Journal of Machine Learning Research 27 (2026) 1-52 Submitted 5/25; Revised 1/26; Published 3/26

Convergence of Decentralized Stochastic Subgradient-based Methods for Nonsmooth Nonconvex Optimization

Siyuan Zhang ZHANGSIYUAN@AMSS.AC.CN
Institute of Computational Mathematics and Scientific/Engineering Computing
Academy of Mathematics and Systems Science
Chinese Academy of Sciences
University of Chinese Academy of Sciences
Beijing, 100190, China

Nachuan Xiao XNCXY@CUHK.EDU.CN
School of Data Science
The Chinese University of Hong Kong, Shenzhen
Shenzhen, 518172, China
Guangdong Provincial Key Laboratory of Big Data Computing, CUHK-Shenzhen,
Shenzhen, 518172, China

Xin Liu LIUXIN@LSEC.CC.AC.CN
Institute of Computational Mathematics and Scientific/Engineering Computing
Academy of Mathematics and Systems Science
Chinese Academy of Sciences
University of Chinese Academy of Sciences
Beijing, 100190, China

Editor: Lam Nguyen

Abstract

In this paper, we focus on the decentralized stochastic subgradient-based methods in minimizing nonsmooth nonconvex functions without Clarke regularity, especially in the decentralized training of nonsmooth neural networks. We propose a general framework that unifies various decentralized subgradient-based methods, such as decentralized stochastic subgradient descent (DSGD), DSGD with gradient-tracking technique (DSGD-T), and DSGD with momentum (DSGD-M). To establish the convergence properties of our proposed framework, we relate the discrete iterates to the trajectories of a continuous-time differential inclusion, which is assumed to have a coercive Lyapunov function with a stable set $\mathcal{A}$. We prove the asymptotic convergence of the iterates to the stable set $\mathcal{A}$ with sufficiently small and diminishing step-sizes. These results provide first convergence guarantees for some well-recognized of decentralized stochastic subgradient-based methods without Clarke regularity of the objective function. Preliminary numerical experiments demonstrate that our proposed framework yields highly efficient decentralized stochastic subgradient-based methods with convergence guarantees in the training of nonsmooth neural networks.

Keywords: Nonsmooth optimization, decentralized stochastic subgradient-based method, random reshuffling, with-replacement sampling, conservative field, Lyapunov function.

1. Introduction

In this paper, we consider the following decentralized optimization problem (DOP) on an undirected, connected graph $\mathtt{G} = (\mathtt{V}, \mathtt{E})$ containing d agents,

$$\min_{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_d \in \mathbb{R}^n} \quad \sum_{i=1}^{d} f_i(\boldsymbol{x}_i), \qquad \text{s. t.} \quad \boldsymbol{x}_i = \boldsymbol{x}_j, \quad \forall (i,j) \in \mathtt{E}. \tag{DOP}$$

Here, for each $i \in [d] := \{1, 2, \ldots, d\}$, $\boldsymbol{x}_i$ refers to the local variable of agent i. Moreover, the set of nodes $\mathtt{V} = [d] := \{1, \ldots, d\}$ represents the set of agents, while the set of edges $\mathtt{E} \subseteq \mathtt{V} \times \mathtt{V}$ represents the set of communication links. An edge $(i,j) \in \mathtt{E}$ refers to that agents i and j are neighbors and can communicate with each other. Furthermore, for any $i \in [d]$, $f_i : \mathbb{R}^n \to \mathbb{R}$ is a locally Lipschitz continuous (possibly nonconvex and nonsmooth) cost function exclusively known to the agent i and takes the empirical-risk formulation (Shapiro et al., 2021). More precisely, for any $i \in [d]$, there exists $F_i : \mathbb{R}^n \times \Omega \to \mathbb{R}$ such that

$$f_i(\boldsymbol{x}) := \frac{1}{|\mathcal{S}_i|} \sum_{s_l \in \mathcal{S}_i} F_i(\boldsymbol{x}; \boldsymbol{s}_l), \tag{1.1}$$

where $\boldsymbol{s}_l$ is a random data vector supported on a local finite set $\mathcal{S}_i$. In real-world applications, the size of $\mathcal{S}_i$ is often sufficiently large, in the sense that computing the exact (sub)gradient of f_i is very costly, which makes stochastic (sub)gradient-based methods inevitable. Problem (DOP) has wide applications in machine learning (Jain et al., 2017; Bottou et al., 2018), signal processing (Mateos and Giannakis, 2012; Cohen et al., 2016) and control (Bolognani et al., 2014), etc.

1.1 Existing Works on Decentralized Stochastic Optimization

Motivated by stochastic gradient descent (SGD), the decentralized stochastic gradient descent (DSGD) (Nedic and Ozdaglar, 2009; Jiang et al., 2017; Lian et al., 2017) is one of the most fundamental methods for solving (DOP). In DSGD, each agent locally averages its model parameters with those of its neighbors and updates along the direction of the negative local stochastic gradient, which can be expressed as

$$\boldsymbol{x}_{i,k+1} = \sum_{j=1}^{d} \boldsymbol{W}(i,j) \boldsymbol{x}_{j,k} - \frac{\eta_k}{|\mathcal{B}_{i,k}|} \sum_{l \in \mathcal{B}_{i,k}} \nabla F_i(\boldsymbol{x}_{i,k}; \boldsymbol{s}_l), \tag{1.2}$$

where the subset $\mathcal{B}_{i,k} \subseteq \mathcal{S}_i$ refers to the mini-batch selection of the samples. Therefore, for any $i \in [d]$, the term $\frac{1}{|\mathcal{B}_{i,k}|} \sum_{l \in \mathcal{B}_{i,k}} \nabla F_i(\boldsymbol{x}_{i,k}; \boldsymbol{s}_l)$ refers to the stochastic gradient of f_i at $\boldsymbol{x}_{i,k}$, In addition, we define $\xi_{i,k+1} := \frac{1}{|\mathcal{B}_{i,k}|} \sum_{l \in \mathcal{B}_{i,k}} \nabla F_i(\boldsymbol{x}_{i,k}; \boldsymbol{s}_l) - \nabla f_i(\boldsymbol{x}_{i,k})$ as the corresponding evaluation noise. $\boldsymbol{W} \in \mathbb{R}^{d \times d}$ represents the mixing matrix corresponding to the graph $\mathtt{G} = (\mathtt{V}, \mathtt{E})$. That is, $\boldsymbol{W}$ is symmetric, doubly stochastic, and $\boldsymbol{W}(i,j) = 0$ whenever $(i,j) \notin \mathtt{E}$ (see Section 2.2 for details).

Based on DSGD, a wide range of variants have been developed for solving (DOP). Some works introduce auxiliary variables within DSGD to track the global stochastic gradient of the objective function, such as GNSD (Lu et al., 2019) and DSGT (Pu and Nedić, 2021). To accelerate convergence, some recent works (Yu et al., 2019; Gao and Huang, 2020) propose the decentralized momentum SGD by integrating the heavy-ball momentum (Polyak, 1964) into DSGD. Moreover, Gao et al. (2023) combines the stochastic gradient tracking method with momentum acceleration, Chen et al. (2023b) proposes a general framework for decentralized adaptive methods such as Adam, AdaGrad, and AMSGrad, and show that under suitable conditions, for nonconvex smooth objectives, if the original adaptive method converges, then its decentralized counterpart also converges. In addition, Huang et al. (2023a,b) introduce random reshuffling (RR) (De Sa, 2020) sampling techniques (that is, randomly samples an index i_k of $[N]$ without replacement at each step k in agent i, ensuring that each index appears exactly once per epoch.) into DSGD and DSGT, and provide theoretical evidence that RR outperforms with-replacement sampling. A series of studies (Lian et al., 2017; George et al., 2019; Yuan et al., 2021; Zhang and You, 2019; Xin et al., 2021b) have established theoretical guarantees for above decentralized SGD-type methods with continuously differentiable nonconvex objective functions. Recently, a line of works (Yan et al., 2023; Bylinkin and Beznosikov, 2024; Condat et al., 2024; Xu et al., 2023; Sun et al., 2024; Liu et al., 2025) have focused on developing distributed SGD-type methods that reduce communication and handle asynchrony and heterogeneity. For instance, Condat et al. (2024) couples local training with communication compression to obtain a communication-efficient distributed optimization method with theoretical guarantees with strongly convex objectives. Xu et al. (2023) analyzes asynchronous adaptive stochastic gradient methods, showing that AMSGrad-type schemes can be parallelized under bounded staleness while retaining convergence guarantees with smooth objective functions. From the perspective of federated learning, Sun et al. (2024) studies the role of server momentum and develop a more general server-momentum framework, which highlights how momentum can mitigate the adverse effects of heterogeneous data and system asynchrony with smooth objective functions.

Much attention has also been directed towards decentralized nonsmooth optimization. Some existing works, such as ProxDGD (Zeng and Yin, 2018), NEXT (Di Lorenzo and Scutari, 2016), and PG-EXTRA (Shi et al., 2015b) develop decentralized proximal gradient type methods for (DOP), under the assumption that f_i is composed of a differentiable term g_i and a convex regularization term r_i for all $i \in [d]$. In this context, the objective function is weakly convex and thus is Clarke regular. Several existing works (Di Lorenzo and Scardapane, 2019; Niu et al., 2021; Xin et al., 2021a; Xiao et al., 2023b; Zhou et al., 2025) extend the convergence analysis of Prox-DGD and NEXT from the deterministic case to the stochastic case, under additional assumptions regarding boundedness and heterogeneity. Additionally, Chen et al. (2021) proposes stoDPSM for ρ-weakly convex objectives and shows the global convergence while employing the Moreau envelope for measuring the stationarity. Koloskova et al. (2021) improves the convergence analysis for DSGT and proves its convergence for weakly convex objectives. From the view of distributed gradient flow, Swenson et al. (2022) establishes the global convergence of DSGD for Clarke regular function with decreasing consensus parameter. The detailed comparisons between the existing results on decentralized SGD-based methods in nonsmooth nonconvex optimization are listed in Table 1.

Despite achieving nice theoretical results, all the aforementioned works are only capable of handling nonsmooth but (Clarke) regular (Clarke, 1990) objective functions in

Method	Update scheme	Step-sizes	Conditions on f_i	Conditions on noises
stoDPSM Chen et al. (2021)	GD	Square summable, $\eta_{k+1}/\eta_k \to 1$	ρ-weakly convex, bounded subgradient	WRS, bounded second moment
S-NEXT Di Lorenzo and Scardapane (2019)	GT	Square summable	Proximal, weakly convex, bounded subgradient	WRS, bounded noise
ProxGT Xin et al. (2021a)	GT	Constant	Proximal, weakly convex, mean-squared smoothness	WRS, bounded variance
Prox-DASA Xiao et al. (2023b)	GD/GT	$\eta_k = \Theta(\frac{1}{\sqrt{K}}), k \in [K]$	Proximal, weakly convex	WRS, bounded variance
DEPOSITUM Zhou et al. (2025)	GT-M	$\eta_k = \Theta(\frac{1}{\sqrt{K}}), k \in [K]$	Proximal, weakly convex	WRS, bounded variance
DGF Swenson et al. (2022)	GD	$\eta_k = \Theta(k^{-p}), p \in (0,1]$	Clarke regular, weakly coercive, decreasing consensus parameter	WRS, bounded variance
ME-DOL Sahinoglu and Shahrampour (2024)	GD	$\eta_k = \Theta(\frac{1}{\sqrt{K}}), k \in [K]$	Global Lipschitz continuous	WRS, bounded variance, bounded second moment
Our work (DSM)	GD/GT /GD-M	$\eta_k = o(1/\log(k))$	Path-differentiable, Coercive	WRS, conditionally controlled noises
Our work (DSM)	GD/GT /GD-M	diminishing η_k	Path-differentiable, Coercive	RR

Table 1: A brief comparison of existing decentralized SGD-based methods for nonsmooth nonconvex optimization. Here, "GD", "GT" and "GD-M", "GT-M" are abbreviations of "gradient descent", "gradient tracking", "gradient descent with momentum technique" and "gradient tracking with momentum technique"; "WRS" and "RR" are abbreviations of "with-replacement sampling" and "random reshuffling"; The term "conditionally controlled noise" means that the evaluation noise is bounded conditional on $\mathcal{F}_k = \sigma(\{\boldsymbol{x}_{i,l}, l \leq k, i \in [d]\})$. The concept "path-differentiable" is defined in Definition 2.6.

(DOP), while some of these works evaluate the efficiency of their developed methods in the decentralized training of neural networks. However, nonsmooth activation functions, including ReLU and leaky ReLU, have served as popular building blocks for modeling neural networks, resulting in loss functions that are nonsmooth and lack Clarke regularity. To the best of our knowledge, Sahinoglu and Shahrampour (2024) is the only recent work that investigates the convergence of decentralized SGD-based method without requiring Clarke regularity. They propose a decentralized first-order and zero-order online algorithm called Me-DOL, but it requires several strong assumptions, such as global Lipschitz function and L^2 bounded stochastic subgradients. As a result, how to establish convergence properties for decentralized stochastic subgradient-based methods in nonsmooth, nonconvex settings remains an area for further exploration.

1.2 Existing Works on Stochastic Subgradient-based Methods for Nonsmooth Nonconvex Optimization

In the training of nonsmooth neural networks, one major challenge comes from computing the (stochastic) subgradients of the loss function. In a wide range of deep learning packages, such as TensorFlow, PyTorch, and JAX, automatic differentiation (AD) algorithms have become the default choice for computing subdifferential. Employing the chain rule, AD algorithms recursively construct the so-called "subgradients" through the composition of Jacobians of each block of the neural network, yet they neglect the underlying nonsmoothness. As the chain rule fails for loss functions lacking Clarke regularity, the so-called "subgradient" may not necessarily be contained in its Clarke subdifferential (see examples in Bolte and Pauwels (2020)). This limitation consequently renders the theoretical analyses of existing literature inapplicable to such functions.

To characterize the behavior of AD algorithms, Bolte and Pauwels (2021) introduces the concept of path-differentiable functions, which constitute a broad function class, including semi-algebraic functions, semi-analytic functions, and functions whose graphs are definable in some o-minimal structures (Davis et al., 2020). More importantly, the set of path-differentiable functions is general enough to cover almost all loss functions in neural network tasks, encompassing most of the loss functions without Clarke regularity. The study (Bolte and Pauwels, 2021) further demonstrates that when applied to path-differentiable loss functions, the outputs of AD algorithms are contained within a conservative field, which is a generalization of the Clarke subdifferential. Conservative field admits chain rules for path-differentiable functions, thereby preserving several good properties. For a comprehensive discussion of this concept, readers are directed to Section 2.3.

Based on the concepts of path-differentiable functions and conservative field, some recent works (Bolte and Pauwels, 2021; Castera et al., 2021; Pauwels, 2021; Xiao et al., 2024; Ruszczyński, 2020; Le, 2024; Xiao et al., 2023a; Bolte et al., 2022a) leverage the ordinary differential equation (ODE) approach (Benaïm et al., 2005; Benaïm, 2006; Borkar, 2009; Duchi and Ruan, 2018) to study the behavior of stochastic subgradient-based methods in the non-distributed setting. Although these studies have achieved significant progress in establishing convergence properties for stochastic subgradient-based methods, the theoretical results of many approaches cannot avoid infinitely many spurious critical points due to the gap between the conservative field and the Clarke subdifferential (Bianchi et al., 2022; Bolte et al., 2021). Moreover, extending existing results to the multi-agent scenario is non-trivial, as the update schemes of all the agents are coupled through local communication under decentralized settings. Given the convergence properties that need improvement and limited existing results in the multi-agent scenario, developing a unified framework for establishing convergence properties for decentralized stochastic subgradient-based methods is of great importance and worth exploring.

1.3 A General Framework for Decentralized Stochastic Subgradient-based Methods

We define the collection of the local variables, evaluation noises and noiseless gradients in (1.2) by matrix $\boldsymbol{X}_k$, Ξ_{k+1}, and $\boldsymbol{H}_k$, respectively:

$$\begin{aligned} \boldsymbol{X}_k &:= [\boldsymbol{x}_{1,k}, \boldsymbol{x}_{2,k}, \dots, \boldsymbol{x}_{d,k}], \\ \Xi_{k+1} &:= [\xi_{1,k+1}, \xi_{2,k+1}, \dots, \xi_{d,k+1}], \\ \boldsymbol{H}_k &:= [\nabla f_1(\boldsymbol{x}_{1,k}), \dots, \nabla f_d(\boldsymbol{x}_{d,k})]. \end{aligned} \tag{1.3}$$

Then for any $i \in [d]$, we denote $\Phi_i^{\mathrm{GD}} : \boldsymbol{x} \mapsto \{\nabla f_i(\boldsymbol{x})\}$ as a singleton-valued mapping from $\mathbb{R}^n$ to $2^{\mathbb{R}^n}$, which characterize the noiseless update directions in the i-th agent in the DSGD method. With these notations, the update scheme of DSGD in (1.2) can be reformulated as follows:

$$\begin{aligned} \boldsymbol{X}_{k+1} &= \boldsymbol{X}_k\boldsymbol{W} - \eta_k(\boldsymbol{H}_k + \Xi_{k+1}) \\ &\in \boldsymbol{X}_k\boldsymbol{W} - \eta_k\Big(\Big[\Phi_1^{\mathrm{GD}}(\boldsymbol{x}_{1,k}), \ldots, \Phi_d^{\mathrm{GD}}(\boldsymbol{x}_{d,k})\Big] + \Xi_{k+1}\Big). \end{aligned} \tag{1.4}$$

In light of the reformulated update scheme (1.4), we consider the following unified framework for decentralized stochastic subgradient-based methods,

$$\boldsymbol{Z}_{k+1} = \boldsymbol{Z}_k\boldsymbol{W} - \eta_k(\boldsymbol{H}_k + \Xi_{k+1}). \tag{DSM}$$

Here, $\boldsymbol{Z}_k = [\boldsymbol{z}_{1,k}, \ldots, \boldsymbol{z}_{d,k}] \in \mathbb{R}^{m\times d}$ denotes the collection of local variables (for example, model parameters, momentum, etc.), $\boldsymbol{H}_k \in \mathbb{R}^{m\times d}$ refers to the collection of noiseless update directions, $\Xi_{k+1} \in \mathbb{R}^{m\times d}$ refers to the collection of evaluation noises, and $\eta_k > 0$ denotes the step-size.

Moreover, the noiseless update directions in (DSM) is characterized by a family of locally bounded graph-closed set-valued mappings $\{\Phi_i : i \in [d]\}$ that maps $\mathbb{R}^m$ to the collections of subsets of $\mathbb{R}^m$. More precisely, there exists a sequence of nonnegative real numbers $\{\delta_k\}$ such that

$$\frac{1}{d}\boldsymbol{H}_k\boldsymbol{1}_d \in \mathrm{conv}\left(\frac{1}{d}\sum_{i=1}^{d}\Phi_i^{\delta_k}(\boldsymbol{z}_{i,k})\right), \quad \forall k \in \mathbb{N}. \tag{1.5}$$

Here, $\Phi_i^{\delta}(\boldsymbol{x}) := \{\boldsymbol{y} \in \mathbb{R}^n : \exists \boldsymbol{z}, \text{ s. t. } \|\boldsymbol{z} - \boldsymbol{x}\| < \delta, \inf_{\boldsymbol{w}\in\Phi_i(\boldsymbol{z})} \|\boldsymbol{y} - \boldsymbol{w}\| < \delta\}$. [1] Therefore, it is easy to verify that (DSM) encloses DSGD method in (1.2).

1.4 Contributions

The main contributions of our paper can be summarized as follows.

We propose a unified framework (DSM) that encompasses a variety of decentralized stochastic subgradient-based methods. We verify that decentralized stochastic subgradient descent (DSGD), DSGD with gradient-tracking technique (DSGD-T) and DSGD with momentum (DSGD-M) all fit into the framework (DSM). In particular, a decentralized variant of SignSGD (Bernstein et al., 2018), denoted as DSignSGD, is also enclosed by (DSM).

We establish the convergence guarantees for our framework (DSM). By introducing the following differential inclusion

$$\frac{\mathrm{d}\boldsymbol{z}}{\mathrm{d}t} \in -\mathrm{conv}\left(\frac{1}{d}\sum_{i=1}^{d}\Phi_i(\boldsymbol{z})\right), \tag{1.6}$$

that admits a coercive Lyapunov function ψ and stable set $\mathcal{A}$, we prove that, under mild conditions with controlled noises, the sequence $\{\boldsymbol{Z}_k\}$ generated by (DSM) is uniformly bounded. Then we prove that the iterates $\{\boldsymbol{z}_{i,k}\}, i \in [d]$ asymptotically reach consensus and approximate the trajectories of the differential inclusion (1.6), hence converging to the stable set $\mathcal{A}$ under sufficiently small and diminishing step-size.

Based on our theoretical results of the proposed framework (DSM), we prove that DSGD, DSGD-M, DSGD-T, and DSignSGD achieve global convergence in the random

1. The superscript is used to denote expansion only when applied to set-valued mappings.

reshuffling case and high-probability global convergence in the with-replacement sampling case. We also show that with random initialization, DSGD, DSGD-M and DSignSGD can avoid the spurious critical points in the sense of conservative field in minimizing path-differentiable functions. Although these algorithms are widely used in the training of nonsmooth neural networks, existing theoretical results are limited to Clarke regular objective functions. Our theoretical findings first fill the gap in solving Problem (DOP) with objectives lacking Clarke regularity.

Preliminary numerical experiments show the efficiency of analyzed methods and exhibit the superiority of our proposed method DSignSGD in comparison with DSGD and DSGD-M, hence demonstrating the flexibility and great potential of our framework.

1.5 Organizations

The rest of this paper is organized as follows. In Section 2, we present some basic notations and definitions. In Section 3, we stipulate some assumptions for our proposed framework (DSM), and establish convergence results under different noise settings by linking it with the dynamics of its corresponding differential inclusion. In Section 4, we demonstrate that framework (DSM) covers major existing decentralized stochastic subgradient-based methods, which inherit the convergence properties from (DSM) and avoid infinitely many spurious critical points. Numerical experiments regarding analyzed decentralized subgradient-based methods are presented in Section 5. In the last section, we draw conclusions and discuss possible future research directions.

2. Preliminaries

In this section, we introduce the necessary preliminaries for the subsequent analysis, including the notation system, the mixing matrix, key concepts from nonsmooth analysis, conservative fields, and differential inclusions.

2.1 Notations

We denote $\langle \cdot, \cdot \rangle$ as the standard inner product, while $\| \cdot \|$ as the ℓ_2-norm of a vector or spectral norm of a matrix. We refer to $\| \cdot \|_1$ as ℓ_1-norm of a vector and $\| \cdot \|_{\mathrm{F}}$ as Frobenius norm of a matrix. $\mathbb{B}(\boldsymbol{x}, \delta) := \left\{ \tilde{\boldsymbol{x}} \in \mathbb{R}^n : \|\tilde{\boldsymbol{x}} - \boldsymbol{x}\|^2 \leq \delta^2 \right\}$ refers to the ball centered at $\boldsymbol{x}$ with radius δ. For a given set $\mathcal{Y}, \mathrm{dist}(\boldsymbol{x}, \mathcal{Y})$ denotes the distance between $\boldsymbol{x}$ and a set $\mathcal{Y}$, that is, $\mathrm{dist}(\boldsymbol{x}, \mathcal{Y}) := \arg\min_{y \in \mathcal{Y}} \|\boldsymbol{x} - \boldsymbol{y}\|$, $\mathrm{conv}\, \mathcal{Y}$ denotes the convex hull of $\mathcal{Y}$, and $\mathcal{Y}^d$ denotes d-fold Cartesian product. The notation $\otimes$ stands for the Kronecker product. $\Delta_m := \{(\lambda_0, \ldots, \lambda_m) : \lambda_i \geq 0, \sum_{i=0}^m \lambda_i = 1\}$ stands for the simplex of dimension m.

For any positive sequence $\{\eta_k\}$, let $\lambda_\eta(0) := 0, \lambda_\eta(i) := \sum_{k=0}^{i-1} \eta_k$, and $\Lambda_\eta(t) := \sup \{k \in \mathbb{N} : t \geq \lambda_\eta(k)\}$. More explicitly, $\Lambda_\eta(t) = p$, if $\lambda_\eta(p) \leq t < \lambda_\eta(p+1)$. We define the set-valued mapping $\mathrm{sign} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ as follows:

$$(\mathrm{sign}(\boldsymbol{x}))_i = \begin{cases} \{-1\}, & \boldsymbol{x}_i < 0; \\ [-1, 1], & \boldsymbol{x}_i = 0; \\ \{1\}, & \boldsymbol{x}_i > 0. \end{cases}$$

It is easy to verify that $\text{sign}(\boldsymbol{x})$ is the Clarke subdifferential of the nonsmooth functions $\boldsymbol{x} \mapsto \|\boldsymbol{x}\|_1$.

Moreover, we denote $\mathcal{N}_i$ as the set of neighbor agents of agent i together with $\{i\}$, $\mathbb{R}_+$ as the set of all nonnegative real numbers. For any $N > 0, [N] := \{1, \ldots, N\}$. And let $\mathbf{1}_d \in \mathbb{R}^d$ represent a vector of all 1's, $\boldsymbol{e}_i \in \mathbb{R}^d$ represent $[0, \ldots, 1, \ldots, 0]^\top$, where 1 is the i-th component. For two integers i and j, $i \wedge j$ represents $\min\{i, j\}$.

We denote $(\Omega, \mathcal{F}, \mathbb{P})$ as the probability space. We use $\sigma(\boldsymbol{X})$ to denote the sigma-algebra generated by the random variable $\boldsymbol{X}$. We say $\{\mathcal{F}_k\}_{k\in\mathbb{N}}$ is a filtration if $\{\mathcal{F}_k\}$ is a collection of σ-algebras that satisfies $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}_\infty \subseteq \mathcal{F}$. A stochastic series $\{\xi_k\}$ is called a martingale with respect to a filtration $\{\mathcal{F}_k\}$, if $\{\xi_k\}$ is adapted to the filtration $\{\mathcal{F}_k\}$ and $\mathbb{E}[\xi_{k+1}|\mathcal{F}_k] = \xi_k$, $\forall k \in \mathbb{N}$. Moreover, A stochastic series $\{\xi_k\}$ is referred to as a martingale difference sequence, if $\{\xi_k\}$ is adapted to the filtration $\{\mathcal{F}_k\}$ and $\mathbb{E}[\xi_{k+1}|\mathcal{F}_k] = 0$ holds for all $k \in \mathbb{N}$.

In addition, we define the summation function f of (DOP) as

$$f(\boldsymbol{x}) := \frac{1}{d}\sum_{i=1}^{d} f_i(\boldsymbol{x}). \tag{2.1}$$

2.2 Mixing Matrix

Mixing matrix $\boldsymbol{W} = [\boldsymbol{W}(i,j)] \in \mathbb{R}^{d\times d}$ determines the structure of the communication network and plays an important role in averaging the local information of neighbor agents. Generally, we assume the mixing matrix $\boldsymbol{W}$ satisfies the following properties, which are standard in the literature (Xiao and Boyd, 2004; Sundhar Ram et al., 2010; Lian et al., 2017).

Definition 2.1 *Given a connected graph* $\mathtt{G} = (\mathtt{V}, \mathtt{E})$, *we say* $\boldsymbol{W} \in \mathbb{R}^{d\times d}$ *is a mixing matrix of* $\mathtt{G}$, *if it satisfies the following conditions.*

1. $\boldsymbol{W}$ *is symmetric.*

2. $\boldsymbol{W}$ *is doubly stochastic, namely,* $\boldsymbol{W}$ *is nonnegative and* $\boldsymbol{W}\mathbf{1}_d = \boldsymbol{W}^\top\mathbf{1}_d = \mathbf{1}_d$.

3. $\boldsymbol{W}(i,j) = 0$, *if and only if* $i \neq j$ *and* $(i,j) \notin \mathtt{E}$.

With a given graph $\mathtt{G}$, there are various approaches to choose its corresponding mixing matrix, such as Laplacian-based constant edge weight matrix (Xiao and Boyd, 2004) and Metropolis weight matrix (Xiao et al., 2006). We refer interested readers to Nedić et al. (2018); Shi et al. (2015a); Wang and Liu (2022); Wang et al. (2025) for more details about how to choose the mixing matrix.

Based on Pillai et al. (2005, Perron-Frobenius Theorem), we derive a direct corollary, which illustrates the fundamental spectral property of mixing matrix $\boldsymbol{W}$.

Corollary 2.2 *For any mixing matrix* $\boldsymbol{W} \in \mathbb{R}^{d\times d}$ *that corresponds to a connected graph* $\mathtt{G}$, *all the eigenvalues of* $\boldsymbol{W}$ *lie in* $(-1, 1]$, *and* $\boldsymbol{W}$ *has a single eigenvalue at* 1 *that admits* $\mathbf{1}_d$ *as its eigenvector.*

2.3 Nonsmooth Analysis and Conservative Field

A set-valued mapping $D : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is a mapping from $\mathbb{R}^n$ to set of subsets of $\mathbb{R}^n$. The graph of D is given by $\operatorname{graph} D = \{(\boldsymbol{x}, \boldsymbol{z}) \mid \boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{z} \in D(\boldsymbol{x})\}$. The set-valued mapping is said to be graph-closed (or has closed graph), if $\operatorname{graph} D$ is a closed subset of $\mathbb{R}^n \times \mathbb{R}^n$. D is said to be locally bounded, if for any $\boldsymbol{x} \in \mathbb{R}^n$, there exists a neighborhood $U_{\boldsymbol{x}}$ of $\boldsymbol{x}$ such that $\sup_{\boldsymbol{z} \in D(\boldsymbol{y}), \boldsymbol{y} \in U_{\boldsymbol{x}}} \|\boldsymbol{z}\| < +\infty$. In addition, D is convex-valued, if $D(\boldsymbol{x})$ is a convex subset of $\mathbb{R}^n$ for any $\boldsymbol{x} \in \mathbb{R}^n$.

Definition 2.3 *(Clarke, 1990) For any given locally Lipschitz continuous function $f : \mathbb{R}^n \to \mathbb{R}$, and any $\boldsymbol{x} \in \mathbb{R}^n$, the generalized directional derivative of f at $\boldsymbol{x}$ along the direction $\boldsymbol{d} \in \mathbb{R}^n$, is defined as*

$$f^\circ(\boldsymbol{x}; \boldsymbol{d}) := \limsup_{\boldsymbol{y} \to \boldsymbol{x}, t \downarrow 0} \frac{f(\boldsymbol{y} + t\boldsymbol{d}) - f(\boldsymbol{y})}{t}.$$

The Clarke subdifferential of f at $\boldsymbol{x} \in \mathbb{R}^n$, denoted by $\partial f(\boldsymbol{x})$, is defined as

$$\partial f(\boldsymbol{x}) := \{\boldsymbol{u} \in \mathbb{R}^n : f^\circ(\boldsymbol{x}; \boldsymbol{d}) \geq \langle \boldsymbol{u}, \boldsymbol{d} \rangle, \forall \boldsymbol{d} \in \mathbb{R}^n\}.$$

Notice that ∂f is a convex-valued graph-closed locally bounded set-valued mapping from $\mathbb{R}^n$ to the subsets of $\mathbb{R}^n$. Based upon the concept of generalized directional derivative, we present the definition of Clarke regular functions.

Definition 2.4 *(Clarke, 1990) We say that f is Clarke regular at $\boldsymbol{x} \in \mathbb{R}^n$, if for every direction $\boldsymbol{d} \in \mathbb{R}^n$, the one-sided directional derivative*

$$f^*(\boldsymbol{x}; \boldsymbol{d}) := \lim_{t \downarrow 0} \frac{f(\boldsymbol{x} + t\boldsymbol{d}) - f(\boldsymbol{x})}{t}$$

exists, and $f^*(\boldsymbol{x}; \boldsymbol{d}) = f^\circ(\boldsymbol{x}; \boldsymbol{d})$.

Next, we present a brief introduction to the conservative field, which characterizes the output of AD algorithms in differentiating nonsmooth neural networks.

Definition 2.5 (Conservative field) *Let $D : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be a nonempty set-valued mapping. We call D a conservative field if it is compact-valued and graph-closed, and for any absolutely continuous loop $\gamma : [0, 1] \to \mathbb{R}^n$ satisfying $\gamma(0) = \gamma(1)$, it holds that*

$$\int_0^1 \max_{\boldsymbol{v} \in D(\gamma(t))} \langle \dot{\gamma}(t), \boldsymbol{v} \rangle \mathrm{d}t = 0.$$

It is remarkable to note that any conservative field is locally bounded (Bolte and Pauwels, 2021, Remark 3). We now introduce the definition of the path-differentiable function corresponding to the conservative field.

Definition 2.6 *Let D be a conservative field, we say a function f is a potential for D or f is path-differentiable, provided that f can be formulated as*

$$f(\boldsymbol{x}) = f(\boldsymbol{x}_0) + \int_0^t \langle \dot{\gamma}(s), D(\gamma(s)) \rangle \mathrm{d}s,$$

for any absolutely continuous curve γ with $\gamma(0) = \boldsymbol{x}_0$ and $\gamma(t) = \boldsymbol{x}$. We also say that D is a conservative field for f, denoted as D_f.

As illustrated in Bolte and Pauwels (2021, Corollary 1), for any given path-differentiable function f, ∂f is a conservative field for f. Therefore, the concept of the conservative field can be regarded as an extension of the Clarke subdifferential. More importantly, when conservative field $D_f(\boldsymbol{x})$ is convex-valued, $\mathbf{0} \in \partial f(\boldsymbol{x})$ implies that $0 \in D_f(\boldsymbol{x})$.

In fact, the class of path-differentiable functions is general enough to cover the most objective functions in real-world problems. The well-known Clarke regular functions mentioned in Clarke (1990) and semi-algebraic functions mentioned in Łojasiewicz (1965) are all path-differentiable functions. In Davis et al. (2020); Bolte and Pauwels (2021), the authors identify definable functions (whose graphs are definable in an o-minimal structure) as an important class of path-differentiable functions. Actually, most activation functions and loss functions in deep neural networks are definable functions, including sigmoid, softplus, ReLU, leaky ReLU, hinge loss, etc.

Furthermore, it is important to highlight that definability is preserved under finite summation and composition (Davis et al., 2020). Consequently, for any neural network constructed by definable blocks, its loss function is definable, thus it is a path-differentiable function. Additionally, it is worth noting that the Clarke subdifferential of definable functions is itself definable (Bolte and Pauwels, 2021). Hence, in the case of a neural network constructed by definable blocks, the conservative field corresponding to the AD algorithms is also definable. The following proposition demonstrates that the definability of both f and its conservative field D_f leads to the nonsmooth Morse–Sard property (Bolte et al., 2007).

Proposition 2.7 (Theorem 5 in Bolte and Pauwels (2021)) *Let f be a path-differentiable function that admits D_f as its conservative field. Suppose both f and D_f are definable over $\mathbb{R}^n$, then the set $\left\{ f(\boldsymbol{x}) : \mathbf{0} \in D_f(\boldsymbol{x}) \right\}$ is finite.*

Finally, based on the concept of the conservative field, we present the definition of the critical points for the optimization problem (DOP).

Definition 2.8 *Let f in* (2.1) *be a path-differentiable function that admits D_f as its convex-valued conservative field, then we say $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ is a D_f-critical point of* (DOP)*, if $\boldsymbol{X} = \frac{1}{d}\boldsymbol{X}\mathbf{1}_d\mathbf{1}_d^\top$ and $\mathbf{0} \in D_f(\frac{1}{d}\boldsymbol{X}\mathbf{1}_d)$. Furthermore, we say $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ is a Clarke (or ∂f)-critical point of* (DOP)*, if $\boldsymbol{X} = \frac{1}{d}\boldsymbol{X}\mathbf{1}_d\mathbf{1}_d^\top$ and $\mathbf{0} \in \partial f(\frac{1}{d}\boldsymbol{X}\mathbf{1}_d)$.*

2.4 Differential Inclusion and Stochastic Subgradient Methods

In this part, we introduce fundamental concepts and theories associated with differential inclusion, which is crucial for establishing the convergence properties for stochastic subgradient methods.

Definition 2.9 *Let $D : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be a set-valued mapping. An absolutely continuous curve $\gamma : \mathbb{R}_+ \to \mathbb{R}^n$ is called a solution (or trajectory) to the differential inclusion*

$$\frac{d\boldsymbol{x}}{dt} \in D(\boldsymbol{x}), \tag{2.2}$$

(or a trajectory of D) with initial point $\boldsymbol{x}$, provided that $\gamma(0) = \boldsymbol{x}$ and $\gamma'(t) \in D(\gamma(t))$ holds for almost every $t \in \mathbb{R}_+$.

Definition 2.10 (Lyapunov function) *Let $\mathcal{B} \subset \mathbb{R}^n$ be a closed set. A continuous function $\psi : \mathbb{R}^n \to \mathbb{R}$ is referred to as a Lyapunov function for the differential inclusion* (2.2) *with the stable set $\mathcal{B}$, if for any solution γ to* (2.2) *and any $t > 0$, it holds that*

$$\psi(\gamma(t)) \leq \psi(\gamma(0)).$$

Moreover, whenever $\gamma(0) \notin \mathcal{B}$, it holds for any $t > 0$ that

$$\psi(\gamma(t)) < \psi(\gamma(0)).$$

Definition 2.11 *For any set-valued mapping $D : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ and any $\delta \geq 0$, we denote $D^\delta : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ as $D^\delta(\boldsymbol{x}) = \cup_{\boldsymbol{y} \in \mathbb{B}(\boldsymbol{x},\delta)}(D(\boldsymbol{y}) + \mathbb{B}(\mathbf{0}, \delta))$.*

Definition 2.12 (Perturbed solution) *An absolutely continuous curve $\gamma : \mathbb{R}_+ \to \mathbb{R}^n$ is said to be a perturbed solution of differential inclusion* (2.2)*, if the following two conditions hold:*

- *There exists a locally integrable function $w : \mathbb{R}_+ \to \mathbb{R}^n$, such that*

$$\lim_{t \to +\infty} \sup_{0 \leq t_0 \leq T} \left\| \int_t^{t+t_0} w(s) ds \right\| = 0$$

  holds for any $T > 0$.

- *For above locally integrable w, there exists $\delta : \mathbb{R}_+ \to \mathbb{R}$ such that $\lim_{t \to \infty} \delta(t) = 0$ and $\dot{\gamma}(t) - w(t) \in D^{\delta(t)}(\gamma(t))$.*

Now we consider an iterative sequence corresponding to the differential inclusion (2.2),

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \eta_k(D^{\delta_k}(\boldsymbol{x}_k) + \xi_{k+1}), \tag{2.3}$$

where $\{\eta_k\}$ is a non-summable positive sequence of step-sizes, $\{\delta_k\}$ is a nonnegative sequence, and ξ_{k+1} is a random noise when evaluating $D(\boldsymbol{x}_k)$. A continuous-time interpolated process $u : \mathbb{R}_+ \to \mathbb{R}^n$ induced by (2.3) is defined as

$$u(\lambda_\eta(k) + s) = \boldsymbol{x}_k + \frac{\boldsymbol{x}_{k+1} - \boldsymbol{x}_k}{\eta_k} s, \quad s \in [0, \eta_k).$$

Here, $\lambda_\eta(0) := 0$ and $\lambda_\eta(k) := \sum_{i=0}^{k-1} \eta_i, k \geq 1$.

In what follows, we summarize and present several lemmas, which respectively describe the conditions under which the interpolated process of the sequence $\{\boldsymbol{x}_k\}$ is a perturbed solution of (2.2), how the sequence $\{\boldsymbol{x}_k\}$ remains uniformly bounded, and a result on global convergence.

Lemma 2.13 (Extension of Proposition 1.3 in Benaïm et al. (2005)) *Let $D : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be a locally bounded, graph-closed and convex-valued set-valued mapping. Assume that the following hold:*

1. *For any $T > 0$,*

$$\lim_{s \to \infty} \sup_{s \leq i \leq \Lambda_\eta(\lambda_\eta(s)+T)} \left\| \sum_{k=s}^{i} \eta_k \xi_{k+1} \right\| = 0,$$

   where $\Lambda_\eta(t) := \sup\{k : \lambda_\eta(k) \leq t\}$.

2. $\{\delta_k\}$ *is diminishing, that is,* $\lim_{k\to\infty} \delta_k = 0$.

3. $\sup_{k\in\mathbb{N}} \|\boldsymbol{x}_k\| < +\infty$.

Then the interpolated process of sequence $\{\boldsymbol{x}_k\}$ *is a perturbed solution for* (2.3).

Before exhibiting the results on uniform boundedness and convergence of (2.3), we provide an assumption about differential inclusion (2.2), which is commonly used in the analysis of ODE-based approaches (Benaïm, 2006; Borkar, 2009; Duchi and Ruan, 2018).

Assumption 2.14 1. *There exists a locally Lipschitz continuous Lyapunov function* $\psi : \mathbb{R}^n \to \mathbb{R}$ *for the differential inclusion* (2.2) *with stable set* $\mathcal{A}$.

2. *The set* $\{\psi(\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{A}\}$ *is a finite subset of* $\mathbb{R}$.

Lemma 2.15 (Theorem 3.6 in Xiao et al. (2023a)) *Let* $D : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ *be a locally bounded, graph-closed and convex-valued set-valued mapping, and* $\mathcal{X}_0$ *be any compact subset of* $\mathbb{R}^n$. *Suppose Assumption 2.14 holds and Lyapunov function* ψ *is coercive.*

Then for any given $r > \max\{0, 4\sup_{x\in\mathcal{X}_0\cup\mathcal{A}} \psi(x)\}$, *there exist* $\alpha > 0$, $T > 0$ *such that whenever* $\sup_{k\geq 0} \eta_k \leq \alpha$, $\sup_{k\geq 0} \delta_k \leq \alpha$, *and* $\sup_{s\leq i\leq \Lambda_\eta(\lambda_\eta(s)+T)} \|\sum_{k=s}^{i} \eta_k \xi_{k+1}\| \leq \alpha, \forall s \geq 0$, *the sequence* $\{x_k\}$ *generated by* (2.3) *with* $x_0 \in \mathcal{X}_0$ *is restricted in* $\mathcal{L}_r := \{\boldsymbol{x} \in \mathbb{R}^n : \psi(\boldsymbol{x}) \leq r\}$.

Lemma 2.16 (Summary of Theorem 3.6, Proposition 3.27 in Benaïm et al. (2005)) *Let* $D : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ *be a locally bounded, graph-closed, and convex-valued set-valued mapping. Suppose Assumption 2.14 holds, and* $\{\boldsymbol{x}_k\}$ *is generated by* (2.3) *with any* $\boldsymbol{x}_0 \in \mathbb{R}^n$. *If the interpolated process of* $\{\boldsymbol{x}_k\}$ *is a perturbed solution of* (2.2), *then any cluster point of* $\{\boldsymbol{x}_k\}$ *generated by* (2.3) *lies in* $\mathcal{A}$, *and the sequence* $\{\psi(\boldsymbol{x}_k)\}$ *converges to* $\psi(\boldsymbol{x}^*)$, *for some* $\boldsymbol{x}^* \in \mathcal{A}$.

Except for Lemma 2.16, similar results under slightly different conditions can be found in Borkar (2009); Davis et al. (2020); Duchi and Ruan (2018), while some recent works (Bianchi and Rios-Zertuche, 2021; Bolte et al., 2022a) focus on analyzing the convergence of (2.3) under relaxed conditions. Interested readers could refer to those works for details.

3. A General Framework for Decentralized Stochastic Subgradient-based Methods

In this section, we aim to establish the convergence properties of proposed framework (DSM) under two noise settings. To this end, it is crucial to ensure the uniform boundedness of iterates $\{\boldsymbol{Z}_k\}$. A large amount of literature (Castera et al., 2021; Ruszczyński, 2020; Davis et al., 2020; Le, 2024) regards this condition as a prior assumption when analyzing the convergence of SGD-type algorithms in a single-agent setting. Nevertheless, more recent works (Bianchi et al., 2019, 2022; Josz and Lai, 2024, 2023; Bolte et al., 2024; Xiao et al., 2023a) focus on rigorously establishing the uniform boundedness of the iterative sequence.

Section 3.1 introduces the fundamental definitions and assumptions used in our analysis, and then rigorously demonstrate the uniform boundedness and asymptotic convergence for our framework (DSM). In Sections 3.2 and 3.3, we further discuss about the convergence properties of (DSM) when the evaluation noise is introduced by random reshuffling and with-replacement sampling, respectively.

3.1 Basic Assumptions and Main Results

(1.5) shows that the average update direction $\{\frac{1}{d}\boldsymbol{H}_k\mathbf{1}_d\}$ approximates $\frac{1}{d}\sum_{i=1}^{d}\Phi_i(\boldsymbol{z}_{i,k})$. To constrain the behavior of $\{\Phi_i\}_{i=1}^{d}$, we impose a set of basic assumptions concerning both the framework in Section 1.3 and its associated continuous-time differential inclusion,

$$\frac{\mathrm{d}\boldsymbol{z}}{\mathrm{d}t} \in -\Phi(\boldsymbol{z}) := \mathrm{conv}\left(\frac{1}{d}\sum_{i=1}^{d}\Phi_i(\boldsymbol{z})\right). \tag{3.1}$$

Assumption 3.1 *1. The sequence of step-sizes $\{\eta_k\}$ is positive and satisfies $\sum_{k=0}^{\infty}\eta_k = +\infty$.*

2. *The sequences $\{\boldsymbol{H}_k\}$ and $\{\Xi_{k+1}\}$ are uniformly bounded in $\mathbb{R}^{m\times d}$, whenever $\{\boldsymbol{Z}_k\}$ is bounded in $\mathbb{R}^{m\times d}$.*

3. *There exists a locally Lipschitz continuous and coercive Lyapunov function $\psi : \mathbb{R}^m \to \mathbb{R}$ for the differential inclusion* (3.1) *with stable set $\mathcal{A}$.*

4. *The set $\{\psi(\boldsymbol{z}) : \boldsymbol{z} \in \mathcal{A}\}$ is a finite subset of $\mathbb{R}$.*

Assumption 3.1(1) allows for a flexible choice of the step-size, while Assumption 3.1(2) is reasonable and mild, which is commonly employed in various existing works (Bolte and Pauwels, 2021; Davis et al., 2020; Xiao et al., 2023a, 2024; Castera et al., 2021).

Assumption 3.1(3) clarifies the Lyapunov properties of a differential inclusion with respect to $-\Phi$, which is standard in the literature (Benaïm et al., 2005; Borkar and Borkar, 2008; Bian and Xue, 2009; Davis et al., 2020; Bolte and Pauwels, 2021; Josz and Lai, 2023) and can be satisfied by most appropriately selected set-valued mappings. Since each Φ_i is graph-closed and locally bounded, the convex-valued set-valued mapping Φ is also graph-closed and locally bounded. Then, the coercivity of ψ in Assumption 3.1(3) indicates the compactness of any level set $\mathcal{L}_r := \{x \in \mathbb{R}^n : \psi(x) \leq r\}$.

Assumption 3.1(4) is similar to the Weak Sard's condition in Davis et al. (2020), which stipulates that the set of values of ψ at points outside the stable set is dense in $\mathbb{R}$. As demonstrated in Bolte and Pauwels (2021), Assumption 3.1(4) holds when f is a definable function with the selection $\Phi := \partial f$ and $\psi := f$.

In the following, we present some basic notations and definitions throughout Section 3. For any $M > 0$, the stopping time τ_M is defined as

$$\tau_M := \inf\{k \in \mathbb{N} : \|\boldsymbol{Z}_k\| > M\},$$

and the upper bound of noisy update direction before stopping time is given by

$$\ell_M := \sup_{0\leq k\leq \tau_M} \|\boldsymbol{H}_k + \Xi_{k+1}\|,$$

which is finite followed by Assumption 3.1(2). The second largest singular value of mixing matrix $\boldsymbol{W}$ is

$$\lambda_2 := \left\|\boldsymbol{W}(\boldsymbol{I}_d - \frac{\mathbf{1}_d\mathbf{1}_d^\top}{d})\right\|.$$

Definition 3.2 *For any positive sequence $\{\eta_k\}$, we say $\{\eta_k\}$ is α_{ub}-upper-bounded or is upper-bounded by α_{ub}, if $\sup_{k\in\mathbb{N}} \eta_k \leq \alpha$.*

Definition 3.3 *For any sequence of vectors (or matrices) $\{\xi_k\}$, and given constants $T, \alpha_{ub} > 0$, we say $\{\xi_k\}$ is $(\alpha_{ub}, T, \{\eta_k\})$-controlled, if for any $s \geq 0$, we have*

$$\sup_{s\leq i\leq \Lambda_\eta(\lambda_\eta(s)+T)} \left\| \sum_{k=s}^{i} \eta_k \xi_{k+1} \right\| \leq \alpha_{ub}.$$

Moreover, we say $\{\xi_k\}$ is $(\alpha_{ub}, \alpha, T, \{\eta_k\})$-asymptotically controlled, if it is $(\alpha_{ub}, T, \{\eta_k\})$-controlled and

$$\limsup_{s\to\infty} \sup_{s\leq i\leq \Lambda_\eta(\lambda_\eta(s)+T)} \left\| \sum_{k=s}^{i} \eta_k \xi_{k+1} \right\| \leq \alpha.$$

With above tools in placement, we first present Proposition 3.4 to demonstrate the average iterative sequence of (DSM) before stopping time coincides with the update scheme (2.3).

Proposition 3.4 *Suppose Assumption 3.1 holds, and let $\mathcal{Z}_0$ be a compact subset of $\mathbb{R}^m$. The sequence $\{\boldsymbol{Z}_k\}$ is generated by* (DSM)*, and $\boldsymbol{Z}_0 = \boldsymbol{z}_0 \mathbf{1}_d^\top, \boldsymbol{z}_0 \in \mathcal{Z}_0$. Then the recursion relation of $\{\frac{1}{d}\boldsymbol{Z}_{k\wedge\tau_M}\mathbf{1}_d\}$ can be reformulated as*

$$\frac{1}{d}\boldsymbol{Z}_{(k+1)\wedge\tau_M}\mathbf{1}_d \in \frac{1}{d}\boldsymbol{Z}_{k\wedge\tau_M}\mathbf{1}_d - \eta_k[\Phi^{s_k}(\frac{1}{d}\boldsymbol{Z}_{k\wedge\tau_M}\mathbf{1}_d) + \frac{1}{d}\Xi_{k+1}\mathbf{1}_d]\mathbb{1}_{\tau_M>k}, \tag{3.2}$$

where $s_k := \delta_k + \|\boldsymbol{Z}_{\perp,k\wedge\tau_M}\| \leq \delta_k + \frac{\sup_{0\leq i\leq k\wedge\tau_M} \eta_i \ell_M}{1-\lambda_2}$.

Proof From (DSM), the update scheme for $\{\boldsymbol{Z}_{k\wedge\tau_M}\}$ can be expressed as

$$\boldsymbol{Z}_{(k+1)\wedge\tau_M} = \boldsymbol{Z}_{k\wedge\tau_M} + [\boldsymbol{Z}_{k\wedge\tau_M}(\boldsymbol{W} - \boldsymbol{I}_d) - \eta_k(\boldsymbol{H}_k + \Xi_{k+1})]\mathbb{1}_{\tau_M>k}, \tag{3.3}$$

where $\mathbb{1}_{\tau_M>k}$ is the indicator function for the event $\{\tau_M > k\}$. Alternatively, the update can be written as

$$\boldsymbol{Z}_{(k+1)\wedge\tau_M} = \boldsymbol{Z}_{k\wedge(\tau_M-1)}\boldsymbol{W} - \eta_{k\wedge(\tau_M-1)}(\boldsymbol{H}_{k\wedge(\tau_M-1)} + \Xi_{(k+1)\wedge\tau_M}). \tag{3.4}$$

Define the consensus projection matrix $\boldsymbol{P} := \frac{1}{d}\mathbf{1}_d\mathbf{1}_d^\top$ and the disagreement matrix $\boldsymbol{Z}_{\perp,k} := \boldsymbol{Z}_k(\boldsymbol{I}_d - \boldsymbol{P})$. Then, the orthogonal decomposition holds:

$$\boldsymbol{Z}_k = \boldsymbol{Z}_k\boldsymbol{P} + \boldsymbol{Z}_{\perp,k}.$$

Intuitively, $\boldsymbol{Z}_{\perp,k}$ measures the dissimilarity among all agents' local variables at k-th iteration.

According to (3.4), simple calculations yield that

$$\begin{aligned}\|\boldsymbol{Z}_{\perp,(k+1)\wedge\tau_M}\| &= \left\|(\boldsymbol{Z}_{k\wedge(\tau_M-1)}\boldsymbol{W} - \eta_{k\wedge(\tau_M-1)}(\boldsymbol{H}_{k\wedge(\tau_M-1)} + \Xi_{(k+1)\wedge\tau_M}))(\boldsymbol{I}_d - \boldsymbol{P})\right\| \\ &\leq \left\|\boldsymbol{Z}_{k\wedge(\tau_M-1)}\boldsymbol{W}(\boldsymbol{I}_d - \boldsymbol{P})\right\| + \left\|\eta_{k\wedge(\tau_M-1)}(\boldsymbol{H}_{k\wedge(\tau_M-1)} + \Xi_{(k+1)\wedge\tau_M})(\boldsymbol{I}_d - \boldsymbol{P})\right\| \\ &\leq \left\|\boldsymbol{Z}_{k\wedge(\tau_M-1)}(\boldsymbol{I}_d - \boldsymbol{P})\right\| \|\boldsymbol{W}(\boldsymbol{I}_d - \boldsymbol{P})\| + \left\|\eta_{k\wedge(\tau_M-1)}(\boldsymbol{H}_{k\wedge(\tau_M-1)} + \Xi_{(k+1)\wedge\tau_M})\right\| \\ &\leq \lambda_2 \left\|\boldsymbol{Z}_{\perp,k\wedge(\tau_M-1)}\right\| + \left\|\eta_{k\wedge(\tau_M-1)}(\boldsymbol{H}_{k\wedge(\tau_M-1)} + \Xi_{(k+1)\wedge\tau_M})\right\| \\ &\leq \sum_{i=0}^{k\wedge(\tau_M-1)} \lambda_2^{k\wedge(\tau_M-1)-i}\|\eta_i(\boldsymbol{H}_i + \Xi_{i+1})\| + \lambda_2^{(k+1)\wedge\tau_M}\|\boldsymbol{Z}_{\perp,0}\| \\ &\leq \frac{\sup_{0\leq i\leq k\wedge(\tau_M-1)} \eta_i \ell_M}{1-\lambda_2},\end{aligned} \tag{3.5}$$

where the second inequality follows from the fact that $(\boldsymbol{I}_d - \boldsymbol{P})\boldsymbol{W}(\boldsymbol{I}_d - \boldsymbol{P}) = \boldsymbol{W}(\boldsymbol{I}_d - \boldsymbol{P})$ and the last inequality is due to $\boldsymbol{Z}_{\perp,0} = \boldsymbol{Z}_0 - \frac{\boldsymbol{Z}_0\boldsymbol{1}_d\boldsymbol{1}_d^\top}{d} = 0$. As a result, we obtain that

$$\|\boldsymbol{Z}_{\perp,l\wedge\tau_M}\| \leq \frac{\sup_{0\leq i\leq k\wedge\tau_M} \eta_i \ell_M}{1-\lambda_2}, \text{ for any } l \leq k+1. \tag{3.6}$$

We right-multiply $\frac{\boldsymbol{1}_d}{d}$ to (3.3) to get:

$$\frac{1}{d}\boldsymbol{Z}_{(k+1)\wedge\tau_M}\boldsymbol{1}_d = \frac{1}{d}\boldsymbol{Z}_{k\wedge\tau_M}\boldsymbol{1}_d - \eta_k(\frac{1}{d}\boldsymbol{H}_k\boldsymbol{1}_d + \frac{1}{d}\Xi_{k+1}\boldsymbol{1}_d)\mathbb{1}_{\tau_M>k}. \tag{3.7}$$

Equation (1.5) reveals that $\frac{1}{d}\boldsymbol{H}_k\boldsymbol{1}_d \in \mathrm{conv}\left(\frac{1}{d}\sum_{i=1}^d \Phi_i^{\delta_k}(\boldsymbol{z}_{i,k})\right)$. According to Carathéodory's Theorem, it admits the representation

$$\frac{1}{d}\boldsymbol{H}_k\boldsymbol{1}_d = \sum_{j=0}^m c_j(\frac{1}{d}\sum_{i=1}^d \boldsymbol{y}_{i,k}^j), \quad \boldsymbol{y}_{i,k}^j \in \Phi_i^{\delta_k}(\boldsymbol{z}_{i,k}), c_j \in \Delta_m.$$

On the event $\{\tau_M > k\}$, combining with (3.6) yields that

$$\begin{aligned}&\mathrm{dist}\left(\frac{1}{d}\boldsymbol{H}_k\boldsymbol{1}_d, \mathrm{conv}\left(\frac{1}{d}\sum_{i=1}^d \Phi_i(\frac{1}{d}\boldsymbol{Z}_{k\wedge\tau_M}\boldsymbol{1}_d + \mathbb{B}(\boldsymbol{0}, \delta_k + \|\boldsymbol{Z}_{\perp,k\wedge\tau_M}\|))\right)\right) \\ \leq &\sum_{j=0}^m c_j \mathrm{dist}\left(\frac{1}{d}\sum_{i=1}^d \boldsymbol{y}_{i,k}^j, \frac{1}{d}\sum_{i=1}^d \Phi_i\left(\frac{1}{d}\boldsymbol{Z}_{k\wedge\tau_M}\boldsymbol{1}_d + \mathbb{B}(\boldsymbol{0}, \delta_k + \|\boldsymbol{Z}_{\perp,k\wedge\tau_M}\|)\right)\right) \\ \leq &\sum_{j=0}^m c_j \frac{1}{d}\sum_{i=1}^d \mathrm{dist}\left(\boldsymbol{y}_{i,k}^j, \Phi_i\left(\frac{1}{d}\boldsymbol{Z}_{k\wedge\tau_M}\boldsymbol{1}_d + \mathbb{B}(\boldsymbol{0}, \delta_k + \|\boldsymbol{Z}_{\perp,k\wedge\tau_M}\|)\right)\right) \\ \leq &\sum_{j=0}^m c_j \frac{1}{d}\sum_{i=1}^d \delta_k = \delta_k,\end{aligned}$$

where the first inequality is due to the Jensen's inequality and the second inequality is followed by the triangle inequality. That is to say,

$$\frac{1}{d}\boldsymbol{H}_k\mathbf{1}_d \in \Phi^{s_k}\left(\frac{1}{d}\boldsymbol{Z}_{k\wedge\tau_M}\mathbf{1}_d\right).$$

Plugging this relation into (3.7) completes the proof. ■

Remark 3.5 *It is worth mentioning that the value of λ_2 (that is, the second-largest eigenvalue of the mixing matrix $\boldsymbol{W}$) is a critical factor in quantifying the effect of network topology on the convergence of decentralized methods. As demonstrated in Boyd et al. (2004), λ_2 is widely regarded as a measurement for network connectivity. Proposition 3.4 establishes* (3.6)*, demonstrating that λ_2 governs the rate at which $\boldsymbol{Z}_{\perp,k}$ converges to zero. Consequently, the value of λ_2 determines how rapidly the iterates $\frac{1}{d}\boldsymbol{Z}_{(k+1)\wedge\tau_M}\mathbf{1}_d$ approach the trajectory of the differential inclusion* (3.1)*. For further analysis regarding the influence of network topology in smooth settings, we refer readers to some related works (Vogels et al., 2023; Song et al., 2022).*

Based on Proposition 3.4, Theorem 3.6 establishes that the sequence $\{\boldsymbol{Z}_k\}$ remains uniformly bounded, whenever step-size $\{\eta_k\}$, $\{\delta_k\}$ are upper-bounded by a sufficiently small $\alpha_{ub} > 0$, and $\{\Xi_{k+1}\}$ is $(\alpha_{ub}, T, \{\eta_k\})$-controlled for some $T > 0$.

Theorem 3.6 *Suppose Assumption 3.1 holds, and let $\mathcal{Z}_0$ be a compact subset of $\mathbb{R}^m$. Then for any $r > \max\{0, 4\sup_{x\in\mathcal{Z}_0\cup\mathcal{A}}\psi(x)\}$, there exist $\alpha_{ub} > 0, T > 0$ such that for any α_{ub}-upper-bounded sequences $\{\eta_k\}$ and $\{\delta_k\}$, and any $(\alpha_{ub}, T, \{\eta_k\})$-controlled sequence $\{\Xi_{k+1}\}$, the sequence $\{\boldsymbol{Z}_k\}$ generated by* (DSM) *with $\boldsymbol{Z}_0 = \boldsymbol{z}_0\mathbf{1}_d^\top, \boldsymbol{z}_0 \in \mathcal{Z}_0$ is restricted in $\mathcal{L}_{2r}^d$.*

Proof Recall from Proposition 3.4, we have

$$\frac{1}{d}\boldsymbol{Z}_{(k+1)\wedge\tau_M}\mathbf{1}_d \in \frac{1}{d}\boldsymbol{Z}_{k\wedge\tau_M}\mathbf{1}_d - \eta_k[\Phi^{s_k}(\frac{1}{d}\boldsymbol{Z}_{k\wedge\tau_M}\mathbf{1}_d) + \frac{1}{d}\Xi_{k+1}\mathbf{1}_d]\mathbb{1}_{\tau_M>k}.$$

For any $M > 0$, employing Lemma 2.15, we conclude that there exists $\alpha, T > 0$ such that for any α-upper-bounded sequences $\{\eta_k\}$ and $\{s_k\}$, and any $(\alpha, T, \{\eta_k\})$-controlled sequence $\{\frac{1}{d}\Xi_{k+1}\mathbf{1}_d\mathbb{1}_{\tau_M>k}\}$, the sequence $\{\frac{1}{d}\boldsymbol{Z}_{k\wedge\tau_M}\mathbf{1}_d\}$ is restricted in $\mathcal{L}_r$.

Let L_1 be the Lipschitz constant of ψ in $\mathcal{L}_r + \mathbb{B}(\mathbf{0}, \frac{\alpha\ell_M}{1-\lambda_2})$, and $\alpha_{ub} := \min\{\frac{\alpha}{2}, \frac{\alpha(1-\lambda_2)}{2\ell_M}, \frac{r(1-\lambda_2)}{L_1\ell_M}\}$. It can be readily verified that

$$\psi(\boldsymbol{z}_{i,k\wedge\tau_M}) - \psi(\frac{1}{d}\boldsymbol{Z}_{k\wedge\tau_M}\mathbf{1}_d) \leq L_1\frac{\sup_{0\leq i\leq k\wedge\tau_M}\eta_i\ell_M}{1-\lambda_2} \leq r,$$

for any $i \in [d]$ and any α_{ub}-upper-bounded $\{\eta_k\}$.

Therefore, for any $M > 0$, any α_{ub}-upper-bounded sequences $\{\eta_k\}$, $\{\delta_k\}$, and any $(\alpha_{ub}, T, \{\eta_k\})$-controlled sequence $\{\Xi_{k+1}\}$, it holds that $\{\eta_k\}$ and $\{s_k\}$ are α-upper-bounded, $\{\frac{1}{d}\Xi_{k+1}\mathbf{1}_d\mathbb{1}_{\tau_M>k}\}$ is $(\alpha, T, \{\eta_k\})$-controlled, thus $\{\boldsymbol{z}_{i,k\wedge\tau_M}\}$ is restricted in $\mathcal{L}_{2r}, \forall i \in [d]$. Taking $M > \sup\{\|\boldsymbol{Z}\| : \boldsymbol{Z} \in \mathcal{L}_{2r}^d\}$, from the definition of the stopping time τ_M, we can derive that $\tau_M = +\infty$, then $\boldsymbol{z}_{i,k} = \boldsymbol{z}_{i,k\wedge\tau_M} \in \mathcal{L}_{2r}, \forall i \in [d], k \geq 0$, which completes the proof. ■

The following lemma from Sundhar Ram et al. (2010, Lemma 3.1) concerns the limit property of a convolution-like scalar sequence.

Lemma 3.7 *Let* $\{\gamma_k\}$ *be a sequence of real numbers. If* $0 < a < 1$ *and* $\lim_{k\to\infty}\gamma_k = \gamma$, *then it holds that* $\lim_{k\to\infty}\sum_{\ell=0}^{k} a^{k-\ell}\gamma_\ell = \frac{\gamma}{1-a}$.

Combining Theorem 3.6 and Lemma 3.7, we demonstrate in Theorem 3.8 that the sequence $\{\boldsymbol{Z}_k\}$ asymptotically reaches consensus and converges to the stable set $\mathcal{A}$ with diminishing upper-bounded $\{\eta_k\}$, $\{\delta_k\}$ and $(\alpha_{ub}, 0, T, \{\eta_k\})$ asymptotically controlled noises.

Theorem 3.8 *Suppose Assumption 3.1 holds, and let* $\mathcal{Z}_0$ *be a compact subset of* $\mathbb{R}^m$, $\{\boldsymbol{Z}_k\}$ *be generated by* (DSM) *with* $\boldsymbol{Z}_0 = \boldsymbol{z}_0\boldsymbol{1}_d^\top$, $\boldsymbol{z}_0 \in \mathcal{Z}_0$. *Then there exist* $\alpha_{ub} > 0, T > 0$ *such that for any* α_{ub}*-upper-bounded diminishing sequences* $\{\eta_k\}$ *and* $\{\delta_k\}$, *and any* $(\alpha_{ub}, 0, T, \{\eta_k\})$*-asymptotically controlled sequence* $\{\Xi_{k+1}\}$, *it follows that*

$$\lim_{k\to\infty} \operatorname{dist}(\boldsymbol{Z}_k, \{\boldsymbol{Z} \in \mathbb{R}^{m\times d} : \boldsymbol{Z} = \boldsymbol{z}\boldsymbol{1}^\top, \boldsymbol{z} \in \mathcal{A}\}) = 0,$$

and $\psi(\boldsymbol{z}_{i,k})$ *converges for each* $i \in [d]$.

Proof From Theorem 3.6, there exists r, $\alpha_{\mathrm{ub}}, T > 0$, for any α_{ub}-upper-bounded $\{\eta_k\}$ and $\{\delta_k\}$, and any $(\alpha_{ub}, T, \{\eta_k\})$-controlled sequence $\{\Xi_{k+1}\}$, the sequence $\{\boldsymbol{Z}_k\} \subseteq \mathcal{L}_{2r}^d$. Taking sufficiently large M, (3.2) becomes

$$\frac{1}{d}\boldsymbol{Z}_{k+1}\boldsymbol{1}_d \in \frac{1}{d}\boldsymbol{Z}_k\boldsymbol{1}_d - \eta_k[\Phi^{s_k}(\frac{1}{d}\boldsymbol{Z}_k\boldsymbol{1}_d) + \frac{1}{d}\Xi_{k+1}\boldsymbol{1}_d],$$

where $s_k := \delta_k + \|\boldsymbol{Z}_{\perp,k}\|$. Since $\lim_{k\to\infty}\|\boldsymbol{Z}_{\perp,k}\| \leq \lim_{k\to\infty}\sum_{i=0}^{k-1}\lambda_2^{k-1-i}\|\eta_i(\boldsymbol{H}_i + \Xi_{i+1})\| = 0$, we have $\lim_{k\to\infty} s_k = 0$.

Moreover, $(\alpha_{ub}, 0, T, \{\eta_k\})$-asymptotically controlled sequence $\{\Xi_{k+1}\}$ implies that for any $T' > 0$,

$$\lim_{s\to\infty} \sup_{s\leq i\leq \Lambda_\eta(\lambda_\eta(s)+T')} \left\| \sum_{k=s}^{i} \eta_k \frac{1}{d}\Xi_{k+1}\boldsymbol{1}_d \right\| = 0,$$

then it follows that the interpolated process of $\{\frac{1}{d}\boldsymbol{Z}_k\boldsymbol{1}_d\}$ is a perturbed solution of (3.1).

By combining Lemma 2.16, we can infer that, any cluster point of $\{\frac{1}{d}\boldsymbol{Z}_k\boldsymbol{1}_d\}$ lies in $\mathcal{A}$ and the sequence $\{\psi(\frac{1}{d}\boldsymbol{Z}_k\boldsymbol{1}_d)\}$ converges. As $\lim_{k\to\infty}\|\boldsymbol{Z}_{\perp,k}\| = 0$, we conclude that, any cluster point of $\{\boldsymbol{z}_{i,k}\}$ lies in $\mathcal{A}$ and $\{\psi(\boldsymbol{z}_{i,k})\}$ converges, for any $i \in [d]$. This completes the proof. ■

3.2 Convergence with Random Reshuffling

Given a family of locally bounded, graph-closed set-valued mappings $\{\mathcal{U}_{i,j} : 1 \leq i \leq d, 0 \leq j \leq N-1\}$, we consider framework (DSM) with the following random reshuffled update scheme,

$$\frac{1}{d}\Xi_{k+1}\boldsymbol{1}_d \in \frac{1}{d}\sum_{i=1}^{d}\mathcal{U}_{i,i_k}^{\rho_k}(\boldsymbol{z}_{i,k}) - \frac{1}{d}\boldsymbol{H}_k\boldsymbol{1}_d. \tag{3.8}$$

One can see (3.8) as a detailed characterization of the relationship between the update direction and the iterates $\{z_{i,k}\}$, thereby serving as a replacement for the role of (1.5). Hereafter, we make some assumptions on (3.8) and step-sizes $\{\eta_k\}$ in (DSM).

Assumption 3.9 *1. For any $k_1, k_2 \in \mathbb{N}$, $\eta_{k_1} = \eta_{k_2}$ holds, if $\lfloor \frac{k_1}{N} \rfloor = \lfloor \frac{k_2}{N} \rfloor$.*

2. *The sequence of indexes $\{i_k\}$ is chosen from $\{0, \dots, N-1\}$ by reshuffling. That is, $\{i_k : lN \leq k < (l+1)N\} = \{0, \dots, N-1\}$ holds for any $l \in \mathbb{N}$.*
3. *The sequence $\{\rho_k\}$ is controlled by $\{\eta_i p(\boldsymbol{Z}_i)\}_{i \leq k}$, where $p : \mathbb{R}^{m \times d} \to \mathbb{R}$ is a locally bounded function. Moreover, $\lim_{k\to\infty} \rho_k = 0$, whenever $\lim_{k\to\infty} \eta_k = 0$ and $\{\boldsymbol{Z}_k\}$ is bounded.*
4. *For any $\boldsymbol{z} \in \mathbb{R}^m$, $\frac{1}{N}\sum_{j=0}^{N-1} \mathcal{U}_{i,j}(\boldsymbol{z}) \subseteq \Phi_i(\boldsymbol{z})$.*

Indeed, we can treat each interval $\{k : lN \leq k < (l+1)N\}$ as an epoch. Assumption 3.9(1) reveals that step-sizes $\{\eta_k\}$ remain unchanged within each epoch, which is a practical and mild condition. Assumption 3.9(2) illustrates that the sequence of indexes $\{i_k\}$ is drawn by reshuffling in each epoch. Assumption 3.9(3) shows how the noisy averaged update scheme asymptotically approximates the Minkowski sum of random reshuffled set-valued mappings. The statement $\{\rho_k\}$ is controlled by $\{\eta_i p(\boldsymbol{Z}_i)\}_{i\leq k}$ means that ρ_k is bounded by a linear combination of the components in $\{\eta_i p(\boldsymbol{Z}_i)\}_{i\leq k}$. In particular, ρ_k can be an exponential moving average of $\{\eta_i p(\boldsymbol{Z}_i)\}_{i\leq k}$, given by $\sum_{i=0}^{k} \lambda_2^{k-i}\eta_i p(\boldsymbol{Z}_i)$. In Assumption 3.9(4), we show the noisy update scheme at agent i is randomly reshuffled from N components of Φ_i, which implies that $\frac{1}{Nd}\sum_{i=1}^{d}\sum_{k=0}^{N-1} \mathcal{U}_{i,i_k}(\boldsymbol{z}) \subseteq \Phi(\boldsymbol{z})$, building a bridge between the sequence $\{\frac{1}{d}\boldsymbol{Z}_{kN}\boldsymbol{1}_d\}$ and the differential inclusion (3.1).

It will be straightforward to see from Lemma 3.10 that, in a certain sense, the averaged evaluation noise is "hidden" over each epoch.

Lemma 3.10 *Suppose Assumption 3.1 and 3.9 hold, let $\mathcal{Z}_0$ be any compact set of $\mathbb{R}^m$ and $\boldsymbol{Z}_0 = \boldsymbol{z}_0\boldsymbol{1}^\top, \boldsymbol{z}_0 \in \mathcal{Z}_0$. $\{\boldsymbol{Z}_k\}$ is generated by* (DSM) *and the evaluation noise is introduced by random reshuffling* (3.8). *Then for any $\hat{\varepsilon} \in (0,1)$ and $M > 0$, there exists a constant $\alpha > 0$ such that for any $k \geq 0$, any $\eta_{kN} \in (0, \alpha]$, we have*

$$\frac{1}{d}\boldsymbol{Z}_{(k+1)N\wedge\tau'_M}\boldsymbol{1}_d \in \frac{1}{d}\boldsymbol{Z}_{kN\wedge\tau'_M}\boldsymbol{1}_d - N\eta_{kN}\Phi^{\hat{\varepsilon}}(\frac{1}{d}\boldsymbol{Z}_{kN\wedge\tau'_M}\boldsymbol{1}_d)\mathbb{1}_{\tau'_M > kN}. \tag{3.9}$$

where $\tau'_M := \inf\{\lfloor \frac{k}{N} \rfloor N \in \mathbb{N} : \|\boldsymbol{Z}_k\| > M\}$.

Proof From Assumption 3.9(3), for any $\hat{\varepsilon} \in (0,1)$ and any $M > 0$, there exists $\alpha_0 > 0$ such that $\rho_k < \frac{\hat{\varepsilon}}{3}$, when $\{\eta_k\}$ is upper bounded by α_0, and $\|\boldsymbol{Z}_k\| \leq M + 1$. We now introduce an additional notation:

$$L_M := \sup\left\{\|\mathcal{U}_{i,j}(x)\| : \|x\| \leq M+1, i \in [d], 0 \leq j \leq N-1\right\},$$

and for any $\hat{\varepsilon} \in (0,1)$, define

$$\alpha := \min\left\{\alpha_0, \frac{1}{1 + (L_M + \frac{\hat{\varepsilon}}{3})}, \frac{\hat{\varepsilon}}{3N(1+L_M) + N\hat{\varepsilon}}, \frac{(1-\lambda_2)\hat{\varepsilon}}{3\ell_M}\right\}.$$

Then for any $\eta_{kN} \in (0, \alpha]$ and any $j \in \{0, \ldots, N-1\}$, it follows from (3.5) that

$$\begin{aligned} \|\boldsymbol{Z}_{\perp,(kN+j)\wedge\tau_M}\| &\leq \sum_{i=0}^{(kN+j-1)\wedge(\tau_M-1)} \lambda_2^{(kN+j-1)\wedge(\tau_M-1)-i} \|\eta_i(\boldsymbol{H}_i + \Xi_{i+1})\| + \lambda_2^{(kN+j)\wedge\tau_M} \|\boldsymbol{Z}_{\perp,0}\| \\ &\leq \frac{\sup_{0\leq i\leq(kN+j-1)\wedge(\tau_M-1)} \eta_i \ell_M}{1-\lambda_2} \leq \frac{\hat{\varepsilon}}{3}, \end{aligned} \tag{3.10}$$

Without loss of generality, assuming $\tau_M \geq kN$, the update scheme (3.8) implies that

$$\begin{aligned} \left\| \frac{1}{d}\boldsymbol{Z}_{(kN+j)\wedge\tau_M}\mathbf{1}_d - \frac{1}{d}\boldsymbol{Z}_{kN\wedge\tau_M}\mathbf{1}_d \right\| &\leq \sum_{l=0}^{j\wedge(\tau_M-kN-1)} \eta_{kN+l} \left\| \frac{1}{d}\boldsymbol{H}_{kN+l}\mathbf{1}_d + \frac{1}{d}\Xi_{kN+l+1}\mathbf{1}_d \right\| \\ &\leq \sum_{l=0}^{j\wedge(\tau_M-kN-1)} \eta_{kN+l}(L_M + \frac{\hat{\varepsilon}}{3}) \\ &\leq \frac{(j+1)\hat{\varepsilon}(L_M + \frac{\hat{\varepsilon}}{3})}{3N(1+L_M)+N\hat{\varepsilon}} \leq \frac{\hat{\varepsilon}}{3}. \end{aligned} \tag{3.11}$$

As a result, we derive that $\sup_{0\leq j\leq N-1} \left\| \boldsymbol{z}_{i,(kN+j)\wedge\tau_M} - \frac{1}{d}\boldsymbol{Z}_{kN\wedge\tau_M}\mathbf{1}_d \right\| \leq \frac{2}{3}\hat{\varepsilon}, \forall i \in [d]$.

Therefore, for any $k > 0$ and any $0 \leq l \leq N-1$, the following inclusion holds:

$$\frac{1}{d}\boldsymbol{Z}_{(kN+l+1)\wedge\tau_M}\mathbf{1}_d \in \frac{1}{d}\boldsymbol{Z}_{(kN+l)\wedge\tau_M}\mathbf{1}_d - \eta_{kN+l}\frac{1}{d}\sum_{i=1}^{d}\left[\mathcal{U}_{i,i_{kN+l}}^{\frac{\hat{\varepsilon}}{3}}\left(\frac{1}{d}\boldsymbol{Z}_{kN\wedge\tau_M}\mathbf{1}_d + \mathbb{B}(\mathbf{0}, \frac{2\hat{\varepsilon}}{3})\right)\right]\mathbb{1}_{\tau_M > kN+l}. \tag{3.12}$$

Moreover, the definition of τ_M' implies the following relations:

$$\tau_M' > kN \Leftrightarrow \tau_M > kN + l, \quad \forall 0 \leq l \leq N-1,$$

and

$$kN \wedge \tau_M' \leq (kN+l+1) \wedge \tau_M, \quad \forall 0 \leq l \leq N-1.$$

Recursively employing (3.12) gives us that

$$\begin{aligned} &\frac{1}{d}\boldsymbol{Z}_{(k+1)N\wedge\tau_M'}\mathbf{1}_d \\ &\in \frac{1}{d}\boldsymbol{Z}_{kN\wedge\tau_M'}\mathbf{1}_d - \sum_{l=kN}^{(k+1)N-1} \eta_{kN}\frac{1}{d}\sum_{i=1}^{d}\left[\mathcal{U}_{i,i_l}(\frac{1}{d}\boldsymbol{Z}_{kN\wedge\tau_M'}\mathbf{1}_d + \mathbb{B}(\mathbf{0},\hat{\varepsilon})) + \mathbb{B}(\mathbf{0}, \frac{\hat{\varepsilon}}{3})\right]\mathbb{1}_{\tau_M'>kN} \\ &\subseteq \frac{1}{d}\boldsymbol{Z}_{kN\wedge\tau_M'}\mathbf{1}_d - N\eta_{kN}\left(\frac{1}{d}\sum_{i=1}^{d}\frac{1}{N}\sum_{l=0}^{N-1}\mathcal{U}_{i,l}(\frac{1}{d}\boldsymbol{Z}_{kN\wedge\tau_M'}\mathbf{1}_d + \mathbb{B}(\mathbf{0},\hat{\varepsilon})) + \mathbb{B}(\mathbf{0}, \frac{\hat{\varepsilon}}{3})\right)\mathbb{1}_{\tau_M'>kN} \\ &\subseteq \frac{1}{d}\boldsymbol{Z}_{kN\wedge\tau_M'}\mathbf{1}_d - N\eta_{kN}\Phi^{\hat{\varepsilon}}\left(\frac{1}{d}\boldsymbol{Z}_{kN\wedge\tau_M'}\mathbf{1}_d\right)\mathbb{1}_{\tau_M'>kN}, \end{aligned}$$

which is our desired result. ■

Lemma 3.10 illustrates that $\{\frac{1}{d}\boldsymbol{Z}_{kN\wedge\tau'_M}\mathbf{1}_d\}$ can be viewed as an inexact discretization of differential inclusion (3.1). Then based on Lemma 2.15, we establish the uniform boundedness and asymptotic convergence of $\{\boldsymbol{Z}_k\}$ when adopting (3.8) under Assumption 3.9. As the evaluation noise in (3.9) vanishes over each epoch and is thus asymptotically controlled, Theorems 3.11 and 3.12 can be viewed as special cases of Theorems 3.6 and 3.8.

Theorem 3.11 *Let $\mathcal{Z}_0$ be any compact set of $\mathbb{R}^m$. Suppose Assumption 3.1 and 3.9 hold, and $\boldsymbol{Z}_0 = \boldsymbol{z}_0\mathbf{1}_d^\top$, where $\boldsymbol{z}_0 \in \mathcal{Z}_0$. Then for any $r > \max\{0, 4\sup_{\boldsymbol{z}\in\mathcal{Z}_0\cup\mathcal{A}}\psi(\boldsymbol{z})\}$, there exist $\alpha > 0$, such that for any α-upper-bounded $\{\eta_k\}$, the sequence $\{\boldsymbol{Z}_k\}$ generated by* (DSM) *with* (3.8) *is restricted in $\mathcal{L}_{2r}^d$.*

Proof For any $r > \max\{0, 4\sup_{\boldsymbol{z}\in\mathcal{Z}_0\cup\mathcal{A}}\psi(\boldsymbol{z})\}$, we set L_2 to be the Lipschitz constant of ψ in $\mathcal{L}_r + \mathbb{B}(0,r)$. According to Lemma 2.15, there exists $\alpha_1 \in (0, \frac{3r}{2L_2})$ such that for any α_1-upper bounded $\{\hat{\eta}_k\}$ and $\{\hat{\delta}_k\}$, the sequence generated by

$$\hat{\boldsymbol{x}}_{k+1} \in \hat{\boldsymbol{x}}_k - \hat{\eta}_k\Phi^{\hat{\delta}_k}(\hat{\boldsymbol{x}}_k),$$

satisfies $\{\hat{\boldsymbol{x}}_k\} \subseteq \mathcal{L}_r$. Given the iterative sequence $\{\frac{1}{d}\boldsymbol{Z}_{kN\wedge\tau'_M}\mathbf{1}_d\}$, Lemma 3.10 guarantees that for any $\delta_{kN} \in (0, \alpha_1)$, there exists $\alpha \in (0, \frac{\alpha_1}{N})$ such that whenever $\{\eta_{kN}\}$ is α-upper bounded, the following inclusion holds:

$$\frac{1}{d}\boldsymbol{Z}_{(k+1)N\wedge\tau'_M}\mathbf{1}_d \in \frac{1}{d}\boldsymbol{Z}_{kN\wedge\tau'_M}\mathbf{1}_d - N\eta_{kN}\Phi^{\delta_{kN}}(\frac{1}{d}\boldsymbol{Z}_{kN\wedge\tau'_M}\mathbf{1}_d)\mathbb{1}_{\tau'_M>kN}.$$

Hence, by Lemma 2.15, the sequence $\{\frac{1}{d}\boldsymbol{Z}_{kN\wedge\tau'_M}\mathbf{1}_d\} \subseteq \mathcal{L}_r$.

Without loss of generality, we assume $\tau_M > kN$. Following arguments analogous to those in (3.10)-(3.11), we deduce that

$$\begin{aligned}
&\left\|\boldsymbol{z}_{i,(kN+j)\wedge\tau_M} - \frac{1}{d}\boldsymbol{Z}_{kN\wedge\tau'_M}\mathbf{1}_d\right\| \\
\leq &\left\|\boldsymbol{z}_{i,(kN+j)\wedge\tau_M} - \frac{1}{d}\boldsymbol{Z}_{(kN+j)\wedge\tau_M}\mathbf{1}_d\right\| + \left\|\frac{1}{d}\boldsymbol{Z}_{(kN+j)\wedge\tau_M}\mathbf{1}_d - \frac{1}{d}\boldsymbol{Z}_{kN}\mathbf{1}_d\right\| \\
< &\frac{2\alpha_1}{3} < \frac{r}{L_2}.
\end{aligned}$$

Consequently,

$$\psi(\boldsymbol{z}_{i,(kN+j)\wedge\tau_M}) \leq \psi(\frac{1}{d}\boldsymbol{Z}_{kN\wedge\tau'_M}\mathbf{1}_d) + L_2\left\|\boldsymbol{z}_{i,(kN+j)\wedge\tau_M} - \frac{1}{d}\boldsymbol{Z}_{kN\wedge\tau'_M}\mathbf{1}_d\right\| \leq 2r, \tag{3.13}$$

for any $i \in [d], j \in \{0, \ldots, N-1\}$.

To finish the proof, it suffices to demonstrate that $\tau_M = +\infty$, for any $M > \sup\{\|\boldsymbol{Z}\| : \boldsymbol{Z} \in \mathcal{L}_{2r}^d\}$. If not, then there exists $k_1 \in \mathbb{N}$ with $k_1N \leq \tau_M < (k_1+1)N$. By (3.13), this implies $\boldsymbol{Z}_{\tau_M} \in \mathcal{L}_{2r}^d$. However, since $M > \sup\{\|\boldsymbol{Z}\| : \boldsymbol{Z} \in \mathcal{L}_{2r}^d\}$, the definition of τ_M forces $\|\boldsymbol{Z}_{\tau_M}\| > M$, contradicting $\boldsymbol{Z}_{\tau_M} \in \mathcal{L}_{2r}^d$.

■

Theorem 3.12 *Suppose Assumption 3.1 and 3.9 hold. Let $\mathcal{Z}_0$ be any compact set of $\mathbb{R}^m$, $\{\boldsymbol{Z}_k\}$ be generated by* (DSM) *with* (3.8)*, and $\boldsymbol{Z}_0 = \boldsymbol{z}_0\mathbf{1}_d^\top, \boldsymbol{z}_0 \in \mathcal{Z}_0$. Then there exists $\alpha_{rr} > 0$, such that for any α_{rr}-upper-bounded diminishing $\{\eta_k\}$ in* (DSM)*, it follows that*

$$\lim_{k\to\infty} \operatorname{dist}(\boldsymbol{Z}_k, \{\boldsymbol{Z} \in \mathbb{R}^{m\times d} : \boldsymbol{Z} = \boldsymbol{z}\mathbf{1}^\top, \boldsymbol{z} \in \mathcal{A}\}) = 0,$$

and $\psi(\boldsymbol{z}_{i,k})$ converges for each $i \in [d]$.

Proof Recall from Theorem 3.11 that there exists $\alpha > 0$, if $\{\boldsymbol{Z}_k\}$ iterates by (DSM) and (3.8) with α-upper-bounded sequence $\{\eta_{kN}\}$ and initial point $\boldsymbol{Z}_0 = \boldsymbol{z}_0\mathbf{1}_d^\top, \boldsymbol{z}_0 \in \mathcal{Z}_0$, then $\{\boldsymbol{Z}_k\}$ is restricted in $\mathcal{L}_{2r}^d$ for some $r > 0$. Moreover, by Assumption 3.9(3), we can further select a $\alpha_{\mathrm{rr}} \in (0, \alpha)$ such that $\rho_k \leq 1$ for α_{rr}-upper-bounded sequence $\{\eta_{kN}\}$.

We now analyze the case when $\{\eta_{kN}\}$ is upper bounded by α_{rr}. With a slight abuse of notation, let $L := \sup\left\{\left\|\mathcal{U}_{i,j}(x)\right\| : x \in \mathcal{L}_{2r} + \mathbb{B}(\mathbf{0},1), i \in [d], 0 \leq j \leq N-1\right\}$. Then for any $i \in [d]$ and $j \in \{0, \dots, N-1\}$, we have

$$\begin{aligned}\left\|\boldsymbol{z}_{i,kN+j} - \frac{1}{d}\boldsymbol{Z}_{kN}\mathbf{1}_d\right\| &\leq \left\|\frac{1}{d}\boldsymbol{Z}_{kN+j}\mathbf{1}_d - \frac{1}{d}\boldsymbol{Z}_{kN}\mathbf{1}_d\right\| + \left\|\boldsymbol{Z}_{\perp,kN+j}\right\| \\ &\leq \sum_{l=0}^{j} \eta_{kN+l}\left\|\frac{1}{d}\boldsymbol{H}_{kN+l}\mathbf{1}_d + \frac{1}{d}\Xi_{kN+l+1}\mathbf{1}_d\right\| + \left\|\boldsymbol{Z}_{\perp,kN+j}\right\| \\ &\leq \sum_{l=0}^{j} \eta_{kN+l}(L+1) + \left\|\boldsymbol{Z}_{\perp,kN+j}\right\| \\ &\leq N(L+1)\eta_{kN} + \max_{0\leq j\leq N-1}\left\{\left\|\boldsymbol{Z}_{\perp,kN+j}\right\|\right\}.\end{aligned} \tag{3.14}$$

According to Assumption 3.1(2), $\|\boldsymbol{H}_k + \Xi_{k+1}\|$ is uniformly bounded as $\{\boldsymbol{Z}_k\}$ is restricted in $\mathcal{L}_{2r}^d$. Plugging Lemma 3.7, we have

$$\lim_{k\to\infty} \|\boldsymbol{Z}_{\perp,kN+j}\| \leq \lim_{k\to\infty} \sum_{i=0}^{kN+j} \lambda_2^{kN+j-i}\|\eta_i(\boldsymbol{H}_i + \Xi_{i+1})\| = 0. \tag{3.15}$$

Define $\hat{\varepsilon}_{kN} := N(L+1)\eta_{kN} + \max_{0\leq j\leq N-1}\{\|\boldsymbol{Z}_{\perp,kN+j}\|\} + \max_{0\leq j\leq N-1}\{\rho_{kN+j}\}$. Similar to the arguments in Lemma 3.10, we can achieve that

$$\frac{1}{d}\boldsymbol{Z}_{(k+1)N}\mathbf{1}_d \in \frac{1}{d}\boldsymbol{Z}_{kN}\mathbf{1}_d - N\eta_{kN}\Phi^{\hat{\varepsilon}_{kN}}\left(\frac{1}{d}\boldsymbol{Z}_{kN}\mathbf{1}_d\right). \tag{3.16}$$

Since $\{\eta_{kN}\}$ is diminishing, Assumption 3.9(3) indicates that $\hat{\varepsilon}_{kN} \to 0$ as $k \to \infty$.

Indeed, the update scheme (3.16) corresponds to an inexact subgradient descent method corresponding for the differential inclusion (3.1). Applying Lemma 2.13, we obtain that the interpolated process of $\{\frac{1}{d}\boldsymbol{Z}_{kN}\mathbf{1}_d\}$ is a perturbed solution of (3.1). Then Lemma 2.16 illustrates that

$$\lim_{k\to\infty} \operatorname{dist}(\frac{1}{d}\boldsymbol{Z}_{kN}\mathbf{1}_d, \mathcal{A}) = 0, \textit{ and } \psi(\frac{1}{d}\sum_{i=1}^{d} \boldsymbol{z}_{i,kN}) \textit{ converges},$$

and (3.14) further implies that

$$\lim_{k\to\infty} \operatorname{dist}(\boldsymbol{z}_{i,kN+j}, \mathcal{A}) = 0, \text{ and } \psi(\boldsymbol{z}_{i,kN+j}) \text{ converges, for any } 0 \leq j \leq N-1. \tag{3.17}$$

Combining (3.15) and (3.17) leads to the desired result. ■

3.3 Convergence under With-replacement Sampling

In this part, we handle the circumstance that the evaluation noise $\{\Xi_{k+1}\}$ arises from with-replacement sampling. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the probability space. We consider the following expression of $\{\Xi_{k+1}\}$:

$$\Xi_{k+1} = \chi(\boldsymbol{Z}_k, \xi_k) = [\chi_1(\boldsymbol{z}_{1,k}, \xi_{1,k}), \ldots, \chi_d(\boldsymbol{z}_{d,k}, \xi_{d,k})], \tag{3.18}$$

where each $\chi_i : \mathbb{R}^m \times \Omega \to \mathbb{R}^m$ is a Borel function, $\{\xi_{i,k}\}_{k\geq 0}$ is a sequence of sampling data picked in Ω at agent i. With-replacement sampling means that the sampling data $\{\xi_{i,k}\}_{k\geq 0}$ are randomly drawn from the same probability distribution $\mathcal{P}_i$, which allows $\xi_{i,k}$ to occur multiple times.

We stipulate the following assumptions on framework (DSM) with (3.18).

Assumption 3.13 *1. For any $\boldsymbol{Z} \in \mathbb{R}^{m\times d}$, $\mathbb{E}_\xi[\chi(\boldsymbol{Z}, \xi)] = 0$. Moreover, there exists a locally bounded function $q : \mathbb{R}^{m\times d} \to \mathbb{R}_+$, such that*

$$\|\chi(\boldsymbol{Z}, \xi)\| \leq q(\boldsymbol{Z}), \quad \text{for almost every } \xi \in \Omega^d.$$

2. The sequence of step-sizes $\{\eta_k\}$ satisfies

$$\sum_{k=0}^{\infty} \eta_k = +\infty, \quad \lim_{k\to\infty} \eta_k \log(k) = 0.$$

Recall from Proposition 3.4, we have

$$\frac{1}{d}\boldsymbol{Z}_{(k+1)\wedge\tau_M}\mathbf{1}_d \in \frac{1}{d}\boldsymbol{Z}_{k\wedge\tau_M}\mathbf{1}_d - \eta_k[\Phi^{s_k}(\frac{1}{d}\boldsymbol{Z}_{k\wedge\tau_M}\mathbf{1}_d) + \frac{1}{d}\Xi_{k+1}\mathbf{1}_d]\mathbb{1}_{\tau_M>k}. \tag{3.19}$$

Under Assumption 3.13(1), the sequence $\{\frac{1}{d}\Xi_{k+1}\mathbf{1}_d\mathbb{1}_{\tau_M>k}\}$ is a uniformly bounded martingale difference sequence with respect to $\mathcal{F}_k := \sigma(\{\boldsymbol{Z}_i, i \leq k\})$. With the $o(1/\log(k))$ step-sizes $\{\eta_k\}$ and any given $\iota > 0$, the following lemma illustrates that sequence $\{\frac{1}{d}\Xi_{k+1}\mathbf{1}_d\mathbb{1}_{\tau_M>k}\}$ is (ι, T, η_k)-controlled with arbitrary high probability.

Lemma 3.14 *Suppose Assumption 3.1 and 3.13 hold. Then for any $\varepsilon, T, \iota > 0$, there exists $\beta_0 > 0$, such that for any β_0-upper-bounded sequence $\{\eta_k\}$, it holds that*

$$\mathbb{P}\left(\left\{\exists s \geq 0, \text{ such that } \sup_{s\leq i\leq \Lambda_\eta(\lambda_\eta(s)+T)} \left\|\sum_{k=s}^{i} \eta_k \frac{1}{d}\Xi_{k+1}\mathbf{1}_d\mathbb{1}_{\tau_M>k}\right\| \geq \iota\right\}\right) \leq \varepsilon.$$

Lemma 3.14 is a direct corollary of Benaïm (2006, Proposition 4.2) and Xiao et al. (2023a, Proposition A.5), hence we omit its proof for simplicity. Based on this lemma, we present Theorem 3.15 and 3.16 to show the uniform boundedness and convergence properties of framework (DSM) together with evaluation noise (3.18).

Theorem 3.15 *Suppose Assumption 3.1 and 3.13 hold, and let $\mathcal{Z}_0$ be any compact set of $\mathbb{R}^m$, $\boldsymbol{Z}_0 = \boldsymbol{z}_0\mathbf{1}_d^\top, \boldsymbol{z}_0 \in \mathcal{Z}_0$. The sequence $\{\boldsymbol{Z}_k\}$ is generated by* (DSM)*, and the evaluation noise is introduced by with-replacement sampling* (3.18)*. Then for any $r > \max\{0, 4\sup_{x\in\mathcal{Z}_0\cup\mathcal{A}} \psi(x)\}$ and any $\varepsilon \in (0,1)$, there exists $\beta > 0$, such that for any β-upper-bounded sequence $\{\eta_k\}$ and $\{\delta_k\}$, $\{\boldsymbol{Z}_k\}$ is restricted in $\mathcal{L}_{2r}^d$ with probability at least $1-\varepsilon$.*

Let Ω_0 denote the event that the sequence $\{\frac{1}{d}\Xi_{k+1}\mathbf{1}_d\mathbb{1}_{\tau_M>k}\}$ is (ι, T, η_k)-controlled. Lemma 3.14 gives us that

$$\mathbb{P}(\Omega_0) \geq 1-\varepsilon.$$

Applying Theorem 3.6 over event Ω_0 directly yields Theorem 3.15. Hence, we omit the proof for clarity.

Theorem 3.16 *Suppose Assumption 3.1 and 3.13 hold, $\{\delta_k\}$ in* (1.5) *satisfies $\lim_{k\to\infty}\delta_k = 0$. Let $\mathcal{Z}_0$ be any compact set of $\mathbb{R}^m$, and $\boldsymbol{Z}_0 = \boldsymbol{z}_0\mathbf{1}_d^\top, \boldsymbol{z}_0 \in \mathcal{Z}_0$. Then for any $\varepsilon \in (0,1)$, there exists $\beta > 0$, such that for any $\{\eta_k\}, \{\delta_k\}$ upper-bounded by β, and any $\{\boldsymbol{Z}_k\}$ generated by* (DSM) *with evaluation noise* (3.18)*, we have*

$$\mathbb{P}\left(\lim_{k\to\infty} \mathrm{dist}(\boldsymbol{Z}_k, \{\boldsymbol{Z}\in\mathbb{R}^{m\times d} : \boldsymbol{Z} = \boldsymbol{z}\mathbf{1}^\top, \boldsymbol{z}\in\mathcal{A}\}) = 0\right) \geq 1-\varepsilon,$$

and $\mathbb{P}(\psi(\boldsymbol{z}_{i,k})$ converges$) \geq 1-\varepsilon$ for any $i\in[d]$.

Proof Denote the event $\Omega_1 := \{\{\boldsymbol{Z}_k\} \subseteq \mathcal{L}_{2r}^d\}$. By Theorem 3.6 and 3.15, we have

$$\Omega_1 \supseteq \Omega_0, \quad \mathbb{P}(\Omega_1) \geq \mathbb{P}(\Omega_0) \geq 1-\varepsilon.$$

Notice that $\{\frac{1}{d}\Xi_{k+1}\mathbf{1}_d\}$ is a uniformly bounded martingale difference sequence over Ω_1 and $\eta_k \sim o(1/\log(k))$. Invoking Benaïm et al. (2005, Proposition 1.4, Remark 1.5), we obtain that

$$\lim_{s\to\infty} \sup_{s\leq i\leq \Lambda_\eta(\lambda_\eta(s)+T)} \left\|\sum_{k=s}^{i} \eta_k \frac{1}{d}\Xi_{k+1}\mathbf{1}_d\right\| = 0$$

holds almost surely over both Ω_1 and Ω_0, for any $T > 0$.

Employing Theorem 3.8 over Ω_0, we completes the proof.

■

4. Convergence Properties of Decentralized Stochastic Subgradient-based Methods

In this section, we demonstrate the flexibility of our proposed framework (DSM) by showing that it encloses a wide range of popular decentralized stochastic subgradient-based methods, including DSGD, decentralized generalized SGD with momentum (DGSGD-M), and DSGD with gradient-tracking technique (DSGD-T). More importantly, our theoretical analysis provides convergence guarantees for these methods in the minimization of nonsmooth path-differentiable functions without Clarke regularity. Additionally, Theorem 4.5, 4.8 establish that DSGD and DGSGD-M can avoid spurious critical points.

To facilitate analysis, we need the following assumptions on problem (DOP).

Assumption 4.1 1. *For each $i \in [d]$ and $\boldsymbol{s}_l \in \mathcal{S}_i$, $F_i(\cdot; \boldsymbol{s}_l)$ is a definable function, which admits a definable conservative field $D_{F_i(\cdot;\boldsymbol{s}_l)}$.*

2. *The objective function f is coercive.*

As discussed in Section 2.3, the class of definable functions is wide enough to enclose the loss functions of common nonsmooth neural networks. Moreover, Bolte and Pauwels (2021) demonstrates that there exists a definable conservative field for any definable loss function constructed by definable blocks, within which the results yielded by Automatic Differentiation (AD) algorithms are contained. Henceforth, Assumption 4.1(1) is mild and reasonable in practical scenarios.

For conciseness, we denote $D_{F_{i,l}} := D_{F_i(\cdot;\boldsymbol{s}_l)}$. According to Bolte and Pauwels (2021, Corollary 4) and Assumption 4.1(1),

$$D_{f_i}(\boldsymbol{x}) := \frac{1}{|\mathcal{S}_i|} \sum_{l=1}^{|\mathcal{S}_i|} D_{F_{i,l}}(\boldsymbol{x}). \tag{4.1}$$

is a definable conservative field for f_i. Recall that $f(\boldsymbol{x}) = \frac{1}{d}\sum_{i=1}^{d} f_i(\boldsymbol{x})$ defined in (2.1), we can choose its definable conservative field D_f as

$$D_f(\boldsymbol{x}) := \mathrm{conv}\left(\frac{1}{d}\sum_{i=1}^{d} D_{f_i}(\boldsymbol{x})\right). \tag{4.2}$$

4.1 Decentralized Stochastic Subgradient Descent

In this subsection, we establish the global convergence for decentralized stochastic subgradient descent (DSGD) by showing that it fits into our proposed framework (DSM). The detailed algorithm of DSGD is presented in Algorithm 1. Intuitively, for each $i \in [d]$, the consensus term $\sum_{j\in\mathcal{N}_i} \boldsymbol{W}(i,j)\boldsymbol{x}_{j,k}$ enforces $\boldsymbol{x}_{i,k}$ towards the mean over all the agents (that is, $\frac{1}{d}\sum_{i\in[d]} \boldsymbol{x}_{i,k}$). Moreover, as $\boldsymbol{d}_{i,k}$ refers to the (mini-batch) stochastic subgradient of f_i at $\boldsymbol{x}_{i,k}$, the term $-\eta_k \boldsymbol{d}_{i,k}$ can be viewed as a single SGD step to minimize f_i in the i-th agent.

In Algorithm 1, the evaluation noise is introduced by the selection of mini-batch $\mathcal{B}_{i,k}$. Now, we discuss two different scenarios for selecting mini-batches and step-sizes.

Assumption 4.2 (Random reshuffling)

Algorithm 1 DSGD for solving (DOP).

Input: Initial point $\boldsymbol{x}_0 \in \mathbb{R}^n$ and a mixing matrix $\boldsymbol{W}$.

for all $i \in [d]$ in parallel **do**
  Set $k \leftarrow 0$. Initialize $\boldsymbol{x}_{i,k} = \boldsymbol{x}_0$;
  while not terminated **do**
    Randomly select a mini-batch $\mathcal{B}_{i,k} \subseteq \mathcal{S}_i$;
    Compute $\boldsymbol{d}_{i,k} \in \frac{1}{|\mathcal{B}_{i,k}|} \sum_{\boldsymbol{s}_l \in \mathcal{B}_{i,k}} D_{F_{i,l}}(\boldsymbol{x}_{i,k})$;
    Choose the step-size η_k;
    Communicate and update the local variables

$$\boldsymbol{x}_{i,k+1} = \sum_{j \in \mathcal{N}_i} \boldsymbol{W}(i,j)\boldsymbol{x}_{j,k} - \eta_k \boldsymbol{d}_{i,k};$$

    Set $k \leftarrow k+1$;
  end while
end for
return $\boldsymbol{X}_k := [\boldsymbol{x}_{1,k}, \ldots, \boldsymbol{x}_{d,k}]$.

1. *$\mathcal{B}_{i,k}$ is selected from set $\mathcal{S}_i$ by random reshuffling, meaning that for every epoch index $l \in \mathbb{N}$, the following holds:*

$$\bigcup_{k=lN}^{(l+1)N-1} \mathcal{B}_{i,k} = \mathcal{S}_i.$$

2. *Let $\{\hat{\eta}_k\}$ be a prefixed sequence such that $\sum_{k=0}^{\infty} \hat{\eta}_k = +\infty$, and $\lim_{k\to\infty} \hat{\eta}_k = 0$. Moreover, for any $j_1, j_2 \in \mathbb{N}$, $\hat{\eta}_{j_1} = \hat{\eta}_{j_2}$ holds, if $\lfloor \frac{j_1}{N} \rfloor = \lfloor \frac{j_2}{N} \rfloor$. We set $\eta_k = c\hat{\eta}_k$, $c > 0$ for all $k \in \mathbb{N}$.*

Assumption 4.3 (With-replacement sampling)

1. *$\mathcal{B}_{i,k}$ is selected from set $\mathcal{S}_i$ by with-replacement sampling for each $k \in \mathbb{N}$.*

2. *Let $\{\hat{\eta}_k\}$ be a prefixed sequence satisfying*

$$\sum_{i=0}^{\infty} \hat{\eta}_k = +\infty, \text{ and } \lim_{k\to\infty} \hat{\eta}_k \log(k) = 0.$$

   We set $\eta_k = c\hat{\eta}_k$, $c > 0$ for all $k \in \mathbb{N}$.

Assumption 4.2(1) and 4.3(1) employ different sampling techniques, leading to different forms of evaluation noise. In subsequent analysis, we will demonstrate that they conform to (3.8) and (3.18), respectively.

Assumption 4.2(2) and 4.3(2) both assume that $\{\eta_k\}$ is composed of a scaling parameter c and a fixed diminishing sequence of step-sizes (at most in the rate of $o(1/\log(k))$), hence enabling great flexibility in choosing the step-sizes $\{\eta_k\}$ in practical scenarios.

In the following theorem, we establish the global convergence of Algorithm 1 by showing that Algorithm 1 fits into our proposed framework with $\Phi = D_f$, $\psi = f$ and $\mathcal{A} = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{0} \in D_f(\boldsymbol{x})\}$.

Theorem 4.4 *Suppose Assumption 4.1 holds. Let $\{\boldsymbol{X}_k\}$ be the sequence generated by Algorithm 1.*

(1) When Assumption 4.2 holds, there exists $\alpha_c > 0$, for any $c \in (0, \alpha_c)$, it holds that almost surely, any cluster point of $\{\boldsymbol{X}_k\}$ is a D_f-critical point of (DOP) *and $\{f(\boldsymbol{x}_{i,k}) : k \in \mathbb{N}\}$ converges for any $i \in [d]$.*

(2) When Assumption 4.3 holds, for any $\varepsilon > 0$, there exists $\alpha_c > 0$, for any $c \in (0, \alpha_c)$,

$$\mathbb{P}\left(\lim_{k\to\infty} \operatorname{dist}(\boldsymbol{X}_k, \{\boldsymbol{X} \in \mathbb{R}^{n\times d} : \boldsymbol{X} = \boldsymbol{x}\mathbf{1}^\top, \mathbf{0} \in D_f(\boldsymbol{x})\}) = 0\right) \geq 1 - \varepsilon,$$

and

$$\mathbb{P}(f(\boldsymbol{x}_{i,k}) \textit{ converges }) \geq 1 - \varepsilon,\ \forall i \in [d].$$

Proof As demonstrated in Line 7 of Algorithm 1, the sequence $\{\boldsymbol{X}_k\}$ follows the update scheme,

$$\boldsymbol{X}_{k+1} = \boldsymbol{X}_k\boldsymbol{W} - \eta_k \boldsymbol{D}_k,\ \boldsymbol{D}_k = [\boldsymbol{d}_{1,k}, \ldots, \boldsymbol{d}_{d,k}]. \tag{4.3}$$

Let the filtration $\mathcal{F}_k := \sigma(\{\boldsymbol{X}_l : l \leq k\})$. We set $\boldsymbol{Z}_k := \boldsymbol{X}_k$, $\boldsymbol{H}_k := \mathbb{E}[\boldsymbol{D}_k|\mathcal{F}_k]$ and $\Xi_{k+1} := \boldsymbol{D}_k - \boldsymbol{H}_k$. To demonstrate that Algorithm 1 fits into our framework (DSM), it suffices to check the update scheme (4.3) is a special case of (DSM) satisfying Assumption 3.1, and verify the validity of Assumption 3.9 or 3.13 with different sampling techniques. Assumption 3.1(1) follows immediately from Assumption 4.2(2) or 4.3(2); Assumption 3.1(2) follows quickly from local boundedness of the conservative field $D_{F_{i,l}}$. Nevertheless, how to verify (1.5) varies across different cases. For case (1) in Theorem 4.4, we set $\Phi_i := D_{f_i}$. In this case, the nonnegative sequence $\{\delta_k\}$ in (1.5) is not subject to any other restrictions, hence we may select a sufficiently large one to meet the practical requirement for $\boldsymbol{H}_k$ and Φ_i. This choice makes Assumption (1.5) hold trivially. For case (2), we need to carefully choose a diminishing $\{\delta_k\}$ so that it conforms to the conditions in Theorem 3.16.

Specifically, we note that $\mathbb{E}[\boldsymbol{d}_{i,k}|\mathcal{F}_k] \in \mathbb{E}_{\mathcal{B}_{i,k}}[\frac{1}{|\mathcal{B}_{i,k}|}\sum_{\boldsymbol{s}_l \in \mathcal{B}_{i,k}} D_{F_{i,l}}(\boldsymbol{x}_{i,k})] = D_{f_i}(\boldsymbol{x}_{i,k})$ in case (2), we can conclude that

$$\frac{1}{d}\boldsymbol{H}_k\mathbf{1}_d \in \frac{1}{d}\sum_{i=1}^{d} D_{f_i}(\boldsymbol{x}_{i,k}) \subseteq \operatorname{conv}\left(\frac{1}{d}\sum_{i=1}^{d} D_{f_i}(\boldsymbol{x}_{i,k})\right).$$

Therefore, by choosing $\Phi_i := D_{f_i}$ and $\delta_k := 0$, we have verified the validity of (1.5).

Furthermore, according to Bolte and Pauwels (2021, Section 6) and the path-differentiability of f, it follows that f is the Lyapunov function of the differential inclusion $\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} \in -D_f(\boldsymbol{x})$, and admits the stable set $\{\boldsymbol{x} \in \mathbb{R}^n : \mathbf{0} \in D_f(\boldsymbol{x})\}$. In addition, combining Assumption 4.1(1) and Proposition 2.7 yields $\{f(\boldsymbol{x}) : \mathbf{0} \in D_f(\boldsymbol{x})\}$ is finite in $\mathbb{R}$. As a result, we verify the validity of Assumption 3.1 with $\Phi := D_f$, $\psi := f$, and $\mathcal{A} := \{\boldsymbol{x} \in \mathbb{R}^n : \mathbf{0} \in D_f(\boldsymbol{x})\}$.

With the choice of $\mathcal{U}_{i,j}(\boldsymbol{x}) := \frac{1}{|\mathcal{B}_{i,j}|}\sum_{\boldsymbol{s}_l \in \mathcal{B}_{i,j}} D_{F_{i,l}}(\boldsymbol{x})$, where $\mathcal{B}_{i,j}, 0 \leq j \leq N-1$ are subsets of equal size, and $\rho_k := 0$, Assumption 3.9 is satisfied for case (1); With the choice of $\xi_{i,k} := \{\textit{Draw out } \mathcal{B}_{i,k} \textit{ from agent } i\}$, $\delta_k := 0$ and $\chi_i(\boldsymbol{x}, \xi_{i,k}) \in \frac{1}{|\mathcal{B}_{i,k}|}\sum_{\boldsymbol{s}_l \in \mathcal{B}_{i,k}} D_{F_{i,l}}(\boldsymbol{x}) - D_{f_i}(\boldsymbol{x})$, Assumption 3.13 is satisfied for case (2). Finally, setting $\mathcal{X}_0$ as a compact set which contains $\boldsymbol{x}_0$, and leveraging Theorem 3.12 and 3.16, we completes our proof. ■

The following theorem demonstrates that Algorithm 1 provably avoids spurious critical points. For clarity, we defer the full proof to Section A.

Theorem 4.5 *Suppose Assumption 4.1 holds. Let $\mathcal{X}_0$ be any compact set of $\mathbb{R}^n$, and sequence $\{\boldsymbol{X}_k\}$ be generated by Algorithm 1.*

(1) When Assumption 4.2 holds, there exists $\alpha_c > 0$, a full-measure subset $\mathcal{S}$ of $(0, \alpha_c)$, and a full-measure subset $\mathcal{X}_1$ of $\mathcal{X}_0$, such that for any $c \in \mathcal{S}$ and $\boldsymbol{x}_0 \in \mathcal{X}_1$, it holds that almost surely, any cluster point of $\{\boldsymbol{X}_k\}$ is a Clarke-critical point of (DOP) *and $\{f(\boldsymbol{x}_{i,k}) : k \in \mathbb{N}\}$ converges for any $i \in [d]$.*

(2) When Assumption 4.3 holds, for any $\varepsilon > 0$, there exists $\alpha_c > 0$, a full-measure subset $\mathcal{S}$ of $(0, \alpha_c)$, and a full-measure subset $\mathcal{X}_1$ of $\mathcal{X}_0$, such that for any $c \in \mathcal{S}$ and $\boldsymbol{x}_0 \in \mathcal{X}_1$,

$$\mathbb{P}\left(\lim_{k\to\infty} \mathrm{dist}(\boldsymbol{X}_k, \{\boldsymbol{X} \in \mathbb{R}^{n\times d} : \boldsymbol{X} = \boldsymbol{x}\boldsymbol{1}^\top, \boldsymbol{0} \in \partial f(\boldsymbol{x})\}) = 0\right) \geq 1 - \varepsilon,$$

and

$$\mathbb{P}(f(\boldsymbol{x}_{i,k}) \text{ converges }) \geq 1 - \varepsilon, \ \forall i \in [d].$$

4.2 Decentralized Generalized SGD with Momentum

Momentum is a technique to accelerate the subgradient descent method in the relevant direction and dampen oscillations. In the single-agent setting, Xiao et al. (2023a) investigates the global convergence properties of generalized SGD with momentum, where an auxiliary function $\phi : \mathbb{R}^n \to \mathbb{R}$ is introduced to normalize the update directions of the primal variables.

In this subsection, motivated by the generalized SGD with momentum analyzed in Xiao et al. (2023a), we introduce the decentralized generalized stochastic subgradient descent method with momentum (DGSGD-M) and present the details in Algorithm 2. Different from DSGD in Algorithm 1, Algorithm 2 introduces the auxiliary variables $\{\boldsymbol{y}_{i,k} : k \in \mathbb{N}\}$ to track the first-order moment of the stochastic subgradients of f_i. Therefore, Algorithm 2 requires two rounds of communication to perform local consensus to $\{\boldsymbol{x}_{1,k}, \ldots, \boldsymbol{x}_{d,k}\}$ and $\{\boldsymbol{y}_{1,k}, \ldots, \boldsymbol{y}_{d,k}\}$, respectively. Moreover, similar to the generalized SGD analyzed in Xiao et al. (2023a), Algorithm 2 introduces an auxiliary function ϕ to normalize the update directions of its primal variables $\{\boldsymbol{x}_{1,k}, \ldots, \boldsymbol{x}_{d,k}\}$.

Remark 4.6 *The auxiliary function ϕ in Algorithm 2 determines how the update directions of the variables $\{\boldsymbol{x}_k\}$ are regularized. Hence, Algorithm 2 yields different variants of DSGD-type methods with different choices of the auxiliary function ϕ. For example, when we choose $\phi(\boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{y}\|^2$, Algorithm 2 corresponds to DSGD with momentum; when we select $\phi(\boldsymbol{y}) = \|\boldsymbol{y}\|_1$, Algorithm 2 employs the sign map to regularize the update directions of the variables $\{\boldsymbol{x}_k\}$, which can be viewed as a decentralized variant of SignSGD analyzed in Bernstein et al. (2018), and denoted as DSignSGD. Interested readers could refer to Xiao et al. (2023a) for more details on the choices of the auxiliary function ϕ.*

Furthermore, recognizing signSGD as a limiting variant of Adam (Chen et al., 2023a), we extend our analysis in Appendix C to demonstrate how Adam-type methods can be adapted to

Algorithm 2 DGSGD-M for solving (DOP).

Input: Initial point $\boldsymbol{x}_0 \in \mathbb{R}^n$, initial descent direction $\boldsymbol{y}_0 \in \mathbb{R}^n$, momentum parameter $\tau > 0$, an auxiliary function $\phi : \mathbb{R}^n \to \mathbb{R}$, and a mixing matrix $\boldsymbol{W}$.

for all $i \in [d]$ in parallel **do**
  Set $k \leftarrow 0$. Initialize $\boldsymbol{x}_{i,k} = \boldsymbol{x}_0, \boldsymbol{y}_{i,k} = \boldsymbol{y}_0$;
  while not terminated **do**
    Choose the step-size η_k;
    Communicate and update the local variable

$$\boldsymbol{x}_{i,k+1} \in \sum_{j \in \mathcal{N}_i} \boldsymbol{W}(i,j)\boldsymbol{x}_{j,k} - \eta_k \partial\phi(\boldsymbol{y}_{i,k});$$

    Randomly select a mini-batch $\mathcal{B}_{i,k+1} \subseteq \mathcal{S}_i$;
    Compute $\boldsymbol{d}_{i,k+1} \in \frac{1}{|\mathcal{B}_{i,k+1}|} \sum_{\boldsymbol{s}_l \in \mathcal{B}_{i,k+1}} D_{F_{i,l}}(\boldsymbol{x}_{i,k+1})$;
    Communicate and update local descent direction

$$\boldsymbol{y}_{i,k+1} = (1 - \tau\eta_k) \sum_{j \in \mathcal{N}_i} \boldsymbol{W}(i,j)\boldsymbol{y}_{j,k} + \tau\eta_k \boldsymbol{d}_{i,k+1};$$

    Set $k \leftarrow k+1$;
  end while
end for
return $\boldsymbol{X}_k := [\boldsymbol{x}_{1,k}, \ldots, \boldsymbol{x}_{d,k}]$.

our framework, hence establishing convergence guarantees for these methods under the nonsmooth nonconvex settings.

Let $\boldsymbol{Y}_k := [\boldsymbol{y}_{1,k}, \ldots, \boldsymbol{y}_{d,k}]$ at k-th iteration in Algorithm 2. The convergence properties of this algorithm are established in Theorem 4.7, within which we concern about two different noise settings and step-sizes in Assumption 4.2 and 4.3.

Theorem 4.7 *Suppose Assumption 4.1 holds. $\{\boldsymbol{X}_k\}$ is generated by Algorithm 2, and auxiliary function ϕ is convex and admits a unique minimizer at $\mathbf{0}$.*

(1) When Assumption 4.2 holds, there exists $\alpha_c > 0$, for any $c \in (0, \alpha_c)$, it holds that almost surely, any cluster point of $\{\boldsymbol{X}_k\}$ is a D_f-critical point of (DOP) *and $\{f(\boldsymbol{x}_{i,k}) : k \in \mathbb{N}\}$ converges for any $i \in [d]$.*

(2) When Assumption 4.3 holds, for any $\varepsilon > 0$, there exists $\alpha_c > 0$, for any $c \in (0, \alpha_c)$,

$$\mathbb{P}\left(\lim_{k\to\infty} \mathrm{dist}(\boldsymbol{X}_k, \{\boldsymbol{X} \in \mathbb{R}^{n\times d} : \boldsymbol{X} = \boldsymbol{x}\boldsymbol{1}^\top, \mathbf{0} \in D_f(\boldsymbol{x})\}) = 0\right) \geq 1 - \varepsilon,$$

and

$$\mathbb{P}(f(\boldsymbol{x}_{i,k}) \text{ converges }) \geq 1 - \varepsilon, \ \forall i \in [d].$$

Proof From Line 5 and Line 8 of Algorithm 2, we can conclude that the sequences $\{\boldsymbol{X}_k\}$ and $\{\boldsymbol{Y}_k\}$ follow the update scheme,

$$\begin{aligned} &\boldsymbol{X}_{k+1} \in \boldsymbol{X}_k\boldsymbol{W} - \eta_k\boldsymbol{T}_k,\ \boldsymbol{T}_k \in [\partial\phi(\boldsymbol{y}_{1,k}),\ldots,\partial\phi(\boldsymbol{y}_{d,k})],\\ &\boldsymbol{D}_{k+1} = [\boldsymbol{d}_{1,k+1},\ldots,\boldsymbol{d}_{d,k+1}],\\ &\boldsymbol{Y}_{k+1} = (1-\tau\eta_k)\boldsymbol{Y}_k\boldsymbol{W} + \tau\eta_k\boldsymbol{D}_{k+1}. \end{aligned} \tag{4.4}$$

Let the filtration $\mathcal{F}_k := \sigma(\{\boldsymbol{X}_l,\ \boldsymbol{Y}_l : l \le k\})$. Then by setting

$$\boldsymbol{Z}_k := \begin{bmatrix} \boldsymbol{X}_k \\ \boldsymbol{Y}_k \end{bmatrix}, \boldsymbol{H}_k := \mathbb{E}\left[\begin{bmatrix} \boldsymbol{T}_k \\ \tau\boldsymbol{Y}_k\boldsymbol{W}-\tau\boldsymbol{D}_{k+1} \end{bmatrix} | \mathcal{F}_k\right], \text{ and } \Xi_{k+1} := \begin{bmatrix} \boldsymbol{T}_k \\ \tau\boldsymbol{Y}_k\boldsymbol{W}-\tau\boldsymbol{D}_{k+1} \end{bmatrix} - \boldsymbol{H}_k,$$

we can conclude that the update scheme (4.4) can be reshaped as $\boldsymbol{Z}_{k+1} = \boldsymbol{Z}_k\boldsymbol{W} - \eta_k(\boldsymbol{H}_k + \Xi_{k+1})$, which aligns with the update format of (DSM).

Now we check the validity of (1.5) and Assumption 3.1, and then show Assumption 3.9 and 3.13 are satisfied for case (1) and (2), respectively. From the local boundedness of D_{f_i}, we can conclude that $\{\boldsymbol{D}_k\}$ is uniformly bounded almost surely whenever $\{\boldsymbol{X}_k\}$ is bounded. Together with the update scheme of $\boldsymbol{Y}_k$, we can conclude that $\|\boldsymbol{Y}_{k+1}\| \le (1-\tau\eta_k)\|\boldsymbol{Y}_k\| + \tau\eta_k\|\boldsymbol{D}_{k+1}\|$, hence there exists a constant $M_Y > 0$ such that $\sup_{k\ge0}\|\boldsymbol{Y}_k\| \le \max\{\|\boldsymbol{Y}_0\|, \sup_{k\ge0}\|\boldsymbol{D}_k\|\} < M_Y$, if $\{\boldsymbol{X}_k\}$ is bounded. Then combining with local boundedness of $\partial\phi$, it can be referred that $\{\boldsymbol{T}_k\}$ is uniformly bounded almost surely whenever $\{\boldsymbol{X}_k\}$ is bounded. This illustrates that Assumption 3.1(2) holds.

Moreover, we set $\Phi_i(\boldsymbol{x},\boldsymbol{y}) := \begin{bmatrix} \partial\phi(\boldsymbol{y}) \\ \tau\boldsymbol{y}-\tau D_{f_i}(\boldsymbol{x}) \end{bmatrix}$, then (1.5) holds automatically for case (1) with sufficiently large δ_k. For case (2), we define δ_k as an exponential moving average of $\{\eta_j\|\boldsymbol{T}_j\|\}_{j\le k}$, expressed as

$$\delta_k := \eta_k\|\boldsymbol{T}_k\| + 2\sum_{j=0}^{k-1}\lambda_2^{k-1-j}\eta_j\|\boldsymbol{T}_j\|.$$

From this, we can deduce that $\frac{1}{d}\boldsymbol{H}_k\mathbf{1}_d \in \frac{1}{d}\sum_{i=1}^{d}\Phi_i^{\delta_k}(\boldsymbol{x}_{i,k},\boldsymbol{y}_{i,k})$ for case (2).

Furthermore, from the definition of Φ_i and (1.5), we attain that $\Phi(\boldsymbol{x},\boldsymbol{y}) = \begin{bmatrix} \partial\phi(\boldsymbol{y}) \\ \tau\boldsymbol{y}-\tau D_f(\boldsymbol{x}) \end{bmatrix}$. Then together with Xiao et al. (2023a, Proposition 4.5) and the path-differentiability of f, we can conclude that $\psi(\boldsymbol{x},\boldsymbol{y}) = f(\boldsymbol{x}) + \frac{1}{\tau}\phi(\boldsymbol{y})$ is the Lyapunov function of the differential inclusion $\left(\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t},\frac{\mathrm{d}\boldsymbol{y}}{\mathrm{d}t}\right) \in -\Phi(\boldsymbol{x},\boldsymbol{y})$, and admits $\{(\boldsymbol{x},\boldsymbol{y}) \in \mathbb{R}^n\times\mathbb{R}^n : \mathbf{0} \in D_f(\boldsymbol{x}), \boldsymbol{y} = \mathbf{0}\}$ as its stable set. Leveraging the definability of f and D_f, it follows that the set $\{\psi(\boldsymbol{x},\boldsymbol{y}) \in \mathbb{R}^n\times\mathbb{R}^n : \mathbf{0} \in D_f(\boldsymbol{x}), \boldsymbol{y} = \mathbf{0}\}$ is finite in $\mathbb{R}$. Therefore, Assumption 3.1 holds for Algorithm 2.

We set $\mathcal{U}_{i,j}(\boldsymbol{x},\boldsymbol{y}) := \begin{bmatrix} \partial\phi(\boldsymbol{y}) \\ \tau\boldsymbol{y}-\tau\frac{1}{|\mathcal{B}_{i,j}|}\sum_{\boldsymbol{s}_l\in\mathcal{B}_{i,j}} D_{F_{i,l}}(\boldsymbol{x}) \end{bmatrix}$, where $\mathcal{B}_{i,j}$ for $0 \le j \le N-1$ are subsets of equal size, and set $\rho_k := \eta_k\|\boldsymbol{T}_k\| + 2\sum_{j=0}^{k-1}\lambda_2^{k-1-j}\eta_j\|\boldsymbol{T}_j\|$. Then (4.4) is a special form of (DSM) with (3.8). Moreover, Assumption 4.2 implies that Assumption 3.9 is satisfied for case (1). With the choice of $\xi_{i,k} := \{\textit{Draw out } \mathcal{B}_{i,k} \textit{ from agent } i\}$, and

$$\chi_i(\boldsymbol{x},\boldsymbol{y},\xi_{i,k}) \in \begin{bmatrix} 0 \\ \frac{-\tau}{|\mathcal{B}_{i,k}|}\sum_{\boldsymbol{s}_l\in\mathcal{B}_{i,k}} D_{F_{i,l}}(\boldsymbol{x},\boldsymbol{y}) - D_{f_i}(\boldsymbol{x},\boldsymbol{y}) \end{bmatrix},$$

(4.4) is a particular form of (DSM) with (3.18), and Assumption 3.13 is satisfied for case (2). Plugging Lemma 3.7 and Theorem 3.15, we can prove that $\delta_k \to 0$, as $k \to \infty$ with arbitrary high probability. [2] Employing Theorem 3.12 and 3.16, we can derive the results in Theorem 4.7.

■

Theorem 4.8 illustrates how Algorithm 2 avoids spurious critical points, whose proof is presented in Appendix A for simplicity. We also provide a numerical example in Appendix B to demonstrate that DSGD-m can avoid spurious critical points.

Theorem 4.8 *Suppose Assumption 4.1 holds and $\boldsymbol{W}$ is a non-singular mixing matrix. $\mathcal{X}_0$ is a compact set of $\mathbb{R}^n$, sequence $\{\boldsymbol{X}_k\}$ is generated by Algorithm 2, and auxiliary function ϕ is convex which admits a unique minimizer at $\mathbf{0}$.*

(1) When Assumption 4.2 holds, there exists $\alpha_c > 0$, a full-measure subset $\mathcal{S}$ of $(0, \alpha_c)$, and a full-measure subset $\mathcal{K}$ of $\mathcal{X}_0 \times \mathbb{R}^n$, such that for any $c \in \mathcal{S}$ and $(\boldsymbol{x}_0, \boldsymbol{y}_0) \in \mathcal{K}$, it holds that almost surely, any cluster point of $\{\boldsymbol{X}_k\}$ is a Clarke-critical point of (DOP) *and $\{f(\boldsymbol{x}_{i,k}) : k \in \mathbb{N}\}$ converges for any $i \in [d]$.*

(2) When Assumption 4.3 holds, for any $\varepsilon > 0$, there exists $\alpha_c > 0$, a full-measure subset $\mathcal{S}$ of $(0, \alpha_c)$, and a full-measure subset $\mathcal{K}$ of $\mathcal{X}_0 \times \mathbb{R}^n$, such that for any $c \in \mathcal{S}$ and $(\boldsymbol{x}_0, \boldsymbol{y}_0) \in \mathcal{K}$,

$$\mathbb{P}\left(\lim_{k\to\infty} \mathrm{dist}(\boldsymbol{X}_k, \{\boldsymbol{X} \in \mathbb{R}^{n\times d} : \boldsymbol{X} = \boldsymbol{x}\mathbf{1}^\top, \mathbf{0} \in \partial f(\boldsymbol{x})\}) = 0\right) \geq 1 - \varepsilon,$$

and

$$\mathbb{P}(f(\boldsymbol{x}_{i,k}) \text{ converges }) \geq 1 - \varepsilon, \ \forall i \in [d].$$

4.3 Decentralized SGD with Gradient-tracking Technique

The decentralized SGD method with gradient-tracking technique (DSGD-T) is another widely employed stochastic (sub)gradient-based method in decentralized learning. Algorithm 3 exhibits the detailed implementations of DSGD-T in minimizing nonsmooth path-differentiable functions. Different from the DSGD in Algorithm 1, DSGD-T introduces a sequence of auxiliary variables to track the average of local subgradients over all the agents. To the best of our knowledge, all the existing works focus on analyzing the convergence of DSGD-T with Clarke regular objective functions, and the convergence of DSGD-T for nonsmooth functions without Clarke regularity remains unexplored.

In this subsection, we prove that DSGD-T can be enclosed by our proposed framework (DSM) with (3.8) or (3.18). Therefore, based on our theoretical results exhibited in Theorem 3.12 and 3.16, we can establish the global convergence of DSGD-T in the minimization of nonsmooth path-differentiable functions.

In the rest of this subsection, we denote $\boldsymbol{V}_k = [\boldsymbol{v}_{1,k}, \ldots, \boldsymbol{v}_{d,k}]$ at k-th iteration in Algorithm 3. We condense the assertion that DSGD-T can fit in our framework (DSM) with different noise settings, along with convergence analysis, into a single theorem.

2. Notice that Theorem 3.15 does not require δ_k to diminish.

Theorem 4.9 *Suppose Assumption 4.1 holds. Let $\{\boldsymbol{X}_k\}$ be the sequence generated by Algorithm 3.*

(1) *When Assumption 4.2 holds, there exists $\alpha_c > 0$, for any $c \in (0, \alpha_c)$, it holds that almost surely, any cluster point of $\{\boldsymbol{X}_k\}$ is a D_f-critical point of* (DOP) *and $\{f(\boldsymbol{x}_{i,k}) : k \in \mathbb{N}\}$ converges for any $i \in [d]$.*

(2) *When Assumption 4.3 holds, for any $\varepsilon > 0$, there exists $\alpha_c > 0$, for any $c \in (0, \alpha_c)$,*

$$\mathbb{P}\left(\lim_{k\to\infty} \mathrm{dist}(\boldsymbol{X}_k, \{\boldsymbol{X} \in \mathbb{R}^{n\times d} : \boldsymbol{X} = \boldsymbol{x}\boldsymbol{1}^\top, \boldsymbol{0} \in D_f(\boldsymbol{x})\}) = 0\right) \geq 1 - \varepsilon$$

and

$$\mathbb{P}(f(\boldsymbol{x}_{i,k}) \textit{ converges }) \geq 1 - \varepsilon,\ \forall i \in [d].$$

Algorithm 3 DSGD-T for solving (DOP)

Input: Initial point $\boldsymbol{x}_0 \in \mathbb{R}^n$, initial descent direction $\boldsymbol{V}_0 = [\boldsymbol{v}_{1,0}, \ldots, \boldsymbol{v}_{d,0}]$ satisfying $\boldsymbol{V}_0\boldsymbol{1}_d = \sum_{i=1}^d \boldsymbol{d}_{i,0}$, where $\boldsymbol{d}_{i,0} \in \frac{1}{|\mathcal{B}_{i,0}|}\sum_{\boldsymbol{s}_l\in\mathcal{B}_{i,0}} D_{F_{i,l}}(\boldsymbol{x}_{i,0})$, $\mathcal{B}_{i,0}$ is a mini-batch of $\mathcal{S}_i$, and a mixing matrix $\boldsymbol{W}$.

1: **for** all $i \in [d]$ in parallel **do**

2: Set $k \leftarrow 0$. Initialize $\boldsymbol{x}_{i,k} = \boldsymbol{x}_0$;

3: **while** not terminated **do**

4: Choose the step-size η_k;

5: Communicate and update the local variable

$$\boldsymbol{x}_{i,k+1} = \sum_{j\in\mathcal{N}_i} \boldsymbol{W}(i,j)\boldsymbol{x}_{j,k} - \eta_k \boldsymbol{v}_{i,k};$$

6: Randomly select a mini-batch $\mathcal{B}_{i,k+1} \subseteq \mathcal{S}_i$;

7: Compute $\boldsymbol{d}_{i,k+1} \in \frac{1}{|\mathcal{B}_{i,k+1}|}\sum_{\boldsymbol{s}_l\in\mathcal{B}_{i,k+1}} D_{F_{i,l}}(\boldsymbol{x}_{i,k+1})$;

8: Communicate and update local descent direction

$$\boldsymbol{v}_{i,k+1} = \sum_{j\in\mathcal{N}_i} \boldsymbol{W}(i,j)(\boldsymbol{v}_{j,k} + \boldsymbol{d}_{j,k+1} - \boldsymbol{d}_{j,k});$$

9: Set $k \leftarrow k+1$;

10: **end while**

11: **end for**

12: **return** $\boldsymbol{X}_k := [\boldsymbol{x}_{1,k}, \ldots, \boldsymbol{x}_{d,k}]$.

Proof From Line 5 - Line 8 of Algorithm 3, the sequences $\{\boldsymbol{X}_k\}$ and $\{\boldsymbol{V}_k\}$ are updated by the following scheme,

$$\begin{aligned} &\boldsymbol{X}_{k+1} = \boldsymbol{X}_k\boldsymbol{W} - \eta_k\boldsymbol{V}_k,\ \boldsymbol{D}_{k+1} = [\boldsymbol{d}_{1,k+1}, \ldots, \boldsymbol{d}_{d,k+1}], \\ &\boldsymbol{V}_{k+1} = (\boldsymbol{V}_k + \boldsymbol{D}_{k+1} - \boldsymbol{D}_k)\boldsymbol{W}. \end{aligned} \tag{4.5}$$

Let the filtration $\mathcal{F}_k := \sigma(\{\boldsymbol{X}_l : l \leq k\})$. By setting $\boldsymbol{Z}_k := \boldsymbol{X}_k$, $\boldsymbol{H}_k := \mathbb{E}[\boldsymbol{V}_k|\mathcal{F}_k]$ and $\Xi_{k+1} := \boldsymbol{V}_k - \boldsymbol{H}_k$, we can conclude that (4.5) coincides with the framework (DSM).

Then we aim to check the validity of Assumption 3.1. Due to the local boundedness of any conservative field, there exists $M_D > 0$ such that $\sup_{k\geq 0} \|\boldsymbol{D}_k\| \leq M_D$ almost surely when $\{\boldsymbol{X}_k\}$ is restricted in some bounded set. Then by the update scheme of $\{\boldsymbol{V}_k\}$ in (4.5), it holds that

$$\boldsymbol{V}_{k+1} = \boldsymbol{D}_{k+1}\boldsymbol{W} + \sum_{i=1}^{k} \boldsymbol{D}_i(\boldsymbol{W} - \boldsymbol{I}_d)\boldsymbol{W}^{k+1-i} - \boldsymbol{D}_0\boldsymbol{W}^{k+1} + \boldsymbol{V}_0\boldsymbol{W}^{k+1}.$$

Then from Corollary 2.2, there exists a constant $\alpha \in (0,1)$, such that $\left\|(\boldsymbol{W} - \boldsymbol{I}_d)\boldsymbol{W}^k\right\| \leq 2\alpha^k$ holds for any $k \geq 0$. Therefore, almost surely, it holds that

$$\begin{aligned}
\sup_{k\geq 0} \|\boldsymbol{V}_k\| \leq{}& \sup_{k\geq 0} \|\boldsymbol{D}_{k+1}\boldsymbol{W} + \sum_{i=1}^{k} \boldsymbol{D}_i(\boldsymbol{W} - \boldsymbol{I}_d)\boldsymbol{W}^{k+1-i} - \boldsymbol{D}_0\boldsymbol{W}^{k+1} + \boldsymbol{V}_0\boldsymbol{W}^{k+1}\| \\
&\leq 3\sup_{k\geq 0} \|\boldsymbol{D}_k\| + \left(\sup_{k\geq 0} \|\boldsymbol{D}_k\|\right) \cdot \sum_{i=1}^{k} \left\|(\boldsymbol{W} - \boldsymbol{I}_d)\boldsymbol{W}^{k+1-i}\right\| \\
&\leq 3M_D + \sum_{i=1}^{\infty} 2M_D\alpha^i \leq \left(3 + \frac{2}{1-\alpha}\right) M_D.
\end{aligned}$$

Therefore, we can conclude that $\{\boldsymbol{V}_k\}$ is uniformly bounded almost surely when $\{\boldsymbol{X}_k\}$ is bounded, hence verifying the validity of Assumption 3.1(2).

Assumption 3.1(1) follows quickly from Assumption 4.2 or 4.3. (1.5) holds vacuously with sufficiently large δ_k and $\Phi_i := D_{f_i}$ for case (1). Moreover for case (2), followed by $\mathbb{E}[\boldsymbol{d}_{i,k}|\mathcal{F}_k] \in \mathbb{E}_{\mathcal{B}_{i,k}}[\frac{1}{|\mathcal{B}_{i,k}|}\sum_{\boldsymbol{s}_l\in\mathcal{B}_{i,k}} D_{F_{i,l}}(\boldsymbol{x}_{i,k})] = D_{f_i}(\boldsymbol{x}_{i,k})$, it is easy to deduce that

$$\frac{1}{d}\boldsymbol{H}_k\boldsymbol{1}_d = \mathbb{E}[\frac{1}{d}\boldsymbol{V}_k\boldsymbol{1}_d|\mathcal{F}_k] = \mathbb{E}[\frac{1}{d}\boldsymbol{D}_k\boldsymbol{1}_d|\mathcal{F}_k] \in \frac{1}{d}\sum_{i=1}^{d} D_{f_i}(\boldsymbol{x}_{i,k}), \forall k \in \mathbb{N}.$$

Therefore, by choosing $\delta_k := 0$ and $\Phi_i := D_{f_i}$ for all $i \in [d]$, we verify the validity of (1.5).

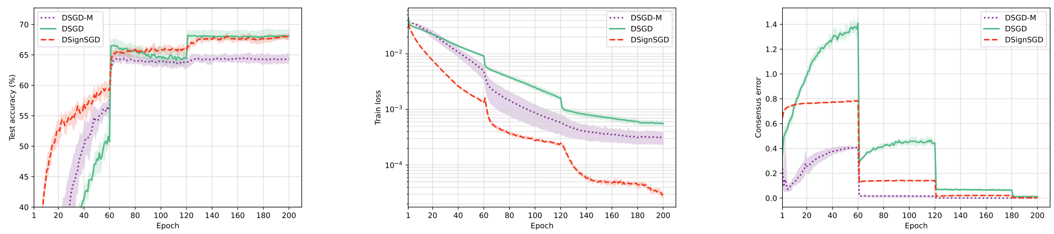
Furthermore, according to Bolte and Pauwels (2021, Section 6), we can assert that f is the Lyapunov function of the differential inclusion $\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} \in -D_f(\boldsymbol{x})$, and admits the stable set $\{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{0} \in D_f(\boldsymbol{x})\}$. In addition, the definability of f and D_f illustrates that $\{f(\boldsymbol{x}) : \boldsymbol{0} \in D_f(\boldsymbol{x})\}$ is finite in $\mathbb{R}$. As a result, we have verified the validity of Assumption 3.1 with $\Phi := D_f$, $\psi := f$, and $\mathcal{A} := \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{0} \in D_f(\boldsymbol{x})\}$.

Following the same construction of $\mathcal{U}_{i,k}(\boldsymbol{x})$, ρ_k, $\xi_{i,k}$ and χ_i as in the proof of Theorem 4.4, we could verify Assumption 3.9 and 3.13 hold for case (1) and case (2), respectively. Henceforth, the update scheme (4.5) is a variant of (DSM) with either (3.8) or (3.18). Applying our theoretical results in Theorem 3.12 and 3.16, we completes the proof.

■

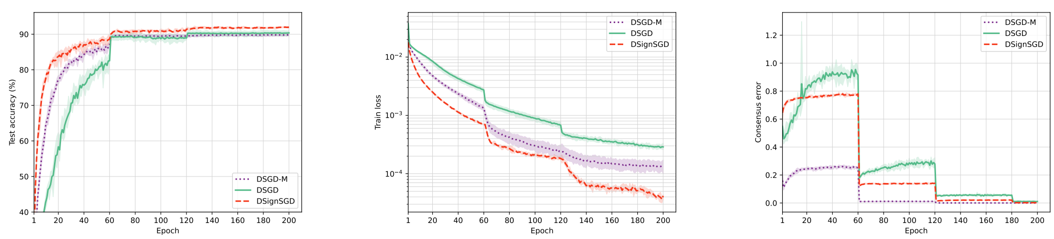
5. Numerical Experiments

In this section, we evaluate the numerical performance of our analyzed decentralized stochastic subgradient-based methods in Section 4 for decentralized training nonsmooth



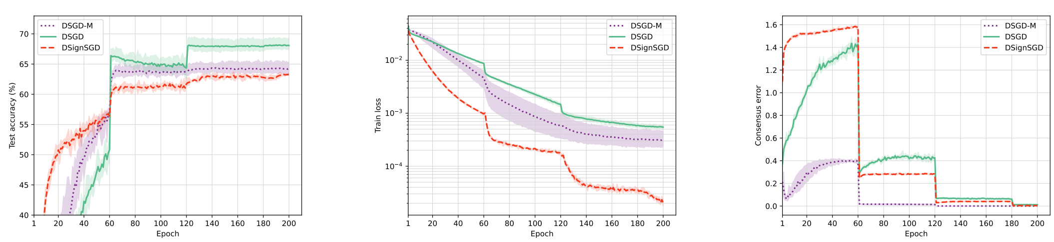
(a) Test Accuracy (b) Train Loss (c) Consensus Error

Figure 1: Numerical performance comparison of DSGD, DSGD-M, and DSignSGD in training ResNet50 on CIFAR-100 dataset using random reshuffling strategy.



(a) Test Accuracy (b) Train Loss (c) Consensus Error

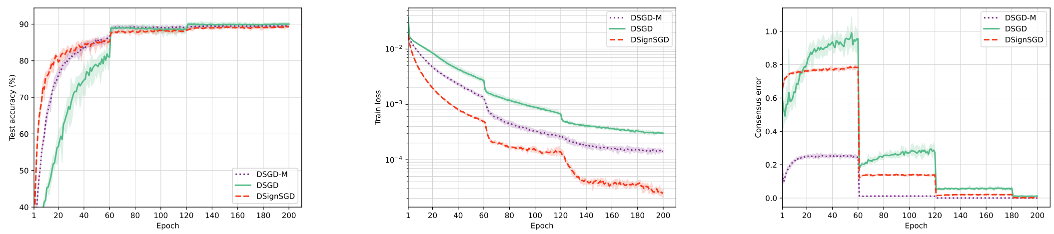
Figure 2: Numerical performance comparison of DSGD, DSGD-M and DSignSGD in training ResNet50 on CIFAR-10 dataset using random reshuffling strategy.



(a) Test Accuracy (b) Train Loss (c) Consensus Error

Figure 3: Numerical performance comparison of DSGD, DSGD-M, and DSignSGD in training ResNet50 on CIFAR-100 dataset using with-replacement sampling strategy.

neural networks. All numerical experiments in this section are conducted on a workstation equipped with two Intel(R) Xeon(R) Processors Gold 5317 CPUs (at 3.00GHz, 18M Cache) and 8 NVIDIA GeForce RTX 4090 GPUs under Ubuntu 20.04.1. We implement all decentralized algorithms with Python 3.8 and PyTorch 1.13.0 using NCCL 2.14.3 (CUDA 11.7) as the communication backend.



(a) Test Accuracy (b) Train Loss (c) Consensus Error

Figure 4: Numerical performance comparison of DSGD, DSGD-M, and DSignSGD in training ResNet50 on CIFAR-10 dataset using with-replacement sampling strategy.

5.1 Experiment Settings

We conduct a series of image classification tasks with CIFAR-10 and CIFAR-100 datasets (Krizhevsky et al., 2009), which both consist of 50000 training samples and 10000 test samples. ResNet-50 (He et al., 2016) is employed as our nonsmooth neural network. The decentralized network topology is set as ring-structure with 8 agents, and the mixing matrix is chosen as the Metropolis weight matrix (Xiao et al., 2006). In each experiment, we divide the training dataset into 8 equal parts, with each part serving as the local dataset for each agent. During any epoch of training, the local dataset of each agent itself remains unchanged, while the mini-batch is selected by random reshuffling or with-replacement sampling to ensure compliance with the sampling scheme in Assumption 4.2 or 4.3.

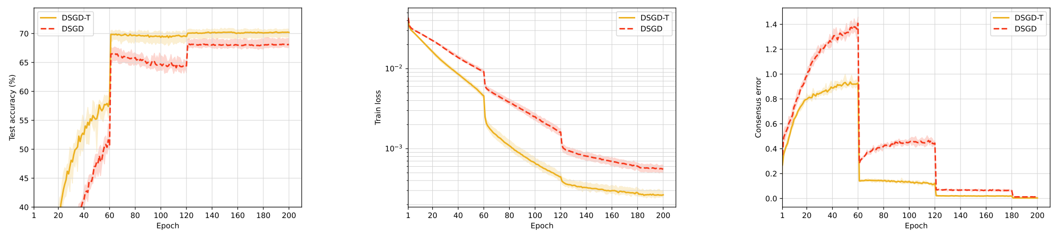
Moreover, we set the number of epochs as 200 and choose the batch size as 128. The strategy for selecting step-size is designed to reduce it three times, specifically at the 60-th, 120-th, and 160-th epochs, with a decay factor of 0.2.

5.2 Performance Comparison between DSGD, DSGD-M, and DSignSGD

We first evaluate the numerical performance of DSGD (Algorithm 1), DSGD-M (Algorithm 2 with $\phi = \frac{1}{2}\|\boldsymbol{y}\|^2$), and DSignSGD (Algorithm 2 with $\phi = \|\boldsymbol{y}\|_1$). All compared methods are executed five times with varying random seeds. In each test instance, all these compared methods are initialized with the same randomly generated initial points.

In Algorithm 2, we choose the momentum parameter τ as $\frac{0.1}{\eta_0}$, and the initial descent direction is taken as $\boldsymbol{Y}_0 = (1-\tau\eta_0)\boldsymbol{D}_0\boldsymbol{W} + \tau\eta_0\boldsymbol{D}_0$. Additionally, we set the initial step-size η_0 of DSGD and DSGD-M to be 0.2, and the counterpart of DSignSGD to be 10^{-4}.

We present the numerical results of the above-mentioned decentralized stochastic subgradient-based methods in Figures 1- 4, including the test accuracy, train loss, and consensus error under two different sampling strategies. It is worth mentioning that the consensus error at the k-th iteration is measured by $\frac{1}{\sqrt{d}}\|\boldsymbol{X}_k(\boldsymbol{I}_d - \boldsymbol{P})\|_{\mathrm{F}}$. As illustrated in Figures 1-4, all the compared decentralized stochastic subgradient-based methods are able to train the ResNet-50 network to a high accuracy. Moreover, we can conclude from Figure 1(a) and 2(a) that when we employ the random reshuffling strategy, DSignSGD achieves similar test accuracy as DSGD and better test accuracy compared to DSGD-M, while showing



(a) Test Accuracy (b) Train Loss (c) Consensus Error

Figure 5: Numerical performance comparison of DSGD and DSGD-T in training ResNet50 on CIFAR-100 dataset using random reshuffling strategy.

superior performance to DSGD and DSGD-M in the aspect of train loss from Figure 1(b) and 2(b). When using with-replacement sampling, DSGD achieves higher test accuracy than that of DSGD-M and DSignSGD, but exhibits the worst training-loss-performance from Figure 3- 4.

Furthermore, in all the compared methods, the consensus error diminishes to 0 throughout the iterations. These results illustrate that our framework is able to yield efficient stochastic subgradient-based methods for decentralized training tasks.

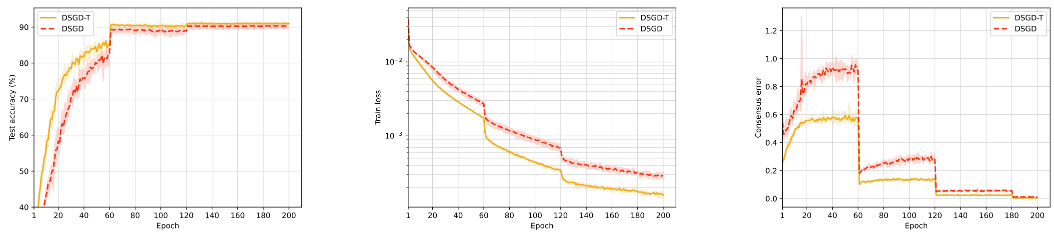
5.3 Performance Comparison between DSGD and DSGD-T

In this subsection, we evaluate the numerical performance of DSGD (Algorithm 1) and DSGD-T (Algorithm 3). Similar to the settings in Section 5.2, we repeat the numerical experiments for 5 times with varying random seeds for DSGD and DSGD-T. For each random seed, DSGD and DSGD-T are executed with the same initial point $\boldsymbol{x}_0$. The initial step-size is taken as 0.2 for both algorithms.

The numerical results on the CIFAR-100 and CIFAR-10 datasets are presented in Figures 5- 8, respectively. It can be concluded that DSGD-T demonstrates slightly better performance than DSGD in all aspects under all sampling strategies, especially in train loss. Therefore, we can conclude that DSGD-T achieves high efficiency in decentralized training tasks while enjoying convergence guarantees from our framework (DSM).

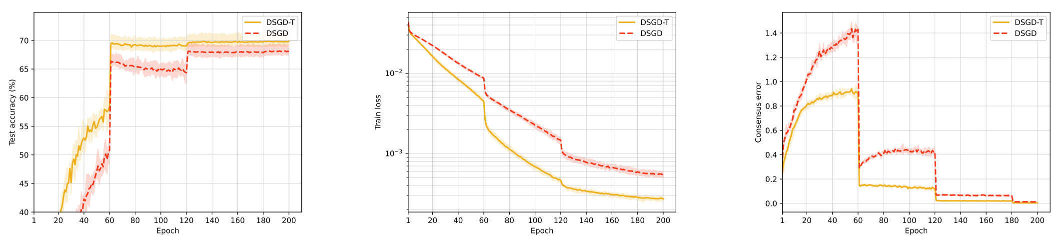
6. Conclusion

Decentralized stochastic subgradient-based methods play an important role in the decentralized training of deep neural networks, where the widely employed nonsmooth building blocks result in non-Clarke regular loss functions. As nearly all of the existing works on decentralized optimization focus on the Clarke regular functions, there is a gap between theoretical analysis and implementation in real-world training tasks. Additionally, despite the development of various acceleration techniques to improve the performance of decentralized stochastic subgradient methods, the convergence guarantees of these methods in training tasks are largely absent.



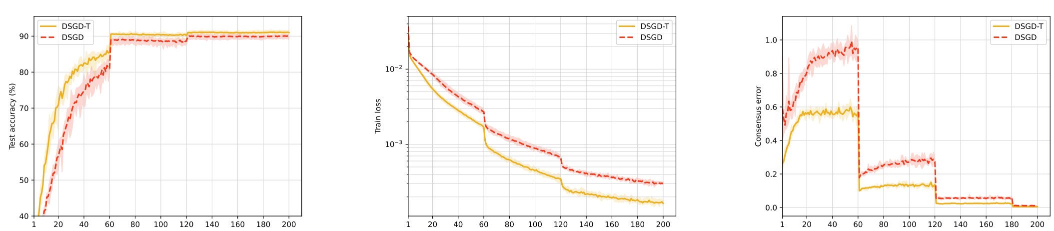
(a) Test Accuracy (b) Train Loss (c) Consensus Error

Figure 6: Numerical performance comparison of DSGD and DSGD-T in training ResNet50 on CIFAR-10 dataset using random reshuffling strategy.



(a) Test Accuracy (b) Train Loss (c) Consensus Error

Figure 7: Numerical performance comparison of DSGD and DSGD-T in training ResNet50 on CIFAR-100 dataset using with-replacement sampling strategy.



(a) Test Accuracy (b) Train Loss (c) Consensus Error

Figure 8: Numerical performance comparison of DSGD and DSGD-T in training ResNet50 on CIFAR-10 dataset using with-replacement sampling strategy.

To address these issues, we propose a unified framework (DSM) for analyzing the global convergence of stochastic subgradient-based methods in nonsmooth decentralized optimization. We conduct consensus analysis and relate the average iterates of (DSM) to the trajectories of its corresponding noiseless differential inclusion. Then, under two different evaluation noise settings, we establish convergence properties of (DSM), in the sense that $\{z_{i,k}\}$ asymptotically reaches consensus and converges to the stable set of the corresponding noiseless differential inclusion.

Moreover, we show our proposed framework (DSM) is general enough to enclose the implementations of various acceleration techniques for decentralized stochastic subgradient methods, including momentum and gradient-tracking techniques. In particular, we demonstrate that a wide range of decentralized stochastic subgradient-based methods, such as DSGD, DSGD-M, and DSGD-T, together with two common sampling schemes, are all encompassed by our proposed framework. Therefore, our theoretical results, for the first time, provide the global convergence guarantees for these decentralized stochastic subgradient-based methods in nonsmooth nonconvex optimization. Furthermore, based on our proposed framework, we develop a new method named DSignSGD, where sign-mapping is employed to regularize the update directions. Preliminary numerical results verify the validity of our developed theoretical results and exhibit the high efficiency of these decentralized stochastic subgradient-based methods enclosed in our framework (DSM).

Our work sets several directions for future research in the field of nonsmooth nonconvex decentralized optimization. Existing ODE approaches are limited to establishing asymptotic convergence properties or the convergence rate for the continuous-time trajectories of the corresponding differential inclusion Bolte et al. (2022b, 2025). Consequently, deriving explicit convergence rates for various nonsmooth decentralized stochastic subgradient methods (such as DSGD, DSGT, and their momentum variants) remains a significant challenge. While our analysis establishes the convergence of DSGT in nonsmooth nonconvex settings, it reveals that DSGT and DSGD are corresponding to the same limiting differential inclusion. Empirically, however, Figures 5-8 demonstrate that DSGT consistently converges faster than DSGD. Bridging this gap between theory and practice, especially in establishing explicit convergence rates that distinguish between DSGT and DSGD in the nonsmooth nonconvex regime, remains a critical direction for future research.

Further attention is also needed on the convergence of decentralized methods with time-varying networks, as well as those that incorporate asynchronous or compressed communication. Additionally, we are also eager to explore routes that presume relaxed conditions for convergence analysis of decentralized methods in solving (DOP). For instance, closed-measure approaches (Bianchi and Rios-Zertuche, 2021; Bolte et al., 2022a; Le, 2024) are developed for establishing the convergence properties of stochastic subgradient-based methods, with relaxed conditions on the step-sizes and evaluation noises. How to employ these closed-measure approaches for analyzing decentralized stochastic subgradient-based methods is worthy of investigating.

Acknowledgments

The work of Professor Xin Liu was supported in part by the National Key R&D Program of China (2023YFA1009300), the National Natural Science Foundation of China (12125108, 12288201), and CAS-Croucher Funding Scheme for Joint Laboratories "CAS AMSS-PolyU Joint Laboratory of Applied Mathematics: Nonlinear Optimization Theory, Algorithms and Applications". The work of Nachuan Xiao was supported by Guangdong Basic and Applied Basic Research Foundation (2026A1515030067).

Appendix A. Proofs of Theorem 4.5 and 4.8

In this part, we devote to showing the avoidance of spurious critical points for Algorithm 1 and 2. In other words, these algorithms can avoid points that are critical for a specific conservative field but are not Clarke-critical points. We begin with the definition of the concept of almost everywhere $\mathcal{C}^1$ set-valued mapping.

Definition A.1 *A measurable mapping $q : \mathbb{R}^{nd} \to \mathbb{R}^{nd}$ is almost everywhere $\mathcal{C}^1$ if for almost every $\boldsymbol{z} \in \mathbb{R}^{nd}$, q is locally continuously differentiable in a neighborhood of z.*

Moreover, a set-valued mapping $\mathcal{Q} : \mathbb{R}^{nd} \rightrightarrows \mathbb{R}^{nd}$ is almost everywhere $\mathcal{C}^1$ if there exists an almost everywhere $\mathcal{C}^1$ mapping $q : \mathbb{R}^{nd} \to \mathbb{R}^{nd}$ such that for almost every $\boldsymbol{z} \in \mathbb{R}^{nd}$, $\mathcal{Q}(\boldsymbol{z}) = \{q(\boldsymbol{z})\}$.

A.1 Proof of Theorem 4.5

For ease of presentation, we define

$$F_{i,\mathcal{B}_{i,k+1}}(\boldsymbol{x}) := \frac{1}{|\mathcal{B}_{i,k+1}|} \sum_{\boldsymbol{s}_l \in \mathcal{B}_{i,k+1}} F_i(\boldsymbol{x}; \boldsymbol{s}_l), \quad \text{and} \quad D_{F_i,\mathcal{B}_{i,k+1}}(\boldsymbol{x}) := \frac{1}{|\mathcal{B}_{i,k+1}|} \sum_{\boldsymbol{s}_l \in \mathcal{B}_{i,k+1}} D_{F_i}(\boldsymbol{x}; \boldsymbol{s}_l).$$

Define the set-valued mappings $\mathcal{U}_k : \mathbb{R}^{nd} \rightrightarrows \mathbb{R}^{nd}$ as

$$\mathcal{U}_k(\boldsymbol{x}) := \left(D_{F_1,\mathcal{B}_{1,k+1}}(\boldsymbol{x}_{(1)}), \ldots, D_{F_d,\mathcal{B}_{d,k+1}}(\boldsymbol{x}_{(d)}) \right)^\top.$$

Note that $(F_i(\cdot, \boldsymbol{s}_l), D_{F_i}(\cdot, \boldsymbol{s}_l))$ has a $\mathcal{C}^r$ variational stratification (Bolte and Pauwels, 2021, Theorem 4), hence the mappings $\{\mathcal{U}_k\}$ are almost everywhere $\mathcal{C}^1$ over $\mathbb{R}^{nd}$, denoted by $\mathcal{U}_k(\boldsymbol{x}) = \{u_k(\boldsymbol{x})\}$, for a.e. $\boldsymbol{x} \in \mathbb{R}^{nd}$. Furthermore, let

$$\mathcal{Y} := \left\{ \boldsymbol{x} \in \mathbb{R}^{nd} : D_{F_i}(\boldsymbol{x}; \boldsymbol{s}_l) \neq \{\nabla F_i(\boldsymbol{x}, \boldsymbol{s}_l)\}, \text{ for some } i \in [N] \text{ and some } \boldsymbol{s}_l \in \mathcal{S}_i \right\},$$

which is a zero-measure subset.

With a slight abuse of notation, we vectorize the sequence $\{\boldsymbol{X}_k\}$ by column-major order, and denote as $\boldsymbol{x}_k \in \mathbb{R}^{nd}$. Then the vectorized sequence $\{\boldsymbol{x}_k\}$ generated by Algorithm 1 satisfies

$$\boldsymbol{x}_{k+1} \in (\boldsymbol{W} \otimes \boldsymbol{I}_n)\boldsymbol{x}_k - c\hat{\eta}_k \mathcal{U}_k(\boldsymbol{x}_k). \tag{A.1}$$

Set $\boldsymbol{P} := \boldsymbol{W} \otimes \boldsymbol{I}_n$, $\mathcal{Q}_k := \hat{\eta}_k \mathcal{U}_k$, and $q_k := \hat{\eta}_k u_k$. The Jacobian of the mapping $\boldsymbol{P}\boldsymbol{x} - cq_k(\boldsymbol{x})$ is

$$\boldsymbol{J}_{c,k}(\boldsymbol{x}) = \boldsymbol{W} \otimes \boldsymbol{I}_n - c\hat{\eta}_k \mathrm{Diag}(\nabla^2 F_{1,\mathcal{B}_{1,k+1}}(\boldsymbol{x}_{(1)}), \ldots, \nabla^2 F_{d,\mathcal{B}_{d,k+1}}(\boldsymbol{x}_{(d)})). \tag{A.2}$$

The determinant of $\boldsymbol{J}_{c,k}(\boldsymbol{x})$ is a zero polynomial of c if and only if there exists $\boldsymbol{v} = [\boldsymbol{v}_1^\top, \ldots, \boldsymbol{v}_d^\top]^\top$ such that $\boldsymbol{v} \in \mathrm{Ker}(\boldsymbol{W} \otimes \boldsymbol{I}_n)$ and $\boldsymbol{v}_i \in \mathrm{Ker}(\nabla^2 F_{i,\mathcal{B}_{i,k+1}}(\boldsymbol{x}_{(i)}))$ for any $i \in [d]$. As each $F_i(\cdot, \boldsymbol{s}_l)$ is a definable function, we can infer that such $(\boldsymbol{x}_{(1)}, \ldots, \boldsymbol{x}_{(d)})$ constitutes a zero-measure subset of $\mathbb{R}^{nd}$. We refer to Γ_k as the full-measure subset of $\mathbb{R}^{nd}$, in which $\mathcal{U}_k(\boldsymbol{x}) = \{u_k(\boldsymbol{x})\}$, and $\det(\boldsymbol{J}_{c,k}(\boldsymbol{x}))$ is a non-trival nd-th order polynomial of c. This indicates the set

$$\{c \in \mathbb{R} : \det(\boldsymbol{J}_{c,k}(\boldsymbol{x})) = 0\}$$

is zero-measure in $\mathbb{R}$. According to Fubini's Theorem, there exists a full-measure subset $\mathcal{S}_k \subseteq \mathbb{R}_+$, such that for any $c \in \mathcal{S}_k$, $\{\boldsymbol{x} \in \Gamma_k : \det(\boldsymbol{J}_{c,k}(\boldsymbol{x})) \neq 0\}$ is a full-measure subset in $\mathbb{R}^{nd}$.

Applying the inverse function theorem, we can conclude that for any $c \in \mathcal{S}_k$ and any $\boldsymbol{x} \in \{x \in \Gamma_k : \det(\boldsymbol{J}_{c,k}(\boldsymbol{x})) \neq 0\}$, $\boldsymbol{P}\boldsymbol{x} - cq_k(\boldsymbol{x})$ is a local diffeomorphism in a neighborhood of $\boldsymbol{x}$, denoted by $\mathbb{B}(\boldsymbol{x}, \delta_{\boldsymbol{x}})$. As a result,

$$\{(\boldsymbol{P} - cq_k)(\mathbb{B}(\boldsymbol{x}^i, \delta_{\boldsymbol{x}^i}))\}_{\boldsymbol{x} \in \{x \in \Gamma_k : \det(\boldsymbol{J}_{c,k}(\boldsymbol{x})) \neq 0\}}$$

is an open covering of $\{\boldsymbol{x} \in \Gamma_k : \det(\boldsymbol{J}_{c,k}(\boldsymbol{x})) \neq 0\}$. Based on Lindelof's lemma (Kelley, 2017), there exists countable $\{\boldsymbol{x}^i\}_{i \in \mathbb{N}_+} \subseteq \{\boldsymbol{x} \in \Gamma_k : \det(\boldsymbol{J}_{c,k}(\boldsymbol{x})) \neq 0\}$, such that

$$\{\boldsymbol{x} \in \Gamma_k : \det(\boldsymbol{J}_{c,k}(\boldsymbol{x})) \neq 0\} \subseteq \bigcup_{i \in \mathbb{N}_+} (\boldsymbol{P} - cq_k)(\mathbb{B}(\boldsymbol{x}^i, \delta_{\boldsymbol{x}^i})).$$

Given any zero-measure subset $A \subseteq \mathbb{R}^{nd}$, since $\boldsymbol{P} - cq_k$ is a diffeomorphism over $\mathbb{B}(\boldsymbol{x}^i, \delta_{\boldsymbol{x}^i})$, it can be deduced that the set $\{\boldsymbol{z} \in \mathbb{B}(\boldsymbol{x}^i, \delta_{\boldsymbol{x}^i}), \boldsymbol{P}\boldsymbol{z} - cq_k(\boldsymbol{z}) \in A\}$ is zero-measure. Therefore, the set $\bigcup_{i \in \mathbb{N}_+} \{\boldsymbol{z} \in \mathbb{B}(\boldsymbol{x}^i, \delta_{\boldsymbol{x}^i}), \boldsymbol{P}\boldsymbol{z} - cq_k(\boldsymbol{z}) \in A\}$ is a zero-measure subset, which yields that

$$\{\boldsymbol{z} \in \mathbb{R}^{nd} : \boldsymbol{P}\boldsymbol{z} - cq_k(\boldsymbol{z}) \in A\} \subset \bigcup_{i \in \mathbb{N}_+} \{\boldsymbol{z} \in \mathbb{B}(\boldsymbol{x}^i, \delta_{\boldsymbol{x}^i}), \boldsymbol{P}\boldsymbol{z} - cq_k(\boldsymbol{z}) \in A\} \cup \{\boldsymbol{z} \in \Gamma_k : \det(\boldsymbol{J}_{c,k}(\boldsymbol{z})) \neq 0\}^c$$

is zero-measure. That is to say, for any $c \in \mathcal{S}_k$, the set $\{\boldsymbol{z} \in \mathbb{R}^{nd} : \boldsymbol{P}\boldsymbol{z} - cq_k(\boldsymbol{z}) \in A\}$ is zero-measure.

Let $\mathcal{S} := \cap_{k \geq p} \mathcal{S}_k$, $A_0 := A = \mathcal{Y}, A_{k+1} := \{\boldsymbol{x} \in \mathbb{R}^{nd} : \boldsymbol{P}\boldsymbol{z} - cq_k(\boldsymbol{z}) \in A\}$, and $\hat{A} := \cup_{k \geq 0} A_k$. It can be inferred that $\mathcal{S}$ and $\hat{A}^c$ are full-measure subsets in $\mathbb{R}$ and $\mathbb{R}^{nd}$, respectively. For any $c \in \mathcal{S}$, and any $\boldsymbol{x} \in \hat{A}^c$, we have $\boldsymbol{P}\boldsymbol{x} - cq_k(\boldsymbol{x}) \in A^c$. Employing Fubini's theorem, the set $\{c \in \mathbb{R}, \boldsymbol{x} \in \mathbb{R}^{nd}, \cup_{k \geq 0}\{\boldsymbol{P}\boldsymbol{x} - cq_k(\boldsymbol{x})\} \subseteq A^c\}$ is full-measure in $\mathbb{R} \times \mathbb{R}^{nd}$, and thus, there exists a full-measure subset $\hat{\mathcal{S}} \times \hat{\mathcal{K}}$, whenever $(c, \boldsymbol{x}_0) \in \hat{\mathcal{S}} \times \hat{\mathcal{K}}$, it follows that $\cup_{k \geq 0}\{\boldsymbol{P}\boldsymbol{x} - cq_k(\boldsymbol{x})\} \subseteq A^c$, that is, $\{\boldsymbol{x}_k\} \in \mathcal{Y}^c$.

Furthermore, the update scheme of vectorized sequence $\{\boldsymbol{x}_k\}$ can be reformulated as

$$\begin{cases} \boldsymbol{g}_k = \left(\nabla F_{1,\mathcal{B}_{1,k+1}}(\boldsymbol{x}_{1,k}), \ldots, \nabla F_{d,\mathcal{B}_{d,k+1}}(\boldsymbol{x}_{d,k})\right)^\top, \\ \boldsymbol{x}_{k+1} \in (\boldsymbol{W} \otimes \boldsymbol{I}_n)\boldsymbol{x}_k - c\hat{\eta}_k \boldsymbol{g}_k. \end{cases}$$

Therefore, by applying Theorem 4.4 with $D_f = \partial f$, we can choose $\mathcal{S} = \hat{\mathcal{S}} \cap (0, \alpha_c)$ and $\mathcal{K} = \hat{\mathcal{K}} \cap \mathcal{X}_0$, such that the claim in Theorem 4.5 holds.

A.2 Proof of Theorem 4.8

Let the set-valued mappings $\mathcal{U}_{1,k} : \mathbb{R}^{2nd} \rightrightarrows \mathbb{R}^{nd}, \mathcal{U}_{2,k} : \mathbb{R}^{2nd} \rightrightarrows \mathbb{R}^{2nd}$ be defined as

$$\begin{aligned} \mathcal{U}_{1,k}(\boldsymbol{x}, \boldsymbol{y}) &:= \left(\partial\hat{\phi}(\boldsymbol{y}), \boldsymbol{0}\right), \\ \mathcal{U}_{2,k}(\boldsymbol{x}, \boldsymbol{y}) &:= \left(\boldsymbol{0}, \tau(\boldsymbol{W} \otimes \boldsymbol{I}_n)\boldsymbol{y} - \tau\left(D_{F_1,\mathcal{B}_{1,k+1}}(\boldsymbol{x}_{(1)}), \ldots, D_{F_d,\mathcal{B}_{d,k+1}}(\boldsymbol{x}_{(d)})\right)^\top\right), \end{aligned}$$

where $\hat{\phi}(\boldsymbol{y}) := (\phi(\boldsymbol{y}_{(1)}), \dots, \phi(\boldsymbol{y}_{(d)}))^\top$, $\phi : \mathbb{R}^n \to \mathbb{R}$ is convex and admits a unique minimizer at $\mathbf{0}$. Notice that $(F_i(\cdot, \boldsymbol{s}_l), D_{F_i(\cdot,\boldsymbol{s}_l)})$ has a $\mathcal{C}^r$ variational stratification (Bolte and Pauwels, 2021, Theorem 4), and $\hat{\phi}$ is differentiable on a full-measure subset of $\mathbb{R}^{nd}$, thus the mappings $\{\mathcal{U}_{j,k}\}, j = 1, 2$ are almost everywhere $\mathcal{C}^1$ over $\mathbb{R}^{2nd}$, the corresponding almost-everywhere-$\mathcal{C}^1$ mappings are denoted by $u_{j,k}(\boldsymbol{z}), j = 1, 2$. Furthermore, let

$$\mathcal{Y} := \Big\{(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^{2nd} : \hat{\phi}(\boldsymbol{y}) \text{ is not differentiable }, D_{F_{i,l}}(\boldsymbol{x}) \neq \{\nabla F_i(\boldsymbol{x}, \boldsymbol{s}_l)\}, \text{ for some } i \in [N] \text{ and } \boldsymbol{s}_l \in \mathcal{S}_i\Big\},$$

which is a zero-measure subset in $\mathbb{R}^{2nd}$.

With the same notations in Section A.1, the vectorized sequence $\{(\boldsymbol{x}_k, \boldsymbol{y}_k)\}$ generated by Algorithm 2 satisfies

$$\begin{aligned} (\boldsymbol{x}_{k+\frac{1}{2}}, \boldsymbol{y}_{k+\frac{1}{2}}) &\in (\boldsymbol{W} \otimes \boldsymbol{I}_n \times \boldsymbol{I}_{nd})(\boldsymbol{x}_k, \boldsymbol{y}_k) - c\hat{\eta}_k \mathcal{U}_{1,k}(\boldsymbol{x}_k, \boldsymbol{y}_k), \\ (\boldsymbol{x}_{k+1}, \boldsymbol{y}_{k+1}) &\in (\boldsymbol{I}_{nd} \times \boldsymbol{W} \otimes \boldsymbol{I}_n)(\boldsymbol{x}_{k+\frac{1}{2}}, \boldsymbol{y}_{k+\frac{1}{2}}) - c\hat{\eta}_k \mathcal{U}_{2,k}(\boldsymbol{x}_{k+\frac{1}{2}}, \boldsymbol{y}_{k+\frac{1}{2}}). \end{aligned} \tag{A.3}$$

Set $\boldsymbol{P}_1 := \boldsymbol{W} \otimes \boldsymbol{I}_n \times \boldsymbol{I}_{nd}$, $\boldsymbol{P}_2 := \boldsymbol{I}_{nd} \times \boldsymbol{W} \otimes \boldsymbol{I}_n$, then the update scheme of vectorized sequence $\{(\boldsymbol{x}_k, \boldsymbol{y}_k)\}$ is

$$\begin{aligned} (\boldsymbol{x}_{k+1}, \boldsymbol{y}_{k+1}) &= \boldsymbol{P}_2(\boldsymbol{P}_1(\boldsymbol{x}_k, \boldsymbol{y}_k)) - c\boldsymbol{P}_2(\hat{\eta}_k \mathcal{U}_{1,k}(\boldsymbol{x}_k, \boldsymbol{y}_k)) - c\hat{\eta}_k \mathcal{U}_{2,k}(\boldsymbol{P}_1(\boldsymbol{x}_k, \boldsymbol{y}_k) - c\hat{\eta}_k \mathcal{U}_{1,k}(\boldsymbol{x}_k, \boldsymbol{y}_k)) \\ &:= \boldsymbol{P}(\boldsymbol{x}_k, \boldsymbol{y}_k) - c\mathcal{Q}_k(\boldsymbol{x}_k, \boldsymbol{y}_k), \end{aligned}$$

where $\mathcal{Q}_k$ is almost everywhere $\mathcal{C}^1$ over $\mathbb{R}^{2nd}$.

The Jacobian of mapping $\boldsymbol{P}_1(\boldsymbol{x}, \boldsymbol{y}) - c\hat{\eta}_k u_{1,k}(\boldsymbol{x}, \boldsymbol{y})$ is

$$\boldsymbol{J}_{1c}(\boldsymbol{x}, \boldsymbol{y}) = \begin{pmatrix} \boldsymbol{W} \otimes \boldsymbol{I}_n & 0 \\ 0 & \boldsymbol{I}_{nd} \end{pmatrix} - c\hat{\eta}_k \begin{pmatrix} 0 & \nabla^2 \phi(\boldsymbol{y}) \\ 0 & 0 \end{pmatrix},$$

where $\nabla^2 \phi(\boldsymbol{y})$ is well-defined over a full-measure set. And the Jacobian of the mapping $\boldsymbol{P}_2(\boldsymbol{x}, \boldsymbol{y}) - c\hat{\eta}_k u_{2,k}(\boldsymbol{x}, \boldsymbol{y})$ is

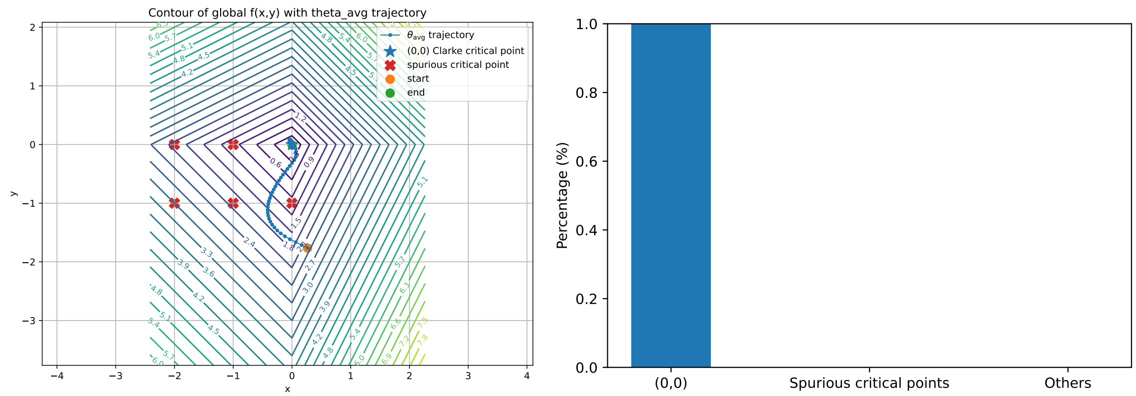
$$\boldsymbol{J}_{2c}(\boldsymbol{x}, \boldsymbol{y}) = \begin{pmatrix} \boldsymbol{I}_{nd} & 0 \\ 0 & \boldsymbol{W} \otimes \boldsymbol{I}_n \end{pmatrix} - c\hat{\eta}_k \begin{pmatrix} 0 & 0 \\ \tau \mathrm{Diag}(\nabla^2 F_{1,\mathcal{B}_{1,k+1}}(\boldsymbol{x}_{(1)}), \dots, \nabla^2 F_{d,\mathcal{B}_{d,k+1}}(\boldsymbol{x}_{(d)})) & -\tau \boldsymbol{W} \otimes \boldsymbol{I}_n \end{pmatrix}.$$

Henceforth, the Jacobian of $\boldsymbol{P}(\boldsymbol{x}, \boldsymbol{y}) - c\mathcal{Q}_k(\boldsymbol{x}, \boldsymbol{y})$ is $\boldsymbol{J}_{2c}(\boldsymbol{x}, \boldsymbol{y})\boldsymbol{J}_{1c}(\boldsymbol{x}, \boldsymbol{y})$.

When $\boldsymbol{W}$ is invertible, then the determinant of $\boldsymbol{J}_2(\boldsymbol{x}, \boldsymbol{y})\boldsymbol{J}_1(\boldsymbol{x}, \boldsymbol{y})$ is a non-trivial polynomial of c for almost everywhere $(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^{2nd}$. Leveraging similar proof techniques in Section A.1, we can derive that there exists a full-measure subset $\hat{\mathcal{S}} \subseteq \mathbb{R}_+$, such that for any $c \in \hat{\mathcal{S}}$, there exists a full-measure subset $\hat{\mathcal{K}}$ of $\mathbb{R}^{nd} \times \mathbb{R}^{nd}$ with the property that, whenever we choose $(\boldsymbol{x}_0, \boldsymbol{y}_0) \in \hat{\mathcal{K}}$, it satisfies that $\{(\boldsymbol{x}_k, \boldsymbol{y}_k)\} \cap \mathcal{Y} = \varnothing$. Consequently, the update scheme of vectorized sequence $(\boldsymbol{x}_k, \boldsymbol{y}_k)$ can be reformulated as

$$\begin{cases} \boldsymbol{x}_{k+1} \in (\boldsymbol{W} \otimes \boldsymbol{I}_n)\boldsymbol{x}_k - c\hat{\eta}_k \nabla \phi(\boldsymbol{y}_k), \\ \boldsymbol{g}_{k+1} = \left(\nabla F_{1,\mathcal{B}_{1,k+1}}(\boldsymbol{x}_{1,k+1}), \dots, \nabla F_{d,\mathcal{B}_{d,k+1}}(\boldsymbol{x}_{d,k+1})\right)^\top, \\ \boldsymbol{y}_{k+1} \in (1 - c\tau\hat{\eta}_k)(\boldsymbol{W} \otimes \boldsymbol{I}_n)\boldsymbol{y}_k + c\tau\hat{\eta}_k \boldsymbol{g}_{k+1}. \end{cases}$$

Therefore, by applying Theorem 4.7 with $D_f = \partial f$, we can choose $\mathcal{S} = \hat{\mathcal{S}} \cap (0, \alpha_c), j = 1, 2$ and $\mathcal{K} = \hat{\mathcal{K}} \cap (\mathcal{X}_0 \times \mathbb{R}^{nd})$, such that the assertion in Theorem 4.8 holds.



(a) Iterate trajectory

(b) The percentages on different types of the final results

Figure 9: Numerical results of Algorithm 2 on a toy example.

Appendix B. Synthetic Numerical Examples on Avoiding Spurious Critical Points

In this part, we present numerical experiments on a synthetic example to illustrate that, under random initialization, Algorithm 2 converges to a Clarke-critical point with probability 1.

We consider a four-node ring network with a non-singular mixing matrix

$$W = \frac{1}{3}\begin{pmatrix} 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \end{pmatrix}.$$

Moreover, the local objective functions are defined as

$$\begin{aligned} f_1(x,y) &= \mathrm{ReLU}(x) + \mathrm{ReLU}(y), \\ f_2(x,y) &= \mathrm{ReLU}(x) + \mathrm{ReLU}(y), \\ f_3(x,y) &= \mathrm{ReLU}(-x) + \mathrm{ReLU}(-y), \\ f_4(x,y) &= \mathrm{ReLU}(-x-1) + (x+1) - \mathrm{ReLU}(x+1) + \mathrm{ReLU}(-x-2) + (x+2) - \mathrm{ReLU}(x+2) \\ &\quad + \mathrm{ReLU}(-y-1) + (y+1) - \mathrm{ReLU}(y+1), \end{aligned}$$

respectively.

Note that $f_4(x,y)$ is identically zero. However, by following the same theoretical analysis in Bolte et al. (2022b), automatic differentiation (AD) yields $\partial_x f_4(-1,y) = 1, \partial_x f_4(-2,y) = 1$, and $\partial_y f_4(x,-1) = 1$ for any $x \in \mathbb{R}$. As a result, AD introduces spurious stationary points for $f = f_1 + f_2 + f_3 + f_4$, which are

$$(-1,0),\ (0,-1),\ (-2,0),\ (-1,-1),\ (-2,-1).$$

In our experiments, we randomly initialize $(x_0, y_0) \sim \mathcal{N}((0,0), I_2)$ in Algorithm 2, and in each epoch we compute only one full-batch local gradient using AD. Figure 9 reports a randomly sampled iterate trajectory and the percentages over 100 independent runs that converge to different critical points. A run is declared to converge to a critical point if the squared distance from its final iterate to that point is below 10^{-4}. As shown in Figure 9, almost all runs generated by Algorithm 2 converge to the true Clarke-critical point $(0, 0)$, which is consistent with Theorem 4.8.

Appendix C. Developing Decentralized Generalized ADAM Methods with Convergence Guarantees

Algorithm 4 DGAdam for solving (DOP).

Input: Initial point $\boldsymbol{x}_0 \in \mathbb{R}^n$, initial momentum terms $\boldsymbol{m}_0, \boldsymbol{v}_0 \in \mathbb{R}^n$, momentum parameters $\tau_1, \tau_2 > 0$, an auxiliary function $\phi : \mathbb{R}^n \to \mathbb{R}$, and a mixing matrix $\boldsymbol{W}$.

for all $i \in [d]$ in parallel **do**
  Set $k \leftarrow 0$. Initialize $\boldsymbol{x}_{i,k} = \boldsymbol{x}_0, \boldsymbol{m}_{i,k} = \boldsymbol{m}_0$;
  while not terminated **do**
    Choose the step-size η_k;
    Communicate and update the local variable

$$\boldsymbol{x}_{i,k+1} \in \sum_{j \in \mathcal{N}_i} \boldsymbol{W}(i,j)\boldsymbol{x}_{j,k} - \eta_k \frac{1}{\sqrt{\boldsymbol{v}_{i,k} + \epsilon}} \odot (\boldsymbol{m}_{i,k} + \rho \boldsymbol{d}_{i,k});$$

    Compute $\boldsymbol{d}_{i,k+1} \in \frac{1}{|\mathcal{B}_{i,k+1}|} \sum_{\boldsymbol{s}_l \in \mathcal{B}_{i,k+1}} D_{F_{i,l}}(\boldsymbol{x}_{i,k+1})$;
    Randomly select a mini-batch $\mathcal{B}_{i,k} \subseteq \mathcal{S}_i$;
    Communicate and update momentum

$$\boldsymbol{m}_{i,k+1} = (1 - \tau_1 \eta_k) \sum_{j \in \mathcal{N}_i} \boldsymbol{W}(i,j)\boldsymbol{m}_{j,k} + \tau_1 \eta_k \boldsymbol{d}_{i,k+1};$$

    Communicate and update second-moment estimate by one of the formats

$$\boldsymbol{v}_{i,k+1} \in (1 - \tau_2 \eta_k) \sum_{j \in \mathcal{N}_i} \boldsymbol{W}(i,j)\boldsymbol{v}_{j,k} + \tau_2 \eta_k \boldsymbol{d}_{i,k+1} \odot \boldsymbol{d}_{i,k+1}; \qquad \triangleleft \text{ Adam}$$

$$\boldsymbol{v}_{i,k+1} \in (1 - \tau_2 \eta_k) \sum_{j \in \mathcal{N}_i} \boldsymbol{W}(i,j)\boldsymbol{v}_{j,k} + \tau_2 \eta_k \mathrm{Diag}(\frac{\mathbf{1}\mathbf{1}^\top}{b_{i,1}}, \ldots, \frac{\mathbf{1}\mathbf{1}^\top}{b_{i,s}})\boldsymbol{d}_{i,k+1} \odot \boldsymbol{d}_{i,k+1}; \triangleleft \text{ Adam-mini}$$

    Set $k \leftarrow k + 1$;
  end while
end for
return $\boldsymbol{X}_k := [\boldsymbol{x}_{1,k}, \ldots, \boldsymbol{x}_{d,k}]$.

Adam (Kingma and Ba, 2014) uses a momentum term and an exponential moving average of squared gradients as an adaptive scaling factor, thereby applying coordinate-wise adaptive step sizes and updating parameters with bias correction. After this work, a number of efficient variants are developed, such as AdaBelief (Zhuang et al., 2020), AMSGrad

(Reddi et al., 2019), decentralized AMSGrad (Chen et al., 2023b), etc. In the single-agent setting, (Xiao et al., 2024; Ding et al., 2023) investigate the global convergence properties of Adam-family methods for nonsmooth objectives via a delicately constructed Lyapunov function.

In this part, we present a decentralized generalized Adam method in Algorithm 4, which fits into (DSM). Notice $b_{i,1}, \ldots, b_{i,s}$ in Line 9 represent the block sizes in Adam-mini at node i.

Let $\boldsymbol{M}_k := [\boldsymbol{m}_{1,k}, \ldots, \boldsymbol{m}_{d,k}]$, $\boldsymbol{V}_k := [\boldsymbol{v}_{1,k}, \ldots, \boldsymbol{v}_{d,k}]$. We have the following theorem illustrating the convergence properties of Algorithm 4.

Theorem C.1 *Suppose Assumption 4.1 holds. $\{\boldsymbol{X}_k\}$ is generated by Algorithm 4.*

(1) When Assumption 4.2 holds, there exists $\alpha_c > 0$, for any $c \in (0, \alpha_c)$, it holds that almost surely, any cluster point of $\{\boldsymbol{X}_k\}$ is a D_f-critical point of (DOP) *and $\{f(\boldsymbol{x}_{i,k}) : k \in \mathbb{N}\}$ converges for any $i \in [d]$.*

(2) When Assumption 4.3 holds, for any $\varepsilon > 0$, there exists $\alpha_c > 0$, for any $c \in (0, \alpha_c)$,

$$\mathbb{P}\left(\lim_{k\to\infty} \operatorname{dist}(\boldsymbol{X}_k, \{\boldsymbol{X} \in \mathbb{R}^{n\times d} : \boldsymbol{X} = \boldsymbol{x}\boldsymbol{1}^\top, \boldsymbol{0} \in D_f(\boldsymbol{x})\}) = 0\right) \geq 1 - \varepsilon,$$

and

$$\mathbb{P}(f(\boldsymbol{x}_{i,k}) \text{ converges }) \geq 1 - \varepsilon,\ \forall i \in [d].$$

Proof The compact format of Algorithm 4 with Adam update can be written as

$$\begin{aligned}
\boldsymbol{X}_{k+1} &\in \boldsymbol{X}_k\boldsymbol{W} - \eta_k \frac{1}{\sqrt{\boldsymbol{V}_k + \epsilon \boldsymbol{I}}} \odot (\boldsymbol{M}_k + \rho \boldsymbol{D}_k),\\
\boldsymbol{D}_{k+1} &= [\boldsymbol{d}_{1,k+1}, \ldots, \boldsymbol{d}_{d,k+1}],\\
\boldsymbol{M}_{k+1} &= (1 - \tau_1\eta_k)\boldsymbol{M}_k\boldsymbol{W} + \tau_1\eta_k\boldsymbol{D}_{k+1},\\
\boldsymbol{V}_{k+1} &\in (1 - \tau_2\eta_k)\boldsymbol{V}_k\boldsymbol{W} + \tau_2\eta_k\boldsymbol{D}_{k+1} \odot \boldsymbol{D}_{k+1}.
\end{aligned} \tag{C.1}$$

However, if we directly define $\boldsymbol{Z}_k := \begin{bmatrix} \boldsymbol{X}_k \\ \boldsymbol{M}_k \\ \boldsymbol{V}_k \end{bmatrix}$, we cannot find a function Φ_i that satisfies Assumption 3.1. To address this issue, we introduce an auxiliary update scheme to Algorithm 4 parameterized by $K > 0$.

$$\begin{aligned}
\boldsymbol{X}_{k+1} &\in \boldsymbol{X}_k\boldsymbol{W} - \eta_k \frac{1}{\sqrt{\boldsymbol{V}_k + \epsilon \boldsymbol{I}}} \odot (\boldsymbol{M}_k + \rho \boldsymbol{D}_k),\\
\boldsymbol{D}_{k+1} &= [\boldsymbol{d}_{1,k+1}, \ldots, \boldsymbol{d}_{d,k+1}],\\
\boldsymbol{M}_{k+1} &= (1 - \tau_1\eta_k)\boldsymbol{M}_k\boldsymbol{W} + \tau_1\eta_k\boldsymbol{D}_{k+1},\\
\frac{1}{d}\boldsymbol{V}_{k+1}\boldsymbol{1}_d\boldsymbol{1}_d^\top &\in (1 - \tau_2\eta_k)\frac{1}{d}\boldsymbol{V}_{k+1}\boldsymbol{1}_d\boldsymbol{1}_d^\top\boldsymbol{W} + \tau_2\eta_k\frac{1}{d}\boldsymbol{D}_{k+1} \odot \boldsymbol{D}_{k+1}\boldsymbol{1}_d\boldsymbol{1}_d^\top \mathbb{1}_K(\boldsymbol{V}_k),
\end{aligned} \tag{C.2}$$

where $\mathbb{1}_K$ denotes the indication function of the subset $\{\boldsymbol{V} : \|\boldsymbol{V}\| \leq K\}$. Denote the filtration $\mathcal{F}_k := \sigma(\{\boldsymbol{X}_l : l \leq k\})$, $D_{F_{\mathcal{B}_{i,k+1}}}(\boldsymbol{x}_{i,k+1}) := \frac{1}{|\mathcal{B}_{i,k+1}|}\sum_{\boldsymbol{s}_l \in \mathcal{B}_{i,k+1}} D_{F_{i,l}}(\boldsymbol{x}_{i,k+1})$, and let

$$\mathcal{V}_i(\boldsymbol{x}) := \begin{cases} \frac{1}{N}\sum_{k=0}^{N-1} D_{F_{\mathcal{B}_{i,k+1}}}(\boldsymbol{x}) \odot D_{F_{\mathcal{B}_{i,k+1}}}(\boldsymbol{x}) & \text{for case (1)},\\ \mathbb{E}[D_{F_{\mathcal{B}_{i,k+1}}}(\boldsymbol{x}) \odot D_{F_{\mathcal{B}_{i,k+1}}}(\boldsymbol{x}) | \mathcal{F}_k] & \text{for case (2)}, \end{cases}$$

and

$$\boldsymbol{Z}_k := \begin{bmatrix} \boldsymbol{X}_k \\ \boldsymbol{M}_k \\ \frac{1}{d}\boldsymbol{V}_k \mathbf{1}_d \mathbf{1}_d^\top \end{bmatrix}, \boldsymbol{H}_k := \mathbb{E}\left[\begin{bmatrix} \frac{1}{\sqrt{\boldsymbol{V}_k + \epsilon \boldsymbol{I}}} \odot (\boldsymbol{M}_k + \rho \boldsymbol{D}_k) \\ \tau_1 \boldsymbol{M}_k \boldsymbol{W} - \tau_1 \boldsymbol{D}_{k+1} \\ \tau_2 \frac{1}{d}\boldsymbol{V}_k \mathbf{1}_d \mathbf{1}_d^\top - \tau_2 \frac{1}{d}\boldsymbol{D}_{k+1} \odot \boldsymbol{D}_{k+1} \mathbf{1}_d \mathbf{1}_d^\top \mathbb{1}_K(\boldsymbol{V}_k) \end{bmatrix} |\mathcal{F}_k\right],$$

$$\Xi_{k+1} := \begin{bmatrix} \frac{1}{\sqrt{\boldsymbol{V}_k + \epsilon \boldsymbol{I}}} \odot (\boldsymbol{M}_k + \rho \boldsymbol{D}_k) \\ \tau_1 \boldsymbol{M}_k \boldsymbol{W} - \tau_1 \boldsymbol{D}_{k+1} \\ \tau_2 \frac{1}{d}\boldsymbol{V}_k \mathbf{1}_d \mathbf{1}_d^\top - \tau_2 \frac{1}{d}\boldsymbol{D}_{k+1} \odot \boldsymbol{D}_{k+1} \mathbf{1}_d \mathbf{1}_d^\top \mathbb{1}_K(\boldsymbol{V}_k) \end{bmatrix} - \boldsymbol{H}_k.$$

Then (C.2) can be reshaped as $\boldsymbol{Z}_{k+1} \in \boldsymbol{Z}_k \boldsymbol{W} - \eta_k(\boldsymbol{H}_k + \Xi_{k+1})$, which aligns with the update format of (DSM).

From the local boundedness of D_{f_i}, we can conclude that $\{\boldsymbol{D}_k\}$ is uniformly bounded almost surely whenever $\{\boldsymbol{Z}_k\}$ is bounded, which further leads to the local boundedness of $\boldsymbol{H}_k$ and Ξ_{k+1}, that is, Assumption 3.1(2) holds. Moreover, we set

$$\Phi_i^{(K)}(\boldsymbol{x}, \boldsymbol{m}, \boldsymbol{v}) := \begin{bmatrix} (|\boldsymbol{v}| + \epsilon)^{-1/2} \odot (\boldsymbol{m} + \rho D_{f_i}(\boldsymbol{x})) \\ \tau_1 \boldsymbol{m} - \tau_1 D_{f_i}(\boldsymbol{x}) \\ \tau_2 \boldsymbol{v} - \tau_2 \mathcal{V}_i(\boldsymbol{x}) \mathbb{1}_K(\boldsymbol{v}) \end{bmatrix}.$$

(1.5) holds vacuously for case (1) with a sufficiently large nonnegative δ_k. Notice that $\mathbb{E}[\frac{1}{d}\boldsymbol{D}_{k+1} \odot \boldsymbol{D}_{k+1}\mathbf{1}_d|\mathcal{F}_k] \in \frac{1}{d}\sum_{i=1}^d \mathcal{V}_i^{\|\boldsymbol{x}_{i,k} - \boldsymbol{x}_{i,k+1}\|}(\boldsymbol{x}_{i,k})$. We set δ_k to be an exponential moving average

$$\begin{aligned} \delta_k = {} & \eta_k \frac{\|\boldsymbol{M}_k + \rho \boldsymbol{D}_k\|}{\sqrt{\epsilon}} + 2\sum_{j=0}^{k-1} \lambda_2^{k-1-j} \eta_j \frac{\|\boldsymbol{M}_j + \rho \boldsymbol{D}_j\|}{\sqrt{\epsilon}} \\ & + \tau_2 \eta_{k-1} \|\boldsymbol{D}_k \odot \boldsymbol{D}_k\| + \sum_{j=0}^{k-1} (\Pi_{i=0}^j (1 - \tau_2 \eta_{k-1-i})) \lambda_2^{j+1} \tau_2 \eta_{k-j-2} \|\boldsymbol{D}_{k-1-j} \odot \boldsymbol{D}_{k-j-1}\|, \end{aligned} \tag{C.3}$$

for case (2). Then, we attain that

$$\Phi^{(K)}(\boldsymbol{x}, \boldsymbol{m}, \boldsymbol{v}) = \begin{bmatrix} (|v| + \epsilon)^{-1/2} \odot (m + \rho D_f(\boldsymbol{x})) \\ \tau_1 \boldsymbol{m} - \tau_1 D_f(\boldsymbol{x}) \\ \tau_2 \boldsymbol{v} - \tau_2 \mathrm{conv}\left(\frac{1}{d}\sum_{i=1}^d \mathcal{V}_i(\boldsymbol{x})\right) \mathbb{1}_K(\boldsymbol{v}) \end{bmatrix}.$$

Since set-valued mapping conv $\left(\frac{1}{d}\sum_{i=1}^d \mathcal{V}_i(\boldsymbol{x})\right)$ is a nonnegative, convex compact valued, and graph-closed, by Proposition 5.3 in Xiao et al. (2023a), we have $\psi^{(K)}(\boldsymbol{x}, \boldsymbol{m}, \boldsymbol{v}) = f(\boldsymbol{x}) + \frac{1}{2\tau_1}\langle \boldsymbol{m}, (|\boldsymbol{v}| + \epsilon)^{-1/2} \odot \boldsymbol{m}\rangle + K\max\{0, \|\boldsymbol{v}\| - K\}$ is a coercive Lyapunov function of differential inclusion $(\frac{d\boldsymbol{x}}{dt}, \frac{d\boldsymbol{m}}{dt}, \frac{d\boldsymbol{v}}{dt}) \in -\Phi^{(K)}(\boldsymbol{x}, \boldsymbol{m}, \boldsymbol{v})$, with stable set $\mathcal{A}_K = \{(\boldsymbol{x}, \boldsymbol{m}, \boldsymbol{v}) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n : \mathbf{0} \in D_f(\boldsymbol{x}), \boldsymbol{m} = 0, \|\boldsymbol{v}\| \leq K\}$, which verifies Assumption 3.1-(3). Moreover, Assumption 3.1-(4) directly follows from the definability of f and $\mathcal{A}_K = \{(\boldsymbol{x}, \boldsymbol{m}, \boldsymbol{v}) : \mathbf{0} \in D_{\psi^{(K)}}(\boldsymbol{x}, \boldsymbol{m}, \boldsymbol{v})\}$.

For case (1), we set $\mathcal{U}_{i,j}(\boldsymbol{x}, \boldsymbol{m}, \boldsymbol{v}) := \begin{bmatrix} (|\boldsymbol{v}| + \epsilon)^{-1/2} \odot (\boldsymbol{m} + \rho D_{F_{\mathcal{B}_{i,j}}}(\boldsymbol{x})) \\ \tau_1 \boldsymbol{m} - \tau_1 D_{F_{\mathcal{B}_{i,j}}}(\boldsymbol{x})) \\ \tau_2 \boldsymbol{v} - \tau_2 D_{F_{\mathcal{B}_{i,j+1}}}(\boldsymbol{x}) \odot D_{F_{\mathcal{B}_{i,j+1}}}(\boldsymbol{x}) \mathbb{1}_K(\boldsymbol{v}) \end{bmatrix}$, where $\mathcal{B}_{i,j}, 0 \leq j \leq N-1$ are subsets of equal size. We also set ρ_k similar to δ_k in (C.3). Then (C.2) is a special form of (DSM) with (3.8). Assumption 4.2 togather with the fact $\frac{1}{N}\sum_{i=0}^{N-1} \mathcal{U}_{i,j}(\boldsymbol{x}, \boldsymbol{m}, \boldsymbol{v}) =$

$\Phi_i^{(K)}(\boldsymbol{x}, \boldsymbol{m}, \boldsymbol{v})$ implies that Assumption 3.9 is satisfied for case (1). For case (2), we choose $\xi_{i,k} := \{\textit{Draw out } \mathcal{B}_{i,k} \textit{ from agent } i\}$, and

$$\chi_i(\boldsymbol{x}, \boldsymbol{m}, \boldsymbol{v}, \xi_{i,k}) \in \begin{bmatrix} (|v|+\epsilon)^{-1/2} \odot (\rho D_{F_{\mathcal{B}_{i,k}}}(\boldsymbol{x}) - \rho D_{f_i}(\boldsymbol{x}) \\ \tau_1 D_{f_i}(\boldsymbol{x}) - \tau_1 D_{F_{\mathcal{B}_{i,k}}}(\boldsymbol{x}) \\ \tau_2 \frac{1}{d} \sum_{i=1}^d \mathcal{V}_i(\boldsymbol{x}) \mathbb{1}_K(\boldsymbol{v}) - \tau_2 \frac{1}{d} \sum_{i=1}^d D_{F_{\mathcal{B}_{i,k}}}(\boldsymbol{x}) \odot D_{F_{\mathcal{B}_{i,k}}}(\boldsymbol{x}) \mathbb{1}_K(\boldsymbol{v}) \end{bmatrix}.$$

(C.2) is a particular form of (DSM) with (3.18), and Assumption 3.13 is satisfied for case (2). Plugging Lemma 3.7 and Theorem 3.15, we can prove that $\delta_k, \rho_k \to 0$, as $k \to \infty$ with arbitrary high probability.

Applying Theorem 3.11 and 3.15, we derive that (C.2) is restricted in some bounded set for case (1) and high-probability bounded for case (2) for any $K > 0$, which implies that Algorithm 4 coincides with (C.2) for some sufficiently large K. Employing Theorem 3.12 and 3.16, we get the desired results in Theorem C.1.

For decentralized Adam-mini, we denote $\Sigma_i := \mathrm{Diag}(\frac{\mathbf{1}\mathbf{1}^\top}{b_{i,1}}, \ldots, \frac{\mathbf{1}\mathbf{1}^\top}{b_{i,s}})$, and replace the update schemes of $\frac{1}{d}\boldsymbol{V}_{k+1}\mathbf{1}_d\mathbf{1}_d^\top$ and $\mathcal{V}_i(x)$ by

$$\frac{1}{d}\boldsymbol{V}_{k+1}\mathbf{1}_d\mathbf{1}_d^\top \in (1-\tau_2\eta_k)\frac{1}{d}\boldsymbol{V}_{k+1}\mathbf{1}_d\mathbf{1}_d^\top\boldsymbol{W} + \tau_2\eta_k\frac{1}{d}[\Sigma_1\boldsymbol{d}_{1,k+1}\odot\boldsymbol{d}_{1,k+1}, \ldots, \Sigma_d\boldsymbol{d}_{d,k+1}\odot\boldsymbol{d}_{d,k+1}]\mathbf{1}_d\mathbf{1}_d^\top\mathbb{1}_K(\boldsymbol{V}_k),$$

and

$$\mathcal{V}_i(\boldsymbol{x}) := \begin{cases} \frac{1}{N}\sum_{k=0}^{N-1}\Sigma_i D_{F_{\mathcal{B}_{i,k+1}}}(\boldsymbol{x}) \odot D_{F_{\mathcal{B}_{i,k+1}}}(\boldsymbol{x}) & \text{for case (1),} \\ \mathbb{E}[\Sigma_i D_{F_{\mathcal{B}_{i,k+1}}}(\boldsymbol{x}) \odot D_{F_{\mathcal{B}_{i,k+1}}}(\boldsymbol{x}) | \mathcal{F}_k] & \text{for case (2).} \end{cases}$$

Correspondingly, we add Σ_i before the term $D_{F_{\mathcal{B}_{i,k}}}(\boldsymbol{x}) \odot D_{F_{\mathcal{B}_{i,k}}}(\boldsymbol{x})$ in the expression of $\mathcal{U}_{i,j}$ and χ_i. Then, using the same approach as in the above proof of decentralized Adam, we can obtain the desired results.

■

References

Michel Benaïm. Dynamics of stochastic approximation algorithms. In *Seminaire de probabilites XXXIII*, pages 1–68. Springer, 2006.

Michel Benaïm, Josef Hofbauer, and Sylvain Sorin. Stochastic approximations and differential inclusions. *SIAM Journal on Control and Optimization*, 44(1):328–348, 2005.

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. Signsgd: Compressed optimisation for non-convex problems. In *International Conference on Machine Learning*, pages 560–569. PMLR, 2018.

Wei Bian and Xiaoping Xue. Subgradient-based neural networks for nonsmooth nonconvex optimization problems. *IEEE Transactions on Neural Networks*, 20(6):1024–1038, 2009.

Pascal Bianchi and Rodolfo Rios-Zertuche. A closed-measure approach to stochastic approximation. *arXiv preprint arXiv:2112.05482*, 2021.

Pascal Bianchi, Walid Hachem, and Adil Salim. Constant step stochastic approximations involving differential inclusions: stability, long-run convergence and applications. *Stochastics*, 91(2):288–320, 2019.

Pascal Bianchi, Walid Hachem, and Sholom Schechtman. Convergence of constant step stochastic gradient descent for non-smooth non-convex functions. *Set-Valued and Variational Analysis*, pages 1–31, 2022.

Saverio Bolognani, Ruggero Carli, Guido Cavraro, and Sandro Zampieri. Distributed reactive power feedback control for voltage regulation and loss minimization. *IEEE Transactions on Automatic Control*, 60(4):966–981, 2014.

Jérôme Bolte and Edouard Pauwels. A mathematical model for automatic differentiation in machine learning. *Advances in Neural Information Processing Systems*, 33:10809–10819, 2020.

Jérôme Bolte and Edouard Pauwels. Conservative set valued fields, automatic differentiation, stochastic gradient methods and deep learning. *Mathematical Programming*, 188(1): 19–51, 2021.

Jérôme Bolte, Aris Daniilidis, Adrian Lewis, and Masahiro Shiota. Clarke subgradients of stratifiable functions. *SIAM Journal on Optimization*, 18(2):556–572, 2007.

Jérôme Bolte, Tam Le, Edouard Pauwels, and Tony Silveti-Falls. Nonsmooth implicit differentiation for machine-learning and optimization. *Advances in neural information processing systems*, 34:13537–13549, 2021.

Jérôme Bolte, Edouard Pauwels, and Rodolfo Rios-Zertuche. Long term dynamics of the subgradient method for lipschitz path differentiable functions. *Journal of the European Mathematical Society*, 2022a.

Jerome Bolte, Edouard Pauwels, and Samuel Vaiter. Automatic differentiation of nonsmooth iterative algorithms. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 26404–26417. Curran Associates, Inc., 2022b. URL `https://proceedings.neurips.cc/paper_files/paper/2022/file/a9077da44185792cb63599cc9e0357bc-Paper-Conference.pdf`.

Jérôme Bolte, Tam Le, Éric Moulines, and Edouard Pauwels. Inexact subgradient methods for semialgebraic functions. *arXiv preprint arXiv:2404.19517*, 2024.

Jérôme Bolte, Tam Le, Eric Moulines, and Edouard Pauwels. Inexact subgradient methods for semialgebraic functions: J. bolte et al. *Mathematical Programming*, pages 1–27, 2025.

Vivek S Borkar. *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer, 2009.

Vivek S Borkar and Vivek S Borkar. *Stochastic approximation: a dynamical systems viewpoint*, volume 9. Springer, 2008.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM review*, 60(2):223–311, 2018.

Stephen Boyd, Persi Diaconis, and Lin Xiao. Fastest mixing markov chain on a graph. *SIAM review*, 46(4):667–689, 2004.

Dmitry Bylinkin and Aleksandr Beznosikov. Accelerated methods with compressed communications for distributed optimization problems under data similarity. *ArXiv*, abs/2412.16414, 2024. URL `https://api.semanticscholar.org/CorpusID:274982476`.

Camille Castera, Jérôme Bolte, Cédric Févotte, and Edouard Pauwels. An inertial newton algorithm for deep learning. *The Journal of Machine Learning Research*, 22(1):5977–6007, 2021.

Shixiang Chen, Alfredo Garcia, and Shahin Shahrampour. On distributed nonconvex optimization: Projected subgradient method for weakly convex problems in networks. *IEEE Transactions on Automatic Control*, 67(2):662–675, 2021.

Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Yao Liu, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, et al. Symbolic discovery of optimization algorithms. *arXiv preprint arXiv:2302.06675*, 2023a.

Xiangyi Chen, Belhal Karimi, Weijie Zhao, and Ping Li. On the convergence of decentralized adaptive gradient methods. In *Asian Conference on Machine Learning*, pages 217–232. PMLR, 2023b.

Frank H Clarke. *Optimization and nonsmooth analysis*, volume 5. SIAM, 1990.

Kobi Cohen, Angelia Nedić, and Rayadurgam Srikant. Distributed learning algorithms for spectrum sharing in spatial random access wireless networks. *IEEE Transactions on Automatic Control*, 62(6):2854–2869, 2016.

Laurent Condat, Artavazd Maranjyan, and Peter Richtárik. Locodl: Communication-efficient distributed learning with local training and compression. *arXiv preprint arXiv:2403.04348*, 2024.

Damek Davis, Dmitriy Drusvyatskiy, Sham Kakade, and Jason D Lee. Stochastic subgradient method converges on tame functions. *Foundations of Computational Mathematics*, 20 (1):119–154, 2020.

Christopher M De Sa. Random reshuffling is not always better. *Advances in Neural Information Processing Systems*, 33:5957–5967, 2020.

Paolo Di Lorenzo and Simone Scardapane. Distributed stochastic nonconvex optimization and learning based on successive convex approximation. In *2019 53rd Asilomar Conference on Signals, Systems, and Computers*, pages 1–5. IEEE, 2019.

Paolo Di Lorenzo and Gesualdo Scutari. Next: In-network nonconvex optimization. *IEEE Transactions on Signal and Information Processing over Networks*, 2(2):120–136, 2016.

Kuangyu Ding, Nachuan Xiao, and Kim-Chuan Toh. Adam-family methods with decoupled weight decay in deep learning. *arXiv preprint arXiv:2310.08858*, 2023.

John C Duchi and Feng Ruan. Stochastic methods for composite and weakly convex optimization problems. *SIAM Journal on Optimization*, 28(4):3229–3259, 2018.

Hongchang Gao and Heng Huang. Periodic stochastic gradient descent with momentum for decentralized training. *arXiv preprint arXiv:2008.10435*, 2020.

Juan Gao, Xin-Wei Liu, Yu-Hong Dai, Yakui Huang, and Junhua Gu. Distributed stochastic gradient tracking methods with momentum acceleration for non-convex optimization. *Computational Optimization and Applications*, 84(2):531–572, 2023.

Jemin George, Tao Yang, He Bai, and Prudhvi Gurram. Distributed stochastic gradient method for non-convex problems with applications in supervised learning. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 5538–5543. IEEE, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Kun Huang, Xiao Li, Andre Milzarek, Shi Pu, and Junwen Qiu. Distributed random reshuffling over networks. *IEEE Transactions on Signal Processing*, 71:1143–1158, 2023a.

Kun Huang, Linli Zhou, and Shi Pu. Distributed random reshuffling methods with improved convergence. *arXiv preprint arXiv:2306.12037*, 2023b.

Prateek Jain, Purushottam Kar, et al. Non-convex optimization for machine learning. *Foundations and Trends® in Machine Learning*, 10(3-4):142–363, 2017.

Zhanhong Jiang, Aditya Balu, Chinmay Hegde, and Soumik Sarkar. Collaborative deep learning in fixed topology networks. *Advances in Neural Information Processing Systems*, 30, 2017.

Cédric Josz and Lexiao Lai. Lyapunov stability of the subgradient method with constant step size. *Mathematical Programming*, 202(1):387–396, 2023.

Cédric Josz and Lexiao Lai. Global stability of first-order methods for coercive tame functions. *Mathematical Programming*, 207(1):551–576, 2024.

John L Kelley. *General topology*. Courier Dover Publications, 2017.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014.

Anastasiia Koloskova, Tao Lin, and Sebastian U Stich. An improved analysis of gradient tracking for decentralized machine learning. *Advances in Neural Information Processing Systems*, 34:11422–11435, 2021.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Tam Le. Nonsmooth nonconvex stochastic heavy ball. *Journal of Optimization Theory and Applications*, pages 1–21, 2024.

Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. *Advances in neural information processing systems*, 30, 2017.

Wei Liu, Anweshit Panda, Ujwal Pandey, Christopher Brissette, Yikang Shen, George M. Slota, Naigang Wang, Jiewei Chen, and Yangyang Xu. Compressed decentralized momentum stochastic gradient methods for nonconvex optimization. *Trans. Mach. Learn. Res.*, 2025, 2025. URL `https://api.semanticscholar.org/CorpusID:280545843`.

Songtao Lu, Xinwei Zhang, Haoran Sun, and Mingyi Hong. Gnsd: A gradient-tracking based nonconvex stochastic algorithm for decentralized optimization. In *2019 IEEE Data Science Workshop (DSW)*, pages 315–321. IEEE, 2019.

Gonzalo Mateos and Georgios B Giannakis. Distributed recursive least-squares: Stability and performance analysis. *IEEE Transactions on Signal Processing*, 60(7):3740–3754, 2012.

Angelia Nedic and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009.

Angelia Nedić, Alex Olshevsky, and Michael G Rabbat. Network topology and communication-computation tradeoffs in decentralized optimization. *Proceedings of the IEEE*, 106(5):953–976, 2018.

Youcheng Niu, Huaqing Li, Zheng Wang, Qingguo Lü, Dawen Xia, and Lianghao Ji. A distributed stochastic proximal-gradient algorithm for composite optimization. *IEEE Transactions on Control of Network Systems*, 8(3):1383–1393, 2021.

Edouard Pauwels. Incremental without replacement sampling in nonconvex optimization. *Journal of Optimization Theory and Applications*, 190(1):274–299, 2021.

S Unnikrishna Pillai, Torsten Suel, and Seunghun Cha. The perron-frobenius theorem: some of its applications. *IEEE Signal Processing Magazine*, 22(2):62–75, 2005.

Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17, 1964.

Shi Pu and Angelia Nedić. Distributed stochastic gradient tracking methods. *Mathematical Programming*, 187:409–457, 2021.

Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*, 2019.

Andrzej Ruszczyński. Convergence of a stochastic subgradient method with averaging for nonsmooth nonconvex constrained optimization. *Optimization Letters*, 14(7):1615–1625, 2020.

Emre Sahinoglu and Shahin Shahrampour. Online optimization perspective on first-order and zero-order decentralized nonsmooth nonconvex stochastic optimization. *arXiv preprint arXiv:2406.01484*, 2024.

Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczynski. *Lectures on stochastic programming: modeling and theory*. SIAM, 2021.

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

Wei Shi, Qing Ling, Gang Wu, and Wotao Yin. A proximal gradient algorithm for decentralized composite optimization. *IEEE Transactions on Signal Processing*, 63(22):6013–6023, 2015b.

Zhuoqing Song, Weijian Li, Kexin Jin, Lei Shi, Ming Yan, Wotao Yin, and Kun Yuan. Communication-efficient topologies for decentralized learning with $o(1)$ consensus rate. *Advances in Neural Information Processing Systems*, 35:1073–1085, 2022.

Jianhui Sun, Xidong Wu, Heng Huang, and Aidong Zhang. On the role of server momentum in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 15164–15172, 2024.

S Sundhar Ram, Angelia Nedić, and Venugopal V Veeravalli. Distributed stochastic subgradient projection algorithms for convex optimization. *Journal of optimization theory and applications*, 147:516–545, 2010.

Brian Swenson, Ryan Murray, H Vincent Poor, and Soummya Kar. Distributed stochastic gradient descent: Nonconvexity, nonsmoothness, and convergence to local minima. *The Journal of Machine Learning Research*, 23(1):14751–14812, 2022.

Thijs Vogels, Hadrien Hendrikx, and Martin Jaggi. Beyond spectral gap (extended): The role of the topology in decentralized learning. *arXiv preprint arXiv:2301.02151*, 2023.

Lei Wang and Xin Liu. Decentralized optimization over the stiefel manifold by an approximate augmented lagrangian function. *IEEE Transactions on Signal Processing*, 70: 3029–3041, 2022.

Lei Wang, Le Bao, and Xin Liu. A decentralized proximal gradient tracking algorithm for composite optimization on riemannian manifolds. *Journal of Machine Learning Research*, 26(106):1–37, 2025.

Lin Xiao and Stephen Boyd. Fast linear iterations for distributed averaging. *Systems & Control Letters*, 53(1):65–78, 2004.

Lin Xiao, Stephen Boyd, and Sanjay Lall. Distributed average consensus with time-varying metropolis weights. *Automatica*, 1, 2006.

Nachuan Xiao, Xiaoyin Hu, and Kim-Chuan Toh. Convergence guarantees for stochastic subgradient methods in nonsmooth nonconvex optimization. *arXiv preprint arXiv:2307.10053*, 2023a.

Nachuan Xiao, Xiaoyin Hu, Xin Liu, and Kim-Chuan Toh. Adam-family methods for nonsmooth optimization with convergence guarantees. *Journal of Machine Learning Research*, 25(48):1–53, 2024. URL http://jmlr.org/papers/v25/23-0576.html.

Tesi Xiao, Xuxing Chen, Krishnakumar Balasubramanian, and Saeed Ghadimi. A one-sample decentralized proximal algorithm for non-convex stochastic composite optimization. In *Uncertainty in Artificial Intelligence*, pages 2324–2334. PMLR, 2023b.

Ran Xin, Subhro Das, Usman A Khan, and Soummya Kar. A stochastic proximal gradient framework for decentralized non-convex composite optimization: Topology-independent sample complexity and communication efficiency. *arXiv preprint arXiv:2110.01594*, 2021a.

Ran Xin, Usman A Khan, and Soummya Kar. An improved convergence analysis for decentralized online stochastic non-convex optimization. *IEEE Transactions on Signal Processing*, 69:1842–1858, 2021b.

Yangyang Xu, Yibo Xu, Yonggui Yan, Colin Sutcher-Shepard, Leopold Grinberg, and Jie Chen. Parallel and distributed asynchronous adaptive stochastic gradient methods. *Mathematical Programming Computation*, 15(3):471–508, 2023.

Yonggui Yan, Jiewei Chen, Pin-Yu Chen, Xiaodong Cui, Songtao Lu, and Yangyang Xu. Compressed decentralized proximal stochastic gradient method for nonconvex composite problems with heterogeneous data. In *International Conference on Machine Learning*, 2023. URL https://api.semanticscholar.org/CorpusID:257232829.

Hao Yu, Rong Jin, and Sen Yang. On the linear speedup analysis of communication efficient momentum sgd for distributed non-convex optimization. In *International Conference on Machine Learning*, pages 7184–7193. PMLR, 2019.

Kun Yuan, Yiming Chen, Xinmeng Huang, Yingya Zhang, Pan Pan, Yinghui Xu, and Wotao Yin. Decentlam: Decentralized momentum sgd for large-batch deep training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3029–3039, 2021.

Jinshan Zeng and Wotao Yin. On nonconvex decentralized gradient descent. *IEEE Transactions on signal processing*, 66(11):2834–2848, 2018.

Jiaqi Zhang and Keyou You. Decentralized stochastic gradient tracking for non-convex empirical risk minimization. *arXiv preprint arXiv:1909.02712*, 2019.

Yuan Zhou, Xinli Shi, Xuelong Li, Jiachen Zhong, Guanghui Wen, and Jinde Cao. Decentralized nonconvex composite federated learning with gradient tracking and momentum. *arXiv preprint arXiv:2504.12742*, 2025.

Juntang Zhuang, Tommy Tang, Yifan Ding, Sekhar C Tatikonda, Nicha Dvornek, Xenophon Papademetris, and James Duncan. Adabelief optimizer: Adapting stepsizes by the belief in observed gradients. *Advances in neural information processing systems*, 33: 18795–18806, 2020.

Stanisław Łojasiewicz. Ensembles semi-analytiques. 1965. URL https://api.semanticscholar.org/CorpusID:118074635.