On the Convergence of AdaGrad-Norm for Non-Convex Optimization

Anonymous authors
Paper under double-blind review

Abstract

Adaptive optimizers have achieved significant success in deep learning by dynamically adjusting the learning rate based on iterative gradients. Compared to stochastic gradient descent (SGD), adaptive optimizers converge much faster in various deep-learning tasks. However, as a fundamental adaptive optimizer, the theoretical analysis of AdaGrad-Norm is inadequate, and there are many technical challenges regarding last-iterate convergence and average-iterate convergence rates for general non-convex loss functions. This paper aims to address these limitations and provides a comprehensive analysis of AdaGrad-Norm. We propose novel techniques that avoid the assumption of no saddle points and derive last-iterate convergence for both almost surely and mean-square senses. Furthermore, under milder conditions, we obtain the near-optimal and sub-optimal rates w.r.t averaged iterate in the expected sense and the almost surely sense, respectively. We relax one restrictive assumption of the uniformly bounded stochastic gradient used in existing high-probability convergence analysis. Moreover, the methodologies provided in this paper have the potential to contribute to further research on the convergence properties of other stochastic algorithms.

1 Introduction

Adaptive gradient methods (Duchi et al., 2011; Kingma & Ba, 2015) have achieved tremendous success in many fields of machine learning. It is observed that adaptive optimizers can achieve better performance than vanilla stochastic gradient descent (SGD) (Vaswani et al., 2017; Duchi et al., 2013; Lacroix et al., 2018; Dosovitskiy et al., 2021) in nonconvex optimization, thus become popular in deep learning. The intuitive explanation of its superiority compared to SGD is that the adaptive optimizers automatically adjust the learning rate based on past stochastic gradients.

AdaGrad (Duchi et al., 2011; McMahan & Streeter, 2010), as a fundamental adaptive learning rate algorithm, has attracted significant research attention in recent years. The norm version of AdaGrad (i.e., AdaGrad-Norm) as a single stepsize adaptation method is described as follows:

$$S_n = S_{n-1} + \left\|\nabla g(\theta_n, \xi_n)\right\|^2, \quad \theta_{n+1} = \theta_n - \frac{\alpha_0}{\sqrt{S_n}}\nabla g(\theta_n, \xi_n), \tag{1}$$

where $g(\theta)$ ($\theta \in \mathbb{R}^d$) is the loss function, $S_0 \geq 0$ is a pre-determined constant, α_0 is a positive constant, and $\nabla g(\theta, \xi_n)$ denotes an unbiased estimate of $\nabla g(\theta)$, i.e., $\mathbb{E}_{\xi_n}[\nabla g(\theta, \xi_n) \mid \mathcal{F}_n] = \nabla g(\theta)$ and the sequence $\{\xi_n\}$ is a sequence of independent random variables. We define a σ-filtration $\mathcal{F}_n := \sigma\{\theta_1, \xi_1, \xi_2, ..., \xi_{n-1}\}$. Despite its simple structure, the convergence results of AdaGrad-Norm on non-convex optimization are sparse and far from satisfactory, especially in the last-iterate sense and the average-iterate sense.

Jin et al. (2022) established almost surely convergence of AdaGrad-Norm in the sense of last-iterate, but heavily relied on the unrealistic assumption that the loss function does not have saddle points. This assumption does not hold in most deep learning applications, for example for neural networks with hidden layers. As a result, the analysis provided in Jin et al. (2022) is not applicable to general nonconvex loss functions with saddle points. It is crucial to explore alternative approaches for analyzing the convergence of AdaGrad-Norm in more general scenarios.

In terms of average-iterate convergence rate, most existing theoretical results for AdaGrad-Norm are based on strong assumptions (Ward et al., 2020; Défossez et al., 2020). For example, Ward et al. (2020); Défossez et al. (2020) assumed the uniform upper bound for all stochastic gradients. To the best of our knowledge, this assumption is often violated when stochastic gradients contain Gaussian random noise. Furthermore, even the conventional mini-batch stochastic gradient fails to satisfy this assumption when the loss function is quadratic (Wang et al., 2023). Recently, Faw et al. (2022); Wang et al. (2023) have removed the uniform boundedness assumption of stochastic gradients, however, they only achieve the convergence rates in the high-probability sense.

The goal of this paper is to address the limitations of existing results and provide a comprehensive analysis of the convergence properties of AdaGrad-Norm for general non-convex loss functions.

Technological Challenges. Despite the inherent simplicity structure of AdaGrad-Norm, investigating its convergence and convergence rate under general conditions poses a significant challenge. In this regard, we will highlight several major obstacles, of which only the first one has been effectively tackled in previous research.

(1) Learning rate $\alpha_0/\sqrt{S_n}$ and stochastic gradient $\nabla g(\theta_n, \xi_n)$ in AdaGrad-Norm are conditionally dependent on the σ-filtration $\mathcal{F}_n$. i.e., we cannot replace $\mathbb{E}\left(\frac{\nabla g(\theta_n)^\top \nabla g(\theta_n,\xi_n)}{\sqrt{S_n}}\middle|\mathcal{F}_n\right)$ with $\frac{\|\nabla g(\theta_n)\|^2}{\sqrt{S_n}}$. This challenge has been effectively resolved in (Jin et al., 2022; Faw et al., 2022; Wang et al., 2023). Faw et al. (2022) addressed this issue by scaling down $1/\sqrt{S_n}$ to $1/\sqrt{S_{n-1} + \|\nabla g(\theta_n)\|^2}$. In Jin et al. (2022); Wang et al. (2023), authors transformed $1/\sqrt{S_n}$ into $1/\sqrt{S_{n-1}} + 1/\sqrt{S_{n-1}} - 1/\sqrt{S_n}$ to obtain a new recurrence relation, where the conditional dependence issue no longer exists. The technique employed in Jin et al. (2022) to solve this issue is also utilized in the proof of this paper.

(2) The quadratic error term $\|\nabla g(\theta_n, \xi_n)\|^2/S_n$ generated by AdaGrad-Norm does not exhibit additivity, i.e., $\sum_{n=1}^{+\infty} \|\nabla g(\theta_n, \xi_n)\|^2/S_n = \Theta(\ln S_n) = +\infty$. The traditional proofs for the almost surely convergence at the last iterate, i.e., $\lim_{n\to+\infty} \|\nabla g(\theta_n)\| = 0 \ \ a.s.$, for SGD or SGD with momentum usually requires the summability of the quadratic error term. This is why the classical Robbins-Monro condition (Robbins & Siegmund, 1971; Jin et al., 2022; Lei et al., 2005; Li & Milzarek, 2022), i.e., $\sum_{n=1}^{+\infty} \epsilon_n = +\infty$, $\sum_{n=1}^{+\infty} \epsilon_n^2 < +\infty$, where $\{\epsilon_n\}_{n=1}^{+\infty}$ is the step size of SGD, arises. Under the Robbins-Monro condition and incorporating the boundedness of $\mathbb{E}(\|\nabla g(\theta_n, \xi_n)\|^2|\mathcal{F}_n)$, it is straightforward to establish the summability of this quadratic error term. However, this is not the same for AdaGrad-Norm. Jin et al. (2022) addressed this issue but relied on the assumption of the absence of saddle points in the loss function. Their approach cannot be applied to loss functions that do have saddle points. A detailed explanation of this issue is provided in the proof sketch of Theorem 3.1 in Section 4.

(3) Demonstrating the convergence in mean square of AdaGrad-Norm with respect to the last iterate, denoted as, $\lim_{n\to+\infty} \mathbb{E}\,\|\nabla g(\theta_n)\|^2 = 0$, encounters several challenges. Typically, for traditional SGD, it is easier to prove the last-iterate mean square convergence than the last-iterate almost surely convergence. By taking the mathematical expectation on both sides of the iteration equation associated with the loss function formed by SGD, we obtain

$$\mathbb{E}(g(\theta_{n+1})) \le \mathbb{E}(g(\theta_n)) - \epsilon_n \,\mathbb{E}\,\|\nabla g(\theta_n)\|^2 + \frac{\epsilon_n^2}{2}\,\mathbb{E}\,\|\nabla g(\theta_n, \xi_n)\|^2,$$

where ϵ_n is the learning rate of SGD. Regarding $\mathbb{E}\,\|\nabla g(\theta_n)\|^2$ as a unified quantity, we convert the original stochastic dynamical system into a deterministic dynamical system. Through further analysis, it is straightforward to derive the mean square convergence result in terms of the last iterate. However, this methodology is not applicable to AdaGrad-Norm. Since the learning rate of AdaGrad-Norm is a random variable, it is not allowed to move the step size $\alpha_0/\sqrt{S_n}$ outside the expectation when computing the mathematical expectation. While we may use operations such as *Hölder's Inequality*, it will introduce two inevitable issues. First, *employing Hölder's Inequality* introduces a change in order, which shifts our object from the second moment of the gradient norm to an alternative quantity. Second, after using *Hölder's Inequality*, our learning rate term incorporates the mathematical expectation, resulting in $\alpha_0/\,\mathbb{E}\,\sqrt{S_n}$, which makes the traditional SGD methods no longer inapplicable. Consequently, the only option is to first prove the last-iterate almost surely convergence and then establish the last-iterate mean-square convergence through *The Lebesgue's Dominated Convergence Theorem*. In other words, we need to prove that the expectation of the uniform upper bound of the gradient norm sequence, i.e., $\mathbb{E}(\sup_{n\ge1}\{\|\nabla g(\theta_n)\|^2\})$, is bounded, which is a challenging task. Due to this inherent difficulty, so far there has been no result regarding the mean square convergence of the last iterate of AdaGrad-Norm.

Contribution. In this paper, we overcome the aforementioned challenges, propose novel techniques, and derive convergence results for both last-iterate and average-iterate under mild conditions. Specifically, we make the following contributions:

(1) For the general nonconvex problems, we propose an innovative analytical perspective and demonstrate the almost surely convergence for AdaGrad-Norm at the last iterate. Since our approach does not examine the iterative characteristics of the dynamical system when the gradients are small, the summability of the squared learning rates is not necessary in our case. As a result, we can overcome the second challenge. Furthermore, our analysis does not depend on the no saddle point assumption required in (Jin et al., 2022), which is a significant improvement.

(2) We are the first to demonstrate the last-iterate mean-square convergence of AdaGrad-Norm under mild conditions. We propose a novel approach by splitting the dynamical system of AdaGrad-Norm into multiple sub-processes via many first entrance times. In this way, the expectation of the maximum value is proved to be finite and addresses the third challenge. Our analytical methodology has the potential applications to other algorithms.

(3) Based on the first two convergence results, we obtain a more exact estimate of S_n, i.e., $\mathbb{E}\, S_n = O(n)$. Utilizing this estimate and the martingale difference estimation lemma (Lemma A.4), we prove both the almost sure and expectation convergence rates for AdaGrad-Norm at the average iterate without assuming the uniform boundedness on stochastic gradients. Our results fill the gap in existing research (Ward et al., 2020; Faw et al., 2022; Wang et al., 2023) which only achieved a high probability convergence rate.

1.1 RELATED WORKS

Both Duchi et al. (2011) and McMahan & Streeter (2010) independently proposed AdaGrad for non-convex optimization. Since then, a series of studies have emerged, analyzing the convergence of AdaGrad-Norm on non-convex landscapes (Ward et al., 2020; Li & Orabona, 2019; Zou et al., 2018; Li & Orabona, 2020; Gadat & Gavra, 2020; Défossez et al., 2020; Kavis et al., 2022; Liu et al., 2022; Faw et al., 2022).

2 PROBLEM SETUP AND ASSUMPTIONS

Throughout the paper, we focus on the following nonconvex problem

$$\min_{\theta} g(\theta) \tag{2}$$

where $g : \mathbb{R}^d \to \mathbb{R}$ is a non-negative and continuously differentiable function and satisfies the following assumptions.

Assumption 2.1. *Loss function $g(\theta)$ satisfies the following conditions:*

(1) The loss function $g(\theta)$ is bounded for any θ that belongs to the gradient sub-level $J_\eta := \{\theta \mid \|\nabla g(\theta)\|^2 < \eta\}$ with some $\eta > 0$, i.e., $g(\theta) < +\infty,\ \forall\, \theta \in J_\eta$.

(2) The gradient $\nabla g(\theta)$ is $\mathcal{L}$-Lipschitz continuous, i.e., for any $x, y \in \mathbb{R}^d$,

$$\big\|\nabla g(x) - \nabla g(y)\big\| \le \mathcal{L}\|x - y\|.$$

(3) For two fixed constants $\sigma_0, \sigma_1 \ge 0$, the stochastic gradient $\nabla g(\theta, \xi_n)$ satisfies that

$$\mathbb{E}_{\xi_n}\Big(\big\|\nabla g(\theta, \xi_n)\big\|^2\Big) \le \sigma_0\big\|\nabla g(\theta)\big\|^2 + \sigma_1 \tag{3}$$

for any $\theta \in \mathbb{R}^d$.

Assumption 2.1 is standard in the non-convex analysis and optimization, which can be found in many previous studies (Faw et al., 2022; Wang et al., 2023; Mertikopoulos et al., 2020). Regarding the first condition in Assumption 2.1 [1], we aim to exclude the existence of a stationary point at the point with infinite function value (i.e., $f(x) = \ln(x)\ (x \to +\infty)$), which has been used in the literature (Mertikopoulos et al., 2020). Note that removing this assumption does not bring any substantial difficulties in the proofs of the theorems, but needs extra discussions on the stationary points at which the function value is infinite. However, the stationary points with the infinite function value are quite special and different from the infinitely distant stationary points with finite function values in logistic regression, i.e., $f(x) = e^{-x}\ (x \to +\infty)$. Without this assumption, proving the statement would become extremely lengthy. Moreover, such functions have rarely appeared in machine learning. Therefore, we make the assumption to simplify the proof.

Jin et al. (2022) makes the assumption on the set of the stationary point to prove the last-iterate almost surely convergence of AdaGrad-Norm. However, such an assumption is not realistic. For loss functions with saddle points, unless the function is defined in one-dimensional space $\mathbb{R}$, we can always find an example that does not satisfy the assumption in Jin et al. (2022). In machine learning, except for problems such as linear regression or logistic regression, the existence of saddle points is very common in many applications. Our assumptions enable us to encompass almost any loss function in machine learning. Besides, Assumption 2.1(3) is commonly used in the analysis of SGD. Our Assumption 2.1(3) is much weaker than the uniformly bounded stochastic gradients (i.e., $\|\nabla g(\theta_n, \xi_n)\| < K < +\infty\ a.s.$.) required in Ward et al. (2020).

3 Theoretical Results

In this section, we present the main convergence results of the AdaGrad-Norm algorithm for smooth nonconvex problems. The proof sketch of each result will be provided in Section 4.

The first result below demonstrates the almost surely convergence of AdaGrad-Norm at the last iterate, which is very challenging in the theoretical analysis of gradient-based methods.

Theorem 3.1. *Consider the AdaGrad-Norm algorithm defined in Equation (1), if Assumption 2.1 holds, then for any initial point $\theta_1 \in \mathbb{R}^d$ and $S_0 \geq 0$, we have*

$$\lim_{n\to\infty} \|\nabla g(\theta_n)\| = 0\ \ a.s..$$

The description of the last iterate convergence provides a more accurate comprehension of the convergence properties of AdaGrad-Norm. This is because, in practice, we typically use the last iterate as the output, rather than the average iterate which is commonly studied in theoretical research. To the best of our knowledge, the literature on the convergence of AdaGrad-Norm in the almost sure sense is sparse. Moreover, we do not assume the absence of saddle points in the loss function, which makes a substantial improvement, compared to the results in Jin et al. (2022). The result of Theorem 3.1 is applicable to almost any loss function encountered in the machine learning regime.

The next theorem describes the convergence of the last iterate of the AdaGrad-Norm algorithm in the mean-square sense.

Theorem 3.2. *Consider the AdaGrad-Norm algorithm shown in Equation (1), if Assumption 2.1 holds, then for any initial point $\theta_1 \in \mathbb{R}^d$ and $S_0 \geq 0$, we have*

$$\lim_{n\to\infty} \mathbb{E}\,\|\nabla g(\theta_n)\|^2 = 0.$$

Theorem 3.2 provides the mean-square convergence of the last iterate of AdaGrad-Norm, which is a novel result, unveiling the uniform convergence of gradient norm convergence under the L_2 norm [2]. We are the first to use the approach by splitting the dynamical system of AdaGrad-Norm into multiple sub-processes through many first entrance times. The proof sketch of the method is provided in Section 4. This approach facilitates a deeper comprehension of the properties of AdaGrad-Norm and can also be applied to the study of other algorithms. We would like to clarify that the almost sure convergence does not imply mean-square convergence. To illustrate this, we

[1] Note that Assumption 2.1 (1) only concerns near-stationary points, not regions where $\|\nabla g(\theta)\|$ may be large. Meanwhile, compared to the assumption in Mertikopoulos et al. (2020), our assumption is weaker as we allow the existence of a stationary point at infinity with a finite loss function value.

[2] The L_2 norm of a random variable ζ is defined as $\sqrt{\mathbb{E}\,\|\zeta\|^2}$.

consider a sequence of random variables $\{\zeta_n\}_{n=1}^{+\infty}$, where $\mathbb{P}(\zeta_n = 0) = 1 - 1/n^2$ and $\mathbb{P}(\zeta_n = n^2) = 1/n^2$. According to *The B-C Lemma*, we can easily show that $\lim_{n\to+\infty} \zeta_n = 0$ almost surely. However, by calculating, we can see that $\mathbb{E}(\zeta_n) = 1$ for all $n > 0$.

As a direct byproduct of Theorem 3.2, we can obtain the following more accurate estimation for S_n:

Property 3.1. *Consider the AdaGrad-Norm algorithm in Equation (1), if Assumption 2.1 holds, then for any initial point $\theta_1 \in \mathbb{R}^d$ and $S_0 \geq 0$, we get that*

$$\mathbb{E}\, S_n = O(n),$$

where other hidden constant in $\mathcal{O}(n)$ is uniquely determined by α_0, c, $g(\theta_1)$, $\nabla g(\theta_1)$, and S_0.

Proof. Through Assumption 2.1 (3) and Theorem 3.2, we can clearly obtain the result. □

Faw et al. (2022); Wang et al. (2023) only obtained the estimation for S_n: $\mathbb{E}\sqrt{S_n} = O(\sqrt{n})$ to achieve the high probability result. However, this estimate is not enough to achieve convergence rates in the expectation sense, which is more difficult than in the high probability sense. In Property 3.1, we derive a more accurate estimation $\mathbb{E}\, S_n = O(n)$, rather than $\mathbb{E}\sqrt{S_n} = O(\sqrt{n})$ and achieve a result in the expectation sense in Theorem 3.4.

Furthermore, we present the almost surely convergence rates for the AdaGrad-Norm algorithm. It is worth noting that the convergence rates provided in this paper are based on the average-iterate sense, rather than the last-iterate convergence rates due to the milder conditions. To obtain convergence rates in the last-iterate sense, one usually needs more conditions to measure the relationship between loss function g and its gradient $\|\nabla g\|$, such as Polyak-Łojasiewicz (PL) condition or Kurdyka-Łojasiewicz (KL) condition. Since this paper focuses on the convergence properties of general non-convex functions, we do not provide the convergence rates of the last iterate here. The first convergence rate result is provided in the almost-surely sense.

Theorem 3.3. *Consider the AdaGrad-Norm algorithm in Equation (1), if Assumption 2.1 holds, then for any initial point $\theta_1 \in \mathbb{R}^d$ and $S_0 \geq 0$, we have*

$$\frac{1}{T}\sum_{k=2}^{T} \big\|\nabla g(\theta_k)\big\|^2 = \mathcal{O}\left(\frac{\ln^{\frac{3}{2}+\sigma} T}{\sqrt{T}}\right) \ \ (\forall\, \sigma > 0) \ \ a.s..$$

Theorem 3.3 presents the near-optimal convergence rate $\mathcal{O}\left(\frac{\ln^{\frac{3}{2}+\sigma} T}{\sqrt{T}}\right)$ in the almost-surely sense for AdaGrad-Norm. We are the first to demonstrate that AdaGrad-Norm converges in a near-optimal rate with probability one, while Faw et al. (2022); Wang et al. (2023) solely provide the high probability results. Moreover, unlike in Ward et al. (2020), we do not impose the restrictive requirement that $\|\nabla g(\theta_n, \xi_n)\|$ is uniformly bounded almost surely.

Theorem 3.4. *Consider the AdaGrad-Norm algorithm in Equation (1), if Assumption 2.1 holds, for any initial point $\theta_1 \in \mathbb{R}^d$, $S_0 \geq 0$, then we have*

$$\frac{1}{T}\sum_{n=1}^{\top} \mathbb{E}\,\|\nabla g(\theta_n)\|^2 = O\left(\frac{\ln^{\frac{2}{p}} T}{T^{\frac{1}{p}}}\right), \quad \forall\, p > 2.$$

Theorem 3.4 shows the convergence rate $O\left(\ln^{\frac{2}{p}} T/T^{\frac{1}{p}}\right)$ $(\forall p > 2)$ for AdaGrad-Norm in the expectation sense. Note that the result of Theorem 3.4 in the expectation sense is different but not weaker than the almost surely result in Theorem 3.3 and high-probability results in Faw et al. (2022); Wang et al. (2023). The distinctions between Theorem 3.3 and 3.4 arise because the hidden constant in $O(\cdot)$ is regarded as a random variable. This hidden constant is almost surely bounded, but its distribution is unknown, so its expectation is not necessarily bounded especially when p approaches 2. On the other hand, in Faw et al. (2022); Wang et al. (2023), the authors may also obtain the expected result $\left(\mathbb{E}\sqrt{\sum_{n=1}^{\top} \|\nabla g(\theta_n)\|^2/T}\right)^2 = O(\ln T/\sqrt{T})$. However, this result does not induce the result w.r.t. $\sum_{n=1}^{\top} \mathbb{E}\,\|\nabla g(\theta_n)\|^2/T$ of Theorem 3.4. Ward et al. (2020) has achieved a near-optimal rate but is based on the uniformly bounded stochastic gradients assumption which is much stronger than ours, and this assumption will facilitate the proof. We will provide a detailed explanation of this situation in Appendix D.3.

4 Proof Sketch

In this section, we will describe the proof sketch of Theorems 3.1 and 3.2, summarize the limitations in previous approaches, and clarify the innovativeness of our methods.

4.1 Proof sketch of Theorem 3.1

To demonstrate the almost surely convergence of AdaGrad-Norm (in Theorem 3.1), the main obstacle is to prove that the iterates sequence $\{\theta_n\}_{n=1}^{+\infty}$ will eventually fall within the vicinity of a connected component of the stationary point set J. We then proceed to narrow down the scope of this region to show that θ_n will ultimately converge to the stationary point set J almost surely.

Step 1: We establish a recursive inequality relationship of the loss functions g in adjacent iterative steps θ_i, θ_{i+1}, i.e.,

$$\begin{aligned} g(\theta_{i+1}) - g(\theta_i) \leq & -\frac{\alpha_0\|\nabla g(\theta_i)\|^2}{\sqrt{S_{i-1}}} + \alpha_0 \frac{\|\nabla g(\theta_i)\| \cdot \mathbb{E}\left(\|\nabla g(\theta_i,\xi_i)\|^2 \big| \mathcal{F}_i\right)}{S_{i-1}} + \frac{c\alpha_0^2}{2} \frac{\mathbb{E}\left(\|\nabla g(\theta_i,\xi_i)\|^2 \big| \mathcal{F}_i\right)}{S_i} \\ & + P_i + Q_i + R_i, \end{aligned} \tag{4}$$

where

$$P_i := \alpha_0 \frac{\nabla g(\theta_i)^\top (\nabla g(\theta_i) - \nabla g(\theta_i,\xi_i))}{\sqrt{S_{i-1}}}, \quad Q_i := \alpha_0 \frac{\|\nabla g(\theta_i)\| \cdot (\|\nabla g(\theta_i,\xi_i)\|^2) - \mathbb{E}(\|\nabla g(\theta_i,\xi_i)\|^2|\mathcal{F}_i)}{S_{i-1}}$$

$$R_i := \frac{\mathcal{L}\alpha_0^2}{2} \frac{\|\nabla g(\theta_i,\xi_i)\|^2 - \mathbb{E}\left(\|\nabla g(\theta_i,\xi_i)\|^2 \big| \mathcal{F}_i\right)}{S_i}.$$

It is observed that when the gradient $\|\nabla g(\theta_n)\|$ is relatively large, i.e., $\forall\ u > 0,\ \|\nabla g(\theta_n)\|^2 > u$, the negative term $-\alpha_0\|\nabla g(\theta_i)\|^2/\sqrt{S_{i-1}}$ will dominate the right side of the inequality (4) as the iterations proceeds. The subsequent terms related to the square of the learning rate can be ignored. However, the martingale difference terms $P_i,\ Q_i,\ R_i$ may be positive and affect the negative term. Therefore, next step we aim to prove that the martingale difference term tends to zero.

Step 2: In order to prove the convergence of the martingale difference when $\|\nabla g(\theta_n)\|^2 > u$, i.e., $\sum_{i=1}^{+\infty} \mathbf{1}_{\|\nabla g(\theta_i)\|^2 > u}(P_i + Q_i + R_i)$ converges almost surely. We present the useful lemma below:

Lemma 4.1. *Suppose $\{\theta_n\}$ is a sequence generated by AdaGrad-norm, and Assumptions 2.1 holds. Then for given $S_0 \geq 0$ and for any $\forall n \in \mathbb{N}_+$, $\theta_1 \in \mathbb{R}^d$, and $\epsilon \in (0, \frac{1}{2})$, we have*

$$\sum_{k=3}^{n} \mathbb{E}\left(\frac{\left\|\nabla g(\theta_k)\right\|^2}{S_{k-1}^{\frac{1}{2}+\epsilon}}\right) < +\infty.$$

This lemma was first proved in Jin et al. (2022) with the *no saddle point* condition. However, as we checked, this lemma does not really need this condition. For clarity, the complete proof is provided in Appendix B.1. Based on Lemma 4.1 and the convergence criterion for martingale difference (in Lemma A.2), we can conclude that $\sum_{i=1}^{+\infty} \mathbf{1}_{\|\nabla g(\theta_i)\|^2 > u}(P_i + Q_i + R_i)$ converges almost surely.

In Steps 1-2, when the gradient norm $\|\nabla g(\theta)\|^2$ is larger than any positive number u, the function value of g on each trajectory of AdaGrad-Norm eventually shows a decreasing trend. We expect that the decrease of function value will bring the iterate θ_n back to the region $\|\nabla g(\theta)\|^2 < u$. Then due to the arbitrariness of u, we can claim the convergence of $\|\nabla g\|$. However, in nonconvex optimization, the main challenge is that the decrease of the loss function does not guarantee a corresponding decrease in its gradient. Our approach stands out from that of Jin et al. (2022) since this step. We will explain the inapplicability of Jin et al. (2022) to loss functions with saddle points.

Literature Review: based on loss function with no saddle points. The main result in Jin et al. (2022) is to prove that the gradient sequence $\{\nabla g(\theta_n)\}_{n=1}^{+\infty}$ crosses a given interval (e, o) in a finite number of times. To achieve this, the authors need to demonstrate the difference between $\|\nabla g(\theta_{n+1})\|^2$ and $\|\nabla g(\theta_n)\|^2$ becoming sufficiently small as the iterations progress. Jin et al. (2022) estimated $\|\nabla g(\theta_{n+1})\|^2 - \|\nabla g(\theta_n)\|^2$ through $g(\theta_{n+1}) - g(\theta_n)$ with an additional condition and then applied *Equation* (4). This condition supposes that when θ approaches J_i with sufficient proximity, the inequality $\|\nabla g(\theta)\|^2 \leq 2\mathcal{L}|g(\theta) - g_i|$ holds for a connected component J_i of a stationary point set J and $g_i := g(\theta)_{\theta \in J_i}$. However, this inequality does not hold near saddle points. For any neighborhood around a saddle point, we can always find a point with the same function value as the saddle point, resulting in the right-hand side of inequality being zero while the left-hand side is positive. Therefore, this method can not handle the presence of saddle points in the loss function. Next, we will introduce our method, which can resolve saddle points.

Step 3: The goal of this step is to show that the image set of the stationary points set $g(J) := \{g(\theta) \mid \theta \in J\}$ can be contained within at most a finite number of disjoint open intervals. Moreover, there exists a lower bound for the distance between any two open intervals, and the measure of each interval can be arbitrarily small. The main idea of the proof is as follows. First, we prove that the stationary points set J can only be divided into countably many connected components $\{J_i\}_{i=1}^{+\infty}$. Then, since each point on each connected component has the same value of the loss function, the set $g(J)$ has at most countably many elements. Next, we construct a sequence of disjoint open intervals $\mathcal{Y}_{x,\delta} := \bigcup_{n=1}^{+\infty} \big((x + (n-1)\delta, x + n\delta) \cup (x - (n-1)\delta, x - n\delta)\big)$, and show that there exists an x such that each open interval in $\mathcal{Y}_{x,\delta}$ does not intersect with set $g(J)$ (the proof is given in Appendix 3.1). By considering the continuity of ∇g and g, we conclude that the elements of set $g(J)$ are not dense in any open intervals. This implies that within each open interval in $\mathcal{Y}_{x,\delta}$, there exists at least one open interval $\mathcal{H}_{x,\delta,n}$ that does not contain any value from set $g(J)$. Furthermore, since $g(J)$ is a bounded set, the measure of all these open intervals $\mathcal{H}_{x,\delta,n}$ must have a minimum value δ_1. Because the value of δ can be arbitrarily small, we have achieved the goal of this step.

Next, we establish the result by demonstrating that $\{g(\theta_n)\}$ will eventually fall into one of the aforementioned open intervals.

Step 4: For any $\delta_0 > 0$. We first construct a subsequence $\{k_n\}_{n=1}^{+\infty}$ to record the boundary points between the sets $\|\nabla g(\theta_n)\|^2 \leq \delta_0/2$ and $\|\nabla g(\theta_n)\|^2 > \delta_0/2$ (the definition is provided in Appendix C). By the definition, the gradient of g at $\theta_{k_{2n-1}}$ and $\theta_{k_{2n}}$ must be greater than $\delta_0/2$. We then apply the inequality in *Equation* (4) to obtain

$$g(\theta_{k_{2n}}) \leq g(\theta_{k_{2n-1}+1}) - \sum_{i=k_{2n-1}+1}^{k_{2n}-1} \frac{\alpha_0\|\nabla g(\theta_j)\|^2}{\sqrt{S_{j-1}}} + \sum_{i=k_{2n-1}+1}^{k_{2n}-1} \frac{\alpha_0\|\nabla g(\theta_j)\| \cdot \mathbb{E}\left(\|\nabla g(\theta_j, \xi_j)\|^2 \middle| \mathcal{F}_j\right)}{S_{j-1}}$$

$$+ \sum_{i=k_{2n-1}+1}^{k_{2n}-1} \frac{c\alpha_0^2}{2} \frac{\mathbb{E}\left(\|\nabla g(\theta_j, \xi_j)\|^2 \middle| \mathcal{F}_i\right)}{S_j} + \sum_{i=k_{2n-1}+1}^{k_{2n}-1} \big(P_j + \mathbf{1}(\|\nabla g(\theta_i)\|^2 \geq \delta_0/2) \cdot (Q_j + R_j)\big).$$

In Step 2, the martingale difference sequences are proven to converge. According to *The Cauchy's Convergence Test*, they can be arbitrarily smaller than any given value. Since the distance between the two open intervals in Step 3 is at least δ, as the iteration progresses, $g(\theta_{k_{2n}})$ cannot be greater than $g(\theta_{k_{2n-1}}) + \delta$. We organize the open intervals in Step 3 in descending order based on the function values they encompass. As the iteration progresses, the index of the open interval where $g(\theta_{k_n})$ is located will definitely increase. According to the monotone convergence theorem, we can prove that $g(\theta_{k_n})$ will definitely fall within one of the open intervals. Then, using *Equation* (4) again, the gradient of points between $\theta_{k_{2n-1}}$ and $\theta_{k_{2n}}$ can be bounded by $O(\delta)$. Combining this result and the fact that the gradient of points between $\theta_{k_{2n-2}}$ and $\theta_{k_{2n-1}}$ are bounded by $\delta_0/2$, we can obtain $\limsup_{n\to+\infty} \|\nabla g(\theta_n)\|^2 < O(\delta_0, \delta)$. Based on the arbitrariness of δ_0 and δ, we prove the result.

4.2 Proof sketch of Theorem 3.2

Following the almost surely convergence of Theorem 3.1, we further can demonstrate the mean-square convergence of AdaGrad-Norm. According to *Lebesgue's Dominated Convergence Theorem*, to obtain mean-square convergence, we only need to find a h^* such that $\|\nabla g(\theta_n)\|^2 \leq h^*$ and $\mathbb{E}\left(|h^*|\right) < +\infty$. Since $\|\nabla g(\theta_n)\|^2 \leq \sup_{k\geq 1} \|\nabla g(\theta_k)\|^2$ always holds, our objective is to prove $\mathbb{E}\left(\sup_{k\geq 1} \|\nabla g(\theta_k)\|^2\right) < +\infty$. In this paper, we are the first to utilize the decomposition of the dynamic system formed by AdaGrad-Norm to prove $\mathbb{E}\left(\sup_{k\geq 1} \|\nabla g(\theta_k)\|^2\right) < +\infty$. In each decomposed sub-process, this dynamic system exhibits a form similar to that of the upper martingale.

Then, in each sub-process, we derive a local maximum by a derivation approach of *Doob's inequality*. Finally, we sum up all the local maxima of the sub-processes to obtain the global maximum.

Step 1: We construct a recursive expression with respect to $g^2(\theta_n)$, rather than $g(\theta)$ used in the proof of Theorem 3.1. This is creative and necessary, and further explanation is provided in **Step** 4. First, we obtain the following lemma for $g^2(\theta_n)$:

Lemma 4.2. *Suppose $\{\theta_n\}$ is a sequence generated by AdaGrad-Norm, and Assumption 2.1 holds. Then $\forall\ n \in \mathbb{N}_+, \forall\ \theta_1 \in \mathbb{R}^d, \forall\ u > 0$ as long as $\|\nabla g(\theta_n)\|^2 > u$, the following inequality holds*

$$\begin{aligned} g^2(\theta_{n+1}) - g^2(\theta_n) \leq{} & \alpha_0(M+1)\left(\frac{g(\theta_{n-1})\big\|\nabla g(\theta_{n-1})\big\|^2}{\sqrt{S_{n-1}}} - \frac{g(\theta_n)\big\|\nabla g(\theta_n)\big\|^2}{\sqrt{S_n}}\right) \\ & + \alpha_0(M+1)\frac{\|\nabla g(\theta_n)\|^3 \cdot \|\nabla g(\theta_n,\xi_n)\|}{S_{n-1}} + \mathcal{L}\alpha_0^2(M+1)\frac{\|\nabla g(\theta_n)\|^2 \cdot \|\nabla g(\theta_n,\xi_n)\|}{S_n} \\ & - \frac{\alpha_0 g(\theta_n)\|\nabla g(\theta_n)\|^2}{\sqrt{S_n}} + \left(2\Big(M+\frac{1}{2}\Big)^2\alpha_0^3\mathcal{L}^2 + \Big(M+\frac{1}{2}\Big)\mathcal{L}^2\alpha_0^3\right)\frac{g(\theta_n)\|\nabla g(\theta_{n-1},\xi_{n-1})\|^2}{S_{n-1}^{\frac{3}{2}}} \\ & + \left(4\|\nabla g(\theta_n)\|^2 + 4\mathcal{L}\alpha_0 + 2\mathcal{L}\alpha_0^2 g(\theta_n)\right)\frac{\big\|\nabla g(\theta_n,\xi_n)\big\|^2}{S_n} + X_n, \end{aligned} \tag{5}$$

where $X_n := \frac{2\alpha_0 g(\theta_n)}{\sqrt{S_{n-1}}}\nabla g(\theta_n)^\top\big(\nabla g(\theta_n) - \nabla g(\theta_n,\xi_n)\big)$ and $M := 2\sigma_0 + 2(\sigma_1/u) - 1$.

Lemma 4.2 holds if $\|\nabla g(\theta_n)\|^2$ exceeds a given constant u. However, this condition may not be fulfilled in every iteration. Therefore, we need to consider the process in segments based on whether the gradient norm satisfies the condition $\|\nabla g(\theta)\|^2 > u$. Based on this, we construct a series of first entrance times in the next step.

Step 2: In this step, we create a sequence of stopping times. By Assumption 2.1(1) (let $u := \eta$): for any θ in $\{\theta \mid \|\nabla g(\theta)\|^2 < u\}$, we have $g(\theta)$ is bounded. This implies that there exists u_0 such that $\|\nabla g(\theta)\|^2 > u$ for all θ in $\{\theta \mid \hat{g}(\theta) > u_0\}$, where $\hat{g}_n := g^2(\theta_n) + \alpha_0(M+1)g(\theta_n)\|\nabla g(\theta_{n-1})\|^2/\sqrt{S_{n-1}}$. For any $\lambda > 0$, the aim of our proof is to calculate the probability $\mathbb{P}\left(\max_{1\leq k\leq n}\hat{g}(\theta_k) > \lambda\right)$ in order to obtain the probability $\mathbb{P}\left(\max_{1\leq k\leq n}\|\nabla g(\theta_k)\|^4 > \lambda\right)$. However, when the gradient norm is small, we do not have a recursive iteration formula similar to Lemma 4.2. However, the boundedness is automatically satisfied when the gradient norm is small. Thus, we need to decompose this process according to the following stopping time. We define events $\mathcal{C}_n := \{\|\nabla g(\theta_n)\|^2 > u\} \cap \{u_0 < \hat{g}(\theta_n) < \lambda\}$ and build a series of stopping times $\{\tau_i^{(\lambda)}\}_{i=1}^{+\infty}$ as follow:

$$\begin{aligned} &\tau_1^{(\lambda)} := \min\{k:\ k \geq 1,\ \mathcal{C}_k \text{ occurs}\},\ \ \tau_2^{(\lambda)} := \min\{k : k > \tau_1^{(\lambda)},\ \mathcal{C}_k \text{ does not occur}\}, ..., \\ &\tau_{2m-1}^{(\lambda)} := \min\{k : k > \tau_{2m-2}^{(\lambda)},\ \mathcal{C}_k \text{ occurs}\},\ \ \tau_{2m}^{(\lambda)} := \min\{k : k \geq \tau_{2m-1}^{(\lambda)},\ \mathcal{C}_k \text{ does not occur}\}. \end{aligned}$$

Then we can define another stopping times $\tau := \min\{k :\ \hat{g}(\theta_1) < \lambda, \hat{g}(\theta_2) < \lambda, ...,\ \hat{g}(\theta_k) < \lambda\}$. Between the stopping times $\tau_{2n-2}\wedge\tau$ and $\tau_{2n-1}\wedge\tau$, it is clear that $\|\nabla g(\theta_n)\|^2 < u$, and before time $\tau_{2n-1}\wedge\tau$, $\hat{g}(\theta_n)$ never exceeds u_0. Hence, we only need to calculate $\mathbb{P}(\max_{\tau_{2n-1}<k<\tau_{2n}}\hat{g}(\theta_k) > \lambda)$ and then sum up the above probabilities over n. In the next step, we will focus on calculating $\mathbb{P}(\max_{\tau_{2n-1}<k<\tau_{2n}}\hat{g}(\theta_k) > \lambda)$.

Step 3: To make it easier to understand, let's first ignore the left stopping time at the beginning of this step of the proof. We define events $\mathcal{B}_{i,k}$ and $\mathcal{B}'_{i,k}$ as follows

$$\mathcal{B}_{i,k} := \begin{cases} \{C_i \text{ does not occur}, \mathcal{C}_{i+1} \text{ occurs}, ...,\ \mathcal{C}_k \text{ occurs}\} & \text{for } k \geq i+1, \\ \{C_i \text{ does not occur}\} & \text{for } k \leq i \end{cases} \quad \text{and} \quad \mathcal{B}'_{i,k} := \mathcal{B}_{i,k-1}/\mathcal{B}_{i,k}.$$

Then for any events $\mathcal{X} \in \mathcal{F}_i$, we get that:

$$\begin{aligned} &\mathbb{E}\left(\mathbf{1}_{\mathcal{X}\cap\mathcal{B}'_{i,m+1}}\hat{g}_{m+1}\right) \leq -\,\mathbb{E}\left(\mathbf{1}_{\mathcal{X}\cap\mathcal{B}_{i,m+1}}\hat{g}_{m+1} - \mathbf{1}_{\mathcal{X}\cap\mathcal{B}_{i,m}}\hat{g}_m\right) \\ &+\mathbb{E}\left(\mathbf{1}_{S_{m-1}<4\beta_0^2/\alpha_0^2}\left(\frac{q_0\|\nabla g(\theta_m,\xi_m)\|^{e_0}}{S_m^{r_0}} + \frac{q_1\|\nabla g(\theta_{m-1},\xi_{m-1})\|^{e_1}}{S_{m-1}^{r_1}}\right)\right), \end{aligned}$$

where $e_0 \geq 2,\ e_1 \geq 2,\ r_0 > 0,\ r_1 > 0,\ q_0,\ q_1$, and β_0 are constants. The proof of the above inequality is quite complicated, and due to limited space, it cannot be fully explained here (the specific proof is given in Appendix D from (77) to (84)). We define $\tau^{(0)} := \min\{k : g(\theta_k) \geq \lambda\},\ \ \tau_m^{(0)} := \min\{k : g(\theta_k) \geq \lambda,\ \ k \geq \tau_{2m-1}^{(\lambda)} \wedge \tau\}$. Then, we let $\mathcal{X} = \{\tau \wedge \tau_{2m-1}^{(\lambda)} \wedge n = i\}$ and sum it up to obtain the result considering the left stopping time. We getting

$$\begin{aligned}\mathbb{E}(\hat{g}_{\tau_m^{(0)} \wedge n}) < &\sum_{i=\tau_{2n-2}^{(\lambda)}}^{n-1} \sum_{m=\tau\wedge\tau_{2m-2}^{(\lambda)}\wedge n}^{n-1} \mathbb{E}\left(\mathbf{1}_{\{\tau\wedge\tau_{2m-1}^{(\lambda)}\wedge n=i\}\cap\mathcal{B}'_{i,m+1}}\hat{g}_{m+1}\right)\\ &\leq u_0\big(\,\mathbb{E}(\mathbf{1}_{\tau\wedge\tau_{2m-2}^{(\lambda)}\wedge n}) - \mathbb{E}(\mathbf{1}_{\tau\wedge\tau_{2m}^{(\lambda)}\wedge n})\big)\\ &+ \beta_0 \sum_{m=\tau\wedge\tau_{2m-2}^{(\lambda)}\wedge n}^{\tau\wedge\tau_{2m}^{(\lambda)}\wedge n} \mathbb{E}\left(\mathbf{1}_{S_{m-1}<4\beta_0^2/\alpha_0^2}\left(\frac{q_0\|\nabla g(\theta_m,\xi_m)\|^{e_0}}{S_m^{r_0}} + \frac{q_1\|\nabla g(\theta_{m-1},\xi_{m-1})\|^{e_1}}{S_{m-1}^{r_1}}\right)\right),\end{aligned} \tag{6}$$

where $e_0 \geq 2,\ e_1 \geq 2,\ r_0 > 0,\ r_1 > 0,\ q_0,\ q_1$ and $\delta_0 > 0$ are seven constants. Next, we will generalize the results from time $\tau_{2m-2} \wedge \tau \wedge n$ to time $\tau_{2m} \wedge \tau \wedge n$ to all times.

Step 4: Define $\overline{\|\nabla g(\theta_n)\|^2} := \sup_{1\leq k\leq n} \|\nabla g(\theta_n)\|^2$ and $\overline{g}_n := \sup_{1\leq k\leq n} g^2(\theta_k)$, for any $\lambda \geq u$, we have $\{\overline{\|\nabla g(\theta_n)\|^4} > \lambda\} \subset \{\overline{g}_n > 2c\lambda\} \cap \{\overline{\|\nabla g(\theta_n)\|^2} > u\} \subset \Big\{\sup_{1\leq k\leq n} g_k^2 > 2c\lambda\Big\}$. Using *The Markov's Inequality* and *Equation* (6) gives $\mathbb{P}\left(\overline{\|\nabla g(\theta_n)\|^4} > \lambda\right) \leq \frac{1}{2\mathcal{L}\lambda}\sum_{m=1}^{+\infty}\mathbb{E}(\hat{g}_{\tau_m^{(0)}\wedge n}) \leq \frac{K}{2c\lambda} \leq \frac{T}{\lambda}$, where $T > 0$ is a finite positive constant. The proof of this inequality is provided in Appendix D *Equation* 86. Then we estimate $\mathbb{E}\left(\overline{\|\nabla g(\theta_n)\|^2}\right)$ and achieve that

$$\begin{aligned}&\mathbb{E}\left(\overline{\|\nabla g(\theta_n)\|^2}\right) = \mathbb{E}\left(\sqrt{\overline{\|\nabla g(\theta_n)\|^4}}\right) = u + \int_{\lambda=u}^{+\infty} \lambda^{\frac{1}{2}-1}\,\mathbb{P}\left(\overline{\|\nabla g(\theta_m)\|^4} > \lambda\right)d\lambda\\ &\leq u + T\int_{\lambda=u}^{+\infty} \lambda^{-\frac{3}{2}}d\lambda = u + \frac{2T}{\sqrt{u}} < +\infty.\end{aligned}$$

Now we are able to address the question raised in **Step** 1 why we use $g^2(\theta)$ rather than $g(\theta)$. If $g(\theta)$ is studied in Step 1, we obtain $\mathbb{P}\left(\overline{\|\nabla g(\theta_n)\|^2} > \lambda\right) \leq O(1/\lambda)$ in Step 3, and further achieve that $\mathbb{E}\left(\overline{\|\nabla g(\theta_n)\|^2}\right) < u + \int_{\lambda-u}^{+\infty}\lambda^{-1}d\lambda = +\infty$. Thus, this is not enough to guarantee bounded of $\mathbb{E}\left(\overline{\|\nabla g(\theta_n)\|^2}\right)$. So far we have proven that $\mathbb{E}\left(\sup_{k\geq 1}\|\nabla g(\theta_k)\|^2\right) < +\infty$. By Theorem 3.1 and *The Lebesgue's Dominated Convergence Theorem*, we have proven $\mathbb{E}\left(\|\nabla g(\theta_k)\|^2\right) \to 0$.

4.3 PROOF SKETCH OF THEOREM 3.3 AND THEOREM 3.4

The proofs of Theorems 3.3 and 3.4 are relatively straightforward compared to those of Theorem 3.1 and Theorem 3.2. For brevity, we omit the proof sketch here, and the complete proofs can be found in Appendix D.1 and Appendix D.2.

5 CONCLUSION

In this paper, we effectively address several limitations of the theoretical analysis of AdaGrad-Norm. Specifically, we propose novel techniques that avoid the no saddle points assumption used in (Jin et al., 2022) and establish the last-iterate convergence in both almost surely and mean-square senses. Additionally, we demonstrate the near-optimal and sub-optimal convergence rates concerning the averaged iterate in the expectation sense and the almost surely sense, respectively. Moreover, we mitigate the uniform boundedness assumption on stochastic gradients, commonly used in existing high-probability convergence analysis. Furthermore, our approaches pave the way for exploring the convergence properties of other stochastic algorithms in future research.

References

Alexandre Défossez, Léon Bottou, Francis Bach, and Nicolas Usunier. A simple convergence proof of adam and adagrad. *arXiv preprint arXiv:2003.02395*, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

John Duchi, Michael I Jordan, and Brendan McMahan. Estimation, optimization, and parallelism when data is sparse. *Advances in Neural Information Processing Systems*, 26:2832–2840, 2013.

Alexandre Défossez, Léon Bottou, Francis Bach, and Nicolas Usunier. A simple convergence proof of adam and adagrad. 2020.

Matthew Faw, Isidoros Tziotis, Constantine Caramanis, Aryan Mokhtari, Sanjay Shakkottai, and Rachel Ward. The power of adaptivity in sgd: Self-tuning step sizes with unbounded gradients and affine variance. In *Conference on Learning Theory*, pp. 313–355. PMLR, 2022.

Sébastien Gadat and Ioana Gavra. Asymptotic study of stochastic adaptive algorithm in non-convex landscape. 2020.

Ruinan Jin, Yu Xing, and Xingkang He. On the convergence of mSGD and AdaGrad for stochastic optimization. In *International Conference on Learning Representations*, 2022.

Ali Kavis, Kfir Yehuda Levy, and Volkan Cevher. High probability bounds for a class of nonconvex algorithms with adagrad stepsize. *arXiv preprint arXiv:2204.02833*, 2022.

Diederik P Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Timothée Lacroix, Nicolas Usunier, and Guillaume Obozinski. Canonical tensor decomposition for knowledge base completion. In *International Conference on Machine Learning*, pp. 2863–2872, 2018.

Guo Lei, Cheng Dai-Zhan, and Feng De-Xing. *Introduction to Control Theory: From Basic Concepts to Research Frontiers*. Beijing: Science Press, 2005.

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

Xiaoyu Li and Francesco Orabona. On the convergence of stochastic gradient descent with adaptive stepsizes. In *The 22nd international conference on artificial intelligence and statistics*, pp. 983–992. PMLR, 2019.

Xiaoyu Li and Francesco Orabona. A high probability analysis of adaptive sgd with momentum. *arXiv preprint arXiv:2007.14294*, 2020.

Zijian Liu, Ta Duy Nguyen, Alina Ene, and Huy Nguyen. On the convergence of adagrad (norm) on r^d: Beyond convexity, non-asymptotic rate and acceleration. In *The Eleventh International Conference on Learning Representations*, 2022.

H Brendan McMahan and Matthew Streeter. Adaptive bound optimization for online convex optimization. *arXiv preprint arXiv:1002.4908*, 2010.

Panayotis Mertikopoulos, Nadav Hallak, Ali Kavis, and Volkan Cevher. On the almost sure convergence of stochastic gradient descent in non-convex problems. *Advances in Neural Information Processing Systems*, 33:1117–1128, 2020.

Herbert Robbins and David Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing methods in statistics*, pp. 233–257. Elsevier, 1971.

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.

Bohan Wang, Huishuai Zhang, Zhiming Ma, and Wei Chen. Convergence of adagrad for non-convex objectives: Simple proofs and relaxed assumptions. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 161–190. PMLR, 2023.

Rachel Ward, Xiaoxia Wu, and Leon Bottou. Adagrad stepsizes: Sharp convergence over nonconvex landscapes. *The Journal of Machine Learning Research*, 21(1):9047–9076, 2020.

Fangyu Zou, Li Shen, Zequn Jie, Ju Sun, and Wei Liu. Weighted adagrad with unified momentum. *arXiv preprint arXiv:1808.03408*, 2018.