A COMPREHENSIVE FRAMEWORK FOR ANALYZING THE CONVERGENCE OF ADAM: BRIDGING THE GAP WITH SGD

Anonymous authors
Paper under double-blind review

ABSTRACT

Adaptive Moment Estimation (Adam) is a cornerstone optimization algorithm in deep learning, widely recognized for its flexibility with adaptive learning rates and efficiency in handling large-scale data. However, despite its practical success, the theoretical understanding of Adam's convergence has been constrained by stringent assumptions, such as almost surely bounded stochastic gradients or uniformly bounded gradients, which are more restrictive than those typically required for analyzing stochastic gradient descent (SGD).

In this paper, we introduce a novel and comprehensive framework for analyzing the convergence properties of Adam. This framework offers a versatile approach to establishing Adam's convergence. Specifically, we prove that Adam achieves asymptotic (last iterate sense) convergence in both the almost sure sense and the L_1 sense under the relaxed assumptions typically used for SGD, namely L-smoothness and the ABC inequality. Meanwhile, under the same assumptions, we show that Adam attains non-asymptotic sample complexity bounds similar to those of SGD.

1 INTRODUCTION

Adaptive Moment Estimation (Adam) is one of the most widely used optimization algorithms in deep learning due to its adaptive learning rate properties and efficiency in handling large-scale data (Kingma & Ba, 2014). Despite its widespread use, the theoretical understanding of Adam's convergence is not as advanced as its practical success. Previous studies have often imposed stringent assumptions on the loss function and stochastic gradients, such as uniformly bounded loss functions and almost surely bounded gradients (Reddi et al., 2018b; Zou & Shen, 2019), which are more restrictive than those required for analyzing classical stochastic gradient descent (SGD).

In this paper, we introduce a novel and comprehensive framework for analyzing the convergence properties of Adam. Our framework unifies various aspects of convergence analysis, including non-asymptotic (average iterate sense) sample complexity, asymptotic (last iterate sense) almost sure convergence, and asymptotic L_1 convergence. Crucially, we demonstrate that under this framework, Adam can achieve convergence under the same assumptions typically used for SGD—namely, the L-smooth condition and the ABC inequality (L_2 sense) (Khaled & Richtárik, 2023; Bottou, 2010; Ghadimi & Lan, 2013).

Several recent works have attempted to relax the stringent conditions required for Adam's convergence, each focusing on different aspects of the stochastic gradient assumptions and convergence guarantees. However, limitations still exist in terms of assumptions and the types of convergence results obtained. Table 1 provides the references and a summary of the works and compares the assumptions on stochastic gradients, the resulting complexities, and the convergence properties achieved.

Our approach builds upon these prior works and seeks to offer a more comprehensive and general framework for analysis. In contrast to these previous works, we study Adam under the ABC inequality, which is more general and less restrictive compared to the assumptions made in the previous studies. Our analysis successfully establishes non-asymptotic sample complexity and achieves

Table 1: Comparison of Assumptions and Convergence Results. (♢) The smoothing term μ is often set to small values like 10^{-8} in practice. It is difficult and relevant to avoid the $\mathcal{O}(\text{poly}(\frac{1}{\mu}))$ dependence (Wang et al., 2024a), which our analysis achieves. (♠) The work focuses on learning rates and hyperparameters dependent on the total number of iterations T, leading to results without a $\mathcal{O}(\ln T)$ term. As our asymptotic analysis uses T-independent parameters, terms regarding $\mathcal{O}(\ln T)$ inevitably appear, though our method can be easily extended to T-dependent settings. (♥) These works have weakened the classical L-smooth condition, which is different from the focus of this paper.

Reference	Assumptions on Stochastic Gradient	Sample Complexity	A.S. Convergence	L_1 Convergence
Wang et al. (2024a)♠	Bounded Variance (or Coordinate Weak Growth)	$\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$	No	No
He et al. (2023)♢	Almost Surely Bounded Stochastic Gradient	$\mathcal{O}\left(\text{poly}\left(\frac{1}{\mu}\right)\cdot\frac{\ln T}{\sqrt{T}}\right)$	Yes	Yes
Zou et al. (2019)	L_2 Bounded Stochastic Gradient	$\mathcal{O}\left(\frac{\ln T}{\sqrt{T}}\right)$	No	No
Zhang et al. (2022)	Randomly Reshuffled Stochastic Gradient	$\mathcal{O}\left(\frac{\ln T}{\sqrt{T}}\right)$	No	No
Li et al. (2024)♥♢	Almost Surely Bounded Stochastic Gradient or Sub-Gaussian Variance	$\mathcal{O}\left(\text{poly}\left(\frac{1}{\mu}\right)\cdot\frac{\ln T}{\sqrt{T}}\right)$	No	No
Wang et al. (2024b)♥	Randomly Reshuffled Stochastic Gradient	$\mathcal{O}\left(\frac{\ln T}{\sqrt{T}}\right)$	No	No
Xiao et al. (2024)♥	Almost Surely Bounded Stochastic Gradient	No Result	Yes	No
Our Work	ABC Inequality	$\mathcal{O}\left(\frac{\ln T}{\sqrt{T}}\right)$	Yes	Yes

asymptotic almost sure convergence and L_1 convergence under conditions that align with those required for SGD. This makes our framework theoretically sound and versatile for analyzing multiple convergence properties of Adam. Our framework might also be of independent interest in analyzing different variants of Adam. In summary, our work presents a novel and general theoretical framework for Adam, unifying various convergence properties. This framework demonstrates that Adam's convergence guarantees can be aligned with those of SGD, which justifies the applicability of Adam across a wide range of machine learning problems.

1.1 Related Works

In recent years, the convergence properties of Adam have been extensively studied, with various works focusing on different assumptions about stochastic gradients and the types of convergence guarantees provided. In the following discussion, we categorize and review key contributions based on the different types of stochastic gradient assumptions they employ, as summarized in Table 1.

Bounded Variance and Coordinate Weak Growth: Wang et al. (2024a) considered Adam's convergence under the assumption of bounded variance or coordinate weak growth. The coordinate weak growth condition (Eq. 2) is particularly stringent as it requires that each component of the stochastic gradient satisfies a weak growth inequality, which is stronger than the traditional weak growth condition (Eq. 1) applied to the entire gradient. Under these assumptions, Wang et al. were able to avoid the $\mathcal{O}(1/\mu)$ complexity. However, their work did not focus on analyzing almost sure convergence or L_1 convergence, as the primary emphasis was on the sample complexity of the algorithm's behavior.

Almost Surely Bounded Stochastic Gradients: Several works, including He et al. (2023) and Xiao et al. (2024), have explored Adam's convergence under the assumption that the stochastic gradients are almost surely bounded. This is a particularly strong assumption, as it implies several other commonly made assumptions about stochastic gradients, such as bounded variance, weak growth, coordinate weak growth, and sub-Gaussian properties. The assumption is often impractical in non-convex settings where gradients can become unbounded. Moreover, studies in Wang et al. (2023) have highlighted that this assumption is unrealistic in many common machine learning frameworks, failing to hold even for simple quadratic functions, let alone for deep neural networks. While these works achieved almost sure convergence and, in some cases, L_1 convergence, they did not address the complexity related to the $\mathcal{O}(1/\mu)$ term.

L_2 Bounded Stochastic Gradients: Zou et al. (2019) analyzed Adam under the assumption of L_2 bounded stochastic gradients. Although this condition is milder than the almost surely bounded gradients assumption, it is still stronger than the traditional weak growth condition and the ABC

inequality. In the standard analytical framework, this assumption can at best be weakened to the coordinate weak growth condition, which remains more restrictive than the assumptions typically considered for SGD. At the same time, this work focused on complexity analysis without addressing asymptotic convergence.

Randomly Reshuffled Stochastic Gradients: In other works, such as those by Zhang et al. (2022) and Wang et al. (2024b), the authors considered the case where the stochastic gradients are randomly reshuffled. Randomly reshuffled stochastic gradients represent a special case where the gradients are typically assumed to satisfy certain inequalities almost surely. This reliance on almost sure properties forms a much stronger and more restrictive analytical framework compared to those based on traditional weak growth conditions or the ABC inequality. While these works successfully avoided $\mathcal{O}(1/\mu)$ complexity, they did not focus on analyzing the asymptotic convergence property.

2 PRELIMINARIES

In this section, we introduce the necessary preliminaries and establish the foundational framework for our convergence analysis of the Adam. We begin by recalling the Adam optimization algorithm. We then state the assumptions that will be used throughout our analysis. These assumptions are standard in stochastic optimization and are crucial for deriving our main results. By laying out these assumptions explicitly, we also facilitate a clear comparison with the conditions used in previous works, highlighting the less restrictive nature of our approach.

2.1 ADAM

Adam is an extension of SGD that computes adaptive learning rates for each parameter by utilizing estimates of the first and second moments of the gradients. It combines the advantages of two other extensions of SGD: AdaGrad, which works well with sparse gradients, and RMSProp, which works well in online and non-stationary settings.

Algorithm 1 Adam

Input: Stochastic oracle $\mathcal{O}$, initial learning rate $\eta_t > 0$, initial parameters $w_1 \in \mathbb{R}^d$, initial exponential moving averages $m_0 = 0$, $v_0 = v \cdot \mathbf{1}^\top$ with $v > 0$, hyperparameters $\beta_1, \beta_{2,1} \in [0, 1)$, smoothing term $\mu > 0$, number of iterations T
Output: Final parameter w_T

1: **for** $t = 1$ **to** T **do**
2: Generate conditioner parameter $\beta_{2,t}$;
3: Sample a random data point z_t and compute the stochastic gradient $g_t = \mathcal{O}_f(w_t, z_t)$;
4: Update the second moment estimate: $v_t = \beta_{2,t} v_{t-1} + (1 - \beta_{2,t}) g_t^{\circ 2}$;
5: Update the first moment estimate: $m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$;
6: Compute the adaptive learning rate: $\eta_{v_t} = \eta_t \circ \frac{1}{\sqrt{v_t} + \mu}$;
7: Update the parameters: $w_{t+1} = w_t - \eta_{v_t} \circ m_t$;
8: **end for**

In Adam, the random variables $\{z_t\}_{t \geq 1}$ are mutually independent. The stochastic gradient at iteration t is denoted by g_t. The quantities m_t and v_t represent the exponential moving averages of the first and second moments of the gradients, respectively. The hyperparameters β_1 and $\beta_{2,t}$ control the exponential decay rates for the moment estimates. A small smoothing term μ is introduced to prevent division by zero, and η_{v_t} represents the adaptive learning rate for each parameter.

In terms of notation, all vectors are column vectors unless specified otherwise, and $\mathbf{1}^\top$ denotes a row vector with all elements equal to 1. For vectors $\beta, \gamma \in \mathbb{R}^d$, the Hadamard product (element-wise multiplication) is represented by $\beta \circ \gamma$, and the element-wise square root of a vector $\gamma \in \mathbb{R}^d$ is written as $\sqrt{\gamma}$. Operations such as $\beta + v_0$, $\frac{1}{\beta}$, and $\beta^{\circ 2}$ are performed element-wise. Additionally, the i-th component of a vector β_t is denoted as $\beta_{t,i}$.

When analyzing Adam, $\nabla f(w_t)$ refers to the true gradient of the loss function at iteration t. We define $\mathscr{F}_t = \sigma(g_1, \ldots, g_t)$ as the σ-algebra generated by the stochastic gradients up to iteration t, with $\mathscr{F}_0 = \{\Omega, \emptyset\}$ and $\mathscr{F}_\infty = \sigma\left(\bigcup_{t\geq 1} \mathscr{F}_t\right)$.

2.2 Assumptions

To establish our convergence results, we make the following standard assumptions, focusing on the stochastic gradient conditions. These assumptions are less restrictive than those imposed in some prior works, as highlighted in Table 1.

Assumption 2.1. ***(Bounded from Below Loss Function)*** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a loss function defined on $\mathbb{R}^d$. We assume that there exists a constant $f^* \in \mathbb{R}$ such that for all $w \in \mathbb{R}^d$, the following inequality holds: $f(w) \geq f^*$.*

This assumption ensures that the loss function f is bounded from below, preventing it from decreasing indefinitely during the optimization process.

Assumption 2.2. ***(L-Smoothness)*** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable loss function. We assume that the gradient ∇f is Lipschitz continuous. That is, there exists a constant $L_f \geq 0$ such that for all $w, w' \in \mathbb{R}^d$, the following inequality holds: $\|\nabla f(w) - \nabla f(w')\| \leq L_f \|w - w'\|$, where $\|\cdot\|$ denotes the Euclidean norm. The constant L_f is known as the Lipschitz constant of the gradient.*

Assumption 2.3. ***(ABC Inequality)*** *We assume that the stochastic gradient g_t is an unbiased estimate of the true gradient, i.e., $\mathbb{E}[g_t \mid \mathscr{F}_{t-1}] = \nabla f(w_t)$, and there exist constants $A, B, C \geq 0$ such that for all iterations t, we have: $\mathbb{E}[\|g_t\|^2 \mid \mathscr{F}_{t-1}] \leq A(f(w_t) - f^*) + B\|\nabla f(w_t)\|^2 + C$.*

The ABC inequality provides a bound on the second moment of the stochastic gradients, which is crucial for analyzing the convergence of stochastic optimization algorithms.

2.3 Comparison with Prior Works on Stochastic Gradient Assumptions

Our assumption on the stochastic gradient (Assumption 2.3) is relatively mild compared to those in prior works. Here, we focus on comparing with the traditional weak growth condition, coordinate weak growth assumption, and the almost surely bounded stochastic gradient assumption.

Traditional Weak Growth Condition The traditional weak growth condition (e.g., Bottou et al. (2018); Nguyen et al. (2018)) assumes that there exist constants $B \geq 0$ and $C \geq 0$ such that:

$$\mathbb{E}[\|g_t\|^2 \mid \mathscr{F}_{t-1}] \leq B\|\nabla f(w_t)\|^2 + C. \tag{1}$$

This condition bounds the expected squared norm of the stochastic gradient by a linear function of the squared norm of the true gradient plus a constant. It is stronger than our ABC inequality because it does not include the term involving the function value difference $f(w_t) - f^*$.

Even under this condition, current methods for analyzing Adam encounter significant difficulties. We will explain these challenges in the proof sketch of Lemma 4.1.

Coordinate Weak Growth Assumption Wang et al. (2024a) introduce the *coordinate weak growth* assumption, which requires that each component of the stochastic gradient satisfies a weak growth inequality. Specifically, for each coordinate i, there exist constants $B, C \geq 0$ such that:

$$\mathbb{E}[g_{t,i}^2 \mid \mathscr{F}_{t-1}] \leq B\|\nabla_i f(w_t)\|^2 + C, \tag{2}$$

where $g_{t,i}$ and $\nabla_i f(w_t)$ are the i-th components of g_t and $\nabla f(w_t)$, respectively.

This assumption is stronger than the traditional weak growth condition because it imposes the inequality on each coordinate individually, rather than on the overall gradient.

Almost Surely Bounded Stochastic Gradient Assumption Some prior works, such as He et al. (2023); Xiao et al. (2024), assume that the stochastic gradients are almost surely bounded. That is, there exists a constant $M \geq 0$ such that for all iterations t: $\|g_t\| \leq M$ almost surely. This is a strong assumption, as it requires that the stochastic gradient norm is uniformly bounded almost surely

at all iterations. In practice, especially in non-convex optimization problems, this assumption is often violated (see Wang et al. 2023). For instance, when optimizing deep neural networks, gradient norms can become unbounded due to the complexity and non-linearity of the models. Moreover, this assumption implies that the true gradient is also bounded by M, because $\|\nabla f(w_t)\|^2 \leq \mathbb{E}[\|g_t\|^2 \mid \mathscr{F}_{t-1}] \leq M^2$. Our assumption is clearly weaker than the almost surely bounded stochastic gradient assumption, as we only require a bound on the expected squared norm of the stochastic gradient, which can depend on the current function value and gradient norm, rather than a uniform almost sure bound.

Moreover, assuming almost surely bounded stochastic gradients is hard to satisfy in practice and may not reflect realistic scenarios. As discussed in Wang et al. (2023); Khaled & Richtárik (2023), such assumptions can be unrealistic and limit the applicability of theoretical results.

Next, we introduce a property. We know that when the loss function is L-smooth, the true gradient of the loss function can be controlled by the loss function value $f(w_t) - f^*$ (as shown in Lemma B.1). Therefore, we can simplify the ABC inequality as follows.

Property 1. *Under Assumptions 2.2 and 2.3, for all iterations t, we have:*

$$\mathbb{E}[\|g_t\|^2 \mid \mathscr{F}_{t-1}] \leq (A + 2L_f B)(f(w_t) - f^*) + C.$$

This property demonstrates that the variance of the stochastic gradients can be bounded by the function value difference, which is a key component in our convergence analysis.

2.4 Hyperparameter Settings

In this paper, to avoid overly lengthy proofs, we choose a class of representative parameter settings, as follows:

$$\beta_{2,t} := \begin{cases} 1-\alpha_0, & \text{if } t=1 \\ 1-\frac{1}{t^\gamma}, & \text{if } t \geq 2 \end{cases}, \ \beta_1 \in [0,1),\ \eta_t = \frac{1}{t^{\frac{1}{2}+\delta}},\ \left(\alpha_0 \in [0,1),\ \gamma \in [1, 2\delta+1],\ \delta \in \left[0, \frac{1}{2}\right]\right).$$

Imposing restrictions on Adam's parameters, particularly $\beta_{2,t}$, is necessary to ensure convergence. Early studies (Reddi et al., 2018a) have demonstrated that without appropriate constraints on $\beta_{2,t}$, counterexamples exist where the algorithm fails to converge. Moreover, for the gradient norm to converge to zero, it is essential that $\beta_{2,t}$ approaches 1 (Zou et al., 2019; He et al., 2023), as noted in previous works.

Some studies on complexity allow $\beta_{2,t}$ to be constant. However, these studies typically focus on the algorithm's complexity over a finite number of iterations T. In such cases, the constant value of $1 - \beta_{2,t}$ is inversely related to T, effectively causing $\beta_{2,t}$ to approach 1 as T increases. This is another means of ensuring that $\beta_{2,t}$ asymptotically approaches 1, which is crucial for convergence.

The hyperparameter settings adopted in this paper are representative and have been considered in previous studies (Zou et al., 2019; He et al., 2023). Our configuration includes settings that can achieve near-optimal complexity of $\mathcal{O}(\ln T/\sqrt{T})$. The logarithmic factor $\ln T$ arises because $\beta_{2,t}$ is chosen independent of the total number of iterations T, which is an unavoidable consequence with this class of parameters.

Our choice of hyperparameters simplifies the analysis while capturing the essential behavior of the Adam. Although the proof techniques can be extended to a broader range of parameter settings, this paper focuses primarily on the assumptions related to the convergence of the algorithm rather than an exhaustive exploration of hyperparameter configurations.

3 Theoretical Results

In this section, we establish both non-asymptotic and asymptotic convergence guarantees for the Adam within our smooth non-convex framework, as defined by Assumptions 2.1–2.3. For the non-asymptotic analysis, we derive a sample complexity bound that is independent of $\mathcal{O}(1/\mu)$, providing an explicit bound on the number of iterations required to achieve a specified accuracy. In the asymptotic analysis, we consider two forms of convergence: almost sure convergence and convergence in the L_1 norm. The almost sure convergence result demonstrates that, the gradient norm of almost

every trajectory converges to zero. Meanwhile, the L_1 convergence result reveals that the convergence across different trajectories is uniform with respect to the L_1 norm of the gradient, where the L_1 norm is taken in the sense of the underlying random variable, meaning the expectation of the gradient norm.

3.1 Non-Asymptotic Sample Complexity

We first establish a non-asymptotic bound on the sample complexity of Adam.

Theorem 3.1 (**Non-Asymptotic Sample Complexity**). *Consider the Adam algorithm as specified in Algorithm 2.1, and suppose that Assumptions 2.1–2.3 hold. Then, for any initial point and for $T \geq 1$, the following results hold:*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(w_t)\|] \leq \begin{cases} \mathcal{O}\Big(\frac{1}{T^{\frac{1}{2}-\delta}}\Big), & \text{if } \delta \in (0, \frac{1}{2}], \\ \mathcal{O}\Big(\frac{\ln T}{\sqrt{T}}\Big), & \text{if } \gamma > 1,\ \delta = 0, \\ \mathcal{O}\Big(\frac{\ln^2 T}{\sqrt{T}}\Big), & \text{if } \gamma = 1,\ \delta = 0. \end{cases}$$

The constant hidden in the $\mathcal{O}$ notation depends on the initial point, the constants in our required assumptions (excluding $1/\mu$), and the parameters δ and α_0.

This theorem provides a non-asymptotic rate of convergence for the expected gradient norm, highlighting how the choice of hyperparameters affects the convergence rate.

3.2 Asymptotic Convergence

We now present our main asymptotic convergence results, demonstrating that the gradients of the Adam converge to zero both almost surely and in the L_1 sense under appropriate conditions.

Theorem 3.2 (**Asymptotic Almost Sure Convergence**). *Under Assumptions 2.1–2.3, consider the Adam with hyperparameters specified in Subsection 2.4 with $\gamma > 1$ and $\delta > 0$. Then, the gradients of the Adam converge to zero almost surely, i.e., $\lim_{t\to\infty}\|\nabla f(w_t)\| = 0 \quad a.s.$*

This theorem shows that the gradients evaluated at the iterates converge to zero almost surely, indicating that the algorithm approaches a critical point of the loss function along almost every trajectory.

Remark 1. ***(Almost sure vs L_1 convergence)*** *As stated in the introduction, it is important to note that the almost sure convergence does not imply L_1 convergence. To illustrate this concept, let us consider a sequence of random variables $\{\zeta_n\}_{n\geq 1}$, where $\mathbb{P}(\zeta_n = 0) = 1 - 1/n^2$ and $\mathbb{P}(\zeta_n = n^2) = 1/n^2$. According to the Borel-Cantelli lemma, it follows that $\lim_{n\to+\infty}\zeta_n = 0$ almost surely. However, it can be shown that $\mathbb{E}(\zeta_n) = 1$ for all $n > 0$ by simple calculations.*

Theorem 3.3 (**Asymptotic L_1-Convergence**). *Under Assumptions 2.1–2.3, consider the Adam with hyperparameters specified in Subsection 2.4 with $\gamma > 1$ and $\delta > 0$. Then, the gradients of the Adam converge to zero in the L_1 sense, i.e., $\lim_{t\to\infty}\mathbb{E}[\|\nabla f(w_t)\|] = 0$.*

This result establishes convergence in the mean sense, showing that the expected gradient norm approaches zero as the number of iterations increases. It indicates that the convergence of gradient norms across different trajectories is uniform in the L_1 norm of the random variables.

In previous works (He et al. (2023); Xiao et al. (2024)), the assumption that the stochastic gradients are uniformly bounded, i.e., $\|g_t\| \leq M$ a.s. $(\forall\, t \geq 1)$, or that the gradients themselves are uniformly bounded, i.e., $\|\nabla f(w_t)\| \leq M$ $(\forall\ t \geq 1)$, allows almost sure convergence to directly imply L_1 convergence via the *Lebesgue's Dominated Convergence* theorem. However, in our framework, which deals with potentially unbounded stochastic gradients or gradients, proving L_1 convergence is much more challenging. We will elaborate on this in the next section.

4 Framework for Analyzing Adam

In this section, we present the analytical framework that underpins our convergence analysis for the Adam. Our approach is built upon the insights provided by existing methods, while introducing

new techniques to address the limitations of previous analyses and provide a more comprehensive understanding of Adam's behavior under weaker assumptions. Our core innovations are detailed in Section 4.3.1, Section 4.4, and Section 4.5.

4.1 Key Properties of Adaptive Learning Rates

We begin by characterizing the fundamental properties of the adaptive learning rate sequence η_{v_t}. These properties are critical as they directly influence the behavior of the algorithm and are foundational to our subsequent analysis. By understanding how these properties interact with the algorithm's dynamics, we obtain more insights on the conditions under which Adam converges.

Property 2. *Each element $\eta_{v_t,i}$ of the sequence $\{\eta_{v_t}\}_{t\geq 1} = \{[\eta_{v_t,1}, \eta_{v_t,2}, \ldots, \eta_{v_t,d}]^\top\}_{t\geq 1}$ is monotonically decreasing with respect to t.*

This property ensures that the learning rate becomes progressively smaller as the algorithm progresses, which is a crucial factor in the stability and convergence of Adam.

Property 3. *Each element $\eta_{v_t,i}$ of the sequence $\{\eta_{v_t}\}_{t\geq 1} = \{[\eta_{v_t,1}, \eta_{v_t,2}, \ldots, \eta_{v_t,d}]^\top\}_{t\geq 1}$ satisfies the inequality $t^\gamma v_{t,i} \geq \alpha_1 S_{t,i}$, where we define $\alpha_1 := \min\{1-\alpha_0, \alpha_0\}$, $S_{t,i} := v + \sum_{k=1}^{t} g_{k,i}^2$ for all $t \geq 1$, and $S_{0,i} := v$.*

This property highlights the relationship between the accumulated gradient information $S_{t,i}$ and the adaptive learning rate, ensuring that the latter appropriately scales with the former as iterations proceed.

Remark 4.1. *For the purpose of simplifying the proofs of subsequent theorems, we define two auxiliary parameters: $\Sigma_{v_t} := \sum_{i=1}^{d} v_{t,i}$ and $S_t := \sum_{i=1}^{d} S_{t,i}$. Additionally, for convenience in the subsequent proofs, we define a new initial parameter based on $S_{0,i}$ as $\eta_{v_0,i} = S_{0,i}/\alpha_1 = v/\alpha_1$.*

These definitions of auxiliary parameters help streamline the analysis, making the mathematical expressions more manageable and the proofs more concise.

With the key properties of the adaptive learning rates established, we now turn our attention to analyzing the momentum term, which plays a crucial role in the Adam.

4.2 Handling the Momentum Term

To effectively analyze the momentum term in the Adam, we adopt a classical method introduced by Liu et al. (2020). The momentum term introduces additional complexity in the analysis due to its recursive nature, which can complicate the convergence proofs. To address this, we construct an auxiliary variable u_t that simplifies the analysis by decoupling the momentum term from the update process. This auxiliary variable is defined as follows:

$$u_t := \frac{w_t - \beta_1 w_{t-1}}{1-\beta_1} = w_t + \frac{\beta_1}{1-\beta_1}(w_t - w_{t-1}) = w_t - \frac{\beta_1}{1-\beta_1}\eta_{v_{t-1}} \circ m_{t-1}. \tag{3}$$

The introduction of u_t allows us to handle the momentum term more effectively by transforming the recursive nature of the updates into a more tractable form. Specifically, we can express the relationship between successive iterations of u_t as follows:

$$u_{t+1} - u_t = -\eta_{v_t} \circ g_t + \frac{\beta_1}{1-\beta_1}(\underbrace{\eta_{v_{t-1}} - \eta_{v_t}}_{\Delta_t}) \circ m_{t-1}. \tag{4}$$

This recursive relation is instrumental in breaking down the complex dependencies introduced by the momentum term, which will facilitate the convergence analysis.

4.3 Establishing the Approximate Descent Inequality

In the convergence analysis of stochastic gradient descent (SGD), a fundamental tool is the *approximate descent inequality*, which quantifies the expected decrease in the objective function at each iteration. Specifically, for SGD, the approximate descent inequality is given by:

$$f(w_{t+1}) \leq f(w_t) - \frac{\eta_t}{2}\|\nabla f(w_t)\|^2 + \frac{\eta_t^2 L}{2}\mathbb{E}[\|g_t\|^2 \mid \mathscr{F}_t] + \eta_t \nabla f(w_t)^\top(\nabla f(w_t) - g_t), \tag{5}$$

where η_t is the learning rate, L is the Lipschitz constant, and g_t is the stochastic gradient.

Motivated by the success of this approach in analyzing SGD, we aim to establish a similar approximate descent inequality for the Adam. The goal is to develop a descent inequality that captures the adaptive nature of Adam's learning rates while maintaining the essential structure seen in the analysis of SGD.

To this end, we present the following key result, which forms the cornerstone of our convergence analysis for Adam.

Lemma 4.1 (**Approximate Descent Inequality**). *Consider the sequences $\{w_t\}_{t\geq 1}$, $\{v_t\}_{t\geq 1}$, and $\{u_t\}_{t\geq 1}$ generated by Algorithm 2.1 and Eq. 4. Under Assumptions 2.1–2.3, the following sufficient decrease inequality holds:*

$$\Pi_{\Delta,t+1}\hat{f}(u_{t+1}) - \Pi_{\Delta,t}\hat{f}(u_t) \leq -\frac{5}{16}\Pi_{\Delta,t+1}\sum_{i=1}^{d}\zeta_i(t) + (L_f+1)\Pi_{\Delta,t+1}\sum_{i=1}^{d}\eta_{v_t,i}^2 g_{t,i}^2 + \Pi_{\Delta,t+1}M_t. \tag{6}$$

Here,

$$\begin{aligned}
&\hat{f}(u_t) := f(u_t) + C\sum_{i=1}^{d}\eta_{v_{t-1},i} + \frac{C_2}{(1-\beta_1)\sqrt{v}}\|m_{t-1}\|^2, \;\; \zeta_i(t) := \eta_{v_t,i}(\nabla_i f(w_t))^2,\\
&\Pi_{\Delta,t} := \prod_{k=1}^{t-1}(1+C_1\overline{\Delta}_k)^{-1} \quad (\forall\, t\geq 2), \quad \Pi_{\Delta,1} := 1, \;\; \overline{\Delta}_t := \sum_{i=1}^{d}\mathbb{E}(\Delta_{t,i}\mid \mathscr{F}_{t-1}),\\
&M_t := M_{t,1} + M_{t,2} + M_{t,3}.
\end{aligned} \tag{7}$$

Constants C_1 and C_2 are defined in Eq. 21; $M_{t,1}$ is defined in Eq. 15; $M_{t,2}$ and $M_{t,3}$ are defined in Eq. 16.

This lemma introduces $\Pi_{\Delta,t}\hat{f}(u_t)$ as a new Lyapunov function for the Adam, which plays a crucial role in our analysis. In inequality 6, the term $-\frac{5}{16}\Pi_{\Delta,t+1}\sum_{i=1}^{d}\zeta_i(t)$ can be interpreted as the descent term, representing the expected decrease in the Lyapunov function. The second term, $(L_f+1)\Pi_{\Delta,t+1}\sum_{i=1}^{d}\eta_{v_t,i}^2 g_{t,i}^2$, accounts for the squared error due to the stochastic nature of the gradients. The third term on the right-hand side, $\Pi_{\Delta,t+1}M_t$, is a martingale difference sequence with respect to the filtration $\{\mathscr{F}_t\}_{t\geq 1}$, which, due to its zero expectation, can be considered to have no overall impact on the algorithm's iteration process.

This structure closely resembles the approximate descent inequality commonly used in the analysis of SGD. For comparison, the approximate descent inequality for SGD is given by Eq. 5.

We now proceed to provide the main idea of proving Lemma 4.1 and highlight the key steps and challenges involved in establishing this result for Adam.

To begin with, we calculate the difference in the loss function values between two consecutive auxiliary variables $\{u_t\}_{t\geq 1}$ that we introduced. We obtain the following expression (informal):

$$\begin{aligned}
f(u_{t+1}) - f(u_t) \leq &\underbrace{-\sum_{i=1}^{d}\mathbb{E}\left[\eta_{v_t,i}\nabla f(w_t)g_{t,i}|\mathscr{F}_{t-1}\right]}_{\textit{Term}_{t,1}} + \underbrace{\mathcal{O}\left(\sum_{i=1}^{d}\eta_{v_t,i}^2 g_{t,i}^2\right)}_{\textit{Term}_{t,2}}\\
&\underbrace{+\sum_{i=1}^{d}\mathbb{E}\left[\eta_{v_t,i}\nabla f(w_t)g_{t,i}|\mathscr{F}_{t-1}\right] - \sum_{i=1}^{d}\eta_{v_t,i}\nabla f(w_t)g_{t,i}}_{\textit{Term}_{t,3}} + R_t.
\end{aligned} \tag{8}$$

It can be observed that the above equation is simply a second-order Taylor expansion of $f(u_{t+1}) - f(u_t)$ (since an L-smooth function is almost everywhere twice differentiable). $\textit{Term}_{t,1}$ represents the first-order term, which in general serves as the descent term. $\textit{Term}_{t,2}$ is the quadratic error, and $\textit{Term}_{t,3}$ is a martingale difference sequence. The remaining term R_t is negligible and can be ignored.

In the informal explanation provided in the sketch, these were collectively referred to as remainder terms. For the exact formulation, refer to the detailed proof in Appendix D.1.

While handling the quadratic error term $\textit{Term}_{t,2}$ is relatively straightforward using standard scaling techniques, addressing the first-order term $\textit{Term}_{t,1}$ is more challenging due to the adaptive nature of Adam's learning rates. Specifically, $\eta_{v_t,i}$ and $g_{t,i}$ are both $\mathscr{F}_t$-measurable, which necessitates the introduction of an auxiliary random variable $\tilde{\eta}_{v_t,i} \in \mathscr{F}_{t-1}$ to facilitate the extraction of the learning rate from the conditional expectation. In this paper, we choose the auxiliary random variable $\eta_{v_{t-1},i}$ to approximate $\eta_{v_t,i}$. There are also other forms of this approximation, as discussed by Wang et al. (2023; 2024a). This allows us to rewrite the first-order term as:

$$\begin{aligned}
-\textit{Term}_{t,1} &= -\sum_{i=1}^{d} \mathbb{E}\left[\eta_{v_t,i} \nabla f(w_t) g_{t,i} | \mathscr{F}_{t-1}\right] \\
&= \underbrace{-\sum_{i=1}^{d} \mathbb{E}\left[\eta_{v_{t-1},i} \nabla f(w_t) g_{t,i} | \mathscr{F}_{t-1}\right]}_{\textit{Descent-Term}_t} + \underbrace{\sum_{i=1}^{d} \mathbb{E}\left[(\eta_{v_{t-1},i} - \eta_{v_t,i}) \nabla f(w_t) g_{t,i} | \mathscr{F}_{t-1}\right]}_{\textit{Term}_{t,4}}.
\end{aligned}$$

The presence of $\textit{Term}_{t,4}$ introduces an additional layer of complexity in the analysis, as it reflects the difference between successive adaptive learning rates. Addressing this extra error term is crucial for establishing robust convergence guarantees under the ABC inequality or weak growth conditions. Existing approaches to handling such terms, which often rely on the cancellation of errors through preceding descent terms, fall short in this context. This necessitates a more innovative strategy, which we present in the following section.

4.3.1 Addressing the Extra Error Term: Our Innovative Approach

The term $\textit{Term}_{t,4}$, introduced by the difference between $\eta_{v_{t-1},i}$ and $\eta_{v_t,i}$, presents a significant challenge in the convergence analysis of Adam under the ABC inequality or weak growth conditions. In existing methods, it is common to attempt to cancel out such error terms by leveraging the preceding descent term $\textit{Descent-Term}_t$. However, this approach might not work within the ABC framework. Recent works such as Wang et al. (2023; 2024a) have shown that, under existing techniques, the best one can achieve is a weakened form of the stochastic gradient assumption, namely the coordinate weak growth condition.

To overcome these limitations, we introduce a novel approach to handle $\textit{Term}_{t,4}$. We scale it as follows:

$$\begin{aligned}
\textit{Term}_{t,4} \leq & \frac{1}{2} \sum_{i=1}^{d} \mathbb{E}\left[\eta_{v_{t-1},i} \nabla f(w_t) g_{t,i} | \mathscr{F}_{t-1}\right] + C_1 f(u_t) \cdot \sum_{i=1}^{d} \mathbb{E}[\Delta_{t,i} \mid \mathscr{F}_{t-1}] \\
& + C \sum_{i=1}^{d} \Delta_{t,i} + \underbrace{C \sum_{i=1}^{d} \Big(\mathbb{E}[\Delta_{t,i} \mid \mathscr{F}_{t-1}] - \Delta_{t,i} \Big)}_{\textit{Term}_5},
\end{aligned}$$

where $\Delta_{t,i} := \eta_{v_{t-1},i} - \eta_{v_t,i}$, and $C_1 := \frac{A+2L_f B}{2}(L_f + 1)$. The key term in the inequality is $C_1 f(u_t) \sum_{i=1}^{d} \mathbb{E}[\Delta_{t,i} \mid \mathscr{F}_{t-1}]$, which cannot be easily canceled out by existing methods.

To handle this issue, we move the term $C_1 f(u_t) \sum_{i=1}^{d} \mathbb{E}[\Delta_{t,i} \mid \mathscr{F}_{t-1}]$ to the left-hand side of inequality 8 and combine it with the existing $f(u_t)$ term. This leads to a new iteration inequality of the form:

$$f(u_{t+1}) - (1 + C_1 \overline{\Delta}_t) f(u_t) \leq -\frac{1}{2} \textit{Descent-Term}_t + \textit{M-Term}_t + \textit{Term}_{t,2} + \textit{R-Term}_t. \tag{9}$$

In the inequality $\textit{M-Term}_t = \textit{Term}_{t,3} + \textit{Term}_{t,5}$ is a martingale difference sequence and $\textit{R-Term}_t$ is the (neglectable) remainder term by combining all other terms from the inequalities. To express this inequality in a form resembling a Lyapunov function, we introduce an auxiliary product variable:

$$\Pi_{\Delta,t} := \prod_{k=1}^{t-1} (1 + C_1 \overline{\Delta}_k)^{-1} \quad (\forall\, t \geq 2), \quad \Pi_{\Delta,1} := 1.$$

Multiplying both sides of the inequality by $\Pi_{\Delta,t+1}$, we obtain the following reformulated inequality:

$$\Pi_{\Delta,t+1} f(u_{t+1}) - \Pi_{\Delta,t} f(u_t) \leq -\frac{1}{2}\Pi_{\Delta,t+1} \cdot \textit{Descent-Term}_t + \Pi_{\Delta,t+1} \cdot \textit{M-Term}_t + \Pi_{\Delta,t+1} \cdot \textit{Term}_{t,2} + \Pi_{\Delta,t+1} \cdot \textit{R-Term}_t. \tag{10}$$

This reformulation introduces $\Pi_{\Delta,t}$ as a scaling factor, which, along with the original Lyapunov function, captures the impact of $\textit{Term}_{t,4}$. The resulting inequality closely parallels the approximate descent inequality for SGD, with additional terms accounting for Adam's adaptive nature.

The handling of $\textit{Term}_{t,4}$ in our analysis framework is a significant advancement over existing methods. It allows us to establish stronger convergence guarantees under more general conditions.

4.4 Deriving Sample Complexity and Almost Sure Convergence

After establishing the *Approximate Descent Inequality*, the next step is to derive the sample complexity and almost sure convergence results for Adam. The methodology for obtaining these results largely mirrors the approaches traditionally used in the analysis of SGD. Specifically, the inequality provides a foundation for bounding the expected decrease in the loss function, which can then be used to establish both sample complexity and almost sure Convergence.

However, a key difference in our analysis lies in the introduction of the term $\Pi_{\Delta,t+1}$ within the *Approximate Descent Inequality*. This term introduces a new layer of complexity not present in the standard SGD analysis. In particular, we are required to bound the p-th moment of the reciprocal of this term, i.e., $\mathbb{E}[\Pi_{\Delta,t+1}^{-p}],\ (p \geq 1)$. Due to the unique structure of $\Pi_{\Delta,t+1}$, determining a bound for this p-th moment is a non-trivial task.

To address this challenge, we leverage tools from discrete martingale theory, particularly the *Burkholder's* inequality. It allows us to establish a recursive relationship between the p-th moment $\mathbb{E}[\Pi_{\Delta,t+1}^{-p}]$ and the $p/2$-th moment $\mathbb{E}[\Pi_{\Delta,t+1}^{-p/2}]$. This recursive structure is crucial as it enables us to iteratively bound the higher moments of $\Pi_{\Delta,t+1}^{-1}$.

Once the recursive relationship is established, we apply fundamental theorems from measure theory, such as the *Lebesgue's Monotone Convergence* theorem or the *Lebesgue's Dominated Convergence* theorem, to obtain the final bound on the p-th moment.

The detailed process for bounding $\mathbb{E}[\Pi_{\Delta,t+1}^{-p}]$ can be found in Lemma B.2 and Lemma B.3.

4.5 Establishing Asymptotic L_1 Convergence

Since we have already proved almost sure convergence in Theorem 3.2, it is natural to attempt to prove L_1 convergence via the *Lebesgue's Dominated Convergence* theorem. To achieve this, we need to find a function h that is $\mathscr{F}_\infty$-measurable and satisfies $\mathbb{E}|h| < +\infty$, and such that for all $t \geq 1$, we have $\|\nabla f(w_t)\| \leq |h|$. Since for all t we naturally have $\|\nabla f(w_t)\| \leq \sup_{k\geq 1} \|\nabla f(w_k)\|$, we only need to prove that $\mathbb{E}[\sup_{k\geq 1} \|\nabla f(w_k)\|] < +\infty$.

This task presents a significant challenge because, within our analytical framework, we cannot assume that the gradients are uniformly bounded, which means we cannot directly apply the *Lebesgue's Dominated Convergence* theorem. Instead, we need to utilize advanced techniques from discrete martingale theory, specifically the first hitting time decomposition method, to obtain a bound on this maximal expectation. The detailed process can be found in Appendix E.4.

5 Conclusion

We have introduced a novel and comprehensive framework for analyzing the convergence properties of Adam. Our frame starts with weak assumptions such as the ABC inequality. By identifying the key properties of the learning rate, handling the momentum term, and establishing the approximate descent inequality, the frame concludes the sample complexity, almost surely convergence, and asymptotic L_1 convergence results of Adam. Our techniques overcome the limitations of existing analyses, and show that Adam's convergence guarantees can be aligned with those of SGD, which justifies the applicability of Adam across a wide range of machine learning problems.

REFERENCES

Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers*, pp. 177–186. Springer, 2010.

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

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM journal on optimization*, 23(4):2341–2368, 2013.

Meixuan He, Yuqing Liang, Jinlan Liu, and Dongpo Xu. Convergence of adam for non-convex objectives: Relaxed hyperparameters and non-ergodic case. *arXiv preprint arXiv:2307.11782*, 2023.

Ahmed Khaled and Peter Richtárik. Better theory for SGD in the nonconvex world. *Trans. Mach. Learn. Res.*, 2023, 2023.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Haochuan Li, Alexander Rakhlin, and Ali Jadbabaie. Convergence of adam under relaxed assumptions. *Advances in Neural Information Processing Systems*, 36, 2024.

Yanli Liu, Yuan Gao, and Wotao Yin. An improved analysis of stochastic gradient descent with momentum. *Advances in Neural Information Processing Systems*, 33:18261–18271, 2020.

Lam Nguyen, Phuong Ha Nguyen, Marten Dijk, Peter Richtárik, Katya Scheinberg, and Martin Takác. Sgd and hogwild! convergence without the bounded gradients assumption. In *International Conference on Machine Learning*, pp. 3750–3758. PMLR, 2018.

Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018a.

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

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.

Bohan Wang, Jingwen Fu, Huishuai Zhang, Nanning Zheng, and Wei Chen. Closing the gap between the upper bound and lower bound of adam's iteration complexity. *Advances in Neural Information Processing Systems*, 36, 2024a.

Bohan Wang, Yushun Zhang, Huishuai Zhang, Qi Meng, Ruoyu Sun, Zhi-Ming Ma, Tie-Yan Liu, Zhi-Quan Luo, and Wei Chen. Provable adaptivity of adam under non-uniform smoothness. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2960–2969, 2024b.

Nachuan Xiao, Xiaoyin Hu, Xin Liu, and Kim-Chuan Toh. Adam-family methods for nonsmooth optimization with convergence guarantees. *Journal of Machine Learning Research*, 25(48):1–53, 2024.

Yushun Zhang, Congliang Chen, Naichen Shi, Ruoyu Sun, and Zhi-Quan Luo. Adam can converge without any modification on update rules. *Advances in neural information processing systems*, 35:28386–28399, 2022.

Difan Zou and Li Shen. Improved convergence analysis of stochastic optimization algorithms for nonconvex optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

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*, pp. 11127–11135, 2019.

CONTENTS

A ESTABLISHING KEY PROPERTIES BETWEEN w_t AND u_t

Here, we establish two key properties that connect the original variable w_t and the auxiliary variable u_t which defined in Section 4.2. These properties are crucial for bounding the changes in the momentum term and linking the function values at different points in the iteration process.

Property 4. *For any iteration step t, the following inequality holds:*

$$\|m_t\|^2 - \|m_{t-1}\|^2 \leq -(1-\beta_1)\|m_{t-1}\|^2 + (1-\beta_1)\|g_t\|^2.$$

This property establishes a bound on the change in the momentum term, which is critical for ensuring that the momentum does not increase indefinitely during the optimization process. Controlling the momentum in this manner is a key aspect of proving convergence.

Property 5. *For any iteration step t, the following inequality holds:*

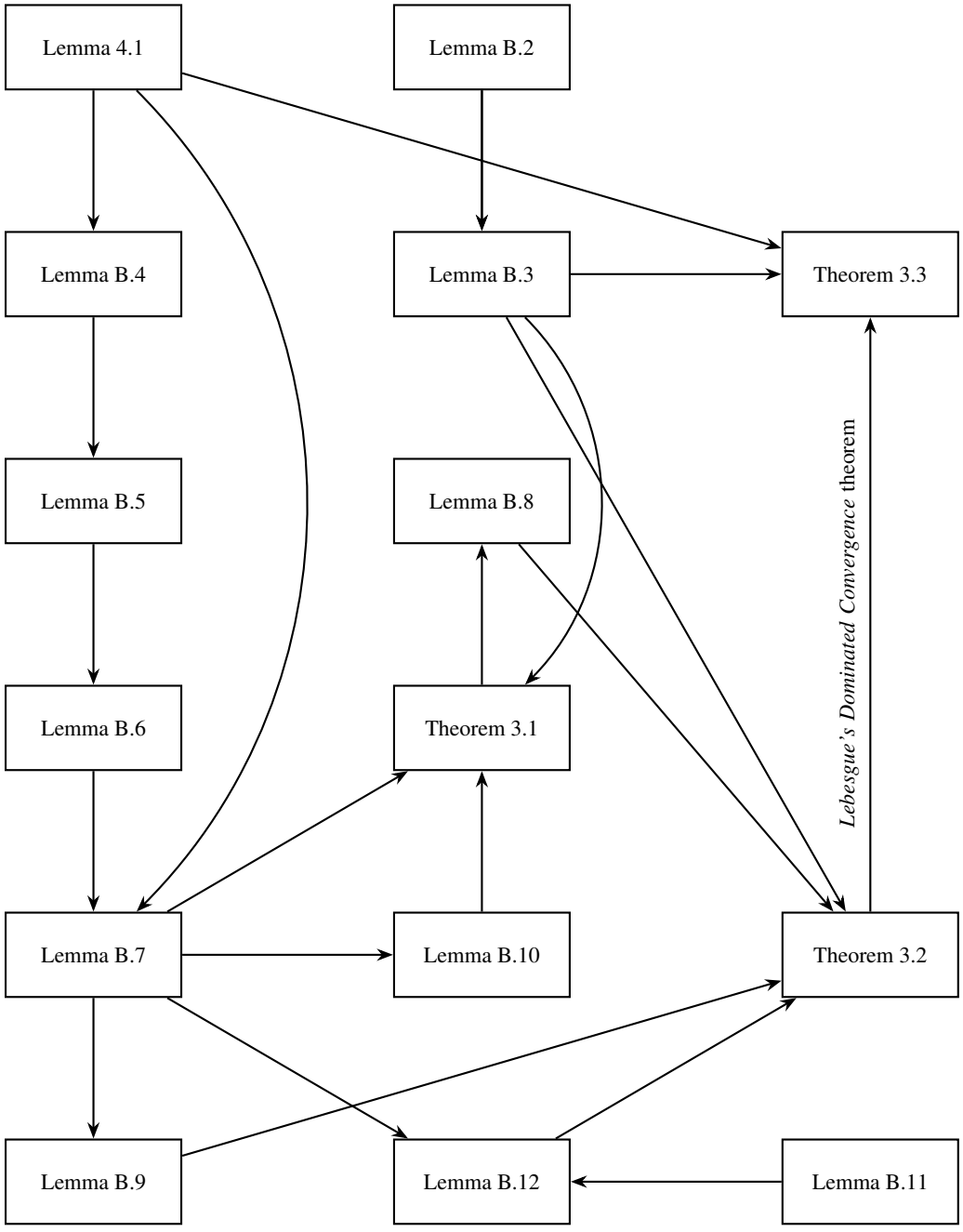
$$f(w_t) \leq (L_f+1)f(u_t) + \frac{(L_f+1)\beta_1^2}{2(1-\beta_1)^2}\left\|\eta_{v_{t-1}} \circ m_{t-1}\right\|^2.$$

This property links the function values at w_t and u_t, providing a foundation for analyzing the convergence of $f(w_t)$. By establishing this relationship, we can relate the behavior of the original variable w_t to the more manageable auxiliary variable u_t, thereby simplifying the overall convergence analysis.

B SUPPORTING LEMMAS

B.1 THEOREM DEPENDENCY GRAPH

In this section, we will supplement several additional supporting lemmas that are crucial to the overall proof. Due to the large number of lemmas, we have combined these lemmas with those in the main text and theorems to create a lemma-theorem dependency graph. Readers can refer to this graph while following the proofs.



B.2 THE BASIC FORM OF ADDITIONAL IMPORTANT LEMMAS

Lemma B.1. *Suppose that $f(x)$ is differentiable and lower bounded $f^* = \inf_{x\in\mathbb{R}^d} f(x) > -\infty$ and $\nabla f(x)$ is Lipschitz continuous with parameter $\mathcal{L} > 0$, then $\forall\, x \in \mathbb{R}^d$, we have*

$$\big\|\nabla f(x)\big\|^2 \le 2\mathcal{L}\big(f(x) - f^*\big).$$

Lemma B.2. *Let $\{(X_n, \mathscr{F}_n)\}_{n\ge1}$ be a non-negative adapted process such that $\sum_{n=1}^{+\infty} X_n = M < +\infty$ almost surely, where M is a finite constant. Define the partial sum of conditional expectations as $\Lambda_T := \sum_{n=1}^{T} \mathbb{E}[X_n \mid \mathscr{F}_{n-1}]$. Then:*

(i) The sequence $\{\Lambda_T\}_{T\ge1}$ converges almost surely, i.e., $\Lambda_T \xrightarrow{a.s.} \Lambda$, where $\Lambda := \sum_{n=1}^{+\infty} \mathbb{E}[X_n \mid \mathscr{F}_{n-1}]$.

(ii) For any $p \ge 1$, the sequence $\{\Lambda_T\}_{T\ge1}$ converges in L_p, i.e., $\lim_{T\to\infty} \mathbb{E}\left[|\Lambda_T - \Lambda|^p\right] = 0$. Meanwhile, the p-th moment of the limit Λ is bounded by a constant $C_\Lambda(p) > 0$, where $C_\Lambda(p) = o(p^{\sqrt{p}})$.

(iii) The arbitrary p-th moment of the random variable e^Λ also exists, and its upper bound depends only on p and M. We denote this upper bound by $C_{e^\Lambda}(p, M)$.

Lemma B.3. *For $\Pi_{\Delta,T+1}$ as defined in Eq. 7, for any $T \ge 0$ and any $p \ge 1$, the p-th moment of its reciprocal is bounded, i.e.,*

$$\mathbb{E}\left[\Pi_{\Delta,T+1}^{-p}\right] < C_{v,d,p} < +\infty,$$

where $C_{v,d,p}$ is a constant that depends only on v, d, and p.

Moreover, we have that $\Pi_{\Delta,\infty}^{-1} := \lim_{t\to+\infty} \Pi_{\Delta,t}^{-1} < +\infty$ a.s., and for any $p \ge 1$, the p-th moment of $\Pi_{\Delta,\infty}^{-1}$ exists, with

$$\mathbb{E}\left[\Pi_{\Delta,\infty}^{-p}\right] \le C_{v,d,p} < +\infty.$$

Lemma B.4. *Consider the Adam in Algorithm 2.1 and suppose that Assumption 2.1∼2.3 hold, then for any initial point, and $T \ge 1$, we have:*

$$\frac{\Pi_{\Delta,t+1}\hat{f}(u_{t+1})}{\sqrt{S_t}} \le \frac{\hat{f}(u_1)}{\sqrt{dv}} + \frac{3(L_f+1)d}{\alpha_1\sqrt{v}} + \sum_{t=1}^{T} \frac{\Pi_{\Delta,t+1}}{\sqrt{S_{t-1}}} M_t. \tag{11}$$

Lemma B.5. *Consider the Adam in Algorithm 2.1 and suppose that Assumption 2.1∼2.3 hold, then for any initial point and $\forall\ \phi > 0$, we have for any $T \ge 1$, the following inequality:*

$$\frac{\Pi_{\Delta,T}\sqrt{S_T}}{(T+1)^\phi} \le \sqrt{dv} + \sum_{t=1}^{T} \Pi_{\Delta,t}\Lambda_{\phi,t}. \tag{12}$$

where

$$\Lambda_{\phi,t} := \frac{\|g_t\|^2}{(t+1)^\phi\sqrt{S_{t-1}}},$$

and S_T is defined in Eq. 4.1.

Lemma B.6. *Consider the Adam as defined in Algorithm 2.1, and suppose that Assumptions 2.1 through 2.3 hold. Then, for any initial point and for all $T \ge 1$, there exists a random variable ζ such that the following results hold:*

(a) $0 \le \zeta < +\infty$ almost surely, and $\mathbb{E}(\zeta)$ is uniformly bounded above by a constant C_ζ, which depends on the initial point and the constants in the required assumptions (excluding $1/\mu$). The explicit form of this upper bound is provided in Eq. 30.

(b) $\sqrt{S_T} \le (T+1)^2\zeta$, and $\ln(S_T) \le \zeta' \ln(T+1)$, where $\zeta' := 8\ln\left(\max\left\{e, \frac{\sqrt{2}\Pi_{\Delta,\infty}^{-1}\zeta}{\sqrt{v}}\right\}\right)$.

Lemma B.7. *Consider the Adam in Algorithm 2.1 and suppose that Assumption 2.1 2.3 hold, then for any initial point, and $T \geq 1$, the following results hold:*

$$\sum_{t=1}^{T} \mathbb{E}\left[\Pi_{\Delta,t+1} \sum_{i=1}^{d} \zeta_i(t)\right] \leq \begin{cases} C_{4,\delta}, & \text{if } \delta \in (0,1] \\ C_5 + C_6\, \mathbb{E}\left[\ln(S_T)\right], & \text{if } \delta = 0 \end{cases},$$

where C_5 and C_6 are two constants that depend on the initial point and the constants in our required assumptions (excluding $1/\mu$), and $C_{4,\delta}$ is a constant that depends on the initial point, δ, and the constants in our required assumptions (excluding $1/\mu$).

Lemma B.8 (Subsequence Convergence). *Under Assumptions 2.1–2.3, consider the Adam (Algorithm 2.1) with hyperparameters as specified in Subsection 2.4, where $\delta > 0$. Then, there exists a subsequence $\{w_{c_t}\}_{t\geq 1}$ such that its gradients converge to zero almost surely, i.e., $\lim_{t\to\infty} \|\nabla f(w_{c_t})\| = 0$ a.s.*

Lemma B.9. *Consider the Adam as defined in Algorithm 2.1, and assume that Assumptions 2.1 through 2.3 hold. Then, for any initial point and for all $T \geq 1$, the following results hold:*

- When $\delta = 0$, we have

$$\sup_{t\geq 1} \frac{\Pi_{\Delta,t+1}(f(w_t) - f^*)}{\ln^2(t+1)} < +\infty \ \text{ a.s., } \ \sup_{T\geq 1} \mathbb{E}\left[\frac{\Pi_{\Delta,t+1}(f(w_t) - f^*)}{\ln^2(t+1)}\right] < M_0 < +\infty,$$

- When $\delta > 0$, we have

$$\sup_{t\geq 1} \Pi_{\Delta,t+1}(f(w_t) - f^*) < +\infty \ \text{ a.s., } \ \sup_{T\geq 1} \mathbb{E}\left[\Pi_{\Delta,t+1}(f(w_t) - f^*)\right] < M_\delta < +\infty,$$

and

$$\sup_{t\geq 1} \Pi_{\Delta,t} \|m_{t-1}\| < +\infty \ \text{ a.s.},,$$

where M_0 and M_δ are two constants that depend on the initial point and the constants in our assumptions (excluding $1/\mu$).

Lemma B.10. *Consider the Adam in Algorithm 2.1 and suppose that Assumption 2.1 2.3 hold, then for any initial point, $T \geq 1,\ i \in [1,d]$, there is*

$$\mathbb{E}(S_T^{3/4}) = \begin{cases} \mathcal{O}(T^{3/4}), & \text{if } \delta \in (0,1] \\ \mathcal{O}(T^{3/4} \ln^{3/2} T), & \text{if } \delta = 0 \end{cases}.$$

where constant hidden in $\mathcal{O}$ depends only on initial point, and the constants in our required assumptions (not includes $1/\mu$).

Lemma B.11. *Under Assumptions 2.1–2.3, consider the Adam (Algorithm 2.1) with the hyperparameters specified in Subsection 2.4. Then, for any $t \geq 1$, the following inequality holds:*

$$\sup_{t\geq 1} \mathbb{E}[\Pi_{\Delta,t+1}\Sigma_{v_t}] < \begin{cases} \mathbb{E}[\Pi_{\Delta,2}\Sigma_{v_1}] + (A + 2L_f B)M_\delta + C, & \text{if } \delta \in (0,1] \\ \mathbb{E}[\Pi_{\Delta,2}\Sigma_{v_1}] + (A + 2L_f B)M_0 \ln^2 t + C, & \text{if } \delta = 0 \end{cases}.$$

Furthermore, if $\lambda > 1$, then we have

$$\sup_{t\geq 1} \mathbb{E}[\Pi_{\Delta,t+1}\Sigma_{v_t}] < \begin{cases} \left((A + 2L_f B)M_\delta + C\right) \sum_{t=1}^{+\infty} \frac{1}{(t+1)^\lambda}, & \text{if } \delta \in (0,1] \\ \left((A + 2L_f B)M_0 + C\right) \sum_{t=1}^{+\infty} \frac{\ln^2 t}{(t+1)^\lambda}, & \text{if } \delta = 0 \end{cases} < +\infty,$$

and the following almost sure bound:

$$\sup_{t\geq 1} \Sigma_{v_t} < +\infty \quad \text{a.s.}$$

Lemma B.12. *Under Assumption 2.1-2.3, consider the Adam with hyperparameters in Subsection 2.4 with $\gamma > 1,\ \delta > 0$. Then for any initial point, the following results hold:*

$$\sum_{t=1}^{+\infty} \eta_t \|\nabla f(w_t)\|^2 < +\infty \ \text{ a.s. and } \ \sum_{t=1}^{+\infty} \eta_t \|\nabla f(u_t)\|^2 < +\infty \ \text{ a.s.}$$

C PROOFS OF VITAL PROPERTIES

C.1 THE PROOF OF PROPERTY 2

Proof. Due to Algorithm 2.1, we know that

$$v_{t+1} = \beta_{2,t+1} v_t + (1-\beta_{2,t+1}) g_{t+1}^{\circ 2} = \Big(1 - \frac{1}{(t+1)^\gamma}\Big) v_t + \frac{1}{(t+1)^\gamma} g_{t+1}^{\circ 2}, \ (\forall\, t \geq 1).$$

which means

$$(t+1)^\gamma v_{t+1,i} = \big((t+1)^\gamma - 1\big) v_{t,i} + g_{t+1,i}^2 \geq t^\gamma v_{t,i}. \tag{13}$$

This implies that $tv_{t,i}$ is monotonically non-decreasing. Subsequently, we can obtain:

$$\eta_{v_t,i} = \frac{\eta_t}{\sqrt{v_{t,i}} + \mu} = \frac{\sqrt{t^\gamma}\eta_t}{\sqrt{t^\gamma v_{t,i}} + \sqrt{t^\gamma}\mu} = \frac{\frac{1}{t^{\delta - \frac{\gamma-1}{2}}}}{\sqrt{t v_{t,i}} + \sqrt{t^\gamma}\mu}.$$

It can be seen that the numerator is monotonically decreasing and greater than 0, while the denominator is monotonically non-increasing and greater than 0. Therefore, overall, we can deduce the monotonic non-increasing property of η_{v_t}. □

C.2 THE PROOF OF PROPERTY 3

Proof. For $v_{1,i}$, we can derive the following estimate:

$$v_{1,i} = \beta_{2,1} v_{0,i} + (1-\beta_{2,1}) g_{1,i}^2 = (1-\alpha_0) v + \alpha_0 g_{1,i}^2 = (1-\alpha_0) v + g_{1,i}^2 - (1-\alpha_0) g_{1,i}^2.$$

It is easy to find that $\alpha_1 S_{1,i} \leq v_{1,i} \leq S_{1,i}$. For $\forall\ k \geq 2$, we back to Eq. 13, acquiring $k^\gamma v_{k,i} \geq (k-1)^\gamma v_{k-1,i} + g_{k,i}^2$. Next, by summing up the above iterative equations, we obtain $\forall\ t \geq 2$,

$$t^\gamma v_{t,i} \geq v_{1,i} + \sum_{k=2}^{t} g_{k,i}^2.$$

Next, combining the estimate for $v_{1,i}$, we obtain $\forall\ t \geq 2$:

$$t^\gamma v_{t,i} \geq (1-\alpha_0) v + \alpha_0 g_{1,i}^2 + \sum_{k=2}^{t} g_{k,i}^2.$$

It is easy to find that $t^\gamma v_{t,i} \geq \alpha_1 S_{t,i}$. With this, we complete the proof. □

C.3 THE PROOF OF PROPERTY 5

Proof. According to Algorithm 2.1, we have the following iterative equations:

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t$$

We take the square of the 2-norm on both sides, yielding

$$\begin{aligned}
\|m_t\|^2 &= \|\beta_1 m_{t-1} + (1-\beta_1) g_t\|^2 \\
&= \beta_1^2 \|m_{t-1}\|^2 + 2\beta_1(1-\beta_1) m_{t-1}^\top g_t + (1-\beta_1)^2 \|g_t\|^2 \\
&\overset{(a)}{\leq} \beta_1 \|m_{t-1}\|^2 + (1-\beta_1)\|g_t\|^2 \\
&\overset{\text{Eq. 13}}{\leq} \beta_1 \|m_{t-1}\|^2 + (1-\beta_1)\|g_t\|^2.
\end{aligned}$$

In step (a), we used the *AM-GM* inequality, i.e.,

$$2\beta_1(1-\beta_1) m_{t-1}^\top g_t \leq \beta_1(1-\beta_1)\|m_{t-1}\|^2 + \beta_1(1-\beta_1)\|g_t\|^2,$$

that is,

$$\|m_t\|^2 - \|m_{t-1}\|^2 \leq -(1-\beta_1)\|m_{t-1}\|^2 + (1-\beta_1)\|g_t\|^2.$$

With this, we complete the proof. □

C.4 THE PROOF OF PROPERTY 5

Proof. Due to

$$|f(w_t) - f(u_t)| = \left|\nabla f(u_t)^\top (w_t - u_t) + \frac{L_f}{2}\|w_t - u_t\|^2\right| \le \|\nabla f(u_t)\|\|w_t - u_t\| + \frac{L_f}{2}\|w_t - u_t\|^2$$
$$\le \frac{1}{2}\|\nabla f(u_t)\|^2 + \frac{L_f + 1}{2}\|w_t - u_t\|^2$$
$$= Lf(u_t) + \frac{(L_f + 1)\beta_1^2}{2(1-\beta_1)^2}\|\eta_{v_{t-1}} \circ m_{t-1}\|^2,$$

we have

$$f(w_t) \le f(u_t) + |f(w_t) - f(u_t)| \le (L_f + 1)f(u_t) + \frac{(L_f + 1)\beta_1^2}{2(1-\beta_1)^2}\|\eta_{v_{t-1}} \circ m_{t-1}\|^2.$$

□

D PROOFS OF THEOREMS AND LEMMAS

D.1 PROOFS OF LEMMA 4.1

Proof. By L-smooth in Assumption 2.2, we have:

$$f(u_{t+1}) - f(u_t) \le \nabla f(u_t)^\top (u_{t+1} - u_t) + \frac{L_f}{2}\|u_{t+1} - u_t\|^2.$$

Then, by substituting the iterative formula for u_t from Eq. 4 into the above inequality, we obtain:

$$\begin{aligned} f(u_{t+1}) - f(u_t) \le & -\sum_{i=1}^{d} \eta_{v_t,i}\nabla_i f(u_t) g_{t,i} + \frac{\beta_1}{1-\beta_1}\sum_{i=1}^{d}\Delta_{t,i}\nabla_i f(u_t) m_{t-1,i} + L_f\sum_{i=1}^{d}\eta_{v_t,i}^2 g_{t,i}^2 \\ & + L_f\Big(\frac{\beta_1}{1-\beta_1}\Big)^2\sum_{i=1}^{d}\Delta_{t,i}^2 m_{t-1,i}^2 \\ \overset{(a)}{=} & \underbrace{-\sum_{i=1}^{d}\eta_{v_t,i}\nabla_i f(w_t) g_{t,i}}_{\Theta_{t,1}} + \underbrace{\sum_{i=1}^{d}(\eta_{v_t,i}(\nabla_i f(w_t) - \nabla_i f(u_t)) g_{t,i})}_{\Theta_{t,2}} \\ & + \frac{\beta_1}{1-\beta_1}\underbrace{\sum_{i=1}^{d}\Delta_{t,i}\nabla_i f(u_t) m_{t-1,i}}_{\Theta_{t,3}} + L_f\sum_{i=1}^{d}\eta_{v_t,i}^2 g_{t,i}^2 \\ & + L_f\Big(\frac{\beta_1}{1-\beta_1}\Big)^2\underbrace{\sum_{i=1}^{d}\Delta_{t,i}^2 m_{t-1,i}^2}_{\Theta_{t,4}}. \end{aligned} \tag{14}$$

Step (a) employs the identity $\nabla_i f(u_t) = \nabla_i f(w_t) + \nabla_i f(u_t) - \nabla_i f(w_t)$. Next, we handle $\Theta_{t,1}$, $\Theta_{t,2}$, $\Theta_{t,3}$ and $\Theta_{t,4}$ separately. First, for $\Theta_{t,1}$, we can perform the following identity transformation:

$$\begin{aligned} \Theta_{t,1} = & -\sum_{i=1}^{d}\eta_{v_t,i}\nabla_i f(w_t) g_{t,i} = -\sum_{i=1}^{d}\eta_{v_{t-1},i}\nabla_i f(w_t) g_{t,i} + \sum_{i=1}^{d}\Delta_{t,i}\nabla_i f(w_t) g_{t,i} \\ = & -\sum_{i=1}^{d}\underbrace{\eta_{v_{t-1},i}(\nabla_i f(w_t))^2}_{\zeta_i(t)} + \underbrace{\sum_{i=1}^{d}\Delta_{t,i}\nabla_i f(w_t) g_{t,i}}_{\Theta_{t,1,1}} + \underbrace{\sum_{i=1}^{d}\eta_{v_{t-1},i}\nabla_i f(w_t)(\nabla_i f(w_t) - g_{t,i})}_{M_{t,1}}, \end{aligned} \tag{15}$$

where $\Delta_{t,i}$ in the above inequality represents the i-th component of the vector Δ_t, which is defined in Eq. 4. It can be observed that we decompose Θ_1 into a descent term $-\sum_{i=1}^{d}\zeta_i(t)$, an error term $\Theta_{t,1,1}$, and a martingale difference term $M_{t,1}$. Next, we will further scale and control the error term $\Theta_{t,1,1}$. Specifically, we have:

$$
\begin{aligned}
\Theta_{t,1,1} &= \sum_{i=1}^{d}\mathbb{E}\left(\Delta_{t,i}\nabla_i f(w_t)g_{t,i} \mid \mathscr{F}_{t-1}\right) \\
&\quad + \underbrace{\sum_{i=1}^{d}\left(\Delta_{t,i}\nabla_i f(w_t)g_{t,i} - \mathbb{E}\left[\Delta_{t,i}\nabla_i f(w_t)g_{t,i} \mid \mathscr{F}_{t-1}\right]\right)}_{M_{t,2}} \\
&\overset{(a)}{<} \sum_{i=1}^{d}\sqrt{\eta_{v_{t-1},i}}\nabla_i f(w_t)\,\mathbb{E}\left[\sqrt{\Delta_{t,i}}g_{t,i} \mid \mathscr{F}_{t-1}\right] + M_{t,2} \\
&\overset{(b)}{\leq} \frac{1}{2}\sum_{i=1}^{d}\eta_{v_{t-1},i}(\nabla_i f(w_t))^2 + \frac{1}{2}\sum_{i=1}^{d}\mathbb{E}^2\left[\sqrt{\Delta_{t,i}}g_{t,i} \mid \mathscr{F}_{t-1}\right] + M_{t,2} \\
&\overset{(c)}{\leq} \frac{1}{2}\sum_{i=1}^{d}\zeta_i(t) + \frac{1}{2}\sum_{i=1}^{d}\mathbb{E}[g_{t,i}^2 \mid \mathscr{F}_{t-1}]\cdot\mathbb{E}[\Delta_{t,i} \mid \mathscr{F}_{t-1}] + M_{t,2} \\
&\leq \frac{1}{2}\sum_{i=1}^{d}\zeta_i(t) + \frac{1}{2}\sum_{i=1}^{d}\mathbb{E}[g_{t,i}^2 \mid \mathscr{F}_{t-1}]\cdot\mathbb{E}[\Delta_{t,i} \mid \mathscr{F}_{t-1}] + M_{t,2} \\
&\leq \frac{1}{2}\sum_{i=1}^{d}\zeta_i(t) + \frac{1}{2}\left(\sum_{i=1}^{d}\mathbb{E}[g_{t,i}^2 \mid \mathscr{F}_{t-1}]\right)\cdot\left(\sum_{i=1}^{d}\mathbb{E}[\Delta_{t,i} \mid \mathscr{F}_{t-1}]\right) + M_{t,2} \\
&\overset{(d)}{\leq} \frac{1}{2}\sum_{i=1}^{d}\zeta_i(t) + \frac{1}{2}\left((A+2L_fB)f(w_t)+C\right)\cdot\left(\sum_{i=1}^{d}\mathbb{E}[\Delta_{t,i} \mid \mathscr{F}_{t-1}]\right) + M_{t,2} \\
&= \frac{1}{2}\sum_{i=1}^{d}\zeta_i(t) + \frac{1}{2}(A+2L_fB)f(w_t)\cdot\left(\sum_{i=1}^{d}\mathbb{E}[\Delta_{t,i} \mid \mathscr{F}_{t-1}]\right) \\
&\quad + C\left(\sum_{i=1}^{d}\mathbb{E}[\Delta_{t,i} \mid \mathscr{F}_{t-1}]\right) + M_{t,2} \\
&= \frac{1}{2}\sum_{i=1}^{d}\zeta_i(t) + \frac{1}{2}(A+2L_fB)f(w_t)\cdot\left(\underbrace{\sum_{i=1}^{d}\mathbb{E}[\Delta_{t,i} \mid \mathscr{F}_{t-1}]}_{\overline{\Delta}_t}\right) + C\sum_{i=1}^{d}\Delta_{t,i} \\
&\quad + \underbrace{C\left(\sum_{i=1}^{d}\Big(\mathbb{E}[\Delta_{t,i} \mid \mathscr{F}_{t-1}] - \Delta_{t,i}\Big)\right)}_{M_{t,3}} + M_{t,2}
\end{aligned}
\tag{16}
$$

In the above derivation, in step (a), we utilized the property of conditional expectation, which states that if random variables $X \in \mathscr{F}_{n-1}$ and $Y \in \mathscr{F}_n$, then $\mathbb{E}[XY|\mathscr{F}_{n-1}] = X\,\mathbb{E}[Y|\mathscr{F}_{n-1}]$. Additionally, we need to note that $\Delta_{t,i} = \sqrt{\Delta_{t,i}}\sqrt{\Delta_{t,i}} < \sqrt{\eta_{v_{t-1}}}\sqrt{\Delta_{t,i}}$ (due to Property 2, we know $\Delta_{t,i} \geq 0$, so taking the square root is well-defined). In step (b), we employed the *AM-GM* inequality, which states $ab \leq \frac{a^2+b^2}{2}$. In step (c), we used the *Cauchy-Schwarz* inequality for conditional expectations: $\mathbb{E}[XY|\mathscr{F}_{n-1}] \leq \sqrt{\mathbb{E}[X^2|\mathscr{F}_{n-1}]\,\mathbb{E}[Y^2|\mathscr{F}_{n-1}]}$. For step (d), we used Property 1. Specifically, we have:

$$
\sum_{i=1}^{d}\mathbb{E}[g_{t,i}^2|\mathscr{F}_{t-1}] = \mathbb{E}[\|g_t\|^2|\mathscr{F}_{t-1}] \leq (A+2L_fB)f(w_t) + C.
$$

Substituting the estimate of $\Theta_{t,1,1}$ back into Eq. 16, we obtain:

$$\Theta_{t,1} = -\frac{1}{2}\sum_{i=1}^{d}\zeta_i(t) + \frac{A+2L_fB}{2}\overline{\Delta}_t \cdot f(w_t) + C\sum_{i=1}^{d}\Delta_{t,i} + \underbrace{M_{t,1}+M_{t,2}+M_{t,3}}_{M_t},$$

Then, we use Property 5 to replace $f(w_t)$ with $f(u_t)$ to obtain:

$$\begin{aligned}\Theta_{t,1} = &-\frac{1}{2}\sum_{i=1}^{d}\zeta_i(t) + \frac{(A+2L_fB)(L_f+1)}{2}\overline{\Delta}_t \cdot f(u_t) + C\sum_{i=1}^{d}\Delta_{t,i} \\ &+ \frac{(L_f+1)\beta_1^2}{2(1-\beta_1)^2}\|\eta_{v_{t-1}} \circ m_{t-1}\|^2 + M_t, \end{aligned} \tag{17}$$

Next, we deal with $\Theta_{t,2}$. Specifically, we have the following derivation:

$$\begin{aligned}\Theta_{t,2} &= \frac{1}{2}\sum_{i=1}^{d}(\eta_{v_t,i}(\nabla_i f(w_t) - \nabla_i f(u_t))g_{t,i}) \\ &\leq \sum_{i=1}^{d}\eta_{v_t,i}^2 g_{t,i}^2 + \frac{1}{2}\sum_{i=1}^{d}(\nabla_i f(w_t) - \nabla_i f(u_t))^2 \\ &= \sum_{i=1}^{d}\eta_{v_t,i}^2 g_{t,i}^2 + \frac{1}{2}\|\nabla f(w_t) - \nabla f(u_t)\|^2 \\ &\leq \sum_{i=1}^{d}\eta_{v_t,i}^2 g_{t,i}^2 + \frac{L_f^2}{2}\|w_t - u_t\|^2 \\ &= \sum_{i=1}^{d}\eta_{v_t,i}^2 g_{t,i}^2 + \frac{\beta_1^2 L_f^2}{2(1-\beta_1)^2}\|\eta_{v_{t-1}} \circ m_{t-1}\|^2. \end{aligned} \tag{18}$$

Next, we deal with $\Theta_{t,3}$, and we obtain:

$$\begin{aligned}\Theta_{t,3} = \sum_{i=1}^{d}\Delta_{t,i}\nabla_i f(u_t)m_{t-1,i} &\overset{\textit{AM-GM inequality}}{\leq} \frac{1}{8}\sum_{i=1}^{d}\Delta_{t,i}(\nabla_i f(u_t))^2 + 2\sum_{i=1}^{d}\Delta_{t,i}m_{t-1,i}^2 \\ &= \frac{1}{8}\sum_{i=1}^{d}\Delta_{t,i}(\nabla_i f(w_t))^2 + \frac{1}{8}\sum_{i=1}^{d}\Delta_{t,i}\big((\nabla_i f(u_t))^2 - (\nabla_i f(w_t))^2\big) + 2\sum_{i=1}^{d}\Delta_{t,i}m_{t-1,i}^2 \\ &\overset{(a)}{\leq} \frac{3}{16}\sum_{i=1}^{d}\Delta_{t,i}(\nabla_i f(w_t))^2 + \frac{3}{8}\sum_{i=1}^{d}(\nabla_i f(u_t) - \nabla_i f(w_t))^2 + 2\sum_{i=1}^{d}\Delta_{t,i}m_{t-1,i}^2 \\ &\overset{(b)}{\leq} \frac{3}{16}\sum_{i=1}^{d}\zeta_i(t) + \frac{3\beta_1^2 L_f^2}{8(1-\beta_1)^2}\|\eta_{v_{t-1}} \circ m_{t-1}\|^2 + 2\sum_{i=1}^{d}\Delta_{t,i}m_{t-1,i}^2. \end{aligned} \tag{19}$$

In step (a), we used the following substitution:

$$\begin{aligned}\frac{1}{8}(\nabla_i f(u_t))^2 - (\nabla_i f(w_t))^2 &= \frac{1}{8}\left(2(\nabla_i f(w_t))(\nabla_i f(u_t) - \nabla_i f(w_t)) + (\nabla_i f(u_t) - \nabla_i f(w_t))^2\right) \\ &\overset{\textit{AM-GM inequality}}{\leq} \frac{1}{16}(\nabla_i f(w_t))^2 + \frac{3}{8}(\nabla_i f(u_t) - \nabla_i f(w_t))^2.\end{aligned}$$

In step (b), we used the L-Smooth condition (Assumption 2.2), i.e.,

$$\begin{aligned}\sum_{i=1}^{d}(\nabla_i f(u_t) - \nabla_i f(w_t))^2 &= \|\nabla f(w_t) - \nabla f(u_t)\|^2 \\ &\leq L_f^2\|w_t - u_t\|^2 \overset{\text{Eq. 4}}{=} \frac{\beta_1^2 L_f}{(1-\beta_1)^2}\|\eta_{v_{t-1}} \circ m_{t-1}\|^2.\end{aligned}$$

For $\Theta_{t,4}$, we have that

$$\Theta_{t,4} = \sum_{i=1}^{d} \Delta_{t,i}^2 m_{t-1,i}^2 \overset{\Delta_{t,i} \leq \eta_{v_{t-1},i}}{<} \sum_{i=1}^{d} \eta_{t,i}^2 m_{t-1,i}^2 = \|\eta_{v_{t-1}} \circ m_{t-1}\|^2. \tag{20}$$

Finally, substituting the estimates of $\Theta_{t,1}$ from Eq. 17, $\Theta_{t,2}$ from Eq. 18, $\Theta_{t,3}$ from Eq. 19, and $\Theta_{t,4}$ from Eq. 20 back into Eq. 14, we obtain:

$$\begin{aligned}
f(u_{t+1}) - f(u_t) \leq & -\frac{5}{16}\sum_{i=1}^{d}\zeta_i(t) + \frac{(A+2L_fB)(L_f+1)}{2}\overline{\Delta}_t \cdot f(u_t) + C\sum_{i=1}^{d}\eta_{v_{t-1},i} - C\sum_{i=1}^{d}\eta_{v_t,i} \\
& + \frac{(L_f+1)\beta_1^2}{2(1-\beta_1)^2}\|\eta_{v_{t-1}} \circ m_{t-1}\|^2 + L_f\Big(\frac{\beta_1}{1-\beta_1}\Big)^2\sum_{i=1}^{d}\eta_{v_t,i}^2 g_{t,i}^2 \\
& + \frac{\beta_1^2 L_f^2}{2(1-\beta_1)^2}\|\eta_{v_{t-1}} \circ m_{t-1}\|^2 + L_f\Big(\frac{\beta_1}{1-\beta_1}\Big)^2\|\eta_{v_{t-1}} \circ m_{t-1}\|^2 \\
& + \frac{3\beta_1^2 L_f^2}{8(1-\beta_1)^2}\|\eta_{v_{t-1}} \circ m_{t-1}\|^2 + 2\sum_{i=1}^{d}\Delta_{t,i}m_{t-1,i}^2 + L_f\sum_{i=1}^{d}\eta_{v_t,i}^2 g_{t,i}^2 \\
& + M_t.
\end{aligned}$$

Then we note that $\Delta_{t,i} < \eta_{v_{t-1},i}$ and $\eta_{v_{t-1},i}^2 < \frac{1}{\sqrt{v}}\eta_{v_{t-1},i}$. After simplification, we obtain:

$$\begin{aligned}
&\Big(f(u_{t+1}) + C\sum_{i=1}^{d}\eta_{v_t,i}\Big) - \Big(f(u_t) + C\sum_{i=1}^{d}\eta_{v_{t-1},i}\Big) \leq -\frac{5}{16}\sum_{i=1}^{d}\zeta_i(t) + C_1\overline{\Delta}_t \cdot f(u_t) \\
&+ C_2\|\sqrt{\eta_{v_{t-1}}} \circ m_{t-1}\|^2 + (L_f+1)\sum_{i=1}^{d}\eta_{v_t,i}^2 g_{t,i}^2 + M_t,
\end{aligned}$$

where

$$\begin{aligned}
C_1 &:= \frac{(A+2L_fB)(L_f+1)}{2}, \\
C_2 &:= \frac{1}{\sqrt{v}} \cdot \Big(\frac{7L_f^2}{8} + \frac{L_f+1}{2}\Big) \cdot \frac{\beta_1^2}{(1-\beta_1)^2} + 2 + \frac{L_f}{\sqrt{v}} \cdot \Big(\frac{\beta_1}{1-\beta_1}\Big)^2.
\end{aligned} \tag{21}$$

We add the following term to both sides of the above inequality:

$$\frac{C_2}{(1-\beta_1)\sqrt{v}}(\|m_t\|^2 - \|m_{t-1}\|^2),$$

to obtain:

$$\begin{aligned}
&\Bigg(\underbrace{f(u_{t+1}) + C\sum_{i=1}^{d}\eta_{v_t,i} + \frac{C_2}{(1-\beta_1)\sqrt{v}}\|m_t\|^2}_{\hat{f}(u_{t+1})}\Bigg) - \Bigg(\underbrace{f(u_t) + C\sum_{i=1}^{d}\eta_{v_{t-1},i} + \frac{C_2}{(1-\beta_1)\sqrt{v}}\|m_{t-1}\|^2}_{\hat{f}(u_t)}\Bigg) \\
&\leq -\frac{5}{16}\sum_{i=1}^{d}\zeta_i(t) + C_1\overline{\Delta}_t \cdot f(u_t) \\
&+ \Big(\frac{C_2}{(1-\beta_1)\sqrt{v}}(\|m_t\|^2 - \|m_{t-1}\|^2) + C_2\|m_{t-1}\|^2\Big) + (L_f+1)\sum_{i=1}^{d}\eta_{v_t,i}^2 g_{t,i}^2 + M_t \\
&\overset{(a)}{\leq} -\frac{5}{16}\sum_{i=1}^{d}\zeta_i(t) + C_1\overline{\Delta}_t \cdot f(u_t) + 0 + (L_f+1)\sum_{i=1}^{d}\eta_{v_t,i}^2 g_{t,i}^2 + M_t.
\end{aligned} \tag{22}$$

In step (a), we mainly used Property 4. Specifically, we have:

$$\frac{C_2}{(1-\beta_1)\sqrt{v}}(\|m_t\|^2 - \|m_{t-1}\|^2) + C_2\|\sqrt{\eta_{v_{t-1}}} \circ m_{t-1}\|^2$$

$$\leq -C_2\|\sqrt{\eta_{v_{t-1}}} \circ m_{t-1}\|^2 + C_2\|\sqrt{\eta_{v_{t-1}}} \circ m_{t-1}\|^2 = 0.$$

We apply an obvious inequality to the second term on the right side of Eq. 22:

$$f(u_t) < f(u_t) + C\sum_{i=1}^{d}\eta_{v_{t-1},i} + \frac{C_2}{(1-\beta_1)\sqrt{v}}\|m_{t-1}\|^2 = \hat{f}(u_t),$$

and then move the expanded term to the left side of the inequality and combine like terms to obtain:

$$\hat{f}(u_{t+1}) - (1 + C_1\overline{\Delta}_t)\hat{f}(u_t) \leq -\frac{5}{16}\sum_{i=1}^{d}\zeta_i(t) + (L_f+1)\sum_{i=1}^{d}\eta_{v_t,i}^2 g_{t,i}^2 + M_t.$$

Next, we construct an auxiliary variable

$$\Pi_{\Delta,t} := \prod_{k=1}^{t-1}(1 + C_1\overline{\Delta}_k)^{-1} \quad (\forall\, t \geq 2),\ \Pi_{\Delta,1} := 1.$$

Multiplying both sides of the above inequality by $\Pi_{\Delta,t+1}$, we obtain

$$\begin{aligned}\Pi_{\Delta,t+1}\hat{f}(u_{t+1}) - \Pi_{\Delta,t}\hat{f}(u_t) \leq -\frac{5}{16}\Pi_{\Delta,t+1}\sum_{i=1}^{d}\zeta_i(t) + (L_f+1)\Pi_{\Delta,t+1}\sum_{i=1}^{d}\eta_{v_t,i}^2 g_{t,i}^2 \\ + \Pi_{\Delta,t+1}M_t.\end{aligned}$$

With this, we complete the proof.

□

D.2 The Proof of Lemma B.1

Proof. For $\forall x \in \mathbb{R}^N$, we define function

$$g(t) = f\left(x + t\frac{x'-x}{\|x'-x\|}\right),$$

where x' is a constant point such that $x'-x$ is parallel to $\nabla f(x)$. By taking the derivative, we obtain

$$g'(t) = \nabla_{x+t\frac{x'-x}{\|x'-x\|}} f\left(x + t\frac{x'-x}{\|x'-x\|}\right)^T \frac{x'-x}{\|x'-x\|}. \tag{23}$$

Through the Lipschitz condition of $\nabla f(x)$, we get $\forall t_1,\ t_2$

$$\begin{aligned}\left|g'(t_1) - g'(t_2)\right| &= \left|\left(\nabla_{x+t\frac{x'-x}{\|x'-x\|}} f\left(x + t_1\frac{x'-x}{\|x'-x\|}\right) - \nabla_{x+t\frac{x'-x}{\|x'-x\|}} f\left(x + t_2\frac{x'-x}{\|x'-x\|}\right)\right)^T \frac{x'-x}{\|x'-x\|}\right| \\ &\leq \left\|\nabla_{x+t\frac{x'-x}{\|x'-x\|}} f\left(x + t_1\frac{x'-x}{\|x'-x\|}\right) - \nabla_{x+t\frac{x'-x}{\|x'-x\|}} f\left(x + t_2\frac{x'-x}{\|x'-x\|}\right)\right\|\left\|\frac{x'-x}{\|x'-x\|}\right\| \leq \mathcal{L}|t_1 - t_2|.\end{aligned}$$

So $g'(t)$ satisfies the Lipschitz condition, and we have $\inf_{t\in\mathbb{R}} g(t) \geq \inf_{x\in\mathbb{R}^N} f(x) > -\infty$. Let $g^* = \inf x \in_{\mathbb{R}} g(x)$, then it holds that for $\forall\, t_0 \in \mathbb{R}$,

$$g(0) - g^* \geq g(0) - g(t_0). \tag{24}$$

By using the *Newton-Leibniz's* formula, we get that

$$g(0) - g(t_0) = \int_{t_0}^{0} g'(\alpha)d\alpha = \int_{t_0}^{0}\big(g'(\alpha) - g'(0)\big)d\alpha + \int_{t_0}^{0} g'(0)d\alpha.$$

Through the Lipschitz condition of g', we get that

$$g(0) - g(t_0) \geq \int_{t_0}^{0} -\mathcal{L}|\alpha - 0|d\alpha + \int_{t_0}^{0} g'(0)d\alpha = \frac{1}{2\mathcal{L}}\big(g'(0)\big)^2.$$

Then we take a special value of t_0. Let $t_0 = -g'(0)/\mathcal{L}$, then we get

$$\begin{aligned} g(0) - g(t_0) &\geq -\int_{t_0}^{0} \mathcal{L}|\alpha| d\alpha + \int_{t_0}^{0} g(0) dt = -\frac{\mathcal{L}}{2}(0 - t_0)^2 + g'(0)(-t_0) \\ &= -\frac{1}{2\mathcal{L}}\big(g'(0)\big)^2 + \frac{1}{\mathcal{L}}\big(g'(0)\big)^2 = \frac{1}{2\mathcal{L}}\big(g'(0)\big)^2. \end{aligned} \tag{25}$$

Substituting Eq. 25 into Eq. 24, we get

$$g(0) - g^* \geq \frac{1}{2\mathcal{L}}\big(g'(0)\big)^2.$$

Due to $g^* \geq f^*$ and $\big(g'(0)\big)^2 = \|\nabla f(x)\|^2$, it follows that

$$\big\|\nabla f(x)\big\|^2 \leq 2\mathcal{L}\big(f(x) - f^*\big).$$

□

D.3 The Proof of Lemma B.2

Proof. (i) Consider the non-negative adapted process $\{X_n, \mathscr{F}_n\}_{n\geq 1}$ and define the partial sum of conditional expectations as $\Lambda_T := \sum_{n=1}^{T} \mathbb{E}[X_n \mid \mathscr{F}_{n-1}]$.

First, we compute the expectation of Λ_T:

$$\mathbb{E}[\Lambda_T] = \mathbb{E}\left[\sum_{n=1}^{T} \mathbb{E}[X_n \mid \mathscr{F}_{n-1}]\right] = \sum_{n=1}^{T} \mathbb{E}[X_n] \leq M.$$

Since X_n are non-negative, we know that Λ_T is a non-decreasing sequence, and considering that $\mathbb{E}(\Lambda_T)$ $(\forall\ T \geq 1)$ is also bounded by M, we can apply the *Lebesgue's Monotone Convergence* theorem.

Thus, Λ_T converges almost surely to a limit Λ:

$$\Lambda := \lim_{T\to\infty} \Lambda_T = \sum_{n=1}^{\infty} \mathbb{E}[X_n \mid \mathscr{F}_{n-1}] \quad \text{a.s.}$$

This concludes that the sequence of conditional expectation sums converges almost surely.

(ii) We begin by normalizing X_n by considering the expression $Y_n = \frac{X_n}{2M}$. According to the *Lebesgue's monotone convergence* theorem, we only need to prove that

$$\forall\, p \geq 1,\ \ \mathbb{E}\left[\sum_{n=1}^{\infty} \mathbb{E}[Y_n|\mathscr{F}_{n-1}]\right]^p := M(p) < +\infty.$$

Next, we proceed with the calculation, and we obtain $\forall\, p \geq 2$, there is:

$$\begin{aligned} M(p) &= \mathbb{E}\left[\sum_{n=1}^{\infty} \mathbb{E}[Y_n|\mathscr{F}_{n-1}]\right]^p = \mathbb{E}\left[\sum_{n=1}^{\infty} Y_n + \sum_{n=1}^{\infty}(\mathbb{E}[Y_n|\mathscr{F}_{n-1}] - Y_n)\right]^p \\ &\overset{(a)}{\leq} \mathbb{E}\left[\frac{1}{2} + \sum_{n=1}^{\infty}(\mathbb{E}[Y_n|\mathscr{F}_{n-1}] - Y_n)\right]^p \overset{(b)}{\leq} 2^{p-1}\left(\frac{1}{2^p} + \mathbb{E}\left[\sum_{n=1}^{\infty}(\mathbb{E}[Y_n|\mathscr{F}_{n-1}] - Y_n)\right]^p\right) \\ &\overset{(c)}{\leq} \frac{1}{2} + 2^{p-1}C_p\,\mathbb{E}\left[\sum_{n=1}^{\infty} |\,\mathbb{E}[Y_n|\mathscr{F}_{n-1}] - Y_n|^2\right]^{p/2} \overset{(d)}{\leq} \frac{1}{2} + 2^{p-1}C_p\,\mathbb{E}\left[\sum_{n=1}^{\infty} |\,\mathbb{E}[Y_n|\mathscr{F}_{n-1}] - Y_n|\right]^{p/2} \\ &\overset{(f)}{\leq} \frac{1}{2} + 2^{p-2}C_p + 2^{\frac{3}{2}p-2}C_p\,\mathbb{E}\left[\sum_{n=1}^{\infty} \mathbb{E}[Y_n|\mathscr{F}_{n-1}]\right]^{p/2} \end{aligned}$$

$$= \frac{1}{2} + 2^{p-2}C_p + 2^{\frac{3}{2}p-2}C_p M(p/2). \tag{26}$$

In the above derivation, inequality (a) requires noting that

$$\sum_{n=1}^{+\infty} Y_n = \frac{1}{2}.$$

Inequality (b) uses the *AM-GM* inequality, specifically,

$$\left(\frac{a+b}{2}\right)^p \leq \frac{a^p + b^p}{2}.$$

Inequality (c) involves using *Burkholder's* inequality [1], where C_p is a constant depending only on p, and its order with respect to p is $\mathcal{O}(p)$. Inequality (d) requires noting that

$$|\,\mathbb{E}[Y_n|\mathscr{F}_{n-1}] - Y_n|^2 \leq |\,\mathbb{E}[Y_n|\mathscr{F}_{n-1}] - Y_n|.$$

By repeatedly iterating Eq. 26 and using the fact that $C_p = \mathcal{O}(p)$, we can finally obtain the following estimate:

$$M(p) = o(p^{\sqrt{p}}),$$

that is,

$$\mathbb{E}[\Lambda^p] = o((2M)^p \cdot p^{\sqrt{p}})$$

(iii) Similarly, we first consider the partial sum sequence Λ_T. Since there exists an obvious upper bound for this sequence that depends on T, given by $\Lambda_T < M \cdot T\ a.s.$, and because the Taylor series of the function $h_p(x) = e^{px}$ at $x = 0$ converges uniformly on every compact subset of $\mathbb{R}$, we can expand e^{Λ_T} using its Taylor series, yielding the following expression:

$$e^{\Lambda_T} = \sum_{n=0}^{\infty} \frac{\Lambda_T^n}{n!}.$$

Since the Taylor series converges uniformly on every compact subset, we can interchange the summation and the expectation operators when taking the expectation on both sides of the above equation. Specifically, we have:

$$\mathbb{E}[e^{p\Lambda_T}] = \mathbb{E}\left[\sum_{n=0}^{\infty} \frac{p^n \Lambda_T^n}{n!}\right] = \sum_{n=0}^{\infty} \frac{p^n \mathbb{E}[\Lambda_T^n]}{n!}.$$

Noting that $\mathbb{E}[\Lambda_T^n] < \mathbb{E}[\Lambda^n] < o((2M)^n \cdot n^{\sqrt{n}})$, we have:

$$\mathbb{E}[e^{p\Lambda_T}] = \sum_{n=0}^{\infty} \frac{p^n \mathbb{E}[\Lambda_T^n]}{n!} \leq \sum_{n=0}^{\infty} \frac{o((2pM)^n \cdot n^{\sqrt{n}})}{n!} = \mathcal{O}\bigg(\sum_{n=0}^{\infty} \frac{(2pM)^n \cdot n^{\sqrt{n}}}{n!}\bigg).$$

We use *Stirling's* approximation [2] to substitute the factorial in the denominator. It is evident that the series inside the $\mathcal{O}$ notation converges and depends only on p and M. Then, by applying the *Lebesgue's Monotone Convergence* theorem, we can prove that $\mathbb{E}[e^{p\Lambda}]$ exists, and its upper bound depends only on p and M.

With this, we complete the proof. □

[1] ***Burkholder's* inequality**: For any martingale $(M_n, \mathscr{F}_n)$ with $M_0 = 0$ almost surely, and for any $1 \leq p < \infty$, there exist constants $c_p > 0$ and $C_p > 0$ depending only on p such that:

$$c_p\, \mathbb{E}[(S(M))^p] \leq \mathbb{E}[(M^*)^p] \leq C_p\, \mathbb{E}[(S(M))^p],$$

where $M^* = \sup_{n\geq 0} |M_n|$ and $S(M) = \left(\sum_{i\geq 1}(M_i - M_{i-1})^2\right)^{1/2}$.

[2] Stirling's approximation can be written as:

$$n! = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\frac{\theta_n}{12n}},$$

where $0 < \theta_n < 1$.

D.4 THE PROOF OF LEMMA B.3

Proof. First, using the expression in Eq. 7, we express $\Pi_{\Delta,T+1}^{-p}$ as follows:

$$\Pi_{\Delta,T+1}^{-p} = \prod_{k=1}^{T}(1 + C_1\overline{\Delta}_k)^p.$$

By applying the logarithmic transformation, we obtain:

$$\Pi_{\Delta,T+1}^{-p} = \exp\left\{p\sum_{k=1}^{T}\ln(1 + C_1\overline{\Delta}_k)\right\} \overset{\ln(1+x)\leq x\ (\forall\ x>0)}{\leq} \exp\left\{pC_1\sum_{k=1}^{T}\overline{\Delta}_k\right\}.$$

Based on the definition of $\overline{\Delta}_k$, we have

$$\overline{\Delta}_k = \sum_{i=1}^{d}\mathbb{E}[\Delta_{k,i}|\mathscr{F}_{k-1}] = \mathbb{E}\left[\sum_{i=1}^{d}\Delta_{k,i}\middle|\mathscr{F}_{k-1}\right].$$

Since $\sum_{i=1}^{d}\Delta_{k,i} > 0$ and

$$\sum_{k=1}^{+\infty}\sum_{i=1}^{d}\Delta_{k,i} < \frac{d}{\sqrt{v}},$$

we can use Lemma B.2 to prove that:

$$\mathbb{E}\left[\exp\left\{pC_1\sum_{k=1}^{T}\overline{\Delta}_k\right\}\right] < C_{v,d,p},$$

which implies

$$\mathbb{E}[\Pi_{\Delta,T+1}^{-p}] < C_{v,d,p}.$$

Next, noting the monotonicity of the sequence $\{\Pi_{\Delta,T+1}^{-1}\}_{T\geq 1}$, we take $T \to +\infty$ and apply the *Lebesgue's Monotone Convergence* theorem to obtain the result regarding $\Pi_{\Delta,\infty}$.

With this, the proof is complete.

□

D.5 THE PROOF OF LEMMA B.4

Proof. Recalling the approximate descent inequality 4.1, we have:

$$\begin{aligned}\Pi_{\Delta,t+1}\hat{f}(u_{t+1}) - \Pi_{\Delta,t}\hat{f}(u_t) \leq &-\frac{5}{16}\Pi_{\Delta,t+1}\sum_{i=1}^{d}\zeta_i(t) + (L_f+1)\Pi_{\Delta,t+1}\sum_{i=1}^{d}\eta_{v_t,i}^2 g_{t,i}^2 \\ &+ \Pi_{\Delta,t+1}M_t.\end{aligned}$$

We multiply both sides of the above equation by $\frac{1}{\sqrt{S_{t-1}}}$, we obtain:

$$\begin{aligned}\frac{\Pi_{\Delta,t+1}\hat{f}(u_{t+1})}{\sqrt{S_t}} - \frac{\Pi_{\Delta,t}\hat{f}(u_t)}{\sqrt{S_{t-1}}} &\leq (L_f+1)\frac{\Pi_{\Delta,t+1}}{\sqrt{S_{t-1}}}\sum_{i=1}^{d}\eta_{v_t,i}^2 g_{t,i}^2 + \frac{\Pi_{\Delta,t+1}}{\sqrt{S_{t-1}}}M_t \\ &\overset{\Pi_{\Delta,t+1}\leq 1}{\leq} (L_f+1)\frac{1}{\sqrt{S_{t-1}}}\sum_{i=1}^{d}\eta_{v_t,i}^2 g_{t,i}^2 + \frac{\Pi_{\Delta,t+1}}{\sqrt{S_{t-1}}}M_t.\end{aligned}$$

We sum the above inequality with respect to the index t from 1 to T, and we obtain:

$$\frac{\Pi_{\Delta,t+1}\hat{f}(u_{t+1})}{\sqrt{S_t}} \leq \frac{\hat{f}(u_1)}{\sqrt{S_0}} + (L_f+1)\sum_{t=1}^{T}\underbrace{\frac{1}{\sqrt{S_{t-1}}}\sum_{i=1}^{d}\eta_{v_t,i}^2 g_{t,i}^2}_{\Gamma_{t,3}} + \sum_{t=1}^{T}\frac{\Pi_{\Delta,t+1}}{\sqrt{S_{t-1}}}M_t. \tag{27}$$

For $\Gamma_{t,3}$, we can perform the following simplification, and we obtain:

$$\begin{aligned}\frac{1}{\sqrt{S_{t-1}}}\sum_{i=1}^{d}\eta_{v_t,i}^2 g_{t,i}^2 &= \frac{1}{\sqrt{S_t}}\sum_{i=1}^{d}\eta_{v_t,i}^2 g_{t,i}^2 + \left(\frac{1}{\sqrt{S_{t-1}}}-\frac{1}{\sqrt{S_t}}\right)\sum_{i=1}^{d}\eta_{v_t,i}^2 g_{t,i}^2 \\ &= \frac{1}{\sqrt{S_t}}\sum_{i=1}^{d}\frac{1}{t^{2\delta}}\frac{g_{t,i}^2}{v_{t,i}} + \left(\frac{1}{\sqrt{S_{t-1}}}-\frac{1}{\sqrt{S_t}}\right)\sum_{i=1}^{d}\frac{1}{t^{2\delta}}\frac{g_{t,i}^2}{v_{t,i}} \\ &\overset{\text{Property 3}}{\leq} \frac{1}{\alpha_1\sqrt{S_t}}\sum_{i=1}^{d}\frac{g_{t,i}^2}{S_{t,i}} + \frac{1}{\alpha_1}\left(\frac{1}{\sqrt{S_{t-1}}}-\frac{1}{\sqrt{S_t}}\right)\sum_{i=1}^{d}\frac{g_{t,i}^2}{S_{t,i}} \\ &\overset{S_t\geq S_{t,i}}{\leq} \frac{1}{\alpha_1}\sum_{i=1}^{d}\frac{g_{t,i}^2}{S_{t,i}^{3/2}} + \frac{d}{\alpha_1}\left(\frac{1}{\sqrt{S_{t-1}}}-\frac{1}{\sqrt{S_t}}\right).\end{aligned}$$

By using the series-integral inequality, we obtain:

$$\begin{aligned}\sum_{t=1}^{T}\Gamma_{t,3} &\leq \frac{1}{\alpha_1}\sum_{i=1}^{d}\sum_{t=1}^{T}\frac{g_{t,i}^2}{S_{t,i}^{3/2}} + \frac{d}{\alpha_1}\sum_{t=1}^{T}\left(\frac{1}{\sqrt{S_{t-1}}}-\frac{1}{\sqrt{S_t}}\right) \\ &\leq \frac{1}{\alpha_1}\sum_{i=1}^{d}\int_{S_{0,i}}^{+\infty}\frac{1}{x^{3/2}}\mathrm{d}x + \frac{d}{\alpha_1\sqrt{S_0}} = \frac{2d}{\alpha_1\sqrt{v}} + \frac{1}{\alpha_1}\sqrt{\frac{d}{v}} \\ &< \frac{3d}{\alpha_1\sqrt{v}}.\end{aligned}$$

Substituting above inequality into Eq. 27, we acquire

$$\frac{\Pi_{\Delta,t+1}\hat{f}(u_{t+1})}{\sqrt{S_t}} \leq \frac{\hat{f}(u_1)}{\sqrt{dv}} + \frac{3(L_f+1)d}{\alpha_1\sqrt{v}} + \sum_{t=1}^{T}\frac{\Pi_{\Delta,t+1}}{\sqrt{S_{t-1}}}M_t.$$

With this, we complete the proof.

□

D.6 The Proof of Lemma B.5

Proof. For any $\phi \in \mathbb{R}$, we consider $\frac{\sqrt{S_T}}{(T+1)^\phi}$, and we obtain:

$$\begin{aligned}\frac{\sqrt{S_T}}{(T+1)^\phi} &= \frac{S_T}{(T+1)^\phi\sqrt{S_T}} = \frac{S_0+\sum_{t=1}^{T}\|g_t\|^2}{(T+1)^\phi\sqrt{S_T}} = \frac{S_0}{(T+1)^\phi\sqrt{S_T}} + \sum_{t=1}^{T}\frac{\|g_t\|^2}{(T+1)^\phi\sqrt{S_T}} \\ &\leq \frac{S_0}{(T+1)^\phi\sqrt{S_T}} + \sum_{t=1}^{T}\frac{\|g_t\|^2}{(T+1)^\phi\sqrt{S_T}} \leq \sqrt{S_0} + \sum_{t=1}^{T}\frac{\|g_t\|^2}{(t+1)^\phi\sqrt{S_{t-1}}} \\ &= \sqrt{dv} + \sum_{t=1}^{T}\frac{\|g_t\|^2}{(t+1)^\phi\sqrt{S_{t-1}}}.\end{aligned}$$

Next, by multiplying both sides of the above inequality by $\Pi_{\Delta,T}$ and noting the monotonicity of $\{\Pi_{\Delta,t}\}_{t\geq 1}$ as well as the fact that $\Pi_{\Delta,T} \leq 1$ for all $T \geq 1$, we immediately obtain the result.

□

D.7 The Proof of Lemma B.6

Proof. We take $\phi = 2$ in Lemma B.5 and bound the expectation of the partial sum $\sum_{t=1}^{T}\Lambda_{2,t}$. We have:

$$\mathbb{E}\left[\sum_{t=1}^{T}\Pi_{\Delta,t}\Lambda_{2,t}\right] = \sum_{t=1}^{T}\mathbb{E}[\Pi_{\Delta,t}\Lambda_{2,t}] = \sum_{t=1}^{T}\mathbb{E}\left[\frac{\Pi_{\Delta,t}\|g_t\|^2}{(t+1)^2\sqrt{S_{t-1}}}\right] = \sum_{t=1}^{T}\mathbb{E}\left[\frac{\Pi_{\Delta,t}\,\mathbb{E}[\|g_t\|^2|\mathscr{F}_{t-1}]}{(t+1)^2\sqrt{S_{t-1}}}\right]$$

$$\overset{\text{Property 1}}{\leq} \sum_{t=1}^{T} \mathbb{E}\left[\frac{(A+2L_fB)\Pi_{\Delta,t}(f(w_t)-f^*)+C}{(t+1)^2\sqrt{S_{t-1}}}\right]$$
$$\overset{\text{Property 5}}{\leq} \sum_{t=1}^{T} \mathbb{E}\left[\frac{(A+2L_fB)\Pi_{\Delta,t}((L_f+1)f(u_t)+\frac{(L_f+1)\beta_1^2}{2(1-\beta_1)^2}\|\eta_{v_{t-1}}\circ m_{t-1}\|^2-f^*)+C}{(t+1)^2\sqrt{S_{t-1}}}\right]$$
$$\leq C_3\sum_{t=1}^{T}\frac{1}{(t+1)^2}\,\mathbb{E}\left[\frac{\Pi_{\Delta,t}\hat{f}(u_t)}{\sqrt{S_{t-1}}}\right]+C_4\sum_{t=1}^{T}\frac{1}{(t+1)^2}, \tag{28}$$

where

$$C_3 := (A+2L_fB)\max\left\{L_f+1, \frac{(L_f+1)\beta_1^2}{2C_2(1-\beta_1)}\right\},\ C_4 := \frac{(A+2L_fB)|f^*|+C}{\sqrt{S_0}}.$$

Based on the results in Lemma B.4, we can compute:

$$\mathbb{E}\left[\frac{\Pi_{\Delta,t}\hat{f}(u_t)}{\sqrt{S_{t-1}}}\right] \leq \frac{\mathbb{E}[\hat{f}(u_1)]}{\sqrt{dv}}+\frac{3(L_f+1)d}{\alpha_1\sqrt{v}}+0.$$

Substitute above result into Eq. 28, and combine

$$\sum_{t=1}^{T}\frac{1}{(t+1)^2}\leq\sum_{t=1}^{+\infty}\frac{1}{t^2}=\frac{\pi^2}{6},$$

we get:

$$\mathbb{E}\left[\sum_{t=1}^{T}\Pi_{\Delta,t}\Lambda_{2,t}\right]\leq\frac{C_3\pi^2\,\mathbb{E}[\hat{f}(u_1)]}{6\sqrt{dv}}+\frac{\pi^2C_3(L_f+1)d}{2\alpha_1\sqrt{v}}+\frac{\pi^2C_4}{6}.$$

It can be observed that the right side of the above inequality is independent of T. Thus, according to the *Lebesgue's Monotone Convergence* theorem, we have

$$\sum_{t=1}^{T}\Pi_{\Delta,t}\Lambda_{2,t}\to\sum_{t=1}^{+\infty}\Pi_{\Delta,t}\Lambda_{2,t}\quad\text{a.s.},$$

and

$$\mathbb{E}\left[\sum_{t=1}^{+\infty}\Pi_{\Delta,t}\Lambda_{2,t}\right]=\lim_{T\to\infty}\mathbb{E}\left[\sum_{t=1}^{T}\Pi_{\Delta,t}\Lambda_{2,t}\right]=\lim_{T\to\infty}\sum_{t=1}^{T}\mathbb{E}[\Pi_{\Delta,t}\Lambda_{2,t}]\leq\frac{C_3\pi^2\,\mathbb{E}[\hat{f}(u_1)]}{6\sqrt{dv}}+\frac{\pi^2C_3(L_f+1)d}{2\alpha_1\sqrt{v}}+\frac{\pi^2C_4}{6}.$$

Next, we set

$$\zeta:=\sqrt{dv}+\sum_{t=1}^{+\infty}\Pi_{\Delta,t}\Lambda_{2,t},$$

and combine Lemma B.5. We get:

$$\sqrt{S_T}\leq\Pi_{\Delta,T}^{-1}(T+1)^2\zeta<\Pi_{\Delta,\infty}^{-1}(T+1)^2\zeta. \tag{29}$$

Meanwhile,

$$\mathbb{E}[\zeta]=\sqrt{dv}+\mathbb{E}\left[\sum_{t=1}^{+\infty}\Lambda_{2,t}\right]\leq\sqrt{dv}+\frac{C_3\pi^2\,\mathbb{E}[\hat{f}(u_1)]}{6\sqrt{dv}}+\frac{\pi^2C_3(L_f+1)d}{2\alpha_1\sqrt{v}}+\frac{\pi^2C_4}{6}. \tag{30}$$

Then through Eq. 29, we have

$$\frac{1}{2}\ln\left(\frac{2S_T}{v}\right)\leq 2\ln(T+1)+\ln\left(\max\left\{e,\frac{\sqrt{2}\sqrt{2}\Pi_{\Delta,\infty}^{-1}\zeta}{\sqrt{v}}\right\}\right)$$
$$=2\ln\left(\max\left\{e,\frac{\sqrt{2}\Pi_{\Delta,\infty}^{-1}\zeta}{\sqrt{v}}\right\}\right)\left(\frac{\ln(T+1)}{\ln\left(\max\left\{e,\frac{\sqrt{2}\Pi_{\Delta,\infty}^{-1}\zeta}{\sqrt{v}}\right\}\right)}+\frac{1}{2}\right)-\ln(\Pi_{\Delta,T}).$$

Next, we note that

$$\ln\left(\max\left\{e, \frac{\sqrt{2}\Pi_{\Delta,\infty}^{-1}\zeta}{\sqrt{v}}\right\}\right) \geq 1, \quad \text{and} \quad \frac{1}{2} \leq \ln(T+1),$$

from which we obtain:

$$\begin{aligned}\frac{1}{2}\ln(S_T) &\leq 2\ln\left(\max\left\{e, \frac{\sqrt{2}\Pi_{\Delta,\infty}^{-1}\zeta}{\sqrt{v}}\right\}\right)\left(\frac{\ln(T+1)}{\ln\left(\max\left\{e, \frac{\sqrt{2}\Pi_{\Delta,\infty}^{-1}\zeta}{\sqrt{v}}\right\}\right)} + \frac{1}{2}\right) \\ &\leq 4\ln\left(\max\left\{e, \frac{\sqrt{2}\Pi_{\Delta,\infty}^{-1}\zeta}{\sqrt{v}}\right\}\right)\ln(T+1).\end{aligned}$$

With this, we complete the proof.

□

D.8 The Proof of Lemma B.7

Proof. For any $T > 0$, taking the expectation on both sides of the recursive inequality in the *Sufficient Decreasing* lemma (Lemma 4.1) and summing the indices from $n = 1$ to T, noting $\mathbb{E}[\Pi_{\Delta,t+1}M_t] = \mathbb{E}[\Pi_{\Delta,t+1}\mathbb{E}[M_t|\mathscr{F}_{t-1}]] = 0$, we obtain:

$$\begin{aligned}\frac{5}{16}\sum_{t=1}^{T}\mathbb{E}\left[\Pi_{\Delta,t+1}\sum_{i=1}^{d}\zeta_i(t)\right] &\leq \hat{f}(u_1) - \Pi_{\Delta,t+1}\hat{f}(u_{t+1}) + (L_f+1)\Pi_{\Delta,t+1}\sum_{t=1}^{T}\mathbb{E}\left(\sum_{i=1}^{d}\eta_{v_t,i}^2 g_{t,i}^2\right) \\ &\quad + 0 \\ &\overset{\hat{f}(u_{t+1})\geq f^*}{\leq} \hat{f}(u_1) - \Pi_{\Delta,t+1}f^* + (L_f+1)\Pi_{\Delta,t+1}\sum_{t=1}^{T}\mathbb{E}\left(\sum_{i=1}^{d}\eta_{v_t,i}^2 g_{t,i}^2\right) \\ &\overset{\Pi_{\Delta,t+1}\leq 1}{\leq} \hat{f}(u_1) + |f^*| + (L_f+1)\sum_{t=1}^{T}\mathbb{E}\Bigg[\underbrace{\Pi_{\Delta,t+1}\sum_{i=1}^{d}\eta_{v_t,i}^2 g_{t,i}^2}_{\Gamma_t}\Bigg].\end{aligned} \tag{31}$$

To prove the conclusion of this lemma, we actually only need to bound $\sum_{t=1}^{T}\mathbb{E}\left[\Gamma_t\right]$. Specifically, we perform the following transformation on Γ_t. We have:

$$\begin{aligned}\Gamma_t = \Pi_{\Delta,t+1}\sum_{i=1}^{d}\eta_{v_t,i}^2 g_{t,i}^2 &= \Pi_{\Delta,t+1}\sum_{i=1}^{d}\frac{\eta_t^2 g_{t,i}^2}{(\sqrt{v_{t,i}}+\mu)^2} \leq \Pi_{\Delta,t+1}\sum_{i=1}^{d}\frac{1}{t^{2\delta}}\frac{g_{t,i}^2}{tv_{t,i}} \\ &\overset{\text{Property 3}}{\leq} \frac{(t+1)^{2\delta}}{t^{2\delta}}\Pi_{\Delta,t+1}\sum_{i=1}^{d}\frac{1}{\alpha_1(t+1)^{2\delta}}\frac{g_{t,i}^2}{S_{t,i}} \overset{(a)}{\leq} \frac{2^{2\delta}}{\alpha_1}\zeta^{\delta}\Pi_{\Delta,t+1}^{1-\delta}\sum_{i=1}^{d}\frac{g_{t,i}^2}{{S_{t,i}}^{1+\frac{\delta}{2}}} \\ &\overset{\Pi_{\Delta,t+1}\leq 1}{\leq} \frac{2^{2\delta}}{\alpha_1}\zeta^{\delta}\sum_{i=1}^{d}\frac{g_{t,i}^2}{{S_{t,i}}^{1+\frac{\delta}{2}}}.\end{aligned} \tag{32}$$

In step (a) of the above derivation, we need to apply Lemma B.6 to $(t+1)^{2\delta}$. Specifically, according to Lemma B.6, we have

$$\sqrt{S_t} \leq (t+1)^2\zeta,$$

which means:

$$\frac{1}{(t+1)^{2\delta}} \leq \frac{\zeta^{\delta}}{(t+1)^{2\delta}S_t^{\frac{\delta}{2}}} \leq \frac{\zeta^{\delta}}{(t+1)^{2\delta}S_{t,i}^{\frac{\delta}{2}}}\Pi_{\Delta,t}^{-1}.$$

Next, with the estimate for Γ_t, we can estimate $\sum_{t=1}^{T}\mathbb{E}[\Gamma_t]$. Specifically, we have:

$$\sum_{t=1}^{T}\mathbb{E}[\Gamma_t] = \mathbb{E}\left[\frac{2^{2\delta}\zeta^{\delta}}{\alpha_1}\sum_{i=1}^{d}\sum_{t=1}^{T}\frac{g_{t,i}^2}{{S_{t,i}}^{1+\frac{\delta}{2}}}\right] \leq \mathbb{E}\left[\frac{2^{2\delta}\zeta^{\delta}}{\alpha_1}\sum_{i=1}^{d}\int_{S_{0,i}}^{S_{T,i}}\frac{1}{x^{1+\frac{\delta}{2}}}\mathrm{d}x\right]$$

$$\leq \begin{cases} \frac{2^{2\delta}d}{\alpha_1 v^{\frac{\delta}{2}}} \mathbb{E}\left[\zeta^{\delta}\right], & \text{if } \delta \in (0,1] \\ \frac{2^{2\delta}}{\alpha_1} \mathbb{E}\left[\ln\left(\frac{S_T}{dv}\right)\right], & \text{if } \delta = 0 \end{cases}$$
$$\overset{(a)}{\leq} \begin{cases} \mathcal{O}(1), & \text{if } \delta \in (0,1] \\ \frac{2^{2\delta}}{\alpha_1} \mathbb{E}\left[\ln\left(\frac{S_T}{dv}\right)\right], & \text{if } \delta = 0 \end{cases}. \tag{33}$$

In step (a), we used the following *Hölder's* inequality to obtain the $\mathcal{O}(1)$ result

$$\mathbb{E}\left[\zeta^{\delta}\right] \leq \left(\mathbb{E}\left[\zeta\right]\right)^{\delta} = C_{\zeta}^{\delta}.$$

Next, we substitute the estimate of $\sum_{t=1}^{T} \mathbb{E}\left[\Gamma_t\right]$ from Eq. 33 back into Eq. 31, and we obtain the result. This completes the proof.

□

D.9 The Proof of Lemma B.9

Proof. We only analyze the case where $\delta = 0$; the case where $\delta > 0$ can be treated using exactly the same analytical approach. Returning to the approximate descent inequality (Lemma 4.1), we have:

$$\Pi_{\Delta,t+1}\hat{f}(u_{t+1}) - \Pi_{\Delta,t}\hat{f}(u_t) \leq -\frac{5}{16}\Pi_{\Delta,t+1}\sum_{i=1}^{d}\zeta_i(t) + (L_f+1)\Pi_{\Delta,t+1}\sum_{i=1}^{d}\eta_{v_t,i}^2 g_{t,i}^2 + \Pi_{\Delta,t+1}M_t.$$

We divide both sides of the above inequality by $\ln^2(t+1)$, and noting that $\ln^2(t+1) < \ln^2(t+2)$, we obtain:

$$\frac{\Pi_{\Delta,t+1}\hat{f}(u_{t+1})}{\ln^2(t+2)} - \frac{\Pi_{\Delta,t}\hat{f}(u_t)}{\ln^2(t+1)} \leq (L_f+1)\frac{\Pi_{\Delta,t+1}}{\ln^2(t+1)}\sum_{i=1}^{d}\eta_{v_t,i}^2 g_{t,i}^2 + \frac{\Pi_{\Delta,t+1}M_t}{\ln^2(t+1)}$$
$$\overset{\Pi_{\Delta,t+1}\leq 1}{\leq} (L_f+1)\underbrace{\frac{1}{\ln^2(t+1)}\sum_{i=1}^{d}\eta_{v_t,i}^2 g_{t,i}^2}_{\Omega_t} + \frac{\Pi_{\Delta,t+1}M_t}{\ln^2(t+1)}. \tag{34}$$

For Ω_t, we can perform the following transformation, and we have:

$$\Omega_t = \frac{1}{\ln^2(t+1)}\sum_{i=1}^{d}\eta_{v_t,i}^2 g_{t,i}^2 = \frac{1}{\ln^2(t+1)}\sum_{i=1}^{d}\frac{g_{t,i}^2}{t v_{t,i}} \leq \frac{1}{\alpha_1}\sum_{i=1}^{d}\frac{g_{t,i}^2}{\ln^2(t+1)S_{t,i}}$$
$$\overset{(a)}{\leq} \frac{1}{\alpha_1}\sum_{i=1}^{d}\frac{(\zeta'^2)\frac{2g_{t,i}^2}{v}}{\ln^2(\frac{2S_{t,i}}{v})\frac{2S_{t,i}}{v}}.$$

In step (a), we use the last result from Lemma B.6, which states:

$$\ln\left(\frac{2S_{t,i}}{v}\right) \leq \ln\left(\frac{2S_t}{v}\right) \leq \zeta' \ln(T+1).$$

Then, using the series-integral inequality, we can bound $\sum_{t=1}^{+\infty}\mathbb{E}[\Omega_t]$, and we obtain:

$$\sum_{t=1}^{+\infty}\mathbb{E}[\Omega_t] \leq \frac{1}{\alpha_1}\sum_{i=1}^{d}\mathbb{E}\left[\sum_{t=1}^{+\infty}\frac{(\zeta'^2)\frac{2g_{t,i}^2}{v}}{\ln^2(\frac{2S_{t,i}}{v})\frac{2S_{t,i}}{v}}\right] < \frac{1}{\alpha_1}\sum_{i=1}^{d}\mathbb{E}\left[\int_2^{+\infty}\frac{\zeta'^2}{x\ln^2 x}\mathrm{d}x\right] = \frac{d\,\mathbb{E}[\zeta'^2]}{\alpha_1 \ln 2}.$$

Next, we use the explicit expression for ζ' given in Lemma B.6 to bound $\mathbb{E}[\zeta'^2]$. We have:

$$\mathbb{E}[\zeta'^2] = \mathbb{E}\left[8\ln\left(\max\left\{e, \frac{\sqrt{2}\Pi_{\Delta,\infty}^{-1}\zeta}{\sqrt{v}}\right\}\right)\right] = 8\,\mathbb{E}\left[\max\left\{1, \ln\left(\frac{\sqrt{2}\Pi_{\Delta,\infty}^{-1}\zeta}{\sqrt{v}}\right)\right\}\right]$$
$$\leq 8\,\mathbb{E}\left[\max\left\{1, \ln\sqrt{2} - \ln\Pi_{\Delta,\infty} + \ln(\zeta) - \ln\sqrt{v}\right\}\right]$$

$$\overset{\text{Lemma B.3 and Lemma B.6}}{<} +\infty.$$

As a result, we have:

$$\sum_{t=1}^{+\infty} \mathbb{E}[\Omega_t] < +\infty. \tag{35}$$

According to the *Lebesgue's Monotone Convergence* theorem, we know that the above result implies:

$$\sum_{t=1}^{+\infty} \mathbb{E}[\Omega_t | \mathscr{F}_{t-1}] < +\infty \text{ a.s.} \tag{36}$$

Next, we take the conditional expectation with respect to $\mathscr{F}_{t-1}$ on both sides of Eq. 34, and we obtain:

$$\mathbb{E}\left[\frac{\Pi_{\Delta,t+1}\hat{f}(u_{t+1})}{\ln^2(t+2)}\middle|\mathscr{F}_{t-1}\right] \leq \frac{\Pi_{\Delta,t}\hat{f}(u_t)}{\ln^2(t+1)} + (L_f+1)\,\mathbb{E}[\Omega_t|\mathscr{F}_{t-1}] + 0. \tag{37}$$

Based on the result from Eq. 36 and the *supermartingale convergence* theorem, we deduce that $\frac{\Pi_{\Delta,t}\hat{f}(u_t)}{\ln^2(t+1)}$ convergence almost surely. Then, according to Property 5, we can bound $f(w_t) - f^*$ using $\hat{f}(u_t)$, i.e.,

$$\begin{aligned} f(w_t) - f^* &\leq (L_f+1)(f(u_t)-f^*) + \frac{(L_f+1)\beta_1^2}{2(1-\beta_1)^2}\|\eta_{v_{t-1}} \circ m_{t-1}\|^2 + L_f f^* \\ &\leq \max\left\{L_f+1, \frac{(L_f+1)\beta_1^2}{2C_2(1-\beta_1)}\right\}\hat{f}(u_t) + L_f f^*. \end{aligned}$$

Then we can acquire our first result. Next, we take the expectation on both sides of Eq. 34, and we obtain:

$$\mathbb{E}\left[\frac{\Pi_{\Delta,t+1}\hat{f}(u_{t+1})}{\ln^2(t+2)}\right] \leq \mathbb{E}\left[\frac{\Pi_{\Delta,t}\hat{f}(u_t)}{\ln^2(t+1)}\right] + (L_f+1)\,\mathbb{E}[\Omega_t] + 0. \tag{38}$$

Based on the convergence result of the expectation summation in Eq. 35 and a simple summation formula for a recursive sequence, we obtain our second result. Thus, the case for $\delta = 0$ has been fully analyzed. For the case where $\delta > 0$, we can reach the conclusion using the same method. This completes the proof.

□

D.10 The Proof of Lemma B.10

Proof. Since the case of $\delta > 0$ is relatively straightforward, we first analyze the scenario where $\delta > 0$. According to the second conclusion for $\delta > 0$ in Lemma B.9, we easily obtain:

$$\mathbb{E}[S_T^{3/4}] = \mathbb{E}[\Pi_{\Delta,T+1}^{-3/4}\Pi_{\Delta,T+1}^{3/4}S_T^{3/4}] \overset{\textit{Hölder's}\text{ inequality}}{\leq} \mathbb{E}^{1/4}[\Pi_{\Delta,T+1}^{-3}]\,\mathbb{E}^{3/4}[\Pi_{\Delta,T+1}S_T].$$

Then according to Lemma B.3, we have $\mathbb{E}[\Pi_{\Delta,T+1}^{-3}] \leq C_{v,d,3}$. For the other term, $\mathbb{E}[\Pi_{\Delta,T+1}S_T]$, we can handle it as follows:

$$\begin{aligned} \mathbb{E}[\Pi_{\Delta,T+1}S_T] &\leq S_0 + \mathbb{E}\left[\Pi_{\Delta,T+1}\sum_{t=1}^{T}\|g_t\|^2\right] \leq dv + \mathbb{E}\left[\sum_{t=1}^{T}\Pi_{\Delta,t+1}\|g_t\|^2\right] \\ &= dv + \sum_{t=1}^{T}\mathbb{E}\left[\Pi_{\Delta,t+1}\,\mathbb{E}[\|g_t\|^2|\mathscr{F}_{t-1}]\right] \\ &\overset{\text{Property 1}}{\leq} dv + \sum_{t=1}^{T}\mathbb{E}\left[\Pi_{\Delta,t+1}((A+2L_fB)(f(w_t)-f^*)+C)\right] \end{aligned}$$

$$\overset{\text{Lemma B.9}}{\leq} dv + ((A + 2L_f B)M_\delta + C)T.$$

This implies that:

$$\mathbb{E}[\sqrt{S_T}] \leq C_{v,d,3}^{1/4}(dv + ((A + 2L_f B)M_\delta + C)T^{3/4} = \mathcal{O}(T^{3/4}).$$

For the case where $\delta = 0$, we use the same approach as in the case of $\delta > 0$ and apply the corresponding conclusion for $\delta = 0$ from Lemma B.9. Thus, we obtain:

$$\mathbb{E}[S_T^{3/4}] = \mathcal{O}(T^{3/4} \ln^{3/2} T).$$

□

D.11 The Proof of Lemma B.11

Proof. We discuss two cases based on the value of λ. In the first case, when $\lambda = 1$, we naturally have:

$$v_{t+1} = \Big(1 - \frac{1}{t+1}\Big)v_t + \frac{1}{t+1} g_t^{\circ 2} \ (\forall\, t \geq 1),$$

that is

$$(t+1)v_{t+1} = tv_t + g_t^{\circ 2}.$$

Summing over all coordinates, we obtain:

$$(t+1)\Sigma_{v_{t+1}} = t\Sigma_{v_t} + \|g_t\|^2. \tag{39}$$

Multiplying both sides of the above equation by $\Pi_{\Delta,t+1}$, and noting that $\Pi_{\Delta,t+1} \geq \Pi_{\Delta,t+2}$, we obtain:

$$(t+1)\Pi_{\Delta,t+2}\Sigma_{v_{t+1}} = t\Pi_{\Delta,t+1}\Sigma_{v_t} + \Pi_{\Delta,t+1}\|g_t\|^2.$$

Taking the expectation on both sides, we obtain:

$$\begin{aligned}
(t+1)\,\mathbb{E}[\Pi_{\Delta,t+2}\Sigma_{v_{t+1}}] &\leq t\,\mathbb{E}[\Pi_{\Delta,t+1}\Sigma_{v_t}] + \mathbb{E}[\Pi_{\Delta,t+1}\|g_t\|^2] \\
&= t\,\mathbb{E}[\Pi_{\Delta,t+1}\Sigma_{v_t}] + \mathbb{E}[\Pi_{\Delta,t+1}\,\mathbb{E}[\|g_t\|^2|\mathscr{F}_{t-1}]] \\
&\overset{\text{Property 1}}{\leq} t\,\mathbb{E}[\Pi_{\Delta,t+1}\Sigma_{v_t}] + (A + 2L_f B)\,\mathbb{E}[\Pi_{\Delta,t+1}(f(w_t) - f^*)] + C \\
&\leq t\,\mathbb{E}[\Pi_{\Delta,t+1}\Sigma_{v_t}] + (A + 2L_f B)\Big(\sup_{t\geq 1}\mathbb{E}[\Pi_{\Delta,t+1}(f(w_t) - f^*)]\Big) + C \\
&\overset{\text{Lemma B.9}}{\leq} t\,\mathbb{E}[\Pi_{\Delta,t+1}\Sigma_{v_t}] + (A + 2L_f B)M_\delta + C.
\end{aligned}$$

By iterating the above inequality, we finally obtain:

$$(t+1)\,\mathbb{E}[\Pi_{\Delta,t+2}\Sigma_{v_{t+1}}] \leq \begin{cases} \mathbb{E}[\Pi_{\Delta,2}\Sigma_{v_1}] + \big((A + 2L_f B)M_\delta + C\big)t, & \text{if } \delta \in (0,1] \\ \mathbb{E}[\Pi_{\Delta,2}\Sigma_{v_1}] + \big((A + 2L_f B)M_0 \ln^2 t + C\big)t, & \text{if } \delta = 0 \end{cases}.$$

This implies that for any $t \geq 1$, we always have:

$$\mathbb{E}[\Pi_{\Delta,t+2}\Sigma_{v_{t+1}}] \leq \begin{cases} \mathbb{E}[\Pi_{\Delta,2}\Sigma_{v_1}] + (A + 2L_f B)M_\delta + C, & \text{if } \delta \in (0,1] \\ \mathbb{E}[\Pi_{\Delta,2}\Sigma_{v_1}] + (A + 2L_f B)M_0 \ln^2 t + C, & \text{if } \delta = 0 \end{cases},$$

that is

$$\sup_{t\geq 1}\mathbb{E}[\Pi_{\Delta,t+1}\Sigma_{v_t}] < \begin{cases} \mathbb{E}[\Pi_{\Delta,2}\Sigma_{v_1}] + (A + 2L_f B)M_\delta + C, & \text{if } \delta \in (0,1] \\ \mathbb{E}[\Pi_{\Delta,2}\Sigma_{v_1}] + (A + 2L_f B)M_0 \ln^2 t + C, & \text{if } \delta = 0 \end{cases}.$$

Next, we discuss the scenario when $\lambda > 1$. In this case, we have the following inequality:

$$\Sigma_{v_{t+1}} \leq \Sigma_{v_t} + \frac{1}{(t+1)^\lambda}\|g_t\|^2.$$

We multiply both sides of the above inequality by $\Pi_{\Delta,t+1}$ and, noting its monotonicity, we obtain:

$$\Pi_{\Delta,t+2}\Sigma_{v_{t+1}} \leq \Pi_{\Delta,t+1}\Sigma_{v_t} + \frac{\Pi_{\Delta,t+1}}{(t+1)^\lambda}\|g_t\|^2. \tag{40}$$

Taking the conditional expectation with respect to $\mathscr{F}_{t-1}$ on both sides of the inequality, we have:

$$\mathbb{E}[\Pi_{\Delta,t+2}\Sigma_{v_{t+1}}|\mathscr{F}_{t-1}] \leq \Pi_{\Delta,t+1}\Sigma_{v_t} + \frac{\Pi_{\Delta,t+1}}{(t+1)^\lambda}\,\mathbb{E}[\Pi_{\Delta,t+1}\|g_t\|^2|\mathscr{F}_{t-1}].$$

According to Property 1 and Lemma B.9, we easily obtain:

$$\begin{aligned}
&\sum_{t=1}^{+\infty}\frac{\Pi_{\Delta,t+1}}{(t+1)^\lambda}\,\mathbb{E}[\Pi_{\Delta,t+1}\|g_t\|^2|\mathscr{F}_{t-1}] \\
&\leq \begin{cases} \Big((A+2L_fB)\sup_{t\geq 1}\big(\Pi_{\Delta,t+1}(f(w_t)-f^*)\big)+C\Big)\cdot\sum_{t=1}^{+\infty}\frac{1}{(t+1)^\lambda}, & \text{if } \delta\in(0,1] \\ \Big((A+2L_fB)\sup_{t\geq 1}\Big(\frac{\Pi_{\Delta,t+1}(f(w_t)-f^*)}{\ln^2 t}\Big)+C\Big)\cdot\sum_{t=1}^{+\infty}\frac{\ln^2 t}{(t+1)^\lambda}, & \text{if } \delta=0 \end{cases} \\
&< +\infty \text{ a.s.}
\end{aligned}$$

By the supermartingale convergence theorem, we easily obtain that $\Pi_{\Delta,t+1}\Sigma_{v_t}$ converges almost surely, which implies that $\sup_{t\geq 1}\Pi_{\Delta,t+1}\Sigma_{v_t} < +\infty$ a.s. According to Lemma B.3, where $\sup_{t\geq 1}\Pi^{-1}_{\Delta,t+1} < +\infty$ a.s., we can immediately deduce that $\sup_{t\geq 1}\Sigma_{v_t} < +\infty$ a.s.. Next, we prove that the expected supremum is finite. Taking the expectation on both sides of Eq. 40, we obtain:

$$\mathbb{E}\left[\Pi_{\Delta,t+2}\Sigma_{v_{t+1}}\right] \leq \mathbb{E}\left[\Pi_{\Delta,t+1}\Sigma_{v_t}\right] + \frac{1}{(t+1)^\lambda}\,\mathbb{E}[\Pi_{\Delta,t+1}\|g_t\|^2].$$

By summing the above recursive inequalities and using the results from Property 1 and Lemma B.9, we can easily prove that

$$\sup_{t\geq 1}\mathbb{E}[\Pi_{\Delta,t+1}\Sigma_{v_t}] < \begin{cases} \big((A+2L_fB)M_\delta+C\big)\sum_{t=1}^{+\infty}\frac{1}{(t+1)^\lambda}, & \text{if } \delta\in(0,1] \\ \big((A+2L_fB)M_0+C\big)\sum_{t=1}^{+\infty}\frac{\ln^2 t}{(t+1)^\lambda}, & \text{if } \delta=0 \end{cases} < +\infty.$$

With this, we complete the proof.

□

D.12 The Proof of Lemma B.12

Proof. According to the result from Lemma B.7, it is straightforward to see that when $\delta > 0$, we have:

$$\begin{aligned}
\sum_{t=2}^{T}\mathbb{E}\left[\Pi_{\Delta,t+1}\frac{\eta_{t-1}\|\nabla f(w_t)\|^2}{\sqrt{\Sigma_{v_{t-1}}+\mu}}\right] &\leq \sum_{t=2}^{T}\mathbb{E}\left[\Pi_{\Delta,t+1}\frac{\eta_{t-1}}{\sqrt{\Sigma_{v_{t-1}}+\mu}}\sum_{i=1}^{d}(\nabla_i f(w_t))^2\right] \\
&\leq \sum_{t=1}^{T}\mathbb{E}\left[\Pi_{\Delta,t+1}\sum_{i=1}^{d}\zeta_i(t)\right] \\
&< C_{4,\delta} < +\infty.
\end{aligned}$$

Next, we apply the *Lebesgue's Monotone Convergence* theorem:

$$\sum_{t=2}^{+\infty}\Pi_{\Delta,t+1}\frac{\eta_{t-1}\|\nabla f(w_t)\|^2}{\sqrt{\Sigma_{v_{t-1}}+\mu}} < +\infty \text{ a.s.}$$

Then, by combining the almost surely boundedness of $\sup_{t\geq 1}\Pi^{-1}_{\Delta,t+1}$ and $\sup_{t\geq 1}\Sigma_{v_t}$ from Lemma B.3 and Lemma B.11, we immediately obtain:

$$\begin{aligned}
\sum_{t=1}^{+\infty}\eta_t\|\nabla f(w_t)\|^2 &\leq \|\nabla f(w_1)\|^2 + \sum_{t=2}^{+\infty}\eta_{t-1}\|\nabla f(w_t)\|^2 \\
&< \|\nabla f(w_1)\|^2 + \Big(\sup_{t\geq 1}\Pi^{-3/2}_{\Delta,t+1}\Big)\cdot\Big(\sqrt{\sup_{t\geq 1}\Pi_{\Delta,t+1}\Sigma_{v_t}}+\mu\Big)\cdot\sum_{t=2}^{+\infty}\Pi_{\Delta,t+1}\frac{\eta_{t-1}\|\nabla f(w_t)\|^2}{\sqrt{\Sigma_{v_{t-1}}}+\mu} \\
&< +\infty \text{ a.s..}
\end{aligned}$$

According to the L-smooth assumption (Assumption 2.2), it is easy to see that

$$|\|\nabla f(w_t)\| - \|\nabla f(u_t)\|| \leq L_f \|w_t - u_t\| = \frac{L_f \beta_1}{1-\beta_1} \|\eta_{v_{t-1}} \circ m_{t-1}\| \leq \frac{L_f \beta_1}{(1-\beta_1) t^{\frac{1}{2}+\delta} \mu} \|m_{t-1}\|,$$

which implies

$$\begin{aligned}
\|\nabla f(u_t)\|^2 &\leq \left(\|\nabla f(w_t)\| + \frac{L_f \beta_1}{(1-\beta_1) t^{\frac{1}{2}+\delta} \mu} \|m_{t-1}\| \right)^2 \\
&\leq 2 \left(\|\nabla f(w_t)\|^2 + \frac{L_f^2 \beta_1^2}{(1-\beta_1)^2 t^{1+2\delta} \mu^2} \|m_{t-1}\|^2 \right) \\
&\leq 2\|\nabla f(w_t)\|^2 + \frac{2 L_f^2 \beta_1^2}{(1-\beta_1)^2 t^{1+2\delta} \mu^2} \left(\sup_{t \geq 1} \|m_t\| \right)^2.
\end{aligned}$$

Thid implies that

$$\sum_{t=1}^{+\infty} \eta_t \|\nabla f(u_t)\|^2 \leq 2 \sum_{t=1}^{+\infty} \eta_t \|\nabla f(w_t)\|^2 + \frac{2 L_f^2 \beta_1^2}{(1-\beta_1)^2 \mu^2} \left(\sup_{t \geq 1} \|m_t\| \right)^2 \sum_{t=1}^{+\infty} \frac{1}{t^{1+2\delta}}$$
$$\overset{\text{Lemma B.9 and } \delta > 0}{<} +\infty \text{ a.s.}$$

With this, we complete the proof.

□

E Proofs of Theorems

E.1 The Proof of Theorem 3.1

Proof. According to Lemma B.7, we have:

$$\sum_{t=1}^{T} \mathbb{E} \left[\Pi_{\Delta,t+1} \sum_{i=1}^{d} \zeta_i(t) \right] \leq \begin{cases} C_{4,\delta}, & \text{if } \delta \in (0,1] \\ C_5 + C_6 \mathbb{E} \left[\ln(S_T) \right], & \text{if } \delta = 0 \end{cases}.$$

According to the monotonicity of $\eta_{v_t,i}$ in Property 2 and the monotonicity of $\Pi_{\Delta,t}$ itself, we obtain the following inequality:

$$\sum_{t=1}^{T} \mathbb{E} \left[\Pi_{\Delta,T+1} \frac{\|\nabla f(w_t)\|^2}{T^{\frac{1}{2}+\delta}(\sqrt{v_T} + \mu)} \right] \leq \sum_{t=1}^{T} \mathbb{E} \left[\Pi_{\Delta,t+1} \sum_{i=1}^{d} \zeta_i(t) \right] \leq \begin{cases} C_{4,\delta}, & \text{if } \delta \in (0,1] \\ C_5 + C_6 \mathbb{E} \left[\ln(S_T) \right], & \text{if } \delta = 0 \end{cases}.$$

For the leftmost part of the above inequality, we apply the *Cauchy-Schwarz* inequality and obtain:

$$\mathbb{E} \left[\Pi_{\Delta,T+1}^{-1} T^{\frac{1}{2}+\delta} (\sqrt{v_T} + \mu) \right] \left(\sum_{t=1}^{T} \mathbb{E} \left[\Pi_{\Delta,T+1} \frac{\|\nabla f(w_t)\|^2}{T^{\frac{1}{2}+\delta}(\sqrt{v_T} + \mu)} \right] \right) \geq \sum_{t=1}^{T} \mathbb{E}[\|\nabla f(w_t)\|],$$

which means

$$\sum_{t=1}^{T} \mathbb{E}[\|\nabla f(w_t)\|] \leq \begin{cases} C_{4,\delta} \mathbb{E} \left[\Pi_{\Delta,T+1}^{-1} T^{\frac{1}{2}+\delta} (\sqrt{v_T} + \mu) \right], & \text{if } \delta \in (0,1] \\ C_5 \mathbb{E} \left[\Pi_{\Delta,T+1}^{-1} T^{\frac{1}{2}+\delta} (\sqrt{v_T} + \mu) \right] + C_6 \mathbb{E} \left[\Pi_{\Delta,T+1}^{-1} T^{\frac{1}{2}+\delta} (\sqrt{v_T} + \mu) \right] \mathbb{E} \left[\ln(S_T) \right], & \text{if } \delta = 0 \end{cases}.$$

Combining the results from Lemma B.10, Lemma B.11 and Lemma B.3, we obtain:

$$\begin{aligned}
\mathbb{E} \left[\Pi_{\Delta,T+1}^{-1} T^{\frac{1}{2}+\delta} (\sqrt{v_T} + \mu) \right] &\leq 2 T^{\frac{1}{2}+\delta} \sqrt{\mathbb{E}[\Pi_{\Delta,T+1}^{-3}]} \sqrt{\mathbb{E}[\Pi_{\Delta,T+1}(v_T + \mu^2)]} \\
&\leq \begin{cases} C_{v,d,3}^{1/2} \mathcal{O}(T^{\frac{1}{2}+\delta}), & \text{if } \gamma > 1 \\ C_{v,d,3}^{1/2} \mathcal{O}(T^{\frac{1}{2}+\delta}), & \text{if } \gamma = 1,\ \delta \in (0,1] \\ C_{v,d,3}^{1/2} \mathcal{O}(\sqrt{T} \ln T) & \text{if } \gamma = 1,\ \delta = 0 \end{cases}.
\end{aligned}$$

and

$$\mathbb{E}[\ln(S_T)] = \frac{4}{3}\,\mathbb{E}[\ln(S_T^{3/4})] \leq \frac{4}{3}\ln(\mathbb{E}[S_T^{3/4}]) = \begin{cases} \mathcal{O}(\ln T), & \text{if } \delta \in (0,1] \\ \mathcal{O}(\ln T) + \mathcal{O}(\ln\ln T), & \text{if } \delta = 0 \end{cases}.$$

Combining the two estimates above, we finally obtain:

$$\sum_{t=1}^{T} \mathbb{E}[\|\nabla f(w_t)\|] \leq \begin{cases} \mathcal{O}(T^{\frac{1}{2}+\delta}), & \text{if } \delta \in (0,1] \\ \mathcal{O}(\sqrt{T}\ln T), & \text{if } \gamma > 1,\ \ \delta = 0 \\ \mathcal{O}(\sqrt{T}\ln^2 T), & \text{if } \gamma = 1,\ \ \delta = 0 \end{cases}.$$

that is

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\|\nabla f(w_t)\|] \leq \begin{cases} \mathcal{O}\Big(\frac{1}{T^{\frac{1}{2}-\delta}}\Big), & \text{if } \delta \in (0,1] \\ \mathcal{O}\Big(\frac{\ln T}{\sqrt{T}}\Big), & \text{if } \gamma > 1,\ \ \delta = 0 \\ \mathcal{O}\Big(\frac{\ln^2 T}{\sqrt{T}}\Big), & \text{if } \gamma = 1,\ \ \delta = 0 \end{cases}.$$

With this, we complete the proof. □

E.2 The Proof of Lemma B.8

Proof. From Theorem 3.1, we have

$$\lim_{T\to+\infty} \min_{1\leq t\leq T} \mathbb{E}\,\|\nabla f(w_t)\| \leq \lim_{T\to+\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\,\|\nabla f(w_t)\| = 0.$$

This implies that there exists a subsequence $\{w_{d_t}\}_{t\geq 1}$ of $\{w_t\}_{t\geq 1}$ such that

$$\lim_{t\to+\infty} \mathbb{E}\,\|\nabla f(w_{d_t})\| = 0.$$

By the *Riesz* theorem, we can find a further subsequence $\{w_{c_t}\}_{t\geq 1}$ from $\{w_{d_t}\}_{t\geq 1}$ such that

$$\lim_{t\to+\infty} \|\nabla f(w_{c_t})\| = 0 \quad \text{a.s.}$$

This completes the proof. □

E.3 The Proof of Theorem 3.2

Proof. According to the L-smooth assumption (Assumption 2.2), it is easy to see that

$$|\|\nabla f(w_t)\| - \|\nabla f(u_t)\|| \leq L_f\|w_t - u_t\| = \frac{L_f\beta_1}{1-\beta_1}\|\eta_{v_{t-1}} \circ m_{t-1}\| \leq \frac{L_f\beta_1}{(1-\beta_1)t^{\frac{1}{2}+\delta}\mu}\|m_{t-1}\|.$$

According to Lemma B.9, we know that $\sup_{t\geq 1}\|m_{t-1}\| < +\infty$ a.s.. This implies that

$$\lim_{t\to+\infty} |\|\nabla f(w_t)\| - \|\nabla f(u_t)\|| \leq \Big(\sup_{t\geq 1}\|m_{t-1}\|\Big) \cdot \lim_{t\to+\infty} \frac{L_f\beta_1}{(1-\beta_1)t^{\frac{1}{2}+\delta}\mu} = 0 \ \text{ a.s.}$$

This implies that we only need to prove $\lim_{t\to+\infty}\|\nabla f(u_t)\| = 0$ a.s.. To achieve this objective, we proceed as follows.

For any $l > 0$, we construct the following stopping time [3] sequence $\{\tau_{l,n}\}_{n\geq 1}$:

$\tau_{l,1} := \min\{t \geq 1 : \|\nabla f(u_t)\| > l\},\ \ \tau_{l,2} := \min\{t > \tau_{l,1} : \|\nabla f(u_t)\| \leq l\},$

...,

$\tau_{l,2k-1} := \min\{t > \tau_{l,2k-2} : \|\nabla f(u_t)\| > l\},\ \ \tau_{l,2k} := \min\{t > \tau_{l,2k-1} : \|\nabla f(u_t)\| \leq l\}.$

[3] In this paper, we adopt the following definition of stopping time: Let τ be a random variable defined on the filtered probability space $(\Omega, \mathscr{F}, (\mathscr{F}_n)_{n\in\mathbb{N}}, \mathbb{P})$ with values in $\mathbb{N}\cup\{+\infty\}$. Then τ is called a stopping time (with respect to the filtration $(\mathscr{F}_n)_{n\in\mathbb{N}}$) if the following condition holds: $\{\tau = n\} \in \mathscr{F}_n$ for all n.

According to the subsequence convergence result in Lemma B.8, we know that when $\tau_{2k-1} < +\infty$ $(\forall\, k \geq 1)$, it must hold that $\tau_{2k} < +\infty$ a.s.. We now discuss two cases:

1. When there exists some $k_0 \geq 1$ such that $\tau_{2k_0-1} = +\infty$, this implies that eventually $\{\|\nabla f(u_t)\|\}_{t\geq 1}$ will remain below l, i.e.,

$$\limsup_{t\to+\infty} \|\nabla f(u_t)\| < l. \tag{41}$$

2. Next, we focus on the second case, where for all τ_{2k-1}, we have $\tau_{2k-1} < +\infty$. In this situation, we examine the behavior of $\sup_{\tau_{2k-1}\leq t<\tau_{2k}} \|\nabla f(u_t)\|$. It is easy to see that:

$$\begin{aligned}
\sup_{\tau_{2k-1}\leq t<\tau_{2k}} \|\nabla f(u_t)\| &\leq l + \sup_{\tau_{2k-1}\leq t<\tau_{2k}} \|\nabla f(u_t)\| - \|\nabla f(u_{\tau_{2k-1}-1})\| \\
&\leq l + \left(\sum_{t=\tau_{2k-1}-1}^{\tau_{2k}-1} \big| \|\nabla f(u_t)\| - \|\nabla f(u_{t-1})\| \big| \right) \\
&\overset{\text{l-smooth}}{\leq} l + \left(L_f \sum_{t=\tau_{2k-1}-1}^{\tau_{2k}-1} \|u_t - u_{t-1}\| \right) \\
&\overset{\text{Eq. 4}}{\leq} l + L_f \underbrace{\left(\sum_{t=\tau_{2k-1}-1}^{\tau_{2k}-1} \|\eta_{v_t} \circ g_t\| \right)}_{\Upsilon_{k,1}} \\
&\quad + \frac{\beta_1 L_f^2}{1-\beta_1} \underbrace{\left(\sum_{t=\tau_{2k-1}-1}^{\tau_{2k}-1} \|\Delta_t \circ m_{t-1}\| \right)}_{\Upsilon_{k,2}}.
\end{aligned}$$

Our next goal is to prove separately that $\limsup_{k\to+\infty} \Upsilon_{k,1} = 0$ a.s. and $\limsup_{k\to+\infty} \Upsilon_{k,2} = 0$ a.s.. For $\Upsilon_{k,1}$, we have:

$$\begin{aligned}
\Upsilon_{k,1} &= \left(\sum_{t=\tau_{2k-1}-1}^{\tau_{2k}-1} \|\eta_{v_t} \circ g_t\| \right) = \left(\sum_{t=\tau_{2k-1}-1}^{\tau_{2k}-1} \sum_{i=1}^{d} \eta_{v_t,i} |g_{t,i}| \right) \\
&= \left(\sum_{t=\tau_{2k-1}-1}^{\tau_{2k}-1} \sum_{i=1}^{d} \frac{\eta_t |g_{t,i}|}{\sqrt{v_t}+\mu} \right) \leq \left(\sum_{t=\tau_{2k-1}-1}^{\tau_{2k}-1} \sum_{i=1}^{d} \frac{\eta_t |g_{t,i}|}{\mu} \right) \\
&= \frac{1}{\mu} \left(\sum_{t=\tau_{2k-1}-1}^{\tau_{2k}-1} \sum_{i=1}^{d} \eta_t\, \mathbb{E}[|g_{t,i}||\mathscr{F}_{t-1}] \right) \\
&\quad + \frac{1}{\mu} \left(\sum_{t=\tau_{2k-1}-1}^{\tau_{2k}-1} \sum_{i=1}^{d} \eta_t (|g_{t,i}| - \mathbb{E}[|g_{t,i}||\mathscr{F}_{t-1}]) \right) \\
&= \frac{1}{\mu} \underbrace{\left(\sum_{t=\tau_{2k-1}-1}^{\tau_{2k}-1} \eta_t \mathbb{E}[|g_t||\mathscr{F}_{t-1}] \right)}_{\Upsilon_{k,1,1}} + \frac{1}{\mu} \underbrace{\left(\sum_{t=\tau_{2k-1}-1}^{\tau_{2k}-1} \sum_{i=1}^{d} \eta_t (|g_{t,i}| - \mathbb{E}[|g_{t,i}||\mathscr{F}_{t-1}]) \right)}_{\Upsilon_{k,1,2}}.
\end{aligned}$$

For $\Upsilon_{k,1,1}$, we have:

$$\begin{aligned}
\Upsilon_{k,1,1} &\overset{\text{Property 1}}{\leq} \left(\sum_{t=\tau_{2k-1}-1}^{\tau_{2k}-1} \eta_t \big((A + 2L_f B)(f(w_t) - f^*) + C \big) \right) \\
&\leq \big((A + 2L_f B) \sup_{t\geq 1} (f(w_t) - f^*) + C \big) \cdot \left(\sum_{t=\tau_{2k-1}-1}^{\tau_{2k}-1} \eta_t \right)
\end{aligned}$$

$$= \big((A+2L_fB)\sup_{t\geq 1}(f(w_t)-f^*)+C\big)\cdot\left(\eta_{\tau_{2k-1}}+\left(\sum_{t=\tau_{2k-1}-1}^{\tau_{2k}}\eta_t\right)\right)$$

$$\overset{(a)}{\leq} \frac{1}{l^2}\big((A+2L_fB)\sup_{t\geq 1}(f(w_t)-f^*)+C\big)\cdot\left(\eta_{\tau_{2k-1}}+\left(\sum_{t=\tau_{2k-1}}^{\tau_{2k}-1}\eta_t\|\nabla f(u_t)\|^2\right)\right).$$

In step (a), this is due to the fact that, over the interval $[\tau_{2k-1},\tau_{2k})$, we always have $\|\nabla f(u_t)\|^2 > l^2$. Based on Lemma B.12, we know that

$$\sum_{t=1}^{+\infty}\eta_t\|\nabla f(u_t)\|^2 < +\infty \quad \text{a.s.}$$

By applying the *Cauchy's* convergence principle, we can prove that

$$\lim_{k\to+\infty}\sum_{t=\tau_{2k-1}}^{\tau_{2k}-1}\eta_t\|\nabla f(u_t)\|^2 = 0 \quad \text{a.s.}$$

On the other hand, it is evident that $\lim_{k\to+\infty}\eta_{\tau_{2k-1}} = 0$. Meanwhile, based on Lemma B.9 and Lemma B.3, we can easily prove that

$$\sup_{t\geq 1}(f(w_t)-f^*) \leq \Big(\sup_{t\geq 1}\Pi_{\Delta,t+1}^{-1}\Big)\cdot\Big(\sup_{t\geq 1}\big(\Pi_{\Delta,t+1}(f(w_t)-f^*)\big)\Big) < +\infty \quad \text{a.s.}$$

Therefore, we have proven that

$$\limsup_{k\to+\infty}\Upsilon_{k,1,1} = \lim_{k\to+\infty}\Upsilon_{k,1,1} = 0.$$

For $\Upsilon_{k,1,2}$, we consider the following martingale difference sequence:

$$\overline{X}_T := \sum_{t=1}^{T}\sum_{i=1}^{d}\eta_t(|g_{t,i}| - \mathbb{E}[|g_{t,i}| \mid \mathscr{F}_{t-1}]).$$

We can compute

$$\sum_{t=1}^{+\infty}\mathbb{E}\left[\left(\sum_{i=1}^{d}\eta_t(|g_{t,i}| - \mathbb{E}[|g_{t,i}| \mid \mathscr{F}_{t-1}])\right)^2\middle|\mathscr{F}_{t-1}\right] \leq d\sum_{t=1}^{+\infty}\eta_t^2\sum_{i=1}^{d}\mathbb{E}[(|g_{t,i}| - \mathbb{E}[|g_{t,i}| \mid \mathscr{F}_{t-1}])^2]$$

$$\leq d\sum_{t=1}^{+\infty}\eta_t^2\sum_{i=1}^{d}\mathbb{E}[\|g_t\|^2 \mid \mathscr{F}_{t-1}]$$

$$\overset{\text{Property 1}}{\leq} d\sum_{t=1}^{+\infty}\eta_t^2\sum_{i=1}^{d}\Big((A+2L_fB)\sup_{t\geq 1}(f(w_t)-f^*)+C\Big)$$

$$\leq d\sum_{i=1}^{d}\left((A+2L_fB)\Big(\sup_{t\geq 1}\Pi_{\Delta,t}\Big)\Big(\sup_{t\geq 1}(f(w_t)-f^*)+C\Big)\right)\cdot\sum_{t=1}^{+\infty}\eta_t^2$$

$$\overset{\text{Lemma B.3 and B.9}}{<} +\infty \quad \text{a.s.}$$

By the *Martingale Convergence* theorem, we obtain

$$\lim_{T\to+\infty}\overline{X}_T = \sum_{t=1}^{+\infty}\sum_{i=1}^{d}\eta_t(|g_{t,i}| - \mathbb{E}[|g_{t,i}| \mid \mathscr{F}_{t-1}]) < +\infty \quad \text{a.s.}$$

Using the *Cauchy's Convergence* principle, we can easily prove that

$$\limsup_{k\to+\infty}\Upsilon_{k,1,2} = \lim_{k\to+\infty}\sum_{t=\tau_{2k-1}-1}^{\tau_{2k}-1}\sum_{i=1}^{d}\eta_t(|g_{t,i}| - \mathbb{E}[|g_{t,i}||\mathscr{F}_{t-1}]) = 0 \quad \text{a.s.}$$

Combining the above two limit proofs for $\Upsilon_{t,1,1}$ and $\Upsilon_{t,1,2}$, we can conclude that

$$\limsup_{k\to+\infty} \Upsilon_{t,1} = 0 \quad \text{a.s.}$$

Next, we will demonstrate that $\lim_{k\to+\infty} \Upsilon_{k,2} = 0$ a.s.. This is relatively easy; we only need to examine the series sum

$$\sum_{t=1}^{+\infty} \|\Delta_t \circ m_{t-1}\|,$$

which can easily be proven to satisfy

$$\sum_{t=1}^{+\infty} \|\Delta_t \circ m_{t-1}\| = \sum_{t=1}^{+\infty}\sum_{i=1}^{d} \Delta_{t,i}^{-1}|m_{t-1,i}| < \sqrt{d}\Big(\sup_{t\geq 1}\Pi_{\Delta,t}\Big)\cdot\Big(\sup_{t\geq 1}\Pi_{\Delta,t}\|m_{t-1}\|\Big)\cdot\sum_{i=1}^{d}\Delta_{t,i} \overset{\text{Lemma B.3 and B.9}}{<} +\infty \quad \text{a.s.}$$

Thus, by the *Cauchy's Convergence* principle, we obtain

$$\limsup_{k\to+\infty} \Upsilon_{k,2} = \lim_{k\to+\infty} \sum_{t=\tau_{2k-1}-1}^{\tau_{2k}-1}\sum_{i=1}^{d} \eta_t(|g_{t,i}| - \mathbb{E}[|g_{t,i}||\mathscr{F}_{t-1}]) = 0 \quad \text{a.s.}$$

Combining the limit results for $\Upsilon_{k,1}$ and $\Upsilon_{k,2}$, we conclude that

$$\limsup_{k\to+\infty} \sup_{\tau_{2k-1}\leq t<\tau_{2k}} \|\nabla f(u_t)\| \leq l + 0 = l.$$

Moreover, combining $\sup_{\tau_{2k}\leq t<\tau_{2k+1}} \|\nabla f(u_t)\| < l$, we can deduce that

$$\limsup_{t\to+\infty} \|\nabla f(u_t)\| \leq l \quad \text{a.s.}$$

Then, due to the arbitrariness of l, we conclude that

$$\limsup_{t\to+\infty} \|\nabla f(u_t)\| = 0 \quad \text{a.s.}$$

This implies that

$$\lim_{t\to+\infty} \|\nabla f(u_t)\| = 0 \quad \text{a.s.}$$

Thus, we complete the proof.

□

E.4 The Proof of Theorem 3.3

Proof. Since we have already proven almost sure convergence in Theorem 3.2, it is natural to attempt to prove L_1 convergence via the *Lebesgue's Dominated Convergence* theorem. To achieve this, we need to find a function h that is $\mathscr{F}_\infty$-measurable and satisfies $\mathbb{E}\,|h| < +\infty$, and such that for all $t \geq 1$, we have $\|\nabla f(w_t)\| \leq |h|$. Since for all t, we naturally have $\|\nabla f(w_t)\| \leq \sup_{k\geq 1}\|\nabla f(w_k)\|$, we only need to prove that $\mathbb{E}[\sup_{k\geq 1}\|\nabla f(w_k)\|] < +\infty$. We proceed to achieve this goal.

Returning to the *Approximate Descent Inequality* (Lemma 4.1), we have:

$$\begin{aligned}\Pi_{\Delta,t+1}\hat{f}(u_{t+1}) - \Pi_{\Delta,t}\hat{f}(u_t) \leq &-\frac{5}{16}\Pi_{\Delta,t+1}\sum_{i=1}^{d}\zeta_i(t) + (L_f+1)\Pi_{\Delta,t+1}\sum_{i=1}^{d}\eta_{v_t,i}^2 g_{t,i}^2 \\ &+ \Pi_{\Delta,t+1}M_t. \end{aligned} \tag{42}$$

For any $\lambda > 0$, define the stopping time τ_λ as the first time the sequence $\{\Pi_{\Delta,t}\hat{f}(u_t)\}_{t\geq 1}$ exceeds λ, i.e.,

$$\tau_\lambda := \min\{t \geq 2 : \Pi_{\Delta,t}\hat{f}(u_t) > \lambda\}.$$

It can be rigorously verified that τ_λ is a stopping time with respect to the filtration $\{\mathscr{F}_t\}_{t\geq 1}$, and satisfies a special property $[\tau_\lambda = n] \in \mathscr{F}_{n-1}$ for all $n \geq 1$. This implies that the preceding time

$\tau_\lambda - 1$ is also a stopping time. Next, for any deterministic time $T \geq 3$, we define $\tau_{\lambda,T} := \tau_\lambda \wedge T$. We then sum the indices of Eq. 42 from 1 to $\tau_{\lambda,T} - 1$. Specifically, we have:

$$\Pi_{\Delta,\tau_{\lambda,T}}\hat{f}(u_{\tau_{\lambda,T}}) \leq \Pi_{\Delta,1}\hat{f}(u_1) + (L_f+1)\sum_{t=1}^{\tau_{\lambda,T}-1}\Pi_{\Delta,t+1}\sum_{i=1}^{d}\eta_{v_t,i}^2 g_{t,i}^2 + \sum_{t=1}^{\tau_{\lambda,T}-1}\Pi_{\Delta,t+1}M_t.$$

Taking the expectation on both sides, we obtain:

$$\begin{aligned}\mathbb{E}\left[\Pi_{\Delta,\tau_{\lambda,T}}\hat{f}(u_{\tau_{\lambda,T}})\right] \leq \mathbb{E}\left[\Pi_{\Delta,1}\hat{f}(u_1)\right] + (L_f+1)\,\mathbb{E}\left[\sum_{t=1}^{\tau_{\lambda,T}-1}\Pi_{\Delta,t+1}\sum_{i=1}^{d}\eta_{v_t,i}^2 g_{t,i}^2\right] \\ + \mathbb{E}\left[\sum_{t=1}^{\tau_{\lambda,T}-1}\Pi_{\Delta,t+1}M_t\right].\end{aligned}$$

Since $\{\Pi_{\Delta,t+1}M_t, \mathscr{F}_{t-1}\}_{t\geq 1}$ is a martingale difference sequence and $\tau_{\lambda,T} \leq T < +\infty$, by *Doob's Stopped* theorem, we know that:

$$\mathbb{E}\left[\sum_{t=1}^{\tau_{\lambda,T}-1}\Pi_{\Delta,t+1}M_t\right] = 0.$$

This implies that:

$$\begin{aligned}\mathbb{E}\left[\Pi_{\Delta,\tau_{\lambda,T}}\hat{f}(u_{\tau_{\lambda,T}})\right] &\leq \mathbb{E}\left[\Pi_{\Delta,1}\hat{f}(u_1)\right] + (L_f+1)\,\mathbb{E}\left[\sum_{t=1}^{\tau_{\lambda,T}-1}\Pi_{\Delta,t+1}\sum_{i=1}^{d}\eta_{v_t,i}^2 g_{t,i}^2\right] \\ &< \mathbb{E}\left[\Pi_{\Delta,1}\hat{f}(u_1)\right] + (L_f+1)\,\mathbb{E}\left[\sum_{t=1}^{T}\Pi_{\Delta,t+1}\sum_{i=1}^{d}\eta_{v_t,i}^2 g_{t,i}^2\right].\end{aligned}$$

According to the estimates of $\sum_{t=1}^{T}\mathbb{E}[\Gamma_t]$ from Eq. 32 to Eq. 33, we easily obtain that when $\delta > 1$, the following holds:

$$\mathbb{E}\left[\Pi_{\Delta,\tau_{\lambda,T}}\hat{f}(u_{\tau_{\lambda,T}})\right] \leq \overline{M} < +\infty,$$

where

$$\overline{M} := \mathbb{E}\left[\Pi_{\Delta,1}\hat{f}(u_1)\right] + \frac{2^{2\delta}d(L_f+1)}{\alpha_1 v^{\frac{\delta}{2}}}C_\zeta^\delta.$$

On the other hand, we easily observe the following event decomposition:

$$\left[\sup_{2\leq t<T}\Pi_{\Delta,t}\hat{f}(u_t) > \lambda\right] = \bigcup_{k=2}^{T-1}[\tau_\lambda = k] = \bigcup_{k=2}^{T-1}[\tau_{\lambda,T} = k].$$

Moreover, since for any $j \neq k$, we have $[\tau_{\lambda,T} = j] \cap [\tau_{\lambda,T} = k] = \emptyset$, it follows that:

$$\begin{aligned}\mathbb{P}\left[\sup_{2\leq t<T}\Pi_{\Delta,t}\hat{f}(u_t) > \lambda\right] = \sum_{k=2}^{T-1}\mathbb{P}[\tau_{\lambda,T} = k] &\overset{\text{Markov's inequality}}{\leq} \frac{1}{\lambda}\sum_{k=2}^{T-1}\mathbb{E}\left[\Pi_{\Delta,k}\hat{f}(u_k)\mathbb{I}_{[\tau_{\lambda,T}=k]}\right] \\ &< \frac{1}{\lambda}\,\mathbb{E}\left[\Pi_{\Delta,\tau_{\lambda,T}}\hat{f}(u_{\tau_{\lambda,T}})\right] \leq \frac{\overline{M}}{\lambda}. \end{aligned} \tag{43}$$

Next, for any $K \geq 1$, we compute $\mathbb{E}\left[\left(\sup_{2\leq t<T}\Pi_{\Delta,t}\hat{f}(u_t)\right)^{3/4} \wedge K\right]$. We have

$$\begin{aligned}\mathbb{E}\left[\left(\sup_{2\leq t<T}\Pi_{\Delta,t}\hat{f}(u_t)\right)^{3/4} \wedge K\right] &= -\int_0^{+\infty} x \;\mathrm{d}\left(\mathbb{P}\left[\left(\sup_{2\leq t<T}\Pi_{\Delta,t}\hat{f}(u_t)\right)^{3/4} \wedge K > x\right]\right) \\ &= -\int_0^{+\infty}\left(\int_0^x 1\mathrm{d}\lambda\right)\mathrm{d}\left(\mathbb{P}\left[\left(\sup_{2\leq t<T}\Pi_{\Delta,t}\hat{f}(u_t)\right)^{3/4} \wedge K > x\right]\right) \\ &\overset{\text{Fubini's theorem}}{=} -\int_0^{+\infty}\left(\int_\lambda^{+\infty} 1\mathrm{d}\left(\mathbb{P}\left[\left(\sup_{2\leq t<T}\Pi_{\Delta,t}\hat{f}(u_t)\right)^{3/4} \wedge K > x\right]\right)\right)\mathrm{d}\lambda\end{aligned}$$

$$
\begin{aligned}
&= \int_0^{+\infty} \mathbb{P}\left[\left(\sup_{2\le t<T} \Pi_{\Delta,t}\hat{f}(u_t)\right)^{3/4} \wedge K > \lambda\right] \mathrm{d}\lambda \\
&\le 1 + \int_1^{+\infty} \mathbb{P}\left[\left(\sup_{2\le t<T} \Pi_{\Delta,t}\hat{f}(u_t)\right)^{3/4} \wedge K > \lambda\right] \mathrm{d}\lambda \\
&= 1 + \int_1^{+\infty} \mathbb{P}\left[\left(\sup_{2\le t<T} \Pi_{\Delta,t}\hat{f}(u_t)\right) \wedge K^{4/3} > \lambda^{4/3}\right] \mathrm{d}\lambda \\
&< 1 + \int_1^{+\infty} \mathbb{P}\left[\left(\sup_{2\le t<T} \Pi_{\Delta,t}\hat{f}(u_t)\right) > \lambda^{4/3}\right] \mathrm{d}\lambda \\
&\overset{\text{Eq. 43}}{<} 1 + \int_1^{+\infty} \frac{\overline{M}}{\lambda^{4/3}} \mathrm{d}\lambda \\
&= 1 + 3\overline{M}.
\end{aligned}
$$

Next, we take $K \to +\infty$ and apply the *Lebesgue's Monotone Convergence* theorem. We get:

$$
\mathbb{E}\left[\left(\sup_{2\le t<T} \Pi_{\Delta,t}\hat{f}(u_t)\right)^{3/4}\right] \le 1 + 3\overline{M}.
$$

Next, by taking $T \to +\infty$ and applying the *Lebesgue's Monotone Convergence* theorem once again, we obtain:

$$
\mathbb{E}\left[\left(\sup_{t\ge 2} \Pi_{\Delta,t}\hat{f}(u_t)\right)^{3/4}\right] \le 1 + 3\overline{M}.
$$

Note that for any finite t, we have $\Pi_{\Delta,t+1} \ge \Pi_{\Delta,\infty}$ (where $\Pi_{\Delta,\infty}$ is defined in Lemma B.3). Thus, we have:

$$
\mathbb{E}\left[\Pi_{\Delta,\infty}^{3/4}\left(\sup_{t\ge 2} \hat{f}(u_t)\right)^{3/4}\right] \le \mathbb{E}\left[\left(\sup_{t\ge 2} \Pi_{\Delta,t}\hat{f}(u_t)\right)^{3/4}\right] \le 1 + 3\overline{M}.
$$

Next, by applying *Hölder's* inequality, we obtain:

$$
\mathbb{E}\left[\left(\sup_{t\ge 2} \hat{f}(u_t)\right)^{1/2}\right] \le \mathbb{E}^{1/3}\left[\Pi_{\Delta,\infty}^{-3/2}\right] \mathbb{E}^{2/3}\left[\Pi_{\Delta,\infty}^{3/4}\left(\sup_{t\ge 2} \hat{f}(u_t)\right)^{3/4}\right] \overset{\text{Lemma B.3}}{\le} C_{v,d,3/2}^{1/3}(1+3\overline{M})^{2/3}.
$$

Then, according to Property 5, we can bound $f(w_t) - f^*$ using $\hat{f}(u_t)$, i.e.,

$$
\begin{aligned}
f(w_t) - f^* &\le (L_f + 1)(f(u_t) - f^*) + \frac{(L_f+1)\beta_1^2}{2(1-\beta_1)^2}\|\eta_{v_{t-1}} \circ m_{t-1}\|^2 + L_f f^* \\
&\le \max\left\{L_f + 1, \frac{(L_f+1)\beta_1^2}{2C_2(1-\beta_1)}\right\}\hat{f}(u_t) + L_f f^*.
\end{aligned}
$$

That means

$$
\mathbb{E}\left[\left(\sup_{t\ge 2}\left(f(w_t) - f^*\right)\right)^{1/2}\right] \le \left(\max\left\{L_f + 1, \frac{(L_f+1)\beta_1^2}{2C_2(1-\beta_1)}\right\}\right)^{1/2} C_{v,d,3/2}^{1/3}(1+3\overline{M})^{2/3} + \sqrt{L_f|f^*|}.
$$

Finally, according to Lemma B.1, we obtain:

$$
\begin{aligned}
\mathbb{E}\left[\sup_{t\ge 2}\|\nabla f(w_t)\|\right] &\le \sqrt{2L_f}\,\mathbb{E}\left[\left(\sup_{t\ge 2}\left(f(w_t) - f^*\right)\right)^{1/2}\right] \\
&< \sqrt{2L_f}\left(\max\left\{L_f + 1, \frac{(L_f+1)\beta_1^2}{2C_2(1-\beta_1)}\right\}\right)^{1/2} C_{v,d,3/2}^{1/3}(1+3\overline{M})^{2/3} \\
&\quad + \sqrt{2L_f^2|f^*|}.
\end{aligned}
$$

By adding the first term, we obtain:

$$
\mathbb{E}\left[\sup_{t\ge 1}\|\nabla f(w_t)\|\right] < \|\nabla f(w_1)\| + \sqrt{2L_f}\left(\max\left\{L_f + 1, \frac{(L_f+1)\beta_1^2}{2C_2(1-\beta_1)}\right\}\right)^{1/2} C_{v,d,3/2}^{1/3}(1+3\overline{M})^{2/3}
$$

$$+ \sqrt{2L_f^2|f^*|} < +\infty.$$

Finally, combining the almost sure convergence result from Theorem 3.2 with the *Lebesgue's Dominated Convergence* theorem, we obtain the L_1 convergence result, namely:

$$\lim_{t \to +\infty} \mathbb{E}[\|\nabla f(w_t)\|] = 0.$$

With this, we complete the proof. □