MOMENTUM MAKES (ADAPTIVE) SUBGRADIENT METHODS OPTIMAL FOR STOCHASTIC NONSMOOTH NONCONVEX OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

SGDM (stochastic gradient descent with momentum) consistently outperforms vanilla SGD across a wide spectrum of modern deep learning tasks, including the training of residual networks and Transformer-based large-scale models. Yet, theoretical understanding of this performance advantage remains incomplete, especially for the nonsmooth and nonconvex settings that dominate real-world deep learning applications. From an online learning perspective, a line of recent studies [16; 26; 57] has established that carefully modified SGDM variants attain optimal convergence rates for locating Goldstein-style stationary points, thereby providing theoretical support for the empirical benefits of momentum acceleration. Despite these fruitful advances, existing studies heavily depend on additional algorithmic modifications such as model randomization or momentum clipping that deviate notably from practical SGDM implementations. In this paper, we pursue an alternative approach to directly analyze the convergence of standard SGDM over a broad family of weakly convex optimization problems. Equipped with squared-geometric-CDF randomized output, we prove that SGDM achieves optimal convergence rates under a mildly relaxed notion of Goldstein stationarity. Our results are valid for constant, decaying, and *adaptive* learning rate schedules alike. Furthermore, as a secondary contribution, our analysis extends to a slightly modified SGDM with one-time momentum reset, yielding novel theoretical insights into its recently observed empirical performance in training large language models and deep reinforcement learning.

1 INTRODUCTION

The main focus of our study is the analysis of stochastic (sub-)gradient descent with momentum (SGDM) and its variants for solving nonsmooth nonconvex machine learning optimization problems. Formally, given a non-negative loss function $\ell : \mathbb{R}^d \times \mathcal{Z} \mapsto \mathbb{R}^+$, the goal is to minimize the following expected risk function:

$$\min_{w \in \mathbb{R}^d} f(w) := \mathbb{E}_{Z \sim \mathcal{D}}[\ell(w; Z)], \tag{1}$$

where $\mathcal{D}$ is a (typically unknown) distribution of data over a measurable space $\mathcal{Z}$. We are interested in the setting where the loss ℓ is Lipschitz continuous with respect to its first argument, and the risk function f belongs to a class of ρ-weakly convex functions, meaning that the mapping $w \mapsto f(w) + \frac{\rho}{2}\|w\|^2$ is convex. This function class is very rich and easy to identify in practice as it encompasses all functions of a composition form $f(w) = h(c(w))$ where h is a convex and Lipschitz function and c is a smooth mapping. Weakly convex optimization has been widely applied across machine learning and signal processing. Representative applications span deep neural network training [9; 26], reinforcement learning [51], robust phase retrieval [17], robust and optimal control [42; 52], and low-rank matrix recovery [12], to name a few.

Stochastic gradient descent (SGD) is the *de facto* algorithm for solving the problem 1. Based on random samples drawn from the underlying data distribution, SGD iteratively updates model parameters in the negative (sub-)gradient direction and scales effectively to large-scale settings. Since its introduction [45], numerous variants of SGD have been developed to enhance both its empirical and theoretical performance. Among these variants, an arguably most popular one is SGD with

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

momentum (SGDM) which uses a momentum term to formulate the update direction by leveraging both current and historical gradient information [41; 46]. More precisely, a generic update rule of SGDM can be expressed as

$$m_t = (1-\beta)m_{t-1} + \beta\hat{g}(w_t; z_t), \tag{2}$$

$$w_{t+1} = w_t - \eta_t m_t, \tag{3}$$

where $\hat{g}(w_t; z_t) \in \partial\ell(w_t; z_t)$ with $z_t \sim \mathcal{D}$. The method is typically initialized with $m_0 = 0$ and $w_1 = 0$. In this paper, we generally allow that the learning rate η_t is potentially relying on the past stochastic subgradients, a standard setting for adaptive SGDM variants like Adam [29], AMSGrad [43] and RMSprop [49]. This family of algorithms are particularly effective for training transformer-based attention models [3; 20; 25; 50; 54].

1.1 Motivation

Despite the practical success of SGDM, our theoretical understanding of its underlying mechanisms remains limited—particularly in nonsmooth nonconvex settings, and with regard to the benefits afforded by *momentum*, the core component that distinguishes SGDM from vanilla SGD. Unlike in the smooth case, identifying an ϵ-stationary point (or even a neighborhood around it) for a nonsmooth objective in 1 is computationally intractable [17; 32]. For weakly convex functions, this impossibility result motivates the employment of the ϵ-stationarity of their Moreau envelopes as a notion of approximate convergence [17]. It has been demonstrated that both SGD and SGDM attain the same optimal first-order oracle complexity $\mathcal{O}(\epsilon^{-4})$ when optimizing the Moreau envelope of objective function [17; 36]. Yet this convergence result alone is far from sufficient to justify SGDM's practical superiority over vanilla SGD.

For general nonsmooth nonconvex functions, Zhang et al. [56] proposed to use Goldstein (δ, ϵ)-stationarity [23] as a convergence criterion. Building on top of the *online-to-nonconvex conversion* (O2NC) technique developed by Cutkosky et al. [16], it has recently been shown in a series of work that some modified versions of SGDM essentially achieve the complexity $\mathcal{O}(\delta^{-1}\epsilon^{-3})$ for producing (δ, ϵ)-stationary points [2; 26; 57], which is tight even when the subgradients are deterministic [31]. These findings, in fact, provide theoretical evidence to justify the benefits of momentum in two key respects, as outlined below:

- **General nonsmooth nonconvex problems.** In this setting, it is by far still not clear if vanilla SGD can also achieve the optimal rate $\mathcal{O}(\delta^{-1}\epsilon^{-3})$ for finding Goldstein stationary points. Indeed, to our knowledge, a complexity rate of $\mathcal{O}(\delta^{-1}\epsilon^{-4})$ is the best-known (dimension-free) result of SGD for optimizing stochastic Lipschitz functions [56].
- **Weakly convex problems.** For ρ-weakly convex functions, the optimal $\mathcal{O}(\epsilon^{-4})$ rates of SGD for the Moreau envelope imply the $\mathcal{O}(\delta^{-1}\epsilon^{-3})$ rates on Goldstein stationarity with $\delta \leq \epsilon/(2\rho)$ [26]. However, such a translation of rates does not necessarily hold for arbitrary precisions δ (say, $\delta = \sqrt{\epsilon}$, a choice critical for yielding optimal rates in second-order smooth functions [16]).

Despite the promise of O2NC for justifying the benefits of momentum in SGDM optimizers, existing results under this framework require significant algorithmic modifications such as model random interpolation [16], momentum clipping or periodic restarting [26], exponentially random momentum scaling [57], and model exponential moving averages [2], along with their combinations. Though theoretically insightful, these modifications are rarely, if ever, adopted in practical SGDM implementations. A core motivation of this work is to close this theory–practice gap, offering a more rigorous and faithful account of how momentum improves SGD performance for the prevalent nonsmooth nonconvex optimization problems encountered in real applications.

1.2 Our contribution

In this work, we develop a more comprehensive theoretical understanding of how momentum endows SGDM with superior performance over vanilla SGD for ρ-weakly convex optimization. Departing from the conventional O2NC framework, we conduct a novel direct analysis of SGDM for finding mildly relaxed Goldstein stationary points. Based on a newly proposed (μ, ϵ)-L_2-regularized stationarity criterion (Definition 2), we prove that SGDM with geometric-distribution-based randomized output achieves optimal complexity guarantees for constant, decaying, and adaptive learning rate schedules. We summarize our main contributions below.

- **Geometric-CDF-based randomized output generation.** In Algorithm 1, we propose two randomized output selection schemes built upon the cumulative distribution function (CDF) of the geometric distribution. This non-uniform output generation component is critical to our analysis of SGDM for nonsmooth nonconvex problems.
- **Results for constant learning rates.** In this setting, our main result is Theorem 1, whose corollary gives a complexity rate $\mathcal{O}\left(\mu^{1/2}\epsilon^{-4} + \rho^2\mu^{-1/2}\right)$ for producing (μ, ϵ)-L_2-regularized stationary points. Substituting $\mu = \epsilon^2\delta^{-2}$ yields a desirable rate $\mathcal{O}(\delta^{-1}\epsilon^{-3} + \rho^2\delta\epsilon^{-1})$ under a L_2-moment relaxed variant of (δ, ϵ)-Goldstein stationarity (Definition 1). The weak convexity parameter can scale up to $\rho = \mathcal{O}(\mu^{1/2}\epsilon^{-2})$ without degrading the dominant term $\mu^{1/2}\epsilon^{-4}$. In contrast to prior O2NC-based analyses [16; 26; 57] for the constant-learning-rate regime, our results hold *without requiring any modifications to SGDM's iteration update rule*.
- **Results for time-decaying learning rates.** When considering the widely adopted decaying learning rate schedule $\eta_t = \mathcal{O}(t^{-1/2})$, our main result from Theorem 2 and its corollary also establishes the complexity bound $\mathcal{O}\left(\mu^{1/2}\epsilon^{-4} + \rho^2\mu^{-1/2}\right)$ for SGDM. This bound yields the rate $\mathcal{O}(\delta^{-1}\epsilon^{-3} + \rho^2\delta\epsilon^{-1})$ for producing (δ, ϵ)-L_2-Goldstein stationary points.
- **Results for adaptive learning rates.** We further extend our analysis to SGDM with time-decaying and adaptive learning rate schedules. Specifically, in Theorem 3 and its corollary, we establish the following complexity bound for AMSGrad [43] (a convergent variant of Adam):

$$\mathcal{O}\left(\frac{\mu^{1/2}}{\epsilon^4} + \frac{\rho^2}{\mu^{1/2}} + \frac{1}{\varepsilon^2}\log\left(\frac{1}{\varepsilon}\right)\right),$$

  where $\varepsilon > 0$ denotes a small safeguard parameter for numerical stability. Notably, this safeguard parameter may be chosen as small as $\varepsilon = \mathcal{O}(\mu^{-1/4}\epsilon^2)$ such that it does not dominate the overall rate (up to logarithmic factors).

Additionally, in Theorem 4, we prove a lower bound stating that $\Omega(\mu^{1/2}\epsilon^{-4})$ gradient queries are unavoidable for finding a (μ, ϵ)-L_2-regularized stationary point of weakly convex functions. This complexity bound verifies that the $\mathcal{O}(\mu^{1/2}\epsilon^{-4})$ rate is indeed optimal under our relaxed notion of Goldstein stationarity. Finally, as a secondary contribution, we propose a momentum-reset variant of SGDM that zeroes the momentum term exactly once midway through the iterations and extend our analysis to this algorithm. In particular, our corresponding results for AMSGrad shed light on the empirical efficacy of momentum-reset Adam, which has been recently observed in LLM training [25] and deep reinforcement learning (RL) tasks [8]. We defer this body of results to Appendix E.

2 Preliminaries

2.1 Notation and definitions

For a pair of integers $T' \le T$, we use abbreviation $[T' : T] = \{T', T'+1, \cdots, T\}$ and $[T] = [1 : T]$. Let $f^* = \min_{w\in\mathbb{R}^d} f(w)$. We abbreviate $\Delta f(w) := f(w) - f^*$. We adopt the standard big-O notation $\mathcal{O}(\cdot)$ and $\Omega(\cdot)$ with universal constants hidden inside. We denote $\mathrm{dist}(0, V) := \inf_{v\in V}\|v\|$ and $\mathrm{conv}\{V\}$ the convex hull of a vector set V. $\mathcal{P}(V)$ denotes the set of all distributions over a measurable set V. The Clarke subdifferential [13] of a nonsmooth function f at $w \in \mathbb{R}^d$ is denoted by $\partial f(w)$. Let $\mathbb{B}_\delta(w)$ denote the closed ball of radius δ centered on w. The Goldstein δ-subdifferential of a Lipschitz function f at a point $w \in \mathbb{R}^d$ is the convex hull of all Clarke subgradients at points in $\mathbb{B}_\delta(w)$, i.e., $\partial_\delta f(w) := \mathrm{conv}\left\{\cup_{v\in\mathbb{B}_\delta(w)}\partial f(v)\right\}$.

We say that a function $f : \mathbb{R}^d \mapsto \mathbb{R}$ is G-Lipschitz continuous if $|f(w) - f(w')| \le G\|w - w'\|$ for all $w, w' \in \mathbb{R}^d$. f is L-smooth if f is differentiable and ∇f is L-Lipschitz continuous; f is H-second-order-smooth if f is twice differentiable and $\nabla^2 f$ is H-Lipschitz continuous. A Lipschitz continuous function f is said to be ρ-weakly convex for some $\rho > 0$ if $f(w) \ge f(w') + \langle g', w - w'\rangle - \frac{\rho}{2}\|w - w'\|^2,\ \forall w, w' \in \mathbb{R}^d, g' \in \partial f(w')$, or equivalently $f(\cdot) + \frac{\rho}{2}\|\cdot\|^2$ is convex. Among others, a common source of weakly convex functions is the composite form $f = h \circ c$ where h is a convex and Lipschitz continuous function, and c is a smooth mapping [17]. For example, the loss function of neural networks with smooth activation functions (e.g., GeLU) takes this composite form, where h denotes a convex top-layer predictor (e.g., cross-entropy loss) and c a smooth hierarchical feature mapping, rendering the overall loss function weakly convex.

2.2 A RELAXED GOLDSTEIN STATIONARITY CRITERION

The Goldstein stationarity [23] is a standard convergence criterion for nonsmooth nonconvex optimization. Formally, a point w is called a (δ, ϵ)-Goldstein stationary point of a Lipschitz function f if $\operatorname{dist}(0, \partial_\delta f(w)) \leq \epsilon$. We next define a natural relaxation of Goldstein stationarity, without restricting the evaluation points of Clarke subgradients to those within the δ-ball.

Definition 1 (L_2-Goldstein Stationarity)**.** *A point w is called a (δ, ϵ)-L_2-Goldstein stationary point of a Lipschitz function f if there exists $V \subset \mathbb{R}^d$ along with $v \sim P \in \mathcal{P}(V)$ and $g(v) \in \partial f(v)$ a.s. such that $\|\mathbb{E}[g(v)]\| \leq \epsilon$ and $\mathbb{E}[\|v - w\|^2] \leq \delta^2$.*

Remark 1. *Compared with (δ, ϵ)-Goldstein stationarity which restricts V to $\mathbb{B}_\delta(w)$, the proposed (δ, ϵ)-L_2-Goldstein stationarity permits elements in V to be arbitrarily large. It maintains, however, tightly controlled average proximity to w via the L_2-moment constraint $\mathbb{E}\left[\|v - w\|^2\right] \leq \delta^2$.*

Further, inspired by Zhang and Cutkosky [57], we introduce a regularized version of (δ, ϵ)-L_2-Goldstein stationarity, which turns out to be more convenient for analyzing SGDM.

Definition 2 (L_2-Regularized Stationarity)**.** *A point w is said to be a (μ, ϵ)-L_2-regularized stationary point of a Lipschitz function f if $\|\partial_{+\mu} f(w)\| \leq \epsilon$, where*

$$\|\partial_{+\mu} f(w)\| := \inf_{V \subset \mathbb{R}^d, v \sim P \in \mathcal{P}(V), g(v) \in \partial f(v) \ a.s.} \sqrt{\|\mathbb{E}\left[g(v)\right]\|^2 + \mu \mathbb{E}\left[\|v - w\|^2\right]}. \tag{4}$$

Further, w is said to be an unbiased (μ, ϵ)-L_2-regularized stationary point if $\|\partial^\star_{+\mu} f(w)\| \leq \epsilon$, where

$$\left\|\partial^\star_{+\mu} f(w)\right\| := \inf_{V \subset \mathbb{R}^d, v \sim P \in \mathcal{P}(V), \mathbb{E}[v] = w, g(v) \in \partial f(v) \ a.s.} \sqrt{\|\mathbb{E}\left[g(v)\right]\|^2 + \mu \mathbb{E}\left[\|v - w\|^2\right]}. \tag{5}$$

Remark 2. *In contrast to the notion of (c, ϵ)-stationarity [57, Definition 2.2], our (μ, ϵ)-L_2-regularized stationarity, as defined in 4, relaxes the unbiasedness constraint that $\mathbb{E}[v] = w$. This modification affords greater flexibility for optimization settings without having to use model exponential moving average (EMA). In addition, we introduce an unbiased formulation defined with 5 to accommodate the mildly more stringent EMA-based configurations.*

Several key properties of the (μ, ϵ)-L_2-regularized stationarity criterion, including its equivalence to (δ, ϵ)-L_2-Goldstein stationarity and its implications for smooth objective functions, can be found in Appendix A.

2.3 KEY ASSUMPTIONS

We now formally state the key structural assumptions on the loss function, risk function, and their subgradients for Problem 1.

Assumption 1 (Lipschitz continuity and weak convexity)**.** *For any $z \in \mathcal{Z}$, the loss function $\ell(\cdot; z)$ is G-Lipschitz with respect to its first argument. Moreover, the risk function f is ρ-weakly convex.*

Assumption 2 (Unbiased stochastic oracle)**.** *For each $w \in \mathbb{R}^d$, it holds that $g(w) = \mathbb{E}_{Z \sim \mathcal{D}}[\hat{g}(w; Z)] \in \partial f(w)$, where $\hat{g}(w; z) \in \partial \ell(w; z)$ for any $z \in \mathcal{Z}$.*

Assumption 3 (Bounded loss)**.** *For any $z \in \mathcal{Z}$, the loss function satisfies $\ell(\cdot; z) \in [0, M]$.*

Assumptions 1 and 2 are commonly adopted in the convergence analysis of stochastic subgradient methods for nonsmooth nonconvex optimization [17; 26; 36]. The bounded-loss condition in Assumption 3 is primarily employed for the analysis under time-decaying learning rate schedules. Such a condition can be realized in practice via capped loss functions adopted for LLM training [44].

3 ALGORITHM AND ANALYSIS

3.1 SGDM WITH SQUARED-GEOMETRIC-CDF RANDOMIZED OUTPUT

We begin by introducing a novel non-uniform randomized output generation mechanism for SGDM based on the geometric distribution, which is critical to our analysis under nonsmooth nonconvex settings. As outlined in Algorithm 1, we just adopt the standard SGDM update rules given in 2 and 3. For output generation, we propose two optional stopping-index selection schemes based on the cumulative distribution function (CDF) of geometric distribution, as described below:

Algorithm 1: SGDM with Squared-Geometric-CDF Randomized Output (SGDM-Geo^2)

Input : $T \in \mathbb{N}$, $\beta \in (0,1)$, $t_0 \in [0 : T-1]$, initial vectors $m_0 = w_1 = 0$.
for $t = 1, ..., T$ **do**
  Randomly sample $z_t \sim \mathcal{D}$ and obtain $\hat{g}(w_t; z_t) \in \partial \ell(w_t; z_t)$;
  Update the momentum term m_t according to 2;
  Update the model iterate w_{t+1} according to 3;
end

(**Option-I**) `/* Squared-geometric-CDF randomized output */`
Randomly select S from $[T]$ with probability $\mathbb{P}(S = t) \propto (1 - (1-\beta)^t)^2$.

(**Option-II**) `/* Nested squared-geometric-CDF randomized output */`
Randomly select R from $[t_0 + 1 : T]$ with probability $\mathbb{P}(R = t) \propto \sum_{\tau=t+1}^{T} (1 - (1-\beta)^{\tau - t})^2$;
Conditioned on $R = t$, randomly select S from $[t+1 : T]$ with conditional probability $\mathbb{P}(S = \tau \mid R = t) \propto (1 - (1-\beta)^{\tau - t})^2$.

Output: w_S.

- **Option-I: Squared-geometric-CDF randomized output.** Given a total iteration budget T, we sample the stopping index $S \in [T]$ from a truncated distribution whose unnormalized probability mass is proportional to the squared CDF of a geometric distribution with parameter β, i.e., $\mathbb{P}(S = t) \propto (1 - (1-\beta)^t)^2$. We term w_S the squared-geometric-CDF randomized output, which we will demonstrate facilitates the analysis of SGDM under constant learning rates.
- **Option-II: Nested squared-geometric-CDF randomized output.** Given T and a prescribed time step t_0, we further propose the following nested squared-geometric-CDF sampling procedure for output generation. First, we randomly sample an index R from $[t_0 + 1 : T]$ with probability $\mathbb{P}(R = t) \propto \sum_{\tau=t+1}^{T} (1 - (1-\beta)^{\tau - t})^2$. Conditioned on the sampled value $R = t \in [t_0 + 1 : T]$, we then draw a second random index $S \in [t+1 : T]$ with conditional probability $\mathbb{P}(S = \tau \mid R = t) \propto (1 - (1-\beta)^{\tau - t})^2$ and output w_S. This nested output sampling scheme proves useful for analyzing SGDM under time-decaying learning rates. Alternatively, the time-decaying case may be addressed by introducing a momentum-reset mechanism into SGDM, which resets the momentum term to zero at a single time instance $t_0 \in [T]$. With this design, we only require the squared-geometric-CDF randomized output mechanism analogous to Option-I. Due to space constraints, we defer this alternative momentum-reset SGDM algorithm along with its convergence analysis and performance evaluation to Appendix E.

Under both squared-geometric-CDF output sampling schemes above, more recent iterates are assigned a higher probability of being selected as the final output. Notably, the hyperparameter t_0 is solely employed to generate outputs after the prescribed iterations and exerts no influence on the intermediate updates. We refer to Algorithm 1 as SGDM-Geo^2 throughout this paper.

Next, we proceed to analyze the convergence rate of SGDM-Geo^2 for weakly convex problems under constant, time-decaying, and adaptive learning rate schedules.

3.2 Results under constant learning rates

We first consider SGDM-Geo^2 with a fixed learning rate $\eta_t \equiv \eta$. In this case, we have the following main result on the convergence rate of Algorithm 1 with randomized output generated via Option-I. See Appendix C.1 for a proof of this result.

Theorem 1. *Suppose Assumption 1 and 2 hold. Let $T \geq 8$ and $\beta \in \left[\frac{4}{T}, \frac{1}{2}\right]$. Let $\eta_t \equiv \eta$ with $\eta \leq \frac{1}{2\rho}$. Then for any $\mu \leq \frac{\beta^2}{6\eta^2(1-\beta)}$, the randomized output w_S produced by SGDM-Geo^2 (Algorithm 1, Option-I) satisfies*

$$\mathbb{E}\left[\left\|\partial_{+\mu} f(w_S)\right\|^2\right] \leq \frac{8\Delta f(w_0)}{\eta T} + 12\beta G^2.$$

Remark 3. *Within the O2NC framework, similar convergence guarantees have also been established for SGDM with constant learning rates, yet they require additional modifications to the momentum term such as exponential random scaling [57] or periodic restarting [26]. Notably, our*

bound in Theorem 1 applies to the vanilla SGDM without any modifications in iteration rule, making it substantially more aligned with practical implementations.

Based on Theorem 1, we further establish the iteration complexity rate of SGDM-Geo^2 for obtaining (μ,ϵ)-L_2-regularized stationary points. Detailed proofs are deferred to Appendix C.2.

Corollary 1. *Suppose Assumption 1 and 2 hold. Let $\mu,\epsilon>0$ denote the target parameters for L_2-regularized stationarity, and let T denote the total iteration budget. Set $\eta_t\equiv\eta=\frac{1}{T^{1/2}\mu^{1/4}}$ and $\beta=\frac{\sqrt{6}\mu^{1/4}}{T^{1/2}}$. Assume T is sufficiently large such that*

$$T\geq\frac{16(2\Delta f(w_0)+3\sqrt{6}G^2)^2\mu^{1/2}}{\epsilon^4}+\frac{3+\rho^2}{\mu^{1/2}}+24\mu^{1/2}.$$

Then the randomized output w_S produced by SGDM-Geo^2 (Algorithm 1, Option-I) satisfies $\mathbb{E}\left[\|\partial_{+\mu}f(w_S)\|\right]\leq\epsilon$.

A few remarks on Corollary 1 are in order.

Remark 4. *In view of Proposition 1, let us set $\mu=\delta^{-2}\epsilon^2$ for a prescribed precision $\delta>0$. Then the bound in Corollary 1 yields an iteration complexity with leading rate $\mathcal{O}\left(\delta^{-1}\epsilon^{-3}+\rho^2\delta\epsilon^{-1}\right)$ for producing (δ,ϵ)-L_2-Goldstein stationary points. The overall complexity is dominated by the leading term $\mathcal{O}(\delta^{-1}\epsilon^{-3})$, which matches the optimal rate known for Goldstein stationarity [16]. As we will show in Theorem 4, the $\mathcal{O}(\mu^{1/2}\epsilon^{-4})$ rate is indeed tight under (μ,ϵ)-L_2-regularized stationarity.*

Remark 5. *Remarkably, the weak convexity parameter can scale as large as $\rho=\mathcal{O}(\mu^{1/2}\epsilon^{-2})$ without degrading the dominant term $\mathcal{O}(\mu^{1/2}\epsilon^{-4})$ in the composite bound. We regard this as another advantage of SGDM over SGD for weakly convex optimization.*

Remark 6. *The choice of hyper-parameters η,β are uniquely determined by user-specified quantities, including iterate budget T and stationarity precision parameters μ,ϵ. Since T must scale as $\mathcal{O}(\mu^{1/2}\epsilon^{-4})$, we require $\eta=\mathcal{O}(\mu^{-1/2}\epsilon^2)$ and $\beta=\mathcal{O}(\epsilon^2)$.*

3.3 Results under time-decaying learning rates

We next analyze the convergence of SGDM-Geo^2 under practically prevalent time-decaying learning rates. Specifically, we adopt the schedule $\eta_t=\frac{\eta}{\sqrt{t+1}}$, where $\eta>0$ denotes a tunable constant. We state our main theoretical result for this setting below, with detailed proofs deferred to Appendix C.3.

Theorem 2. *Suppose Assumption 1, 2 and 3 hold. Let $T\geq 48$ and $\beta\in\left[\frac{24}{T},\frac{1}{2}\right]$, and set $\eta_t=\frac{\eta}{\sqrt{t+1}}$ with $\eta>0$. Assume $\eta_{t_0}\leq\frac{1}{2\rho}$ for some $t_0\leq\frac{T}{2}-3$. Then for any $\mu\leq\frac{\beta^2T}{6\eta^2(1-\beta)}$, the randomized output w_S produced by SGDM-Geo^2 (Algorithm 1, Option-II) satisfies*

$$\mathbb{E}\left[\|\partial_{+\mu}f(w_S)\|^2\right]\leq\frac{25}{T}\left(\frac{3G^2}{\beta^2(t_0+1)}+\frac{6M}{\beta\eta\sqrt{t_0+1}}+6G^2\beta T\log\left(\frac{T}{t_0}\right)+\frac{5M\sqrt{T}}{\eta}\right).$$

As a direct consequence of Theorem 2, the following corollary further establishes the complexity rates of SGDM-Geo^2 for obtaining (μ,ϵ)-L_2-regularized stationary points under proper choice of t_0. Its proof is deferred to Appendix C.4.

Corollary 2. *Suppose Assumption 1, 2 and 3 hold. Let $\mu,\epsilon>0$ denote the target parameters for L_2-regularized stationarity and T denote the total iteration budget. Set $\eta_t=\frac{1}{\mu^{1/4}\sqrt{t+1}}$, $t_0=\left\lfloor\frac{T}{3}\right\rfloor$ and $\beta=\frac{\sqrt{6}\mu^{1/4}}{T^{1/2}}$. Assume T is sufficiently large such that*

$$T\geq\frac{625(17G^2+5M)^2\mu^{1/2}}{\epsilon^4}+\frac{75}{\epsilon^2}\left(\frac{G^2}{2\mu^{1/2}}+\sqrt{2}M\right)+\frac{96+12\rho^2}{\mu^{1/2}}+24\mu^{1/2}.$$

Then the randomized output w_S produced by SGDM-Geo^2 (Algorithm 1, Option-II) satisfies $\mathbb{E}\left[\|\partial_{+\mu}f(w_S)\|\right]\leq\epsilon$.

Remark 7. *Similar to the constant learning rate setting, our sample complexity bounds in Corollary 2 are governed by the rate $\mathcal{O}(\mu^{1/2}\epsilon^{-4})$, which is tight in μ,ϵ. Furthermore, the weak convexity parameter ρ can scale up to $\mathcal{O}(\mu^{1/2}\epsilon^{-2})$ while preserving this optimality.*

While the results in this subsection are established for the learning rate $\eta_t = \frac{\eta}{\sqrt{t+1}}$, we emphasize that analogous conclusions apply to the bias-corrected learning rate $\eta_t = \frac{\eta}{1-(1-\beta)^t}$. We omit the details here and present a more general result for Adam-style algorithms in the next subsection.

3.4 Results under adaptive learning rates

We further consider SGDM-Geo^2 equipped with adaptive learning rates, where η_t is determined by aggregating the norms of historical stochastic subgradients. This is a standard design adopted by Adam [29] and its variants [35; 43]. Building upon our previous analysis under time-decaying learning rates, we focus specifically on AMSGrad [43] which is a variant of Adam whose learning rate is both adaptive and monotonically decreasing across iterations. The AMSGrad update rule is stated as follows:

$$m_t = (1-\beta_1)m_{t-1} + \beta_1 \hat{g}(w_t; z_t), \tag{6}$$

$$v_t = (1-\beta_2)v_{t-1} + \beta_2 \|\hat{g}(w_t; z_t)\|^2,\ \hat{v}_t = \max\left\{\hat{v}_{t-1}, \frac{v_t}{1-(1-\beta_2)^t}\right\}, \tag{7}$$

$$w_{t+1} = w_t - \eta_t m_t,\ \eta_t = \frac{\eta}{(\sqrt{\hat{v}_t}+\varepsilon)(1-(1-\beta_1)^t)}, \tag{8}$$

where β_1 is the momentum parameter, β_2 is the decay rate of the EMA of the squared norm of subgradients, and ε is a small safeguard value introduced to avoid numerical instability. Typically, we set $v_0 = \hat{v}_0 = 0$. For analytical purposes, we will shortly propose initializing $\hat{v}_0$ as some deterministic value satisfying $\hat{v}_0 \le G^2$. Compared to the original version [43], we explicitly incorporate bias correction and a safeguard mechanism, two key components widely adopted in practical Adam-style optimizers. For simplicity, we focus on the vector-based formulation of AMSGrad, though our analysis readily extends to its coordinate-wise counterpart.

Since the AMSGrad update rule described above fits within the general SGDM framework, with momentum parameter $\beta = \beta_1$ and learning rate η_t specified by 8, we instantiate Algorithm 1 to yield the AMSGrad variant equipped with the nested randomized output Option-II. We now present our main convergence result for AMSGrad in weakly convex optimization under the L_2-regularized stationarity metric. Detailed proofs are deferred to Appendix C.5.

Theorem 3. *Suppose Assumption 1, 2 and 3 hold. Let $T \ge 48$ and $\beta_1 \in \left[\frac{24}{T}, \frac{1}{2}\right]$. Choose η_t according to 8 for arbitrary $\eta > 0$. Further assume $\eta_{t_0} \le \frac{1}{2\rho}$, $(1-\beta_1)^{t_0} \le 0.05$ for some $t_0 \le \frac{T}{2} - 3$, and $\varepsilon \ge G\sqrt{2\beta_2}$. Then for any $\mu \le \frac{(G+\varepsilon)^2\beta_1^2}{6\eta^2(1-\beta_1)}$, the randomized output w_S produced by SGDM-Geo^2 (Algorithm 1, Option-II) satisfies*

$$\begin{aligned}&\mathbb{E}\left[\|\partial_{+\mu} f(w_S)\|^2\right]\\ \le&\frac{84(G+\varepsilon)^2G^2}{T^2\beta_1^2(\hat{v}_0+\varepsilon^2)} + \frac{72(G+\varepsilon)^2M}{T(\sqrt{\hat{v}_0}+\varepsilon)\eta} + \frac{96(G+\varepsilon)^2\beta_1}{T}\left(\frac{1}{\beta_2} + \frac{2}{\beta_2}\log\left(\frac{2}{\beta_2}\right) + \frac{T-t_0}{1-\beta_2}\right).\end{aligned}$$

The following corollary further establishes the iteration complexity of Algorithm 1 (with $t_0 \approx \frac{T}{3}$) under the adaptive learning rate schedule used on AMSGrad. Its proof is deferred to Appendix C.6.

Corollary 3. *Suppose Assumption 1, 2 and 3 hold. Let $\mu, \epsilon > 0$ denote the target parameters for L_2-regularized stationarity and T denote the total iteration budget. Let $\varepsilon > 0$ be an arbitrary safeguard value. Set $t_0 = \lfloor \frac{T}{3} \rfloor$, $\eta = \frac{\sqrt{\hat{v}_0}+\varepsilon}{\mu^{1/4}T^{1/2}}$, $\beta_1 = \frac{\sqrt{6}\mu^{1/4}}{T^{1/2}}$, and $\beta_2 = \frac{\log T}{T}$. Assume T is sufficiently large such that*

$$T \ge \frac{4(G+\varepsilon)^4\mu^{1/2}}{\epsilon^4}\left(\frac{72M}{\hat{v}_0+\varepsilon^2} + 706\right)^2 + \frac{28(G+\varepsilon)^2G^2}{(\hat{v}_0+\varepsilon^2)\epsilon^2\mu^{1/2}} + \frac{96+5\rho^2}{\mu^{1/2}} + 24\mu^{1/2} + \frac{8G^2}{\varepsilon^2}\log\left(\frac{G}{\varepsilon}\right).$$

Then the randomized output w_S produced by SGDM-Geo^2 (Algorithm 1, Option-II) satisfies $\mathbb{E}\left[\|\partial_{+\mu} f(w_S)\|\right] \le \epsilon$.

Remark 8. *Without loss of generality, we assume $\hat{v}_0 = \mathcal{O}(1)$ (say, $\hat{v}_0 = \|g(0)\|$). Since T must scale as $\mathcal{O}(\mu^{1/2}\epsilon^{-4})$, this necessitates the choices $\eta = \mathcal{O}(\mu^{-1/2}\epsilon^2)$, $\beta_1 = \mathcal{O}(\epsilon^2)$ and $\beta_2 = \tilde{\mathcal{O}}(\mu^{-1/2}\epsilon^4)$. This well aligns with typical practical hyperparameter selections satisfying $\beta_1 \gg \beta_2$ (e.g., 0.1 versus 0.001). Moreover, the safeguard value may be set as small as $\varepsilon = \mathcal{O}(\mu^{-1/4}\epsilon^2)$ without dominating the optimal rate, up to logarithmic factors.*

3.5 LOWER BOUND ANALYSIS

In this subsection, we further show that the rate $\mathcal{O}(\mu^{1/2}\epsilon^{-4})$ established in Corollary 1, 2, and 3 is optimal for stochastic weakly convex optimization under the (μ,ϵ)-L_2-regularized stationarity criterion. To prove this optimality, we establish that $\Omega(\mu^{1/2}\epsilon^{-4})$ gradient queries are necessary to obtain a (μ,ϵ)-L_2-regularized stationary point, even for smooth objective functions — a subclass of weakly convex functions. The theorem below formally characterizes the stochastic oracle complexity for computing such (μ,ϵ)-L_2-regularized stationary point under weak convexity. Detailed proofs are deferred to Appendix C.7.

Theorem 4. *For any constants γ, ρ, G, μ, ϵ satisfying $G \geq \gamma^{1/2}\mu^{1/4}$ and $\rho \geq \mu^{1/2}$, there exist an objective function f and a stochastic gradient estimator $\hat{g}$ such that f is ρ-weakly convex and G-Lipschitz, $f(w_0) - \inf f \leq \gamma$, and $\mathbb{E}_{Z\sim\mathcal{D}}\|\nabla f(w) - \hat{g}(w;Z)\|^2 \leq G^2$. Moreover, any randomized algorithm with oracle access only to $\hat{g}$ requires $\Omega(\gamma G^2\mu^{1/2}\epsilon^{-4})$ iterations to find an (μ,ϵ)-L_2-regularized stationary point of f.*

Remark 9. *Based on Proposition 1, substituting $\mu = \delta^{-2}\epsilon^2$ into Theorem 4 gives the lower bound $\Omega(\gamma G^2\delta^{-1}\epsilon^{-3})$ for finding (δ,ϵ)-L_2-Goldstein stationary points of weakly convex functions.*

3.6 COMPARISON WITH PRIOR RESULTS

Table 1 compares our SGDM results with existing results for stochastic subgradient-based methods for nonsmooth nonconvex optimization along four dimensions: objective structure, convergence criteria and iteration complexity, implications for second-order smooth functions, and compatibility with learning rate schedules. A few remarks are in order.

- **Objective function structure.** The objective functions examined in the theoretical analysis are weakly convex for SGDM-Geo^2, SGD [17], SGDM [36], and D-O2NC [26]; Lipschitz continuous for SINGD [56]; and well-behaved (e.g., f is differentiable everywhere) for O2NC [16]. Notably, the algorithms considered herein for weakly convex functions are fully deterministic, aside from stochastic gradient estimation. By contrast, the SINGD method for Lipschitz functions and the vanilla O2NC method for well-behaved functions both require randomization via auxiliary uniform interpolation on iterates, which creates discrepancies relative to practical SGDM implementations.
- **Convergence criteria and iteration complexity.** For SGD and SGDM, convergence is characterized via ϵ-stationarity of the Moreau envelope of the objective function, where the complexity rate $\mathcal{O}(\epsilon^{-4})$ is optimal, as recently verified in [34]. For SINGD, O2NC and D-O2NC, convergence is assessed under (δ,ϵ)-Goldstein stationarity: the dominant rate $\mathcal{O}(\delta^{-1}\epsilon^{-3})$ achieved by O2NC and D-O2NC is optimal while SINGD yields a suboptimal rate of $\mathcal{O}(\delta^{-1}\epsilon^{-4})$. For our proposed SGDM-Geo^2, we employ the (δ,ϵ)-L_2-Goldstein stationarity criterion, which is equivalent to (μ,ϵ)-L_2-regularized stationarity under the mapping $\mu = \delta^{-2}\epsilon^2$. Its complexity rate $\mathcal{O}(\delta^{-1}\epsilon^{-3})$ is tight in the sense of Remark 4 and Remark 9.
- **Implications for second-order smooth functions.** The complexity bounds of O2NC, D-O2NC and SGDM-Geo^2 (with EMA output) naturally recover the optimal $\mathcal{O}(\epsilon^{-3.5})$ rate for finding an ϵ-stationary point under second-order smoothness. In contrast, SGD, SGDM, and SINGD do not automatically enjoy this property, per the results summarized in Table 1.
- **Compatibility with learning rate schedules.** Our analysis for SGDM-Geo^2 not only attains optimal complexity rates, but also accommodates constant, time-decaying, and adaptive learning rate schedules simultaneously — a distinctive advantage over the compared baseline approaches.

4 FURTHER RELATED WORK

Since the early developments of Polyak [40]; Nesterov [37], optimization algorithms based on momentum techniques have become popular in both theory and practice [10; 11]. For smooth and convex problems, momentum-based gradient acceleration algorithms have been shown to achieve the optimal complexity rate [38]. The effectiveness of such techniques in the deterministic setting has inspired researchers to use momentum for stochastic optimization algorithms [15; 47; 53]. Most theoretical studies on SGDM and its variants focus on characterizing the sample complexity bound

Method	Objective	Stationarity	Complexity	$\rightsquigarrow$ 2nd-sm	C-	D-	A-lr
SGD [17]	ρ-wcvx	ϵ-mes	$\mathcal{O}\left(\frac{\rho}{\epsilon^4}\right)$	–	✓	✓	✗
SGDM [36]	ρ-wcvx	ϵ-mes	$\mathcal{O}\left(\frac{\rho^2}{\epsilon^4}\right)$	–	✓	✓	✗
SINGD [56]	Lip	(δ,ϵ)-gs	$\mathcal{O}\left(\frac{1}{\delta\epsilon^4}\right)$	–	✗	✗	✓
O2NC [16]	wbvd	(δ,ϵ)-gs	$\mathcal{O}\left(\frac{1}{\delta\epsilon^3}\right)$	$\mathcal{O}\left(\frac{1}{\epsilon^{3.5}}\right)$	✓	✗	✗
D-O2NC [26]	ρ-wcvx	(δ,ϵ)-gs	$\mathcal{O}\left(\frac{1}{\delta\epsilon^3}+\frac{\rho^{7/3}\delta^{4/3}}{\epsilon^{2/3}}\right)$	$\mathcal{O}\left(\frac{1}{\epsilon^{3.5}}\right)$	✓	✗	✗
SGDM-Geo^2	ρ-wcvx	(δ,ϵ)-L_2-gs	$\mathcal{O}\left(\frac{1}{\delta\epsilon^3}+\frac{\rho^2\delta}{\epsilon}\right)$	$\mathcal{O}\left(\frac{1}{\epsilon^{3.5}}\right)$	✓	✓	✓

Table 1: Theoretical comparison of relevant results for stochastic subgradient methods in nonsmooth nonconvex optimization. Abbreviations: *mes*: Moreau-envelope stationarity; *gs*: Goldstein stationarity; *2nd-sm*: second-order smoothness; *C-/D-/A-lr*: constant/decaying/adaptive learning rate; *wcvx*: weakly convex; *Lip*: Lipschitz, *wbvd*: well-behaved.

for convex or smooth nonconvex functions, settings in which algorithms such as SGD have already attained the minimax optimal convergence rate [5; 36; 58]. Recently, the study on the benefits of momentum for enhancing stochastic optimization methods in nonsmooth nonconvex problems has become an emerging topic. Under the generalized smoothness conditions [55], it was shown by Crawshaw et al. [14] that momentum is beneficial for improving SGD. From the perspective of online learning, several recent works have shown that SGDM and its variants achieve optimal sample complexity for nonsmooth nonconvex optimization problems [2; 4; 26; 39], which has not been possible via the existing analysis of SGD methods.

Our study builds on a recent line of research focused on efficient algorithms with finite-time complexity guarantees for finding Goldstein-style stationary points—an area that has garnered growing interest in nonsmooth nonconvex optimization [19; 28; 30; 48; 56]. Particularly, Cutkosky et al. [16] introduced the O2NC framework which for the first time established the optimal rate for stochastic nonsmooth nonconvex optimization. By instantiating different online learners within this framework, it is possible to recover several popular optimizers including SGDM [57], Adam [2], schedule-free SGD [1], and Muon [27].

Specifically, for a rich class of weakly convex functions, a vast body of asymptotic convergence results has been established for stochastic optimization algorithms [21; 22]. In recent years, the finite-time non-asymptotic rates were achieved by Davis and Grimmer [18]; Davis and Drusvyatskiy [17]; Mai and Johansson [36], demonstrating that various SGD/SGDM algorithms can attain the complexity rate $\mathcal{O}(\epsilon^{-4})$, which is tight even for deterministic gradient methods [34], for generating an ϵ-stationary point of the Moreau envelope of the objective function.

5 Conclusion

This work advances the theoretical understanding of momentum benefits for stochastic subgradient methods in weakly convex optimization. Departing from the dominant O2NC analysis paradigm, we provide, to our knowledge, the first direct complexity analysis of standard SGDM for computing relaxed Goldstein stationary points. We show that SGDM equipped with squared-geometric-CDF randomized output achieves optimal iteration complexity across constant, decaying, and adaptive learning rate schedules — a favorable property absent in vanilla SGD. While our core theory is established under L_2-regularized stationarity, all results naturally extend to its unbiased variant, which better accommodates the EMA mechanism prevalent in modern adaptive optimizers.

Open questions and future works. Several open questions remain regarding the tightness and generality of our results. For SGDM, it remains unclear whether optimal rates can be preserved under the original Goldstein stationarity without relaxation. For SGD, it is likewise unknown whether optimal complexity is achievable even under relaxed Goldstein stationarity. Furthermore, while Theorem 2 and 3 derive favorable complexity rates for SGDM with decaying and adaptive learning rates, these results rely on a bounded-loss condition. Whether this restrictive global assumption can be eliminated remains an open issue. We leave these promising directions for future investigation.

AI USE STATEMENT

In this work, we used generative AI tools for manuscript language polishing, including refining English grammar and improving the readability of non-descriptive textual content. We did not use generative AI tools for research conception, theoretical derivation, result analysis, or the writing of core manuscript sections. All AI-generated outputs were thoroughly revised and validated by the authors to ensure technical accuracy, contextual consistency, and compliance with academic standards. We take full responsibility for the final manuscript content, including all text, claims, and artifacts produced with generative AI assistance. All authors are ultimately accountable for the integrity and scientific validity of this work.

ETHICS STATEMENT

This work focuses on the theoretical analysis of SGDM for nonsmooth nonconvex optimization. It involves no human-subject studies, collection of sensitive personal data, or animal experiments. No ethical risks are present, and no ethical review is required for this research.

REPRODUCIBILITY STATEMENT

This work is primarily theoretical. All assumptions, algorithm definitions, and full proofs for our main theorems and remarks are provided in the main text and appendix. The appendix further includes preliminary experimental results, accompanied by complete documentation of experimental setups, hyperparameters, and evaluation protocols. No external datasets or custom source code are required to verify our theoretical claims.

REFERENCES

[1] Kwang Jun Ahn, Gagik Magakyan, and Ashok Cutkosky. General framework for online-to-nonconvex conversion: Schedule-free sgd is also effective for nonconvex optimization. In *International Conference on Machine Learning (ICML)*, 2025.

[2] Kwangjun Ahn and Ashok Cutkosky. Adam with model exponential moving average is effective for nonconvex optimization. *Advances in Neural Information Processing Systems*, 37: 94909–94933, 2024.

[3] Kwangjun Ahn, Xiang Cheng, Minhak Song, Chulhee Yun, Ali Jadbabaie, and Suvrit Sra. Linear attention is (maybe) all you need (to understand transformer optimization). In *The Twelfth International Conference on Learning Representations*, 2024.

[4] Kwangjun Ahn, Gagik Magakyan, and Ashok Cutkosky. General framework for online-to-nonconvex conversion: Schedule-free sgd is also effective for nonconvex optimization. *arXiv preprint arXiv:2411.07061*, 2024.

[5] Ahmet Alacaoglu, Yura Malitsky, Panayotis Mertikopoulos, and Volkan Cevher. A new regret analysis for adam-type algorithms. In *International conference on machine learning*, pages 202–210. PMLR, 2020.

[6] Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Ayush Sekhari, and Karthik Sridharan. Second-order information in non-convex stochastic optimization: Power and limitations. In *Conference on Learning Theory*, pages 242–299. PMLR, 2020.

[7] Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1):165–214, 2023.

[8] Kavosh Asadi, Rasool Fakoor, and Shoham Sabach. Resetting the optimizer in deep rl: An empirical study. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

[9] Hilal Asi and John C Duchi. The importance of better models in stochastic optimization. *Proceedings of the National Academy of Sciences*, 116(46):22924–22930, 2019.

[10] A. Beck. *First-Order Methods in Optimization*. MOS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics, 2017. ISBN 9781611974997. URL https://books.google.com/books?id=wrk4DwAAQBAJ.

[11] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Found. Trends Mach. Learn.*, 8(3-4):231–357, 2015.

[12] Vasileios Charisopoulos, Yudong Chen, Damek Davis, Mateo Díaz, Lijun Ding, and Dmitriy Drusvyatskiy. Low-rank matrix recovery with composite optimization: good conditioning and rapid convergence. *Foundations of Computational Mathematics*, 21(6):1505–1593, 2021.

[13] Frank H Clarke. *Optimization and nonsmooth analysis*. SIAM, 1990.

[14] Michael Crawshaw, Mingrui Liu, Francesco Orabona, Wei Zhang, and Zhenxun Zhuang. Robustness to unbounded smoothness of generalized signsgd. *Advances in neural information processing systems*, 35:9955–9968, 2022.

[15] Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex sgd. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, pages 15236–15245, 2019.

[16] Ashok Cutkosky, Harsh Mehta, and Francesco Orabona. Optimal stochastic non-smooth non-convex optimization through online-to-non-convex conversion. In *International Conference on Machine Learning (ICML)*, pages 6643–6670. PMLR, 2023.

[17] Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM J. Optim.*, 29(1):207–239, 2019.

[18] Damek Davis and Benjamin Grimmer. Proximally guided stochastic subgradient method for nonsmooth, nonconvex problems. *SIAM Journal on Optimization*, 29(3):1908–1930, 2019.

[19] Damek Davis, Dmitriy Drusvyatskiy, Yin Tat Lee, Swati Padmanabhan, and Guanghao Ye. A gradient sampling method with complexity guarantees for lipschitz functions in high and low dimensions. *Advances in neural information processing systems*, 35:6692–6703, 2022.

[20] Aaron Defazio, Xingyu Yang, Ahmed Khaled, Konstantin Mishchenko, Harsh Mehta, and Ashok Cutkosky. The road less scheduled. *Advances in Neural Information Processing Systems*, 37:9974–10007, 2024.

[21] John C. Duchi and Feng Ruan. Stochastic methods for composite and weakly convex optimization problems. *SIAM J. Optim.*, 28(4):3229–3259, 2018.

[22] Yu M Ermol'ev and VI Norkin. Stochastic generalized gradient method for nonconvex nonsmooth stochastic optimization. *Cybernetics and Systems Analysis*, 34(2):196–215, 1998.

[23] Allen A Goldstein. Optimization of lipschitz continuous functions. *Mathematical Programming*, 13(1):14–22, 1977.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, NV, 2016.

[25] Tianjin Huang, Ziquan Zhu, Gaojie Jin, Lu Liu, Zhangyang Wang, and Shiwei Liu. Spam: Spike-aware adam with momentum reset for stable llm training. In *International Conference on Learning Representations (ICLR)*, 2025.

[26] Fanfan Ji and Xiao-Tong Yuan. Derandomized online-to-non-convex conversion for stochastic weakly convex optimization. In *International Conference on Learning Representations (ICLR)*, 2026.

[27] Ruichen Jiang, Zakaria Mhammedi, Mehryar Mohri, and Aryan Mokhtari. Adaptive matrix online learning through smoothing with guarantees for nonsmooth nonconvex optimization. In *Conference on Learning Theory (COLT)*, pages 3782–3824, 2026.

[28] Michael Jordan, Guy Kornowski, Tianyi Lin, Ohad Shamir, and Manolis Zampetakis. Deterministic nonsmooth nonconvex optimization. In *Conference on Learning Theory*, pages 4570–4597. PMLR, 2023.

[29] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, San Diego, CA, 2015.

[30] Siyu Kong and Adrian S Lewis. Lipschitz minimization and the goldstein modulus. *Mathematical Programming*, pages 1–30, 2025.

[31] Guy Kornowski. The complexity of finding stationary points in nonsmooth nonconvex optimization. *arXiv preprint arXiv:2609.17780*, 2026.

[32] Guy Kornowski and Ohad Shamir. Oracle complexity in nonsmooth nonconvex optimization. *Journal of Machine Learning Research*, 23:1–43, 2022.

[33] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. *Technical Report, University of Toronto*, 2009.

[34] Jiajin Li and Siyu Pan. Optimal deterministic oracle complexity for weakly convex optimization. *arXiv preprint arXiv:2608.03246*, 2026.

[35] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, New Orleans, LA, 2019.

[36] Vien V Mai and Mikael Johansson. Convergence of a stochastic gradient method with momentum for non-smooth non-convex optimization. In *International Conference on Machine Learning (ICML)*, 2020.

[37] Yurii Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. In *Dokl akad nauk Sssr*, volume 269, page 543, 1983.

[38] Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

[39] Quan Nguyen. How to set β_1, β_2 in adam: An online learning perspective. *arXiv e-prints*, pages arXiv–2510, 2025.

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

[41] Boris Teodorovich Polyak. *Introduction to optimization*. New York, Optimization Software, 1987.

[42] Spyridon Pougkakiotis and Dionysios S Kalogerias. A zeroth-order proximal stochastic gradient method for weakly convex stochastic optimization. *SIAM Journal on Scientific Computing*, 45(5):2679–2702, 2023. doi: 10.1137/22M1486306.

[43] Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018.

[44] Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

[45] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.

[46] Andrzej Ruszczynski and Wojciech Syski. Stochastic approximation method with gradient averaging for unconstrained problems. *IEEE Transactions on Automatic Control*, 28(12):1097–1105, 1983.

[47] Ilya Sutskever, James Martens, George E Dahl, and Geoffrey E Hinton. On the importance of initialization and momentum in deep learning. In *International Conference on Machine Learning*, volume 28, pages 1139–1147. JMLR. org, 2013.

[48] Lai Tian and Anthony Man-Cho So. No dimension-free deterministic algorithm computes approximate stationarities of lipschitzians. *Mathematical Programming*, 208(1):51–74, 2024.

[49] Tijmen Tieleman and Geoffrey Hinton. RmsProp: Divide the gradient by a running average of its recent magnitude. *COURSERA Neural Networks Mach. Learn*, 17, 2012.

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)*, pages 5998–6008, Long Beach, CA, 2017.

[51] Qiuhao Wang, Chin Pang Ho, and Marek Petrik. Policy gradient in robust mdps with global convergence guarantee. In *International Conference on Machine Learning*, pages 35763–35797. PMLR, 2023.

[52] Yuto Watanabe, Feng-Yi Liao, and Yang Zheng. Weak convexity and proximal bundle methods for nonsmooth policy optimization in robust control. *arXiv preprint arXiv:2609.06202*, 2026.

[53] Yang You, Igor Gitman, and Boris Ginsburg. Scaling sgd batch size to 32k for imagenet training. 2017.

[54] Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large batch optimization for deep learning: Training bert in 76 minutes. In *International Conference on Learning Representations*, 2020.

[55] Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*, 2020.

[56] Jingzhao Zhang, Hongzhou Lin, Stefanie Jegelka, Suvrit Sra, and Ali Jadbabaie. Complexity of finding stationary points of nonconvex nonsmooth functions. In *International Conference on Machine Learning*, pages 11173–11182. PMLR, 2020.

[57] Qinzi Zhang and Ashok Cutkosky. Random scaling and momentum for non-smooth nonconvex optimization. In *Forty-first International Conference on Machine Learning (ICML)*, pages 58780 – 58799. PMLR, 2024.

[58] Zhiming Zhou, Qingru Zhang, Guansong Lu, Hongwei Wang, Weinan Zhang, and Yong Yu. Adashift: Decorrelation and convergence of adaptive learning rate methods. In *International Conference on Learning Representations*, 2019.

A SOME PROPERTIES OF (μ, ϵ)-L_2-REGULARIZED STATIONARITY

We present in this appendix section some key properties of (μ, ϵ)-L_2-regularized stationarity to further motivate its definition for nonsmooth nonconvex optimization.

A.1 CONNECTIONS BETWEEN L_2-REGULARIZED AND L_2-GOLDSTEIN STATIONARITY

The following result establishes equivalence between (μ, ϵ)-L_2-regularized stationarity and (δ, ϵ)-L_2-Goldstein stationarity.

Proposition 1. *Let $\delta, \epsilon, \mu > 0$ be arbitrary positive values. Consider a G-Lipschitz function f.*

(a) If w is a (δ, ϵ)-L_2-Goldstein stationary point, then it admits a $\left(\frac{\epsilon^2}{\delta^2}, \sqrt{2}\epsilon\right)$-$L_2$-regularized stationary point.

(b) If w is a (μ, ϵ)-L_2-regularized stationary point, then it admits a $\left(\frac{\epsilon}{\mu^{1/2}}, \epsilon\right)$-$L_2$-Goldstein stationary point.

Proof. Part(a): Let w be a (δ, ϵ)-L_2-Goldstein stationary point of f. Then according to Definition 1, there exists $V \subset \mathbb{R}^d$ and some probability distribution of $v \sim P \in \mathcal{P}(V)$ such that $\|\mathbb{E}[g(v)]\| \leq \epsilon$ with $g(v) \in \partial f(v)$ and $\mathbb{E}[\|v - w\|^2] \leq \delta^2$. Consider $\mu = \frac{\epsilon^2}{\delta^2}$. It follows that

$$\|\partial_{+\mu} f(w)\| \leq \sqrt{\|\mathbb{E}[g(v)]\|^2 + \mu \mathbb{E}[\|v - w\|^2]} \leq \sqrt{\epsilon^2 + \mu\delta^2} = \sqrt{2}\epsilon.$$

Then by Definition 2 w must be a $\left(\frac{\epsilon^2}{\delta^2}, \sqrt{2}\epsilon\right)$-$L_2$-regularized stationary point of f.

Part(b): Let w be a (μ, ϵ)-L_2-regularized stationary point of f. Then according to Definition 2 with Equation 4, there exists subset $V \subset \mathbb{R}^d$ and some distribution of $v \sim P \in \mathcal{P}(V)$such that

$$\|\partial_{+\mu} f(w)\|^2 = \|\mathbb{E}[g(v)]\|^2 + \mu \mathbb{E}[\|v - w\|^2] \leq \epsilon^2,$$

where $g(v) \in \partial f(v)$ almost surely. Consequently, we have $\|\mathbb{E}[g(v)]\| \leq \epsilon$ and $\mathbb{E}[\|v - w\|^2] \leq \frac{\epsilon^2}{\mu}$, which implies that w is a $\left(\frac{\epsilon}{\mu^{1/2}}, \epsilon\right)$-$L_2$-Goldstein stationary point. □

Remark 10. *We note that the result in Part (b) also extends to the unbiased (μ, ϵ)-L_2-regularized stationarity as defined in Definition 2 with Equation 5.*

A.2 MONOTONICITY OF $\|\partial_{+\mu} f(w)\|$ IN TERMS OF μ

We have the following lemma which shows the monotonicity of $\|\partial_{+\mu} f(w)\|$ with respect to the regularization strength parameter μ.

Proposition 2. *Let f be a Lipschitz function. Then for any $w \in \mathbb{R}^d$ and $0 < \mu_1 \leq \mu_2$, it holds that $\|\partial_{+\mu_1} f(w)\| \leq \|\partial_{+\mu_2} f(w)\|$.*

Proof. Let $\varepsilon > 0$ be arbitrary. By Definition 2 we know that there exists a subset $V_2(\varepsilon) \subset \mathbb{R}^d$ and some distribution of $v \sim P \in \mathcal{P}(V_2(\varepsilon))$ such that

$$\|\partial_{+\mu_2} f(w)\| \geq \sqrt{\|\mathbb{E}\left[g(v)\right]\|^2 + \mu_2 \mathbb{E}\left[\|v - w\|^2\right]} - \varepsilon.$$

Again, by Definition 2 and the condition $\mu_1 \leq \mu_2$ we can see that

$$\begin{aligned}
\|\partial_{+\mu_1} f(w)\| \leq & \sqrt{\|\mathbb{E}\left[g(v)\right]\|^2 + \mu_1 \mathbb{E}\left[\|v - w\|^2\right]} \\
\leq & \sqrt{\|\mathbb{E}\left[g(v)\right]\|^2 + \mu_2 \mathbb{E}\left[\|v - w\|^2\right]} \\
\leq & \|\partial_{+\mu_2} f(w)\| + \varepsilon.
\end{aligned}$$

By noting that ε can be arbitrarily small, we must have $\|\partial_{+\mu_1} f(w)\| \leq \|\partial_{+\mu_2} f(w)\|$. □

A.3 IMPLICATIONS FOR SMOOTH OBJECTIVES

A core motivation for introducing L_2-regularized stationarity is to simplify the analysis of SGDM while preserving the favorable properties of Goldstein stationarity. This motivation is analogous to that of the (c, ϵ)-stationarity [57], which has proven effective for analyzing O2NC methods. In parallel to [57, Lemma 2.3], we next demonstrate that L_2-regularized stationary points reduce to Clarke stationary points under smooth or second-order smooth objective functions.

Proposition 3. *Suppose f is L-smooth. If $\|\partial_{+\mu} f(w)\| \le \epsilon$ for $\mu = L^2$, then $\|\nabla f(w)\| \le 2\epsilon$.*

Proof. Since $\|\partial_{+\mu} f(w)\| \le \epsilon$ and $\mu = L^2$, by definition it holds that there exist a set $V \subset \mathbb{R}^d$ and a probability measure $v \sim P \in \mathcal{P}(V)$ such that

$$\|\mathbb{E}[\nabla f(v)]\|^2 + L^2 \mathbb{E}\left[\|v - w\|^2\right] \le \epsilon^2,$$

which immediately implies

$$\|\mathbb{E}[\nabla f(v)]\| \le \epsilon, \quad \mathbb{E}[\|v - w\|] \le \sqrt{\mathbb{E}[\|v - w\|^2]} \le \frac{\epsilon}{L}.$$

It follows that

$$\begin{aligned}\|\nabla f(w)\| \le& \|\nabla f(w) - \mathbb{E}[\nabla f(v)]\| + \|\mathbb{E}[\nabla f(v)]\| \\ \le& \mathbb{E}[\|\nabla f(w) - \nabla f(v)\|] + \|\mathbb{E}[\nabla f(v)]\| \\ \le& L\mathbb{E}[\|w - v\|] + \|\mathbb{E}[\nabla f(v)]\| \le 2\epsilon,\end{aligned}$$

where the last but one step uses the L-smoothness of f. This proves the desired bound. □

Remark 11. *In view of Proposition 3, with $\mu = \mathcal{O}(1)$, our $\mathcal{O}(\mu^{1/2}\epsilon^{-4})$ rate established for SGDM implies an optimal $\mathcal{O}(\epsilon^{-4})$ rate for smooth objectives in Clarke stationarity [6].*

Proposition 4. *Suppose f is H-second-order-smooth.*

(a) If $\|\partial_{+\mu} f(w)\| \le \epsilon$ for $\mu = \frac{H\epsilon}{2}$, then there exists a set $V \subset \mathbb{R}^d$ and a distribution $v \sim P \in \mathcal{P}(V)$ such that $\|\nabla f(\bar{v})\| \le 2\epsilon$, where $\bar{v} = \mathbb{E}[v]$, and $\|w - \bar{v}\| \le \sqrt{\frac{2\epsilon}{H}}$.

(b) If $\left\|\partial^{\star}_{+\mu} f(w)\right\| \le \epsilon$ for $\mu = \frac{H\epsilon}{2}$, then $\|\nabla f(w)\| \le 2\epsilon$.

Proof. Part(a): First, we consider the condition $\|\partial_{+\mu} f(w)\| \le \epsilon$ with $\mu = \frac{H\epsilon}{2}$. Under this condition, there must exist a set $V \subset \mathbb{R}^d$ and a probability distribution $v \sim P \in \mathcal{P}(V)$ such that

$$\|\mathbb{E}[\nabla f(v)]\|^2 + \frac{H\epsilon}{2}\mathbb{E}\left[\|v - w\|^2\right] \le \epsilon^2,$$

which immediately implies that

$$\|\mathbb{E}[\nabla f(v)]\| \le \epsilon, \quad \mathbb{E}[\|v - w\|] \le \sqrt{\mathbb{E}[\|v - w\|^2]} \le \sqrt{\frac{2\epsilon}{H}}.$$

Let $\bar{v} = \mathbb{E}[v]$. By Taylor expansion, there exists some $u \in \mathrm{conv}\{w, v\}$ such that $\nabla f(v) = \nabla f(\bar{v}) + \nabla^2 f(\bar{v})(v - \bar{v}) + \frac{1}{2}(v - \bar{v})^\top \nabla^3 f(u)(v - \bar{v})$. Consequently, we can show that

$$\begin{aligned}\|\nabla f(\bar{v})\| \le& \|\mathbb{E}[\nabla f(v)]\| + \|\mathbb{E}[\nabla f(\bar{v})] - \nabla f(v)\| \\ =& \|\mathbb{E}[\nabla f(v)]\| + \|\mathbb{E}[\nabla f(v) - \nabla f(\bar{v})]\| \\ =& \|\mathbb{E}[\nabla f(v)]\| + \left\|\mathbb{E}\left[\nabla^2 f(\bar{v})(v - \bar{v}) + \frac{1}{2}(v - \bar{v})^\top \nabla^3 f(u)(v - \bar{v})\right]\right\| \\ =& \|\mathbb{E}[\nabla f(v)]\| + \left\|\mathbb{E}\left[\frac{1}{2}(v - \bar{v})^\top \nabla^3 f(u)(v - \bar{v})\right]\right\| \\ \overset{\zeta_1}{\le}& \|\mathbb{E}[\nabla f(v)]\| + \mathbb{E}\left[\frac{H}{2}\|v - \bar{v}\|^2\right] \\ \le& \|\mathbb{E}[\nabla f(v)]\| + \frac{H}{2}\mathbb{E}\left[\|v - w\|^2\right] \le 2\epsilon,\end{aligned}$$

where the step "ζ_1" uses the assumption that f is H-second-order-smooth. Additionally, we note that $\|w - \bar{v}\| \leq \mathbb{E}[\|w - v\|] \leq \sqrt{\frac{2\epsilon}{H}}$.

Part(b): Next, we consider the condition $\left\|\partial^{\star}_{+\mu} f(w)\right\| \leq \epsilon$ with $\mu = \frac{H\epsilon}{2}$. In this case, since $\mathbb{E}[v] = w$, following the same reasoning as above we can show that $\|\nabla f(w)\| \leq 2\epsilon$. □

Remark 12. *For the SGDM algorithm and its variants considered in this paper, the set $V \subset \mathbb{R}^d$ and a probability measure $v \sim P \in \mathcal{P}(V)$ appearing in part (a) of Proposition 4 are typically induced by the intermediate iterates generated by SGDM. This point will become clearer in the subsequent analysis.*

Remark 13. *In view of Proposition 4, setting $\mu = \mathcal{O}(\epsilon)$, our derived $\mathcal{O}(\mu^{1/2}\epsilon^{-4})$ iteration complexity for SGDM yields the optimal $\mathcal{O}(\epsilon^{-3.5})$ rate for second-order smooth objectives, as established in [7].*

B Preliminary lemmas

B.1 Additional notation

We start by introducing some key quantities that feature prominently in our analysis. For any $t \geq 1$, by Assumption 2 we have $g(w_t) := \mathbb{E}[\hat{g}(w_t; z_t) \mid \mathcal{F}_{t-1}] \in \partial f(w_t)$. For any index pair (t', t) satisfying $t' \leq t - 1$, let us consider the following EMA of subgradients $\{g(w_\tau)\}_{\tau=t'+1}^{t}$:

$$\bar{g}_{[t'+1:t]} := \frac{1}{1-(1-\beta)^{t-t'}} \sum_{\tau=t'+1}^{t} (1-\beta)^{t-\tau} \beta g(w_\tau) = \sum_{\tau=t'+1}^{t} q_{\tau,[t'+1:t]} g(w_\tau),$$

where $q_{\tau,[t'+1:t]} := \frac{(1-\beta)^{t-\tau}\beta}{1-(1-\beta)^{t-t'}}$ is the weight of $g(w_\tau)$ for any $\tau \in [t'+1:t]$. Intuitively, $\bar{g}_{[t'+1:t]}$ is convex combination of past subgradients desirable for L_2-regularized stationarity analysis. Correspondingly, the EMA of squared distances for the sequence $\{w_\tau\}_{\tau=t'+1}^{t}$ with respect to a target vector w is defined as:

$$V^2_{[t'+1:t]}(w) = \sum_{\tau=t'+1}^{t} q_{\tau,[t'+1:t]} \|w_\tau - w\|^2. \tag{9}$$

The quantity $V^2_{[t'+1:t]}(w)$ is treated as a measure of the iterate variance around w, and is likewise employed for the L_2-regularized stationarity analysis. Finally, we define the following quantity associated with the squared-geometric-CDF sampling strategy used in our algorithms:

$$A_{[t'+1:t]} := \sum_{\tau=t'+1}^{t} (1-(1-\beta)^{\tau-t'})^2. \tag{10}$$

B.2 Two key lemmas

We now derive a generic upper bound for the weighted sum of $\left\|\bar{g}_{[t_0+1:t]}\right\|^2$ and $V^2_{[t_0+1:t]}(w_t)$, which plays a fundamental role in the subsequent stationarity analysis of Algorithm 1 across a range of learning rate schedules. Let us consider the filtration $\mathcal{F}_t = \mathcal{S}\{z_1, z_2, ..., z_t\}$ where $\mathcal{S}\{\cdot\}$ denotes the sigma field. In general, we allow the learning rate η_t to be stochastic and $\mathcal{F}_t$-measurable.

Lemma 1. *Suppose Assumption 1 and 2 hold. Consider some $T \geq 1$ and $t_0 < T$. Let $\{a_t\}_{t\in[T]}$ be a deterministic sequence such that $a_t \leq \eta_t \leq \eta_{t-1}$ for all $t \in [T]$ almost surely. Assume $\eta_{t_0} \leq \frac{1}{2\rho}$ and $\beta \leq 0.5$. Then it holds that*

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=t_0+1}^{T} \left(1-(1-\beta)^{t-t_0}\right)^2 \left(a_t^2 \left\|\bar{g}_{[t_0+1:t]}\right\|^2 + \frac{\beta^2}{6(1-\beta)} V^2_{[t_0+1:t]}(w_t)\right)\right] \\
\leq &\mathbb{E}\left[\frac{6\eta_{t_0}^2 \|m_{t_0}\|^2}{\beta} + 4\eta_{t_0}\Delta f(w_{t_0}) + 6\beta \sum_{t=t_0+1}^{T} \eta_{t-1}^2 \|\hat{g}(w_t; z_t)\|^2\right].
\end{aligned}$$

Proof. See Appendix B.4 for a proof of this result. □

Remark 14. *In fact, the weights* $(1-(1-\beta)^{t-t_0})^2$ *appearing on the left-hand side of the first bound in Lemma 1 motivate the squared-geometric-CDF sampling strategy used for output construction in Algorithm 1.*

Lemma 2. *Suppose that Assumption 1 and 2 hold. Consider some* $T \geq 1$ *and* $t_0 < T$*. Let* $\{a_t\}_{t\in[T]}$ *be a deterministic sequence such that* $a_t \leq \eta_t \leq \eta_{t-1}$ *for all* $t \in [T]$ *holds almost surely. Suppose that* $\eta_{t_0} \leq \frac{1}{2\rho}$ *and* $\beta \leq 0.5$*. Then it holds that*

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=t_0+1}^{T}\sum_{\tau=t+1}^{T}\left(1-(1-\beta)^{\tau-t}\right)^2\left(a_\tau^2\left\|\bar{g}_{[t+1:\tau]}\right\|^2+\frac{\beta^2}{6(1-\beta)}V^2_{[t+1:\tau]}(w_\tau)\right)\right]\\
\leq&\mathbb{E}\left[\frac{6\eta_{t_0}^2}{\beta^2}\|m_{t_0}\|^2+\frac{12\eta_{t_0}\Delta f(w_{t_0})}{\beta}+6\sum_{t=t_0+1}^{T}\eta_{t-1}^2\|\hat{g}(w_t;z_t)\|^2+4\sum_{t=t_0+1}^{T}\eta_t\Delta f(w_t)\right.\\
&\left.+6\beta\sum_{t=t_0+1}^{T}\sum_{\tau=t+1}^{T}\eta_{\tau-1}^2\|\hat{g}(w_\tau;z_\tau)\|^2\right].
\end{aligned}$$

Proof. See Appendix B.5 for a proof of this result. □

B.3 Some technical lemmas

We further present a few technical lemmas to be used in the proof of Lemma 1. The following lemma quantifies the progress made by the algorithm from the momentum reset time step t_0 to any given time step $t > t_0$.

Lemma 3. *Assume that Assumption 1 holds. Consider an index* $t \geq 1$ *and* $t_0 \in [0 : t-1]$*. Suppose that* $\eta_{t_0} \leq \frac{1}{2\rho}$*,* $\eta_\tau \leq \eta_{\tau-1}$ *for all* $\tau \in [t_0+1 : t]$*, and* $\beta \leq 0.5$*. Then it holds that*

$$\begin{aligned}
&\eta_t^2\|m_t\|^2+\sum_{\tau=t_0+1}^{t}\frac{(1-\beta)^{t-\tau}\beta}{4}\|w_\tau-w_{\tau-1}\|^2\\
\leq&(1-\beta)^{t-t_0}\eta_{t_0}^2\|m_{t_0}\|^2+2\beta(1-\beta)\sum_{\tau=t_0+1}^{t}(1-\beta)^{t-\tau}\eta_{\tau-1}\left(f(w_{\tau-1})-f(w_\tau)\right)\\
&+\beta^2\sum_{\tau=t_0+1}^{t}(1-\beta)^{t-\tau}\eta_{\tau-1}^2\|\hat{g}(w_\tau;z_\tau)\|^2+2\beta(1-\beta)\sum_{\tau=t_0+1}^{t}(1-\beta)^{t-\tau}\eta_{\tau-1}^2\langle m_{\tau-1},u_\tau\rangle.
\end{aligned}\tag{11}$$

Proof. See Appendix D.1 for its proof. □

The following lemma provides a lower bound for the squared norm of momentum term m_t appearing on the LHS of the bound in Lemma 3.

Lemma 4. *Assume that Assumption 1 holds. Then for any* $t \geq 1$ *and* $t_0 \in [0, t-1]$*, it holds that*

$$\|m_t\|^2 \geq \frac{(1-(1-\beta)^{t-t_0})^2}{2}\left\|\bar{g}_{[t_0+1:t]}\right\|^2-2(1-\beta)^{2(t-t_0)}\|m_{t_0}\|^2-2\beta^2\left\|\sum_{\tau=t_0+1}^{t}(1-\beta)^{t-\tau}u_\tau\right\|^2.$$

Proof. See Appendix D.2 for its proof. □

We further establish in the following lemma a lower bound for the second term on the LHS of the bound in Lemma 3.

Lemma 5. *Suppose that $\beta \leq 0.5$. Then for any $T \geq 1$ and $t_0 \in [0, : T-1]$, it holds that*

$$\sum_{t=t_0+1}^{T} \left(1-(1-\beta)^{t-t_0}\right) V^2_{[t_0+1:t]}(w_t) \leq \frac{3(1-\beta)}{\beta} \sum_{t=t_0+1}^{T} \sum_{\tau=t_0+1}^{t} (1-\beta)^{t-\tau} \|w_\tau - w_{\tau-1}\|^2.$$

Proof. See Appendix D.3 for its proof. □

Let us denote $u_t := \hat{g}(w_t; z_t) - g(w_t)$. We have the following concentration result.

Lemma 6. *Under Assumption 1 and Assumption 2. Then for any index $t \geq 1$ and $t_0 \in [0 : t-1]$, it holds that*

$$\mathbb{E}\left[\eta_{t-1}^2 \langle m_{t-1}, u_t \rangle\right] = 0,$$

and

$$\mathbb{E}\left[\left\|\sum_{\tau=t_0+1}^{t} (1-\beta)^{t-\tau} u_\tau\right\|^2\right] \leq \mathbb{E}\left[\sum_{\tau=t_0+1}^{t} (1-\beta)^{t-\tau} \|\hat{g}(w_\tau; z_\tau)\|^2\right].$$

Proof. See Appendix D.4 for its proof. □

The following lemma provides lower and upper bounds for $A_{[t+1:T]}$ defined in 10.

Lemma 7. *For any $t \in [0 : T-1]$, it holds that $T - t - \frac{2}{\beta} \leq A_{[t+1:T]} \leq T - t$.*

Proof. See Appendix D.5 for a proof of this result. □

Lemma 8. *Suppose that the learning rate sequence $\{\eta_t\}_{t\geq 1}$ is non-increasing. For any pair of integers $t_1 < t_2$, it holds that*

$$\sum_{t=t_1+1}^{t_2} \sum_{\tau=t_1+1}^{t} (1-\beta)^{t-\tau} \eta_{\tau-1}(f(w_{\tau-1}) - f(w_\tau)) \leq \frac{1-(1-\beta)^{t_2-t_1}}{\beta} \eta_{t_1} \Delta f(w_{t_1}),$$

and

$$\sum_{t=t_1+1}^{t_2} \sum_{\tau=t_1+1}^{t} (1-\beta)^{t-\tau} \eta_{\tau-1}^2 \|\hat{g}(w_\tau; z_\tau)\|^2 \leq \frac{1-(1-\beta)^{t_2-t_1}}{\beta} \sum_{t=t_1+1}^{t_2} \eta_{t-1}^2 \|\hat{g}(w_t; z_t)\|^2.$$

Proof. See Appendix D.6 for a proof of this result. □

B.4 Proof of Lemma 1

With the technical lemmas stated in the previous subsection available, we are in the position to prove Lemma 1.

Proof. Consider a fixed index $t \in [t_0+1 : T]$. Since $\eta_t \geq a_t$, it follows from Lemma 4 that

$$\begin{aligned}
&\eta_t^2 \|m_t\|^2 \\
\geq &a_t^2 \|m_t\|^2 \\
\geq &\frac{(1-(1-\beta)^{t-t_0})^2 a_t^2}{2} \left\|\bar{g}_{[t_0+1:t]}\right\|^2 - 2(1-\beta)^{2(t-t_0)} a_t^2 \|m_{t_0}\|^2 - 2\beta^2 a_t^2 \left\|\sum_{\tau=t_0+1}^{t} (1-\beta)^{t-\tau} u_\tau\right\|^2.
\end{aligned}$$

By substituting the above bound into the bound 11 in Lemma 3 and rearranging the terms we get

$$
\begin{aligned}
&\frac{(1-(1-\beta)^{t-t_0})^2 a_t^2}{2}\left\|\bar{g}_{[t_0+1:t]}\right\|^2+\frac{\beta}{4}\sum_{\tau=t_0+1}^{t}(1-\beta)^{t-\tau}\|w_\tau-w_{\tau-1}\|^2\\
\leq&(1-\beta)^{t-t_0}\eta_{t_0}^2\|m_{t_0}\|^2+2\beta(1-\beta)\sum_{\tau=t_0+1}^{t}(1-\beta)^{t-\tau}\eta_{\tau-1}\left(f(w_{\tau-1})-f(w_\tau)\right)\\
&+\beta^2\sum_{\tau=t_0+1}^{t}(1-\beta)^{t-\tau}\eta_{\tau-1}^2\|\hat{g}(w_\tau;z_\tau)\|^2+2(1-\beta)\beta\sum_{\tau=t_0+1}^{t}(1-\beta)^{t-\tau}\eta_{\tau-1}^2\langle m_{\tau-1},u_\tau\rangle\\
&+2a_t^2(1-\beta)^{2(t-t_0)}\|m_{t_0}\|^2+2\beta^2a_t^2\left\|\sum_{\tau=t_0+1}^{t}(1-\beta)^{t-\tau}u_\tau\right\|^2\\
\leq&3(1-\beta)^{t-t_0}\eta_{t_0}^2\|m_{t_0}\|^2+2\beta(1-\beta)\sum_{\tau=t_0+1}^{t}(1-\beta)^{t-\tau}\eta_{\tau-1}\left(f(w_{\tau-1})-f(w_\tau)\right)\\
&+\beta^2\sum_{\tau=t_0+1}^{t}(1-\beta)^{t-\tau}\eta_{\tau-1}^2\|\hat{g}(w_\tau;z_\tau)\|^2+2(1-\beta)\beta\sum_{\tau=t_0+1}^{t}(1-\beta)^{t-\tau}\eta_{\tau-1}^2\langle m_{\tau-1},u_\tau\rangle\\
&+2\beta^2a_t^2\left\|\sum_{\tau=t_0+1}^{t}(1-\beta)^{t-\tau}u_\tau\right\|^2,
\end{aligned}
$$

where in the last step we have used the almost sure relation $a_t \leq \eta_t \leq \eta_{t_0}$ for any $t \geq t_0$, and $\beta \in (0,1)$. Taking expectation on both sides of the above inequality and applying Lemma 6 yield

$$
\begin{aligned}
&\mathbb{E}\left[\frac{(1-(1-\beta)^{t-t_0})^2 a_t^2}{2}\left\|\bar{g}_{[t_0+1:t]}\right\|^2+\frac{\beta}{4}\sum_{\tau=t_0+1}^{t}(1-\beta)^{t-\tau}\|w_\tau-w_{\tau-1}\|^2\right]\\
\leq&\mathbb{E}\left[3(1-\beta)^{t-t_0}\eta_{t_0}^2\|m_{t_0}\|^2+2\beta(1-\beta)\sum_{\tau=t_0+1}^{t}(1-\beta)^{t-\tau}\eta_{\tau-1}\left(f(w_{\tau-1})-f(w_\tau)\right)\right.\\
&\left.+\beta^2\sum_{\tau=t_0+1}^{t}(1-\beta)^{t-\tau}\eta_{\tau-1}^2\|\hat{g}(w_\tau;z_\tau)\|^2+2\beta^2a_t^2\sum_{\tau=t_0+1}^{t}(1-\beta)^{t-\tau}\|\hat{g}(w_\tau;z_\tau)\|^2\right]\\
\leq&\mathbb{E}\left[3(1-\beta)^{t-t_0}\eta_{t_0}^2\|m_{t_0}\|^2+2\beta(1-\beta)\sum_{\tau=t_0+1}^{t}(1-\beta)^{t-\tau}\eta_{\tau-1}\left(f(w_{\tau-1})-f(w_\tau)\right)\right.\\
&\left.+3\beta^2\sum_{\tau=t_0+1}^{t}(1-\beta)^{t-\tau}\eta_{\tau-1}^2\|\hat{g}(w_\tau;z_\tau)\|^2\right],
\end{aligned}
$$

where in the last step we have used the almost sure relation $a_t \leq \eta_t \leq \eta_{t'}$ for any $t \geq t'$. Summing the above inequality over $t = t_0 + 1, \cdots, T$ with a little algebraic manipulation yields

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=t_0+1}^{T}\left(1-(1-\beta)^{t-t_0}\right)^2 a_t^2\left\|\bar{g}_{[t_0+1:t]}\right\|^2+\frac{\beta}{2}\underbrace{\sum_{t=t_0+1}^{T}\sum_{\tau=t_0+1}^{t}(1-\beta)^{t-\tau}\|w_\tau-w_{\tau-1}\|^2}_{I_1}\right]\\
\leq&\mathbb{E}\left[\frac{6\eta_{t_0}^2\|m_{t_0}\|^2}{\beta}+4(1-\beta)\beta\underbrace{\sum_{t=t_0+1}^{T}\sum_{\tau=t_0+1}^{t}(1-\beta)^{t-\tau}\eta_{\tau-1}\left(f(w_{\tau-1})-f(w_\tau)\right)}_{I_2}\right.\\
&\left.+6\beta^2\underbrace{\sum_{t=t_0+1}^{T}\sum_{\tau=t_0+1}^{t}(1-\beta)^{t-\tau}\eta_{\tau-1}^2\|\hat{g}(w_\tau;z_\tau)\|^2}_{I_3}\right].
\end{aligned} \tag{12}$$

We next proceed to bound the components I_1, I_2 and I_3 in the above. For the term I_1, we can apply Lemma 5 to obtain

$$\begin{aligned}
I_1=&\sum_{t=t_0+1}^{T}\sum_{\tau=t_0+1}^{t}(1-\beta)^{t-\tau}\|w_\tau-w_{\tau-1}\|^2\\
\geq&\frac{\beta}{3(1-\beta)}\sum_{t=t_0+1}^{T}\left(1-(1-\beta)^{t-t_0}\right)V_{[t_0+1:t]}^2(w_t)\\
\geq&\frac{\beta}{3(1-\beta)}\sum_{t=t_0+1}^{T}\left(1-(1-\beta)^{t-t_0}\right)^2V_{[t_0+1:t]}^2(w_t).
\end{aligned} \tag{13}$$

Regarding the term I_2, we can apply the first bound in Lemma 8 to show that

$$\begin{aligned}
I_2=&\sum_{t=t_0+1}^{T}\sum_{\tau=t_0+1}^{t}(1-\beta)^{t-\tau}\eta_{\tau-1}(f(w_{\tau-1})-f(w_\tau))\\
\leq&\frac{(1-(1-\beta)^{T-t_0})\eta_{t_0}\Delta f(w_{t_0})}{\beta}\leq\frac{\eta_{t_0}\Delta f(w_{t_0})}{\beta}.
\end{aligned} \tag{14}$$

Further, the terms I_3 can be bounded via applying the second bound in Lemma 8 as follows:

$$\begin{aligned}
I_3=&\sum_{t=t_0+1}^{T}\sum_{\tau=t_0+1}^{t}(1-\beta)^{t-\tau}\eta_{\tau-1}^2\|\hat{g}(w_\tau;z_\tau)\|^2\\
\leq&\frac{1-(1-\beta)^{T-t_0}}{\beta}\sum_{t=t_0+1}^{T}\eta_{t-1}^2\|\hat{g}(w_t;z_t)\|^2\leq\frac{1}{\beta}\sum_{t=t_0+1}^{T}\eta_{t-1}^2\|\hat{g}(w_t;z_t)\|^2.
\end{aligned} \tag{15}$$

By substituting 13, 14, 15 to 12 we obtain

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=t_0+1}^{T}\left(1-(1-\beta)^{t-t_0}\right)^2\left(a_t^2\left\|\bar{g}_{[t_0+1:t]}\right\|^2+\frac{\beta^2}{6(1-\beta)}V_{[t_0+1:t]}^2(w_t)\right)\right]\\
\leq&\mathbb{E}\left[\frac{6\eta_{t_0}^2\|m_{t_0}\|^2}{\beta}+4\eta_{t_0}\Delta f(w_{t_0})+6\beta\sum_{t=t_0+1}^{T}\eta_{t-1}^2\|\hat{g}(w_t;z_t)\|^2\right],
\end{aligned}$$

which is the desired bound. The proof is completed. □

B.5 Proof of Lemma 2

Proof. We first note that $\eta_t \leq \frac{1}{2\rho}$ for all $t \geq t_0$ due to the almost sure monotonicity of η_t. Then, it follows from Lemma 1 that the following holds for any fixed $t \geq t_0$:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{\tau=t+1}^{T}\left(1-(1-\beta)^{\tau-t}\right)^2\left(\eta_\tau^2\left\|\bar{g}_{[t+1:\tau]}\right\|^2+\frac{\beta^2}{6(1-\beta)}V^2_{[t+1:\tau]}(w_\tau)\right)\right]\\
\leq&\mathbb{E}\left[\frac{6\eta_t^2\|m_t\|^2}{\beta}+4\eta_t\Delta f(w_t)+6\beta\sum_{\tau=t+1}^{T}\eta_{\tau-1}^2\|\hat{g}(w_\tau;z_\tau)\|^2\right],
\end{aligned}$$

Summing the above inequality over $t \in [t_0+1, T]$ we get

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=t_0+1}^{T}\sum_{\tau=t+1}^{T}\left(1-(1-\beta)^{\tau-t}\right)^2\left(\eta_\tau^2\left\|\bar{g}_{[t+1:\tau]}\right\|^2+\frac{\beta^2}{6(1-\beta)}V^2_{[t+1:\tau]}(w_\tau)\right)\right]\\
\leq&\mathbb{E}\left[\frac{6}{\beta}\underbrace{\sum_{t=t_0+1}^{T}\eta_t^2\|m_t\|^2}_{I_1}+4\sum_{t=t_0+1}^{T}\eta_t\Delta f(w_t)+6\beta\sum_{t=t_0+1}^{T}\sum_{\tau=t+1}^{T}\eta_{\tau-1}^2\|\hat{g}(w_\tau;z_\tau)\|^2\right].
\end{aligned}\tag{16}$$

We next upper bound the expectation of the term I_1 in the above inequality. To this end, we can apply Lemma 3 to show that for all $t \geq t_0+1$,

$$\begin{aligned}
&\eta_t^2\|m_t\|^2\\
\leq&(1-\beta)^{t-t_0}\eta_{t_0}^2\|m_{t_0}\|^2+2\beta(1-\beta)\sum_{\tau=t_0+1}^{t}(1-\beta)^{t-\tau}\eta_{\tau-1}\left(f(w_{\tau-1})-f(w_\tau)\right)\\
&+\beta^2\sum_{\tau=t_0+1}^{t}(1-\beta)^{t-\tau}\eta_{\tau-1}^2\|\hat{g}(w_\tau;z_\tau)\|^2+2\beta(1-\beta)\sum_{\tau=t_0+1}^{t}(1-\beta)^{t-\tau}\eta_{\tau-1}^2\langle m_{\tau-1},u_\tau\rangle.
\end{aligned}$$

Taking expectation on both sides of the above inequality and applying Lemma 6 yield

$$\begin{aligned}
&\mathbb{E}\left[\eta_t^2\|m_t\|^2\right]\\
\leq&\mathbb{E}\left[(1-\beta)^{t-t_0}\eta_{t_0}^2\|m_{t_0}\|^2+2\beta(1-\beta)\sum_{\tau=t_0+1}^{t}(1-\beta)^{t-\tau}\eta_{\tau-1}\left(f(w_{\tau-1})-f(w_\tau)\right)\right.\\
&\left.+\beta^2\sum_{\tau=t_0+1}^{t}(1-\beta)^{t-\tau}\eta_{\tau-1}^2\|\hat{g}(w_\tau;z_\tau)\|^2\right].
\end{aligned}$$

By summing the above inequality over $t = t_0 + 1, \cdots, T$ we get

$$
\begin{aligned}
&\mathbb{E}[I_1] \\
=&\mathbb{E}\left[\sum_{t=t_0+1}^{T} \eta_t^2 \|m_t\|^2\right] \\
\leq&\mathbb{E}\left[\sum_{t=t_0+1}^{T} (1-\beta)^{t-t_0}\eta_{t_0}^2\|m_{t_0}\|^2 + 2\beta(1-\beta)\sum_{t=t_0+1}^{T}\sum_{\tau=t_0+1}^{t}(1-\beta)^{t-\tau}\eta_{\tau-1}\left(f(w_{\tau-1}) - f(w_\tau)\right)\right. \\
&\left.+\beta^2\sum_{t=t_0+1}^{T}\sum_{\tau=t_0+1}^{t}(1-\beta)^{t-\tau}\eta_{\tau-1}^2\|\hat{g}(w_\tau;z_\tau)\|^2\right] \\
\leq&\mathbb{E}\left[\frac{(1-\beta)(1-(1-\beta)^{T-t_0})}{\beta}\eta_{t_0}^2\|m_{t_0}\|^2 + 2\beta(1-\beta)\frac{1-(1-\beta)^{T-t_0}}{\beta}\eta_{t_0}\Delta f(w_{t_0})\right. \\
&\left.+\beta^2\frac{1-(1-\beta)^{T-t_0}}{\beta}\sum_{t=t_0+1}^{T}\eta_{t-1}^2\|\hat{g}(w_t;z_t)\|^2\right] \\
\leq&\mathbb{E}\left[\frac{\eta_{t_0}^2}{\beta}\|m_{t_0}\|^2 + 2\eta_{t_0}\Delta f(w_{t_0}) + \beta\sum_{t=t_0+1}^{T}\eta_{t-1}^2\|\hat{g}(w_t;z_t)\|^2\right],
\end{aligned}
\tag{17}
$$

where in the last but one inequality we have used Lemma 8.

By substituting equation 17 into equation 16 we arrive at the following desired bound:

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=t_0+1}^{T}\sum_{\tau=t+1}^{T}\left(1-(1-\beta)^{\tau-t}\right)^2\left(a_\tau^2\left\|\bar{g}_{[t+1:\tau]}\right\|^2 + \frac{\beta^2}{6(1-\beta)}V_{[t+1:\tau]}^2(w_\tau)\right)\right] \\
\leq&\mathbb{E}\left[\frac{6\eta_{t_0}^2}{\beta^2}\|m_{t_0}\|^2 + \frac{12\eta_{t_0}\Delta f(w_{t_0})}{\beta} + 6\sum_{t=t_0+1}^{T}\eta_{t-1}^2\|\hat{g}(w_t;z_t)\|^2 + 4\sum_{t=t_0+1}^{T}\eta_t\Delta f(w_t)\right. \\
&\left.+6\beta\sum_{t=t_0+1}^{T}\sum_{\tau=t+1}^{T}\eta_{\tau-1}^2\|\hat{g}(w_\tau;z_\tau)\|^2\right].
\end{aligned}
$$

The proof is completed. □

C Proofs in Section 3

Based on the generic upper bounds in Lemma 1 and Lemma 2, we can prove our main convergence results of SGDM for weakly convex functions under constant, time-decaying, and adaptive learning rate schedules.

C.1 Proof of Theorem 1

Proof. Let us consider $\eta_t = a_t \equiv \eta$. Since $\eta \leq \frac{1}{2\rho}$, it must hold that $\eta_0 \leq \frac{1}{2\rho}$ and $a_t \leq \eta_t \leq \frac{1}{2\rho}$ almost surely for all $t \geq 1$, and obviously a_t is deterministic. Let us consider $t_0 = 0$ in Algorithm 1 and recall that $m_0 = 0$. Then we can apply Lemma 1 to obtain

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\left(1-(1-\beta)^t\right)^2\left(\eta^2\left\|\bar{g}_{[t]}\right\|^2 + \frac{\beta^2}{6(1-\beta)}V_{[t]}^2(w_t)\right)\right] \\
\leq&\mathbb{E}\left[4\eta\Delta f(w_0) + 6\beta\eta^2\sum_{t=1}^{T}\|\hat{g}(w_t;z_t)\|^2\right] \leq 4\eta\Delta f(w_0) + 6\beta\eta^2G^2T,
\end{aligned}
$$

where we have used the G-Lipschitzness of loss function. In view of the above, we have

$$\mathbb{E}\left[\sum_{t=1}^{T}\frac{(1-(1-\beta)^t)^2}{A_{[T]}}\left(\eta^2\left\|\bar{g}_{[t]}\right\|^2+\frac{\beta^2}{6(1-\beta)}V_{[t]}^2(w_t)\right)\right]$$
$$\leq\frac{1}{A_{[T]}}\left(4\eta\Delta f(w_0))+6\beta\eta^2G^2T\right)\leq\frac{8\eta\Delta f(w_0)}{T}+12\beta\eta^2G^2,$$

where in the last step we have used $A_{[T]}\geq T-\frac{2}{\beta}\geq T/2$ according to Lemma 7 and the condition $\beta\geq\frac{4}{T}$. Finally, based on the above inequality and the definition of random index S distributed over $[T]$ we arrive at

$$\mathbb{E}\left[\left\|\bar{g}_{[S]}\right\|^2+\frac{\beta^2}{6\eta^2(1-\beta)}V_{[S]}^2(w_S)\right]\leq\frac{8\Delta f(w_0)}{\eta T}+12\beta G^2.$$

Since $\mu\leq\mu'=\frac{\beta^2}{6\eta^2(1-\beta)}$, in view of Proposition 2 we get

$$\left\|\partial_{+\mu}f(w_S)\right\|^2\leq\left\|\partial_{+\mu'}f(w_S)\right\|^2\leq\left\|\bar{g}_{[S]}\right\|^2+\mu'V_{[S]}^2(w_S).$$

Combining the preceding two inequalities leads to the following desired bound:

$$\mathbb{E}\left[\left\|\partial_{+\mu}f(w_S)\right\|^2\right]\leq\frac{8\Delta f(w_0)}{\eta T}+12\beta G^2.$$

The proof is completed. □

C.2 Proof of Corollary 1

Proof. Under the condition of T, we must have $T\geq\frac{3}{\mu^{1/2}}+24\mu^{1/2}>8$ and

$$\eta=\frac{1}{T^{1/2}\mu^{1/4}}\leq\frac{1}{2\rho},\ \ \beta=\frac{\sqrt{6}\mu^{1/4}}{T^{1/2}}\in\left[\frac{4}{T},\frac{1}{2}\right].$$

Again with the given choice of β,η, it can be readily shown that $\frac{\beta^2}{6\eta^2(1-\beta)}\geq\frac{\beta^2}{6\eta^2}=\mu$. Then, the conditions of Theorem 1 are fulfilled in our setting, and we can thus apply it to obtain that

$$\mathbb{E}\left[\left\|\partial_{+\mu}f(w_S)\right\|^2\right]\leq\frac{8\Delta f(w_0)}{\eta T}+12\beta G^2=\frac{4\mu^{1/4}(2\Delta f(w_0)+3\sqrt{6}G^2)}{T^{1/2}}\leq\epsilon^2,$$

where in the last step we have used the condition on T. The desired bound then follows from the fact $\mathbb{E}\left[\left\|\partial_{+\mu}f(w_S)\right\|\right]\leq\sqrt{\mathbb{E}\left[\left\|\partial_{+\mu}f(w_S)\right\|^2\right]}\leq\epsilon$. This proves the desired bound. □

C.3 Proof of Theorem 2

Proof. Given the iteration budget T, let us consider $a_t\equiv\frac{\eta}{\sqrt{T+1}}$ for any $t\in[T]$. Then based on the conditions in the theorem, it must hold deterministically that $\eta_{t_0}\leq\frac{1}{2\rho}$, and $a_t\leq\eta_t\leq\eta_{t-1}$ for all

$t \in [T]$. We can apply Lemma 2 to obtain

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=t_0+1}^{T}\sum_{\tau=t+1}^{T}\left(1-(1-\beta)^{\tau-t}\right)^2\left(\frac{\eta^2}{T+1}\left\|\bar{g}_{[t+1:\tau]}\right\|^2+\frac{\beta^2}{6(1-\beta)}V^2_{[t+1:\tau]}(w_\tau)\right)\right]\\
\leq&\mathbb{E}\left[\frac{6\eta^2\|m_{t_0}\|^2}{\beta^2(t_0+1)}+\frac{12\eta\Delta f(w_{t_0})}{\beta\sqrt{t_0+1}}+6\eta^2\sum_{t=t_0+1}^{T}\frac{\|\hat{g}(w_t;z_t)\|^2}{t}+4\eta\sum_{t=t_0+1}^{T}\frac{\Delta f(w_t)}{\sqrt{t+1}}\right.\\
&\left.+6\beta\eta^2\sum_{t=t_0+1}^{T}\sum_{\tau=t+1}^{T}\frac{\|\hat{g}(w_\tau;z_\tau)\|^2}{\tau}\right]\\
\overset{\zeta_1}{\leq}&\frac{6\eta^2G^2}{\beta^2(t_0+1)}+\frac{12\eta M}{\beta\sqrt{t_0+1}}+6\eta^2\sum_{t=t_0+1}^{T}\frac{G^2}{t}+4\eta\sum_{t=t_0+1}^{T}\frac{M}{\sqrt{t+1}}+6\beta\eta^2\sum_{t=t_0+1}^{T}\sum_{\tau=t+1}^{T}\frac{G^2}{\tau}\\
\overset{\zeta_2}{\leq}&\frac{6\eta^2G^2}{\beta^2(t_0+1)}+\frac{12\eta M}{\beta\sqrt{t_0+1}}+6\eta^2G^2(1+\beta T)\sum_{t=t_0+1}^{T}\frac{1}{t}+4\eta\sum_{t=t_0+1}^{T}\frac{M}{\sqrt{t+1}}\\
\leq&\frac{6\eta^2G^2}{\beta^2(t_0+1)}+\frac{12\eta M}{\beta\sqrt{t_0+1}}+6G^2\eta^2(1+\beta T)\log\left(\frac{T}{t_0}\right)+8\eta M\sqrt{T+1}\\
\leq&\frac{6\eta^2G^2}{\beta^2(t_0+1)}+\frac{12\eta M}{\beta\sqrt{t_0+1}}+12G^2\eta^2\beta T\log\left(\frac{T}{t_0}\right)+8\eta M\sqrt{T+1},
\end{aligned}
$$

where the inequality "ζ_1" uses the fact $\|m_{t_0}\| \leq G$ due to the G-Lipschitzness of the loss function and $\Delta f(w_t) = f(w_t) - f^* \leq M$ by Assumption 3, and the inequality "ζ_2" uses the fact that $\sum_{\tau=t+1}^{T}\frac{1}{\tau} \leq (T-t)\frac{1}{t+1} \leq \frac{T}{t}$, and in the last step we used the fact $\beta T \geq 1$. Recall that $A_{[t+1:T]} := \sum_{\tau=t+1}^{T}\left(1-(1-\beta)^{\tau-t}\right)^2$ according to 10. Then, in view of the above bound we have

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=t_0+1}^{T}\frac{A_{[t+1:T]}}{\sum_{t=t_0+1}^{T}A_{[t+1:T]}}\sum_{\tau=t+1}^{T}\frac{\left(1-(1-\beta)^{\tau-t}\right)^2}{A_{[t+1:T]}}\left(\frac{\eta^2}{T+1}\left\|\bar{g}_{[t+1:\tau]}\right\|^2+\frac{\beta^2}{6(1-\beta)}V^2_{[t+1:\tau]}(w_\tau)\right)\right]\\
=&\frac{1}{\sum_{t=t_0+1}^{T}A_{[t+1:T]}}\mathbb{E}\left[\sum_{t=t_0+1}^{T}\sum_{\tau=t+1}^{T}\left(1-(1-\beta)^{\tau-t}\right)^2\left(\frac{\eta^2}{T+1}\left\|\bar{g}_{[t+1:\tau]}\right\|^2+\frac{\beta^2}{6(1-\beta)}V^2_{[t+1:\tau]}(w_\tau)\right)\right]\\
\leq&\frac{1}{\sum_{t=t_0+1}^{T}A_{[t+1:T]}}\left(\frac{6\eta^2G^2}{\beta^2(t_0+1)}+\frac{12\eta M}{\beta\sqrt{t_0+1}}+12G^2\eta^2\beta T\log\left(\frac{T}{t_0}\right)+8\eta M\sqrt{T+1}\right)\\
\leq&\frac{24}{T^2}\left(\frac{3\eta^2G^2}{\beta^2(t_0+1)}+\frac{6\eta M}{\beta\sqrt{t_0+1}}+6G^2\eta^2\beta T\log\left(\frac{T}{t_0}\right)+4\eta M\sqrt{T+1}\right),
\end{aligned}
$$

where in the last step we have used Lemma 7 and the conditions $\beta \geq \frac{24}{T}$ and $T \geq 2t_0+6$ such that $\sum_{t=t_0+1}^{T}A_{[t+1:T]} \geq \sum_{t=t_0+1}^{T}(T-t-\frac{2}{\beta}) = \frac{(T-t_0-1)(T-t_0)}{2} - \frac{2(T-t_0)}{\beta} \geq \frac{(T-t_0)^2}{3} \geq \frac{T^2}{12}$. Finally, based on the above inequality and the definition of random indices R, S we arrive at

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\bar{g}_{[R+1:S]}\right\|^2+\frac{\beta^2T}{6\eta^2(1-\beta)}V^2_{[R+1:S]}(w_S)\right]\\
\leq&\mathbb{E}\left[\left\|\bar{g}_{[R+1:S]}\right\|^2+\frac{\beta^2(T+1)}{6\eta^2(1-\beta)}V^2_{[R+1:S]}(w_S)\right]\\
\leq&\frac{24(T+1)}{T^2}\left(\frac{3G^2}{\beta^2(t_0+1)}+\frac{6M}{\beta\eta\sqrt{t_0+1}}+6G^2\beta T\log\left(\frac{T}{t_0}\right)+\frac{4M\sqrt{T+1}}{\eta}\right)\\
\leq&\frac{25}{T}\left(\frac{3G^2}{\beta^2(t_0+1)}+\frac{6M}{\beta\eta\sqrt{t_0+1}}+6G^2\beta T\log\left(\frac{T}{t_0}\right)+\frac{5M\sqrt{T}}{\eta}\right),
\end{aligned}
$$

where in the last step we have used the condition $T \geq 48$. Since $\mu \leq \mu' = \frac{\beta^2T}{6\eta^2(1-\beta)}$, in view of Proposition 2 we get that the following holds for any index pair (t,τ) satisfying $t+1 \leq \tau$:

$$
\left\|\partial_{+\mu}f(w_\tau)\right\|^2 \leq \left\|\partial_{+\mu'}f(w_\tau)\right\|^2 \leq \left\|\bar{g}_{[t+1:\tau]}\right\|^2+\mu' V^2_{[t+1:\tau]}(w_\tau).
$$

Combining the preceding two inequalities leads to the following desired bound:

$$\mathbb{E}\left[\|\partial_{+\mu} f(w_S)\|^2\right] \leq \frac{25}{T}\left(\frac{3G^2}{\beta^2(t_0+1)} + \frac{6M}{\beta\eta\sqrt{t_0+1}} + 6G^2\beta T\log\left(\frac{T}{t_0}\right) + \frac{5M\sqrt{T}}{\eta}\right).$$

The proof is completed. □

C.4 Proof of Corollary 2

Proof. Let us set $\eta = \frac{1}{\mu^{1/4}}$ in Theorem 2. Under the condition of T we must have $T \geq 3\mu^{-1/2} + 24\mu^{1/2} > 9$ and $\beta = \frac{\sqrt{6}\mu^{1/4}}{T^{1/2}} \in \left[\frac{24}{T}, \frac{1}{2}\right]$. Again with the given choice of β, η, it can be readily shown that $\frac{\beta^2 T}{6\eta^2(1-\beta)} \geq \frac{\beta^2 T}{6\eta^2} = \mu$. Let us consider $t_0 = \left\lfloor \frac{T}{3} \right\rfloor$. The condition on T ensures that

$$\eta_{t_0} = \frac{1}{\mu^{1/4}\sqrt{t_0+1}} \leq \frac{\sqrt{3}}{\mu^{1/4}T^{1/2}} \leq \frac{1}{2\rho}.$$

Under the aforementioned hyperparameter choice, all conditions required for Theorem 2 hold, which allows us to apply the corresponding bound and show that

$$\begin{aligned}
&\mathbb{E}\left[\|\partial_{+\mu} f(w_S)\|^2\right] \\
\leq &\frac{25}{T}\left(\frac{3G^2}{\beta^2(t_0+1)} + \frac{6M}{\beta\eta\sqrt{t_0+1}} + 6G^2\beta T\log\left(\frac{T}{t_0}\right) + \frac{5M\sqrt{T}}{\eta}\right) \\
\leq &\frac{25}{T}\left(\frac{3G^2}{2\mu^{1/2}} + 3\sqrt{2}M + 6\sqrt{6}G^2\mu^{1/4}\sqrt{T}\log(3) + 5M\mu^{1/4}\sqrt{T}\right) \\
= &\frac{75}{T}\left(\frac{G^2}{2\mu^{1/2}} + \sqrt{2}M\right) + \frac{25\mu^{1/4}}{T^{1/2}}\left(17G^2 + 5M\right) \leq \epsilon^2,
\end{aligned}$$

where the last "$\leq$" is due to the condition on T such that the two terms on the LHS are respectively no larger than $\frac{\epsilon^2}{2}$. The desired bound then follows from the fact $\mathbb{E}\left[\|\partial_{+\mu} f(w_S)\|\right] \leq \sqrt{\mathbb{E}\left[\|\partial_{+\mu} f(w_S)\|^2\right]} \leq \epsilon$. The proof is completed. □

C.5 Proof of Theorem 3

We first present a few technical lemmas to be used in the proof of Theorem 3.

Lemma 9. *The learning rate sequence $\{\eta_t\}_{t\geq 0}$ is non-increasing. Moreover, if the loss function is G-Lipschitz continuous and $\hat{v}_0 \leq G^2$, then it holds that*

$$\frac{\eta}{G+\varepsilon} \leq \eta_t \leq \frac{\eta}{(\sqrt{\hat{v}_0}+\varepsilon)(1-(1-\beta_1)^t)}, \ \forall t \geq 1.$$

Proof. See Appendix D.7 for a proof of this result. □

Lemma 10. *Let $\beta \in (0,1)$ and $\{x_t\}_{t\in[T]}$ be a set of positive numbers with $x_t \leq G$. Consider the sequence defined by $a_t = (1-\beta)a_{t-1} + \beta x_t^2$ and suppose that $a_0 \leq G^2$. Then for any $t_0 \in [0 : T-1]$ and any $\varepsilon \geq \sqrt{2\beta}G$, it holds that*

$$\sum_{t=t_0+1}^{T} \frac{x_t^2}{a_{t-1}+\varepsilon^2} \leq \frac{1}{\beta} + \frac{2}{\beta}\log\left(\frac{2}{\beta}\right) + \frac{T-t_0}{1-\beta}.$$

Proof. See Appendix D.8 for a proof of this result. □

The following basic lemma is useful in our analysis for solving inequalities $\frac{\log T}{T^q} \leq \epsilon$ with some $q > 0$ (e.g., $q = 1$ or 0.5).

Lemma 11. *Let $a, q > 0$ be some positive real values. If $x^q \geq \max\{2, 2a\log(2a)\}$, then $\frac{x^q}{\log(x)} \geq qa$.*

Proof. See Appendix D.9 for a proof of this result. □

Now we are in the position to prove Theorem 3.

Proof of Theorem 3. Let us consider $a_t \equiv \frac{\eta}{G+\varepsilon}$. Since $\eta_{t_0} \leq \frac{1}{2\rho}$ and $a_t \leq \eta_t \leq \eta_{t-1}$ (based on Lemma 9), we can apply Lemma 2 to obtain

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=t_0+1}^{T}\sum_{\tau=t+1}^{T}\left(1-(1-\beta_1)^{\tau-t}\right)^2\left(\frac{\eta^2}{(G+\varepsilon)^2}\left\|\bar{g}_{[t+1:\tau]}\right\|^2+\frac{\beta_1^2}{6(1-\beta_1)}V^2_{[t+1:\tau]}(w_\tau)\right)\right]\\
\leq&\mathbb{E}\left[\frac{6\eta_{t_0}^2\|m_{t_0}\|^2}{\beta_1^2}+\frac{12\eta_{t_0}\Delta f(w_{t_0})}{\beta_1}+6\sum_{t=t_0+1}^{T}\eta_{t-1}^2\|\hat{g}(w_t;z_t)\|^2+4\sum_{t=t_0+1}^{T}\eta_t\Delta f(w_t)\right.\\
&\left.+6\beta_1\sum_{t=t_0+1}^{T}\sum_{\tau=t+1}^{T}\eta_{\tau-1}^2\|\hat{g}(w_\tau;z_\tau)\|^2\right]\\
\overset{\zeta_1}{\leq}&\mathbb{E}\left[\frac{6\eta^2G^2}{\beta_1^2(\sqrt{\hat{v}_0}+\varepsilon)^2(1-(1-\beta_1)^{t_0})^2}+\frac{12\eta M}{\beta_1(\sqrt{\hat{v}_0}+\varepsilon)(1-(1-\beta_1)^{t_0})}+6\eta^2\sum_{t=t_0+1}^{T}\frac{\|\hat{g}(w_t;z_t)\|^2}{(\sqrt{\hat{v}_{t-1}}+\varepsilon)^2(1-(1-\beta_1)^{t-1})^2}\right.\\
&\left.+\sum_{t=t_0+1}^{T}\frac{4\eta M}{(\sqrt{\hat{v}_0}+\varepsilon)(1-(1-\beta_1)^t)}+6\beta_1\eta^2\sum_{t=t_0+1}^{T}\sum_{\tau=t+1}^{T}\frac{\|\hat{g}(w_\tau;z_\tau)\|^2}{(\sqrt{\hat{v}_{\tau-1}}+\varepsilon)^2(1-(1-\beta_1)^{\tau-1})^2}\right]\\
\overset{\zeta_2}{\leq}&\mathbb{E}\left[\frac{7\eta^2G^2}{\beta_1^2(\sqrt{\hat{v}_0}+\varepsilon)^2}+\frac{13\eta M}{\beta_1(\sqrt{\hat{v}_0}+\varepsilon)}+7\eta^2\sum_{t=t_0+1}^{T}\frac{\|\hat{g}(w_t;z_t)\|^2}{\left(\sqrt{\frac{v_{t-1}}{1-(1-\beta_2)^{t-1}}}+\varepsilon\right)^2}+\sum_{t=t_0+1}^{T}\frac{5\eta M}{\sqrt{\hat{v}_0}+\varepsilon}\right.\\
&\left.+7\beta_1\eta^2\sum_{t=t_0+1}^{T}\sum_{\tau=t+1}^{T}\frac{\|\hat{g}(w_\tau;z_\tau)\|^2}{\left(\sqrt{\frac{v_{\tau-1}}{1-(1-\beta_2)^{\tau-1}}}+\varepsilon\right)^2}\right]\\
\overset{\zeta_3}{\leq}&\mathbb{E}\left[\frac{7\eta^2G^2}{\beta_1^2(\sqrt{\hat{v}_0}+\varepsilon)^2}+\frac{5\eta M}{\sqrt{\hat{v}_0}+\varepsilon}\left(\frac{3}{\beta_1}+T-t_0\right)+7\eta^2\sum_{t=t_0+1}^{T}\frac{\|\hat{g}(w_t;z_t)\|^2}{\left(\sqrt{v_{t-1}}+\varepsilon\right)^2}+7\beta_1\eta^2\sum_{t=t_0+1}^{T}\sum_{\tau=t+1}^{T}\frac{\|\hat{g}(w_\tau;z_\tau)\|^2}{\left(\sqrt{v_{\tau-1}}+\varepsilon\right)^2}\right]\\
\leq&\mathbb{E}\left[\frac{7\eta^2G^2}{\beta_1^2(\hat{v}_0+\varepsilon^2)}+\frac{5\eta M}{\sqrt{\hat{v}_0}+\varepsilon}\left(\frac{3}{\beta_1}+T-t_0\right)+7\eta^2\sum_{t=t_0+1}^{T}\frac{\|\hat{g}(w_t;z_t)\|^2}{v_{t-1}+\varepsilon^2}+7\beta_1\eta^2\sum_{t=t_0+1}^{T}\sum_{\tau=t+1}^{T}\frac{\|\hat{g}(w_\tau;z_\tau)\|^2}{v_{\tau-1}+\varepsilon^2}\right]\\
\overset{\zeta_4}{\leq}&\frac{7\eta^2G^2}{\beta_1^2(\hat{v}_0+\varepsilon^2)}+\frac{5\eta M}{\sqrt{\hat{v}_0}+\varepsilon}\left(\frac{3}{\beta_1}+T-t_0\right)+7\eta^2\left(\frac{1}{\beta_2}+\frac{2}{\beta_2}\log\left(\frac{2}{\beta_2}\right)+\frac{T-t_0}{1-\beta_2}\right)\\
&+7\beta_1\eta^2\sum_{t=t_0+1}^{T}\left(\frac{1}{\beta_2}+\frac{2}{\beta_2}\log\left(\frac{2}{\beta_2}\right)+\frac{T-t}{1-\beta_2}\right)\\
\leq&\frac{7\eta^2G^2}{\beta_1^2(\hat{v}_0+\varepsilon^2)}+\frac{5\eta M}{\sqrt{\hat{v}_0}+\varepsilon}\left(\frac{3}{\beta_1}+T-t_0\right)+7\eta^2\left(\frac{1}{\beta_2}+\frac{2}{\beta_2}\log\left(\frac{2}{\beta_2}\right)+\frac{T-t_0}{1-\beta_2}\right)\\
&+7\beta_1\eta^2(T-t_0)\left(\frac{1}{\beta_2}+\frac{2}{\beta_2}\log\left(\frac{2}{\beta_2}\right)+\frac{T-t_0}{2(1-\beta_2)}\right)\\
\leq&\frac{7\eta^2G^2}{\beta_1^2(\hat{v}_0+\varepsilon^2)}+\frac{6\eta MT}{\sqrt{\hat{v}_0}+\varepsilon}+8\eta^2\beta_1T\left(\frac{1}{\beta_2}+\frac{2}{\beta_2}\log\left(\frac{2}{\beta_2}\right)+\frac{T-t_0}{1-\beta_2}\right),
\end{aligned}
\tag{18}
$$

where we have used in "ζ_1" the restart setting $\|m_{t_0}\| \leq G$, the boundedness of f, and Lemma 9; in "ζ_2" the condition $(1-\beta_1)^t \leq 0.05$ for $t \geq t_0$ and the fact $\hat{v}_t \geq \frac{v_t}{1-(1-\beta_2)^t}$ by definition, and in "ζ_3" the fact $1-(1-\beta_2)^t \in (0,1)$, in "$\zeta_4$" we have applied Lemma 10 with β, a_t, x_t substantialized by

$\beta_2, v_t, \|\hat{g}(w_t; z_t)\|$, respectively, and in the last step we have used the condition $\beta_1 \geq \frac{24}{T}$. In view of 18, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=t_0+1}^{T} \frac{A_{[t+1:T]}}{\sum_{t=t_0+1}^{T} A_{[t+1:T]}} \sum_{\tau=t+1}^{T} \frac{(1-(1-\beta)^{\tau-t})^2}{A_{[t+1:T]}} \left(\frac{\eta^2}{(G+\varepsilon)^2}\left\|\bar{g}_{[t+1:\tau]}\right\|^2 + \frac{\beta_1^2}{6(1-\beta_1)} V^2_{[t+1:\tau]}(w_\tau)\right)\right] \\
=&\frac{1}{\sum_{t=t_0+1}^{T} A_{[t+1:T]}} \mathbb{E}\left[\sum_{t=t_0+1}^{T} \sum_{\tau=t+1}^{T} (1-(1-\beta)^{\tau-t})^2 \left(\frac{\eta^2}{(G+\varepsilon)^2}\left\|\bar{g}_{[t+1:\tau]}\right\|^2 + \frac{\beta_1^2}{6(1-\beta_1)} V^2_{[t+1:\tau]}(w_\tau)\right)\right] \\
\leq&\frac{1}{\sum_{t=t_0+1}^{T} A_{[t+1:T]}} \left(\frac{7\eta^2 G^2}{\beta_1^2(\hat{v}_0+\varepsilon^2)} + \frac{6\eta MT}{\sqrt{\hat{v}_0}+\varepsilon} + 8\eta^2\beta_1 T\left(\frac{1}{\beta_2} + \frac{2}{\beta_2}\log\left(\frac{2}{\beta_2}\right) + \frac{T-t_0}{1-\beta_2}\right)\right) \\
\leq&\frac{12}{T^2}\left(\frac{7\eta^2 G^2}{\beta_1^2(\hat{v}_0+\varepsilon^2)} + \frac{6\eta MT}{\sqrt{\hat{v}_0}+\varepsilon} + 8\eta^2\beta_1 T\left(\frac{1}{\beta_2} + \frac{2}{\beta_2}\log\left(\frac{2}{\beta_2}\right) + \frac{T-t_0}{1-\beta_2}\right)\right),
\end{aligned}$$

where in the last but step we have used Lemma 7 and the conditions $\beta \geq \frac{24}{T}$ and $T \geq 2t_0 + 6$ such that $\sum_{t=t_0+1}^{T} A_{[t+1:T]} \geq \sum_{t=t_0+1}^{T}(T - t - \frac{2}{\beta}) = \frac{(T-t_0-1)(T-t_0)}{2} - \frac{2(T-t_0)}{\beta} \geq \frac{(T-t_0)^2}{3} \geq \frac{T^2}{12}$. Finally, based on the above inequality and the definition of random indices R, S we arrive at

$$\begin{aligned}
&\mathbb{E}\left[\left\|\bar{g}_{[R+1:S]}\right\|^2 + \frac{(G+\varepsilon)^2\beta_1^2}{6\eta^2(1-\beta_1)} V^2_{[R+1:S]}(w_S)\right] \\
\leq&\frac{12(G+\varepsilon)^2}{T^2\eta^2}\left(\frac{7\eta^2 G^2}{\beta_1^2(\hat{v}_0+\varepsilon^2)} + \frac{6\eta MT}{\sqrt{\hat{v}_0}+\varepsilon} + 8\eta^2\beta_1 T\left(\frac{1}{\beta_2} + \frac{2}{\beta_2}\log\left(\frac{2}{\beta_2}\right) + \frac{T-t_0}{1-\beta_2}\right)\right) \\
=&\frac{84(G+\varepsilon)^2 G^2}{T^2\beta_1^2(\hat{v}_0+\varepsilon^2)} + \frac{72(G+\varepsilon)^2 M}{T(\sqrt{\hat{v}_0}+\varepsilon)\eta} + \frac{96(G+\varepsilon)^2\beta_1}{T}\left(\frac{1}{\beta_2} + \frac{2}{\beta_2}\log\left(\frac{2}{\beta_2}\right) + \frac{T-t_0}{1-\beta_2}\right).
\end{aligned}$$

Since $\mu \leq \mu' = \frac{(G+\varepsilon)^2\beta_1^2}{6\eta^2(1-\beta_1)}$, in view of Proposition 2 we get that the following holds for any index pair (t, τ) satisfying $t + 1 \leq \tau$:

$$\|\partial_{+\mu} f(w_\tau)\|^2 \leq \|\partial_{+\mu'} f(w_\tau)\|^2 \leq \left\|\bar{g}_{[t+1:\tau]}\right\|^2 + \mu' V^2_{[t+1:\tau]}(w_\tau).$$

Combining the preceding two inequalities leads to the following desired bound:

$$\begin{aligned}
&\mathbb{E}\left[\|\partial_{+\mu} f(w_S)\|^2\right] \\
\leq&\frac{84(G+\varepsilon)^2 G^2}{T^2\beta_1^2(\hat{v}_0+\varepsilon^2)} + \frac{72(G+\varepsilon)^2 M}{T(\sqrt{\hat{v}_0}+\varepsilon)\eta} + \frac{96(G+\varepsilon)^2\beta_1}{T}\left(\frac{1}{\beta_2} + \frac{2}{\beta_2}\log\left(\frac{2}{\beta_2}\right) + \frac{T-t_0}{1-\beta_2}\right).
\end{aligned}$$

The proof is completed. □

C.6 Proof of Corollary 3

Proof. Consider $\eta = \frac{\sqrt{\hat{v}_0}+\varepsilon}{\mu^{1/4}T^{1/2}}$ and $t_0 = \lfloor \frac{T}{3} \rfloor$ in Theorem 3. Under the condition of T, we must have $T > \frac{96}{\mu^{1/2}} + 24\mu^{1/2} > 48$ and

$$\beta_1 = \frac{\sqrt{6}\mu^{1/4}}{T^{1/2}} \in \left[\frac{24}{T}, \frac{1}{2}\right], \ (1-\beta_1)^{t_0} \leq \left(1 - \frac{24}{T}\right)^{T/3} \leq e^{-8} < 0.05.$$

Consequently, based on Lemma 9 and the condition on T it can be verified that

$$\eta_{t_0} = \frac{\eta}{(\sqrt{\hat{v}_0}+\varepsilon)(1-(1-\beta_1)^{t_0})} \leq \frac{1.05\eta}{\sqrt{\hat{v}_0}+\varepsilon} = \frac{1.05}{\mu^{1/4}T^{1/2}} \leq \frac{1}{2\rho}.$$

Again with the given choice of β_1, η, it can be readily shown that

$$\frac{(G+\varepsilon)^2\beta_1^2}{6\eta^2(1-\beta_1)} \geq \frac{(G+\varepsilon)^2\beta_1^2}{6\eta^2} = \frac{(G+\varepsilon)^2\mu}{(\sqrt{\hat{v}_0}+\varepsilon)^2} \geq \mu.$$

Moreover, since $T \geq \max\left\{8, \frac{8G^2}{\varepsilon^2}\log\left(\frac{G}{\varepsilon}\right)\right\}$, according to Lemma 11 with $a = \frac{G^2}{2\varepsilon^2}$ and $q = 1$ we must have $G\sqrt{2\beta_2} = G\sqrt{\frac{2\log(T)}{T}} \leq \varepsilon$.

In view of the above arguments, the conditions of Theorem 3 are fulfilled in our setting, and we can thus apply it to obtain

$$\begin{aligned}
&\mathbb{E}\left[\|\partial_{+\mu} f(w_S)\|^2\right] \\
\leq &\frac{84(G+\varepsilon)^2 G^2}{T^2\beta_1^2(\hat{v}_0+\varepsilon^2)} + \frac{72(G+\varepsilon)^2 M}{T(\sqrt{\hat{v}_0}+\varepsilon)\eta} + \frac{96(G+\varepsilon)^2\beta_1}{T}\left(\frac{1}{\beta_2} + \frac{2}{\beta_2}\log\left(\frac{2}{\beta_2}\right) + \frac{T-t_0}{1-\beta_2}\right) \\
\overset{\zeta_1}{\leq} &\frac{14(G+\varepsilon)^2 G^2}{T\mu^{1/2}(\hat{v}_0+\varepsilon^2)} + \frac{72(G+\varepsilon)^2 M\mu^{1/4}}{(\sqrt{\hat{v}_0}+\varepsilon)^2 T^{1/2}} + \frac{288\sqrt{6}(G+\varepsilon)^2\mu^{1/4}}{T^{1/2}} \\
\leq &\frac{14(G+\varepsilon)^2 G^2}{T\mu^{1/2}(\hat{v}_0+\varepsilon^2)} + \frac{(G+\varepsilon)^2\mu^{1/4}}{T^{1/2}}\left(\frac{72M}{\hat{v}_0+\varepsilon^2} + 706\right) \leq \epsilon^2,
\end{aligned}$$

where in "ζ_1" we have used the choice of $\beta_2 = \frac{\log(T)}{T}$ and $t_0 = \lfloor \frac{T}{3} \rfloor$ which imply $\frac{1}{\beta_2} + \frac{2}{\beta_2}\log\left(\frac{2}{\beta_2}\right) = \frac{T}{\log(T)} + \frac{2T}{\log(T)}\log\left(\frac{2T}{\log(T)}\right) \leq 2.26T$ and $\frac{T-t_0}{1-\beta_2} \leq \frac{2T/3}{1-\log(48)/(48)} < 0.73T$, and in the last step we have used the condition on T such that the two terms on the LHS are respectively no larger than $\frac{\epsilon^2}{2}$. The desired bound then follows from the fact $\mathbb{E}\left[\|\partial_{+\mu} f(w_S)\|\right] \leq \sqrt{\mathbb{E}\left[\|\partial_{+\mu} f(w_S)\|^2\right]} \leq \epsilon$. This proves the desired bound. □

C.7 Proof of Theorem 4

Our proof follows the reduction argument from smooth stationarity to Goldstein-style stationarity [16; 57], which relies on the following lower-bound construction lemma for smooth stochastic optimization problems.

Lemma 12 (Cutkosky et al. [16]). *For any constants L, γ, σ, ϵ, there exists an objective function f and stochastic gradient estimator $\hat{g}$ satisfying*

(1) f is L-smooth and $\sqrt{L\gamma}$-Lipschitz, $f(w_0) - \inf f \leq \gamma$, and $\mathbb{E}_{Z\sim\mathcal{D}}\|\nabla f(w) - \hat{g}(w;Z)\|^2 \leq \sigma^2$;

(2) any randomized algorithm relying on $\hat{g}$ requires $\Omega(\gamma\sigma^2 L\epsilon^{-4})$ iterations to find an ϵ-stationary point $\hat{w}$ of f such that $\|\nabla f(\hat{w})\| \leq \epsilon$.

Proof of Theorem 4. Given arbitrary γ, ρ, G, μ, ϵ satisfying $G \geq \gamma^{1/2}\mu^{1/4}$ and $\rho \geq \mu^{1/2}$, we set $L = \mu^{1/2}$ and $\sigma = G$. It is valid that $G \geq \sqrt{L\gamma}$ and $\rho \geq L$. By the lower bound in Lemma 12, there exits a function f that is L-smooth, G-Lipschitz, satisfies $f(w_0) - \inf f \leq \gamma$, and $\mathbb{E}\|\nabla f(w) - g(w;z)\|^2 \leq G^2$. Moreover, finding an ϵ-stationary point of f requires $\Omega(\gamma G^2\mu^{1/2}\epsilon^{-4})$ iterations.

Clearly, the above hard function f must be ρ-weakly convex as it is L-smooth. Recall that $\mu = L^2$ by construction. In view of Proposition 3 in Appendix A.3, any (μ,ϵ)-L_2-regularized stationary point of f is also an 2ϵ-stationary point of f. Consequently, the iteration complexity lower bound $\Omega(\gamma G^2\mu^{1/2}\epsilon^{-4})$ also applies to the task of finding (μ,ϵ)-L_2-regularized stationary points of f. □

D PROOFS OF TECHNICAL LEMMAS

D.1 PROOF OF LEMMA 3

Proof. For any $t \geq t_0 + 1$, it can be shown that

$$
\begin{aligned}
&\|m_t\|^2 = \|(1-\beta)m_{t-1} + \beta\hat{g}(w_t; z_t)\|^2 \\
=&(1-\beta)^2\|m_{t-1}\|^2 + 2\beta(1-\beta)\langle m_{t-1}, \hat{g}(w_t; z_t)\rangle + \beta^2\|\hat{g}(w_t; z_t)\|^2 \\
=&(1-\beta)^2\|m_{t-1}\|^2 + 2\beta(1-\beta)\langle m_{t-1}, g(w_t)\rangle + \beta^2\|\hat{g}(w_t; z_t)\|^2 + 2\beta(1-\beta)\langle m_{t-1}, u_t\rangle \\
=&(1-\beta)^2\|m_{t-1}\|^2 + \frac{2\beta(1-\beta)}{\eta_{t-1}}\langle w_{t-1} - w_t, g(w_t)\rangle + \beta^2\|\hat{g}(w_t; z_t)\|^2 + 2\beta(1-\beta)\langle m_{t-1}, u_t\rangle \\
\overset{\zeta_1}{\leq}&(1-\beta)^2\|m_{t-1}\|^2 + \frac{2\beta(1-\beta)}{\eta_{t-1}}\left(f(w_{t-1}) - f(w_t) + \frac{\rho}{2}\|w_t - w_{t-1}\|^2\right) + \beta^2\|\hat{g}(w_t; z_t)\|^2 \\
&+ 2\beta(1-\beta)\langle m_{t-1}, u_t\rangle \\
=&(1-\beta)^2\|m_{t-1}\|^2 + \beta(1-\beta)\rho\eta_{t-1}\|m_{t-1}\|^2 + \frac{2\beta(1-\beta)}{\eta_{t-1}}\left(f(w_{t-1}) - f(w_t)\right) + \beta^2\|\hat{g}(w_t; z_t)\|^2 \\
&+ 2\beta(1-\beta)\langle m_{t-1}, u_t\rangle \\
\overset{\zeta_2}{\leq}&(1-\beta)^2\|m_{t-1}\|^2 + \frac{\beta(1-\beta)}{2}\|m_{t-1}\|^2 + \frac{2\beta(1-\beta)}{\eta_{t-1}}\left(f(w_{t-1}) - f(w_t)\right) + \beta^2\|\hat{g}(w_t; z_t)\|^2 \\
&+ 2\beta(1-\beta)\langle m_{t-1}, u_t\rangle \\
\overset{\zeta_3}{\leq}&\left(1 - \frac{5\beta}{4}\right)\|m_{t-1}\|^2 + \frac{2\beta(1-\beta)}{\eta_{t-1}}\left(f(w_{t-1}) - f(w_t)\right) + \beta^2\|\hat{g}(w_t; z_t)\|^2 + 2\beta(1-\beta)\langle m_{t-1}, u_t\rangle,
\end{aligned}
$$

where in "ζ_1" we have used the ρ-weak-convexity of f, in "ζ_2" the condition $\eta_{t-1} \leq \eta_{t_0} \leq \frac{1}{2\rho}$, and in "$\zeta_3$" the condition $\beta \leq 0.5$. Consequently, since $\eta_t \leq \eta_{t-1}$, we have

$$
\begin{aligned}
&\eta_t^2\|m_t\|^2 \leq \eta_{t-1}^2\|m_t\|^2 \\
\leq&\left(1 - \frac{5\beta}{4}\right)\eta_{t-1}^2\|m_{t-1}\|^2 + 2\beta(1-\beta)\eta_{t-1}\left(f(w_{t-1}) - f(w_t)\right) + \beta^2\eta_{t-1}^2\|\hat{g}(w_t; z_t)\|^2 \\
&+ 2\beta(1-\beta)\eta_{t-1}^2\langle m_{t-1}, u_t\rangle \\
=&(1-\beta)\eta_{t-1}^2\|m_{t-1}\|^2 - \frac{\beta}{4}\|w_t - w_{t-1}\|^2 + 2\beta(1-\beta)\eta_{t-1}\left(f(w_{t-1}) - f(w_t)\right) \\
&+ \beta^2\eta_{t-1}^2\|\hat{g}(w_t; z_t)\|^2 + 2\beta(1-\beta)\eta_{t-1}^2\langle m_{t-1}, u_t\rangle.
\end{aligned}
$$

Unfolding the above inequality over $\tau = t, t-1, ..., t_0 + 1$ and rearranging the terms yield

$$
\begin{aligned}
&\eta_t^2\|m_t\|^2 + \sum_{\tau=t_0+1}^{t}\frac{(1-\beta)^{t-\tau}\beta}{4}\|w_\tau - w_{\tau-1}\|^2 \\
\leq&(1-\beta)^{t-t_0}\eta_{t_0}^2\|m_{t_0}\|^2 + 2\beta(1-\beta)\sum_{\tau=t_0+1}^{t}(1-\beta)^{t-\tau}\eta_{\tau-1}\left(f(w_{\tau-1}) - f(w_\tau)\right) \\
&+ \beta^2\sum_{\tau=t_0+1}^{t}(1-\beta)^{t-\tau}\eta_{\tau-1}^2\|\hat{g}(w_\tau; z_\tau)\|^2 + 2\beta(1-\beta)\sum_{\tau=t_0+1}^{t}(1-\beta)^{t-\tau}\eta_{\tau-1}^2\langle m_{\tau-1}, u_\tau\rangle.
\end{aligned}
$$

The proof is completed. □

D.2 PROOF OF LEMMA 4

Proof. Based on the definition, m_t can be expressed as

$$
m_t = (1-\beta)^{t-t_0}m_{t_0} + \sum_{\tau=t_0+1}^{t}(1-\beta)^{t-\tau}\beta\hat{g}(w_\tau; z_\tau).
$$

Then we have

$$
\begin{aligned}
&\|m_t\|^2 \\
=&\left\|(1-\beta)^{t-t_0} m_{t_0} + \sum_{\tau=t_0+1}^{t} (1-\beta)^{t-\tau}\beta\hat{g}(w_\tau; z_\tau)\right\|^2 \\
=&\left\|(1-\beta)^{t-t_0} m_{t_0} + \sum_{\tau=t_0+1}^{t} (1-\beta)^{t-\tau}\beta\left(g(w_\tau) + u_\tau\right)\right\|^2 \\
\overset{\zeta_1}{\geq}&\frac{1}{2}\left\|\sum_{\tau=t_0+1}^{t} (1-\beta)^{t-\tau}\beta g(w_\tau)\right\|^2 - \left\|(1-\beta)^{t-t_0} m_{t_0} + \sum_{\tau=t_0+1}^{t} (1-\beta)^{t-\tau}\beta u_\tau\right\|^2 \\
\geq&\frac{1}{2}\left\|\sum_{\tau=t_0+1}^{t} (1-\beta)^{t-\tau}\beta g(w_\tau)\right\|^2 - 2\left\|(1-\beta)^{t-t_0} m_{t_0}\right\|^2 - 2\left\|\sum_{\tau=t_0+1}^{t} (1-\beta)^{t-\tau}\beta u_\tau\right\|^2 \\
=&\frac{(1-(1-\beta)^{t-t_0})^2}{2}\left\|\bar{g}_{[t_0+1:t]}\right\|^2 - 2(1-\beta)^{2(t-t_0)}\|m_{t_0}\|^2 - 2\beta^2\left\|\sum_{\tau=t_0+1}^{t} (1-\beta)^{t-\tau} u_\tau\right\|^2,
\end{aligned}
\tag{19}
$$

where in "ζ_1" we have used the fact $\|x+y\|^2 \geq \frac{\|x\|^2}{2} - \|y\|^2$. The proof is completed. □

D.3 Proof of Lemma 5

We first establish The following lemma that provides an upper bound for the quantity $V^2_{[t_0+1:t]}(w_t)$, as defined in 9, in terms of the adjacent iterate distances $\{\|w_\tau - w_{\tau-1}\|^2\}_{\tau=t_0+1}^{t}$.

Lemma 13. *Suppose that $\beta \leq 0.5$. For any $t \geq 1$ and $t_0 \in [0 : t-1]$, it holds that*

$$
V^2_{[t_0+1:t]}(w_t) \leq \sum_{\tau=t_0+1}^{t} \frac{2(t-\tau+1)(1-\beta)^{t-\tau+1} + \beta^{-1}(1-\beta)^{t-\tau+1}}{(1-(1-\beta)^{t-t_0}}\|w_\tau - w_{\tau-1}\|^2.
$$

Proof. See Appendix D.10 for a proof of this result. □

Remark 15. *It is noteworthy that the bound in Lemma 13 also holds for $V^2_{[t_0+1:t]}(\bar{w}_{[t_0+1:t]})$ with respect to the model EMA $\bar{w}_{[t_0+1:t]} := \sum_{\tau=t_0+1}^{t} q_{\tau,[t_0+1:t]} w_\tau$. Indeed, given this key bound, all our main results readily extend to the model EMA case. For the sake of simplicity, we do not explore this direction further and instead focus on analyzing the convergence of the raw model iterates.*

Now we are ready to prove Lemma 5.

Proof of Lemma 5. Under the given conditions, we can apply Lemma 13 to obtain

$$
\begin{aligned}
&\sum_{t=t_0+1}^{T} \left(1-(1-\beta)^{t-t_0}\right) V^2_{[t_0+1:t]}(w_t) \\
\leq&\sum_{t=t_0+1}^{T} \left(1-(1-\beta)^{t-t_0}\right) \sum_{\tau=t_0+1}^{t} \frac{2(t-\tau+1)(1-\beta)^{t-\tau+1} + \beta^{-1}(1-\beta)^{t-\tau+1}}{(1-(1-\beta)^{t-t_0}}\|w_\tau - w_{\tau-1}\|^2 \\
=&2\underbrace{\sum_{t=t_0+1}^{T}\sum_{\tau=t_0+1}^{t} (t-\tau+1)(1-\beta)^{t-\tau+1}\|w_\tau - w_{\tau-1}\|^2}_{I_1} + \frac{1-\beta}{\beta}\sum_{t=t_0+1}^{T}\sum_{\tau=t_0+1}^{t}(1-\beta)^{t-\tau}\|w_\tau - w_{\tau-1}\|^2.
\end{aligned}
\tag{20}
$$

We next proceed to upper bound the term I_1.

$$
\begin{aligned}
I_1 &= \sum_{t=t_0+1}^{T} \sum_{\tau=t_0+1}^{t} (t-\tau+1)(1-\beta)^{t-\tau+1} \|w_\tau - w_{\tau-1}\|^2 \\
&= \sum_{\tau=t_0+1}^{T} \|w_\tau - w_{\tau-1}\|^2 \sum_{t=\tau}^{T} (t-\tau+1)(1-\beta)^{t-\tau+1} \\
&= \sum_{\tau=t_0+1}^{T} \|w_\tau - w_{\tau-1}\|^2 \sum_{i=1}^{T-\tau+1} i(1-\beta)^{i} \\
&= \sum_{\tau=t_0+1}^{T} \|w_\tau - w_{\tau-1}\|^2 (1-\beta) \frac{\mathrm{d}}{\mathrm{d}\beta} \left(- \sum_{i=1}^{T-\tau+1} (1-\beta)^{i} \right) \\
&= \sum_{\tau=t_0+1}^{T} \|w_\tau - w_{\tau-1}\|^2 (1-\beta) \frac{\mathrm{d}}{\mathrm{d}\beta} \left(\frac{\beta - 1 + (1-\beta)^{T-\tau+2}}{\beta} \right) \\
&= \sum_{\tau=t_0+1}^{T} \|w_\tau - w_{\tau-1}\|^2 \left(\frac{(1-\beta)^2 (1-(1-\beta)^{T-\tau+1})}{\beta^2} + \frac{(1-\beta)(1-(T-\tau+2)(1-\beta)^{T-\tau+1})}{\beta} \right) \\
&\overset{\zeta_1}{\leq} \sum_{\tau=t_0+1}^{T} \|w_\tau - w_{\tau-1}\|^2 \left(\frac{(1-\beta)^2 (1-(1-\beta)^{T-\tau+1})}{\beta^2} + \frac{(1-\beta)(1-(1-\beta)^{T-\tau+1})}{\beta} \right) \\
&= \sum_{\tau=t_0+1}^{T} \|w_\tau - w_{\tau-1}\|^2 \frac{(1-\beta)(1-(1-\beta)^{T-\tau+1})}{\beta^2} \\
&= \frac{(1-\beta)}{\beta} \sum_{\tau=t_0+1}^{T} \|w_\tau - w_{\tau-1}\|^2 \sum_{t=\tau}^{T} (1-\beta)^{t-\tau} \\
&= \frac{1-\beta}{\beta} \sum_{t=t_0+1}^{T} \sum_{\tau=t_0+1}^{t} (1-\beta)^{t-\tau} \|w_\tau - w_{\tau-1}\|^2,
\end{aligned}
$$

where in "ζ_1" we have used $T-\tau+2 \geq 1$ for all $\tau \in [t_0+1:T]$. By substituting the above bound of I_1 into 20 we obtain that

$$
\sum_{t=t_0+1}^{T} \left(1-(1-\beta)^{t-t_0}\right) V^2_{[t_0+1:t]}(w_t) \leq \frac{3(1-\beta)}{\beta} \sum_{t=t_0+1}^{T} \sum_{\tau=t_0+1}^{t} (1-\beta)^{t-\tau} \|w_\tau - w_{\tau-1}\|^2.
$$

The proof is completed. □

D.4 Proofs of Lemma 6

Proof. For any $t \geq 1$, since m_{t-1} and η_{t-1} are both $\mathcal{F}_{t-1}$-measurable and Assumption 2 is valid, we must have

$$
\mathbb{E}\left[\eta^2_{t-1} \langle m_{t-1}, u_t \rangle \mid \mathcal{F}_{t-1}\right] = \eta^2_{t-1} \langle m_{t-1}, \mathbb{E}\left[\hat{g}(w_t; z_t) - g(w_t) \mid \mathcal{F}_{t-1}\right] \rangle = 0.
$$

Moreover, the sequence $\{u_\tau\}_{\tau \geq 1}$ is a vector-valued martingale difference sequence with respect to the filtration $\{\mathcal{F}_\tau\}_{\tau \geq 1}$, and thus we can show that

$$
\begin{aligned}
&\mathbb{E}\left[\left\| \sum_{\tau=t_0+1}^{t} (1-\beta)^{t-\tau} u_\tau \right\|^2 \right] = \mathbb{E}\left[\sum_{\tau=t_0+1}^{t} (1-\beta)^{2(t-\tau)} \|u_\tau\|^2 \right] \\
=&\mathbb{E}\left[\sum_{\tau=t_0+1}^{t} (1-\beta)^{2(t-\tau)} \left\|\hat{g}(w_\tau; z_\tau) - g(w_\tau)\right\|^2 \right] \leq \mathbb{E}\left[\sum_{\tau=t_0+1}^{t} (1-\beta)^{2(t-\tau)} \left\|\hat{g}(w_\tau; z_\tau)\right\|^2 \right] \\
\leq&\mathbb{E}\left[\sum_{\tau=t_0+1}^{t} (1-\beta)^{t-\tau} \left\|\hat{g}(w_\tau; z_\tau)\right\|^2 \right],
\end{aligned}
$$

where in the last step we have used the fact $(1-\beta)^2 \leq (1-\beta)$. The proof is completed. □

D.5 Proof of Lemma 7

Proof. By definition we have $A_{[t+1:T]} = \sum_{\tau=t+1}^{T} \left(1-(1-\beta)^{\tau-t}\right)^2 \leq T-t$. Moreover,

$$A_{[t+1:T]} = \sum_{\tau=t+1}^{T} (1-(1-\beta)^{\tau-t})^2$$
$$\geq \sum_{\tau=t+1}^{T} (1-2(1-\beta)^{\tau-t}) = T-t-\frac{2(1-\beta)(1-(1-\beta)^{T-t})}{\beta} \geq T-t-\frac{2}{\beta}.$$

The proof is completed. □

D.6 Proof of Lemma 8

Proof. It can be shown that

$$\sum_{t=t_1+1}^{t_2}\sum_{\tau=t_1+1}^{t}(1-\beta)^{t-\tau}\eta_{\tau-1}(f(w_{\tau-1})-f(w_\tau))$$
$$=\sum_{t=1}^{t_2-t_1}(1-\beta)^{t-1}\sum_{\tau=t_1+1}^{t_2-t+1}\eta_{\tau-1}(f(w_{\tau-1})-f(w_\tau))$$
$$=\sum_{t=1}^{t_2-t_1}(1-\beta)^{t-1}\left(\eta_{t_1}f(w_{t_1})-\eta_{t_2-t}f(w_{t_2-t+1})+\sum_{\tau=t_1+1}^{t_2-t}f(w_\tau)\left(\eta_\tau-\eta_{\tau-1}\right)\right)$$
$$\overset{\zeta_1}{\leq}\sum_{t=1}^{t_2-t_1}(1-\beta)^{t-1}\left(\eta_{t_1}f(w_{t_1})-\eta_{t_2-t}f^*+\sum_{\tau=t_1+1}^{t_2-t}f^*\left(\eta_\tau-\eta_{\tau-1}\right)\right)$$
$$=\sum_{t=1}^{t_2-t_1}(1-\beta)^{t-1}\eta_{t_1}(f(w_{t_1})-f^*)=\frac{\left(1-(1-\beta)^{t_2-t_1}\right)\eta_{t_1}\Delta f(w_{t_1})}{\beta},$$

where in "ζ_1" we have used $f(\cdot)\geq f^*$ and $\eta_\tau \leq \eta_{\tau-1}$. This proves the first desired bound.

In the meanwhile, we can show that

$$\sum_{t=t_1+1}^{t_2}\sum_{\tau=t_1+1}^{t}(1-\beta)^{t-\tau}\eta_{\tau-1}^2\|\hat{g}(w_\tau;z_\tau)\|^2=\sum_{\tau=t_1+1}^{t_2}\eta_{\tau-1}^2\|\hat{g}(w_\tau;z_\tau)\|^2\sum_{t=\tau}^{t_2}(1-\beta)^{t-\tau}$$
$$=\sum_{\tau=t_1+1}^{t_2}\frac{(1-(1-\beta)^{t_2-\tau+1})\eta_{\tau-1}^2\|\hat{g}(w_\tau;z_\tau)\|^2}{\beta}\leq\frac{1-(1-\beta)^{t_2-t_1}}{\beta}\sum_{t=t_1+1}^{t_2}\eta_{t-1}^2\|\hat{g}(w_t;z_t)\|^2,$$

where in the last step we have used $\tau \geq t_1+1$. This proves the second desired bound. □

D.7 Proof of Lemma 9

Proof. By definition we have $\hat{v}_t \geq \hat{v}_{t-1}$. Also, it clear that $1-(1-\beta_1)^t \geq 1-(1-\beta_1)^{t-1}$. Therefore, we must have

$$\eta_t=\frac{\eta}{(\sqrt{\hat{v}_t}+\varepsilon)(1-(1-\beta_1)^t)}\leq\frac{\eta}{(\sqrt{\hat{v}_{t-1}}+\varepsilon)(1-(1-\beta_1)^{t-1})}=\eta_{t-1}.$$

This proves the monotonicity of learning rates.

Further, suppose that the loss function is G-Lipschitz continuous, we can prove by induction that $\hat{v}_t \leq G^2$. Indeed, it is obviously true for $\hat{v}_0 \leq G^2$. Assume that $\hat{v}_{t-1} \leq G^2$ for some $t \geq 1$. Then based on the definition of $\hat{v}_t$ and note that $\frac{v_t}{1-(1-\beta_2)^t} \leq G^2$, we must have $\hat{v}_t \leq G^2$. Therefore, it is valid that

$$\eta_t=\frac{\eta}{(\sqrt{\hat{v}_t}+\varepsilon)(1-(1-\beta_1)^t)}\geq\frac{\eta}{\sqrt{\hat{v}_t}+\varepsilon}\geq\frac{\eta}{G+\varepsilon}.$$

On the other side, by definition we have $\hat{v}_t \geq \hat{v}_0$. It follows that

$$\eta_t = \frac{\eta}{(\sqrt{\hat{v}_t}+\varepsilon)(1-(1-\beta_1)^t)} \leq \frac{\eta}{(\sqrt{\hat{v}_0}+\varepsilon)(1-(1-\beta_1)^t)}.$$

The proof is completed. □

D.8 Proof of Lemma 10

Proof. Consider any fixed $t \geq t_0+1$. First, we note that

$$a_t = (1-\beta)^t a_0 + \sum_{\tau=1}^{t}(1-\beta)^{t-\tau}\beta x_\tau^2 \leq (1-\beta)^t G^2 + \sum_{\tau=1}^{t}(1-\beta)^{t-\tau}\beta G^2 = G^2.$$

Let us define $b_t = (1-\beta)^{T-t}\beta(x_t^2+0.5\varepsilon^2)$. Then it can be shown that

$$\begin{aligned}
&\frac{b_t}{\sum_{\tau=t_0+1}^{t} b_\tau}\\
=&\frac{(1-\beta)^{T-t}\beta(x_t^2+0.5\varepsilon^2)}{\sum_{\tau=t_0+1}^{t}(1-\beta)^{T-\tau}\beta(x_\tau^2+0.5\varepsilon^2)}\\
=&\frac{\beta(x_t^2+0.5\varepsilon^2)}{\sum_{\tau=t_0+1}^{t}(1-\beta)^{t-\tau}\beta(x_\tau^2+0.5\varepsilon^2)}\\
=&\frac{\beta(x_t^2+0.5\varepsilon^2)}{\sum_{\tau=t_0+1}^{t-1}(1-\beta)^{t-\tau}\beta x_\tau^2+\beta x_t^2+0.5(1-(1-\beta)^{t-t_0})\varepsilon^2}\\
=&\frac{\beta(x_t^2+0.5\varepsilon^2)}{(1-\beta)a_{t-1}+\beta x_t^2+0.5(1-(1-\beta)^{t-t_0})\varepsilon^2}\\
\geq&\frac{\beta x_t^2}{a_{t-1}+\beta G^2+0.5\varepsilon^2} \geq \frac{\beta x_t^2}{a_{t-1}+\varepsilon^2},
\end{aligned}$$

where in the last step we have used $2\beta G^2 \leq \varepsilon^2$. Therefore, we have

$$\begin{aligned}
&\sum_{t=t_0+1}^{T}\frac{\beta x_t^2}{a_{t-1}+\varepsilon^2}\\
\leq&\sum_{t=t_0+1}^{T}\frac{b_t}{\sum_{\tau=t_0+1}^{t} b_\tau} = 1+\sum_{t=t_0+2}^{T}\frac{b_t}{\sum_{\tau=t_0+1}^{t} b_\tau} \leq 1+\int_{b_{t_0+1}}^{\sum_{\tau=t_0+1}^{T} b_\tau}\frac{1}{x}\mathrm{d}x = 1+\log\left(\frac{\sum_{\tau=t_0+1}^{T} b_\tau}{b_{t_0+1}}\right)\\
=&1+\log\left(\frac{a_T+0.5(1-(1-\beta)^{T-t_0})\varepsilon^2}{(1-\beta)^{T-t_0-1}\beta(x_{t_0+1}^2+0.5\varepsilon^2)}\right) \leq 1+\log\left(\frac{G^2+0.5\varepsilon^2}{0.5(1-\beta)^{T-t_0-1}\beta\varepsilon^2}\right)\\
=&1+\log\left(\frac{2G^2+\varepsilon^2}{(1-\beta)^{T-t_0-1}\beta\varepsilon^2}\right)\\
\overset{\zeta_1}{\leq}&1+\log\left(\frac{\varepsilon^2\beta^{-1}+\varepsilon^2}{(1-\beta)^{T-t_0-1}\beta\varepsilon^2}\right)\\
=&1+\log\left(\frac{1+\beta}{(1-\beta)^{T-t_0-1}\beta^2}\right) = 1+\log\left(\frac{1+\beta}{\beta^2}\right)+(T-t_0-1)\log\left(\frac{1}{1-\beta}\right)\\
\leq&1+\log\left(\frac{2}{\beta^2}\right)+(T-t_0-1)\log\left(1+\frac{\beta}{1-\beta}\right) \leq 1+2\log\left(\frac{2}{\beta}\right)+\frac{(T-t_0)\beta}{1-\beta},
\end{aligned}$$

where in "ζ_1" we have used $2\beta G^2 \leq \varepsilon^2$ and in the last step we have used $\beta \leq 1$ and $\log(1+x) \leq x$, $\forall x \geq 0$. This implies the desired bound. The proof is completed. □

D.9 Proof of Lemma 11

Proof. Let us first consider the case $a \in (0,1]$. Since $y \geq \log(y)$ for all $y > 0$, we must have $\frac{x^q}{\log(x^q)} \geq 1 \geq a$ given that $x^q \geq 2$. Next, we consider the complementary case $a > 1$. Then it can

be verified that

$$\frac{x^q}{q\log x}=\frac{x^q}{\log(x^q)}=\frac{x^q}{\log(2a)+\log(x^q/(2a))}\geq\frac{x^q}{\log(2a)+x^q/(2a)}\geq a,$$

where in the last step we have used the condition $x^q\geq 2a\log(2a)$. The desired bound is directly implied from the above inequality. □

D.10 Proof of Lemma 13

Proof. Fix an index $\tau\in[t_0+1:t]$, we must have

$$\begin{aligned}\|w_\tau-w_t\|^2=&\left\|\sum_{i=\tau+1}^{t}(w_{i-1}-w_i)\right\|^2\\ \leq&(t-\tau)\sum_{i=\tau+1}^{t}\|w_i-w_{i-1}\|^2\leq(t-\tau+1)\sum_{i=\tau}^{t}\|w_i-w_{i-1}\|^2.\end{aligned}$$

Consequently, we have

$$\begin{aligned}&V^2_{[t_0+1:t]}(w_t)\\ =&\sum_{\tau=t_0+1}^{t}q_{\tau,[t_0+1:t]}\|w_\tau-w_t\|^2\\ =&\sum_{\tau=t_0+1}^{t}\frac{(1-\beta)^{t-\tau}\beta}{1-(1-\beta)^{t-t_0}}\|w_\tau-w_t\|^2\\ \leq&\sum_{\tau=t_0+1}^{t}\frac{(1-\beta)^{t-\tau}\beta(t-\tau+1)}{1-(1-\beta)^{t-t_0}}\sum_{i=\tau}^{t}\|w_i-w_{i-1}\|^2\\ =&\sum_{i=t_0+1}^{t}\frac{\beta\|w_i-w_{i-1}\|^2}{1-(1-\beta)^{t-t_0}}\sum_{\tau=t_0+1}^{i}(1-\beta)^{t-\tau}(t-\tau+1)\\ =&\sum_{i=t_0+1}^{t}\frac{\beta\|w_i-w_{i-1}\|^2}{1-(1-\beta)^{t-t_0}}\frac{\mathrm{d}}{\mathrm{d}\beta}\left(-\sum_{\tau=t_0+1}^{i}(1-\beta)^{t-\tau+1}\right)\\ =&\sum_{i=t_0+1}^{t}\frac{\beta\|w_i-w_{i-1}\|^2}{1-(1-\beta)^{t-t_0}}\frac{\mathrm{d}}{\mathrm{d}\beta}\left(-\frac{(1-\beta)^{t-i+1}\left(1-(1-\beta)^{i-t_0}\right)}{\beta}\right)\\ =&\sum_{i=t_0+1}^{t}\frac{\|w_i-w_{i-1}\|^2}{1-(1-\beta)^{t-t_0}}\times\\ &\left(\frac{\left((t-i+1)(1-\beta)^{t-i}-(t-t_0+1)(1-\beta)^{t-t_0}\right)\beta+(1-\beta)^{t-i+1}\left(1-(1-\beta)^{i-t_0}\right)}{\beta}\right)\\ \overset{\zeta_1}{\leq}&\sum_{i=t_0+1}^{t}\frac{\|w_i-w_{i-1}\|^2}{1-(1-\beta)^{t-t_0}}\left(\frac{(t-i+1)(1-\beta)^{t-i}\beta+(1-\beta)^{t-i+1}}{\beta}\right)\\ =&\sum_{i=t_0+1}^{t}\frac{\|w_i-w_{i-1}\|^2}{1-(1-\beta)^{t-t_0}}\left((t-i+1)(1-\beta)^{t-i}+\frac{(1-\beta)^{t-i+1}}{\beta}\right)\\ \leq&\sum_{i=t_0+1}^{t}\frac{\|w_i-w_{i-1}\|^2}{1-(1-\beta)^{t-t_0}}\left(2(t-i+1)(1-\beta)^{t-i+1}+\frac{(1-\beta)^{t-i+1}}{\beta}\right),\end{aligned}$$

where in "ζ_1" we simply drop non-positive terms, and in the last inequality we have used $\beta\leq 0.5$ such that $1\leq 2(1-\beta)$. This proves the second desired bound. □

Algorithm 2: Momentum-Reset SGDM (MR-SGDM)

Input : $T \in \mathbb{N}$, $\beta \in (0,1)$, $t_0 \in [0 : T-1]$, initial vectors $m_0 = w_1 = 0$.
for $t = 1, ..., T$ **do**
  Randomly sample $z_t \sim \mathcal{D}$ and obtain $\hat{g}(w_t; z_t) \in \partial \ell(w_t; z_t)$;
  Update the momentum term m_t according to 2;
  `/* Momentum reset */`
  if $t \equiv t_0$ **then**
    Reset $m_t = 0$;
  end
  Update the model iterate w_{t+1} according to 3;
end
`/* Squared-geometric-CDF randomized output */`
Randomly select an index S from $[t_0 + 1 : T]$ with probability $\mathbb{P}(S = t) \propto (1 - (1-\beta)^{t-t_0})^2$.
Output : w_S.

E MOMENTUM-RESET SGDM AND ANALYSIS

In this section, we introduce momentum-reset SGDM (MR-SGDM), a minor variant of standard SGDM that resets the momentum term to zero exactly once over the course of iterations. We then analyze its convergence for weakly convex problems under the L_2-regularized stationarity criterion. The core proof strategies developed for this modified framework closely mirror those for standard SGDM presented in the main manuscript.

Motivation. The primary motivation for introducing momentum reset to SGDM is to enable tight convergence analysis both time-decaying and adaptive learning rate schedules. Beyond facilitating analysis, this momentum-reset technique has also recently demonstrated strong empirical performance in the training of large language models (LLMs) and deep reinforcement learning (RL) tasks [8; 25]. Our convergence results for MR-SGDM thus furnish a rigorous theoretical foundation for explaining the empirical success of such kind of momentum-reset techniques.

E.1 MOMENTUM-RESET SGDM

As outlined in Algorithm 2, MR-SGDM follows exactly the generic SGDM update rules stated in 2 and 3, with the only difference being that the momentum term m_t is reset to zero at a user-specified time step $t = t_0$. The algorithm's output is randomly sampled from a truncated squared-geometric-CDF distribution over $[t_0 + 1 : T]$, analogous to Option-I in Algorithm 1.More precisely, given a total iteration budget T, we sample the stopping index $S \in [t_0 + 1 : T]$ from a truncated distribution whose unnormalized probability mass is proportional to the squared CDF of a geometric distribution with parameter β, i.e., $\mathbb{P}(S = t) \propto (1 - (1-\beta)^{t-t_0})^2$. Under this non-uniform sampling scheme, more recent iterates are assigned a higher probability of being selected as the final output.

Algorithm 2 can also be interpreted as a warm-start variant of SGDM: at time step t_0, the momentum update is restarted from scratch, and the iterate is initialized with w_{t_0}. In particular, when setting $t_0 = 0$, this procedure reduces to the vanilla SGDM. For comparison, the proposed MR-SGDM algorithm differs fundamentally from the periodically restarted O2NC framework introduced by Ji and Yuan [26], which resets the momentum term periodically at fixed iteration intervals.

In the following two subsections, we establish the convergence of MR-SGDM under decaying and adaptive learning rates. The convergence of MR-SGDM with constant learning rates can be recovered by setting $t_0 = 0$ in Algorithm 2, and the proof is identical to that of Algorithm 1 with Option-I.

E.2 RESULTS UNDER TIME-DECAYING LEARNING RATES

We analyze the convergence of Algorithm 2 with time-decaying learning rate schedule $\eta_t = \frac{\eta}{\sqrt{t+1}}$, where $\eta > 0$ is a fixed constant to be determined. The following is our main result in this setting.

Theorem 5. *Suppose that Assumption 1, 2 and 3 hold. Suppose that $T \geq 8$ and $\beta \in \left[\frac{4}{T}, \frac{1}{2}\right]$. Let $\eta_t = \frac{\eta}{\sqrt{t+1}}$ with $\eta > 0$. Suppose that $\eta_{t_0} \leq \frac{1}{2\rho}$ for some $t_0 \in \left[\left\lfloor \frac{T}{2} \right\rfloor\right]$. Then for any $\mu \leq \frac{\beta^2 T}{6\eta^2(1-\beta)}$, the randomized output w_S generated by Algorithm 2 with t_0 satisfies*

$$\mathbb{E}\left[\left\|\partial_{+\mu} f(w_S)\right\|^2\right] \leq \frac{5T}{T-2t_0}\left(\frac{2M}{\eta\sqrt{t_0+1}} + 3\beta G^2 \log\left(\frac{T}{t_0}\right)\right).$$

Proof. See Appendix F.1 for a proof of this result. □

As a direct consequence of Theorem 5, the following corollary further establishes the complexity rates of Algorithm 2 with suitable choice of t_0.

Corollary 4. *Suppose that Assumption 1, 2 and 3 hold. Let $\mu, \epsilon > 0$ be the target parameters of L_2-regularized stationarity and T be the total budget of iterates. Set $\eta_t = \frac{1}{\mu^{1/4}\sqrt{t+1}}$, $\beta = \frac{\sqrt{6}\mu^{1/4}}{T^{1/2}}$, and $t_0 = \left\lfloor \frac{T}{4} \right\rfloor$. Suppose that T is sufficiently large such that*

$$T \geq \frac{(176G^2 + 40M)^2\mu^{1/2}}{\epsilon^4} + \frac{3 + 16\rho^2}{\mu^{1/2}} + 24\mu^{1/2}.$$

Then the randomized output generated by Algorithm 2 satisfies $\mathbb{E}\left[\left\|\partial_{+\mu} f(w_S)\right\|\right] \leq \epsilon$.

Proof. See Appendix F.2 for its proof. □

Remark 16. *The sample complexity bounds in Corollary 4 are dominated by the rate $\mathcal{O}(\mu^{1/2}\epsilon^{-4})$, which is tight with respect to μ, ϵ according to Theorem 4. Furthermore, the weak convexity parameter ρ is allowed to scale up to $\mathcal{O}(\mu^{1/2}\epsilon^{-2})$ in both bounds without sacrificing optimality.*

E.3 Results under adaptive learning rates

We further analyze the convergence of Algorithm 2 when substantialized to AMSGrad using the adaptive learning rate schedule 8.

Theorem 6. *Suppose that Assumption 1, 2 and 3 hold. Suppose that $T \geq 8$ and $\beta_1 \in \left[\frac{4}{T}, \frac{1}{2}\right]$. Let η_t be chosen according to 8 with an arbitrary $\eta > 0$. Consider some $t_0 \in \left[\left\lfloor \frac{T}{2} \right\rfloor\right]$. Suppose that $\eta_{t_0} \leq \frac{1}{2\rho}$ and $(1-\beta_1)^{t_0} \leq 0.5$. Suppose that $\hat{v}_0 \leq G^2$ and $\varepsilon \geq G\sqrt{2\beta_2}$. Then for any $\mu \leq \frac{(G+\varepsilon)^2\beta_1^2}{6\eta^2(1-\beta_1)}$, the randomized output w_S generated by Algorithm 2 satisfies*

$$\mathbb{E}\left[\left\|\partial_{+\mu} f(w_S)\right\|^2\right] \leq \frac{16(G+\varepsilon)^2 M}{(T-2t_0)(\sqrt{\hat{v}_0}+\varepsilon)\eta} + \frac{48(G+\varepsilon)^2\beta_1}{T-2t_0}\left(\frac{1}{\beta_2} + \frac{2}{\beta_2}\log\left(\frac{2}{\beta_2}\right) + \frac{T-t_0}{1-\beta_2}\right).$$

Proof. See Appendix F.3 for a proof of this result. □

The following corollary further establishes the iteration complexity of Algorithm 2 (with $t_0 \approx \frac{3T}{8}$) under the adaptive learning rate schedule of AMSGrad.

Corollary 5. *Suppose that Assumption 1, 2 and 3 hold. Let $\mu, \epsilon > 0$ be the target parameters of L_2-regularized stationarity and T be the total budget of iterates. Let $\varepsilon > 0$ be an arbitrary safeguard value. Set $t_0 = \left\lfloor \frac{3T}{8} \right\rfloor$, $\eta = \frac{\sqrt{\hat{v}_0}+\varepsilon}{\mu^{1/4}T^{1/2}}$, $\beta_1 = \frac{\sqrt{6}\mu^{1/4}}{T^{1/2}}$, and $\beta_2 = \frac{\log T}{T}$. Suppose that T is sufficiently large such that*

$$T \geq \frac{64^2(G+\varepsilon)^4\mu^{1/2}}{\epsilon^4}\left(\frac{M}{(\sqrt{\hat{v}_0}+\varepsilon)^2} + 45\right)^2 + \frac{3+16\rho^2}{\mu^{1/2}} + 24\mu^{1/2} + \frac{8G^2}{\varepsilon^2}\log\left(\frac{G}{\varepsilon}\right).$$

Then the randomized output generated by Algorithm 2 satisfies $\mathbb{E}\left[\left\|\partial_{+\mu} f(w_S)\right\|\right] \leq \epsilon$.

Proof. See Appendix F.4 for a proof of this result. □

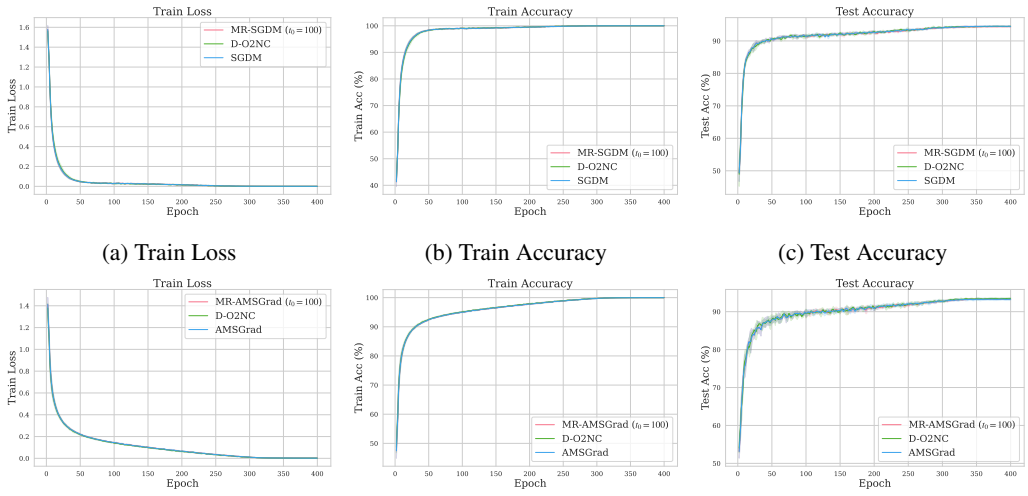
Remark 17. *Recently, the momentum-reset scheme has been observed to effectively enhance both the training stability and resource efficiency of Adam optimizers in LLM training and deep RL tasks [25; 8]. Yet, the theoretical understanding of this phenomenon remains limited. Our convergence results for momentum-reset AMSGrad, in turn, provide theoretical evidence to justify the empirical success of the momentum-reset technique in deep learning.*

E.4 PRELIMINARY EXPERIMENTAL RESULTS

To validate the effectiveness of MR-SGDM and MR-AMSGrad, this appendix presents and analyzes the corresponding experimental results in detail. All experiments are conducted on the CIFAR-10 image [33] classification task using the ResNet-101 architecture [24]. The details of the training setup are provided in the corresponding sections. To facilitate reproducibility, we also report the hardware and software configurations used in our experiments. All experiments are run on a server equipped with an AMD EPYC 7K62 48-core CPU, four NVIDIA RTX 4090D GPUs with 24 GB of memory each, and 256 GB of system RAM. Our implementation is based on Python 3.12 and PyTorch 2.6.

E.4.1 COMPARISON WITH THE BASELINE OPTIMIZERS

We first present a comparison between our method and the baseline algorithms. As baselines, we consider SGDM, AMSGrad, and D-O2NC [26], where D-O2NC restarts the momentum multiple times during the training phase. In our implementation of D-O2NC, we restart the momentum once per epoch. For SGDM, we adopt an initial learning rate of 0.01 with `CosineAnnealingLR`, a momentum coefficient of 0.1, and a weight decay of 5×10^{-4}. For AMSGrad, we use an initial learning rate of 0.001 with `CosineAnnealingLR`, momentum parameters of $(0.1, 0.001)$, and a weight decay of 5×10^{-4}. For MR-SGDM and MR-AMSGrad, we restart the momentum at the 100th epoch. All models are trained for 400 epochs with a batch size of 256 (each epoch taking approximately *50 seconds*). We record the training loss, training accuracy, and test accuracy at each epoch during training. The experimental results are shown in Figure 1.
Across both SGDM and AMSGrad, the training loss, training accuracy, and test accuracy curves of MR-SGDM/MR-AMSGrad and D-O2NC almost perfectly coincide with those of the corresponding baseline optimizers. This indicates that neither the proposed momentum-restarting strategy nor D-O2NC induces any noticeable change in the optimization dynamics or the final generalization performance.



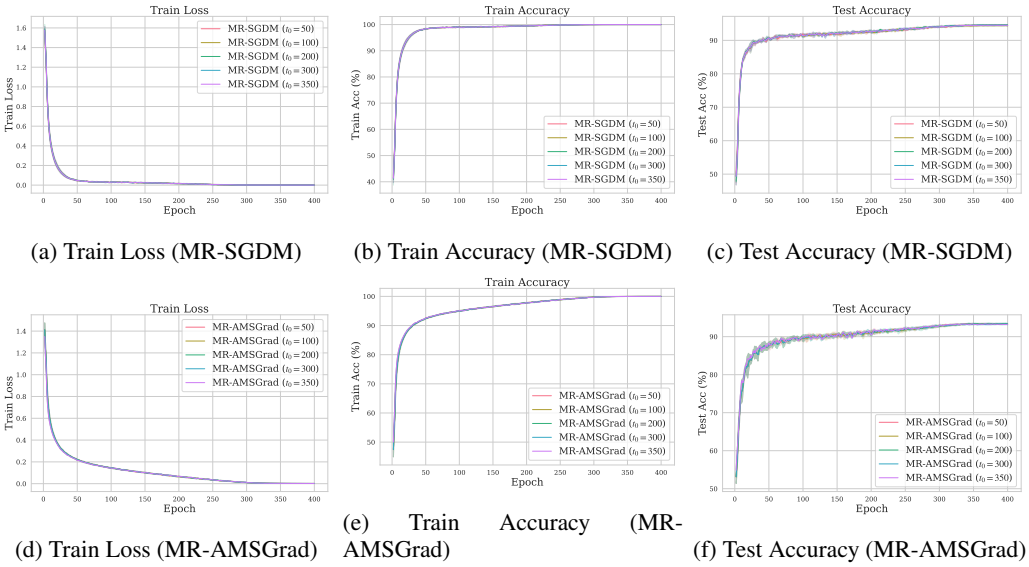
(a) Train Loss (b) Train Accuracy (c) Test Accuracy

(d) Train Loss (e) Train Accuracy (f) Test Accuracy

Figure 1: Experimental results on various optimizers, the top row reports results based on SGDM, while the bottom row shows results based on AMSGrad.

E.4.2 RESULTS ON VARIOUS RESTARTING POINTS

In this section, we investigate the impact of restarting the momentum at different epochs on the final model performance. Specifically, we set $t_0 \in \{50, 100, 200, 300, 350\}$, and apply the resetting strategy to both SGDM and AMSGrad, while keeping all other experimental settings identical to those described above. The results are shown in Figure 2. For both MR-SGDM and MR-AMSGrad,



(a) Train Loss (MR-SGDM) (b) Train Accuracy (MR-SGDM) (c) Test Accuracy (MR-SGDM)

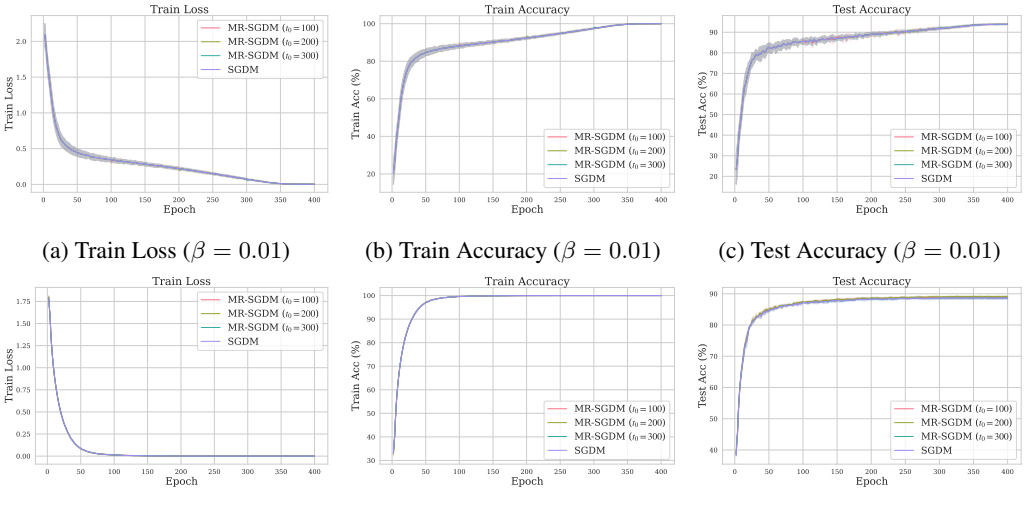
(d) Train Loss (MR-AMSGrad) (e) Train Accuracy (MR-AMSGrad) (f) Test Accuracy (MR-AMSGrad)

Figure 2: Experimental results of different t_0.

the training loss and accuracy curves corresponding to different values of t_0 almost perfectly overlap, indicating that the choice of t_0 has negligible influence on the optimization dynamics.

E.4.3 Results of other hyperparameters

In the previous experiments, we set momentum coefficient $\beta = 0.1$ and the initial learning rate lr to 0.01 for SGDM. In this section, we study two additional configurations: $\beta = 0.01$ with an initial learning rate of 0.01, and $\beta = 0.1$ with an initial learning rate of 0.001. We choose $t_0 \in \{100, 200, 300\}$, and the results are shown in Figure 3. Across different choices of the t_0 and SGDM hyperparameters, the training and test curves of MR-SGDM almost perfectly match those of SGDM, indicating that the proposed momentum restarting strategy has negligible impact on both optimization dynamics and final generalization performance.



(a) Train Loss ($\beta = 0.01$) (b) Train Accuracy ($\beta = 0.01$) (c) Test Accuracy ($\beta = 0.01$)

(d) Train Loss (lr=0.001) (e) Train Accuracy (lr=0.001) (f) Test Accuracy (lr=0.001)

Figure 3: Experimental results of various hyperparameter settings.

F PROOFS IN SECTION E

F.1 PROOF OF THEOREM 5

Proof. Let us consider $a_t \equiv \frac{\eta}{\sqrt{T+1}}$ for any $t \in [T]$. Then based on the conditions in the theorem, it must hold deterministically that $\eta_{t_0} \leq \frac{1}{2\rho}$, and $a_t \leq \eta_t \leq \eta_{t-1}$ for all $t \in [T]$. We can apply Lemma 1 to obtain

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=t_0+1}^{T}\left(1-(1-\beta)^{t-t_0}\right)^2\left(\frac{\eta^2}{T+1}\left\|\bar{g}_{[t_0+1:t]}\right\|^2+\frac{\beta^2}{6(1-\beta)}V^2_{[t_0+1:t]}(w_t)\right)\right]\\
\leq&\mathbb{E}\left[\frac{6\eta^2\|m_{t_0}\|^2}{\beta(t_0+1)}+\frac{4\eta\Delta f(w_{t_0})}{\sqrt{t_0+1}}+6\beta\eta^2\sum_{t=t_0+1}^{T}\frac{\|\hat{g}(w_t;z_t)\|^2}{t}\right] \qquad (21)\\
\leq&\frac{4\eta M}{\sqrt{t_0+1}}+6\beta G^2\eta^2\log\left(\frac{T}{t_0}\right),
\end{aligned}$$

where in the second inequality we have used the restart setting $m_{t_0}=0$ in Algorithm 2, $\Delta f(w_{t_0}) = f(w_{t_0})-f^* \leq M$ by Assumption 3, as well as the G-Lipschitzness of the loss function. In view of the above, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=t_0+1}^{T}\frac{\left(1-(1-\beta)^{t-t_0}\right)^2}{A_{[t_0+1:T]}}\left(\frac{\eta^2}{T+1}\left\|\bar{g}_{[t_0+1:t]}\right\|^2+\frac{\beta^2}{6(1-\beta)}V^2_{[t_0+1:t]}(w_t)\right)\right]\\
\leq&\frac{1}{A_{[t_0+1:T]}}\left(\frac{4\eta M}{\sqrt{t_0+1}}+6\beta G^2\eta^2\log\left(\frac{T}{t_0}\right)\right)\\
\leq&\frac{4}{T-2t_0}\left(\frac{2\eta M}{\sqrt{t_0+1}}+3\beta G^2\eta^2\log\left(\frac{T}{t_0}\right)\right),
\end{aligned}$$

where in the last step we have used $A_{[t_0+1:T]} \geq T-t_0-\frac{2}{\beta} \geq \frac{T}{2}-t_0 > 0$ according to Lemma 7, as well as the conditions on β and t_0. Finally, based on the above inequality and the definition of random index S distributed over $[t_0+1:T]$ we arrive at

$$\begin{aligned}
&\mathbb{E}\left[\left\|\bar{g}_{[t_0+1:S]}\right\|^2+\frac{\beta^2 T}{6\eta^2(1-\beta)}V^2_{[t_0+1:S]}(w_S)\right]\\
\leq&\mathbb{E}\left[\left\|\bar{g}_{[t_0+1:S]}\right\|^2+\frac{\beta^2(T+1)}{6\eta^2(1-\beta)}V^2_{[t_0+1:S]}(w_S)\right]\\
\leq&\frac{4(T+1)}{T-2t_0}\left(\frac{2M}{\eta\sqrt{t_0+1}}+3\beta G^2\log\left(\frac{T}{t_0}\right)\right)\\
\leq&\frac{5T}{T-2t_0}\left(\frac{2M}{\eta\sqrt{t_0+1}}+3\beta G^2\log\left(\frac{T}{t_0}\right)\right),
\end{aligned}$$

where in the last step we have used $T \geq 4$. Since $\mu \leq \mu' = \frac{\beta^2 T}{6\eta^2(1-\beta)}$, in view of Proposition 2 we get

$$\left\|\partial_{+\mu}f(w_S)\right\|^2 \leq \left\|\partial_{+\mu'}f(w_S)\right\|^2 \leq \left\|\bar{g}_{[t_0+1:S]}\right\|^2+\mu' V^2_{[t_0+1:S]}(w_S).$$

Combining the preceding two inequalities leads to the following desired bound:

$$\mathbb{E}\left[\left\|\partial_{+\mu}f(w_S)\right\|^2\right] \leq \frac{5T}{T-2t_0}\left(\frac{2M}{\eta\sqrt{t_0+1}}+3\beta G^2\log\left(\frac{T}{t_0}\right)\right).$$

The proof is completed. □

F.2 PROOF OF COROLLARY 4

Proof. Let us set $t_0 = \left\lfloor \frac{T}{4} \right\rfloor$ in Algorithm 2, and set $\eta = \frac{1}{\mu^{1/4}}$ in Theorem 5. Under the condition of T we must have $T \geq \mu^{-1/2}+24\mu^{1/2} > 8$ and $\beta = \frac{\sqrt{6}\mu^{1/4}}{T^{1/2}} \in \left[\frac{4}{T}, \frac{1}{2}\right]$. Again with the given

choice of β, η, it can be readily shown that $\frac{\beta^2 T}{6\eta^2(1-\beta)} \geq \frac{\beta^2 T}{6\eta^2} = \mu$. Also under the condition of T, it can be verified that $\eta_{t_0} = \frac{1}{\mu^{1/4}\sqrt{t_0+1}} \leq \frac{2}{\mu^{1/4}T^{1/2}} \leq \frac{1}{2\rho}$. Therefore, all conditions of Theorem 5 are fulfilled under the given choice of hyper-parameters, and we can thus apply the corresponding bound to obtain

$$\begin{aligned}\mathbb{E}\left[\|\partial_{+\mu} f(w_S)\|^2\right] \leq & \frac{5T}{(T-2t_0)}\left(\frac{3G^2}{\beta(t_0+1)} + \frac{2M}{\eta\sqrt{t_0+1}} + 3\beta G^2 \log\left(\frac{eT}{t_0+1}\right)\right) \\ \leq & \frac{40M\mu^{1/4}}{T^{1/2}} + \frac{30\sqrt{6}\log(4e)G^2\mu^{1/4}}{T^{1/2}} \\ \leq & \frac{\mu^{1/4}}{T^{1/2}}\left(40M + 176G^2\right) \leq \epsilon^2,\end{aligned}$$

where in the last step we have used the condition on T. The desired bound then follows from the fact $\mathbb{E}\left[\|\partial_{+\mu} f(w_S)\|\right] \leq \sqrt{\mathbb{E}\left[\|\partial_{+\mu} f(w_S)\|^2\right]} \leq \epsilon$. This gives the desired bound in part(b). The proof is completed. □

F.3 Proof of Theorem 6

Proof. Let us consider $a_t \equiv \frac{\eta}{G+\varepsilon}$. Since $\eta_{t_0} \leq \frac{1}{2\rho}$ and $a_t \leq \eta_t \leq \eta_{t-1}$ hold almost surely (based on Lemma 9), we can apply Lemma 1 to obtain

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=t_0+1}^{T}\left(1-(1-\beta_1)^{t-t_0}\right)^2\left(\frac{\eta^2}{(G+\varepsilon)^2}\left\|\bar{g}_{[t_0+1:t]}\right\|^2 + \frac{\beta_1^2}{6(1-\beta_1)}V^2_{[t_0+1:t]}(w_t)\right)\right] \\ \leq & \mathbb{E}\left[\frac{6\eta_{t_0}^2\|m_{t_0}\|^2}{\beta_1} + 4\eta_{t_0}\Delta f(w_{t_0}) + 6\beta_1\sum_{t=t_0+1}^{T}\eta_{t-1}^2\|\hat{g}(w_t;z_t)\|^2\right] \\ \overset{\zeta_1}{\leq} & \mathbb{E}\left[\frac{4\eta M}{(\sqrt{\hat{v}_0}+\varepsilon)(1-(1-\beta_1)^{t_0})} + 6\beta_1\eta^2\sum_{t=t_0+1}^{T}\frac{\|\hat{g}(w_t;z_t)\|^2}{(\sqrt{\hat{v}_{t-1}}+\varepsilon)^2(1-(1-\beta_1)^{t-1})^2}\right] \\ \overset{\zeta_2}{\leq} & \mathbb{E}\left[\frac{8\eta M}{\sqrt{\hat{v}_0}+\varepsilon} + 6\beta_1\eta^2\sum_{t=t_0+1}^{T}\frac{4\|\hat{g}(w_t;z_t)\|^2}{\left(\sqrt{\frac{v_{t-1}}{1-(1-\beta_2)^{t-1}}}+\varepsilon\right)^2}\right] \\ \overset{\zeta_3}{\leq} & \mathbb{E}\left[\frac{8\eta M}{\sqrt{\hat{v}_0}+\varepsilon} + 24\beta_1\eta^2\sum_{t=t_0+1}^{T}\frac{\|\hat{g}(w_t;z_t)\|^2}{\left(\sqrt{v_{t-1}}+\varepsilon\right)^2}\right] \\ \leq & \mathbb{E}\left[\frac{8\eta M}{\sqrt{\hat{v}_0}+\varepsilon} + 24\beta_1\eta^2\sum_{t=t_0+1}^{T}\frac{\|\hat{g}(w_t;z_t)\|^2}{v_{t-1}+\varepsilon^2}\right] \\ \leq & \frac{8\eta M}{\sqrt{\hat{v}_0}+\varepsilon} + 24\beta_1\eta^2\left(\frac{1}{\beta_2} + \frac{2}{\beta_2}\log\left(\frac{2}{\beta_2}\right) + \frac{T-t_0}{1-\beta_2}\right),\end{aligned} \tag{22}$$

where we have used in "ζ_1" the restart setting $m_{t_0} = 0$, the boundedness of f by Assumption 3, and Lemma 9, in "ζ_2" the condition $(1-\beta_1)^t \leq 0.5$ for $t \geq t_0$ and the fact $\hat{v}_t \geq \frac{v_t}{1-(1-\beta_2)^t}$ by definition, and in "ζ_3" the fact $1-(1-\beta_2)^t \in (0,1)$, while in the last step we have have applied Lemma 10 with β, a_t, x_t substantialized by $\beta_2, v_t, \|\hat{g}(w_t;z_t)\|$, respectively. In view of 22, we have

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=t_0+1}^{T}\frac{\left(1-(1-\beta_1)^{t-t_0}\right)^2}{A_{[t_0+1:T]}}\left(\frac{\eta^2}{(G+\varepsilon)^2}\left\|\bar{g}_{[t_0+1:t]}\right\|^2 + \frac{\beta_1^2}{6(1-\beta_1)}V^2_{[t_0+1:t]}(w_t)\right)\right] \\ \leq & \frac{1}{A_{[t_0+1:T]}}\left(\frac{8\eta M}{\sqrt{\hat{v}_0}+\varepsilon} + 24\beta_1\eta^2\left(\frac{1}{\beta_2} + \frac{2}{\beta_2}\log\left(\frac{2}{\beta_2}\right) + \frac{T-t_0}{1-\beta_2}\right)\right) \\ \leq & \frac{16}{T-2t_0}\left(\frac{\eta M}{\sqrt{\hat{v}_0}+\varepsilon} + 3\beta_1\eta^2\left(\frac{1}{\beta_2} + \frac{2}{\beta_2}\log\left(\frac{2}{\beta_2}\right) + \frac{T-t_0}{1-\beta_2}\right)\right),\end{aligned}$$

where in the last step we have used $A_{[t_0+1:T]} \geq T - t_0 - \frac{2}{\beta_1} \geq \frac{T}{2} - t_0 > 0$ according to Lemma 7, as well as the conditions on β_1 and t_0. Finally, based on the above inequality and the definition of random index S distributed over $[t_0 + 1 : T]$ we arrive at

$$\begin{aligned}
&\mathbb{E}\left[\left\|\bar{g}_{[t_0+1:S]}\right\|^2 + \frac{(G+\varepsilon)^2\beta_1^2}{6\eta^2(1-\beta_1)}V^2_{[t_0+1:S]}(w_S)\right] \\
\leq &\frac{16(G+\varepsilon)^2}{(T-2t_0)\eta^2}\left(\frac{\eta M}{\sqrt{\hat{v}_0}+\varepsilon} + 3\beta_1\eta^2\left(\frac{1}{\beta_2} + \frac{2}{\beta_2}\log\left(\frac{2}{\beta_2}\right) + \frac{T-t_0}{1-\beta_2}\right)\right) \\
= &\frac{16(G+\varepsilon)^2 M}{(T-2t_0)(\sqrt{\hat{v}_0}+\varepsilon)\eta} + \frac{48(G+\varepsilon)^2\beta_1}{T-2t_0}\left(\frac{1}{\beta_2} + \frac{2}{\beta_2}\log\left(\frac{2}{\beta_2}\right) + \frac{T-t_0}{1-\beta_2}\right).
\end{aligned}$$

Since $\mu \leq \mu' = \frac{(G+\varepsilon)^2\beta_1^2}{6\eta^2(1-\beta_1)}$, in view of Proposition 2 we get

$$\left\|\partial_{+\mu} f(w_S)\right\|^2 \leq \left\|\partial_{+\mu'} f(w_S)\right\|^2 \leq \left\|\bar{g}_{[t_0+1:S]}\right\|^2 + \mu' V^2_{[t_0+1:S]}(w_S).$$

Combining the preceding two inequalities leads to the following desired bound:

$$\mathbb{E}\left[\left\|\partial_{+\mu} f(w_S)\right\|^2\right] \leq \frac{16(G+\varepsilon)^2 M}{(T-2t_0)(\sqrt{\hat{v}_0}+\varepsilon)\eta} + \frac{48(G+\varepsilon)^2\beta_1}{T-2t_0}\left(\frac{1}{\beta_2} + \frac{2}{\beta_2}\log\left(\frac{2}{\beta_2}\right) + \frac{T-t_0}{1-\beta_2}\right).$$

The proof is completed. □

F.4 Proof of Corollary 5

Proof. Consider $\eta = \frac{\sqrt{\hat{v}_0}+\varepsilon}{\mu^{1/4}T^{1/2}}$ and $t_0 = \left\lfloor \frac{3T}{8} \right\rfloor$ in Theorem 6. Under the condition of T, we must have $T > \frac{3}{\mu^{1/2}} + 24\mu^{1/2} > 8$ and

$$\beta_1 = \frac{\sqrt{6}\mu^{1/4}}{T^{1/2}} \in \left[\frac{4}{T}, \frac{1}{2}\right], \ (1-\beta_1)^{t_0} \leq \left(1 - \frac{4}{T}\right)^{3T/8} \leq e^{-3/2} \leq 0.5.$$

Consequently, based on Lemma 9 and the condition on T it can be verified that

$$\eta_{t_0} = \frac{\eta}{(\sqrt{\hat{v}_0}+\varepsilon)(1-(1-\beta_1)^{t_0})} \leq \frac{2\eta}{\sqrt{\hat{v}_0}+\varepsilon} = \frac{2}{\mu^{1/4}T^{1/2}} \leq \frac{1}{2\rho}.$$

Again with the given choice of β_1, η, it can be readily shown that

$$\frac{(G+\varepsilon)^2\beta_1^2}{6\eta^2(1-\beta_1)} \geq \frac{(G+\varepsilon)^2\beta_1^2}{6\eta^2} = \frac{(G+\varepsilon)^2\mu}{(\sqrt{\hat{v}_0}+\varepsilon)^2} \geq \mu.$$

Moreover, since $T \geq \max\left\{8, \frac{8G^2}{\varepsilon^2}\log\left(\frac{G}{\varepsilon}\right)\right\}$, according to Lemma 11 with $a = \frac{G^2}{2\varepsilon^2}$ and $q = 1$ we must have $G\sqrt{2\beta_2} = G\sqrt{\frac{2\log(T)}{T}} \leq \varepsilon$. In view of the above arguments, the conditions of Theorem 6 are fulfilled in our setting, and we can thus apply it to obtain

$$\begin{aligned}
&\mathbb{E}\left[\left\|\partial_{+\mu} f(w_S)\right\|^2\right] \\
\leq &\frac{16(G+\varepsilon)^2 M}{(T-2t_0)(\sqrt{\hat{v}_0}+\varepsilon)\eta} + \frac{48(G+\varepsilon)^2\beta_1}{T-2t_0}\left(\frac{1}{\beta_2} + \frac{2}{\beta_2}\log\left(\frac{2}{\beta_2}\right) + \frac{T-t_0}{1-\beta_2}\right) \\
\overset{\zeta_1}{\leq} &\frac{64(G+\varepsilon)^2 M}{(\sqrt{\hat{v}_0}+\varepsilon)\eta T} + 1152(G+\varepsilon)^2\beta_1 \\
= &\frac{64(G+\varepsilon)^2 M\mu^{1/4}}{(\sqrt{\hat{v}_0}+\varepsilon)^2 T^{1/2}} + \frac{1152\sqrt{6}(G+\varepsilon)^2\mu^{1/4}}{T^{1/2}} \\
= &\frac{64(G+\varepsilon)^2\mu^{1/4}}{T^{1/2}}\left(\frac{M}{(\sqrt{\hat{v}_0}+\varepsilon)^2} + 45\right) \leq \epsilon^2,
\end{aligned}$$

where in "ζ_1" we have used the choice of $\beta_2 = \frac{\log(T)}{T}$ and $t_0 = \left\lfloor \frac{3T}{8} \right\rfloor$ which imply $\frac{1}{\beta_2} + \frac{2}{\beta_2}\log\left(\frac{2}{\beta_2}\right) = \frac{T}{\log(T)} + \frac{2T}{\log(T)}\log\left(\frac{2T}{\log(T)}\right) \leq 5T$ and $\frac{T-t_0}{1-\beta_2} \leq \frac{5T/8}{1-\log(8)/8} \leq T$, and in the last step we have used the condition on T. The desired bound then follows from the fact $\mathbb{E}\left[\left\|\partial_{+\mu} f(w_s)\right\|\right] \leq \sqrt{\mathbb{E}\left[\left\|\partial_{+\mu} f(w_s)\right\|^2\right]} \leq \epsilon$. This proves the desired bound. □