F_S(A(S))\right]| \leq \frac{6L^2\Gamma_k^d}{\mu mn}.$$

**Remark 6** *Theorem 4 presents generalization error bounds that match that of SGD (Hardt, Recht, and Singer 2016) and DSGD (Le Bars et al. 2024) in the strongly convex case. This is the first generalization bound for a ZO algorithm that aligns with the result of SGD. The key to getting improved results is the new decomposition technique used in Lemma 2, which allows us to control the ZO growth recursion.*

**Theorem 5 (Non-convex Case)** *Assume that the loss function $f(\cdot, z) \in [0,1]$ is L-lipschitz (Assumption 1) and $\beta$-smooth (Assumption 2). Let $A(S) = \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i^T$ be the final consensus model produced by algorithm ZO-DSGD with mixing matrix $\mathbf{W}^t$ satisfying Assumption 4. When $\eta_t \leq \frac{C}{(t+1)\Gamma_K^d}, \epsilon \leq \frac{L\Gamma_k^d}{\beta m n d^{3/2}}$, ZO-DSGD has a bounded expected generalization error:*

$$|\mathbb{E}_{A,S}\left[F(A(S)) - F_S(A(S))\right]| \leq \left(1+\frac{1}{\beta C}\right)\left(3L^2C\right)^{\frac{1}{\beta C+1}}\frac{T^{\frac{\beta C}{\beta C+1}}}{m^{\frac{1}{\beta C+1}}n}.$$

**Remark 7** *Theorem 5 presents the generalization error of ZO-DSGD in the non-convex case. The obtained generalization bound is of order $\mathcal{O}\left(\frac{T^a}{m^{1-a}n}\right)$, where $a = \frac{\beta C}{\beta C+1} \in (0,1)$. This bound aligns with the result for DSGD derived recently in Le Bars et al. (2024).*

Note that the generalization bounds in Theorem 3, 4, and 5 are topology-independent. It indicates that the communication topology has no impact on the generalization error of the consensus model learned by ZO-DSGD. This conclusion aligns with previous generalization analysis of DSGD (Le Bars et al. 2024; Wang and Chen 2024). Nonetheless, considering that the optimization error is influenced by the topology, the final excess risk still depends on the topology.

## Generalization Error for Local Models

In the last section, we provide generalization bounds for the average model $\mathbf{x}^t = \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i^t$ with respect to global generalization error $F(\mathbf{x}^T) - F_S(\mathbf{x}^T)$. However, the global generalization bound fails to characterize the individual generalization performance $F_i(\mathbf{x}_i^T) - F_{S_i}(\mathbf{x}_i^T)$ of the local model at each client. Considering that obtaining the average model $\mathbf{x}^t$ is challenging when the communication is limited or unstable, we provide the generalization bounds for local models $\{\mathbf{x}_i^t\}_{i=1}^m$ in this section. The corresponding results are practical, especially in decentralized systems with changing topology. Specifically, we provide generalization bounds for local models in the convex, strongly convex, and non-convex cases. The obtained results reflect the influence of the communication topology on the generalization performance.

**Theorem 6 (Convex Case)** *Assume that the loss function $f(\cdot, z)$ is convex, L-Lipschitz (Assumption 1) and $\beta$-smooth (Assumption 2). Let $\mathbf{W}^{T-1:t}$ denote the matrix product as defined in Definition 4, and let $A_i(S) = \mathbf{x}_i^T$ represent the local model at the $i$-th client after $T$ iterations of ZO-DSGD. Assume $\epsilon \leq \frac{L\Gamma_k^d}{\beta m n d^{3/2}}$, $\eta_t \leq \frac{2K}{\beta(K+d-1)}$, and that the mixing matrix $\mathbf{W}^t$ satisfies Assumption 4. For any $i \in [m]$, the local generalization error associated with $A_i(S)$ is bounded by:*

$$|\mathbb{E}_{S,A}\left[F_i(A_i(S)) - F_{S_i}(A_i(S))\right]| \leq \frac{2L^2\Gamma_K^d}{n}\sum_{t=0}^{T-1}\eta_t\left(\mathbf{W}_{i,i}^{T-1:t} + \frac{1}{m}\right),$$

*where $F_i(A_i(S)) = F_i(\mathbf{x}_i^T)$ and $F_{S_i}(A_i(S)) = F_{S_i}(\mathbf{x}_i^T)$ represent the local population risk and local empirical risk at each client, respectively.*

**Remark 8** *Theorem 6 provides the generalization error bounds for local models $\{\mathbf{x}_i^T\}_{i=1}^m$ learned by ZO-DSGD, which holds for decentralized systems with time-varying topology. Compared to the topology-independent bounds derived for global generalization error, the results in Theorem 6 capture the influence of communication topology on local generalization error $\mathbb{E}_{S,A}\left[F_i(A_i(S)) - F_{S_i}(A_i(S))\right]$. In the special case that the communication network is complete (fully connected) at each step, the topology-dependent term satisfies $\mathbf{W}_{i,i}^{T-1:t} = \frac{1}{m}$. The local generalization bounds can be simplified as $\frac{4L^2\Gamma_K^d}{mn}\sum_{t=0}^{T-1}\eta_t$ in this case, which matches the optimal rate for centralized SGD with convex loss function (Zhang et al. 2022).*

To further analyze the influence of communication topology on the average local generalization error of ZO-DSGD, we provide the following Corollary.

**Corollary 7** *Under the same conditions of Theorem 6, the average local generalization error is bounded by:*

$$\left|\mathbb{E}_{S,A}\left[\frac{1}{m}\sum_{i=1}^{m}\left(F_i(A_i(S)) - F_{S_i}(A_i(S))\right)\right]\right| \leq \frac{2L^2\Gamma_K^d}{mn}\sum_{t=0}^{T-1}\eta_t\left(\mathrm{Tr}(\mathbf{W}^{T-1:t})+1\right),$$

*where $\mathrm{Tr}(\mathbf{W}^{T-1:t})$ is the trace of matrix $\mathbf{W}^{T-1:t}$.*

**Remark 9** *In the setting of centralized federated learning, we have $\mathrm{Tr}(\mathbf{W}^{T-1:t}) = 1$. Corollary 7 has a worst-case upper bound $\mathcal{O}(\frac{T\Gamma_K^d}{n})$, since $\mathrm{Tr}(\mathbf{W}^{T-1:t}) \leq m$ by the property of doubly stochastic matrix. If the topology at step $T-1$ is fully connected, i.e., $\mathbf{W}^{T-1} = \frac{1}{m}\cdot\mathbf{1}_m\mathbf{1}_m^\top$, we obtain $\mathrm{Tr}(\mathbf{W}^{T-1:t}) = 1$ based on the double stochastic matrix property. In this case, the bound improves to $\mathcal{O}(\frac{T\Gamma_K^d}{mn})$.*

**Theorem 8 (Strongly Convex Case)** *Assume that the loss function $f(\cdot, z)$ is strongly convex, L-Lipschitz (Assumption 1) and $\beta$-smooth (Assumption 2). Let $\mathbf{W}^{T-1:t}$ be the matrix product defined in definition 4 and $A_i(S) = \mathbf{x}_i^T$ be the local model at $i$-th client after $T$ iterations of ZO-DSGD with mixing matrix $\mathbf{W}^t$ satisfying Assumption 4. Let $t_0 = \frac{2\beta(K+d-1)}{\mu K}$, apply $\eta_t = \frac{2}{\mu t}$ for $t \geq t_0$, and $\eta_t \leq \frac{K}{\beta(K+d-1)}$ for $t < t_0$. With $\epsilon \leq \frac{L\Gamma_k^d}{\beta m n d^{3/2}}$, for any $i \in [m]$, the local generalization error is bounded by:*

$$\left|\mathbb{E}_{S,A}\left[F_i(A_i(S)) - F_{S_i}(A_i(S))\right]\right| \leq \frac{6L^2\Gamma_K^d}{\mu n T}\sum_{t=t_0}^{T-1}\left(\mathbf{W}_{i,i}^{T-1:t} + \frac{1}{m}\right),$$

*where $F_i(A_i(S)) = F_i(\mathbf{x}_i^T)$ and $F_{S_i}(A_i(S)) = F_{S_i}(\mathbf{x}_i^T)$ represent the local population risk and local empirical risk at each client, respectively.*

**Remark 10** *Theorem 8 provides the generalization error bounds for the local model $\mathbf{x}_i^T$ learned by $i$-th client. Compared with the consensus model, the generalization error of local models is directly related to the average local model stability. It establishes the local generalization bounds of ZO-DSGD for strongly convex loss functions with a monotonically non-increasing learning rate. For the special case of centralized ZO-SGD, the topology-dependent term satisfies $\mathbf{W}_{i,i}^{T-1:t} = \frac{1}{m}$, and the corresponding bound will be simplified as $\frac{12L^2\Gamma_K^d}{\mu mn}$.*

**Corollary 9** *Under the same conditions of Theorem 8, the average local generalization error is bounded by:*

$$\left|\mathbb{E}_{S,A}\left[\frac{1}{m}\sum_{i=1}^{m}\left(F_i(A_i(S)) - F_{S_i}(A_i(S))\right)\right]\right| \leq \frac{6L^2\Gamma_K^d}{\mu mn}\sum_{t=t_0}^{T-1}\left(\frac{\mathrm{Tr}(\mathbf{W}^{T-1:t})+1}{T}\right).$$

**Remark 11** *The results in corollary 9 can extend to the extreme setting where each client operates local SGD steps without any inter-client communication. In this case, Corollary 9 has a worst-case upper bound $\mathcal{O}(\frac{\Gamma_K^d}{\mu n})$, since $\mathrm{Tr}(\mathbf{W}^{T-1:t}) = m$. Similar to Corollary 7, as long as the communication matrix at the last step is fully connected, i.e., $\mathbf{W}^{T-1} = \frac{1}{m}\cdot\mathbf{1}_m\mathbf{1}_m^\top$, the bound improves to $\mathcal{O}(\frac{\Gamma_K^d}{\mu mn})$.*

In the non-convex case, we have the following results.

**Theorem 10 (Non-convex case)** *Assume that the loss function $f(\cdot, z) \in [0,1]$ is L-Lipschitz (Assumption 1) and $\beta$-smooth (Assumption 2). Let $\mathbf{W}^{T-1:t}$ be the double product defined in definition 4 and $A_i(S) = \mathbf{x}_i^T$ be the local model at $i$-th client after $T$ iteration of ZO-DSGD with mixing matrix $\mathbf{W}^t$ satisfying Assumption 4. With monotonically non-increasing learning rate $\eta_t \leq \frac{C}{(t+1)\Gamma_K^d}$ and $\epsilon \leq \frac{L\Gamma_k^d}{\beta m n d^{3/2}}$, for any iteration index $t_0$, the local generalization error is bounded by:*

$$\left|\mathbb{E}_{S,A}\left[F_i(A_i(S)) - F_{S_i}(A_i(S))\right]\right| \leq \frac{t_0}{n} + \frac{2CL^2T^{\beta C}}{n}\sum_{t=t_0}^{T-1}\left(\frac{\mathbf{W}_{i,i}^{T-1:t} + \frac{1}{m}}{(t+1)^{C\beta+1}}\right),$$

*where $F_i(A_i(S)) = F_i(\mathbf{x}_i^T)$ and $F_{S_i}(A_i(S)) = F_{S_i}(\mathbf{x}_i^T)$ represent the local population risk and local empirical risk at each client, respectively.*

**Corollary 11** *Under the same conditions of Theorem 10, with monotonically non-increasing learning rate $\eta_t \leq \frac{C}{(\mathrm{Tr}(\mathbf{W}^{T-1:t})+1)(t+1)\Gamma_k^d}$ and $\epsilon \leq \frac{L\Gamma_k^d\,\mathrm{Tr}(\mathbf{W}^{T-1:t})}{\beta m n d^{3/2}}$, then it follows that*

$$\left|\mathbb{E}_{S,A}\left[\frac{1}{m}\sum_{i=1}^{m}\left(F_i(A_i(S)) - F_{S_i}(A_i(S))\right)\right]\right| \leq \mathcal{O}\left(\frac{T^{\frac{\beta C}{\beta C+1}}}{m^{\frac{1}{\beta C+1}}n}\right).$$

**Remark 12** *Theorem 10 and corollary 11 provide the generalization error bounds for local models learned by decentralized zeroth-order optimization methods. The generalization error bounds depend on the communication matrix product $\mathbf{W}^{T-1:t}$ defined in Definition 4.*

| Algorithms | Reference | Learning Rate | Assumption | Bound |
|---|---|---|---|---|
| SGD | Hardt, Recht, and Singer (2016) | $\eta_t \leq \frac{2}{\beta}$ | C | $\mathcal{O}\left(\frac{T}{n}\right)$ |
| SGD | Hardt, Recht, and Singer (2016) | $\eta_t \leq \frac{1}{\beta}$ | SC | $\mathcal{O}\left(\frac{1}{\mu m n}\right)$ |
| SGD | Hardt, Recht, and Singer (2016) | $\eta_t \leq \frac{C}{t+1}$ | NC | $\mathcal{O}\left(\frac{T^{\frac{\beta C}{\beta C+1}}}{n}\right)$ |
| DSGD | Richards and Rebeschini (2020) | $\eta_t \leq \frac{2}{\beta}$ | C | $\mathcal{O}\left(\frac{T}{mn}\right)$ |
| DSGD | Richards and Rebeschini (2020) | $\eta_t \leq \frac{1}{\beta}$ | SC | $\mathcal{O}\left(\frac{1}{\mu m n}\right)$ |
| DSGD | Le Bars et al. (2024) | $\eta_t \leq \frac{C}{t+1}$ | NC | $\mathcal{O}\left(\frac{T^{\frac{\beta C}{\beta C+1}}}{m^{\frac{1}{\beta C+1}} n}\right)$ |
| ZO-SGD | Liu et al. (2024) | $\eta_t \leq \frac{C}{(t+1)}$ | C | $\mathcal{O}\left(\frac{T}{n}\right)$ |
| ZO-SGD | Nikolakakis et al. (2022) | $\eta_t \leq \frac{C}{(t+1)\Gamma_K^d}$ | NC | $\mathcal{O}\left(\frac{T^{\frac{\beta C}{\beta C+1}}}{n}\right)$ |
| ZO-DSGD | Ours | $\eta_t \leq \frac{2K}{\beta(K+d-1)}$ | C | $\mathcal{O}\left(\frac{T}{mn}\right)$ |
| ZO-DSGD | Ours | $\eta_t \leq \frac{K}{\beta(K+d-1)}$ | SC | $\mathcal{O}\left(\frac{1}{\mu m n}\right)$ |
| ZO-DSGD | Ours | $\eta_t \leq \frac{C}{(t+1)\Gamma_K^d}$ | NC | $\mathcal{O}\left(\frac{T^{\frac{\beta C}{\beta C+1}}}{m^{\frac{1}{\beta C+1}} n}\right)$ |
| ZO-DSGD* | Ours | $\eta_t \leq \frac{2K}{\beta(K+d-1)}$ | C | $\mathcal{O}\left(\frac{\sum_{t=0}^{T-1} C_{\text{topo}}^t}{mn}\right)$ |
| ZO-DSGD* | Ours | $\eta_t = \min(\frac{2}{\mu t}, \frac{K}{\beta(K+d-1)})$ | SC | $\mathcal{O}\left(\frac{\sum_{t=0}^{T-1}(C_{\text{topo}}^t/T)}{\mu m n}\right)$ |
| ZO-DSGD* | Ours | $\eta_t \leq \frac{C}{(t+1)\Gamma_k^d C_{\text{topo}}^t}$ | NC | $\mathcal{O}\left(\frac{T^{\frac{\beta C}{\beta C+1}}}{m^{\frac{1}{\beta C+1}} n}\right)$ |

Table 1: Summary of the generalization error bounds for stochastic gradient algorithms and zeroth-order stochastic algorithms. Here, $T$ represents the total number of iterations, $m$ is the number of clients, $n$ denotes the sample size at each client, $C$ is a constant, $\beta$ is the smoothness constant, NC, C and SC are the abbreviations of nonconvex, convex, and strongly convex, respectively, and $C_{\text{topo}}^t = \text{Tr}(\mathbf{W}^{T-1:t}) + 1$ characterizes the properties of the decentralized topology, * represents the results for local models.

## Related Work

We list the most relevant works on SGD, DSGD, and ZO-SGD in Table 1. For the generalization analysis of ZO algorithms, Nikolakakis et al. (2022) and Liu et al. (2024) establish generalization bounds of order $\mathcal{O}(T^{\frac{\beta C}{\beta C+1}}/n)$ for ZO-SGD in the non-convex cases. However, there is still a gap between the generalization bounds of ZO-SGD and SGD in the convex and strongly convex cases (Hardt, Recht, and Singer 2016; Liu et al. 2024). Our generalization bounds on the consensus model bridge this gap. Our results indicate that in order to achieve the same convergence rate as the SGD algorithm, Zero-Order SGD requires a smaller learning rate $(K/(\beta(K+d-1))$ for ZO-DSGD compared to $1/\beta$ for SGD). For the generalization Analysis of DSGD, existing works mainly focus on the global performance of the average consensus model (Richards and Rebeschini 2020; Le Bars et al. 2024). We analyze not only the generalization error of the average model in ZO-DSGD but also the generalization error for local models. Our results show that the generalization bounds for the consensus model are topology-independent, and the generalization bounds for local models are topology-dependent, which reveals the influence of communication topology on the ZO-DSGD.

## Conclusion and Limitations

In this paper, we present a systematic analysis of the generalization error of ZO-DSGD, which covers strongly convex, convex, and non-convex cases. The generalization bounds for the global consensus model are derived based on the uniform stability tools. The obtained generalization bounds match that of centralized SGD and DSGD in the strongly convex and convex. To the best of our knowledge, this is the first work on generalization bounds of zeroth-order optimization algorithms that also recovers the results of SGD. Moreover, we propose a new average stability framework for the generalization analysis of individual local models and derive the generalization bounds reflecting the explicit influence of communication topology. Our results provide a theoretical foundation for understanding the generalization performance of decentralized zeroth-order optimization algorithms. The limitation of our work lies in the assumption of Lipschitz and smoothness. In practice, the Lipschitz and smoothness assumptions may not hold. In the future, we plan to extend our analysis to more general settings, such as non-Lipschitz or non-smooth optimization problems. Moreover, the extension to the minibatch setting is also an interesting direction for future work.

## Acknowledgments

This research was supported by Beijing Natural Science Foundation (No.4222029), National Natural Science Foundation of China (No.62476277, No.6207623), CCF-ALIMAMA TECH Kangaroo Fund (No.CCF-ALIMAMA OF 2024008), and Huawei-Renmin University joint program on Information Retrieval. We also acknowledge the support provided by the fund for building worldclass universities (disciplines) of Renmin University of China and by the funds from Beijing Key Laboratory of Big Data Management and Analysis Methods, Gaoling School of Artificial Intelligence, Renmin University of China, from Engineering Research Center of Next-Generation Intelligent Search and Recommendation, Ministry of Education, from Intelligent Social Governance Interdisciplinary Platform, Major Innovation and Planning Interdisciplinary Platform for the "DoubleFirst Class" Initiative, Renmin University of China, from Public Policy and Decision-making Research Lab of Renmin University of China, and from Public Computing Cloud, Renmin University of China.

## References

Allen-Zhu, Z.; and Hazan, E. 2016. Variance Reduction for Faster Non-Convex Optimization. In *International Conference on Machine Learning*, volume 48 of *JMLR Workshop and Conference Proceedings*, 699–707. JMLR.org.

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

Chen, A.; Zhang, Y.; Jia, J.; Diffenderfer, J.; Parasyris, K.; Liu, J.; Zhang, Y.; Zhang, Z.; Kailkhura, B.; and Liu, S. 2024. DeepZero: Scaling Up Zeroth-Order Optimization for Deep Model Training. In *The Twelfth International Conference on Learning Representations*.

Chen, J.; Chen, H.; Gu, B.; and Deng, H. 2023. Fine-grained theoretical analysis of federated zeroth-order optimization. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Deng, X.; Sun, T.; Li, S.; and Li, D. 2023. Stability-Based Generalization Analysis of the Asynchronous Decentralized SGD. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6): 7340–7348.

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(5): 2788–2806.

Elisseeff, A.; Evgeniou, T.; and Pontil, M. 2005. Stability of Randomized Learning Algorithms. *Journal of Machine Learning Research*, 6: 55–79.

Fang, W.; Yu, Z.; Jiang, Y.; Shi, Y.; Jones, C. N.; and Zhou, Y. 2022. Communication-Efficient Stochastic Zeroth-Order Optimization for Federated Learning. *IEEE Transactions on Signal Processing*, 70: 5058–5073.

Gu, B.; Wei, X.; Zhang, H.; Chang, Y.; and Huang, H. 2024. Obtaining Lower Query Complexities Through Lightweight Zeroth-Order Proximal Gradient Algorithms. *Neural Computation*, 36: 897–935.

Hardt, M.; Recht, B.; and Singer, Y. 2016. Train Faster, Generalize Better: Stability of Stochastic Gradient Descent. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, 1225–1234. JMLR.org.

Koloskova, A.; Lin, T.; Stich, S. U.; and Jaggi, M. 2020. Decentralized Deep Learning with Arbitrary Communication Compression. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Le Bars, B.; Bellet, A.; Tommasi, M.; Scaman, K.; and Neglia, G. 2024. Improved Stability and Generalization Guarantees of the Decentralized SGD Algorithm. In *ICML 2024-The Forty-first International Conference on Machine Learning*.

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

Lei, Y.; and Ying, Y. 2020. Fine-Grained Analysis of Stability and Generalization for Stochastic Gradient Descent. In *Proceedings of the 37th International Conference on Machine Learning*, 5809–5819. PMLR.

Li, X.; Yang, W.; Wang, S.; and Zhang, Z. 2019. Communication-Efficient Local Decentralized SGD Methods. *arXiv preprint arXiv:1910.09126*.

Li, Z.; Ying, B.; Liu, Z.; and Yang, H. 2024. Achieving Dimension-Free Communication in Federated Learning via Zeroth-Order Optimization. *arXiv preprint arXiv:2405.15861*.

Ling, Z.; Chen, D.; Yao, L.; Li, Y.; and Shen, Y. 2024. On the Convergence of Zeroth-Order Federated Tuning for Large Language Models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*, 1827–1838. ACM.

Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; and Neubig, G. 2023. Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Computing Surveys*, 55(9): 1–35.

Liu, X.; Zhang, H.; Gu, B.; and Chen, H. 2024. General Stability Analysis for Zeroth-Order Optimization Algorithms. In *The Twelfth International Conference on Learning Representations*.

Martínez Beltrán, E. T.; Pérez, M. Q.; Sánchez, P. M. S.; Bernal, S. L.; Bovet, G.; Pérez, M. G.; Pérez, G. M.; and Celdrán, A. H. 2023. Decentralized Federated Learning: Fundamentals, State of the Art, Frameworks, Trends, and Challenges. *IEEE Communications Surveys & Tutorials*, 25(4): 2983–3013.

Nesterov, Y.; and Spokoiny, V. 2017. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2): 527–566.

Nikolakakis, K.; Haddadpour, F.; Kalogerias, D.; and Karbasi, A. 2022. Black-Box Generalization: Stability of

Zeroth-Order Learning. *Advances in Neural Information Processing Systems*, 35: 31525–31541.

Qiu, Y.; Shanbhag, U.; and Yousefian, F. 2023. Zeroth-Order Methods for Nondifferentiable, Nonconvex, and Hierarchical Federated Optimization. *Advances in Neural Information Processing Systems*, 36: 3425–3438.

Reddi, S.; Charles, Z.; Zaheer, M.; Garrett, Z.; Rush, K.; Konečný, J.; Kumar, S.; and McMahan, H. B. 2020. Adaptive Federated Optimization. *arXiv preprint arXiv:2003.00295*.

Reddi, S. J.; Hefny, A.; Sra, S.; Poczos, B.; and Smola, A. 2016. Stochastic Variance Reduction for Nonconvex Optimization. In *International Conference on Machine Learning*, 314–323. PMLR.

Richards, D.; and Rebeschini, P. 2020. Graph-Dependent Implicit Regularisation for Distributed Stochastic Subgradient Descent. *Journal of Machine Learning Research*, 21(34): 1–44.

Shamir, O. 2017. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *Journal of Machine Learning Research*, 18(52): 1–11.

Shamir, O.; and Srebro, N. 2014. Distributed stochastic optimization and learning. In *2014 52nd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 850–857. IEEE.

Sun, T.; Li, D.; and Wang, B. 2021. Stability and Generalization of Decentralized Stochastic Gradient Descent. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(11): 9756–9764.

Tang, H.; Lian, X.; Yan, M.; Zhang, C.; and Liu, J. 2018. $D^2$: Decentralized Training over Decentralized Data. In Dy, J.; and Krause, A., eds., *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 4848–4856. PMLR.

Wang, J.; and Chen, H. 2024. Towards Stability and Generalization Bounds in Decentralized Minibatch Stochastic Gradient Descent. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(14): 15511–15519.

Zhang, Y.; Zhang, W.; Bald, S.; Pingali, V.; Chen, C.; and Goswami, M. 2022. Stability of sgd: Tightness analysis and improved bounds. In *Uncertainty in artificial intelligence*, 2364–2373. PMLR.

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

Zhong, H.; Deng, Z.; Su, W. J.; Wu, Z. S.; and Zhang, L. 2024. Provable multi-party reinforcement learning with diverse human feedback. *arXiv preprint arXiv:2403.05006*.

Zhu, T.; He, F.; Zhang, L.; Niu, Z.; Song, M.; and Tao, D. 2022. Topology-Aware Generalization of Decentralized SGD. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 27479–27503. PMLR.