Journal of Machine Learning Research 25 (2024) 1-53 Submitted 5/23; Revised 2/24; Published 2/24

Adam-family Methods for Nonsmooth Optimization with Convergence Guarantees

Nachuan Xiao XNC@LSEC.CC.AC.CN
Institute of Operations Research and Analytics
National University of Singapore
3 Research Link, Singapore, 117602

Xiaoyin Hu[*] HXY@AMSS.AC.CN
School of Computer and Computing Science
Hangzhou City University
Hangzhou, China, 310015

Xin Liu LIUXIN@LSEC.CC.AC.CN
State Key Laboratory of Scientific and Engineering Computing
Academy of Mathematics and Systems Science, Chinese Academy of Sciences
Beijing, China, 100190

Kim-Chuan Toh MATTOHKC@NUS.EDU.SG
Department of Mathematics and Institute of Operations Research and Analytics
National University of Singapore
10 Lower Kent Ridge Road, Singapore, 119076

Editor: Krishnakumar Balasubramanian

Abstract

In this paper, we present a comprehensive study on the convergence properties of Adam-family methods for nonsmooth optimization, especially in the training of nonsmooth neural networks. We introduce a novel two-timescale framework that adopts a two-timescale updating scheme, and prove its convergence properties under mild assumptions. Our proposed framework encompasses various popular Adam-family methods, providing convergence guarantees for these methods in training nonsmooth neural networks. Furthermore, we develop stochastic subgradient methods that incorporate gradient clipping techniques for training nonsmooth neural networks with heavy-tailed noise. Through our framework, we show that our proposed methods converge even when the evaluation noises are only assumed to be integrable. Extensive numerical experiments demonstrate the high efficiency and robustness of our proposed methods.

Keywords: nonsmooth optimization, stochastic subgradient methods, Adam, nonconvex optimization, gradient clipping

[*] Corresponding author

1. Introduction

In this paper, we consider the following unconstrained nonlinear optimization problem

$$\min_{x\in\mathbb{R}^n} \quad f(x), \tag{UNP}$$

where f is nonconvex, locally Lipschitz continuous and possibly nonsmooth over $\mathbb{R}^n$.

The optimization problem in the form of UNP has numerous important applications in machine learning and data science, especially in training deep neural networks. In these applications of UNP, we usually only have access to the stochastic evaluations of the exact gradients of f. The stochastic gradient descent (SGD) is one of the most popular methods for solving UNP, and incorporating the momentum terms to SGD for acceleration is also very popular in practice. In SGD, the updating rule depends on the stepsizes (i.e., learning rates), where all of the coordinates of the variable x are equipped with the same stepsize. Recently, a variety of accelerated versions for SGD are proposed. In particular, the widely used Adam algorithm (Kingma and Ba, 2015) is developed based on the adaptive adjustment of the coordinate-wise stepsizes and the incorporation of momentum terms in each iteration. These enhancements have led to its high efficiency in practice. Motivated by Adam, a number of efficient Adam-family methods are developed, such as AdaBelief (Zhuang et al., 2020), AMSGrad (Reddi et al., 2018), NAdam (Dozat, 2016), Yogi (Zaheer et al., 2018), etc.

Towards the convergence properties of these Adam-family methods, Kingma and Ba (2015) shows the convergence properties for Adam with constant stepsize in minimizing a Lipschitz continuously differentiable objective function f. Then a great number of existing works are conducted to establish the convergence properties of Adam-family methods, see (De et al., 2018; Zaheer et al., 2018; Zou et al., 2019; Barakat and Bianchi, 2021; Guo et al., 2021; Shi et al., 2021; Zhang et al., 2022; Wang et al., 2022) for more information. Some of these existing works (Zou et al., 2019; Guo et al., 2021; Shi et al., 2021; Barakat and Bianchi, 2021; Wang et al., 2022; Zhang et al., 2022) adopt diminishing stepsizes to ensure the almost surely convergence to stationary points of f for Adam, while some other existing works (De et al., 2018; Zaheer et al., 2018) fix the stepsize as constant and show that the sequence converges to a neighborhood of the stationary points of f. A comparison of the results of these existing works is presented in Table 1.

Despite extensive studies on Adam-family methods, most existing works focus on the cases where f is differentiable over $\mathbb{R}^n$, as depicted in Table 1. However, nonsmooth activation functions, including ReLU and leaky ReLU, are very popular choices in building neural networks in practice (Ming et al., 2018; Fu et al., 2020, 2021; Wang et al., 2023), and most of the existing works (e.g., the works listed in Table 1) test their analyzed Adam-family methods on the neural networks built by nonsmooth activation functions. As highlighted in (Bolte and Pauwels, 2021; Bianchi et al., 2022), when we build a neural network by nonsmooth blocks, the corresponding loss function is typically nonsmooth and not Clarke regular. Consequently, although numerous existing works establish the convergence properties for Adam and its variants, their results are not applicable to the analysis of these Adam-family methods in training nonsmooth neural networks. This naturally leads us to the following question:

	Beyond differentiability	Beyond global Lipschitz	Nesterov mo.	A.s. convergence	Stepsize
Our work	✓	✓	✓	✓	$o(1/\log(k))$
Dozat (2016)	✗	✗	✓	✗	No convergence result
De et al. (2018)	✗	✗	✗	✗	Constant
Zou et al. (2019)	✗	✗	✗	✓	$O(k^{-s}), s \in (0,1]$
Zaheer et al. (2018)	✗	✗	✗	✗	Constant
Guo et al. (2021)	✗	✗	✗	✓	Square-summable
Shi et al. (2021)	✗	✗	✗	✓	Square-summable
Barakat and Bianchi (2021)	✗	✓	✗	✓	Square-summable
Wang et al. (2022)	✗	✓	✗	✓	Square-summable
Zhang et al. (2022)	✗	✗	✗	✓	Square-summable
Chen et al. (2022)	✗	✗	✗	✓	$O(k^{-s}), s \in (0,1]$

Table 1: Comparison of existing results on the convergence of Adam. Here "mo." is the abbreviation for "momentum". The term "A.s. convergence" refers to whether the sequence converges to stationary points of f rather than to a neighborhood of these stationary points almost surely.

> Do Adam-family methods have any convergence guarantees in minimizing nonsmooth functions under practical settings, especially in training nonsmooth neural networks?

1.1 Challenges from Training Nonsmooth Neural Networks

In training nonsmooth neural networks, one of the major challenges lies in how to differentiate their loss functions. These functions are typically formulated as compositions of elementary blocks that may not be smooth. To address this issue, automatic differentiation (AD) algorithms have been widely adopted in various well-known machine learning packages, such as PyTorch, TensorFlow, JAX, MindSpore, and PaddlePaddle. Based on the chain rule, the AD algorithms can efficiently compute the gradients for those functions expressed through the composition of elementary differentiable blocks. However, as the chain rule fails for Clarke subdifferential, when we differentiate a neural network built from nonsmooth blocks by those AD algorithms, the results may not be contained in the Clarke subdifferential of its loss function. As pointed out in Bolte and Pauwels (2021), most of the existing works ignore this issue. They use AD algorithms in training nonsmooth neural networks, but assume differentiability or weak convexity for the objective functions in their theoretical analysis to bypass these theoretical issues arising from the application of AD algorithms. Based on the chain rule for directional derivatives, some existing works (Barton et al., 2018) propose specifically designed forward mode AD algorithms for evaluating the elements in lexicographic subdifferential (Nesterov, 2005), which is contained in the Clarke subdifferential. However, as described in Bolte et al. (2021), these approaches have expensive computational costs and require significant modifications to the algorithms in existing machine learning packages, and hence are less applicable to practical scenarios.

To understand how AD algorithms differentiate the loss functions of nonsmooth neural networks, Bolte and Pauwels (2021) introduces the concept of the *conservative field* as a generalization of Clarke subdifferential for its corresponding *potential functions*. The class of potential functions includes semi-algebraic functions, semi-analytic functions, and functions whose graphs are definable in some o-minimal structures, hence covering the objective functions in a wide range of real-world applications (Davis et al., 2020; Bolte and Pauwels, 2021). The conservative field preserves the validity of the chain rule for nonsmooth functions, explaining the results generated by AD algorithms from various popular numerical libraries such as PyTorch, TensorFlow, JAX, etc. Based on the concept of conservative field, we can characterize the stationarity and design algorithms for the unconstrained nonsmooth optimization, especially when the objective function is differentiated by AD algorithms. Interested readers can refer Bolte and Pauwels (2021) for more detailed properties of the conservative field.

The theoretical properties of the conservative field enable us to investigate the convergence properties of stochastic subgradient algorithms, especially when applied to train nonsmooth neural networks with AD algorithms. Some existing frameworks (Benaïm et al., 2005; Davis et al., 2020; Bolte and Pauwels, 2021) establish the convergence properties for stochastic subgradient methods by analyzing the limiting behaviour of their corresponding differential inclusions, and prove that these methods converge to stationary points of f in the sense of its corresponding conservative field $\mathcal{D}_f$. Based on these frameworks, some recent works (Davis et al., 2020; Bolte and Pauwels, 2021; Hu et al., 2023) prove the convergence properties for SGD and proximal SGD. Moreover, Castera et al. (2021) proposes the inertial Newton algorithm (INNA), which can be regarded as a variant of SGD with heavy-ball momentum. Additionally, Ruszczyński (2020); Le (2023) show the convergence property of SGD with heavy-ball momentum for nonsmooth nonconvex functions from the Norkin class. For Adam and its variants, some existing works (Da Silva and Gazeau, 2020; Barakat and Bianchi, 2021; Gadat and Gavra, 2022) established the convergence properties of Adam for Lipschitz smooth f by analyzing the limiting behaviour of its corresponding differential equation. However, their approaches rely on some time-dependent differential equations, which are challenging to be extended to nonsmooth cases based on the frameworks in (Benaïm et al., 2005; Davis et al., 2020; Bolte and Pauwels, 2021). To the best of our knowledge, no existing work addresses the convergence properties of Adam-family methods for nonsmooth optimization.

Furthermore, Bolte and Pauwels (2021) demonstrate that the Clarke subdifferential is a subset of the conservative field for any potential functions. For nonsmooth neural networks, the conservative fields associated with AD algorithms may introduce infinitely many spurious stationary points (Bolte et al., 2021; Bianchi et al., 2022). Therefore, when we design stochastic subgradient methods based on the conservative field, the results in some existing frameworks (Benaïm et al., 2005; Davis et al., 2020; Bolte and Pauwels, 2021) can only ensure the convergence to stationary points in the sense of conservative field. As demonstrated in (Bolte et al., 2021; Bianchi et al., 2022), these results fail to guarantee the convergence to meaningful stationary points of f. To this end, Bianchi et al. (2022) establishes that under mild assumptions with randomized initial points and stepsizes, SGD can find Clarke stationary points for nonsmooth neural networks almost surely. However, their analysis is limited to SGD without any momentum term, and how to extend their results to Adam-family methods remains an open question.

1.2 Challenges from Heavy-tailed Evaluation Noises

Another challenge for solving UNP lies in the noises when evaluating the stochastic subgradient of the objective function. The evaluation noises in a great number of existing works are assumed to have finite second-order moment or even uniformly bounded, for the sake of convenience when analyzing their theoretical properties. However, in various

	Beyond differentiability	Assumption on noises	Adaptive stepsize	Heavy-ball mo.	Nesterov mo.	Convergence
Our work	✓	L^1	✓	✓	✓	✓
Zhang et al. (2020a)	✗	L^∞	✗	✗	✗	✓
Gorbunov et al. (2020)	✗	L^2	✗	✗	✗	✓
Zhang et al. (2020b)	✗	L^s for $s \in (1, 2]$	✗	✗	✗	✓
Zhang et al. (2020b)	✗	No convergence result	✓	✗	✗	✗
Mai and Johansson (2021)	✓$^{(a)}$	L^2	✗	✗	✗	✓
Qian et al. (2021)	✗	L^∞	✗	✗	✗	✓
Elesedy and Hutter (2023)	✗	L^2	✗	✗	✗	✓
Reisizadeh et al. (2023)	✗	L^2	✗	✗	✗	✓

Table 2: Comparison of existing convergence results of stochastic (sub)gradient methods with gradient clipping techniques. Here "mo." is the abbreviation for "momentum". (a): The proof techniques in Mai and Johansson (2021) relies on weak convexity on f and cannot be applied to non-regular cases.

machine learning tasks, such as classification models (Mahoney and Martin, 2019; Simsekli et al., 2019, 2020; Camuto et al., 2021; Wan et al., 2023) and language models (Zhang et al., 2020a,b), some recent works (Simsekli et al., 2019; Zhang et al., 2020b) illustrate that the evaluation noises of the stochastic subgradients could be heavy-tailed (i.e., only have bounded s-order moment for some $s \in [1, 2)$ (Zhang et al., 2020a)). As illustrated in Zhang et al. (2020a), the heavy-tailed evaluation noises have a higher probability of producing extreme values or outliers when compared to normal distributions, and hence may undermine the performance of SGD in these tasks. Even in finite-sum settings, the frequently occurred extreme values in the evaluation of stochastic subgradients can result in extremely large variance (Simsekli et al., 2019). These results explain the empirical observations in training neural networks, including the long-standing failure cases of SGD methods in training recurrent neural networks (Pascanu et al., 2012), and the superior performance of adaptive methods over SGD methods in training language models (Zhang et al., 2020b).

To address the challenges in solving UNP with heavy-tailed evaluation noises, the gradient clipping technique has been developed. Gradient clipping normalizes the stochastic gradient, thus preventing extreme values in evaluating the stochastic gradients that can cause instability or divergence in the optimization algorithms. With the gradient clipping technique, some recent works (Zhang et al., 2020a) show that SGD converges when the evaluation noises are bounded in L^s (i.e., the noises $\{\xi_k\}$ satisfy $\sup_{k\geq 0} \mathbb{E}[||\xi_k||^s] < +\infty$) for some $s \in (1, 2)$. Table 2 exhibits the related works on the convergence properties of stochastic subgradient methods with gradient clipping (Zhang et al., 2020a; Gorbunov et al., 2020; Zhang et al., 2020b; Mai and Johansson, 2021; Qian et al., 2021; Elesedy and Hutter, 2023; Reisizadeh et al., 2023). As illustrated in Table 2, all of these existing works rely on the weak convexity of the objective function f, hence they are not applicable for training nonsmooth neural networks.

Moreover, most of these existing works focus on the standard SGD method without the momentum term. Although Zhang et al. (2020b); Pan and Li (2023) introduce stochastic Adam-family methods with gradient clipping, they did not provide any convergence guarantee for their proposed method. Therefore, a significant gap exists between the existing theoretical analysis (Zhang et al., 2020a; Gorbunov et al., 2020; Zhang et al., 2020b; Mai and Johansson, 2021; Qian et al., 2021; Elesedy and Hutter, 2023; Reisizadeh et al., 2023) and practical implementations for stochastic subgradient methods with heavy-tailed noise, and how to fill that gap is challenging and remains unexplored.

1.3 Contributions

In this paper, we aim to establish the convergence properties of Adam-family methods for nonsmooth optimization, especially in the context of training nonsmooth neural networks. To this end, we employ the concept of the conservative field to characterize how the objective function f is differentiated, and consider the following set-valued mapping $\mathcal{G} : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \rightrightarrows \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n$,

$$\mathcal{G}(x,m,v) := \left\{ \begin{bmatrix} (|v| + \varepsilon)^{-\gamma} \odot (m + \alpha d) \\ \tau_1 m - \tau_1 d \\ \tau_2 v - \tau_2 u \end{bmatrix} : d \in \mathcal{D}_f(x),\ u \in \mathcal{U}(x,m,v) \right\}. \tag{1}$$

Here $\mathcal{D}_f$ refers to the conservative field that characterizes how we differentiate the objective function f, and $\alpha, \gamma, \varepsilon, \tau_1, \tau_2$ are hyper-parameters. Moreover, $\odot$ and $(\cdot)^\gamma$ refer to the element-wise multiplication and power, respectively. Furthermore, $\mathcal{U} : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is a set-valued mapping that determines how the estimator v_k is updated. Then we propose the following generalized framework for Adam-family methods (AFM),

$$(x_{k+1}, m_{k+1}, v_{k+1}) = (x_k, m_k, v_k) - \eta_k (d_{x,k}, d_{m,k}, d_{v,k}) - \theta_k (\xi_{x,k}, \xi_{m,k}, \xi_{v,k}). \tag{AFM}$$

In (AFM), $(d_{x,k}, d_{m,k}, d_{v,k})$ denotes the updating direction, which is an approximated evaluation for $\mathcal{G}(x_k, m_k, v_k)$, while $(\xi_{x,k}, \xi_{m,k}, \xi_{v,k})$ refers to the evaluation noise. Moreover, $\{\eta_k\}$ and $\{\theta_k\}$ are the two-timescale stepsizes for updating directions and evaluation noises respectively, in the sense that they may satisfy $\eta_k/\theta_k \to 0$ as $k \to \infty$.

We prove that under mild conditions, any cluster point of the sequence x_k generated by our proposed framework (AFM) with stepsizes in the order of $o(1/\log(k))$ is a $\mathcal{D}_f$-stationary point of f. Furthermore, we establish that under mild conditions with randomly chosen initial points and stepsizes, almost surely, any cluster point of the sequence x_k is a Clarke stationary point of f, independent of the chosen conservative field $\mathcal{D}_f$.

Based on our proposed framework (AFM), we demonstrate that our proposed framework can be employed to analyze the convergence properties for a class of Adam-family methods with diminishing stepsize, including Adam, AdaBelief, AMSGrad, NAdam, and Yogi. We prove that these Adam-family methods converge to stationary points of f in both senses of conservative field and Clarke subdifferential under mild conditions, thus providing theoretical guarantees for their performance in training nonsmooth neural networks with AD algorithms.

Another application of our proposed framework (AFM) lies in investigating the convergence properties of stochastic subgradient methods that incorporate the gradient clipping technique. We prove that under heavy-tailed evaluation noises that are only assumed to be integrable, our proposed gradient clipping methods conform to the proposed framework (AFM). As a result, the convergence properties of these gradient clipping methods directly follow those established for our proposed framework (AFM), under mild conditions.

Furthermore, we perform extensive numerical experiments to evaluate the performance of our proposed Adam-family methods. By comparing with the implementations of Adam-family methods in PyTorch that utilize fixed stepsize in updating their momentum terms and variance estimators, we demonstrate that our proposed Adam-family methods achieve similar accuracy and training loss. Moreover, when the evaluation noises are heavy-tailed, the numerical examples demonstrate that our proposed Adam-family methods outperform existing approaches in terms of training efficiency and robustness.

1.4 Organization

The rest of this paper is organized as follows. In Section 2, we define the notations used throughout the paper and present some essential concepts related to probability theory, nonsmooth analysis and differential inclusion. In Section 3, we focus on the analysis of the convergence properties of our proposed framework (AFM), in both senses of the conservative field and Clarke subdifferential. Section 4 illustrates the application of our framework (AFM) in establishing the convergence properties for a class of Adam-family methods, including Adam, AdaBelief, AMSGrad, NAdam, and Yogi, under practical settings with mild conditions. Section 5 demonstrates another application of our framework (AFM) by illustrating the convergence properties of stochastic subgradient methods with gradient clipping technique under heavy-tailed evaluation noises. In Section 6, we present the results of our numerical experiments that investigate the performance of our proposed Adam-family methods for training nonsmooth neural networks. Finally, we conclude the paper in the last section.

2. Preliminary

2.1 Basic Notations

For any vectors x and y in $\mathbb{R}^n$ and $\delta \in \mathbb{R}$, we denote $x \odot y$, x^δ, x/y, $|x|$, $x + \delta$ as the vectors whose i-th entries are respectively given by $x_i y_i$, x_i^δ, x_i/y_i, $|x_i|$ and $x_i + \delta$. Moreover, for any sets $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^n$, we denote $\mathcal{X} \odot \mathcal{Y} := \{x \odot y : x \in \mathcal{X}, y \in \mathcal{Y}\}$, $(\mathcal{X})^p := \{x^p : x \in \mathcal{X}\}$ and $|\mathcal{X}| := \{|x| : x \in \mathcal{X}\}$. In addition, for any $z \in \mathbb{R}^n$, we denote $z + \mathcal{X} := \{z\} + \mathcal{X}$ and $z \odot \mathcal{X} := \{z\} \odot \mathcal{X}$.

We define the set-valued mappings $\mathrm{sign} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ and $\widetilde{\mathrm{sign}} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ as follows: For any $x \in \mathbb{R}^n$,

$$(\mathrm{sign}(x))_i = \begin{cases} \{-1\} & x_i < 0; \\ [-1, 1] & x_i = 0; \\ \{1\} & x_i > 0. \end{cases}, \quad \text{and} \quad \left(\widetilde{\mathrm{sign}}(x)\right)_i = \begin{cases} \{-1\} & x_i < 0; \\ \{0\} & x_i = 0; \\ \{1\} & x_i > 0. \end{cases}$$

Then it is easy to verify that $\widetilde{\mathrm{sign}}(x) \odot \mathrm{sign}(x) = (\widetilde{\mathrm{sign}}(x))^2$ holds for any $x \in \mathbb{R}^n$.

In addition, we denote $\mathbb{R}^n_+ := \{x \in \mathbb{R}^n : x_i \geq 0 \text{ for any } 1 \leq i \leq n\}$. Moreover, μ^d refers to the Lebesgue measure on $\mathbb{R}^d$, and when the dimension d is clear from the context, we write the Lebesgue measure as μ for brevity. Furthermore, we say a measurable set A is zero-measure if $\mu(A) = 0$, and A is full-measure if $\mu(A^c) = 0$.

2.2 Probability Theory

In this subsection, we present some essential concepts from probability theory, which are necessary for the proofs in this paper.

Definition 1 *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. We say $\{\mathcal{F}_k\}_{k\in\mathbb{N}}$ is a filtration if $\{\mathcal{F}_k\}$ is a collection of σ-algebras that satisfies $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}_\infty \subseteq \mathcal{F}$.*

Definition 2 *We say that a stochastic series $\{\xi_k\}$ is a martingale if the following conditions hold.*

- *The sequence of random vectors $\{\xi_k\}$ is adapted to the filtration $\{\mathcal{F}_k\}$. That is, for any $k \geq 0$, ξ_k is measurable with respect to the σ-algebra $\mathcal{F}_k$.*
- *The equation $\mathbb{E}[\xi_{k+1}|\mathcal{F}_k] = \xi_k$ holds almost surely for every $k \geq 0$.*

Definition 3 *We say that a stochastic series $\{\xi_k\}$ is a supermartingale if $\{\xi_k\}$ is adapted to the filtration $\{\mathcal{F}_k\}$ and $\mathbb{E}[\xi_{k+1}|\mathcal{F}_k] \leq \xi_k$ holds almost surely for every $k \geq 0$.*

Definition 4 *We say that a stochastic series $\{\xi_k\}$ is a martingale difference sequence if the following conditions hold.*

- *The sequence of random vectors $\{\xi_k\}$ is adapted to the filtration $\{\mathcal{F}_k\}$.*
- *For each $k \geq 1$, almost surely, it holds that $\mathbb{E}[|\xi_k|] < +\infty$ and $\mathbb{E}\left[\xi_k|\mathcal{F}_{k-1}\right] = 0$.*

The following proposition plays an important role in establishing the convergence properties for our proposed framework (AFM). In this proposition, we improve the results in (Benaïm, 2006, Proposition 4.4) and demonstrate that with appropriately chosen $\{\eta_k\}$ and $\{\theta_k\}$, the uniform boundedness of the martingale difference sequence $\{\xi_k\}$ leads to the validity of the regularity conditions in (Benaïm et al., 2005, Section 1.5).

Proposition 1 *Suppose $\{\eta_k\}$ and $\{\theta_k\}$ are two diminishing positive sequences of real numbers that satisfy*

$$\lim_{k\to+\infty} \frac{\theta_k^2}{\eta_k} \log(k) = 0.$$

Let $\lambda_0 := 0$, $\lambda_i := \sum_{k=0}^{i-1} \eta_k$, and $\Lambda(t) := \sup\{k \geq 0 : t \geq \lambda_k\}$. Then for any $T > 0$, and any uniformly bounded martingale difference sequence $\{\xi_k\}$, almost surely, it holds that

$$\lim_{s\to+\infty} \sup_{s\leq i\leq \Lambda(\lambda_s+T)} \left\| \sum_{k=s}^{i} \theta_k \xi_k \right\| = 0.$$

Proof Since the martingale difference sequence $\{\xi_k\}$ is uniformly bounded, ξ_k is sub-Gaussian for any $k \geq 0$. Then there exists a constant $M > 0$ such that for any $w \in \mathbb{R}^n$, it holds for any $k \geq 0$ that

$$\mathbb{E}\left[\exp\left(\langle w, \xi_{k+1}\rangle\right)|\mathcal{F}_k\right] \leq \exp\left(\frac{M}{2}\|w\|^2\right),$$

holds almost surely. Therefore, for any $w \in \mathbb{R}^n$ and any $C > 0$, let

$$Z_i := \exp\left[\left\langle Cw, \sum_{k=s}^{i} \theta_k \xi_k \right\rangle - \frac{MC^2}{2} \sum_{k=s}^{i} \theta_k^2 \|w\|^2\right].$$

Then for any $i \geq 0$, we have that $\mathbb{E}[Z_{i+1}|\mathcal{F}_i] \leq Z_i$. Hence for any $\delta > 0$, and any $C > 0$, it holds that

$$\begin{aligned}
&\mathbb{P}\left(\sup_{s\leq i\leq \Lambda(\lambda_s+T)} \left\langle w, \sum_{k=s}^{i} \theta_k \xi_k \right\rangle > \delta\right) = \mathbb{P}\left(\sup_{s\leq i\leq \Lambda(\lambda_s+T)} \left\langle Cw, \sum_{k=s}^{i} \theta_k \xi_k \right\rangle > C\delta\right)\\
\leq{}& \mathbb{P}\left(\sup_{s\leq i\leq \Lambda(\lambda_s+T)} Z_i > \exp\left(C\delta - \frac{MC^2}{2}\sum_{k=s}^{\Lambda(\lambda_s+T)} \theta_k^2 \left\|w\right\|^2\right)\right)\\
\leq{}& \exp\left(\left(\frac{M}{2}\left\|w\right\|^2 \sum_{k=s}^{\Lambda(\lambda_s+T)} \theta_k^2\right) C^2 - \delta C\right).
\end{aligned}$$

Here the second inequality holds since $\{Z_k\}$ is a nonnegative super-martingale and $\mathbb{E}[Z_s] \leq 1$. Then from the arbitrariness of C, set $C = \frac{\delta}{M\|w\|^2 \sum_{k=s}^{\Lambda(\lambda_s+T)} \theta_k^2}$, it holds that

$$\mathbb{P}\left(\sup_{s\leq i\leq \Lambda(\lambda_s+T)} \left\langle w, \sum_{k=s}^{i} \theta_k \xi_k \right\rangle > \delta\right) \leq \exp\left(\frac{-\delta^2}{2M\left\|w\right\|^2 \sum_{k=s}^{\Lambda(\lambda_s+T)} \theta_k^2}\right).$$

From the arbitrariness of w, there exists constants C_1 and C_2 that only depend on n such that

$$\mathbb{P}\left(\sup_{s\leq i\leq \Lambda(\lambda_s+T)} \left\|\sum_{k=s}^{i} \theta_k \xi_k\right\| > \delta\right) \leq C_1 \exp\left(\frac{-\delta^2}{2MC_2 \sum_{k=s}^{\Lambda(\lambda_s+T)} \theta_k^2}\right) \leq C_1 \exp\left(\frac{-\delta^2}{2MC_2 T \frac{\theta_{k'}^2}{\eta_{k'}}}\right),$$

holds for some $k' \in [s, \Lambda(\lambda_s + T)]$. Here $\{\eta_k\}$ refers to the diminishing sequence of real numbers as defined in the condition of this proposition.

Therefore, for any $j \geq 0$, there exists $k_j \in [\Lambda(jT), \Lambda((j+1)T)]$, such that

$$\begin{aligned}
&\sum_{j=0}^{+\infty} \mathbb{P}\left(\sup_{\Lambda(jT)\leq i\leq \Lambda(jT+T)} \left\|\sum_{k=s}^{i} \theta_k \xi_k\right\| \geq \delta\right)\\
\leq{}& \sum_{j=0}^{+\infty} C_1 \exp\left(\frac{-\delta^2}{2MC_2 T \eta_{k_j}^{-1} \theta_{k_j}^2}\right) \leq \sum_{k=0}^{+\infty} 2C_1 \exp\left(\frac{-\delta^2}{2MC_2 T \frac{\theta_k^2}{\eta_k}}\right) < +\infty.
\end{aligned}$$

Here the last inequality holds from the fact that $\lim_{k\to+\infty} \frac{\theta_k^2}{\eta_k} \log(k) = 0$. Therefore, we can conclude that

$$\lim_{j\to+\infty} \mathbb{P}\left(\sup_{\Lambda(jT)\leq i\leq \Lambda(jT+T)} \left\|\sum_{k=\Lambda(jT)}^{i} \theta_k \xi_k\right\| \geq \delta\right) = 0,$$

holds almost surely for any $\delta > 0$. Then the arbitrariness of δ illustrates that almost surely, we have

$$\lim_{j\to+\infty} \sup_{\Lambda(jT)\leq i\leq \Lambda(jT+T)} \left\|\sum_{k=\Lambda(jT)}^{i} \theta_k \xi_k\right\| = 0.$$

Finally, notice that for any $jT \leq s \leq jT + T$, it holds that

$$\begin{aligned}
&\sup_{s \leq i \leq \Lambda(\lambda_s + T)} \left\| \sum_{k=\Lambda(jT)}^{i} \theta_k \xi_k \right\| \\
\leq{}& 2 \sup_{\Lambda(jT) \leq i \leq \Lambda(jT+T)} \left\| \sum_{k=\Lambda(jT)}^{i} \theta_k \xi_k \right\| + \sup_{\Lambda((j+1)T) \leq i \leq \Lambda((j+2)T)} \left\| \sum_{k=\Lambda(jT+T)}^{i} \theta_k \xi_k \right\|.
\end{aligned}$$

Then we achieve that

$$\lim_{s \to +\infty} \sup_{s \leq i \leq \Lambda(\lambda_s + T)} \left\| \sum_{k=s}^{i} \theta_k \xi_k \right\| = 0,$$

holds almost surely. Hence we complete the proof. ■

2.3 Nonsmooth Analysis

2.3.1 Clarke Subdifferential

In this part, we introduce the concept of Clarke subdifferential (Clarke, 1990), which plays an important role in characterizing the stationarity and designing efficient algorithms for nonsmooth optimization problems.

Definition 5 (Clarke (1990)) *For any given locally Lipschitz continuous function $f : \mathbb{R}^n \to \mathbb{R}$ and any $x \in \mathbb{R}^n$, the generalized directional derivative of f at x in the direction $d \in \mathbb{R}^n$, denoted by $f^\circ(x; d)$, is defined as*

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

Then the generalized gradient or the Clarke subdifferential of f at x, denoted by $\partial f(x)$, is defined as

$$\partial f(x) := \left\{ w \in \mathbb{R}^n : \langle w, d \rangle \leq f^\circ(x; d), \text{ for all } d \in \mathbb{R}^n \right\}.$$

Then based on the concept of generalized directional derivative, we present the definition of (Clarke) regular functions.

Definition 6 (Clarke (1990)) *For any given locally Lipschitz continuous function $f : \mathbb{R}^n \to \mathbb{R}$ and any $x \in \mathbb{R}^n$, we say that f is (Clarke) regular at $x \in \mathbb{R}^n$ if for every direction $d \in \mathbb{R}^n$, the one-sided directional derivative*

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

exists and $f^\star(x; d) = f^\circ(x; d)$.

2.3.2 CONSERVATIVE FIELD

In this part, we present a brief introduction on the conservative field, which can be applied to characterize how the nonsmooth neural networks are differentiated by AD algorithms.

Definition 7 *A set-valued mapping $\mathcal{D} : \mathbb{R}^n \rightrightarrows \mathbb{R}^s$ is a mapping from $\mathbb{R}^n$ to a collection of subsets of $\mathbb{R}^s$. $\mathcal{D}$ is said to have closed graph (or $\mathcal{D}$ is graph-closed) if the graph of $\mathcal{D}$, defined by*

$$\mathrm{graph}(\mathcal{D}) := \{(w, z) \in \mathbb{R}^n \times \mathbb{R}^s : w \in \mathbb{R}^n, z \in \mathcal{D}(w)\},$$

is a closed subset of $\mathbb{R}^n \times \mathbb{R}^s$.

Definition 8 *A set-valued mapping $\mathcal{D} : \mathbb{R}^n \rightrightarrows \mathbb{R}^s$ is said to be locally bounded if, for any $x \in \mathbb{R}^n$, there is a neighborhood V_x of x such that $\cup_{y \in V_x} \mathcal{D}(y)$ is bounded.*

Definition 9 (Aumann's integral) *Let $(\Theta, \mathcal{F}, P)$ be a measurable space, and $\mathcal{D} : \mathbb{R}^n \times \Theta \rightrightarrows \mathbb{R}^n$ be a measurable set-valued mapping. Then for all $x \in \mathbb{R}^n$, the integral of $\mathcal{D}$ with respect to P is defined as*

$$\mathbb{E}_{s\sim P}\left[\mathcal{D}(x,s)\right] := \left\{ \int_\Theta \chi(x,s)\, \mathrm{d}P(s) :\ \chi(x,\cdot) \textit{ is integrable, and } \chi(x,s) \in \mathcal{D}(x,s) \textit{ for any } s \in \Theta \right\}.$$

The following lemma illustrates that the composition of two locally bounded graph-closed set-valued mappings is locally bounded and graph-closed. Therefore, we can easily verify the graph-closeness for the composition of set-valued mappings, which plays an important role in our theoretical analysis.

Lemma 1 (Lemma 2.5 in (Xiao et al., 2023)) *Suppose $\mathcal{D}_1 : \mathbb{R}^n \rightrightarrows \mathbb{R}^s$ and $\mathcal{D}_2 : \mathbb{R}^d \rightrightarrows \mathbb{R}^n$ are two locally bounded graph-closed set-valued mappings, then their composition $\mathcal{D}_1 \circ \mathcal{D}_2$ is locally bounded and graph-closed.*

In the following definitions, we present the definition for the conservative field and its corresponding potential function.

Definition 10 *An absolutely continuous curve is a continuous mapping $\gamma : \mathbb{R}_+ \to \mathbb{R}^n$ whose derivative γ' exists almost everywhere in $\mathbb{R}_+$ and $\gamma(t) - \gamma(0)$ equals to the Lebesgue integral of γ' between 0 and t for all $t \in \mathbb{R}_+$, i.e.,*

$$\gamma(t) = \gamma(0) + \int_0^t \gamma'(u)\mathrm{d}u, \qquad \textit{for all } t \in \mathbb{R}_+.$$

Definition 11 *Let $\mathcal{D}$ be a set-valued mapping from $\mathbb{R}^n$ to subsets of $\mathbb{R}^n$. Then we call $\mathcal{D}$ as a conservative field whenever it has closed graph, nonempty compact valued, and for any absolutely continuous curve $\gamma : [0,1] \to \mathbb{R}^n$ satisfying $\gamma(0) = \gamma(1)$, we have*

$$\int_0^1 \max_{v \in \mathcal{D}(\gamma(t))} \langle \gamma'(t), v \rangle \,\mathrm{d}t = 0, \tag{2}$$

where the integral is understood in the Lebesgue sense.

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

Definition 12 *Let $\mathcal{D}$ be a conservative field in $\mathbb{R}^n$. Then with any given $x_0 \in \mathbb{R}^n$, we can define a function $f : \mathbb{R}^n \to \mathbb{R}$ through the path integral*

$$f(x) = f(x_0) + \int_0^1 \max_{v \in \mathcal{D}(\gamma(t))} \langle \gamma'(t), v \rangle \, \mathrm{d}t = f(x_0) + \int_0^1 \min_{v \in \mathcal{D}(\gamma(t))} \langle \gamma'(t), v \rangle \, \mathrm{d}t, \tag{3}$$

for any absolutely continuous curve γ that satisfies $\gamma(0) = x_0$ and $\gamma(1) = x$. Then f is called a potential function for $\mathcal{D}$, and we also say $\mathcal{D}$ admits f as its potential function, or that $\mathcal{D}$ is a conservative field for f.

The following two lemmas characterize the relationship between the conservative field and the Clarke subdifferential.

Lemma 2 (Theorem 1 in Bolte and Pauwels (2021)) *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a potential function that admits $\mathcal{D}_f$ as its conservative field. Then $\mathcal{D}_f(x) = \{\nabla f(x)\}$ almost everywhere.*

Lemma 3 (Corollary 1 in Bolte and Pauwels (2021)) *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a potential function that admits $\mathcal{D}_f$ as its conservative field. Then ∂f is a conservative field for f, and for all $x \in \mathbb{R}^n$, it holds that*

$$\partial f(x) \subseteq \mathrm{conv}\, \left(\mathcal{D}_f(x)\right).$$

Definition 13 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a potential function that admits $\mathcal{D}_f$ as its conservative field, then we say x is a $\mathcal{D}_f$-stationary point of f if $0 \in \mathcal{D}_f(x)$. In particular, we say x is a ∂f-stationary point of f if $0 \in \partial f(x)$.*

It is worth mentioning that the class of potential functions is general enough to cover the objectives in a wide range of real-world problems. As shown in Davis et al. (2020, Section 5.1), any Clarke regular function is a potential function. Another important function class is the definable functions (i.e. the functions whose graphs are definable in an *o*-minimal structure) (Davis et al., 2020, Definition 5.10). As demonstrated in Van den Dries and Miller (1996), any definable function is also a potential function (Davis et al., 2020; Bolte and Pauwels, 2021). To characterize the definable functions, the Tarski–Seidenberg theorem (Bierstone and Milman, 1988) shows that any semi-algebraic function is definable. Moreover, Wilkie (1996) shows there exists an *o*-minimal structure that contains both the graph of the exponential function and all semi-algebraic sets. As a result, numerous common activation and loss functions, including sigmoid, softplus, ReLU, ℓ_1-loss, MSE loss, hinge loss, logistic loss, and cross-entropy loss, are all definable. Additionally, Bolte et al. (2021) reveals that parameterized solutions in a broad class of optimizations are definable.

Additionally, it should be noted that definability is preserved under finite summation and composition (Davis et al., 2020; Bolte and Pauwels, 2021). As a result, for any neural network built from definable blocks, its loss function is definable and thus is a potential function. Moreover, the Clarke subdifferential of definable functions are definable (Bolte and Pauwels, 2021). Therefore, for any neural network built from definable blocks, the conservative field corresponding to the AD algorithms is definable. The following proposition shows that the definability of f and $\mathcal{D}_f$ leads to the nonsmooth Morse–Sard property (Bolte et al., 2007) for UNP.

Proposition 2 (Theorem 5 in Bolte and Pauwels (2021)) *Let f be a potential function that admits $\mathcal{D}_f$ as its conservative field. Suppose both f and $\mathcal{D}_f$ are definable over $\mathbb{R}^n$, then the set $\{f(x) : 0 \in \mathcal{D}_f(x)\}$ is finite.*

2.4 Differential Inclusion and Stochastic Subgradient Methods

In this subsection, we introduce some fundamental concepts related to the differential equation (i.e., differential inclusion) that are essential for the proofs presented in this paper.

Definition 14 *For any locally bounded set-valued mapping $\mathcal{D} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ that is nonempty compact convex valued and has closed graph. We say the absolutely continuous path $x(t)$ in $\mathbb{R}^n$ is a solution for the differential inclusion*

$$\frac{\mathrm{d}x}{\mathrm{d}t} \in \mathcal{D}(x), \tag{4}$$

with initial point x_0 if $x(0) = x_0$, and $\dot{x}(t) \in \mathcal{D}(x(t))$ holds for almost every $t \geq 0$.

Definition 15 *For any given set-valued mapping $\mathcal{D} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ and any constant $\delta \geq 0$, the set-valued mapping $\mathcal{D}^\delta$ is defined as*

$$\mathcal{D}^\delta(x) := \{w \in \mathbb{R}^n : \exists z \in \mathbb{B}_\delta(x), \operatorname{dist}(w, \mathcal{D}(z)) \leq \delta\}. \tag{5}$$

For the differential inclusion (4), we introduce the concept on its perturbed solution in the following definition.

Definition 16 (Definition II in Benaïm et al. (2005)) *We say an absolutely continuous function γ is a perturbed solution to* (4) *if there exists a locally integrable function $u : \mathbb{R}_+ \to \mathbb{R}^n$, such that*

- *For any $T > 0$, it holds that $\lim\limits_{t\to+\infty} \sup\limits_{0\leq l\leq T} \left\| \int_t^{t+l} u(s)\, \mathrm{d}s \right\| = 0$.*
- *There exists $\delta : \mathbb{R}_+ \to \mathbb{R}$ such that $\lim\limits_{t\to+\infty} \delta(t) = 0$ and*

$$\dot{\gamma}(t) - u(t) \in \mathcal{D}^{\delta(t)}(\gamma(t)).$$

Now consider the sequence $\{x_k\}$ generated by the following updating scheme,

$$x_{k+1} = x_k + \eta_k(d_k + \xi_k), \tag{6}$$

where $\{\eta_k\}$ is a diminishing positive sequence of real numbers. We define the (continuous-time) interpolated process of $\{x_k\}$ generated by (6) as follows.

Definition 17 *The (continuous-time) interpolated process of $\{x_k\}$ generated by* (6) *is the mapping $w : \mathbb{R}_+ \to \mathbb{R}^n$ such that*

$$w(\lambda_i + s) := x_{i-1} + \frac{s}{\lambda_i - \lambda_{i-1}} \left(x_i - x_{i-1} \right), \quad s \in [0, \eta_i).$$

Here $\lambda_0 := 0$, and $\lambda_i := \sum_{k=0}^{i-1} \eta_k$.

The following lemma is an extension of (Benaïm et al., 2005, Proposition 1.3). Compared with (Benaïm et al., 2005, Proposition 1.3), Lemma 4 allows for inexact evaluations of the set-valued mapping $\mathcal{D}$, and shows that the interpolated process of $\{x_k\}$ from (6) is a perturbed solution of the differential inclusion (4).

Lemma 4 *Let $\mathcal{D} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be a locally bounded set-valued mapping that is nonempty compact convex valued with closed graph. Suppose the following conditions hold in* (6)*:*

1. *For any $T > 0$, it holds that*
$$\lim_{s \to +\infty} \sup_{s \leq i \leq \Lambda(\lambda_s + T)} \left\| \sum_{k=s}^{i} \eta_k \xi_k \right\| = 0.$$
2. *There exist a positive sequence $\{\delta_k\}$ such that $\lim_{k \to +\infty} \delta_k = 0$ and $d_k \in \mathcal{D}^{\delta_k}(x_k)$.*
3. $\sup_{k \geq 0} \|x_k\| < +\infty$, $\sup_{k \geq 0} \|d_k\| < +\infty$.

Then the interpolated process of $\{x_k\}$ is a perturbed solution for (4).

Proof Let $w : \mathbb{R}_+ \to \mathbb{R}^n$ denote the interpolated process for (6). Then define $u : \mathbb{R}_+ \to \mathbb{R}^n$ as

$$u(\lambda_j + s) := \xi_j, \quad \text{for any } j \geq 0.$$

Therefore, for any $t > 0$, it holds that

$$\dot{w}(t) = u(t) + d_{\Lambda(t)}.$$

Let

$$\delta(t) := \left\| w(t) - x_{\Lambda(t)} \right\| + \sup_{k \geq \Lambda(t)} \delta_k.$$

Then it holds that

$$\dot{w}(t) - u(t) \in \mathcal{D}^{\delta(t)}(w(t)).$$

From the definition of $\delta(t)$, it holds that

$$\limsup_{t \to +\infty} \delta(t) \leq \limsup_{t \to +\infty} \left(\eta_{\Lambda(t)} \left\| \xi_{\Lambda(t)} + d_{\Lambda(t)} \right\| + \sup_{k \geq \Lambda(t)} \delta_k \right) = 0.$$

In addition, from the definition of $u(t)$, we achieve

$$\lim_{t \to +\infty} \sup_{0 \leq l \leq T} \left\| \int_t^{t+l} u(s) \mathrm{d}s \right\| = \lim_{s \to +\infty} \sup_{s \leq i \leq \Lambda(\lambda_s + T)} \left\| \sum_{k=s}^{i} \eta_k \xi_k \right\| = 0.$$

Therefore, from Definition 16, we can conclude that w is a perturbed solution for (4). This completes the proof. ■

3. A General Framework for Convergence Properties

3.1 Convergence to $\mathcal{D}_f$-stationary Points

In this subsection, we aim to show the convergence properties of the proposed abstract framework (AFM). We first make the following assumptions on f.

Assumption 1 *For the problem UNP, we assume f is locally Lipschitz continuous, bounded from below. Moreover, there exists a compact convex valued mapping $\mathcal{D}_f : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ that has closed graph such that*

1. *f is a potential function that admits $\mathcal{D}_f$ as its conservative field.*
2. *The set $\{f(x) : 0 \in \mathcal{D}_f(x)\}$ has empty interior in $\mathbb{R}$. That is, the complementary of $\{f(x) : 0 \in \mathcal{D}_f(x)\}$ is dense in $\mathbb{R}$.*

As discussed in Section 2.3, Assumption 1(1) is satisfied in a wide range of applications of UNP. Moreover, Assumption 1(2) is the weak Sard's theorem, which has been shown to be a mild assumption, as demonstrated in (Davis et al., 2020; Castera et al., 2021).

Furthermore, we make the following assumptions on the framework (AFM).

Assumption 2

1. *The parameters satisfy $\alpha, \gamma \geq 0$, and $\varepsilon, \tau_1, \tau_2 > 0$.*
2. *The sequences of iterates $\{x_k\}$, $\{m_k\}$ and $\{v_k\}$ are almost surely bounded, i.e.,*

$$\sup_{k\geq 0} \|x_k\| + \|m_k\| + \|v_k\| < +\infty$$

holds almost surely.

3. *$\mathcal{U}$ is a locally bounded set-valued mapping that is convex compact valued with closed graph. Moreover, there exists a constant $\kappa \geq 0$ such that for any $x, m, v \in \mathbb{R}^n$, it holds that*

$$\widetilde{\mathrm{sign}}(v) \odot \mathcal{U}(x, m, v) \geq \kappa |v| \geq 0.$$

4. *$(d_{x,k}, d_{m,k}, d_{v,k})$ is an approximated evaluation for $\mathcal{G}(x_k, m_k, v_k)$ in the sense that there exists a positive sequence of real numbers $\{\delta_k\}$ such that $\lim_{k\to+\infty} \delta_k = 0$ and*

$$(d_{x,k}, d_{m,k}, d_{v,k}) \in \mathcal{G}^{\delta_k}(x_k, m_k, v_k).$$

5. *$\{(\xi_{x,k}, \xi_{m,k}, \xi_{v,k})\}$ is a uniformly bounded martingale sequence. That is, almost surely, it holds for any $k \geq 1$ that*

$$\mathbb{E}[(\xi_{x,k}, \xi_{m,k}, \xi_{v,k})|\mathcal{F}_{k-1}] = 0, \quad \text{and} \quad \sup_{k\geq 0} \|(\xi_{x,k}, \xi_{m,k}, \xi_{v,k})\| < +\infty.$$

6. *The stepsizes $\{\eta_k\}$ and $\{\theta_k\}$ are positive and satisfy*

$$\sum_{k=0}^{+\infty} \eta_k = +\infty, \ \sum_{k=0}^{+\infty} \theta_k = +\infty, \ \lim_{k\to+\infty} \eta_k \log(k) = 0, \text{ and } \lim_{k\to+\infty} \frac{\theta_k^2}{\eta_k} \log(k) = 0.$$

Here are some comments for Assumption 2. Assumption 2(2) assumes that the generated sequence $\{(x_k, m_k, v_k)\}$ and the updating directions $\{(d_{x,k}, d_{m,k}, d_{v,k})\}$ are uniformly bounded, which is a common assumption in various existing works (Benaïm et al., 2005; Benaïm, 2006; Davis et al., 2020; Bolte and Pauwels, 2021; Castera et al., 2021). Assumption 2(3) enforces regularity conditions on the set-valued mapping $\mathcal{U}$, which are satisfied in a wide range of adaptive stochastic gradient methods such as Adam, AdaBelief, AMSGrad, NAdam, Yogi, as discussed later in Section 4. Assumption 2(4) illustrates how $(d_{x,k}, d_{m,k}, d_{v,k})$ approximates $\mathcal{G}(x_k, m_k, v_k)$, which is a mild assumption commonly used in existing works (Benaïm et al., 2005; Benaïm, 2006; Bolte and Pauwels, 2021; Castera et al., 2021). In addition, Assumption 2(5) is a prevalent assumption in various existing works (Bolte and Pauwels, 2021; Castera et al., 2021).

Furthermore, Assumption 2(6) allows for a flexible choice of the stepsize η_k in (AFM), enabling it to be set in the order of $o(1/\log(k))$. It is easy to verify that a simple choice of $\theta_k = \eta_k$ satisfies Assumption 2(6). Hence the framework (AFM) includes those cases where the evaluation noises are uniformly bounded. More importantly, Assumption 2(6) allows for a two-timescale scheme for (AFM) in the sense that we can choose $\{\theta_k\}$ satisfying $\theta_k/\eta_k \to +\infty$, since we can always set $\theta_k = \eta_k \left(\eta_k \log(k)\right)^{-s}$ for any $s \in (0, \frac{1}{2})$. As shown in Section 4, the two-timescale framework is crucial for developing adaptive subgradient methods with gradient clipping techniques when the evaluation noises are only assumed to be integrable. The two-timescale updating scheme assumed in Assumption 2 distinguishes our proposed framework (AFM) from existing frameworks in (Benaïm et al., 2005; Davis et al., 2020; Bolte and Pauwels, 2021).

Remark 1 *It is worth mentioning that some existing works have investigated the conditions to ensure the sequence of iterates to be uniformly bounded almost surely, including the proximal stochastic subgradient descent method (Davis et al., 2020), the SGD method with constant stepsize (Bianchi et al., 2022), and noiseless heavy-ball SGD method (Josz and Lai, 2023). However, their analysis is limited within specific subgradient methods, making it challenging to establish similar results for adaptive methods.*

On the other hand, when f is assumed to be differentiable, Barakat and Bianchi (2021); Gadat and Gavra (2022); Li and Milzarek (2022) utilize the Robbins-Siegmund theorem (Robbins and Siegmund, 1971) to prove that the function values of their merit functions are uniformly bounded almost surely. Then based on the coercivity of their employed merit functions, they establish the uniform boundedness of the sequence generated by adaptive methods. However, when f is assumed to be nonsmooth and nonconvex, it is challenging to estimate the decrease of the objective function over the iterations. As a result, their proof techniques cannot be applied to prove the uniform boundedness of $\{(x_k, m_k, v_k)\}$ within the context of the framework (AFM) *under nonsmooth settings. How to establish the uniform boundedness for the sequence $\{(x_k, m_k, v_k)\}$ under the framework* (AFM) *still remains open.*

To prove the convergence properties of (AFM), we consider the following differential inclusion

$$\left(\frac{\mathrm{d}x}{\mathrm{d}t}, \frac{\mathrm{d}m}{\mathrm{d}t}, \frac{\mathrm{d}v}{\mathrm{d}t}\right) \in -\mathcal{G}(x, m, v). \tag{7}$$

Let the function $\phi : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be defined as,

$$\phi(x, m, v) := f(x) + \frac{1}{2\tau_1} \left\langle m, (|v| + \varepsilon)^{-\gamma} \odot m \right\rangle,$$

and let the set $\mathcal{B}$ be chosen as $\mathcal{B} := \{(x, m, v) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n : 0 \in \mathcal{D}_f(x), m = 0\}$. Then we have the following lemma to illustrate the relationship between $\mathcal{D}_f$-stationary points of f and $\mathcal{B}$. The proof for Lemma 5 is straightforward, hence we omit it for simplicity.

Lemma 5 *The function ϕ is locally Lipschitz continuous over $\mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n$, and $\{f(x) : 0 \in \mathcal{D}_f(x)\} = \{\phi(x, m, v) : (x, m, v) \in \mathcal{B}\}$.*

In the following lemma, we illustrate that ϕ is a potential function whenever f is a potential function for $\mathcal{D}_f$, and investigate the expression of the conservative field for ϕ. Lemma 6 directly follows from the expressions for ϕ and the validity of the chain rule for the conservative field (Bolte and Pauwels, 2021), hence we omit the proof for Lemma 6 for simplicity.

Lemma 6 *Suppose f is a potential function that admits $\mathcal{D}_f$ as its conservative field, then ϕ is a potential function that admits the conservative field $\mathcal{D}_\phi$ defined by*

$$\mathcal{D}_\phi(x, m, v) = \begin{bmatrix} \mathcal{D}_f(x) \\ \frac{1}{\tau_1}(|v| + \varepsilon)^{-\gamma} \odot m \\ -\frac{\gamma}{2\tau_1} m^2 \odot (|v| + \varepsilon)^{-\gamma-1} \odot \mathrm{sign}(v) \end{bmatrix}.$$

The following proposition illustrates that ϕ is a Lyapunov function for $\mathcal{B}$ with respect to the differential inclusion (7). When f is assumed to be differentiable, similar Lyapunov functions have been proposed in (Barakat et al., 2021; Barakat and Bianchi, 2021; Gadat and Gavra, 2022).

Proposition 3 *Suppose Assumption 1 and Assumption 2 hold with $(1 - \kappa)\gamma\tau_2 \leq 2\tau_1$. For any $(x_0, m_0, v_0) \notin \mathcal{B}$, let $(x(t), m(t), v(t))$ be any trajectory of the differential inclusion* (7) *with initial point (x_0, m_0, v_0). Then for any $t > 0$, it holds that*

$$\phi(x(t), m(t), v(t)) < \phi(x(0), m(0), v(0)).$$

That is, ϕ is the Lyapunov function for $\mathcal{B}$ with respect to the differential inclusion (7).

Proof From the definition of (7), for any $(x(t), m(t), v(t))$ that is a trajectory of the differential inclusion (7), there exists measurable functions $l_f(s)$ and $l_v(s)$ such that for almost every $s \geq 0$, $l_f(s) \in \mathcal{D}_f(x(s))$, $l_v(s) \in \mathcal{U}(x(s), m(s), v(s))$, and

$$(\dot{x}(s), \dot{m}(s), \dot{v}(s)) = - \begin{bmatrix} (|v(s)| + \varepsilon)^{-\gamma} \odot \left(m(s) + \alpha l_f(s)\right) \\ \tau_1 m(s) - \tau_1 l_f(s) \\ \tau_2 v(s) - \tau_2 l_v(s) \end{bmatrix}.$$

Therefore, from the expression of $\mathcal{D}_\phi$, we can conclude that

$$\begin{aligned}
&\left\langle (\dot{x}(s), \dot{m}(s), \dot{v}(s)), \mathcal{D}_\phi(x(s), m(s), v(s)) \right\rangle \\
=\ & \left\langle \mathcal{D}_f(x(s)), -(|v(s)|+\varepsilon)^{-\gamma} \odot \left(m(s) + \alpha l_f(s)\right) \right\rangle \\
&+ \frac{1}{\tau_1} \left\langle (|v(s)|+\varepsilon)^{-\gamma} \odot m(s), -\tau_1 m(s) + \tau_1 l_f(s) \right\rangle \\
&+ \frac{-\gamma}{2\tau_1} \left\langle m(s)^2 \odot (|v(s)|+\varepsilon)^{-\gamma-1} \odot \widetilde{\mathrm{sign}}(v(s)), (-\tau_2 v(s) + \tau_2 l_v(s)) \right\rangle \\
\ni\ & -\alpha \left\langle l_f(s), (|v(s)|+\varepsilon)^{-\gamma} \odot l_f(s) \right\rangle \\
&- \left\langle m(s), (|v(s)|+\varepsilon)^{-\gamma} \odot m(s) \right\rangle + \frac{\gamma\tau_2}{2\tau_1} \left\langle m(s)^2, (|v(s)|+\varepsilon)^{-\gamma-1} \odot |v(s)| \right\rangle \\
&- \frac{\gamma\tau_2}{2\tau_1} \left\langle m(s)^2 \odot (|v(s)|+\varepsilon)^{-\gamma-1}, \widetilde{\mathrm{sign}}(v(s)) \odot l_v(s) \right\rangle.
\end{aligned}$$

Notice that for any $v \in \mathbb{R}^n$,

$$(|v|+\varepsilon)^{-\gamma-1} \odot |v| = (|v|+\varepsilon)^{-\gamma} - \varepsilon(|v|+\varepsilon)^{-\gamma-1},$$

we have that

$$\begin{aligned}
&- \left\langle m(s), (|v(s)|+\varepsilon)^{-\gamma} \odot m(s) \right\rangle + \frac{\gamma\tau_2}{2\tau_1} \left\langle m(s)^2, (|v(s)|+\varepsilon)^{-\gamma-1} \odot |v(s)| \right\rangle \\
=\ & -\left(1 - \frac{\gamma\tau_2}{2\tau_1}\right) \left\langle m(s), (|v(s)|+\varepsilon)^{-\gamma} \odot m(s) \right\rangle - \frac{\varepsilon\gamma\tau_2}{2\tau_1} \left\langle m(s), (|v(s)|+\varepsilon)^{-\gamma-1} \odot m(s) \right\rangle.
\end{aligned} \tag{8}$$

Moreover, Assumption 2(3) illustrates that $\widetilde{\mathrm{sign}}(v(s)) \odot U(x(s), m(s), v(s)) \geq \kappa |v(s)|$. Hence for any $s > 0$, we have

$$\begin{aligned}
&\left\langle m(s)^2 \odot (|v(s)|+\varepsilon)^{-\gamma-1}, \widetilde{\mathrm{sign}}(v(s)) \odot l_v(s) \right\rangle \\
\geq\ & \kappa \left\langle m(s)^2 \odot (|v(s)|+\varepsilon)^{-\gamma-1}, |v(s)| \right\rangle \geq 0.
\end{aligned} \tag{9}$$

Additionally, under Assumption 2, let the positive constant δ_γ be defined as

$$\delta_\gamma = \begin{cases} 1, & \gamma = 0; \\ \frac{\gamma\tau_2}{2\tau_1}, & \gamma > 0. \end{cases}$$

Then when $\gamma = 0$, it holds that

$$\begin{aligned}
&\left(1 - \frac{(1-\kappa)\gamma\tau_2}{2\tau_1}\right) \left\langle m(s), (|v(s)|+\varepsilon)^{-\gamma} \odot m(s) \right\rangle = \left\langle m(s), (|v(s)|+\varepsilon)^{-\gamma} \odot m(s) \right\rangle \\
\geq\ & \varepsilon \left\langle m(s), (|v(s)|+\varepsilon)^{-\gamma-1} \odot m(s) \right\rangle = \varepsilon\delta_\gamma \left\langle m(s), (|v(s)|+\varepsilon)^{-\gamma-1} \odot m(s) \right\rangle
\end{aligned}$$

On the other hand, when $\gamma > 0$, it holds from the definition of δ_γ that

$$\frac{\varepsilon\gamma\tau_2}{2\tau_1} \left\langle m(s), (|v(s)|+\varepsilon)^{-\gamma-1} \odot m(s) \right\rangle = \varepsilon\delta_\gamma \left\langle m(s), (|v(s)|+\varepsilon)^{-\gamma-1} \odot m(s) \right\rangle.$$

Therefore, we can conclude that under Assumption 2, it holds for any $\gamma \geq 0$ that

$$\begin{aligned}&\left(1 - \frac{(1-\kappa)\gamma\tau_2}{2\tau_1}\right) \left\langle m(s), (|v(s)| + \varepsilon)^{-\gamma} \odot m(s) \right\rangle + \frac{\varepsilon\gamma\tau_2}{2\tau_1} \left\langle m(s), (|v(s)| + \varepsilon)^{-\gamma-1} \odot m(s) \right\rangle \\ &\geq \varepsilon\delta_\gamma \left\langle m(s), (|v(s)| + \varepsilon)^{-\gamma-1} \odot m(s) \right\rangle .\end{aligned} \tag{10}$$

As a result, for any $s \geq 0$, we have

$$\begin{aligned}&\inf_{d_\phi \in \mathcal{D}_\phi(x(s),m(s),v(s))} \left\langle (\dot{x}(s), \dot{m}(s), \dot{v}(s)), d_\phi \right\rangle \\ \leq\; & -\alpha \left\langle l_f(s), (|v(s)| + \varepsilon)^{-\gamma} \odot l_f(s) \right\rangle \\ & - \left\langle m(s), (|v(s)| + \varepsilon)^{-\gamma} \odot m(s) \right\rangle + \frac{\gamma\tau_2}{2\tau_1} \left\langle m(s)^2, (|v(s)| + \varepsilon)^{-\gamma-1} \odot |v(s)| \right\rangle \\ & - \frac{\gamma\tau_2}{2\tau_1} \left\langle m(s)^2 \odot (|v(s)| + \varepsilon)^{-\gamma-1}, \widetilde{\mathrm{sign}}(v(s)) \odot l_v(s) \right\rangle \\ \leq\; & - \left\langle m(s), (|v(s)| + \varepsilon)^{-\gamma} \odot m(s) \right\rangle + \frac{\gamma\tau_2}{2\tau_1} \left\langle m(s)^2, (|v(s)| + \varepsilon)^{-\gamma-1} \odot |v(s)| \right\rangle \\ & - \frac{\gamma\tau_2}{2\tau_1} \left\langle m(s)^2 \odot (|v(s)| + \varepsilon)^{-\gamma-1}, \widetilde{\mathrm{sign}}(v(s)) \odot l_v(s) \right\rangle \\ \leq\; & -\varepsilon\delta_\gamma \left\langle m(s), (|v(s)| + \varepsilon)^{-\gamma-1} \odot m(s) \right\rangle .\end{aligned}$$

Here the last inequality directly follows (8), (9), and (10).

Then for any $t > 0$, from Definition 12, we have

$$\begin{aligned}&\phi(x(t), m(t), v(t)) - \phi(x(0), m(0), v(0)) \\ =\; & \int_0^t \inf_{d_\phi \in \mathcal{D}_\phi(x(s),m(s),v(s))} \left\langle (\dot{x}(s), \dot{m}(s), \dot{v}(s)), d_\phi \right\rangle \mathrm{d}s \\ \leq\; & -\int_0^t \varepsilon\delta_\gamma \left\langle m(s), (|v(s)| + \varepsilon)^{-\gamma-1} \odot m(s) \right\rangle \mathrm{d}s \leq 0.\end{aligned}$$

Therefore, for any $s_1 > s_2 \geq 0$, we have

$$\phi(x(s_1), m(s_1), v(s_1)) \leq \phi(x(s_2), m(s_2), v(s_2)), \tag{11}$$

which illustrates that $\phi(x(t), m(t), v(t))$ is non-increasing for any $t > 0$.

Now we prove that ϕ is a Lyapunov function for $\mathcal{B}$. When $(x, m, v) \notin \mathcal{B}$, we first consider the cases where $m(0) \neq 0$. From the continuity of the path $(x(t), m(t), v(t))$, there exists $T > 0$ such that $m(t) \neq 0$ for any $t \in [0, T]$. Therefore, we can conclude that for any $t > 0$,

$$\begin{aligned}&\phi(x(t), m(t), v(t)) - \phi(x(0), m(0), v(0)) \\ \leq\; & -\int_0^t \varepsilon\delta_\gamma \left\langle m(s), (|v(s)| + \varepsilon)^{-\gamma-1} \odot m(s) \right\rangle \\ \leq\; & -\int_0^{\min\{t,T\}} \varepsilon\delta_\gamma \left\langle m(s), (|v(s)| + \varepsilon)^{-\gamma-1} \odot m(s) \right\rangle < 0.\end{aligned}$$

On the other hand, when $(x(0), m(0), v(0)) \notin \mathcal{B}$ with $m(0) = 0$, then it holds that $0 \notin \mathcal{D}_f(x(0))$. Therefore, by the hyperplane separation theorem, there exists $w \in \mathbb{R}^n$ such that $\|w\| = 1$ and $\inf_{d \in \mathcal{D}_f(x(0))} \langle d, w \rangle > 0$. Then from the outer-semicontinuity of $\mathcal{D}_f$ and the continuity of $m(t)$, there exists a constant $c > 0$ and time $T > 0$ such that

$$\inf_{d \in \mathcal{D}_f(x(t))} \langle d, w \rangle > c, \quad \text{and} \quad \|m(t)\| \leq \frac{c}{2}, \quad \forall t \in [0, T].$$

As a result, for any $t \in [0, T]$, we have

$$\begin{aligned} \langle m(t), w \rangle = \int_0^t \langle \dot{m}(s), w \rangle \, \mathrm{d}s = \int_0^t \tau_1 \left\langle m(s) - \mathcal{D}_f(x(s)), w \right\rangle \mathrm{d}s \\ \leq \int_0^t -\frac{\tau_1 c}{2} \mathrm{d}s = -\frac{\tau_1 c t}{2} < 0. \end{aligned}$$

Then for any $t > 0$, it holds that

$$\int_0^t \|m(s)\|^2 \, \mathrm{d}s \geq \int_0^t \langle m(s), w \rangle^2 \, \mathrm{d}s > \int_0^t \left(\frac{\tau_1 c s}{2} \right)^2 \mathrm{d}s > 0.$$

As a result, let $M := \sup_{t \in [0,T]} \| |v(t)| + \varepsilon \|$, we achieve

$$\begin{aligned} &\phi(x(t), m(t), v(t)) - \phi(x(0), m(0), v(0)) \\ \leq & -\int_0^{\min\{t,T\}} \varepsilon \delta_\gamma \left\langle m(s), (|v(s)| + \varepsilon)^{-\gamma - 1} \odot m(s) \right\rangle \\ \leq & -\frac{\varepsilon \delta_\gamma}{M^{\gamma+1}} \int_0^{\min\{t,T\}} \|m(s)\|^2 \, \mathrm{d}s \ < \ 0. \end{aligned}$$

Thus we can conclude that

$$\phi(x(t), m(t), v(t)) < \phi(x(0), m(0), v(0)),$$

holds whenever $(x, m, s) \notin \mathcal{B}$. Hence, ϕ is a Lyapunov function for the set $\mathcal{B}$ with respect to the differential inclusion (7). This completes the proof. ■

Proposition 3 requires the parameters in the framework (AFM) to satisfy $(1 - \kappa)\gamma\tau_2 \leq 2\tau_1$. When κ is 0, the Assumption 2(3) becomes $\widetilde{\mathrm{sign}}(v) \odot \mathcal{U}(x, m, v) \geq 0$ for any $x, m, v \in \mathbb{R}^n$. As demonstrated in Section 4, this condition can be satisfied by numerous popular Adam-family methods, including Adam, AdaBelief, AMSGrad, NAdam, and Yogi. Moreover, it can be proven that the corresponding set-valued mapping $\mathcal{U}$ in AMSGrad satisfies Assumption 2(3) with $\kappa = 1$. Consequently, as discussed later in Corollary 2, AMSGrad converges with any combinations of $\tau_1, \tau_2 > 0$ within the framework (AFM).

Based on Proposition 3 and (Benaïm et al., 2005, Proposition 3.27), the following theorem illustrates the global almost sure convergence properties for (AFM).

Theorem 1 *For any sequence $\{(x_k, m_k, v_k)\}$ generated by* (AFM)*. Suppose Assumption 1 and Assumption 2 hold, and the parameters in* (AFM) *satisfy $(1 - \kappa)\gamma\tau_2 \leq 2\tau_1$. Then almost surely, any cluster point of $\{x_k\}$ lies in $\{x \in \mathbb{R}^n : 0 \in \mathcal{D}_f(x)\}$, and the sequence $\{f(x_k)\}$ converges.*

Proof Firstly, we aim to prove the uniform boundedness of the sequence $\{(d_{x,k}, d_{m,k}, d_{v,k})\}$. From Assumption 2(4), it is easy to verify that $\lim_{k\to+\infty} \mathrm{dist}(m_k - d_{m,k}, \tau_1 \mathcal{D}_f^{\delta_k}(x_k)) = 0$. Moreover, Assumption 2(2) assumes that the sequence of iterates $\{(x_k, m_k, v_k)\}$ is uniformly bounded almost surely. Then as the local boundedness of $\mathcal{D}_f$ and $\mathcal{U}$ implies the local boundedness of $\mathcal{G}$, we can conclude that the sequence $\{(d_{x,k}, d_{m,k}, d_{v,k})\}$ is uniformly bounded almost surely. In addition, as $\mathcal{D}_f$ and $\mathcal{U}$ are assumed to be convex-valued, it holds that the set-valued mapping $\mathcal{G}$ is also convex-valued.

Moreover, Proposition 1 controls the summation of the noises in the framework (AFM) (i.e., $\sum_{k=j}^{i} \theta_k(\xi_{x,k}, \xi_{m,k}, \xi_{v,k})$). From Proposition 1 and Assumption 2(5), for any $T > 0$, almost surely, we have that,

$$\lim_{j\to+\infty} \sup_{j\leq i\leq \Lambda(\lambda_j+T)} \left\| \sum_{k=j}^{i} \theta_k(\xi_{x,k}, \xi_{m,k}, \xi_{v,k}) \right\| = 0.$$

Together with Assumption 2(2), we can conclude that for almost every $\omega \in \Omega$, the interpolation for $\{(x_k(\omega), m_k(\omega), v_k(\omega))\}$ is a perturbed solution for the differential inclusion (7). Then Proposition 3 illustrates that ϕ is a Lyapunov function for the set $\mathcal{B}$ with respect to the differential inclusion (7).

Let $M_{m,\omega} := \sup_{k\geq 0} \|m_k(\omega)\|$, $M_{v,\omega} := \sup_{k\geq 0} \|v_k(\omega)\|$, $M_{x,\omega} := \sup_{k\geq 0} \|x_k(\omega)\|$ and $\mathcal{C}_\omega := \mathbb{B}_{M_{x,\omega}}(0) \times \mathbb{B}_{M_{m,\omega}}(0) \times \mathbb{B}_{M_{v,\omega}}(0)$, then the set $\mathcal{B} \cap \mathcal{C}_\omega$ is a compact set. Then for almost every $\omega \in \Omega$, (Benaïm et al., 2005, Proposition 3.27) illustrates that any cluster point of $\{(x_k(\omega), m_k(\omega), v_k(\omega))\}$ lies in $\mathcal{B} \cap \mathcal{C}_\omega$, and $\{\phi(x_k(\omega), m_k(\omega), v_k(\omega))\}$ converges. From Lemma 5, we can conclude that for almost every $\omega \in \Omega$, any limit point of $\{x_k(\omega)\}$ lies in $\{x : 0 \in \mathcal{D}_f(x)\}$, and $\{f(x_k(\omega))\}$ converges. Hence we complete the proof. ■

3.2 Convergence to ∂f-stationary Points with Random Initialization

In training nonsmooth neural networks, the conservative fields associated with AD algorithms may introduce infinitely many spurious stationary points (Bolte and Pauwels, 2020; Bolte et al., 2021; Bianchi et al., 2022). To address these issues, several existing works Bolte and Pauwels (2020); Bianchi et al. (2022) demonstrate that the SGD method can avoid the spurious stationary points introduced by conservative field almost surely. For the SGD method with diminishing stepsizes and mini-batch random sampling, Bolte and Pauwels (2020) prove that it can avoid these spurious stationary points almost surely. Moreover, Bianchi et al. (2022) prove that with randomly chosen initial points and stepsizes, the vanilla SGD method with constant stepsizes converges to a neighborhood of the ∂f-stationary points of UNP. Their results guarantee that the vanilla SGD method is able to yield meaningful stationary points in training nonsmooth neural works, regardless of the chosen conservative field.

In this subsection, to establish similar properties for Adam-family methods, we adopt the techniques from (Bolte and Pauwels, 2020; Bianchi et al., 2022) and extend their results to analyze the following abstract framework that employs diminishing stepsizes,

$$z_{k+1} - z_k \in -c s_k \mathcal{Q}_k(z_k, \omega_k), \tag{12}$$

where $z_k \in \mathbb{R}^d$ refers to the iteration points, $s_k \in \mathbb{R}$ refers to the stepsizes, $c \in \mathbb{R}$ is a scaling parameter for the stepsizes, Ξ is a probability space and $\{\omega_k\} \subset \Xi$ characterizes the stochasticity in (12). Moreover, for any $k \geq 0$, $\mathcal{Q}_k : \mathbb{R}^d \times \Xi \rightrightarrows \mathbb{R}^d$ is a set-valued mapping. Furthermore, for almost every ω, we assume the set-valued mapping $\mathcal{Q}_k(\cdot, \omega) : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is almost everywhere $\mathcal{C}^1$ for any $k \geq 0$.

Definition 18 *A measurable mapping $q : \mathbb{R}^d \to \mathbb{R}^d$ is almost everywhere $\mathcal{C}^1$ if for almost every $z \in \mathbb{R}^d$, q is locally continuously differentiable in a neighborhood of z.*

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

For any $k \geq 0$, let $q_k : \mathbb{R}^d \times \Xi \to \mathbb{R}^d$ be the mapping that is locally continuously differentiable for almost every $z \in \mathbb{R}^d$ and $\mathcal{Q}_k(z, \omega) = \{q_k(z, \omega)\}$ holds for almost every $\omega \in \Xi$. Then we define the mapping $T_{s,\omega,k}$ as

$$T_{s,\omega,k}(z) = z - s q_k(z, \omega).$$

The following proposition illustrates that for almost every $s \in \mathbb{R}$, the mapping $T_{s,\omega,k}^{-1}$ maps zero-measure subsets into the zero-measure subsets over $\mathbb{R}^d$.

Proposition 4 *Suppose $q_k(\cdot, \omega)$ is almost everywhere $\mathcal{C}^1$ for almost every $\omega \in \Xi$ and any $k \geq 0$. Then for almost every $\omega \in \Xi$, there exists a full-measure subset S_k of $\mathbb{R}$ such that for any $s \in S_k$ and any zero-measure set $A \subset \mathbb{R}^d$, the subset*

$$\{z \in \mathbb{R}^d : T_{s,\omega,k}(z) \in A\} \tag{13}$$

is zero-measure.

Proof From Definition 18, there exists a full-measure subset $\Gamma_\phi \subseteq \mathbb{R}^d$ such that for almost every $\omega \in \Omega$ and any $z \in \Gamma_\phi$, $q_k(\cdot, \omega)$ is continuously differentiable in a neighborhood of $z \in \Gamma_\phi$.

For any fixed $k \geq 0$, $\omega \in \Omega$, and $z \in \Gamma_\phi$, denote the Jacobian of q_k with respect to z as J_{q_k}. Then the Jacobian of $T_{s,\omega,k}$ can be expressed as

$$J_{T_{s,\omega,k}}(z) = I_n - s J_{q_k}(z, \omega).$$

Notice that for any fixed $z \in \Gamma_\phi$, $\det(J_{T_{s,\omega,k}}(z))$ is a non-trivial d-th order polynomial of s in $\mathbb{R}$, hence its roots are zero-measure in $\mathbb{R}$. Therefore, we can conclude that for any $\omega \in \Omega$ and any $z \in \Gamma_\phi$,

$$\{s \in \mathbb{R} : \det(J_{T_{s,\omega,k}}(z)) = 0\}$$

is zero-measure in $\mathbb{R}$. From Fubini's theorem, we can conclude that for almost every given $\omega \in \Omega$, there exists a full-measure subset S_k of $\mathbb{R}$ such that for any $s \in S_k$,

$$\Gamma_{\phi,s,\omega,k} := \{z \in \Gamma_\phi : \det(J_{T_{s,\omega,k}}(z)) \neq 0\} \tag{14}$$

is full-measure in $\mathbb{R}^d$.

By inverse function theorem, for any $k \geq 0$, almost every given $\omega \in \Omega$, $s \in S_k$ and $z \in \Gamma_{\phi,s,\omega,k}$, the mapping $T_{s,\omega,k}$ is a local diffeomorphism in a neighborhood of z. Let $\tilde{V}_z := \mathbb{B}_z(\tilde{\delta}_z)$, where $\tilde{\delta}_z > 0$ and $T_{s,\omega,k}$ is a local diffeomorphism in $\mathcal{B}_z(\tilde{\delta}_z)$ (i.e., $T_{s,\omega,k}$ is continuously differentiable and has non-singular Jacobian over $\mathcal{B}_z(\tilde{\delta}_z)$). Therefore, $V_z := T_{s,\omega,k}(\mathcal{B}_z(\tilde{\delta}_z))$ is an open set for any $z \in \Gamma_{\phi,s,\omega,k}$.

Notice that $\{V_z\}_{z\in\Gamma_{\phi,s,\omega,k}}$ is an open cover for $\Gamma_{\phi,s,\omega,k}$. Based on Lindelof's lemma (Kelley, 2017) and the fact that $\mathbb{R}^d$ is a second-countable space, there exists $\{z_i\}_{i\in\mathbb{N}_+} \subset \Gamma_{\phi,s,\omega,k}$ such that $\Gamma_{\phi,s,\omega,k} \subseteq \bigcup_{i\in\mathbb{N}_+} V_{z_i} \subseteq \bigcup_{i\in\mathbb{N}_+} \tilde{V}_{z_i}$. Given any zero-measure set $A \subset \mathbb{R}^d$, for any $i \in \mathbb{N}_+$, since $T_{s,\omega,k}$ is a local diffeomorphism in $\tilde{V}_{z_i}$, we can conclude that the set $\{z \in \tilde{V}_{z_i} : T_{s,\omega,k}(z) \in A\}$ is zero-measure. Then the set

$$\bigcup_{i\in\mathbb{N}_+} \{z \in \tilde{V}_{z_i} : T_{s,\omega,k}(z) \in A\}$$

is zero-measure, hence the set $\{z \in \Gamma_{\phi,s,\omega,k} : T_{s,\omega,k}(z) \in A\}$ is zero-measure. Combined with the fact that $\Gamma_{\phi,s,k}$ is a full-measure subset of $\mathbb{R}^d$, we can conclude that the set $\{z \in \mathbb{R}^d : T_{s,\omega,k}(z) \in A\}$ satisfies

$$\{z \in \mathbb{R}^d : T_{s,\omega,k}(z) \in A\} \subset \Gamma^c_{\phi,s,\omega,k} \cup \{z \in \Gamma_{\phi,s,\omega,k} : T_{s,\omega,k}(z) \in A\}.$$

Then from the definition of $\Gamma_{\phi,s,\omega,k}$ in (14), for almost every $\omega \in \Xi$, for any $s \in S_k$, the set $\Gamma^c_{\phi,s,\omega,k}$ is zero-measure. This completes the proof. ■

Based on Proposition 4, the following proposition illustrates that for almost every prefixed $\omega \in \Xi$ and $s \in \mathbb{R}_+$, the mapping $T^{-1}_{s,\omega,0} \circ T^{-1}_{s,\omega,1} \circ \cdots T^{-1}_{s,\omega,k}$ maps zero-measure subsets of $\mathbb{R}^d$ into zero-measure subsets for all $k \geq 0$.

Proposition 5 *Suppose $\mathcal{Q}_k(\cdot,\omega)$ is almost everywhere $\mathcal{C}^1$ for almost every $\omega \in \Xi$ and any $k \geq 0$. Then for any given zero-measure set $A \subset \mathbb{R}^d$, for almost every $\omega \in \Xi$, there exists a full-measure subset $S_{init,\omega}$ of $\mathbb{R}^d \times \mathbb{R}$ such that for any sequence $\{z_k\}$ that employ the following update scheme*

$$z_{k+1} = z_k - s\mathcal{Q}_k(z_k, \omega), \tag{15}$$

with $(z_0, s) \in S_{init,\omega}$, we have that

$$\{z_k : k \geq 0\} \bigcap A = \emptyset.$$

Proof According to Proposition 4, given any zero-measure subset $A \subset \mathbb{R}^d$, for any $k \geq 0$ and $\omega \in \Xi$, there exists a full-measure subset $S_{\omega,k} \subseteq \mathbb{R}_+$ such that for any $s \in S_{\omega,k}$ and any zero-measure subset A of $\mathbb{R}^d$, the set $\{z \in \mathbb{R}^d : z - s\mathcal{Q}_k(z,\omega) \cap A \neq \emptyset\}$ is a zero-measure subset of $\mathbb{R}^d$.

Then let $S_\omega := \cup_{k\geq 0}\mathcal{S}_{\omega,k}$, it is easy to verify that for any $s \in S_\omega$, the set $\{z \in \mathbb{R}^d : z - s\mathcal{Q}_i(z) \cap A \neq \emptyset$ for some $i \geq 0\}$ is a zero-measure subset of $\mathbb{R}^d$. As a result, let $\tilde{Y}_{0,\omega,s} = A$, and recursively define

$$\tilde{Y}_{k+1,\omega,s} = \{z \in \mathbb{R}^d : (z - s\mathcal{Q}_i(z,\omega)) \cap (\tilde{Y}_{k,\omega,s} \cup A) \neq \emptyset \text{ for some } i \geq 0\}.$$

Then Proposition 4 illustrates that $\tilde{Y}_{k,\omega,s}$ is a zero-measure subset of $\mathbb{R}^d$ for any $k \geq 0$. Moreover, from the definition of $\tilde{Y}_{k,\omega,s}$, we can conclude that for any $j \geq 0$, any $s \in S_\omega$, and for any sequence $\{z_k\}$ that follows equation (15) with an initial condition $z_0 \notin \tilde{Y}_{j,\omega,s}$, it holds true that $\{z_k : k \leq j\} \cap A = \emptyset$.

Let $Y_{\omega,s} = (\cup_{k\geq 0}\tilde{Y}_{k,\omega,s})^c$, then for any $k \geq 0$, any $s \in S_\omega$, and any $z_0 \in Y_{\omega,s}$, we have that $Y_{\omega,s}$ is a zero-measure subset of $\mathbb{R}^d$ and the sequence $\{z_k\} \cap (\tilde{Y}_{k,\omega,s} \cup A) = \emptyset$. Then from Fubini's theorem, the subset $\{(z_0, s) : s \in S_\omega^c \text{ or } z_0 \in Y_{\omega,s}^c\}$ is a zero-measure subset of $\mathbb{R}^d \times \mathbb{R}_+$. As a result, $S_{init,\omega} = \{(z_0, s) : s \in S_\omega, z_0 \in Y_{\omega,s}\}$ is a full-measure subset of $\mathbb{R}^d \times \mathbb{R}_+$. Moreover, from the choices of S_ω and $Y_{\omega,s}$, for any $(z_0, s) \in S_{init,\omega}$, the sequence $\{z_k\}$ that follows the scheme in (15) satisfies $\{z_k\} \cap A = \emptyset$. This completes the proof. ■

Next, we present the following theorem to illustrate that under mild assumptions, with random initialization for the initial point z_0 and the scaling parameter c, the sequence generated by (12) can avoid any zero-measure subset A of $\mathbb{R}^d$. As shown later in Section 4 and Section 5, Theorem 2 directly implies the almost sure convergence to ∂f-stationary points of UNP for the analyzed stochastic subgradient methods.

Theorem 2 *Suppose $\mathcal{Q}_k(\cdot, \omega)$ is almost everywhere $\mathcal{C}^1$ for almost every $\omega \in \Xi$ and any $k \geq 0$. Then for any zero-measure subset $A \subset \mathbb{R}^d$, there exists a full-measure subset $S_{init} \subseteq \mathbb{R}^d \times \mathbb{R}$ and $S_\omega \subseteq \Omega$, such that for any $(z_0, c) \in S_{init}$, almost surely in Ξ, it holds that the sequence $\{z_k\}$ generated by* (12) *satisfies $\{z_k\} \subset A^c$.*

Proof As illustrated in Proposition 5, for almost every $\omega \in \Xi$, there exists a full-measure subset $S_{init,\omega}$ of $\mathbb{R}^d \times \mathbb{R}_+$ such that for any $(z_0, c) \in S_{init,\omega}$, almost surely in Ξ, it holds that the sequence $\{z_k\}$ generated by (12) satisfies $\{z_k\} \subset A^c$.

Notice that for almost every $\omega \in \Xi$, the set $S_{init,\omega}$ is a full-measure subset of $\mathbb{R}^d \times \mathbb{R}_+$. Then by applying Fubini's theorem, the set $T = \{(\omega, z_0, s) : \omega \in \Xi,\ (z_0, s) \in S_{init,\omega}\}$ is a full-measure subset of $\Xi \times \mathbb{R}^d \times \mathbb{R}_+$. Then by applying Fubini's theorem again, we can conclude that there exists a full-measure subset S_{init} of $\mathbb{R}^d$ such that for any $(z_0, s) \in S_{init}$, the subset $\{\omega : (\omega, z_0, s) \in T\}$ is a full-measure subset of Ξ. This completes the proof. ■

4. Applications: Convergence Guarantees for Adam-family Methods

In this section, we establish the convergence properties of ADAM, AMSGrad, Yogi and AdaBelief for solving UNP based on our proposed framework when the objective function f takes the following form,

$$f(x) := \mathbb{E}_{s\sim P}[f_s(x)]. \tag{16}$$

Here $(\Theta, \mathcal{F}, P)$ is a probability space, where Θ refers to the sample space, $\mathcal{F}$ is the corresponding σ-algebra, and P is the probability distribution. Throughout this section, we make the following assumptions on f.

Assumption 3 *There exists a measurable set-valued mapping $\mathcal{D} : \mathbb{R}^n \times \Theta \rightarrow \mathbb{R}^n$ such that*

1. *The mapping $(x, s) \mapsto f_s(x)$ is measurable over $\mathbb{R}^n \times \Theta$;*

2. *For almost every $s \in \mathbb{R}^n$, $x \mapsto \mathcal{D}(x,s)$ is a definable conservative field that admits f_s as its potential function. Moreover, there exists a measurable mapping $\chi : \mathbb{R}^n \times \Theta \to \mathbb{R}^n$ such that $\chi(x,s) \in \mathcal{D}(x,s)$ for all $x \in \mathbb{R}^n$ and almost every $s \in \Theta$.*

3. *There exists an integrable function $p_\Theta : \Theta \to \mathbb{R}_+$ such that for all $x \in \mathbb{R}^n$ and $s \in \Theta$, it holds that*
$$\sup_{d \in \mathcal{D}(x,s)} \|d\| \leq p_\Theta(s).$$

4. *The set $\{f(x) : 0 \in \mathrm{conv}\ (\mathbb{E}_{s\sim P}[\mathcal{D}(x,s)])\}$ has empty interior in $\mathbb{R}$.*

Based on the results from (Bolte et al., 2023, Theorem 3.10), the integral of $\mathcal{D}$ with respect to the measure P (i.e., $\mathbb{E}_{s\sim P}[\mathcal{D}(x,s)]$ as defined in Definition 9) is a conservative field that admits f as its potential function. As a result, in this section, we choose the conservative field $\mathcal{D}_f$ in the framework (AFM) as

$$\mathcal{D}_f(x) = \mathrm{conv}\ (\mathbb{E}_{s\sim P}[\mathcal{D}(x,s)]).$$

Moreover, the mapping χ defined in Assumption 3(2) is called a (measurable) selection of the set-valued mapping $\mathcal{D}$. From Definition 9, it is easy to verify that $\mathbb{E}_{s\sim P}[\chi(x,s)] \in \mathcal{D}_f(x)$ holds for all $x \in \mathbb{R}^n$.

Remark 2 *It is worth mentioning that Assumption 3 is mild in practice. For the loss function of any neural network that is built from definable blocks, Bolte and Pauwels (2021) show that the results returned by the AD algorithms are contained within a definable conservative field. This illustrates that Assumption 3(1)-(3) are easy to be satisfied in practice.*

Moreover, Bolte and Pauwels (2021, Theorem 5) illustrates that the set $\{f(x) : 0 \in \mathcal{D}_f(x)\}$ is finite whenever both f and $\mathcal{D}_f$ are definable. As discussed in (Bolte and Pauwels, 2021), when the set Θ is finite, the f can be expressed in a finite-sum formulation. Under such settings, both f and $\mathcal{D}_f$ are definable whenever both f_s and $\mathcal{D}(\cdot,s)$ are definable for any $s \in \Theta$. On the other hand, for the cases where Θ contains infinitely many elements, Bolte et al. (2023, Theorem 4.8) guarantees the definability of f and $\mathcal{D}_f$ under appropriate conditions. In particular, Bolte et al. (2023) shows that when we assume $\Theta = \mathbb{R}^q$ for some $q > 0$, the probability measure P has a semi-algebraic density function, and $\mathcal{D}$ is assumed to be convex-valued and semi-algebraic, then $\mathcal{D}_f$ is semi-algebraic. These results illustrate that Assumption 3(4) is also mild in practice.

Inspired by the pioneering works (Barakat and Bianchi, 2021; Barakat et al., 2021; Gadat and Gavra, 2022), we consider a class of Adam-family methods with diminishing stepsizes for minimizing f over $\mathbb{R}^n$. The detailed algorithm is presented in Algorithm 1. In step 6 of Algorithm 1, different Adam-family methods employ different schemes for updating the estimator $\{v_k\}$, which is characterized by a specific mapping $R_{\mathcal{U}} : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$. Table 3 summarizes the updating rules for Adam, AdaBelief, AMSGrad, NAdam and Yogi, their corresponding set-valued mappings $\mathcal{U}$ in the framework (AFM), and the settings for the parameters α and κ.

To establish the convergence properties for Algorithm 1, we make some mild assumptions as follows.

Algorithm 1 Stochastic subgradient-based Adam for nonsmooth optimization problems.

```
Require: Initial point $x_0 \in \mathbb{R}^n$, $m_0 \in \mathbb{R}^n$ and $v_0 \in \mathbb{R}^n_+$, parameters $\alpha \geq 0$, and $\tau_1, \tau_2, \varepsilon > 0$,
    and $\chi$ as a selection of stochastic subgradients;
Set $k = 0$;
while not terminated do
    Independently sample $s_k \sim P$, and compute $g_k = \chi(x_k, s_k)$;
    Choose the stepsize $\eta_k$;
    Update the momentum term by $m_{k+1} = (1 - \tau_1 \eta_k) m_k + \tau_1 \eta_k g_k$;
    Update the estimator $v_{k+1}$ from $g_k$, $m_{k+1}$ and $v_k$ by
        $$v_{k+1} = v_k - \tau_2 \eta_k R_{\mathcal{U}}(g_k, m_{k+1}, v_k).$$
    Compute the scaling parameters $\rho_{m,k+1}$ and $\rho_{v,k+1}$;
    Update $x_k$ by
        $$x_{k+1} = x_k - \eta_k (\rho_{v,k+1} |v_{k+1}| + \varepsilon)^{-\frac{1}{2}} \odot (\rho_{m,k+1} m_{k+1} + \alpha g_k);$$
    $k = k + 1$;
end while
Return $x_k$.
```

Assumption 4 1. *The sequence $\{x_k\}$ is almost surely bounded. That is,*

$$\sup_{k \geq 0} \|x_k\| < +\infty$$

holds almost surely.

2. *The sequence of stepsizes $\{\eta_k\}$ is positive and satisfies*

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

3. *The scaling parameters $\{\rho_{m,k}\}$ and $\{\rho_{v,k}\}$ satisfy*

$$\lim_{k \to +\infty} \rho_{m,k} = 1, \quad \lim_{k \to +\infty} \rho_{v,k} = 1.$$

4. *There exists a constant $M_\Theta > 0$ such that $p_\Theta(s) \leq M_\Theta$ holds for almost every $s \in \Theta$. Here $p_\Theta(s)$ is the auxiliary function defined in Assumption 3(3).*

Remark 3 *For the set-valued mapping $\mathcal{S}$ in Table 3, it is worth mentioning that under Assumption 4(4), for any $x \in \mathbb{R}^n$ and almost every $s \in \Theta$, we have $\sup_{d \in \mathcal{S}(x,s)} \|d\| \leq (p_\Theta(s))^2$, and $p_\Theta(x)^2$ is integrable in Θ. Therefore, based on (Shapiro and Xu, 2007, Theorem 2), we can conclude that $\mathbb{E}_{s \sim P}[\mathcal{S}(x,s)]$ has closed graph and is compact valued. Hence the set-valued mapping $\mathcal{U}$ corresponding to Adam has closed graph and is compact valued. Similarly, based on (Shapiro and Xu, 2007, Theorem 2) and Assumption 4(4), it is easy to verify that all the set-valued mappings $\mathcal{U}$ listed in Table 3 has closed graph and is compact valued, and $\mathbb{E}_{s \sim P}[R_{\mathcal{U}}(\mathcal{D}(x,s), m, v)] \in v - \mathcal{U}(x,m,v)$ holds for any $(x,m,v) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n_+$.*

Adam-family methods	Expression of $R_{\mathcal{U}}$ for updating $\{v_k\}$	Corresponding $\mathcal{U}$ in (AFM)	α and κ
Adam (Kingma and Ba, 2015)	$R_{\mathcal{U}}(g,m,v) = v - g^2$	$\mathrm{sign}(v) \odot \mathcal{S}(x)$	$\alpha = 0, \kappa = 0$
AdaBelief (Zhuang et al., 2020)	$R_{\mathcal{U}}(g,m,v) = v - (g-m)^2$	$\mathrm{sign}(v) \odot \tilde{\mathcal{S}}(x,m,v)$	$\alpha = 0, \kappa = 0$
AMSGrad (Reddi et al., 2018)	$R_{\mathcal{U}}(g,m,v) = v - \max\{v, g^2\}$	$\mathrm{sign}(v) \odot (\mathbb{E}_{s\sim P}[\max\{\lvert v\rvert, \mathcal{S}(x,s)\}])$	$\alpha = 0, \kappa = 1$
NAdam (Dozat, 2016)	$R_{\mathcal{U}}(g,m,v) = v - g^2$	$\mathrm{sign}(v) \odot \mathcal{S}(x)$	$\alpha > 0, \kappa = 0$
Yogi (Zaheer et al., 2018)	$R_{\mathcal{U}}(g,m,v) = v - \mathrm{sign}(v-g^2) \odot g^2$	$\mathrm{sign}(v) \odot (\mathbb{E}_{s\sim P}[\{\lvert v\rvert - \mathrm{sign}(\lvert v\rvert - d) \odot d : d \in \mathcal{S}(x,s)\}])$	$\alpha = 0, \kappa = 0$

Table 3: Different updating schemes for $\{v_k\}$ in Step 6 of Algorithm 1 that describe Adam, AdaBelief, AMSGrad, NAdam and Yogi. Here $\mathcal{S}(x,s) := \mathrm{conv}\left(\{d \odot d : d \in \mathcal{D}(x,s)\}\right)$, $\mathcal{S}(x) := \mathbb{E}_{s\sim P}[\mathcal{S}(x,s)]$, and $\tilde{\mathcal{S}}(x,m,v) := \mathbb{E}_{s\sim P}[\{(d-m)^2 : d \in \mathcal{D}(x,s)\}]$.

The following lemma illustrates that the sequence $\{(x_k, m_k, v_k)\}$ is uniformly bounded under Assumption 4.

Lemma 7 *For any sequence $\{x_k\}$ generated by Algorithm 1. Suppose Assumption 3 and Assumption 4 hold, and the sequence $\{v_k\}$ follows the schemes in Table 3. Then almost surely, it holds that*

$$\sup_{k\geq 0} \|x_k\| + \|m_k\| + \|v_k\| < +\infty.$$

Proof Based on Assumption 3(3), Assumption 4(1) and Assumption 4(4), we have

$$\sup_{k\geq 0} \|g_k\| \leq \sup_{k\geq 0} \sup_{d\in\mathcal{D}(x_k,s_k)} \|d\| \leq \sup_{k\geq 0} p_\Theta(s_k) \leq M_\Theta. \tag{17}$$

From the updating rule in Algorithm 1, it holds for any $k \geq 0$ that

$$\|m_{k+1}\| \leq (1-\tau_1\eta_k)\|m_k\| + \tau_1\eta_k\|g_k\| \leq \max\left\{\|m_0\|, \sup_{k\geq 0}\|g_k\|\right\}. \tag{18}$$

Therefore, we can conclude that $\sup_{k\geq 0}\|m_k\| < +\infty$.

Next, we prove that the sequence $\{v_k\}$ is uniformly bounded for all the updating schemes in Table 3.

Adam and NAdam: For any $k \geq 0$, it holds that

$$\|v_{k+1}\| \leq (1-\tau_2\eta_k)\|v_k\| + \tau_2\eta_k\left\|g_k^2\right\| \leq \max\left\{\|v_0\|, \sup_{k\geq 0}\left\|g_k^2\right\|\right\}.$$

Therefore, we can conclude that $\sup_{k\geq 0}\|v_k\| < +\infty$.

AdaBelief: For any $k \geq 0$, it holds that

$$\|v_{k+1}\| \leq (1-\tau_2\eta_k)\|v_k\| + \tau_2\eta_k\left\|(g_k - m_{k+1})^2\right\| \leq \max\left\{\|v_0\|, \sup_{k\geq 0}\left\|(g_k - m_{k+1})^2\right\|\right\}.$$

Therefore, we can conclude that $\sup_{k\geq 0}\|v_k\| < +\infty$.

AMSGrad: For any $k \geq 0$, it holds that

$$\sup_{k\geq 0}\|v_{k+1}\| = \left\|v_k + \tau_2\eta_k\max\{0, g_k^2 - v_k\}\right\| \leq \max\{\|v_0\|, \sup_{k\geq 0}\left\|g_k^2\right\|\} < +\infty.$$

Yogi: For any $k \geq 0$, it holds that

$$\|v_{k+1}\| \leq \max\left\{\|v_k\|, (1+\tau_2\eta_k)\left\|g_k^2\right\|\right\}.$$

Therefore, we can conclude that

$$\sup_{k\geq 0}\|v_{k+1}\| \leq \max\left\{\|v_0\|, \sup_{k\geq 0}(1+\tau_2\eta_k)\left\|g_k^2\right\|\right\} < +\infty.$$

Combined with the above inequalities, we can conclude that for any of the updating schemes in Table 3, it holds that

$$\sup_{k\geq 0}\|x_k\| + \|m_k\| + \|v_k\| < +\infty.$$

This completes the proof. ■

Next, we establish the convergence properties for Algorithm 1 by relating it to the framework (AFM). The following corollary demonstrates that Algorithm 1 fits the framework (AFM) when choosing the updating scheme for the estimators $\{v_k\}$ specified in Table 3. Consequently, the convergence properties of Algorithm 1 directly follow from Theorem 1.

Corollary 1 *For any sequence $\{x_k\}$ generated by Algorithm 1. Suppose Assumption 3 and Assumption 4 hold, the sequence $\{v_k\}$ follows the schemes in Table 3, and $(1-\kappa)\tau_2 \leq 4\tau_1$. Then almost surely, every cluster point of $\{x_k\}$ is a $\mathcal{D}_f$-stationary point of f and the sequence $\{f(x_k)\}$ converges.*

Proof We first check the validity of Assumption 2. Lemma 7 implies that Assumption 2(2) holds. Moreover, as discussed in (Bolte et al., 2023, Theorem 3.10), $\mathcal{D}_f$ is a conservative field that admits f as its potential function. Then it follows from Lemma 1 that $\mathcal{U}$ has closed graph and is locally bounded, hence Assumption 2(3) holds with $\mathcal{U}$ chosen as the formulations in Table 3 and $\gamma = \frac{1}{2}$. In addition, Lemma 7 illustrates that $\|m_{k+1} - m_k\| + \|v_{k+1} - v_k\|$ converges to 0 as k goes to infinity. Then combined with Assumption 4(3), we can conclude that Assumption 2(4) holds. Furthermore, Assumption 2(5) directly follows from (17), and Assumption 2(6) directly follows from Assumption 4(2) by choosing $\theta_k = \eta_k$.

On the other hand, as the results in (Bolte et al., 2023, Theorem 3.10) show that under Assumption 3(1)-(3), f is a potential function that admits $\mathcal{D}_f$ as its conservative field. Then the validity of Assumption 1 directly follows from Assumption 3.

Therefore, we can conclude that Algorithm 1 follows the framework in (AFM). Then from Theorem 1, we can conclude that almost surely, every cluster point of $\{x_k\}$ is a $\mathcal{D}_f$-stationary point of f and the sequence $\{f(x_k)\}$ converges. This completes the proof. ■

Since κ can be set to 1 in AMSGrad, the following corollary illustrates that AMSGrad converges with any combination of the parameters $\tau_1, \tau_2 > 0$. This improves the results in Reddi et al. (2018), where f is assumed to be differentiable, while the stepsizes are chosen as $\eta_k = \mathcal{O}(k^{-\frac{1}{2}})$.

Corollary 2 *For any sequence $\{x_k\}$ generated by Algorithm 1 with AMSGrad updating scheme in Table 3. Suppose Assumption 3 and Assumption 4 hold. Then almost surely, every cluster point of $\{x_k\}$ is a $\mathcal{D}_f$-stationary point of f and the sequence $\{f(x_k)\}$ converges.*

Finally, the following corollary demonstrates that under mild assumptions, with almost every initial point and stepsize in Algorithm 1, the generated sequence $\{x_k\}$ can find the stationary points in the sense of Clarke subdifferential almost surely. Therefore, although AD algorithms may introduce spurious stationary points for f, Algorithm 1 can avoid such spurious stationary points for almost every choice of initial points and stepsizes.

Corollary 3 *Suppose Assumption 3 holds. Moreover, for any sequence $\{x_k\}$ generated by Algorithm 1 with the update schemes in Table 3, we assume that*

1. *There exists a prefixed positive sequence $\{\nu_k\}$ and parameters $0 < c_{\min} < c_{\max}$, such that the stepsizes $\{\eta_k\}$ in Algorithm 1 are set as $\eta_k = c\nu_k$ for any $k \geq 0$ with some $c \in (c_{\min}, c_{\max})$.*
2. *There exists a non-empty open subset $K \subset \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n_+$ such that Assumption 4 holds with any $(x_0, m_0, v_0, c) \in K \times (c_{\min}, c_{\max})$.*

Then for almost every $(x_0, m_0, v_0, c) \in K \times (c_{\min}, c_{\max})$, it holds almost surely that every cluster point of $\{x_k\}$ is a ∂f-stationary point of f and the sequence $\{f(x_k)\}$ converges.

Proof Let $A = \{x \in \mathbb{R}^n : \mathcal{D}_f(x) \neq \partial f(x)\}$. Then it holds from Bolte and Pauwels (2021) that $\mathcal{A}$ is a zero-measure subset of $\mathbb{R}^n$. Therefore, the set $\{(x, m, v) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n_+ : x \in A\}$ is zero-measure in $\mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n_+$.

Let

$$Q_k^{(1)}(x, m, v, s_k) = \begin{bmatrix} 0 \\ \tau_1 m - \tau_1 \mathcal{D}(x, s_k) \\ 0 \end{bmatrix}, \quad Q_k^{(2)}(x, m, v, s_k) = \begin{bmatrix} 0 \\ 0 \\ \tau_2 R_{\mathcal{U}}(\mathcal{D}(x, s_k), m, v) \end{bmatrix},$$

and

$$Q_k^{(3)}(x, m, v, s_k) = \begin{bmatrix} (\rho_{v,k+1}|v| + \varepsilon)^{-\frac{1}{2}} \odot (\rho_{m,k+1} m + \alpha \mathcal{D}(x, s_k)) \\ 0 \\ 0 \end{bmatrix}.$$

For almost every $s \in \Theta$, f_s is a definable function that admits $x \mapsto \mathcal{D}(x, s)$ as its conservative field, Bolte and Pauwels (2021, Theorem 4) illustrates that the set-valued mapping $x \mapsto \mathcal{D}(x, s)$ is almost everywhere $\mathcal{C}^1$. Moreover, from the update schemes from Table 3, we can conclude that all the listed $R_{\mathcal{U}}$ is semi-algebraic. Therefore, the set-valued mappings $\mathcal{Q}_k^{(1)}$, $\mathcal{Q}_k^{(2)}$ and $\mathcal{Q}_k^{(3)}$ are almost everywhere $\mathcal{C}^1$ for almost every $s \in \Theta$ and any $k \geq 0$. Moreover, Algorithm 1 can be expressed as

$$\begin{aligned}
\left(x_{k+\frac{1}{3}}, m_{k+\frac{1}{3}}, v_{k+\frac{1}{3}}\right) &\in (x_k, m_k, v_k) - \eta_k Q_k^{(1)}(x_k, m_k, v_k, s_k), \\
\left(x_{k+\frac{2}{3}}, m_{k+\frac{2}{3}}, v_{k+\frac{2}{3}}\right) &\in (x_{k+\frac{1}{3}}, m_{k+\frac{1}{3}}, v_{k+\frac{1}{3}}) - \eta_k Q_k^{(2)}\left(x_{k+\frac{1}{3}}, m_{k+\frac{1}{3}}, v_{k+\frac{1}{3}}, s_k\right), \\
(x_{k+1}, m_{k+1}, v_{k+1}) &\in \left(x_{k+\frac{2}{3}}, m_{k+\frac{2}{3}}, v_{k+\frac{2}{3}}\right) - \eta_k Q_k^{(3)}\left(x_{k+\frac{2}{3}}, m_{k+\frac{2}{3}}, v_{k+\frac{2}{3}}, s_k\right).
\end{aligned}$$

From Theorem 2 we can conclude that there exists a full-measure subset S_{init} of $K \times [c_{\min}, c_{\max}]$ such that for any $(x_0, m_0, v_0, c) \in S_{init}$, almost surely, it holds that $\{(x_k, m_k, v_k) : k = 0, \frac{1}{3}, \frac{2}{3}, 1, ...\} \subset A^c$, implying that $\mathbb{E}_{s_k \sim P}[g_k] = \mathbb{E}_{s_k \sim P}[\chi(x_k, s_k)] \in \partial f(x_k)$ holds for any $k \geq 0$. Therefore, by fixing $\mathcal{D}_f$ as ∂f in Theorem 1, we can conclude that every cluster point of $\{x_k\}$ is a ∂f-stationary point of f and the sequence $\{f(x_k)\}$ converges almost surely. This completes the proof. ■

5. Applications: Gradient Clipping for Stochastic Subgradient Methods

In this section, we present stochastic subgradient methods with gradient clipping technique to solve UNP, under the assumption that the evaluation noises for subgradients are heavy-tailed. Then we prove the convergence properties for our proposed methods based on the framework (AFM).

For a clearer and more comprehensive presentation of our proposed methods, we follow the settings and notations in Section 4 throughout this section. In particular, we assume the objective function f in UNP can be expressed as

$$f(x) = \mathbb{E}_{s \sim P}[f_s(x)],$$

where $(\Theta, \mathcal{F}, P)$ is the probability space.

For any given compact and convex subset $\mathcal{S} \subset \mathbb{R}^n$ such that 0 lies in its interior, we define the clipping mapping $\mathrm{Clip}_{(\cdot)}(\cdot) : \mathbb{R}_+ \times \mathbb{R}^n \to \mathbb{R}^n$ as,

$$\mathrm{Clip}_C(g) := \arg\min_{x \in C\mathcal{S}} \|x - g\|, \tag{19}$$

where $C\mathcal{S} = \{Cs : s \in \mathcal{S}\}$. Intuitively, the clipping mapping avoids the extreme values in evaluating the gradients by restricting them in a compact region, and hence helps enforce the convergence of stochastic subgradient methods with heavy-tailed evaluation noises.

The explicit expression of the clipping mapping depends on the choice of the convex compact set $\mathcal{S}$. When $\mathcal{S}$ is chosen as the n-dimensional hypercube $[-1, 1]^n$, the corresponding clipping mapping Clip_C can be expressed as $\mathrm{Clip}_C(g) = \min\{\max\{g, -C\}, C\}$, which is a coordinate-wise mapping and can be computed easily. Furthermore, when $\mathcal{S}$ is chosen as the unit ball in $\mathbb{R}^n$, the corresponding clipping mapping becomes $\mathrm{Clip}_C(g) = g \cdot \min\left\{1, \frac{C}{\|g\|}\right\}$, which coincides with the clipping mapping employed in (Zhang et al., 2020a). Note that it is not a coordinate-wise mapping. It is worth mentioning that the convergence properties of our analyzed stochastic subgradient methods are independent of the specific choice of $\mathcal{S}$. Hence we do not specify the choice of $\mathcal{S}$ in the clipping mapping $\mathrm{Clip}_C(\cdot)$ in the rest of this section.

5.1 SGD with Gradient Clipping

In this subsection, we consider a general framework of SGD that incorporates the gradient clipping technique to deal with heavy-tailed evaluating noises in its stochastic subgradients:

$$\boxed{\begin{aligned}&\text{Sample } s_k \sim P \text{ and choose } g_k = \chi(x_k, s_k),\\ &\hat{g}_k = \mathrm{Clip}_{C_k}(g_k),\\ &m_{k+1} = (1-\tau_1\eta_k)m_k + \tau_1\eta_k\hat{g}_k,\\ &x_{k+1} = x_k - \eta_k(m_{k+1} + \alpha\hat{g}_k).\end{aligned}} \tag{SGD-C}$$

Here χ is a selection of the set-valued mapping $\mathcal{D}$, as defined in Assumption 3(2). Moreover, τ_1 and α refer to the parameters for heavy-ball momentum and Nesterov momentum, respectively. Therefore, compared to existing works that concentrate on the convergence of standard SGD without any momentum term, the updating scheme in SGD-C encompasses various popular variants of SGD, including SGD (Bolte and Pauwels, 2021; Bianchi et al., 2022), and its momentum accelerated variants (Nesterov, 2003; Loizou and Richtárik, 2017; Castera et al., 2021).

To establish the convergence properties for SGD-C method, we make the following assumptions.

Assumption 5 1. *The parameters satisfy $\alpha \geq 0$, $\tau_1 > 0$.*

2. *The sequence $\{x_k\}$ is almost surely bounded. That is,*

$$\sup_{k\geq 0} \|x_k\| < +\infty$$

holds almost surely.

3. *The stepsizes $\{\eta_k\}$ and clipping parameters $\{C_k\}$ are positive and satisfy*

$$\sum_{k=0}^{+\infty} \eta_k = +\infty, \quad \lim_{k\to+\infty} \eta_k \log(k) = 0, \quad \lim_{k\to+\infty} C_k = +\infty, \quad \textit{and} \quad \lim_{k\to+\infty} C_k^2\eta_k \log(k) = 0.$$

Different from the existing works, in this section, we only assume the evaluation noises to be integrable in Assumption 3(3), without any further assumptions of the uniform boundedness such as Assumption 4(4). As far as we know, such an assumption is among the weakest ones in the relevant literature (Zhang et al., 2020a; Gorbunov et al., 2020; Zhang et al., 2020b; Mai and Johansson, 2021; Qian et al., 2021; Elesedy and Hutter, 2023; Reisizadeh et al., 2023). Moreover, Assumption 5(3) is mild, as we can always choose $C_k = C_0 \left(\eta_k \log(k)\right)^{-\frac{1}{3}}$ in SGD-C.

Let the σ-algebras $\{\mathcal{F}_k\}$ be chosen as $\mathcal{F}_k = \sigma(x_j, m_j, g_j, s_j : j < k)$, $d_k = \mathbb{E}_{s_k\sim P}[\hat{g}_k]$, and $\xi_k = \frac{\hat{g}_k - d_k}{C_k}$. Then $\hat{g}_k = d_k + C_k\xi_k$. Hence the update scheme in SGD-C can be rewritten as

$$\begin{aligned} m_{k+1} &= (1-\tau_1\eta_k)m_k + \tau_1\eta_k(d_k + C_k\xi_k),\\ x_{k+1} &= x_k - \eta_k(m_{k+1} + \alpha(d_k + C_k\xi_k)). \end{aligned}$$

Here $(-(m_{k+1} + \alpha d_k), \tau_1(-m_k + d_k))$ is regarded as the noiseless update direction for (x_k, m_k), while $(\alpha C_k\xi_k, \tau_1 C_k\xi_k)$ can be interpreted as the corresponding evaluation noises. We first present Lemma 8 to exhibit some basic properties of the sequence $\{d_k\}$ and $\{\xi_k\}$.

Lemma 8 *Suppose Assumption 3 and Assumption 5(1) hold, then there exists a sequence of nonnegative constants $\{\delta_k\}$ such that $\lim_{k\to+\infty}\delta_k = 0$,*

$$d_k \in \mathcal{D}_f^{\delta_k}(x_k), \quad \forall k \geq 0,$$

and $\{\xi_k\}$ is a uniformly bounded martingale difference sequence.

Proof Let $\varepsilon_{\mathcal{S}} = \min_{x\notin\mathcal{S}} \|x\|$, and $M_{\mathcal{S}} = \max_{x\in\mathcal{S}} \|x\|$. Then from the definition of $\hat{g}_k$, it holds for any $C > 0$ and $k \geq 0$ that

$$\begin{aligned}
&\mathbb{E}_{s_k\sim P}[\|g_k - \mathrm{Clip}_C(g_k)\|] = \mathbb{E}_{s_k\sim P}\left[\|g_k - \mathrm{Clip}_C(g_k)\| \cdot \mathbb{1}_{\{g_k\notin C\mathcal{S}\}}\right]\\
\leq{}& \mathbb{E}_{s_k\sim P}\left[\|g_k\| \cdot \mathbb{1}_{\{\|g_k\|\geq C\varepsilon_{\mathcal{S}}\}}\right] \leq \mathbb{E}_{s_k\sim P}\left[\|g_k\| \cdot \mathbb{1}_{\{p_\Theta(s_k)\geq C\varepsilon_{\mathcal{S}}\}}\right]\\
\leq{}& \mathbb{E}_{s_k\sim P}\left[p_\Theta(s_k) \cdot \mathbb{1}_{\{p_\Theta(s_k)\geq C\varepsilon_{\mathcal{S}}\}}\right].
\end{aligned}$$

As a result, from the fact that p_Θ is integrable over Θ, it holds that

$$\lim_{C\to+\infty} \sup_{k\geq 0} \mathbb{E}_{s_k\sim P}\left[\|g_k - \mathrm{Clip}_C(g_k)\|\right] = 0.$$

Therefore, let $\delta_k = \mathbb{E}_{s_k\sim P}\left[\left\|g_k - \mathrm{Clip}_{C_k}(g_k)\right\|\right]$, then it is easy to verify that $\lim_{k\to+\infty}\delta_k = 0$. Moreover, from the definition of δ_k, we have

$$\begin{aligned}
&\mathrm{dist}\left(d_k, \mathcal{D}_f(x_k)\right) \leq \|\mathbb{E}_{s_k\sim P}[\hat{g}_k] - \mathbb{E}_{s_k\sim P}[g_k]\| = \left\|\mathbb{E}_{s_k\sim P}\left[g_k - \mathrm{Clip}_{C_k}(g_k)\right]\right\|\\
\leq{}& \mathbb{E}_{s_k\sim P}\left[\left\|g_k - \mathrm{Clip}_{C_k}(g_k)\right\|\right] = \delta_k.
\end{aligned}$$

Furthermore, from the definition of $\{\xi_k\}$, it holds for any $k \geq 0$ that $\|\hat{g}_k\| \leq C_k M_{\mathcal{S}}$ almost surely. Then we can conclude that almost surely, $\sup_{k\geq 0} \|\xi_k\| \leq M_{\mathcal{S}}$. Moreover, as s_k is chosen independently from $\{s_0, ..., s_{k-1}\}$, it holds that $\mathbb{E}[\xi_k|\mathcal{F}_{k-1}] = \mathbb{E}_{s_k\sim P}[\xi_k] = 0$. Therefore, we can conclude that $\{\xi_k\}$ is a uniformly bounded martingale difference sequence. This completes the proof. ■

Next, we show that the sequence $\{m_k\}$ is almost surely uniformly bounded in the following proposition, even if the corresponding evaluation noises are not uniformly bounded. The proof for Proposition 6 is presented in Appendix A.

Proposition 6 *Suppose Assumption 3 and Assumption 5 hold, then we have*

$$\sup_{k\geq 0} \|m_k\| < +\infty.$$

In the following theorem, we demonstrate that the framework SGD-C conforms to the framework (AFM), and directly establish its convergence properties based on Theorem 1 and Theorem 2.

Theorem 3 *Let $\{x_k\}$ be the sequence generated by SGD-C. Suppose Assumption 3 and Assumption 5 hold. Then, almost surely, any cluster point of $\{x_k\}$ is a $\mathcal{D}_f$-stationary point of f, and the sequence $\{f(x_k)\}$ converges.*

Proof Let $d_k = \mathbb{E}_{s_k \sim P}[\hat{g}_k]$, then from Lemma 8, it holds that $\lim_{k\to+\infty} \mathrm{dist}\left(d_k, \mathcal{D}_f(x_k)\right) = 0$, and $\{\xi_k\}$ is a uniformly bounded martingale difference sequence.

Then we set $\theta_k = C_k\eta_k$ in the framework (AFM). With $\gamma = 0$, $\varepsilon = 1$, and $\mathcal{U}(x, m, v) = \{0\}$ in (1), SGD-C can be reformulated as the following scheme,

$$(x_{k+1}, m_{k+1}, v_{k+1}) = (x_k, m_k, v_k) - \eta_k(d_{x,k}, d_{m,k}, d_{v,k}) - \theta_k(\alpha\xi_k, -\tau_1\xi_k, 0),$$

where $(d_{x,k}, d_{m,k}, d_{v,k}) = (m_k + \alpha d_k, \tau_1 m_k - \tau_1 d_k, \tau_2 v)$.

Next, we check the validity of Assumption 2. Assumption 5(1)-(2) and Proposition 6 imply Assumption 2(1)-(2), while Assumption 2(3) holds as we set $\mathcal{U}(x, m, v) = \{0\}$ in (AFM). Moreover, from the uniform boundedness of $\{(x_k, m_k)\}$ and Lemma 8, it is easy to verify that there exists a diminishing sequence $\{\delta_k\}$ such that $(d_{x,k}, d_{m,k}, d_{v,k}) \in \mathcal{G}^\delta(x_k, m_k, v_k)$ with $\gamma = 0$, $\varepsilon = 1$, and $\mathcal{U}(x, m, v) = \{0\}$ in (1). Then the validity of Assumption 2(4) is guaranteed. In addition, Assumption 2(5) follows from the fact that $\{\xi_k\}$ is a martingale difference sequence, as illustrated in Lemma 8. Furthermore, Assumption 2(6) directly follows Assumption 5(4) with $\theta = C_k\eta_k$. Therefore, from Theorem 1, we can conclude that any cluster point of $\{x_k\}$ is a $\mathcal{D}_f$-stationary point of f, and the sequence $\{f(x_k)\}$ converges. This completes the proof. ■

The following theorem illustrates that under mild assumptions with almost every initial points and stepsizes, any sequence generated by SGD-C converges to ∂f-stationary points of f, hence avoids the spurious stationary points introduced by conservative field $\mathcal{D}_f$.

Theorem 4 *Suppose Assumption 3 holds. Moreover, for the sequence $\{x_k\}$ generated by SGD-C, we assume that*

1. *There exists a prefixed positive sequence $\{\nu_k\}$ and the parameters $0 < c_{\min} < c_{\max}$ such that the stepsizes $\{\eta_k\}$ are chosen as $\eta_k = c\nu_k$ for any $k \geq 0$ with some $c \in (c_{\min}, c_{\max})$.*

2. *There exists a non-empty open subset K of $\mathbb{R}^n \times \mathbb{R}^n$ such that Assumption 5 holds with any $(x_0, m_0, c) \in K \times (c_{\min}, c_{\max})$.*

Then for almost every $(x_0, m_0, c) \in K \times (c_{\min}, c_{\max})$, it holds almost surely that every cluster point of $\{x_k\}$ is a ∂f-stationary point of f and the sequence $\{f(x_k)\}$ converges.

Proof For almost every $s \in \Theta$, since f_s is definable, we can conclude that the set-valued mapping $x \mapsto \partial f_s(x)$ is almost everywhere $\mathcal{C}^1$. Then we consider the following set-valued mappings

$$\mathcal{Q}_k^{(1)}(x, m, s_k) = \begin{bmatrix} 0 \\ -\tau_1 m + \tau_1 \mathrm{Clip}_{C_k}\left(\mathcal{D}(x, s_k)\right) \end{bmatrix}$$

and

$$\mathcal{Q}_k^{(2)}(x, m, s_k) = \begin{bmatrix} -m - \alpha \mathrm{Clip}_{C_k}\left(\mathcal{D}(x, s_k)\right) \\ 0 \end{bmatrix}.$$

From the definability of f_s and $\mathcal{D}(\cdot, s)$, $\mathcal{D}(\cdot, s)$ is almost everywhere $\mathcal{C}^1$ over $\mathbb{R}^n$ (Bolte and Pauwels, 2021, Theorem 4). As a result, from the continuity of clipping mapping Clip_C, we can conclude that for any $k \geq 0$ and almost every $s \in \Theta$, both $\mathcal{Q}^{(1)}(\cdot,\cdot,s)$ and $\mathcal{Q}^{(2)}(\cdot,\cdot,s)$ are almost everywhere $\mathcal{C}^1$ over $\mathbb{R}^n \times \mathbb{R}^n$.

From the expression of $\mathcal{Q}^{(1)}$ and $\mathcal{Q}^{(2)}$, the update scheme in SGD-C can be reshaped as

$$\begin{aligned}
&(x_{k+\frac{1}{2}}, m_{k+\frac{1}{2}}) \in (x_k, m_k) - \eta_k \mathcal{Q}_k^{(1)}(x_k, m_k, s_k),\\
&(x_{k+1}, m_{k+1}) \in (x_{k+\frac{1}{2}}, m_{k+\frac{1}{2}}) - \eta_k \mathcal{Q}_k^{(2)}(x_{k+\frac{1}{2}}, m_{k+\frac{1}{2}}, s_k).
\end{aligned}$$

Notice that the set $A := \{(x,m) \in \mathbb{R}^n \times \mathbb{R}^n : \mathcal{D}_f(x) \neq \partial f(x)\}$ is zero-measure in $\mathbb{R}^n \times \mathbb{R}^n$ as illustrated in Bolte and Pauwels (2021). Therefore, Theorem 2 illustrates that for almost every $(x_0, m_0, c) \in K \times (c_{\min}, c_{\max})$, it holds almost surely that the sequence $\{(x_k, m_k)\}$ generated by SGD-C satisfies $\{(x_k, m_k)\} \subset A^c$.

From the definition of A, we can conclude that $\mathcal{D}_f(x) = \partial f(x)$ holds for any $x \in A^c$. Therefore, for any sequence $\{(x_k, m_k)\}$ with those initial points $(x_0, m_0) \in K$ and scaling parameter $c \in (c_{\min}, c_{\max})$, the corresponding conservative field $\mathcal{D}_f$ can be directly chosen as ∂f since $\mathcal{D}_f(x_k) = \partial f(x_k)$ holds for any $k \geq 0$. Therefore, Theorem 1 illustrates that with those initial points $(x_0, m_0) \in K$ and scaling parameter $c \in (c_{\min}, c_{\max})$, any cluster point of $\{x_k\}$ is a ∂f-stationary point and the sequence $\{f(x_k)\}$ converges. This completes the proof. ■

5.2 Adam-family Method with Gradient Clipping

In this subsection, we consider developing an Adam-family method (ADAM-C) that employs the gradient clipping technique for solving UNP under heavy-tailed evaluation noises. Then we establish its convergence properties based on our proposed framework (AFM). The detailed method is summarized by the following update scheme.

$$\boxed{\begin{aligned}
&\text{Sample } s_k \sim P \text{ and choose } g_k = \chi(x_k, s_k),\\
&\hat{g}_k = \mathrm{Clip}_{C_k}(g_k),\\
&m_{k+1} = (1 - \tau_1 \eta_k) m_k + \tau_1 \eta_k \hat{g}_k,\\
&v_{k+1} = (1 - \tau_2 \eta_k) v_k + \tau_2 \eta_k |\hat{g}_k|,\\
&x_{k+1} = x_k - \eta_k (\rho_{v,k+1} |v_{k+1}| + \varepsilon)^{-1} \odot (\rho_{m,k+1} m_{k+1} + \alpha g_k).
\end{aligned}} \tag{ADAM-C}$$

Here the χ is a selection of the set-valued mapping $\mathcal{D}$, as defined in Assumption 3(2). Moreover, the estimator v_k is updated for tracking the first-order moment of $|g_k|$.

It is worth mentioning that the estimators $\{v_k\}$ in ADAM-C adopt a different update scheme as those in the original Adam (Kingma and Ba, 2015), since the evaluation noises are assumed to be heavy-tailed. In the original Adam, the estimators v_k estimate the noise level of each coordinate by tracking the second-order moment of the stochastic subgradients g_k. However, when the evaluation noise of g_k is assumed to be heavy-tailed, $\mathbb{E}[\mathrm{Clip}_C(g_k)^2]$ may diverge to infinity as $C \to +\infty$. As a result, the sequence $\{v_k\}$ may not be uniformly bounded in Adam under heavy-tailed noises, leading to the absence of convergence guarantees.

To estimate the noise level of each coordinate under heavy-tailed evaluation noises, we consider tracking the first-order moment of $\{|g_k|\}$ by $\{v_k\}$, and employ the $(|v_{k+1}| + \varepsilon)^{-1}$ as the coordinate-wise adaptive stepsizes. Under Assumption 3, $\mathbb{E}_{s_k \sim P}[|g_k|]$ exists and takes finite values almost surely. As a result, the estimators $\{v_k\}$ in ADAM-C can be proved to be uniformly bounded, which is crucial in establishing the convergence properties for ADAM-C based on the framework (AFM).

To establish the convergence properties for ADAM-C, we make the following assumptions.

Assumption 6 *1. The parameters satisfy $\alpha \geq 0$, $\tau_1, \tau_2, \varepsilon > 0$ and $\tau_2 \leq 2\tau_1$.*

2. The sequence $\{x_k\}$ is almost surely bounded, i.e.,

$$\sup_{k\geq 0} \|x_k\| < +\infty$$

holds almost surely.

3. The stepsizes $\{\eta_k\}$ and clipping parameters $\{C_k\}$ are positive and satisfy

$$\sum_{k=0}^{+\infty} \eta_k = +\infty, \quad \lim_{k\to+\infty} \eta_k \log(k) = 0, \quad \lim_{k\to+\infty} C_k = +\infty, \quad \textit{and} \quad \lim_{k\to+\infty} C_k^2 \eta_k \log(k) = 0.$$

4. The scaling parameters $\{\rho_{m,k}\}$ and $\{\rho_{v,k}\}$ satisfy

$$\lim_{k\to+\infty} \rho_{m,k} = 1, \quad \lim_{k\to+\infty} \rho_{v,k} = 1.$$

We first present Proposition 7 to illustrate that the uniform boundedness of $\{x_k\}$ implies the uniformly boundedness of $\{m_k\}$ and $\{v_k\}$. The proof of Proposition 7 follows the same techniques as in Proposition 6, hence we omit its proof for simplicity.

Proposition 7 *Suppose Assumption 3 and Assumption 6 hold. Then we have $\sup_{k\geq 0} \|m_k\| + \|v_k\| < +\infty$.*

Next, we present the following theorem that illustrates the convergence properties of ADAM-C.

Theorem 5 *Let $\{x_k\}$ be the sequence generated by ADAM-C. Suppose Assumption 3 and Assumption 6 hold. Then, almost surely, any cluster point of $\{x_k\}$ is a $\mathcal{D}_f$-stationary point of f, and the sequence $\{f(x_k)\}$ converges.*

Proof Let $\mathcal{W}(x) = \mathrm{conv}\,(\mathbb{E}_s[|\mathcal{D}(x,s)|])$. Then following Assumption 3(3), we can conclude that for any $x \in \mathbb{R}^n$ and almost every $s \in \Theta$, it holds that $\sup_{d\in|\mathcal{D}(x,s)|} \|d\| \leq p_\Theta(s)$. Then together with (Shapiro and Xu, 2007, Theorem 2) and the fact that the set-valued mapping $x \mapsto |\mathcal{D}(x,s)|$ is locally bounded and graph closed for almost every $s \in \Theta$, we can conclude that $\mathcal{W}(x)$ is convex compact valued, graph closed and locally bounded.

Then with $\mathcal{U}(x,m,v) = \mathcal{W}(x)$ in (1), we can show that ADAM-C fits in the framework (AFM). Moreover, similar to the proof in Theorem 3, we can verify the validity of Assumption 2. Then from Theorem 1 we can conclude that any cluster point of $\{x_k\}$ is a

$\mathcal{D}_f$-stationary point of f and the sequence $\{f(x_k)\}$ converges. This completes the proof. ■

Similar to Theorem 4, we can show that under mild assumptions with almost every initial point and stepsize, any sequence generated by ADAM-C is capable of finding ∂f-stationary points of f, regardless of the chosen conservative field in ADAM-C.

Theorem 6 *Suppose Assumption 3 holds. Moreover, for the sequence $\{(x_k, m_k, v_k)\}$ generated by ADAM-C, we assume that*

1. *There exists a prefixed sequence $\{\nu_k\}$ and the parameters $0 < c_{\min} < c_{\max}$ such that the stepsizes $\{\eta_k\}$ in ADAM-C are chosen as $\eta_k = c\nu_k$ for any $k \geq 0$ with some $c \in (c_{\min}, c_{\max})$.*
2. *There exists a non-empty open subset K of $\mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n_+$ such that Assumption 6 holds with any $(x_0, m_0, v_0, c) \in K \times (c_{\min}, c_{\max})$.*

Then for almost every $(x_0, m_0, v_0, c) \in K \times (c_{\min}, c_{\max})$, it holds almost surely that every cluster point of $\{x_k\}$ is a ∂f-stationary point of f and the sequence $\{f(x_k)\}$ converges .

Proof Let $A := \{(x, m, v) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n : \mathcal{D}_f(x) \neq \partial f(x)\}$, and for any $k \geq 0$, we define the set-valued mappings

$$\mathcal{Q}_k^{(1)}(x, m, v, s) := \begin{bmatrix} 0 \\ \tau_1 m - \tau_1 \mathrm{Clip}_{C_k}(\mathcal{D}(x, s)) \\ 0 \end{bmatrix}, \quad \mathcal{Q}_k^{(2)}(x, m, v, s) := \begin{bmatrix} 0 \\ 0 \\ \tau_2 v - \tau_2 |\mathrm{Clip}_{C_k}(\mathcal{D}(x, s))| \end{bmatrix},$$

and

$$\mathcal{Q}_k^{(3)}(x, m, v, s) := \begin{bmatrix} (\rho_{v,k+1}|v| + \varepsilon)^{-1} \odot (\rho_{m,k+1} m + \alpha \mathrm{Clip}(\mathcal{D}(x, s))) \\ 0 \\ 0 \end{bmatrix}.$$

Then for any $k \geq 0$ and almost every $s \in \Theta$, $\mathcal{Q}_k^{(1)}(\cdot, \cdot, \cdot, s)$, $\mathcal{Q}_k^{(2)}(\cdot, \cdot, \cdot, s)$, and $\mathcal{Q}_k^{(3)}(\cdot, \cdot, \cdot, s)$ are almost everywhere $\mathcal{C}^1$ in $\mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n$. More importantly, the update scheme in (ADAM-C) can be reshaped as

$$\begin{aligned}
\left(x_{k+\frac{1}{3}}, m_{k+\frac{1}{3}}, v_{k+\frac{1}{3}}\right) &\in (x_k, m_k, v_k) - \eta_k Q_k^{(1)}(x_k, m_k, v_k, s_k), \\
\left(x_{k+\frac{2}{3}}, m_{k+\frac{2}{3}}, v_{k+\frac{2}{3}}\right) &\in (x_{k+\frac{1}{3}}, m_{k+\frac{1}{3}}, v_{k+\frac{1}{3}}) - \eta_k Q_k^{(2)}\left(x_{k+\frac{1}{3}}, m_{k+\frac{1}{3}}, v_{k+\frac{1}{3}}, s_k\right), \\
(x_{k+1}, m_{k+1}, v_{k+1}) &\in \left(x_{k+\frac{2}{3}}, m_{k+\frac{2}{3}}, v_{k+\frac{2}{3}}\right) - \eta_k Q_k^{(3)}\left(x_{k+\frac{2}{3}}, m_{k+\frac{2}{3}}, v_{k+\frac{2}{3}}, s_k\right).
\end{aligned}$$

Notice that the set A is zero-measure in $\mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n$, following the same techniques as Theorem 4, we can prove that for almost every $(x_0, m_0, v_0, c) \in K \times (c_{\min}, c_{\max})$, it holds that $\{(x_k, m_k, v_k) : k = 0, \frac{1}{3}, \frac{2}{3}, 1, ...\} \subset A^c$. As a result, for almost every $(x_0, m_0, v_0, c) \in K \times (c_{\min}, c_{\max})$, we can choose the conservative field $\mathcal{D}_f$ in ADAM-C as ∂f in Theorem 5. Therefore, Theorem 1 illustrates that every cluster point of $\{x_k\}$ is a ∂f-stationary point of f and the sequence $\{f(x_k)\}$ converges almost surely. This completes the proof. ■

Remark 4 *Following the updating schemes in Table 3, we can also choose the updating scheme for the estimators $\{v_k\}$ in ADAM-C as one of the followings.*

- **AdaBelief-C:** $v_{k+1} = (1 - \tau_2\eta_k)v_k + \tau_2\eta_k|\hat{g}_k - m_{k+1}|$;
- **AMSGrad-C:** $v_{k+1} = v_k + \tau_2\eta_k \max\{0, |\hat{g}_k| - v_k\}$;
- **Yogi-C:** $v_{k+1} = v_k - \tau_2\eta_k \mathrm{sign}(v_k - |\hat{g}_k|) \odot |\hat{g}_k|$.

Then for these stochastic adaptive subgradient methods with gradient clipping, we can establish the same convergence properties by following the same proof routines as those in Theorem 5 and Corollary 6, hence we omit these proofs for simplicity.

6. Numerical Experiments

In this section, we evaluate the numerical performance of our analyzed Adam-family methods for training nonsmooth neural networks. All the numerical experiments in this section are conducted on a server equipped with an Intel Xeon 6342 CPU and a NVIDIA GeForce RTX 3090 GPU, running Python 3.8 and PyTorch 1.9.0.

6.1 Comparison with Implementations in PyTorch

In this subsection, we evaluate the numerical performance of Algorithm 1 by comparing it with the Adam-family methods available in PyTorch and torch-optimizer packages. In view of the great popularity of Adam-family methods in training nonsmooth neural networks, we aim to investigate whether we can preserve their high efficiency while providing convergence guarantees with minimal modifications to their implementations.

It is important to note that the Adam-family methods in PyTorch can be viewed as Algorithm 1 with a fixed $\eta_k = \eta_0$ in updating the momentum terms $\{m_k\}$ and estimators $\{m_k\}$ (i.e., Steps 5-6 in Algorithm 1). Moreover, $\beta_1 := 1 - \tau_1\eta_0$ and $\beta_2 := 1 - \tau_2\eta_0$ are commonly referred to as the momentum parameters for these Adam-family methods in PyTorch. To our best knowledge, these Adam-family methods with constant stepsizes in updating the momentum terms $\{m_k\}$ and estimators $\{m_k\}$ do not have any convergence guarantees in training nonsmooth neural networks. More importantly, some existing works (Reddi et al., 2018; Zhang et al., 2022) illustrate that Adam may diverge when $\beta_1 < \sqrt{\beta_2}$ and f is assumed to be differentiable.

In our numerical experiments, we investigate the performance of these compared Adam-family methods on training ResNet-50 (He et al., 2016) for image classification tasks on the CIFAR-10 and CIFAR-100 data sets (Krizhevsky et al., Toronto, ON, Canada, 2009). We set the batch size to 128 for all test instances and select the regularization parameter ε as $\varepsilon = 10^{-15}$. Furthermore, at the k-th epoch, we choose the stepsize as $\eta_k = \frac{\eta_0}{\sqrt{k+1}}$ for all the tested algorithms. Following the settings in Castera et al. (2021), we use a grid search to find a suitable initial stepsize η_0 and parameters τ_1, τ_2 for the Adam-family methods provided in PyTorch. We select the initial stepsize η_0 from $\{k_1 \times 10^{-k_2} : k_1 = 1,3,5,7,9,\ k_2 = 3,4,5\}$, and choose the parameters $\tau_1,\ \tau_2$ from $\{0.1/\eta_0, 0.05/\eta_0, 0.01/\eta_0, 0.005/\eta_0, 0.001/\eta_0\}$, to find a combination of (η_0, τ_1, τ_2) that yields the most significant increase in accuracy after 20 epochs. All other parameters for these Adam-family methods in PyTorch remain fixed at their default values.

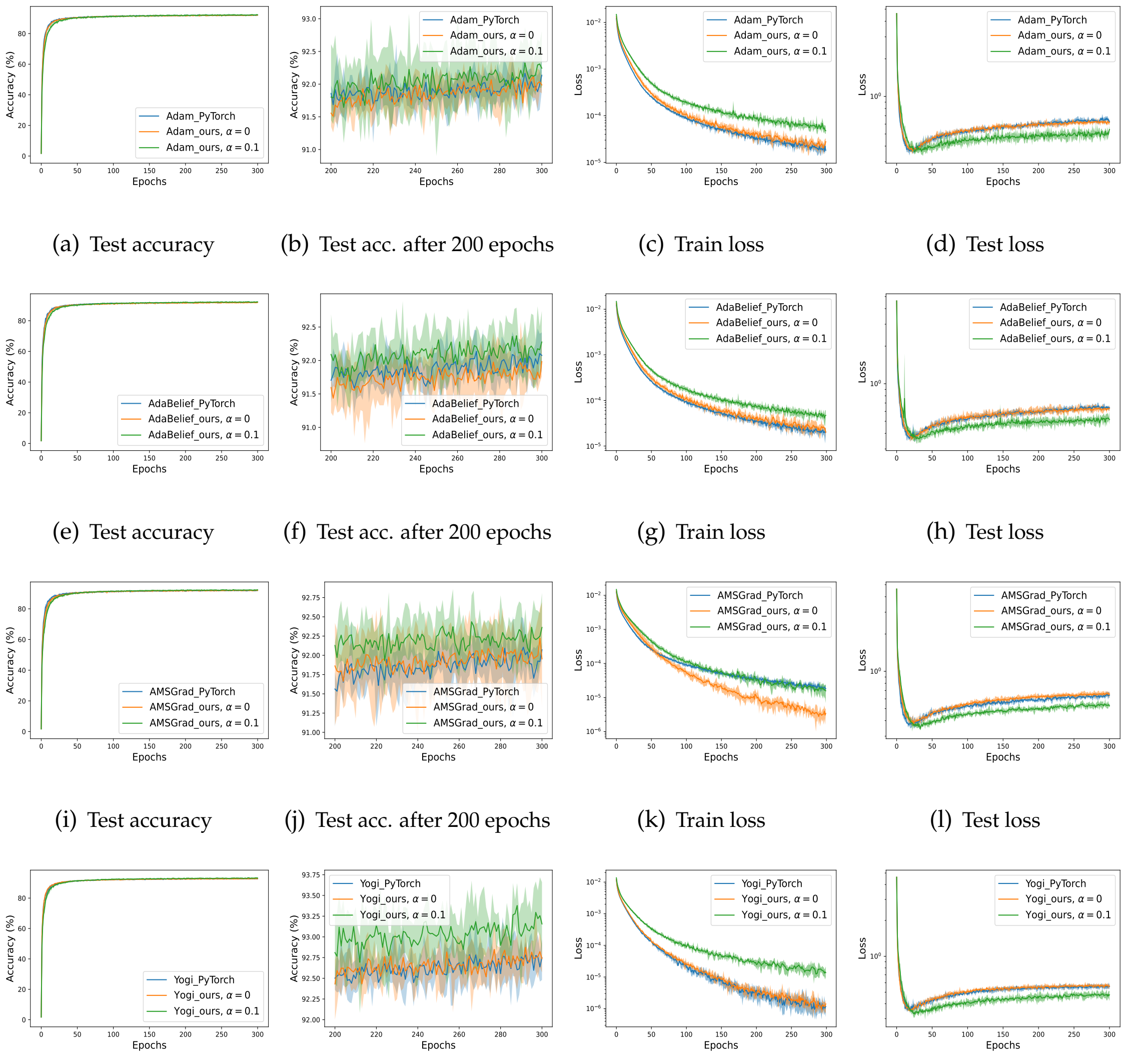
For our proposed algorithms (i.e., Adam-family methods with diminishing stepsizes for $\{m_k\}$ and $\{v_k\}$ as in Algorithm 1), we keep all other parameters the same as those available in PyTorch, as we aim to perform minimal modifications to their released counterparts. Moreover, to investigate the performance of our proposed Adam-family methods with the Nesterov momentum term, in each test instance, we choose the Nesterov momentum parameter α as 0 and 0.1, respectively. We run each test instance five times with different random seeds. In each test instance, all compared methods are tested using the same random seed and initialized with the same random weights by the default initialization function in PyTorch.

The numerical results are presented in Figure 1 and Figure 2. These figures demonstrate that our proposed Adam-family methods with diminishing stepsizes exhibit the same performance as the existing Adam-family methods available in PyTorch and torch-optimizer packages. These empirical results highlight the effectiveness of our proposed Adam-family methods, as they achieve comparable performance to their widely used counterparts in the community. Furthermore, we note that the integration of Nesterov momentum can potentially lead to increased accuracy and reduced test loss across all tested Adam-family methods in our numerical experiments, especially in the classification tasks on the CIFAR-10 data set. These empirical results, together with our presented theoretical analysis, demonstrate that by simply choosing diminishing stepsizes for the momentum terms and estimators in existing Adam-family methods, we can preserve their high performance in practice while benefiting from the convergence guarantees in training nonsmooth neural networks.

6.2 Gradient Clipping

In this subsection, we evaluate the numerical performance of our proposed stochastic subgradient methods with gradient clipping technique by comparing them to the optimizers provided by PyTorch. We conduct numerical experiments using the LeNet (LeCun et al., 1998) for the classification task on the MNIST data set (LeCun, 1998). Following the settings in (Grandvalet et al., 1997; Maaten et al., 2013), the training samples are randomly perturbed by noise following the Levy stable distribution, with the stability parameter of 1.1, the skewness parameter of 1, and the scale of 0.2. Consequently, the imposed perturbation noise exhibits zero mean without finite second-order moment (hence it is heavy-tailed). In addition, we test the numerical performance of these compared stochastic subgradient methods in training language models (Vaswani et al., 2017), and present the results in Appendix C.

In our numerical experiments, we set the batch size to 64 for all test instances and select the regularization parameter $\varepsilon = 10^{-15}$ for all the Adam-family methods. Moreover, at the k-th epoch, we choose the stepsize as $\eta_k = \frac{\eta_0}{\sqrt{k+1}}$ for all tested algorithms. For all compared optimizers, we choose the initial stepsize η_0 and the momentum parameters τ_1, τ_2 using the same grid search method as in Section 6.1, and retain all other parameters at their default values for the optimizers in PyTorch. We run each test instance 5 times with different random seeds. In each test instance, all compared algorithms are tested using the same random seed and initialized with the same random weights by the default initialization function in PyTorch.



(a) Test accuracy (b) Test acc. after 200 epochs (c) Train loss (d) Test loss

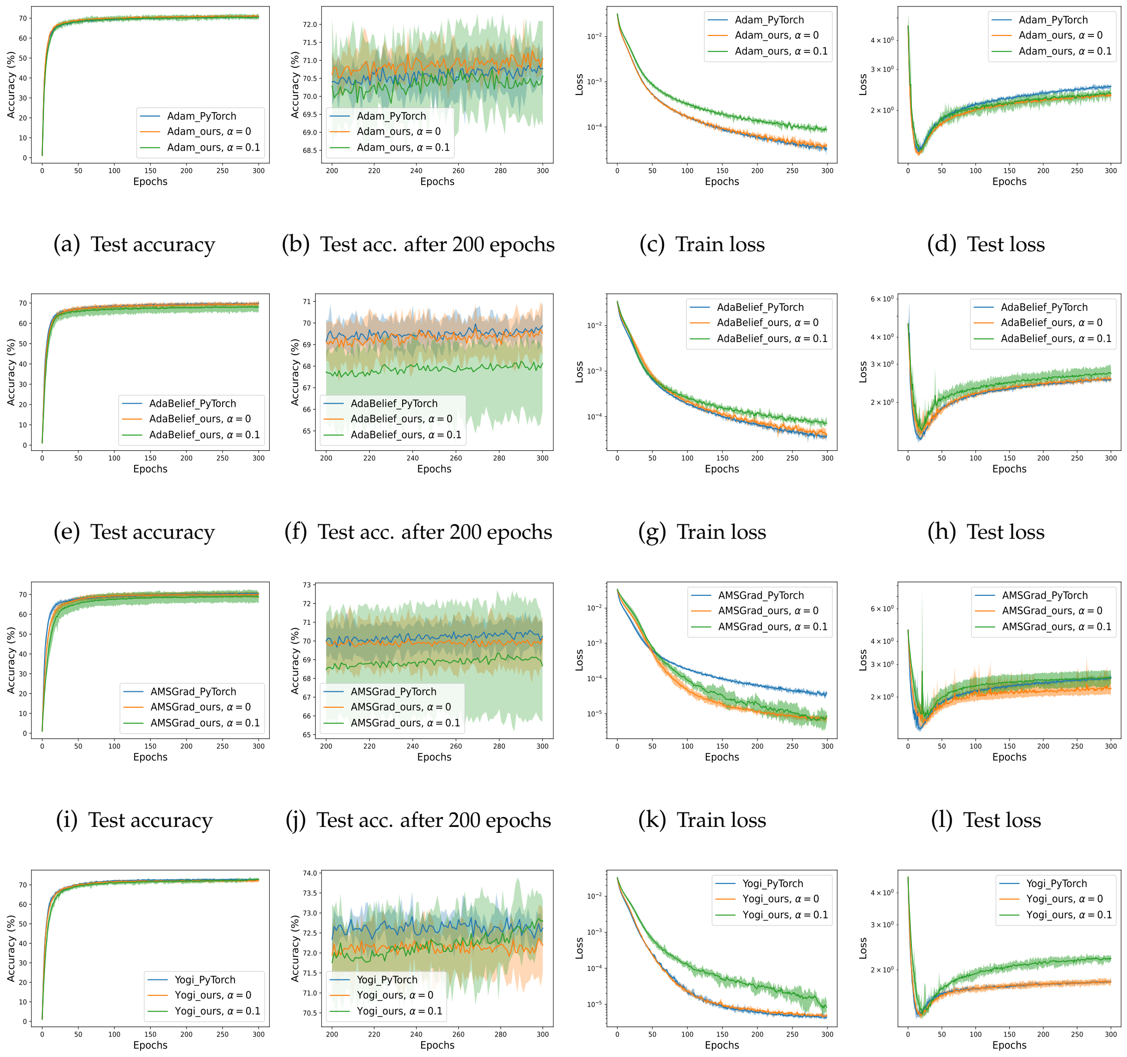
(e) Test accuracy (f) Test acc. after 200 epochs (g) Train loss (h) Test loss

(i) Test accuracy (j) Test acc. after 200 epochs (k) Train loss (l) Test loss

(m) Test accuracy (n) Test acc. after 200 epochs (o) Train loss (p) Test loss

Figure 1: Test results on CIFAR-10 data set with ResNet50. Here “acc.” is the abbreviation of “accuracy”.

The numerical results are presented in Figure 3. These figures indicate that our proposed SGD-C and ADAM-C converge successfully and achieve high accuracy. In contrast, without gradient clipping, SGD fails to converge and Adam converges much slower than ADAM-C. Moreover, compared with SGD-C, ADAM-C achieves improved accuracy and a faster decrease in the loss curve. Therefore, we can conclude that with the gradient clipping technique, our proposed Adam-family method (12) outperforms SGD-C and the Adam provided in PyTorch. These observations further demonstrate the great potential of our proposed stochastic subgradient methods with gradient clipping in solving UNP in the presence of heavy-tailed noises.



(a) Test accuracy (b) Test acc. after 200 epochs (c) Train loss (d) Test loss

(e) Test accuracy (f) Test acc. after 200 epochs (g) Train loss (h) Test loss

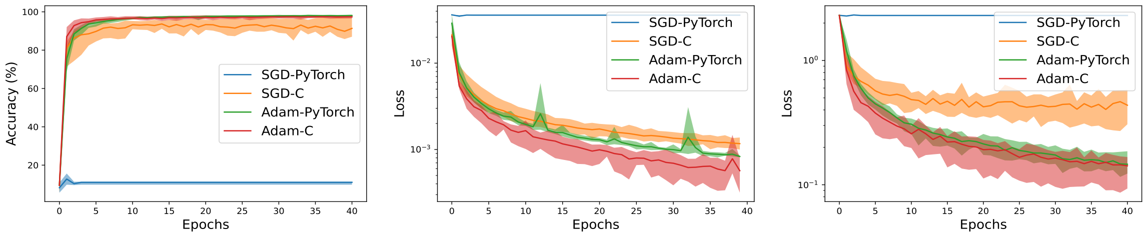
(i) Test accuracy (j) Test acc. after 200 epochs (k) Train loss (l) Test loss

(m) Test accuracy (n) Test acc. after 200 epochs (o) Train loss (p) Test loss

Figure 2: Test results on CIFAR-100 data set with ResNet50. Here "acc." is the abbreviation of "accuracy".

7. Conclusion

Adam-family methods are powerful tools for nonsmooth optimization, especially in training neural networks. However, as most of the neural networks are built from nonsmooth blocks, their loss functions are typically nonsmooth and not Clarke regular, thus leading to great difficulties in analyzing the convergence properties for these methods. Additionally, the presence of heavy-tailed evaluation noises in numerous applications of UNP poses significant challenges in designing efficient algorithms and establishing theoretical guarantees for UNP.



(a) Test accuracy

(b) Train loss

(c) Test loss

Figure 3: Test results on MNIST data set with LeNet.

The primary contributions of this paper can be summarized as follows:

- **A novel framework for Adam-family methods**
  To establish convergence properties for Adam-family methods, we first introduce a two-timescale framework (AFM) that assigns different stepsizes to the updating directions and evaluation noises, respectively. Then we establish convergence properties for (AFM) in the sense of conservative field under mild assumptions. Furthermore, we prove that under mild assumptions with almost every initialized stepsize and initial point, any cluster point of the sequence generated by our proposed framework is a Clarke stationary point of the objective function. These results provide theoretical guarantees for our proposed framework (AFM). In particular, although AD algorithms may introduce spurious stationary points to UNP, we prove that our proposed framework (AFM) can avoid these spurious stationary points for almost every initial point and stepsize.

- **Convergence properties for Adam-family methods**
  We show that Adam, AdaBelief, AMSGrad, NAdam and Yogi, when equipped with diminishing stepsizes, follow our proposed framework (AFM). Consequently, through our established results for (AFM), we provide a convergence analysis for these Adam-family methods under mild assumptions in the sense of both conservative field and Clarke subdifferential. These results are applicable to a wide range of neural network training problems, hence providing convergence guarantees for the application of these Adam-family methods in training nonsmooth neural networks.

- **Gradient clipping technique for heavy-tailed noises**
  We develop stochastic subgradient methods that incorporate the gradient clipping technique based on our proposed framework. Under mild assumptions and appropriately chosen clipping parameters, we show that these stochastic subgradient methods conform to our proposed framework (AFM) even when the evaluation noises are only assumed to be integrable. Therefore, by employing the gradient clipping technique to tackle heavy-tailed evaluation noises, a wide range of stochastic subgradient methods can be developed with guaranteed convergence properties for solving UNP.

Furthermore, we conduct extensive numerical experiments to illustrate that our proposed Adam-family methods are as efficient as the widely employed Adam-family methods provided by PyTorch. Additionally, preliminary numerical experiments demonstrate the high efficiency and robustness of our proposed stochastic subgradient methods with gradient clipping in training neural networks with heavy-tailed evaluation noises. Therefore, we can conclude that our results have provided theoretical guarantees for Adam-family methods in practical settings, especially when the neural networks are nonsmooth or the evaluation noises are heavy-tailed.

Future research questions of this work include establishing convergence rates and complexity results for Adam-family methods in minimizing nonsmooth and non-regular functions, which are extremely challenging to tackle. Most existing works focus on SGD with the exact evaluation of the Clarke subdifferential, or only consider the convergence rate of the trajectories of the corresponding noiseless differential inclusions (Castera et al., 2021). To the best of our knowledge, there is no existing result for establishing the complexity for stochastic subgradient methods in the form of Algorithm 1 when f is only assumed to be a potential function. Furthermore, as the gradient clipping technique is widely employed in various natural language processing tasks, future works of this paper could investigate the performance of our proposed stochastic subgradient methods with gradient clipping in these real-world applications of UNP.

Acknowledgement

We thank the reviewers for their valuable suggestions that have significantly contributed to the improvement of our paper.

The research of Nachuan Xiao and Kim-Chuan Toh is supported by Academic Research Fund Tier 3 grant call (MOE-2019-T3-1-010). The research of Xiaoyin Hu is supported by the National Natural Science Foundation of China (No. 12301408), Zhejiang Provincial Natural Science Foundation of China under Grant (No. LQ23A010002), Scientific research project of Zhejiang Provincial Education Department (Y202248716), and the advanced computing resources provided by the Supercomputing Center of HZCU. The research of Xin Liu is supported in part by the National Natural Science Foundation of China (12125108, 12226008, 12021001, 12288201, 11991021), Key Research Program of Frontier Sciences, Chinese Academy of Sciences (ZDBS-LY-7022), and CAS AMSS-PolyU Joint Laboratory of Applied Mathematics.

Appendix A. Proof for Proposition 6

In this section, we present the proof for Proposition 6. We begin our proof with the following auxiliary proposition, which employs a similar proof technique as (Xiao et al., 2023, Proposition 4.10).

Proposition 8 *Suppose $\{\xi_k\}$ is a sequence of uniformly bounded martingale difference sequence, $\{\eta_k\}$ and $\{C_k\}$ are positive sequences that satisfies*

$$\lim_{k\to+\infty} \eta_k = 0, \quad \lim_{k\to+\infty} C_k = +\infty, \quad \textit{and} \quad \lim_{k\to+\infty} C_k^2 \eta_k \log(k) = 0.$$

Then almost surely, it holds that

$$\lim_{k \to +\infty} \sum_{i=0}^{k} \left(\eta_i \prod_{j=i+1}^{k} (1-\eta_j) \right) C_i \xi_i = 0.$$

Proof Let $z_k = \sum_{i=0}^{k} \left(\eta_i \prod_{j=i+1}^{k} (1-\eta_j) \right) C_i\xi_i$ and $z_0 = 0$, $\rho_{k,i} := C_i \eta_i \prod_{j=i+1}^{k}(1-\eta_j)$ and $\rho_{k,k} := C_k \eta_k$, then there exists $K > 0$ such that $|\rho_{k,i}| \leq C_i \eta_i$ holds for any $k \geq i \geq K$. Without loss of generality, we assume that $\rho_{k,i} \geq 0$ holds for any $k \geq i \geq K$. Moreover, from the expression of z_k, we can conclude that

$$z_k = \sum_{i=0}^{k} \rho_{k,i} \xi_i.$$

Since the martingale difference sequence $\{\xi_{m,k}\}$ is uniformly bounded, it holds that $\xi_{m,k}$ is sub-Gaussian for any $k \geq 0$. Then there exists a constant $M > 0$ such that for any $w \in \mathbb{R}^n$, it holds for any $k \geq 0$ that

$$\mathbb{E}\left[\exp\left(\langle w, \xi_{m,k+1} \rangle\right) | \mathcal{F}_k\right] \leq \exp\left(\frac{M}{2}\|w\|^2\right).$$

Therefore, for any $s > K$, $T > 0$, $w \in \mathbb{R}^n$ and any $C > 0$, let

$$Z_i := \exp\left[\left\langle Cw, \sum_{k=s}^{i} \rho_{\Lambda(\lambda_s+T),k} \xi_{m,k} \right\rangle - \frac{MC^2}{2} \sum_{k=s}^{i} \rho^2_{\Lambda(\lambda_s+T),k} \|w\|^2 \right],$$

where $\Lambda(0) := 0$, $\Lambda(i) := \sum_{k=0}^{i-1} \eta_k$, and $\Lambda(t) := \sup\{k \geq 0 : t \geq \Lambda(k)\}$. Then for any $i \geq s$, we have that $\mathbb{E}[Z_{i+1}|\mathcal{F}_i] \leq Z_i$. Hence for any $\delta > 0$, and any $C > 0$, it holds that

$$\begin{aligned}
&\mathbb{P}\left(\sup_{s \leq i \leq \Lambda(\lambda_s+T)} \left\langle w, \sum_{k=s}^{i} \rho_{\Lambda(\lambda_s+T),k} \xi_k \right\rangle > \delta \right) \\
&= \mathbb{P}\left(\sup_{s \leq i \leq \Lambda(\lambda_s+T)} \left\langle Cw, \sum_{k=s}^{i} \rho_{\Lambda(\lambda_s+T),k} \xi_k \right\rangle > C\delta \right) \\
&\leq \mathbb{P}\left(\sup_{s \leq i \leq \Lambda(\lambda_s+T)} Z_i > \exp\left(C\delta - \frac{MC^2}{2} \sum_{k=s}^{\Lambda(\lambda_s+T)} \rho^2_{\Lambda(\lambda_s+T),k} \|w\|^2 \right)\right) \\
&\leq \exp\left(\left(\frac{M}{2}\|w\|^2 \sum_{k=s}^{\Lambda(\lambda_s+T)} \rho^2_{\Lambda(\lambda_s+T),k}\right) C^2 - \delta C\right).
\end{aligned}$$

Here the second inequality holds since $\{Z_i\}$ is nonnegative, $\mathbb{E}[Z_{i+1}|\mathcal{F}_i] \leq Z_i$ holds for any $i \geq 0$ and $\mathbb{E}[Z_s] \leq 1$. Then from the arbitrariness of C, we can set $C = \frac{\delta}{M\|w\|^2 \sum_{k=s}^{\Lambda(\lambda_s+T)} \rho^2_{\Lambda(\lambda_s+T),k}}$ to obtain that

$$\mathbb{P}\left(\sup_{s \leq i \leq \Lambda(\lambda_s+T)} \left\langle w, \sum_{k=s}^{i} \rho_{\Lambda(\lambda_s+T),k} \xi_k \right\rangle > \delta \right) \leq \exp\left(\frac{-\delta^2}{2M\|w\|^2 \sum_{k=s}^{\Lambda(\lambda_s+T)} \rho^2_{\Lambda(\lambda_s+T),k}}\right).$$

From the arbitrariness of w and the fact that $\rho_{\Lambda(\lambda_s+T),k} \leq \eta_k$, we can deduce that there exists constants C_1, C_2 that only depend on n, such that

$$\begin{aligned}
&\mathbb{P}\left(\sup_{s\leq i\leq \Lambda(\lambda_s+T)} \left\|\sum_{k=s}^{i} \rho_{\Lambda(\lambda_s+T),k}\xi_{m,k}\right\| > \delta\right) \\
&\leq C_1 \exp\left(\frac{-\delta^2}{2C_2M\sum_{k=s}^{\Lambda(\lambda_s+T)} \rho^2_{\Lambda(\lambda_s+T),k}}\right) \leq C_1 \exp\left(\frac{-\delta^2}{2C_2M\sum_{k=s}^{\Lambda(\lambda_s+T)} C_k^2\eta_k^2}\right) \\
&\leq C_1 \exp\left(\frac{-\delta^2}{2C_2M\eta_{k'}C_{k'}^2\sum_{k=s}^{\Lambda(\lambda_s+T)} \eta_k}\right) \leq \exp\left(\frac{-\delta^2}{2MT\eta_{k'}C_{k'}^2}\right),
\end{aligned}$$

holds for some $k' \in [s, \Lambda(\lambda_s + T)]$.

Therefore, for any $j \geq 0$, there exists $k_j \in [\Lambda(jT), \Lambda((j+1)T)]$, such that

$$\begin{aligned}
&\sum_{j=0}^{+\infty} \mathbb{P}\left(\sup_{\Lambda(jT)\leq i\leq \Lambda(jT+T)} \left\|\sum_{k=\Lambda(jT)}^{i} \rho_{\Lambda(jT+T),k}\xi_k\right\| > \delta\right) \\
&\leq \sum_{j=0}^{+\infty} \exp\left(\frac{-\delta^2}{2MT\eta_{k_j}C_{k_j}^2}\right) \leq \sum_{k=0}^{+\infty} \exp\left(\frac{-\delta^2}{2MT\eta_k C_k^2}\right) < +\infty.
\end{aligned} \tag{20}$$

Here the last inequality holds from the fact that $\lim_{k\to+\infty} \eta_k C_k^2 \log(k) = 0$.

Therefore, let $\mathcal{E}_j$ denote the event $\left\{\sup_{\Lambda(jT)\leq i\leq \Lambda(jT+T)} \left\|\sum_{k=\Lambda(jT)}^{i} \rho_{\Lambda(jT+T),k}\xi_k\right\| > \delta\right\}$. From the Borel-Cantelli lemma and (20), we can conclude that $\mathbb{P}\left(\lim_{j\to+\infty} \cap_{j=1}^{+\infty} \cup_{l=j}^{+\infty} \mathcal{E}_l\right) = 0$. Therefore, we can conclude that, almost surely,

$$\lim_{j\to+\infty} \sup_{\Lambda(jT)\leq i\leq \Lambda(jT+T)} \left\|\sum_{k=\Lambda(jT)}^{i} \rho_{\Lambda(jT+T),k}\xi_k\right\| \leq \delta.$$

Then the arbitrariness of δ illustrates that, almost surely, we have

$$\lim_{j\to+\infty} \sup_{\Lambda(jT)\leq i\leq \Lambda(jT+T)} \left\|\sum_{k=\Lambda(jT)}^{i} \rho_{\Lambda(jT+T),k}\xi_k\right\| = 0. \tag{21}$$

Notice that for any $j \geq 0$ such that $\Lambda(jT) \geq K$, it holds that

$$z_{\Lambda(jT+T)} = \left(\prod_{k=\Lambda(jT)}^{\Lambda(jT+T)} (1-\eta_k)\right) z_{\Lambda(jT)} + \sum_{k=\Lambda(jT)}^{\Lambda(jT+T)} \rho_{\Lambda(jT+T),k}\xi_k,$$

which illustrates that almost surely,

$$\left\|z_{\Lambda(jT+T)}\right\| \leq \exp(-T)\left\|z_{\Lambda(jT)}\right\| + \left\|\sum_{k=\Lambda(jT)}^{\Lambda(jT+T)} \rho_{\Lambda(jT+T),k}\xi_{m,k}\right\|, \tag{22}$$

hence $\lim_{j\to+\infty}\left\|z_{\Lambda(jT)}\right\| = 0$.

Finally, for any i such that $\Lambda(jT) \leq i \leq \Lambda(jT+T)$, it holds that

$$\begin{aligned}
\left\|z_{\Lambda(jT+T)}\right\| ={}& \left\|\left(\prod_{k=i}^{\Lambda(jT+T)}(1-\eta_k)\right) z_i + \sum_{k=i}^{\Lambda(jT+T)} \rho_{\Lambda(jT+T),k}\xi_k\right\| \\
\geq{}& \exp(-T)\left\|z_i\right\| - \left\|\sum_{k=i}^{\Lambda(jT+T)} \rho_{\Lambda(jT+T),k}\xi_k\right\| \\
\geq{}& \exp(-T)\left\|z_i\right\| - 2\sup_{\Lambda(jT)\leq i\leq\Lambda(\lambda_s+T)}\left\|\sum_{k=\Lambda(jT)}^{i} \rho_{\Lambda(\lambda_s+T),k}\xi_k\right\|.
\end{aligned}$$

As a result, we have

$$\begin{aligned}
&\sup_{\Lambda(jT)\leq i\leq\Lambda(jT+T)}\left\|z_i\right\| \\
\leq{}& \exp(T)\left(\left\|z_{\Lambda(jT)}\right\| + 2\sup_{\Lambda(jT)\leq i\leq\Lambda(\lambda_s+T)}\left\|\sum_{k=\Lambda(jT)}^{i} \rho_{\Lambda(\lambda_s+T),k}\xi_k\right\|\right).
\end{aligned} \tag{23}$$

Combining (21), (22), and (23) together, we achieve that,

$$\begin{aligned}
&\limsup_{k\to+\infty}\left\|z_k\right\| = \lim_{j\to+\infty}\sup_{\Lambda(jT)\leq i\leq\Lambda(jT+T)}\left\|z_i\right\| \\
\leq{}& \exp(T)\lim_{j\to+\infty}\left\|z_{\Lambda(jT)}\right\| + 2\exp(T)\lim_{j\to+\infty}\sup_{\Lambda(jT)\leq i\leq\Lambda(\lambda_s+T)}\left\|\sum_{k=\Lambda(jT)}^{i} \rho_{\Lambda(\lambda_s+T),k}\xi_k\right\| \\
={}& 0.
\end{aligned}$$

holds almost surely. This completes the proof.

■

With Proposition 8, we now present the proof for Proposition 6.

Proof of Proposition 6:

For any $k \geq 0$, the m_{k+1} in SGD-C can be expressed as

$$\begin{aligned}
m_{k+1} ={}& \sum_{i=0}^{k}\left(\prod_{j=i+1}^{k}(1-\tau_1\eta_j)\right)\tau_1\eta_i\hat{g}_i \\
={}& \sum_{i=0}^{k}\left(\prod_{j=i+1}^{k}(1-\tau_1\eta_j)\right)\tau_1\eta_i(d_i + C_i\xi_i) \\
={}& \sum_{i=0}^{k}\left(\prod_{j=i+1}^{k}(1-\tau_1\eta_j)\right)\tau_1\eta_i d_i + \sum_{i=0}^{k}\left(\prod_{j=i+1}^{k}(1-\tau_1\eta_j)\right)\tau_1\eta_i C_i\xi_i.
\end{aligned}$$

Here we set $\prod_{i+1}^{i}(1-\tau_1\eta_j) = 1$ for simplicity.

As illustrated in Proposition 8, almost surely, it holds that

$$\lim_{k\to+\infty}\sum_{i=0}^{k}\left(\prod_{j=i+1}^{k}(1-\tau_1\eta_j)\right)\tau_1\eta_i C_i\xi_i = 0,$$

which implies that $\sup_{k\geq 0}\left\|\sum_{i=0}^{k}\left(\prod_{j=i+1}^{k}(1-\tau_1\eta_j)\right)\tau_1\eta_i C_i\xi_i\right\| < +\infty$.

On the other hand, Lemma 8 illustrates that there exists a nonnegative diminishing sequence $\{\delta_k\}$ such that $d_k \in \mathcal{D}_f^{\delta_k}(x_k)$ holds for any $k \geq 0$. Then from the local boundedness of $\mathcal{D}_f$ and the uniform boundedness of the sequence $\{x_k\}$, we have that $\sup_{k\geq 0}\|d_k\| < +\infty$ holds almost surely. Then for any $k > 0$, it holds that

$$\left\|\sum_{i=0}^{k}\left(\prod_{j=i+1}^{k}(1-\tau_1\eta_j)\right)\tau_1\eta_i d_i\right\| \leq \sum_{i=0}^{k}\left(\prod_{j=i+1}^{k}(1-\tau_1\eta_j)\right)\tau_1\eta_i\|d_i\| \leq \sup_{0\leq i\leq k}\|d_i\|.$$

Then we can conclude that

$$\sup_{k\geq 0}\|m_k\| \leq \sup_{k\geq 0}\|d_k\| + \sup_{k\geq 0}\left\|\sum_{i=0}^{k}\left(\prod_{j=i+1}^{k}(1-\tau_1\eta_j)\right)\tau_1\eta_i C_i\xi_i\right\| < +\infty.$$

This completes the proof of Proposition 6.

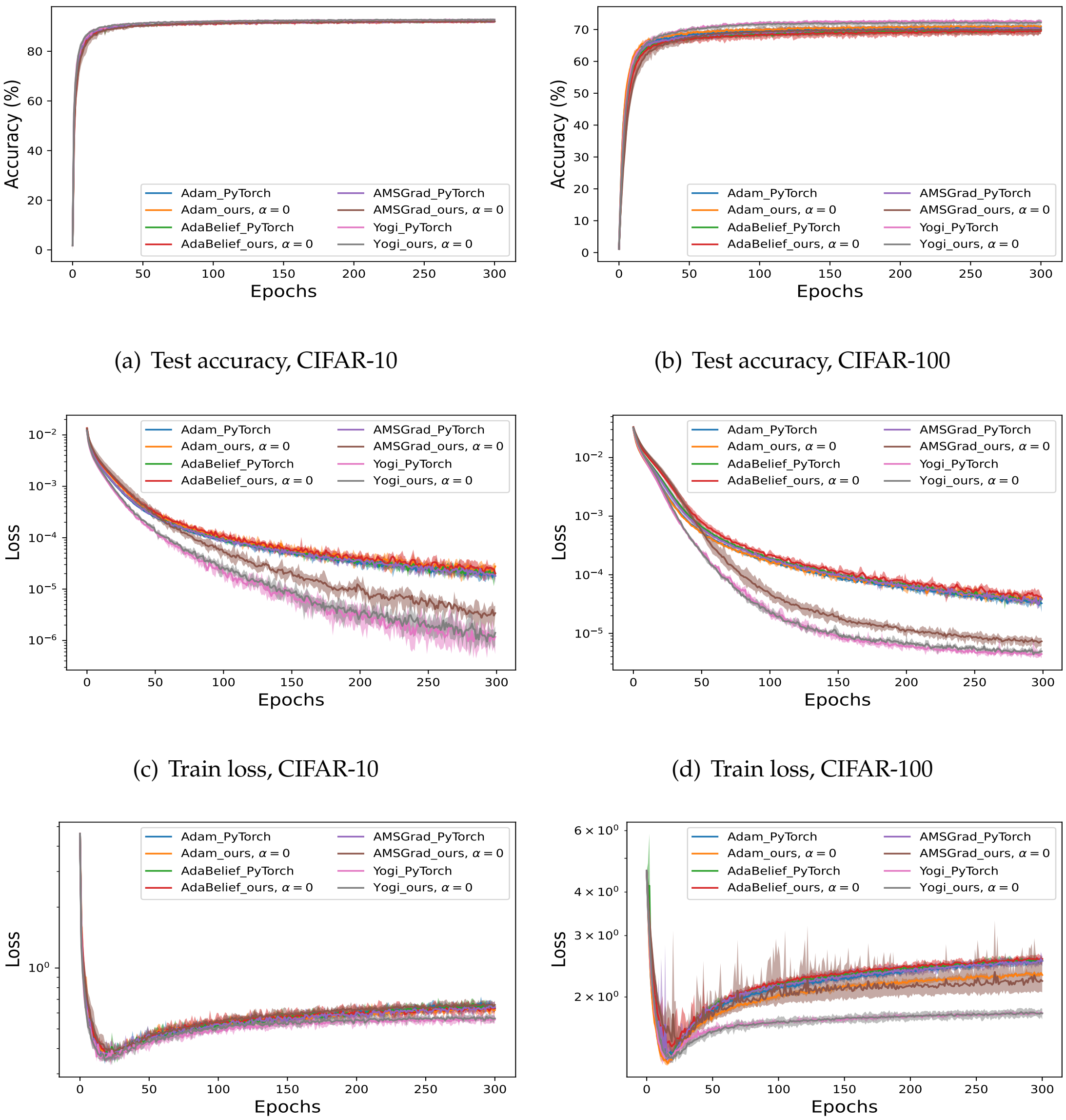
Appendix B. Supplementary Numerical Experiments for Section 6.1

In this section, we present the supplementary results for the numerical experiments in Section 6.1. In Figure 4, we present the performance of all the compared Adam-family methods, in the aspects of test accuracy, test error, and train error. In particular, the curves of all the compared Adam-family methods are plotted in a single subfigure in Figure 4, for a better illustration on the performances of different Adam-family methods.

Appendix C. Supplementary Numerical Experiments for Section 6.2

In this section, we present numerical experiments on testing the efficiency of our proposed stochastic subgradient methods with clipping on the natural language processing (NLP) tasks. In these numerical experiments, different from the settings in Section 6.2, we do not introduce any corruption to the training samples.

We first evaluate the performance of (SGD-C) and (ADAM-C) by training language-translation model (Vaswani et al., 2017) on the Multi30k data set (Elliott et al., 2016). In our numerical experiments, we choose the language translation model as the Seq2Seq network with transformer proposed by (Vaswani et al., 2017). Similar to the settings in Section 6.2, we set the batch size to 128 for all test instances and select the regularization parameter $\varepsilon = 10^{-15}$ for all the Adam-family methods. Moreover, at the k-th epoch, we choose the stepsize as $\eta_k = \frac{\eta_0}{\sqrt{k+1}}$ for all tested algorithms. For all compared optimizers, we choose the initial stepsize η_0 and the momentum parameters τ_1, τ_2 using the same grid search method as in Section 6.1, and retain all other parameters at their default values for the optimizers in PyTorch. We run each test instance 5 times with different random seeds. In each test



(a) Test accuracy, CIFAR-10

(b) Test accuracy, CIFAR-100

(c) Train loss, CIFAR-10

(d) Train loss, CIFAR-100

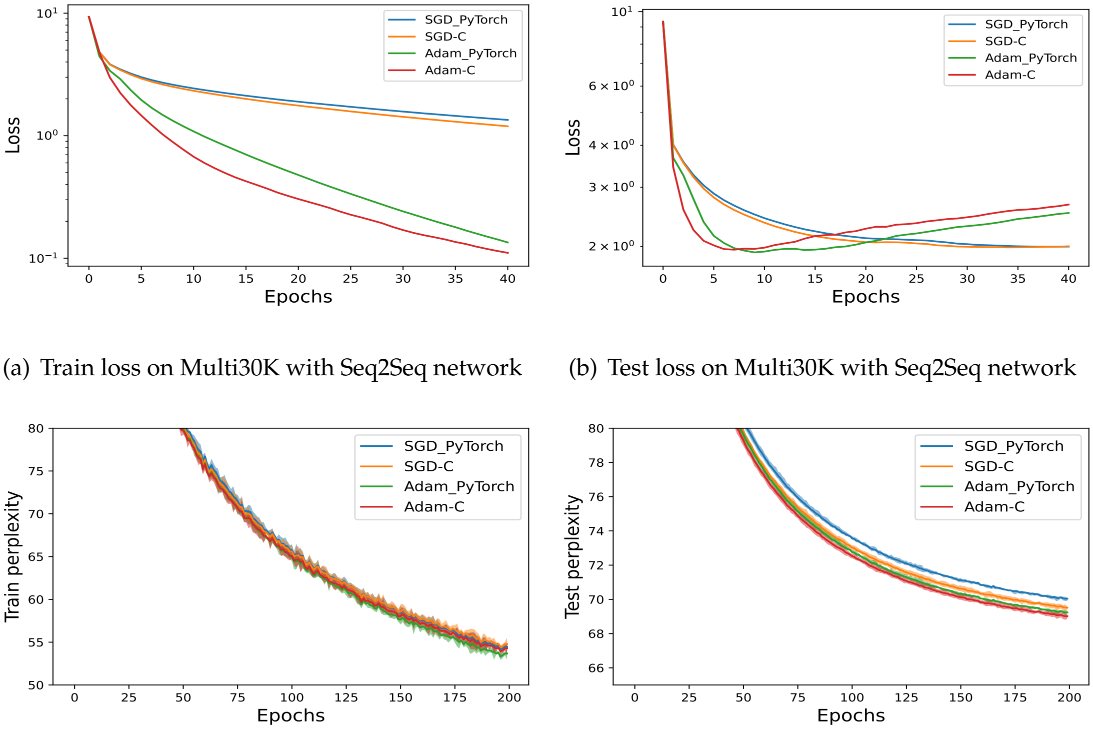
(e) Test loss, CIFAR-10

(f) Test loss, CIFAR-100

Figure 4: Test results on CIFAR data sets with ResNet50.

instance, all compared algorithms are tested using the same random seed and initialized with the same random weights by the default initialization function in PyTorch.

Then we evaluate the efficiency of all the compared methods in training 3-layer long short-term memory (LSTM) models. In all the numerical experiments, we consistently train our models for 200 epochs while employing a batch size of 128. These settings adhere to the commonly used experimental setup for training LSTM models, as demonstrated in previous works (Zhuang et al., 2020).



(a) Train loss on Multi30K with Seq2Seq network

(b) Test loss on Multi30K with Seq2Seq network

(c) Train perplexity on Penn Treebank with LSTM

(d) Test perplexity on Penn Treebank with LSTM

Figure 5: Numerical results on NLP tasks.

Figure 5 exhibits the results of our numerical experiments with error bars. Notably, although we run each compared method for 5 times with different random seeds, the loss curves seem to be very close. As depicted in Figure 5, the method outlined in (SGD-C) slightly outperforms the default SGD method in PyTorch. Moreover, the ADAM-C method achieves slightly lower training loss than the build-in Adam method in PyTorch, although the test loss for ADAM-C is slightly worse than that of Adam-PyTorch. These observations illustrate that even when the training samples are finite, the stochastic subgradient methods with gradient clipping exhibit slightly better performance in the training of language models. Combined with the results in Section 6.2, our numerical experiments results illustrate the potential of our proposed stochastic subgradient methods with gradient clipping technique for solving UNP.

References

Anas Barakat and Pascal Bianchi. Convergence and dynamical behavior of the ADAM algorithm for nonconvex stochastic optimization. *SIAM Journal on Optimization*, 31(1): 244–274, 2021.

Anas Barakat, Pascal Bianchi, Walid Hachem, and Sholom Schechtman. Stochastic optimization with momentum: convergence, fluctuations, and traps avoidance. *Electronic Journal of Statistics*, 15(2):3892–3947, 2021.

Paul I Barton, Kamil A Khan, Peter Stechlinski, and Harry AJ Watson. Computationally relevant generalized derivatives: theory, evaluation and applications. *Optimization Methods and Software*, 33(4-6):1030–1072, 2018.

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.

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.

Edward Bierstone and Pierre D Milman. Semianalytic and subanalytic sets. *Publications Mathématiques de l'IHÉS*, 67:5–42, 1988.

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, 2021.

Jérôme Bolte, Tam Le, and Edouard Pauwels. Subgradient sampling for nonsmooth nonconvex minimization. *SIAM Journal on Optimization*, 33(4):2542–2569, 2023.

Alexander Camuto, Xiaoyu Wang, Lingjiong Zhu, Chris Holmes, Mert Gurbuzbalaban, and Umut Simsekli. Asymmetric heavy tails and implicit bias in gaussian noise injections. In *International Conference on Machine Learning*, pages 1249–1260. PMLR, 2021.

Camille Castera, Jérôme Bolte, Cédric Févotte, and Edouard Pauwels. An inertial Newton algorithm for deep learning. *Journal of Machine Learning Research*, 22(134):1–31, 2021.

Congliang Chen, Li Shen, Fangyu Zou, and Wei Liu. Towards practical Adam: Nonconvexity, convergence theory, and mini-batch acceleration. *Journal of Machine Learning Research*, 23:1–47, 2022.

Frank H Clarke. *Optimization and Nonsmooth Analysis*, volume 5. SIAM, 1990.

André Belotto Da Silva and Maxime Gazeau. A general system of differential equations to model first-order adaptive algorithms. *The Journal of Machine Learning Research*, 21(1): 5072–5113, 2020.

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.

Soham De, Anirbit Mukherjee, and Enayat Ullah. Convergence guarantees for RMSProp and ADAM in non-convex optimization and an empirical comparison to nesterov acceleration. *arXiv preprint arXiv:1807.06766*, 2018.

Timothy Dozat. Incorporating Nesterov Momentum into Adam. In *Proceedings of the 4th International Conference on Learning Representations*, pages 1–4, 2016.

Bryn Elesedy and Marcus Hutter. U-clip: On-average unbiased stochastic gradient clipping. *arXiv preprint arXiv:2302.02971*, 2023.

D. Elliott, S. Frank, K. Sima'an, and L. Specia. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, 2016.

Qiankun Fu, Yue Zhang, Jiangming Liu, and Meishan Zhang. DRTS parsing with structure-aware encoding and decoding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6818–6828, Online, July 2020. Association for Computational Linguistics.

Qiankun Fu, Linfeng Song, Wenyu Du, and Yue Zhang. End-to-end AMR coreference resolution. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4204–4214, Online, August 2021. Association for Computational Linguistics.

Sébastien Gadat and Ioana Gavra. Asymptotic study of stochastic adaptive algorithms in non-convex landscape. *The Journal of Machine Learning Research*, 23(1):10357–10410, 2022.

Eduard Gorbunov, Marina Danilova, and Alexander Gasnikov. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. *Advances in Neural Information Processing Systems*, 33:15042–15053, 2020.

Yves Grandvalet, Stéphane Canu, and Stéphane Boucheron. Noise injection: Theoretical prospects. *Neural Computation*, 9(5):1093–1108, 1997.

Zhishuai Guo, Yi Xu, Wotao Yin, Rong Jin, and Tianbao Yang. A novel convergence analysis for algorithms of the Adam family. *NeurIPS OPT Workshop*, 2021.

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 (CVPR)*, pages 770–778, 2016.

Xiaoyin Hu, Nachuan Xiao, Xin Liu, and Kim-Chuan Toh. An improved unconstrained approach for bilevel optimization. *SIAM Journal on Optimization*, 33(4):2801–2829, 2023.

Cédric Josz and Lexiao Lai. Global stability of first-order methods for coercive tame functions. *Mathematical Programming*, pages 1–26, 2023.

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

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *In Proceedings of the 3rd International Conference for Learning Representations*, 2015.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Technical report*, Toronto, ON, Canada, 2009.

Tam Le. Nonsmooth nonconvex stochastic heavy ball. *arXiv preprint arXiv:2304.13328*, 2023.

Yann LeCun. The mnist database of handwritten digits, 1998. URL `http://yann.lecun.com/exdb/mnist/`.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Xiao Li and Andre Milzarek. A unified convergence theorem for stochastic optimization methods. *Advances in Neural Information Processing Systems*, 35:33107–33119, 2022.

Nicolas Loizou and Peter Richtárik. Linearly convergent stochastic heavy ball method for minimizing generalization error. *NIPS-Workshop on Optimization for Machine Learning*, 2017.

Laurens Maaten, Minmin Chen, Stephen Tyree, and Kilian Weinberger. Learning with marginalized corrupted features. In *International Conference on Machine Learning*, pages 410–418. PMLR, 2013.

Michael Mahoney and Charles Martin. Traditional and heavy tailed self regularization in neural network models. In *International Conference on Machine Learning*, pages 4284–4293. PMLR, 2019.

Vien V Mai and Mikael Johansson. Stability and convergence of stochastic gradient clipping: Beyond lipschitz continuity and smoothness. In *International Conference on Machine Learning*, pages 7325–7335. PMLR, 2021.

Zhao-Yan Ming, Jingjing Chen, Yu Cao, Ciarán Forde, Chong-Wah Ngo, and Tat Seng Chua. Food photo recognition for dietary tracking: System and experiment. In *MultiMedia Modeling: 24th International Conference, MMM 2018, Bangkok, Thailand, February 5-7, 2018, Proceedings, Part II 24*, pages 129–141. Springer, 2018.

Yu Nesterov. Lexicographic differentiation of nonsmooth functions. *Mathematical Programming*, 104(2):669–700, 2005.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer Science & Business Media, 2003.

Yan Pan and Yuanzhi Li. Toward understanding why Adam converges faster than sgd for transformers. *arXiv preprint arXiv:2306.00204*, 2023.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. Understanding the exploding gradient problem. *CoRR, abs/1211.5063*, 2(417):1, 2012.

Jiang Qian, Yuren Wu, Bojin Zhuang, Shaojun Wang, and Jing Xiao. Understanding gradient clipping in incremental gradient methods. In *International Conference on Artificial Intelligence and Statistics*, pages 1504–1512. PMLR, 2021.

Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of Adam and beyond. *In 6th International Conference on Learning Representations (ICLR)*, 2018.

Amirhossein Reisizadeh, Haochuan Li, Subhro Das, and Ali Jadbabaie. Variance-reduced clipping for non-convex optimization. *arXiv preprint arXiv:2303.00883*, 2023.

Herbert Robbins and David Siegmund. A convergence theorem for nonnegative almost supermartingales and some applications. In *Optimizing Methods in Statistics*, pages 233–257. Elsevier, 1971.

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

Alexander Shapiro and Huifu Xu. Uniform laws of large numbers for set-valued mappings and subdifferentials of random functions. *Journal of Mathematical Analysis and Applications*, 325(2):1390–1399, 2007.

Naichen Shi, Dawei Li, Mingyi Hong, and Ruoyu Sun. Rmsprop converges with proper hyperparameter. In *International Conference on Learning Representation*, 2021.

Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, pages 5827–5837. PMLR, 2019.

Umut Simsekli, Lingjiong Zhu, Yee Whye Teh, and Mert Gurbuzbalaban. Fractional underdamped langevin dynamics: Retargeting SGD with momentum under heavy-tailed gradient noise. In *International Conference on Machine Learning*, pages 8970–8980. PMLR, 2020.

Lou Van den Dries and Chris Miller. Geometric categories and o-minimal structures. *Duke Mathematical Journal*, 84(2):497–540, 1996.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

Yijun Wan, Abdellatif Zaidi, and Umut Simsekli. Implicit compressibility of overparametrized neural networks trained with heavy-tailed SGD. *arXiv preprint arXiv:2306.08125*, 2023.

Bohan Wang, Yushun Zhang, Huishuai Zhang, Qi Meng, Zhi-Ming Ma, Tie-Yan Liu, and Wei Chen. Provable adaptivity in Adam. *arXiv preprint arXiv:2208.09900*, 2022.

Jun Wang, Chengfeng Zhou, Songchang Chen, Jianwu Hu, Minghui Wu, Xudong Jiang, Chenming Xu, and Dahong Qian. Chromosome detection in metaphase cell images using morphological priors. *IEEE Journal of Biomedical and Health Informatics*, 2023.

Alex J Wilkie. Model completeness results for expansions of the ordered field of real numbers by restricted pfaffian functions and the exponential function. *Journal of the American Mathematical Society*, 9(4):1051–1094, 1996.

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

Manzil Zaheer, Sashank Reddi, Devendra Sachan, Satyen Kale, and Sanjiv Kumar. Adaptive methods for nonconvex optimization. *Advances in Neural Information Processing Systems*, 31, 2018.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. *In 8th International Conference on Learning Representations (ICLR)*, 2020a.

Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? *Advances in Neural Information Processing Systems*, 33:15383–15393, 2020b.

Yushun Zhang, Congliang Chen, Naichen Shi, Ruoyu Sun, and Zhi-Quan Luo. Adam can converge without any modification on update rules. In *Advances in Neural Information Processing Systems*, 2022.

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.

Fangyu Zou, Li Shen, Zequn Jie, Weizhong Zhang, and Wei Liu. A sufficient condition for convergences of Adam and RMSProp. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11127–11135, 2019.