The Convergence Behavior of Adam under Heavy-Tailed Noise

Yijiang Pang[1]

[1]Michigan State University, East Lansing, USA

Abstract

We establish the first convergence guarantees for the plain vector-form *Adam* optimizer under heavy-tailed stochastic noise. While several Adam variants are known to achieve optimal iteration complexity in bounded-variance nonconvex optimization, little is understood about their behavior when stochastic gradients admit only a bounded p-th central moment for some $p \in (1, 2]$, a setting increasingly observed in modern deep learning. To address this gap, we generalize the recent online-to-nonconvex conversion framework to accommodate heavy-tailed martingale-difference noise. Building on this generalized framework, we develop a discounted regret analysis for Adam, without restrictive parameter coupling. Our results show that Adam converges to (ρ, ϵ)-stationary points under heavy-tailed noise. However, it exhibits a suboptimal iteration complexity and p-dependent convergence, a suboptimality that persists even in the bounded-variance case ($p = 2$). When the domain radius is known and used to control the online-learner output, a standard setup in related literature, the convergence rate improves to match the optimal complexity. These findings provide new theoretical insight into the robustness and limitations of Adam in heavy-tailed regimes.

1 INTRODUCTION

Adaptive gradient methods, most notably Adam [Kingma, 2014], have become the de facto optimization tools for training modern deep neural networks. Despite their remarkable empirical success, the theoretical understanding of Adam-type algorithms remains incomplete, particularly in stochastic environments where gradient noise deviates from classical bounded-variance assumptions.

Recent empirical and theoretical studies [Battash et al., 2024, Ahn et al., 2023, Garg et al., 2021] suggest that stochastic gradients in deep learning often exhibit heavy-tailed behavior. Instead of satisfying a bounded-variance condition, the noise may only admit a bounded p-th central moment for some $p \in (1, 2]$. Such heavy-tailed phenomena have been observed in large-scale language model training and other high-dimensional learning problems. In these regimes, analyses relying on sub-Gaussian or bounded-variance assumptions are no longer applicable, and even classical stochastic gradient descent (SGD) [Robbins and Monro, 1951] may suffer from instability or divergence [Zhang et al., 2020].

Over the past few years, substantial progress has been made in understanding both non-adaptive and adaptive methods under heavy-tailed noise. Clipped SGD, normalized gradient methods, and sign-based optimizers have been shown to provide improved robustness in heavy-tailed settings [Zhang et al., 2020, Liu et al., 2023, Sun et al., 2024, He et al., 2025]. On the adaptive side, recent works analyze clipped or modified variants of AdaGrad and Adam-type methods and establish (high-probability) convergence guarantees [Chezhegov et al., 2024, Fradin et al., 2025]. More recently, classical algorithms have been shown to achieve optimal convergence guarantees in convex settings and near-optimal rates for non-smooth non-convex optimization under heavy-tailed noise [Liu, 2025].

Nevertheless, convergence guarantees for the Adam update under heavy-tailed noise, without algorithmic modification, remain largely unexplored. In fact, until very recently, rigorous convergence results for Adam-type methods under heavy-tailed stochastic gradients were unavailable [Yu et al., 2026]. Even in the bounded-variance case ($p = 2$), many existing analyses focus on simplified or modified variants of Adam rather than the exact algorithm used in practice.

A major conceptual development in recent years is the online-to-nonconvex conversion framework Cutkosky et al. [2023], Ahn et al. [2024b], Zhang and Cutkosky [2024], which establishes a principled connection between adap-

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:5244–5265.

tive optimization and online learning. We briefly recall the fundamental setup.

Online Linear Optimization (OLO) proceeds over T rounds. At round t, the learner selects $\mathbf{z}_t \in \mathbb{R}^d$ and then observes a linear loss $\ell_t(\cdot) := \langle \mathbf{v}_t, \cdot \rangle$, where $\mathbf{v}_t$ is a cost vector revealed by the environment. The goal is to minimize the *static regret* $\mathrm{Regret}_T(\mathbf{u}) := \sum_{t=1}^{T} \langle \mathbf{v}_t, \mathbf{z}_t - \mathbf{u} \rangle$, for any comparator $\mathbf{u}$ in a domain $\mathcal{D}$. To connect online learning with non-convex optimization, recent works introduce the notion of β*-discounted regret*. The discounted loss is defined as $\ell_t^{[\beta]}(\cdot) = \langle \beta^{-t} \mathbf{v}_t, \cdot \rangle$, leading to the discounted regret

$$\mathrm{Regret}_T^{[\beta]}(\mathbf{u}) = \sum_{t=1}^{T} \langle \beta^{T-t} \mathbf{v}_t, \mathbf{z}_t - \mathbf{u} \rangle.$$

Within this framework, effective nonconvex optimizers correspond to online learners achieving small discounted regret. This perspective provides structural insight into adaptive methods. For example, interpreting Adam as a Follow-the-Regularized-Leader (FTRL) procedure clarifies its update structure and helps explain its empirical advantages over SGD when problem properties are unknown [Ahn and Cutkosky, 2024]. Moreover, the conversion framework has enabled sharp iteration complexity results for Adam. However, existing analyses often rely on technical simplifications, for example, removing bias-correction terms $1/(1-\beta_1^t)$ and $1/(1-\beta_2^t)$ or imposing parameter couplings such as $\beta_2 = \beta_1^2$ for analytical tractability. While these assumptions facilitate proof techniques, they depart from the exact algorithm used in practice and partially obscure its intrinsic behavior. Furthermore, the current conversion framework does not directly accommodate heavy-tailed noise.

Given these developments, a central question remains open:

What convergence guarantees can be established for the Adam update, without restrictive modifications? This question becomes even more compelling in heavy-tailed settings.

Our contributions. In this work, we provide the first convergence guarantees for the Adam update under heavy-tailed stochastic noise. Our main contributions are summarized as follows:

- **Heavy-tail conversion theory.** We generalize the online-to-nonconvex conversion framework to accommodate heavy-tailed martingale-difference noise.
- **Vector-form Adam regret analysis.** We establish a complete discounted regret analysis for the exact vector-form Adam update without enforcing restrictive parameter coupling: (a) we only require $\beta_1 \leq \beta_2$; (b) the numerical stabilizer ϵ in Adam's denominator is treated as a fixed small constant and is not required to scale with either G or σ.
- **Heavy-tailed nonconvex convergence.** We show that Adam converges to (ρ, ϵ)-stationary points under heavy-tailed noise. The resulting complexity is p-dependent and suboptimal relative to the optimal nonsmooth stochastic rate; this suboptimality remains even in the bounded-variance case $p = 2$.
- **Optimal rates under known domain radius.** When the domain radius is known and used to clip the online-learner output, a standard assumption in related literature, the iteration complexity improves to match optimal rates.

2 RELATED WORK

Heavy-tailed stochastic optimization Heavy-tailed gradient noise, typically modeled by assuming only a bounded p-th central moment for some $p \in (1, 2]$, has been widely studied as a realistic alternative to bounded-variance assumptions in deep learning. In non-adaptive optimization, a large body of work develops robust methods for this regime. Zhang et al. [2019] analyzed clipped SGD and established convergence guarantees together with lower bounds under heavy-tailed noise. Gorbunov et al. [2020] and Nguyen et al. [2023] strengthened these results to high-probability guarantees and studied accelerated variants. Beyond explicit clipping, Gorbunov et al. [2020] and Yu et al. [2026] analyzed normalization-based and sign-based methods, showing that magnitude suppression can achieve optimal or near-optimal rates without explicit truncation. For adaptive methods, theoretical guarantees are comparatively limited. Chezhegov et al. [2024] studied clipped AdaGrad- and Adam-type algorithms and established high-probability convergence guarantees under heavy-tailed noise. Ilboudo et al. [2023] proposed robust moment estimators based on heavy-tailed modeling and analyzed their stability. These works demonstrate that adaptive methods can be made robust in heavy-tailed regimes, but require explicit algorithmic modifications.

Online-to-nonconvex conversion and heavy-tailed online learning A major recent development in adaptive optimization is the online-to-nonconvex conversion framework. Cutkosky et al. [2023], Ahn et al. [2024b], Zhang and Cutkosky [2024], Ahn et al. [2024a] connected nonconvex stationarity guarantees to regret bounds of online learners through discounted regret. These works established optimal iteration complexity in the bounded-variance setting for carefully designed online learners with momentum-type updates. Parallel to these developments, heavy-tailed online learning has also been studied. Zhang and Cutkosky [2022] established high-probability regret guarantees via gradient truncation techniques for online convex optimization. Liu [2025] established regret guarantees for classical online convex optimization algorithms without explicit truncation.

3 PRELIMINARIES

In this section, we introduce the assumptions and basic notions used throughout the paper. Our assumptions largely follow prior work on non-smooth non-convex stochastic optimization and adaptive methods [Cutkosky et al., 2023, Ahn and Cutkosky, 2024, Zhang and Cutkosky, 2024], while explicitly allowing only bounded p-th moments of the stochastic gradient noise. We also recall the notion of (ρ, ϵ)-stationarity, which serves as the optimality criterion in our analysis. This notion is standard in recent studies of non-smooth non-convex stochastic optimization [Zhang and Cutkosky, 2024, Ahn and Cutkosky, 2024, Zhang et al., 2019, Jordan et al., 2023, Tian et al., 2022]. Although (ρ, ϵ)-stationarity relaxes Goldstein stationarity [Goldstein, 1977], it retains sufficient structure to recover first-order stationarity under additional smoothness assumptions.

Assumption 3.1. Let $F : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function with the following properties:

- The **function** is bounded below: $\inf_{\mathbf{x}} F(\mathbf{x}) > -\infty$. We define $\Delta := F(\mathbf{x}_0) - \inf_{\mathbf{x}} F(\mathbf{x})$.
- The **function** F is well-behaved, i.e., $\forall \mathbf{x}$ and $\mathbf{y}$, $F(\mathbf{y}) - F(\mathbf{x}) = \int_0^1 \langle \nabla F(\mathbf{x} + t(\mathbf{y} - \mathbf{x}), \mathbf{y} - \mathbf{x}) \rangle dt$.
- The **function** F is G-Lipschitz, i.e., $\forall \mathbf{x}$, $\|\nabla F(\mathbf{x})\| \leq G$.
- For every query point $\mathbf{x}$, the **stochastic gradient** $\mathbf{g} \leftarrow \mathrm{StoGrad}(\mathbf{x}, r)$ satisfies $\mathbb{E}[\mathbf{g} \mid \mathbf{x}] = \nabla F(\mathbf{x}), \quad \mathbb{E}[\|\mathbf{g} - \nabla F(\mathbf{x})\|^p \mid \mathbf{x}] \leq \sigma^p$.

Definition 3.2 ((ρ, ϵ)-stationary point)**.** Supposing $F : \mathbb{R}^d \to \mathbb{R}$ is differentiable. Then $\mathbf{x}$ is a (ρ, ϵ)-stationary point of F if $\|\nabla F(\mathbf{x})\|^{[\rho]} \leq \epsilon$ where $\|\nabla F(\mathbf{x})\|^{[\rho]} := \inf_{p \in \mathcal{P}(\mathbb{R}^d), \mathbb{E}_{\mathbf{y} \sim p}[\mathbf{y}] = \mathbf{x}} \left\{ \|\mathbb{E}[\nabla F(\mathbf{y})]\| + \rho \mathbb{E}[\|\mathbf{y} - \mathbf{x}\|^2] \right\}$.

To facilitate an online-learning interpretation of Adam, we introduce an equivalent assumption on the cost vectors corresponding to the stochastic gradients in Assumption 3.1.

Assumption 3.3. For each $t \in [T]$, let $\mathbf{v}_t \in \mathbb{R}^d$ denote a stochastic cost vector revealed after the history $\mathcal{F}_{t-1}$. Define

$$\boldsymbol{\mu}_t := \mathbb{E}[\mathbf{v}_t \mid \mathcal{F}_{t-1}], \qquad \boldsymbol{\xi}_t := \mathbf{v}_t - \boldsymbol{\mu}_t.$$

Assume that, almost surely,

$$\|\boldsymbol{\mu}_t\| \leq G, \quad \mathbb{E}[\|\boldsymbol{\xi}_t\|^p \mid \mathcal{F}_{t-1}] \leq \sigma^p, \quad \mathbb{E}[\boldsymbol{\xi}_t \mid \mathcal{F}_{t-1}] = 0.$$

Assumption 3.4 (Ball-constrained domain)**.** Let $\mathcal{D} \subseteq \mathbb{R}^d$ denote the Euclidean ball of radius $D > 0$, $\mathcal{D} := \{x \in \mathbb{R}^d : \|x\| \leq D\}$. Any comparator $\mathbf{u} \in \mathcal{D}$ satisfies $\|\mathbf{u}\| \leq D$.

Assumption 3.4 is used to specify the comparator class in the online-learning analysis.

Algorithm 1 Discounted-to-nonconvex conversion algorithm [Zhang and Cutkosky, 2024]

Input: Initial point $\mathbf{x}_0$, T, online learner $\mathcal{A}$ outputting $\mathbf{z}$, and discounting factor $\beta \in (0, 1)$
for $t = 1$ **to** T **do**
  Receive $\mathbf{z}_t$ from $\mathcal{A}$
  Update $\mathbf{x}_t \leftarrow \mathbf{x}_{t-1} + \kappa_t \mathbf{z}_t$, where $\kappa_t \sim \mathrm{Exp}(1)$ i.i.d.
  Compute $\mathbf{g}_t \leftarrow \mathrm{StoGrad}(\mathbf{x}_t, r_t)$
  Send $\ell_t^{[\beta]}(\mathbf{z}_t) := \langle \beta^{-t} \mathbf{g}_t, \mathbf{z}_t \rangle$ to $\mathcal{A}$
end for

Remark 3.5. The symbol D plays two related but distinct roles. For the Adam regret bound, D is only the comparator radius and need not be used by the algorithm. In the known-domain-radius setting [Ahn and Cutkosky, 2024, Liu, 2025], D is additionally available to the algorithm and can be used to control the online-learner output. For example, Ahn and Cutkosky [2024] employ a D-based clipping operation $\mathrm{clip}_D(\mathbf{z}) := \frac{\mathbf{z}}{\|\mathbf{z}\|} \min(D, \|\mathbf{z}\|)$, to control the outputs of the online learner. In our analysis, this distinction explains the difference between the plain-Adam rate and the optimal rate obtained when D-based output control is available.

4 DISCOUNTED-TO-NONCONVEX CONVERSION

We now present a generalized discounted-to-nonconvex conversion framework that accommodates both heavy-tailed and bounded-variance noise. The results of this section are of independent interest beyond the analysis of Adam.

To obtain a (ρ, ϵ)-stationary point, both components in Definition 3.2 must be controlled: (i) the expected gradient norm/smoothed first-order term $\|\mathbb{E}[\nabla F(\mathbf{y})]\|$, and (ii) the variance term $\mathbb{E}[\|\mathbf{y} - \mathbf{x}\|^2]$. We bound these two terms separately.

4.1 BOUNDING THE TWO TERMS

Bounding the expected gradient norm via discounted regret We first establish a heavy-tailed analogue of the classical interaction between expected gradient norm and discounted regret.

The proof follows the standard discounted-to-nonconvex conversion argument [Ahn and Cutkosky, 2024], with a key modification to accommodate heavy-tailed noise. In particular, we invoke the von Bahr–Esseen inequality for martingale differences [von Bahr and Esseen, 1965] to control stochastic error terms without assuming bounded variance. The key result is summarized in Proposition4.2, together with its supporting Lemma 4.1; the proofs are deferred to Appendix A.

Lemma 4.1. *Suppose Assumption 3.3 holds. Then for any $\beta \in (0,1)$, $n \in [T]$, and comparator $\mathbf{u}_n \in \mathcal{D}$,*

$$\mathbb{E}\sum_{t=1}^{n} \beta^{n-t}\langle \boldsymbol{\xi}_t, \mathbf{u}_n\rangle \le 2(1-\beta)^{-\frac{1}{p}}\sigma D.$$

Proposition 4.2. *Suppose F satisfies Assumption 3.1. For any comparator sequence $\{\mathbf{u}_t\} \subseteq \mathcal{D}$, Algorithm 1*

$$\mathbb{E}_\tau \|\mathbb{E}_{\mathbf{y}_\tau} \nabla F(\mathbf{y}_\tau)\| \le 2(1-\beta+\frac{\beta}{T})(1-\beta)^{-\frac{1}{p}}\sigma + \frac{\Delta}{DT} + \frac{1}{DT}\left((1-\beta)\sum_{t=1}^{T}\mathbb{E}\mathrm{Regret}_t^{[\beta]}(\mathbf{u}_t) + \beta\mathbb{E}\mathrm{Regret}_T^{[\beta]}(\mathbf{u}_T)\right)$$

where τ is a random index in $[T]$ with distribution $\mathbb{P}(\tau = t) = \begin{cases} \frac{1-\beta^t}{T} & \text{if} \quad t = 1, \cdots, T-1, \\ \frac{1-\beta^T}{(1-\beta)T} & \text{if} \quad t = T \end{cases}$. Moreover, $\mathbf{y}_t$ is sampled from $\{\mathbf{x}_s\}_{s=1}^t$ according to $\mathbb{P}(\mathbf{y}_t = \mathbf{x}_s) = \frac{(1-\beta)\beta^{t-s}}{1-\beta^t}, s = 1, \ldots, t$.

When $p = 2$, the above result recovers the existing discounted-to-nonconvex conversion under bounded variance (Lemma 7 of Ahn and Cutkosky [2024]).

Bounding the variance term We now bound the second term in Definition 3.2. This result directly adapts existing analyses [Ahn et al., 2024a, Zhang and Cutkosky, 2024] and highlights the role of norm control in achieving optimal guarantees.

Lemma 4.3 (Lemma 8 of Ahn et al. [2024a], Lemma 3.2 of Zhang and Cutkosky [2024])**.** *Using the notations of Proposition 4.2. Let $\mathbb{E}[\mathbf{y}_\tau] = \widetilde{\mathbf{x}}_\tau$, then $\forall t$, it holds that*

$$\rho\mathbb{E}_\tau\mathbb{E}_{\mathbf{y}_\tau}\|\mathbf{y}_\tau - \widetilde{\mathbf{x}}_\tau\|^2 \le \frac{2\rho\beta\sup_{t\in[T]}\mathbb{E}\left[\|\mathbf{z}_t\|^2\right]}{(1-\beta)^2}$$

Thus, controlling $\|\mathbf{z}_t\|$, for example via D-based clipping, directly controls the variance term in the stationarity measure

4.2 OPTIMIZATION GUARANTEE AFTER CONVERSION

Combining Proposition 4.2 and Lemma 4.3, we obtain the following (ρ,ϵ)-stationarity guarantee under heavy-tailed noise:

$$\mathbb{E}_\tau\|\nabla F(\widetilde{\mathbf{x}}_\tau)\|^{[\rho]} \le 2(1-\beta+\frac{\beta}{T})(1-\beta)^{-\frac{1}{p}}\sigma + \frac{1}{DT}\left((1-\beta)\sum_{t=1}^{T}\mathbb{E}\mathrm{Regret}_t^{[\beta]}(\mathbf{u}_t) + \beta\mathbb{E}\mathrm{Regret}_T^{[\beta]}(\mathbf{u}_T)\right) + \frac{\Delta}{DT} + \frac{2\rho\beta\sup_{t\in[T]}\mathbb{E}\left[\|\mathbf{z}_t\|^2\right]}{(1-\beta)^2} \quad (1)$$

From an algorithmic perspective, an effective method should: (i) achieve small discounted regret and (ii) control the magnitude of $\mathbf{z}_t$.

With this heavy-tailed discounted-to-nonconvex conversion framework in place, our next focus is to establish a regret analysis for Adam under heavy-tailed noise. This constitutes the main technical contribution of the paper.

5 ALGORITHM: ADAM

In this section, we present Adam from an online-learning perspective. Algorithm 2 gives an equivalent formulation of Adam as a Follow-the-Regularized-Leader (FTRL) procedure with exponentially weighted costs and a quadratic regularizer. Throughout the paper, we analyze this *vector* version of Adam; an extension to coordinate-wise variants is discussed later.

With the choice of α_t in Algorithm 2[a], the output admits the algebraically equivalent form to the standard vector-form Adam update. For later use, we adopt the shorthand $\mathbf{z}_t = -\eta_t\mathbf{a}_t/(\sqrt{b_t}+\epsilon)$ defined in Algorithm 2[b].

Notation. We summarize the key quantities used throughout the analysis:

- $\mathbf{z}_t$: Incremental / output of online learner Adam
- $\{\alpha_t\}$: step-size sequence
- $\{\eta_t\}$: learning-rate sequence

Organization. This section states three preparatory components used in the discounted regret analysis: heavy-tail moment controls for Adam's adaptive denominator and the discounted cost norms that appear in the regret bound, a β_2-selection rule that yields a deterministic bound on $\|\mathbf{z}_t\|$, and monotonicity of effective step size α_t.

5.1 HEAVY-TAIL CONTROL OF DISCOUNTED NORMS

We next record the heavy-tail moment controls needed in the regret analysis. Unlike bounded-variance analyses, the quantity $\sqrt{b_{t+1}}$ and the discounted norm aggregate $\sqrt{\sum_{s=1}^t \beta^{t-s}\|\mathbf{v}_s\|^2}$ cannot be bounded by second moments. The key observation is that their expectations can still be controlled under a bounded p-th moment assumption by using the concavity of $x^{p/2}$ for $p \le 2$. The resulting bounds quantify the heavy-tail cost through the factor $(1-\beta)^{-(1/p-1/2)}$, which is the price of replacing a second-moment assumption by a p-moment assumption.

Proposition 5.1. *Under Assumption 3.3, for every $t \in [T]$,*

$$\mathbb{E}\sqrt{b_{t+1}} \le G + \sigma\left(\frac{2}{p}\right)^{1/p}(1-\beta_2)^{-\left(\frac{1}{p}-\frac{1}{2}\right)}.$$

Algorithm 2 Online learner: Adam (vector-form)

Input: $\beta \triangleq \beta_1, \beta_2 \in (0,1)$, step-size sequence $\{\alpha_t\}$, learning-rate sequence $\{\eta_t\}_{t\in[T]}$, and $\epsilon > 0$.
for $t = 1$ **to** T **do**
  $\mathbf{z}_t = \arg\min_{\mathbf{z}\in\mathbb{R}^d} \frac{\|\mathbf{z}\|^2}{2\alpha_t} + \sum_{s=1}^{t-1} \langle \beta_1^{-s}\mathbf{v}_s, \mathbf{z}\rangle = -\alpha_t \sum_{s=1}^{t-1} \beta_1^{-s}\mathbf{v}_s$ [a,b]
  Observe loss $\ell_t(\cdot) = \langle \mathbf{v}_t, \cdot\rangle$
end for

[a] Selecting $\alpha_t \triangleq \eta_t \frac{1-\beta_1}{1-\beta_1^{t-1}} \frac{\beta_1^{t-1}}{\sqrt{\frac{1-\beta_2}{1-\beta_2^{t-1}}\sum_{s=1}^{t-1}\beta_2^{t-1-s}\|\mathbf{v}_s\|^2+\epsilon}}$, the incremental $\mathbf{z}_t$ can be formulated as

$$\mathbf{z}_t = -\eta_t \frac{\frac{1-\beta_1}{1-\beta_1^{t-1}}\sum_{s=1}^{t-1}\beta_1^{t-1-s}\mathbf{v}_s}{\sqrt{\frac{1-\beta_2}{1-\beta_2^{t-1}}\sum_{s=1}^{t-1}\beta_2^{t-1-s}\|\mathbf{v}_s\|^2+\epsilon}} \xrightarrow[\text{More familiar Adam update}]{\text{Rolling as EMA}} \begin{cases} \mathbf{m}_t = \beta_1\mathbf{m}_{t-1} + (1-\beta_1)\mathbf{v}_{t-1} \\ v_t = \beta_2 v_{t-1} + (1-\beta_2)\|\mathbf{v}_{t-1}\|^2 \\ \mathbf{z}_t = -\eta_t \frac{\mathbf{m}_t/(1-\beta_1^{t-1})}{\sqrt{v_t/(1-\beta_2^{t-1})}+\epsilon}. \end{cases}$$

[b] For the notation simplification purpose, we further define

$$\mathbf{z}_t \triangleq -\frac{\eta_t\,\mathbf{a}_t}{\sqrt{b_t}+\epsilon} \quad \text{where} \quad \mathbf{a}_t = \frac{1-\beta_1}{1-\beta_1^{t-1}}\sum_{s=1}^{t-1}\beta_1^{t-1-s}\mathbf{v}_s, \quad b_t = \frac{1-\beta_2}{1-\beta_2^{t-1}}\sum_{s=1}^{t-1}\beta_2^{t-1-s}\|\mathbf{v}_s\|^2.$$

The proof of Proposition 5.1 is deferred to Appendix B.

Corollary 5.2. *Assume Assumption 3.3 holds. Then for any $\beta \in (0,1)$ and $t \in [T]$,*

$$\mathbb{E}\sqrt{\sum_{s=1}^{t}\beta^{t-s}\|\mathbf{v}_s\|^2} \le \sqrt{\frac{1-\beta^t}{1-\beta}}\left(G + \sigma\left(\frac{2}{p}\right)^{1/p}(1-\beta)^{-\left(\frac{1}{p}-\frac{1}{2}\right)}\right)$$

5.2 DETERMINISTIC OUTPUT BOUNDS AND EFFECTIVE STEP SIZE CONTROL

In this subsection, we first state a simple parameter condition that yields a deterministic bound on $\|\mathbf{z}_t\|$.

Remark 5.3 (β_2 and $C(\beta)$ selection strategy). Fix $\beta_1 \in (0,1)$. Choose any $\beta_2 \in [\beta_1, 1)$ and a constant $C(\beta) \ge 1$ such that $1-\beta_1 = C(\beta)(1-\beta_2)$. For an integer-valued choice, it suffices to take $C(\beta) \ge \left\lceil \frac{1-\beta_1}{1-\beta_2} \right\rceil$.

The following proposition (Proof detailed in Appendix B) shows that this mild relation between β_1 and β_2 implies a deterministic norm bound for Adam's output, which will be crucial for our in-expectation analysis.

Proposition 5.4. *Under Remark 5.3, Adam's output satisfies the deterministic bound* $\|\mathbf{z}_t\| \le \eta_t\sqrt{C(\beta)}, \quad \forall t \ge 1.$

It is also worth noting that Remark 5.3 and Proposition 5.4 allow $C(\beta)$ to be small, or even equal to 1 when $\beta_2 = \beta_1$

Additionally, the following Proposition 5.5 records the natural monotonicity of the resulting effective step size. Throughout the paper, we take a constant base step size, i.e., $\eta_t \equiv \eta$.

Proposition 5.5. *Let $\beta_1 \le \beta_2$, use $\eta_s = \eta$, and initialize $\alpha_1 := \alpha_2$. Then $\alpha_s = \frac{\eta(1-\beta_1)\beta_1^{s-1}}{(1-\beta_1^{s-1})(\sqrt{b_s}+\epsilon)}, s \ge 2$, is non-increasing: $\alpha_s \ge \alpha_{s+1}$ for all $s \ge 1$.*

6 REGRET AND OPTIMIZATION GUARANTEES

This section presents our main performance guarantees for Adam under heavy-tailed noise. We first establish a discounted regret bound for the Adam online learner. We then combine this regret bound with the heavy-tailed discounted-to-nonconvex conversion (Section 4) to obtain convergence guarantees to (ρ,ϵ)-stationary points. Finally, we show how the guarantee improves when the domain radius D is available to the algorithm and is used to control the online-learner output.

6.1 DISCOUNTED REGRET UNDER HEAVY TAILS

We analyze Adam as an online learner receiving discounted linear losses and producing increments $\mathbf{z}_t$ as in Algorithm 2. The comparator radius D specifies the regret benchmark; in the Adam result, D is not used to modify the algorithm. Our goal is to bound the discounted regret

$$\mathbb{E}\left[\mathrm{Regret}_t^{[\beta]}(\mathbf{u})\right] = \mathbb{E}\left[\sum_{s=1}^{t}\langle \beta_1^{t-s}\mathbf{v}_s,\, \mathbf{z}_s - \mathbf{u}\rangle\right].$$

The result of the regret bound under heavy-tailed noise is summarized in Theorem 6.2, together with its supporting Proposition 6.1.

Proposition 6.1. *Under Assumption 3.3, for* $t \geq T_0 := 2 + \frac{\log(1/(1-\beta_1))}{\log(\sqrt{\beta_2}/\beta_1)}$*, it holds that*

$$\mathbb{E}\left[\sum_{s=1}^{t} \frac{\alpha_s}{2} \beta_1^{t-2s} \|\overline{\mathbf{v}}_s\|^2\right] \leq \eta \frac{2\sqrt{\beta_2}(1-\beta_1)}{\beta_1(1-\beta_2)} \left(G + 2\sigma(1-\beta_2)^{-\left(\frac{1}{p}-\frac{1}{2}\right)}\right)$$

where $\overline{\mathbf{v}}_s := \mathrm{clip}_{\sqrt{\sum_{n=1}^{s-1} \beta_2^{2s-2n}\|\mathbf{v}_n\|^2}}(\mathbf{v}_s)$.

The proof of Proposition 6.1 is deferred to Appendix C.

Theorem 6.2. *Assume Assumption 3.3 holds, and employ the* $\beta_2, C(\beta)$ *selection strategy in Remark 5.3. For any* $t \geq T_0$ *and any comparator* $\mathbf{u} \in \mathcal{D}$ *with* $\|\mathbf{u}\| \leq D$*, if the online learner is Adam with* $\eta = \frac{D}{(1-\beta_1)^{1/4}}$*, then the discounted regret satisfies*

$$\mathbb{E}\left[\mathit{Regret}_t^{[\beta]}(\mathbf{u})\right] \leq \frac{C(R)D\left(G + 2\sigma(1-\beta_2)^{-\left(\frac{1}{p}-\frac{1}{2}\right)} + \epsilon\right)}{(1-\beta_1)^{3/4}}.$$

where $C(R) := \max\left(\frac{1}{2} + \frac{2\sqrt{\beta_2}(1-\beta_1)^{3/2}}{\beta_1(1-\beta_2)} + \sqrt{8C(\beta)}, \frac{1}{2} + \frac{2\sqrt{\beta_2}(1-\beta_1)}{\beta_1(1-\beta_2)} + \sqrt{8}\right)$.

Proof. We first establish the deterministic intermediate regret bound in equation 2, inspired by techniques presented in Orabona [2019], Ahn and Cutkosky [2024], Ahn et al. [2024b], Tim [2021], and then control each component appearing in this bound.

Part 1. Intermediate result (deterministic bound) for $\mathbf{Regret}_t^{[\beta]}(\mathbf{u})$**.** Define $F_s(\mathbf{z}) := \frac{1}{2\alpha_s}\|\mathbf{z}\|^2 + \sum_{n=1}^{s-1}\langle \beta_1^{-n}\mathbf{v}_n, \mathbf{z}\rangle$ and $\alpha_s := \eta_s \frac{(1-\beta_1)\beta_1^{s-1}}{(1-\beta_1^{s-1})(\sqrt{b_s}+\epsilon)}$, so

$$\mathbf{z}_s = \arg\min F_s(\mathbf{z}) = -\alpha_s \sum_{n=1}^{s-1} \beta_1^{-n}\mathbf{v}_n$$

is the Adam update.

Using $\sum(F_s(\mathbf{z}_s) - F_{s+1}(\mathbf{z}_{s+1})) = F_1(\mathbf{z}_1) - F_{t+1}(\mathbf{z}_{t+1})$, $F_1(\mathbf{z}_1) = \min_{\mathbf{x}} \frac{1}{2\alpha_1}\|\mathbf{x}\|^2$, and $F_{t+1}(\mathbf{z}_{t+1}) \leq F_{t+1}(\mathbf{u})$:

$$\sum_{s=1}^{t}\langle \beta_1^{-s}\mathbf{v}_s, \mathbf{z}_s - \mathbf{u}\rangle \leq \frac{\|\mathbf{u}\|^2}{2\alpha_{t+1}} + \sum_{s=1}^{t}\left[\underbrace{F_s(\mathbf{z}_s) - F_{s+1}(\mathbf{z}_{s+1}) + \langle \beta_1^{-s}\mathbf{v}_s, \mathbf{z}_s\rangle}_{\text{Term B}}\right].$$

Using $F_s(\mathbf{z}_{s+1}) - F_{s+1}(\mathbf{z}_{s+1}) = -\langle \beta_1^{-s}\mathbf{v}_s, \mathbf{z}_{s+1}\rangle + \frac{1}{2\alpha_s}\|\mathbf{z}_{s+1}\|^2 - \frac{1}{2\alpha_{s+1}}\|\mathbf{z}_{s+1}\|^2$ and $\alpha_s \geq \alpha_{s+1}$ (Proposition 5.5),

$$\begin{aligned}\text{Term B} &\leq F_s(\mathbf{z}_s) - F_s(\mathbf{z}_{s+1}) + \langle \beta_1^{-s}\mathbf{v}_s, \mathbf{z}_s - \mathbf{z}_{s+1}\rangle \\ &= \underbrace{F_s(\mathbf{z}_s) - F_s(\mathbf{z}_{s+1}) + \langle \beta_1^{-s}\overline{\mathbf{v}}_s, \mathbf{z}_s - \mathbf{z}_{s+1}\rangle}_{\text{B.1}} \\ &\quad + \underbrace{\langle \beta_1^{-s}\mathbf{v}_s - \beta_1^{-s}\overline{\mathbf{v}}_s, \mathbf{z}_s - \mathbf{z}_{s+1}\rangle}_{\text{B.2}}\end{aligned}$$

where $\overline{\mathbf{v}}_s := \mathrm{clip}_{\sqrt{\sum_{n=1}^{s-1} \beta_2^{2s-2n}\|\mathbf{v}_n\|^2}}(\mathbf{v}_s)$.

Term B.1:

- $F_s(\mathbf{z}_s) - F_s(\mathbf{z}_{s+1}) \leq -\frac{\|\mathbf{z}_s - \mathbf{z}_{s+1}\|^2}{2\alpha_s}$ by strong convexity and the variational inequality for the minimizer $\mathbf{z}_s = \arg\min F_s(\mathbf{z})$. (R1)
- Employing Young's inequality, $\langle \beta_1^{-s}\overline{\mathbf{v}}_s, \mathbf{z}_s - \mathbf{z}_{s+1}\rangle \leq \frac{\alpha_s}{2}\beta_1^{-2s}\|\overline{\mathbf{v}}_s\|^2 + \frac{\|\mathbf{z}_s - \mathbf{z}_{s+1}\|^2}{2\alpha_s}$. (R2)

Combining (R1)+(R2): Term B.1 $\leq \frac{\alpha_s}{2}\beta_1^{-2s}\|\overline{\mathbf{v}}_s\|^2$.

Term B.2:

$$\begin{aligned}&\underbrace{\sum_{s=1}^{t}\beta_1^{-s}\langle \mathbf{v}_s - \overline{\mathbf{v}}_s, \mathbf{z}_s - \mathbf{z}_{s+1}\rangle}_{\text{Part B.2}} \\ &\leq \sum_{s=1}^{t}\beta_1^{-s}\|\mathbf{v}_s - \mathrm{clip}_{\sqrt{\sum_{n=1}^{s-1} \beta_2^{2s-2n}\|\mathbf{v}_n\|^2}}(\mathbf{v}_s)\|\|\mathbf{z}_s - \mathbf{z}_{s+1}\| \\ &\overset{(a)}{\leq} 2\max_{s\in[t]}\|\mathbf{z}_s\| \sum_{s=1}^{t}\beta_1^{-s}\left(\|\mathbf{v}_s\| - \sqrt{\sum_{n=1}^{s-1}\beta_2^{2s-2n}\|\mathbf{v}_n\|^2}\right)_+ \\ &\overset{(b)}{\leq} 2\max_{s\in[t]}\|\mathbf{z}_s\| \sum_{s=1}^{t}\left(\beta_1^{-s}\|\mathbf{v}_s\| - \sqrt{\sum_{n=1}^{s-1}\beta_1^{-2n}\|\mathbf{v}_n\|^2}\right)_+ \\ &\leq 2\max_{s\in[t]}\|\mathbf{z}_s\| \sum_{s=1}^{t}\left(\sqrt{\sum_{n=1}^{s}\beta_1^{-2n}\|\mathbf{v}_n\|^2} - \sqrt{\sum_{n=1}^{s-1}\beta_1^{-2n}\|\mathbf{v}_n\|^2}\right) \\ &\leq 2\max_{s\in[t]}\|\mathbf{z}_s\| \sqrt{\sum_{s=1}^{t}\beta_1^{-2s}\|\mathbf{v}_s\|^2},\end{aligned}$$

where (a) uses the identity $\|\mathbf{v} - \mathrm{clip}_R(\mathbf{v})\| = (\|\mathbf{v}\| - R)_+$, and (b) uses $\beta_1 \leq \beta_2$.

Combining the results of Term B.1 and Term B.2, it suffices to have

$$\sum_{s=1}^{t}\langle \beta_1^{-s}\mathbf{v}_s, \mathbf{z}_s - \mathbf{u}\rangle \leq \frac{\|\mathbf{u}\|^2}{2\alpha_{t+1}} + \sum_{s=1}^{t}\frac{\alpha_s}{2}\beta_1^{-2s}\|\overline{\mathbf{v}}_s\|^2 + 2\max_{s\in[t]}\|\mathbf{z}_s\|\sqrt{\sum_{s=1}^{t}\beta_1^{-2s}\|\mathbf{v}_s\|^2}.$$

Multiplying β_1^t on both sides gives

$$\mathrm{Regret}_t^{[\beta]}(\mathbf{u}) \le \frac{\beta_1^t\|\mathbf{u}\|^2}{2\alpha_{t+1}} + \sum_{s=1}^{t}\frac{\alpha_s}{2}\beta_1^{t-2s}\|\overline{\mathbf{v}}_s\|^2 + 2\max_{s\in[t]}\|\mathbf{z}_s\|\sqrt{\sum_{s=1}^{t}\beta_1^{2t-2s}\|\mathbf{v}_s\|^2} \quad (2)$$

where $\overline{\mathbf{v}}_s := \mathrm{clip}_{\sqrt{\sum_{n=1}^{s-1}\beta_2^{2s-2n}\|\mathbf{v}_n\|^2}}(\mathbf{v}_s)$.

Part 2. Final regret bound. Taking expectation over equation 2,

$$\begin{aligned}&\mathbb{E}\left[\mathrm{Regret}_t^{[\beta]}(\mathbf{u})\right]\\ &\le \frac{D^2}{2\eta(1-\beta_1)}\left(G+2\sigma(1-\beta_2)^{-\left(\frac{1}{p}-\frac{1}{2}\right)}\right)\\ &\quad+\eta\frac{2\sqrt{\beta_2}(1-\beta_1)}{\beta_1(1-\beta_2)}\left(G+2\sigma(1-\beta_2)^{-\left(\frac{1}{p}-\frac{1}{2}\right)}\right)\\ &\quad+2\eta\sqrt{C(\beta)}\sqrt{\frac{1}{1-\beta_1^2}}\left(G+2\sigma(1-\beta_1^2)^{-\left(\frac{1}{p}-\frac{1}{2}\right)}\right)\end{aligned}$$

Here, the first term on the right-hand side uses Proposition 5.1, the second term uses Proposition 6.1, and the last term uses Corollary 5.2 and the deterministic increment bound in Proposition 5.4.

The above inequality can be further reformulated as

$$\begin{aligned}&\mathbb{E}\left[\mathrm{Regret}_t^{[\beta]}(\mathbf{u})\right]\\ &\le\left(\frac{D^2}{2\eta(1-\beta_1)}+\eta\frac{2\sqrt{\beta_2}(1-\beta_1)}{\beta_1(1-\beta_2)}+\eta\sqrt{C(\beta)}\sqrt{\frac{8}{1-\beta_1}}\right)\\ &\quad\times\left(G+2\sigma(1-\beta_2)^{-\left(\frac{1}{p}-\frac{1}{2}\right)}+\epsilon\right). \quad (3)\end{aligned}$$

Setting $\eta = \frac{D}{(1-\beta_1)^{1/4}}$ gives

$$\begin{aligned}\mathbb{E}\left[\mathrm{Regret}_t^{[\beta]}(\mathbf{u})\right] &\le \frac{D\left(G+2\sigma(1-\beta_2)^{-\left(\frac{1}{p}-\frac{1}{2}\right)}+\epsilon\right)}{(1-\beta_1)^{3/4}}\\ &\quad\times\left(\frac{1}{2}+\frac{2\sqrt{\beta_2}(1-\beta_1)^{3/2}}{\beta_1(1-\beta_2)}+\sqrt{8C(\beta)}\right),\end{aligned}$$

which concludes the proof. □

Next, we consider the regret bound when the learner output is explicitly controlled by the known radius D.

Corollary 6.3. *Assume Assumption 3.3 holds. If the learner is the constrained-FTRL form of Adam over the ball $\{\|\mathbf{z}\| \le D\}$, equivalently the Adam increment is clipped as $\mathbf{z} \leftarrow \mathrm{clip}_D(\mathbf{z})$, and $\eta = \frac{D}{\sqrt{1-\beta_1}}$, then*

$$\mathbb{E}\left[\mathit{Regret}_t^{[\beta]}(\mathbf{u})\right] \le \frac{C(R)D\left(G+2\sigma(1-\beta_2)^{-\left(\frac{1}{p}-\frac{1}{2}\right)}+\epsilon\right)}{\sqrt{1-\beta_1}}.$$

where $C(R) := \max\left(\frac{1}{2}+\frac{2\sqrt{\beta_2}(1-\beta_1)^{3/2}}{\beta_1(1-\beta_2)}+\sqrt{8C(\beta)}, \frac{1}{2}+\frac{2\sqrt{\beta_2}(1-\beta_1)}{\beta_1(1-\beta_2)}+\sqrt{8}\right)$.

Proof. For the clipped method, $\mathbf{z}_s$ is the FTRL minimizer over the ball $\{\|\mathbf{z}\| \le D\}$, which is equivalent to clipping the unconstrained Adam increment. The same decomposition as in Part 1 of Theorem 6.2 applies, and we begin from equation 3 but substitute deterministic increment bound term $\eta\sqrt{C(\beta)}$ as D

$$\begin{aligned}&\mathbb{E}\left[\mathrm{Regret}_t^{[\beta]}(\mathbf{u})\right]\\ &\le\left(\frac{D^2}{2\eta(1-\beta_1)}+\eta\frac{2\sqrt{\beta_2}(1-\beta_1)}{\beta_1(1-\beta_2)}+\sqrt{\frac{8}{1-\beta_1}}D\right)\\ &\quad\times\left(G+2\sigma(1-\beta_2)^{-\left(\frac{1}{p}-\frac{1}{2}\right)}+\epsilon\right).\end{aligned}$$

Setting $\eta = \frac{D}{\sqrt{1-\beta_1}}$ gives

$$\begin{aligned}\mathbb{E}\left[\mathrm{Regret}_t^{[\beta]}(\mathbf{u})\right] &\le \frac{D\left(G+2\sigma(1-\beta_2)^{-\left(\frac{1}{p}-\frac{1}{2}\right)}+\epsilon\right)}{\sqrt{1-\beta_1}}\\ &\quad\times\left(\frac{1}{2}+\frac{2\sqrt{\beta_2}(1-\beta_1)}{\beta_1(1-\beta_2)}+\sqrt{8}\right).\end{aligned}$$

This concludes the proof. □

In the stationarity results below, ϵ denotes the target accuracy. The numerical stabilizer in the Adam denominator is treated as a fixed small constant and is omitted from the displayed convergence rates to avoid overloading notation.

6.2 CONVERGENCE GUARANTEE UNDER HEAVY TAILS

We now combine the discounted regret bound (Theorem 6.2) with the heavy-tailed discounted-to-nonconvex conversion (Section 4) to obtain a (ρ,ϵ)-stationarity guarantee for Adam. At a high level, the conversion inequality equation 1 mainly involves two terms: (i) a regret term, and (ii) a variance term controlled by $\sup_t \mathbb{E}\|\mathbf{z}_t\|^2$. Theorem 6.2 controls (i), Proposition 5.4 controls (ii), and we choose (β_1, D, T) to balance all contributions.

Theorem 6.4. *Assume F satisfies Assumption 3.1 and fix any $\rho > 0$. Run Algorithm 1 with online learner Adam as in Algorithm 2. Then, for any $p \in (\frac{4}{3}, 2]$, there exists a random index $\tau \in [T]$ such that*

$$\mathbb{E}_\tau\left[\|\nabla F(\widetilde{\mathbf{x}}_\tau)\|^{[\rho]}\right] \le \left(8 + C(R) + 2C(R)C(\beta)^{\frac{1}{p}-\frac{1}{2}}\right)\epsilon,$$

where τ is a random index among $[T]$ such that $\mathbb{P}(\tau = t) = \begin{cases}\frac{1-\beta_1^t}{T} & \text{if} \quad t = 1,\cdots,T-1,\\ \frac{1-\beta_1^T}{(1-\beta_1)T} & \text{if} \quad t = T\end{cases}$. Moreover, the $\mathbf{y}_t$

is randomly distributed over $\{\mathbf{x}_s\}_{s=1}^t$ *as* $\mathbb{P}(\mathbf{y}_t = \mathbf{x}_s) = \frac{(1-\beta_1)\beta_1^{t-s}}{1-\beta_1^t}, s = 1, \ldots, t.$

In particular, it suffices to choose:

$\beta_1 = 1 - \left(\frac{\epsilon}{G+\sigma}\right)^{\frac{4p}{3p-4}}$, β_2 *so that* $1-\beta_1 = C(\beta)(1-\beta_2)$ *for a fixed constant* $C(\beta) \geq 1$, $D = \frac{(1-\beta_1)^{5/4}\epsilon^{1/2}}{C(\beta)\rho^{1/2}}$, *and*

$$T = \mathcal{O}\Big(\max\Big\{ (C(R)G+\sigma)^{\frac{8p-4}{3p-4}} \epsilon^{-\frac{8p-4}{3p-4}} \log\tfrac{G+\sigma}{\epsilon}, \Delta\rho^{\frac{1}{2}}(C(R)G+\sigma)^{\frac{5p}{3p-4}}\epsilon^{-\left(\frac{5p}{3p-4}+\frac{3}{2}\right)} \Big\}\Big).$$

Moreover, in the special case $p = 2$, *the bound further reduces to* $T = \mathcal{O}\Big(\max\Big\{ (C(R)G+\sigma)^6 \epsilon^{-6} \log\tfrac{G+\sigma}{\epsilon}, \Delta\rho^{\frac{1}{2}}(C(R)G+\sigma)^5\epsilon^{-\frac{13}{2}} \Big\}\Big).$

For readability, the full max-form conditions on T, including lower-order and burn-in terms, and the proof of Theorem 6.4, are given in Appendix C.

Discussion of rates. Theorem 6.4 implies that Adam converges in expectation to a (ρ, ϵ)-stationary with an $\mathcal{O}\left(\epsilon^{-3}\right)$ gap compared to the optimal rate under the same framework [Zhang and Cutkosky, 2024, Liu, 2025], in the bounded-variance case $p = 2$. This provides a rigorous convergence guarantee for the *plain* Adam update, albeit at a sub-optimal rate.

6.3 GUARANTEE WITH KNOWN DOMAIN RADIUS

In this case, we use the constrained-FTRL version of Adam over $\{\|\mathbf{z}\| \leq D\}$, equivalently clipping the online-learner output as $\mathbf{z}_t \leftarrow \mathrm{clip}_D(\mathbf{z}_t)$. This directly controls the variance term in the conversion bound and improves the overall iteration complexity.

Corollary 6.5. *Use the same notation and assumptions as in Theorem 6.4. Suppose the Adam online learner is run in constrained form over* $\{\|\mathbf{z}\| \leq D\}$, *equivalently replacing* $\mathbf{z}_t$ *by* $\mathrm{clip}_D(\mathbf{z}_t)$ *in the conversion algorithm. Then, for any* $p \in (1, 2]$, *the method guarantees a* $(\rho, \mathcal{O}(\epsilon))$*-stationary point when choosing*

$\beta_1 = 1 - \left(\frac{\epsilon}{G+\sigma}\right)^{\frac{p}{p-1}}$, β_2 *so that* $1-\beta_1 = C(\beta)(1-\beta_2)$ *for a fixed constant* $C(\beta) \geq 1$, $D = \frac{(1-\beta_1)\epsilon^{1/2}}{\rho^{1/2}}$, *and*

$$T = \mathcal{O}\Big(\max\Big\{ (G+\sigma)^{\frac{p}{p-1}+1} \epsilon^{-\left(\frac{p}{p-1}+1\right)} \log\tfrac{G+\sigma}{\epsilon}, \Delta\rho^{\frac{1}{2}} (C(R)G+\sigma)^{\frac{p}{p-1}} \epsilon^{-\left(\frac{p}{p-1}+\frac{3}{2}\right)} \Big\}\Big).$$

Moreover, in the special case $p = 2$, *the bound further reduces to* $T = \mathcal{O}\Big(\max\Big\{ (G+\sigma)^3 \epsilon^{-3} \log\tfrac{G+\sigma}{\epsilon}, \Delta\rho^{\frac{1}{2}}(G+\sigma)^2\epsilon^{-\frac{7}{2}} \Big\}\Big).$

The proof of Corollary 6.5 is deferred to Appendix C.

Remark 6.6. Corollary 6.5 matches the known optimal heavy-tail exponent for stochastic non-smooth non-convex optimization under the same conversion framework, up to constants and lower-order terms [Zhang and Cutkosky, 2024, Ahn and Cutkosky, 2024, Liu, 2025].

Moreover, under additional smoothness assumptions, these rates translate into $\mathcal{O}(\epsilon^{-4})$ complexity for standard first-order stationarity, and admit natural extensions to coordinate-wise versions of Adam targeting L_1-stationarity under standard coordinate-wise assumptions; we omit the details since the extension follows existing arguments [Ahn and Cutkosky, 2024].

7 CONCLUSION AND DISCUSSION

In this work, we established convergence guarantees for Adam under heavy-tailed noise. Our analysis proceeds by (i) extending the discounted-to-nonconvex conversion framework to heavy-tailed martingale-difference noise, and (ii) developing a discounted regret analysis for exact vector-form Adam. Together, these ingredients yield (ρ, ϵ)-stationarity guarantees under only bounded p-th moments. For plain Adam, the current analysis applies for $p > 4/3$ and gives the leading exponent $\frac{5p}{3p-4} + \frac{3}{2}$ in the target accuracy. When the domain radius is known and used to control the online-learner output, the exponent improves to the optimal heavy-tail rate $\frac{p}{p-1} + \frac{3}{2}$ for all $p \in (1, 2]$. In particular, for $p = 2$, the plain-Adam bound scales as $\mathcal{O}(\epsilon^{-13/2})$, whereas the known-radius version scales as $\mathcal{O}(\epsilon^{-7/2})$.

On the selection of β_2 and numerical stabilizer. Our regret analysis uses a mild relation between β_1 and β_2 ($\beta_1 \leq \beta_2$ in Remark 5.3), written as $\frac{1-\beta_1}{1-\beta_2} \leq C(\beta)$, to obtain deterministic control of the Adam increment. This condition allows $C(\beta)$ to be a small constant and does not require the restrictive coupling. The numerical stabilizer in Adam's denominator is treated as a fixed small constant and is not required to scale with either G or σ.

An open question. Corollary 6.5 shows that explicit norm control of the increment suffices to close the optimality gap, yielding optimal iteration complexity. This raises a natural question: *Can the plain Adam update, or a closely related variant such as AdamW, without access to D for output clipping, still attain optimal convergence rates under heavy-tailed noise?* Answering this question would clarify whether explicit norm control is fundamentally necessary, or merely an artifact of the current analytical techniques.

References

Kwangjun Ahn and Ashok Cutkosky. Adam with model exponential moving average is effective for nonconvex optimization. *arXiv preprint arXiv:2405.18199*, 2024.

Kwangjun Ahn, Xiang Cheng, Minhak Song, Chulhee Yun, Ali Jadbabaie, and Suvrit Sra. Linear attention is (maybe) all you need (to understand transformer optimization). *arXiv preprint arXiv:2310.01082*, 2023.

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*, 2024a.

Kwangjun Ahn, Zhiyu Zhang, Yunbum Kook, and Yan Dai. Understanding adam optimizer via online learning of updates: Adam is ftrl in disguise. *arXiv preprint arXiv:2402.01567*, 2024b.

Barak Battash, Lior Wolf, and Ofir Lindenbaum. Revisiting the noise model of stochastic gradient descent. In *International Conference on Artificial Intelligence and Statistics*, pages 4780–4788. PMLR, 2024.

Savelii Chezhegov, Yaroslav Klyukin, Andrei Semenov, Aleksandr Beznosikov, Alexander Gasnikov, Samuel Horváth, Martin Takáč, and Eduard Gorbunov. Clipping improves adam-norm and adagrad-norm when the noise is heavy-tailed. *arXiv preprint arXiv:2406.04443*, 2024.

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*, pages 6643–6670. PMLR, 2023.

Adrien Fradin, Abdurakhmon Sadiev, Laurent Condat, and Peter Richtárik. Tight lower bounds and optimal algorithms for stochastic nonconvex optimization with heavy-tailed noise. *arXiv preprint arXiv:2512.18713*, 2025.

Saurabh Garg, Joshua Zhanson, Emilio Parisotto, Adarsh Prasad, Zico Kolter, Zachary Lipton, Sivaraman Balakrishnan, Ruslan Salakhutdinov, and Pradeep Ravikumar. On proximal policy optimization's heavy-tailed gradients. In *International Conference on Machine Learning*, pages 3610–3619. PMLR, 2021.

Allen A Goldstein. Optimization of lipschitz continuous functions. *Mathematical Programming*, 13:14–22, 1977.

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

Chuan He, Zhaosong Lu, Defeng Sun, and Zhanwang Deng. Complexity of normalized stochastic first-order methods with momentum under heavy-tailed noise. *arXiv preprint arXiv:2506.11214*, 2025.

Wendyam Eric Lionel Ilboudo, Taisuke Kobayashi, and Takamitsu Matsubara. Adaterm: Adaptive t-distribution estimated robust moments for noise-robust stochastic gradient optimization. *Neurocomputing*, 557:126692, 2023.

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

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

Zijian Liu. Online convex optimization with heavy tails: Old algorithms, new regrets, and applications. *arXiv preprint arXiv:2508.07473*, 2025.

Zijian Liu, Jiawei Zhang, and Zhengyuan Zhou. Breaking the lower bound with (little) structure: Acceleration in non-convex stochastic optimization with heavy-tailed noise. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 2266–2290. PMLR, 2023.

Ta Duy Nguyen, Alina Ene, and Huy L Nguyen. Improved convergence in high probability of clipped gradient methods with heavy tails. *arXiv preprint arXiv:2304.01119*, 2023.

Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

Iosif Pinelis. Best possible bounds of the von bahr–esseen type. *Annals of Functional Analysis*, 6(4):1–29, 2015.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

Tao Sun, Xinwang Liu, and Kun Yuan. Gradient normalization provably benefits nonconvex sgd under heavy-tailed noise. *arXiv preprint arXiv:2410.16561*, page 5, 2024.

Lai Tian, Kaiwen Zhou, and Anthony Man-Cho So. On the finite-time complexity and practical computation of approximate stationarity concepts of lipschitz functions. In *International Conference on Machine Learning*, pages 21360–21379. PMLR, 2022.

van Erven Tim. Why ftrl is better than online mirror descent. *https://www.timvanerven.nl/blog/ftrl-vs-omd/*, 2021.

Bengt von Bahr and Carl-Gustav Esseen. Inequalities for the rth absolute moment of a sum of random variables, 1 $\leq$ r $\leq$ 2. *The Annals of Mathematical Statistics*, pages 299–303, 1965.

Dingzhi Yu, Hongyi Tao, Yuanyu Wan, Luo Luo, and Lijun Zhang. Sign-based optimizers are effective under heavy-tailed noise. *arXiv preprint arXiv:2602.07425*, 2026.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. *arXiv preprint arXiv:1905.11881*, 2019.

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

Jiujia Zhang and Ashok Cutkosky. Parameter-free regret in high probability with heavy tails. *Advances in Neural Information Processing Systems*, 35:8000–8012, 2022.

Qinzi Zhang and Ashok Cutkosky. Random scaling and momentum for non-smooth non-convex optimization. *arXiv preprint arXiv:2405.09742*, 2024.

Supplementary Material

Yijiang Pang[1]

[1]Michigan State University, East Lansing, USA

A MISSING PROOFS OF SECTION 4

Lemma 4.1. *Suppose Assumption 3.3 holds. Then for any $\beta \in (0,1)$, $n \in [T]$, and comparator $\mathbf{u}_n \in \mathcal{D}$,*

$$\mathbb{E}\sum_{t=1}^{n} \beta^{n-t}\langle \boldsymbol{\xi}_t, \mathbf{u}_n\rangle \le 2(1-\beta)^{-\frac{1}{p}}\sigma D.$$

Proof. Using $\|\mathbf{u}_n\| \le D$ and Cauchy–Schwarz, we have

$$\sum_{t=1}^{n} \beta^{n-t}\langle \boldsymbol{\xi}_t, \mathbf{u}_n\rangle = \Big\langle \sum_{t=1}^{n} \beta^{n-t}\boldsymbol{\xi}_t,\ \mathbf{u}_n \Big\rangle \le D\Big\|\sum_{t=1}^{n} \beta^{n-t}\boldsymbol{\xi}_t\Big\|.$$

Taking expectation and applying Jensen's inequality (since $x \mapsto x^p$ is convex for $p \ge 1$) yields

$$\mathbb{E}\sum_{t=1}^{n} \beta^{n-t}\langle \boldsymbol{\xi}_t, \mathbf{u}_n\rangle \le D\,\mathbb{E}\left[\Big\|\sum_{t=1}^{n} \beta^{n-t}\boldsymbol{\xi}_t\Big\|\right] \le D\left(\mathbb{E}\left[\Big\|\sum_{t=1}^{n} \beta^{n-t}\boldsymbol{\xi}_t\Big\|^p\right]\right)^{1/p}. \tag{R1}$$

We now bound the p-moment of the weighted sum using the unbiasedness condition $\mathbb{E}[\boldsymbol{\xi}_t \mid \mathcal{F}_{t-1}] = \mathbf{0}$ and $p \in (1,2]$. Define $\mathbf{d}_t \triangleq \beta^{n-t}\boldsymbol{\xi}_t$. Since β^{n-t} is deterministic, $\mathbb{E}[\mathbf{d}_t \mid \mathcal{F}_{t-1}] = \beta^{n-t}\mathbb{E}[\boldsymbol{\xi}_t \mid \mathcal{F}_{t-1}] = \mathbf{0}$, so $\{\mathbf{d}_t, \mathcal{F}_t\}$ is a martingale-difference sequence.

By a von Bahr–Esseen type inequality for martingale differences for $p \in (1,2]$ [von Bahr and Esseen, 1965, Pinelis, 2015], there exists a constant $C_p \le 2$ such that

$$\begin{aligned}\mathbb{E}\Big\|\sum_{t=1}^{n} \beta^{n-t}\boldsymbol{\xi}_t\Big\|^p = \mathbb{E}\Big\|\sum_{t=1}^{n} \mathbf{d}_t\Big\|^p &\le C_p \sum_{t=1}^{n} \mathbb{E}\|\mathbf{d}_t\|^p = C_p \sum_{t=1}^{n} \beta^{p(n-t)}\,\mathbb{E}\|\boldsymbol{\xi}_t\|^p \\ &\le C_p\,\sigma^p \sum_{k=0}^{n-1} \beta^{pk} \le \frac{C_p\,\sigma^p}{1-\beta^p} \le \frac{C_p\,\sigma^p}{1-\beta}.\end{aligned} \tag{R2}$$

Combining (R1) and (R2) gives

$$\mathbb{E}\sum_{t=1}^{n} \beta^{n-t}\langle \boldsymbol{\xi}_t, \mathbf{u}_n\rangle \le D \cdot \frac{C_p^{1/p}\sigma}{(1-\beta^p)^{1/p}} \le D \cdot \frac{C_p^{1/p}\sigma}{(1-\beta)^{1/p}} \le 2(1-\beta)^{-1/p}\sigma D,$$

where the last inequality uses $C_p \le 2$. □

Lemma A.1. *For $\beta \in (0,1)$, consider the iterates generated as per 1. Then, $\forall n \in [T]$, we have*

$$\mathbb{E}\left[\sum_{t=1}^{n} \beta^{n-t}\big(F(\mathbf{x}_t) - F(\mathbf{x}_{t-1})\big)\right] \le -D\frac{1-\beta^n}{1-\beta}\mathbb{E}\Big\|\mathbb{E}_{\mathbf{y}_n}\nabla F(\mathbf{y}_n)\Big\| + 2(1-\beta)^{-\frac{1}{p}}\sigma D + \mathbb{E}\mathrm{Regret}_n^{[\beta]}(\mathbf{u}_n),$$

where $\mathbf{y}_n$ is randomly distributed over $\{\mathbf{x}_t\}_{t=1}^n$ as $\mathbb{P}(\mathbf{y}_n = \mathbf{x}_t) = \frac{(1-\beta)\beta^{n-t}}{1-\beta^n}, t = 1, \dots, n$.

Proof. Following the fact about exponential random variable, Lemma 3.1 in Zhang and Cutkosky [2024], $\mathbb{E}\big[F(\mathbf{x}_t) - F(\mathbf{x}_{t-1})\big] = \mathbb{E}\langle \nabla F(\mathbf{x}_t), \mathbf{z}_t\rangle$:

$$\begin{aligned}
\mathbb{E}\big[F(\mathbf{x}_t) - F(\mathbf{x}_{t-1})\big] &= \mathbb{E}\langle \nabla F(\mathbf{x}_t), \mathbf{z}_t\rangle \\
&= \mathbb{E}\langle \nabla F(\mathbf{x}_t), \mathbf{u}_n\rangle + \mathbb{E}\langle \nabla F(\mathbf{x}_t) - \mathbf{g}_t, \mathbf{z}_t - \mathbf{u}_n\rangle + \mathbb{E}\langle \mathbf{g}_t, \mathbf{z}_t - \mathbf{u}_n\rangle \\
&= \underbrace{\mathbb{E}\langle \nabla F(\mathbf{x}_t), \mathbf{u}_n\rangle}_{①} + \underbrace{\mathbb{E}\langle \nabla F(\mathbf{x}_t) - \mathbf{g}_t, -\mathbf{u}_n\rangle}_{②} + \underbrace{\mathbb{E}\langle \mathbf{g}_t, \mathbf{z}_t - \mathbf{u}_n\rangle}_{③},
\end{aligned} \tag{4}$$

where we used the fact $\mathbb{E}\langle \nabla F(\mathbf{x}_t) - \mathbf{g}_t, \mathbf{z}_t\rangle = 0$ for the last equality.

We bound the three pieces.

①: Define $\mathbf{u}_n \triangleq -D\frac{\sum_{t=1}^n \beta^{n-t}\nabla F(\mathbf{x}_t)}{\|\sum_{t=1}^n \beta^{n-t}\nabla F(\mathbf{x}_t)\|}$, then

$$\begin{aligned}
\mathbb{E}\sum_{t=1}^{n} \beta^{n-t}\langle \nabla F(\mathbf{x}_t), \mathbf{u}_n\rangle &= \mathbb{E}\left\langle \sum_{t=1}^{n} \beta^{n-t}\nabla F(\mathbf{x}_t),\ -D\frac{\sum_{t=1}^n \beta^{n-t}\nabla F(\mathbf{x}_t)}{\|\sum_{t=1}^n \beta^{n-t}\nabla F(\mathbf{x}_t)\|}\right\rangle \\
&= -D\mathbb{E}\Big\|\sum_{t=1}^{n} \beta^{n-t}\nabla F(\mathbf{x}_t)\Big\| \\
&\le -D\frac{1-\beta^n}{1-\beta}\mathbb{E}\Big\|\mathbb{E}_{\mathbf{y}_n}\nabla F(\mathbf{y}_n)\Big\|
\end{aligned}$$

where the last equality follows by probability conversion: the $\mathbf{y}_n$ is randomly distributed over $\{\mathbf{x}_t\}_{t=1}^n$ as $\mathbb{P}(\mathbf{y}_n = \mathbf{x}_t) = \frac{(1-\beta)\beta^{n-t}}{1-\beta^n}, t = 1, \dots, n$.

②:

$$\mathbb{E}\sum_{t=1}^{n} \beta^{n-t}\langle \nabla F(\mathbf{x}_t) - \mathbf{g}_t, -\mathbf{u}_n\rangle = \mathbb{E}\sum_{t=1}^{n} \beta^{n-t}\langle \boldsymbol{\xi}_t, \mathbf{u}_n\rangle \le 2(1-\beta)^{-\frac{1}{p}}\sigma D,$$

where the last inequality uses the result $\mathbb{E}\sum_{s=1}^{t} \beta^{t-s}\langle \boldsymbol{\xi}_s, \mathbf{u}\rangle \le 2(1-\beta)^{-\frac{1}{p}}\sigma D$ in Lemma 4.1;

③:

$$\mathbb{E}\sum_{t=1}^{n} \beta^{n-t}\langle \mathbf{g}_t, \mathbf{z}_t - \mathbf{u}_n\rangle = \mathbb{E}\mathrm{Regret}_n^{[\beta]}(\mathbf{u}_n).$$

Combining the results in above three steps gives

$$\mathbb{E}\left[\sum_{t=1}^{n} \beta^{n-t}\big(F(\mathbf{x}_t) - F(\mathbf{x}_{t-1})\big)\right] \le -D\frac{1-\beta^n}{1-\beta}\mathbb{E}\Big\|\mathbb{E}_{\mathbf{y}_n}\nabla F(\mathbf{y}_n)\Big\| + 2(1-\beta)^{-\frac{1}{p}}\sigma D + \mathbb{E}\mathrm{Regret}_n^{[\beta]}(\mathbf{u}_n), \tag{5}$$

which completes the proof. □

Proposition 4.2. *Suppose F satisfies Assumption 3.1. For any comparator sequence $\{\mathbf{u}_t\} \subseteq \mathcal{D}$, Algorithm 1*

$$\begin{aligned}
\mathbb{E}_\tau \left\|\mathbb{E}_{\mathbf{y}_\tau}\nabla F(\mathbf{y}_\tau)\right\| &\le 2(1-\beta+\frac{\beta}{T})(1-\beta)^{-\frac{1}{p}}\sigma + \frac{\Delta}{DT} \\
&+ \frac{1}{DT}\left((1-\beta)\sum_{t=1}^{T}\mathbb{E}\mathrm{Regret}_t^{[\beta]}(\mathbf{u}_t) + \beta\mathbb{E}\mathrm{Regret}_T^{[\beta]}(\mathbf{u}_T)\right)
\end{aligned}$$

where τ is a random index in $[T]$ with distribution $\mathbb{P}(\tau = t) = \begin{cases} \frac{1-\beta^t}{T} & \text{if} \quad t = 1, \cdots, T-1, \\ \frac{1-\beta^T}{(1-\beta)T} & \text{if} \quad t = T \end{cases}$. Moreover, $\mathbf{y}_t$ is sampled from $\{\mathbf{x}_s\}_{s=1}^t$ according to $\mathbb{P}(\mathbf{y}_t = \mathbf{x}_s) = \frac{(1-\beta)\beta^{t-s}}{1-\beta^t}, s = 1, \ldots, t$.

Proof. We start from a change of summation and telescoping,

$$\sum_{n=1}^{T}\sum_{t=1}^{n}(1-\beta)\beta^{n-t}\big(F(\mathbf{x}_t) - F(\mathbf{x}_{t-1})\big) = F(\mathbf{x}_T) - F(\mathbf{x}_0) - \sum_{t=1}^{T}\beta^{T-t+1}\big(F(\mathbf{x}_t) - F(\mathbf{x}_{t-1})\big).$$

Rearranging and using $F(\mathbf{x}_0) - F(\mathbf{x}_T) \leq \Delta$ gives

$$-\Delta \leq \mathbb{E}\Bigg[\underbrace{\sum_{n=1}^{T}\sum_{t=1}^{n}(1-\beta)\beta^{n-t}\big(F(\mathbf{x}_t) - F(\mathbf{x}_{t-1})\big)}_{A}\Bigg] + \mathbb{E}\Bigg[\underbrace{\sum_{t=1}^{T}\beta^{T-t+1}\big(F(\mathbf{x}_t) - F(\mathbf{x}_{t-1})\big)}_{B}\Bigg]. \tag{6}$$

Using the result in Lemma A.1, we have that

$$\begin{aligned} -\Delta \leq (1-\beta)\sum_{t=1}^{T}\Bigg(-D\frac{1-\beta^t}{1-\beta}\mathbb{E}\Big\|\mathbb{E}_{\mathbf{y}_t}\nabla F(\mathbf{y}_t)\Big\| + 2(1-\beta)^{-\frac{1}{p}}\sigma D + \mathbb{E}\mathrm{Regret}_t^{[\beta]}(\mathbf{u}_t),\Bigg) \\ + \beta\Bigg(-D\frac{1-\beta^T}{1-\beta}\mathbb{E}\Big\|\mathbb{E}_{\mathbf{y}_T}\nabla F(\mathbf{y}_T)\Big\| + 2(1-\beta)^{-\frac{1}{p}}\sigma D + \mathbb{E}\mathrm{Regret}_T^{[\beta]}(\mathbf{u}_T),\Bigg) \end{aligned}$$

Rearranging, we obtain

$$\begin{aligned} &D\sum_{t=1}^{T}(1-\beta^t)\mathbb{E}\Big\|\mathbb{E}_{\mathbf{y}_t}\nabla F(\mathbf{y}_t)\Big\| + D\frac{\beta-\beta^{T+1}}{1-\beta}\mathbb{E}\left\|\mathbb{E}_{\mathbf{y}_T}\nabla F(\mathbf{y}_T)\right\| \\ &\quad\leq (1-\beta)\sum_{t=1}^{T}\mathbb{E}\left[\mathrm{Regret}_t^{[\beta]}(\mathbf{u}_t)\right] + \beta\mathbb{E}\left[\mathrm{Regret}_T^{[\beta]}(\mathbf{u}_T)\right] + 2((1-\beta)T+\beta)(1-\beta)^{-\frac{1}{p}}\sigma D + \Delta \end{aligned}$$

By probability conversion over left hand side: since $\sum_{t=1}^{T}(1-\beta^t) + \frac{\beta-\beta^{T+1}}{1-\beta} = T$, τ is a random index among $[T]$ such that $\mathbb{P}(\tau = t) = \begin{cases} \frac{1-\beta^t}{T} & \text{if} \quad t = 1, \cdots, T-1, \\ \frac{1-\beta^T}{(1-\beta)T} & \text{if} \quad t = T \end{cases}$. And, expectations are written as $\mathbb{E}_{\mathbf{y}_\tau}[\cdot]$ since we condition on the trajectory for the conversion analysis, and the only remaining randomness is the sampling of $\mathbf{y}_\tau$. It suffices to have

$$DT\mathbb{E}_\tau\left\|\mathbb{E}_{\mathbf{y}_\tau}\nabla F(\mathbf{y}_\tau)\right\| \leq (1-\beta)\sum_{t=1}^{T}\mathbb{E}\left[\mathrm{Regret}_t^{[\beta]}(\mathbf{u}_t)\right] + \beta\mathbb{E}\left[\mathrm{Regret}_T^{[\beta]}(\mathbf{u}_T)\right] + 2((1-\beta)T+\beta)(1-\beta)^{-\frac{1}{p}}\sigma D + \Delta$$

$$\mathbb{E}_\tau\left\|\mathbb{E}_{\mathbf{y}_\tau}\nabla F(\mathbf{y}_\tau)\right\| \leq \frac{1}{DT}\left((1-\beta)\sum_{t=1}^{T}\mathbb{E}\left[\mathrm{Regret}_t^{[\beta]}(\mathbf{u}_t)\right] + \beta\mathbb{E}\left[\mathrm{Regret}_T^{[\beta]}(\mathbf{u}_T)\right]\right) + 2(1-\beta+\frac{\beta}{T})(1-\beta)^{-\frac{1}{p}}\sigma + \frac{\Delta}{DT}.$$

This concludes the proof. □

B MISSING PROOF OF SECTION 5

Proposition 5.5. *Let $\beta_1 \leq \beta_2$, use $\eta_s = \eta$, and initialize $\alpha_1 := \alpha_2$. Then $\alpha_s = \frac{\eta(1-\beta_1)\beta_1^{s-1}}{(1-\beta_1^{s-1})(\sqrt{b_s}+\epsilon)}, s \geq 2$, is non-increasing: $\alpha_s \geq \alpha_{s+1}$ for all $s \geq 1$.*

Proof. Let $c_s := \frac{(1-\beta_1)\beta_1^{s-1}}{(1-\beta_1^{s-1})(\sqrt{b_s}+\epsilon)}, \quad s \geq 2$, and set $c_1 := c_2$. Since $\alpha_s = \eta c_s$, it suffices to prove $c_s \leq c_{s-1}$ for $s \geq 3$. The case $s = 2$ holds by initialization.

Fix $s \geq 3$ and set $k = s - 2$. By bias correction,

$$b_s = w_s b_{s-1} + (1 - w_s)\|\mathbf{v}_{s-1}\|^2, \quad w_s = \frac{\beta_2(1-\beta_2^k)}{1-\beta_2^{k+1}},$$

hence $b_s \geq w_s b_{s-1}$. Define $\phi_k(\beta) := \frac{\beta(1-\beta^k)}{1-\beta^{k+1}}$, so that $w_s = \phi_k(\beta_2)$.

Using $\frac{x+\epsilon}{\sqrt{w_s}x+\epsilon} \leq \frac{1}{\sqrt{w_s}}$ for $x \geq 0$,

$$\frac{c_s}{c_{s-1}} \leq \frac{\beta_1(1-\beta_1^k)}{1-\beta_1^{k+1}} \cdot \frac{\sqrt{b_{s-1}}+\epsilon}{\sqrt{w_s b_{s-1}}+\epsilon} \leq \frac{\phi_k(\beta_1)}{\sqrt{\phi_k(\beta_2)}}.$$

It remains to show $\phi_k(\beta_1)^2 \leq \phi_k(\beta_2)$. Since $\phi_k(\beta) = \frac{\sum_{j=1}^k \beta^j}{\sum_{j=0}^k \beta^j} = 1 - \frac{1}{\sum_{j=0}^k \beta^j}$, we have $0 < \phi_k(\beta) < 1$, and ϕ_k is increasing on $0, 1$. Thus, using $\beta_1 \leq \beta_2$, $\phi_k(\beta_1)^2 \leq \phi_k(\beta_1) \leq \phi_k(\beta_2)$.

hence $c_s/c_{s-1} \leq 1$, so $c_s \leq c_{s-1}$ for all $s \geq 2$. Therefore $\alpha_s = \eta c_s$ is non-increasing. □

Proposition 5.4. *Under Remark 5.3, Adam's output satisfies the deterministic bound* $\|\mathbf{z}_t\| \leq \eta_t\sqrt{C(\beta)}, \qquad \forall t \geq 1.$

Proof. Recall that

$$\mathbf{z}_t = -\eta_t \frac{\sum_{s=1}^{t-1} w_{1,t-1,s}\mathbf{v}_s}{\sqrt{\sum_{s=1}^{t-1} w_{2,t-1,s}\|\mathbf{v}_s\|^2}+\epsilon},$$

where

$$w_{1,t,s} := \frac{(1-\beta_1)\beta_1^{t-s}}{1-\beta_1^t}, \quad w_{2,t,s} := \frac{(1-\beta_2)\beta_2^{t-s}}{1-\beta_2^t}.$$

Then discarding the ϵ term, we obtain

$$\|\mathbf{z}_t\|^2 \leq \eta_t^2 \frac{\left\|\sum_{s=1}^{t-1} w_{1,t-1,s}\mathbf{v}_s\right\|^2}{\sum_{s=1}^{t-1} w_{2,t-1,s}\|\mathbf{v}_s\|^2} \leq \eta_t^2 \frac{\sum_{s=1}^{t-1} w_{1,t-1,s}\|\mathbf{v}_s\|^2}{\sum_{s=1}^{t-1} w_{2,t-1,s}\|\mathbf{v}_s\|^2},$$

where the inequality follows from Jensen's inequality since $\sum_s w_{1,t-1,s} = 1$.

Rewriting the numerator gives

$$\|\mathbf{z}_t\|^2 \leq \eta_t^2 \max_{1\leq s\leq t-1} \frac{w_{1,t-1,s}}{w_{2,t-1,s}}.$$

A direct computation yields

$$\max_{1\leq s\leq t-1} \frac{w_{1,t-1,s}}{w_{2,t-1,s}} = \frac{1-\beta_1}{1-\beta_2}\frac{1-\beta_2^{t-1}}{1-\beta_1^{t-1}} \max_{1\leq s\leq t-1}\left(\frac{\beta_1}{\beta_2}\right)^{t-1-s} \leq \frac{1-\beta_1}{1-\beta_2},$$

where we used $\beta_1 \leq \beta_2$. Finally, Remark 5.3 ensures $\frac{1-\beta_1}{1-\beta_2} \leq C(\beta)$.

It suffice to have $\|\mathbf{z}_t\|^2 \leq \eta_t^2 C(\beta)$. Taking square roots completes the proof. □

Proposition 5.1. *Under Assumption 3.3, for every* $t \in [T]$,

$$\mathbb{E}\sqrt{b_{t+1}} \leq G + \sigma\left(\frac{2}{p}\right)^{1/p}(1-\beta_2)^{-\left(\frac{1}{p}-\frac{1}{2}\right)}.$$

Proof. Recall that

$$b_{t+1} = \sum_{s=1}^{t} w_{s,t}\|\mathbf{v}_s\|^2, \qquad w_{s,t} = \frac{1-\beta_2}{1-\beta_2^t}\beta_2^{t-s}.$$

Since $\mathbf{v}_s = \boldsymbol{\mu}_s + \boldsymbol{\xi}_s$, we have

$$\sqrt{b_{t+1}} = \left(\sum_{s=1}^{t} w_{s,t} \|\boldsymbol{\mu}_s + \boldsymbol{\xi}_s\|^2 \right)^{1/2}.$$

By the triangle inequality in the product Hilbert space,

$$\sqrt{b_{t+1}} \le \left(\sum_{s=1}^{t} w_{s,t} \|\boldsymbol{\mu}_s\|^2 \right)^{1/2} + \left(\sum_{s=1}^{t} w_{s,t} \|\boldsymbol{\xi}_s\|^2 \right)^{1/2}.$$

Since $\|\boldsymbol{\mu}_s\| \le G$ and $\sum_s w_{s,t} = 1$, $\left(\sum_{s=1}^{t} w_{s,t} \|\boldsymbol{\mu}_s\|^2 \right)^{1/2} \le G$. Therefore,

$$\mathbb{E}\sqrt{b_{t+1}} \le G + \mathbb{E} \left(\sum_{s=1}^{t} w_{s,t} \|\boldsymbol{\xi}_s\|^2 \right)^{1/2}.$$

By Lyapunov's inequality,

$$\mathbb{E} \left(\sum_{s=1}^{t} w_{s,t} \|\boldsymbol{\xi}_s\|^2 \right)^{1/2} \le \left[\mathbb{E} \left(\sum_{s=1}^{t} w_{s,t} \|\boldsymbol{\xi}_s\|^2 \right)^{p/2} \right]^{1/p}.$$

Because $p/2 \le 1$, the map $x \mapsto x^{p/2}$ is subadditive on $\mathbb{R}_+$, and hence

$$\left(\sum_{s=1}^{t} w_{s,t} \|\boldsymbol{\xi}_s\|^2 \right)^{p/2} \le \sum_{s=1}^{t} w_{s,t}^{p/2} \|\boldsymbol{\xi}_s\|^p.$$

Taking expectations and using $\mathbb{E}\|\boldsymbol{\xi}_s\|^p \le \sigma^p$ gives

$$\mathbb{E} \left(\sum_{s=1}^{t} w_{s,t} \|\boldsymbol{\xi}_s\|^2 \right)^{p/2} \le \sigma^p \sum_{s=1}^{t} w_{s,t}^{p/2}.$$

Thus

$$\mathbb{E}\sqrt{b_{t+1}} \le G + \sigma \left(\sum_{s=1}^{t} w_{s,t}^{p/2} \right)^{1/p}.$$

It remains to compute the weight sum. We have

$$\sum_{s=1}^{t} w_{s,t}^{p/2} = \left(\frac{1-\beta_2}{1-\beta_2^t} \right)^{p/2} \sum_{s=1}^{t} \beta_2^{(t-s)p/2} = \left(\frac{1-\beta_2}{1-\beta_2^t} \right)^{p/2} \frac{1-\beta_2^{tp/2}}{1-\beta_2^{p/2}}.$$

Therefore,

$$\mathbb{E}\sqrt{b_{t+1}} \le G + \sigma \left[\left(\frac{1-\beta_2}{1-\beta_2^t} \right)^{p/2} \frac{1-\beta_2^{tp/2}}{1-\beta_2^{p/2}} \right]^{1/p}.$$

Since for $q = p/2 \in (0, 1]$ and $x \in [0, 1]$, $1 - x^q \le (1-x)^q$, we get $\frac{1-\beta_2^{tp/2}}{(1-\beta_2^t)^{p/2}} \le 1$. Hence

$$\left(\sum_{s=1}^{t} w_{s,t}^{p/2} \right)^{1/p} \le \frac{(1-\beta_2)^{1/2}}{(1-\beta_2^{p/2})^{1/p}}.$$

Finally, by concavity of $x^{p/2}$, $1 - \beta_2^{p/2} \ge \frac{p}{2}(1-\beta_2)$, so

$$\frac{(1-\beta_2)^{1/2}}{(1-\beta_2^{p/2})^{1/p}} \le \left(\frac{2}{p} \right)^{1/p} (1-\beta_2)^{\frac{1}{2}-\frac{1}{p}}.$$

The claim follows. □

C MISSING PROOFS OF SECTION 6

Proposition 6.1. *Under Assumption 3.3, for $t \geq T_0 := 2 + \frac{\log(1/(1-\beta_1))}{\log(\sqrt{\beta_2}/\beta_1)}$, it holds that*

$$\mathbb{E}\left[\sum_{s=1}^{t} \frac{\alpha_s}{2} \beta_1^{t-2s} \|\overline{\mathbf{v}}_s\|^2\right] \leq \eta \frac{2\sqrt{\beta_2}(1-\beta_1)}{\beta_1(1-\beta_2)} \left(G + 2\sigma(1-\beta_2)^{-\left(\frac{1}{p}-\frac{1}{2}\right)}\right)$$

where $\overline{\mathbf{v}}_s := \mathrm{clip}_{\sqrt{\sum_{n=1}^{s-1} \beta_2^{2s-2n}\|\mathbf{v}_n\|^2}}(\mathbf{v}_s)$.

Proof. The proof consists of three parts.

Part 1. We begin with

$$\begin{aligned}
&\frac{\alpha_s}{2}\beta_1^{-2s}\|\overline{\mathbf{v}}_s\|^2 \\
&\leq \frac{\eta}{2}\frac{1-\beta_1}{1-\beta_1^{s-1}}\frac{\sqrt{1-\beta_2^{s-1}}}{\sqrt{1-\beta_2}}\frac{1}{\sqrt{\beta_2^s}}\frac{\beta_1^{-s-1}\|\overline{\mathbf{v}}_s\|^2}{\sqrt{\sum_{n=1}^{s-1}\beta_2^{-n-1}\|\mathbf{v}_n\|^2}} \\
&\overset{(a)}{\leq} \frac{\eta}{2}\frac{1-\beta_1}{1-\beta_1^{s-1}}\frac{\sqrt{1-\beta_2^{s-1}}}{\sqrt{1-\beta_2}}\frac{\beta_1^{-s-1}}{\beta_2^{-s-1}}\frac{1}{\sqrt{\beta_2^s}}\frac{\sqrt{2}\,\beta_2^{-s-1}\|\overline{\mathbf{v}}_s\|^2}{\sqrt{\beta_2^{-s-1}\|\overline{\mathbf{v}}_s\|^2 + \sum_{n=1}^{s-1}\beta_2^{-n-1}\|\mathbf{v}_n\|^2}} \\
&\overset{(b)}{\leq} \frac{\eta}{2}\frac{1-\beta_1}{\sqrt{1-\beta_2}}\frac{\sqrt{1-\beta_2^{s-1}}}{1-\beta_1^{s-1}}\frac{\beta_1^{-s-1}}{\beta_2^{-s-1}}\frac{1}{\sqrt{\beta_2^s}} \cdot 2\sqrt{2}\Bigg(\sqrt{\beta_2^{-s-1}\|\overline{\mathbf{v}}_s\|^2 + \sum_{n=1}^{s-1}\beta_2^{-n-1}\|\mathbf{v}_n\|^2} - \sqrt{\textstyle\sum_{n=1}^{s-1}\beta_2^{-n-1}\|\mathbf{v}_n\|^2}\Bigg) \\
&\leq \frac{\eta}{2}\frac{1-\beta_1}{\sqrt{1-\beta_2}}\frac{\sqrt{1-\beta_2^{s-1}}}{1-\beta_1^{s-1}}\frac{\beta_1^{-s-1}}{\beta_2^{-s-1}}\frac{1}{\sqrt{\beta_2^s}} \cdot 2\sqrt{2}\underbrace{\left(\sqrt{\textstyle\sum_{n=1}^{s}\beta_2^{-n-1}\|\mathbf{v}_n\|^2} - \sqrt{\textstyle\sum_{n=1}^{s-1}\beta_2^{-n-1}\|\mathbf{v}_n\|^2}\right)}_{=Q_s - Q_{s-1}} \\
&= \underbrace{\sqrt{2}\eta\frac{(1-\beta_1)\sqrt{1-\beta_2^{s-1}}}{(1-\beta_1^{s-1})\sqrt{1-\beta_2}}\frac{\beta_2^{s/2+1}}{\beta_1^{s+1}}}_{=P_s}(Q_s - Q_{s-1})
\end{aligned}$$

where (a) uses $\beta_2^{-s-1}\|\overline{\mathbf{v}}_s\|^2 \leq \sum_{n=1}^{s-1}\beta_2^{s-2n-1}\|\mathbf{v}_n\|^2 \leq \sum_{n=1}^{s-1}\beta_2^{-n-1}\|\mathbf{v}_n\|^2$, which follows from the definition of $\overline{\mathbf{v}}_s$; (b) uses $\frac{x}{\sqrt{x+y}} \leq 2(\sqrt{x+y} - \sqrt{y})$ for $x, y > 0$.

From the result above, we have

$$\sum_{s=1}^{t}\frac{\alpha_s}{2}\beta_1^{t-2s}\|\overline{\mathbf{v}}_s\|^2 \leq \beta_1^t\left(\sum_{s=1}^{t}P_s(Q_s - Q_{s-1})\right) = \beta_1^t\left(P_tQ_t + \sum_{s=1}^{t-1}(P_s - P_{s+1})Q_s\right).$$

Part 2. Next, we show that the remainder $R(t) := \sum_{s=1}^{t-1}(P_s - P_{s+1})Q_s$ is non-positive for $t \geq T_0 := 2 + \frac{\log(1/(1-\beta_1))}{\log(\sqrt{\beta_2}/\beta_1)}$.

Part 2.1. The sequence $\{P_s\}_{s\geq 2}$ decreases and then increases. Extend $h(s) := \log P_s$ to real $s \geq 2$. Let $\lambda_i := -\log\beta_i$ and $\tau := \lambda_1/\lambda_2 \geq 1$. Then

$$h'(s) = \frac{\lambda_2}{2}\left[(2\tau - 1) + \frac{1}{e^y - 1} - \frac{2\tau}{e^{\tau y} - 1}\right], \qquad y := \lambda_2(s-1).$$

After multiplying by the positive quantity $2(e^y - 1)(e^{\tau y} - 1)/\lambda_2$, the sign of $h'(s)$ is the sign of

$$G(y) := (2\tau - 1)(e^y - 1)(e^{\tau y} - 1) + (e^{\tau y} - 1) - 2\tau(e^y - 1).$$

To verify the claimed sign pattern, write $x = e^y \geq 1$. A direct calculation gives

$$G'(y) = x\Phi(x),$$

where

$$\Phi(x) = (2\tau - 1)(\tau + 1)x^\tau - 2\tau(\tau - 1)x^{\tau-1} - (4\tau - 1).$$

Moreover,

$$\Phi(1) = -\tau < 0, \qquad \Phi(x) \to \infty,$$

and

$$\Phi'(x) = \tau x^{\tau-2}\left((2\tau - 1)(\tau + 1)x - 2(\tau - 1)^2\right) > 0 \quad (x \geq 1),$$

because

$$(2\tau - 1)(\tau + 1) - 2(\tau - 1)^2 = 5\tau - 3 > 0.$$

Thus G' has exactly one positive zero. Since $G(0) = 0$, $G'(0) < 0$, and $G(y) \to \infty$, it follows that G has exactly one positive zero. Hence h' changes sign once, from negative to positive.

Therefore h first decreases and then increases on $[2, \infty)$. Consequently, there exists an integer $s_\star \geq 2$ such that $P_{s+1} \geq P_s$, we have

$$P_{s+1} < P_s \quad (2 \leq s < s_\star), \qquad P_{s+1} \geq P_s \quad (s \geq s_\star).$$

Part 2.2. Write

$$\rho := \frac{\sqrt{\beta_2}}{\beta_1} > 1, \qquad C_L := \frac{\sqrt{2}\eta(1 - \beta_1)\beta_2}{\beta_1}.$$

From the definition of P_s,

$$P_s \geq C_L\rho^s, \qquad P_2 = \frac{C_L\rho^2}{1 - \beta_1}.$$

Hence $t \geq T_0$ implies $P_t \geq P_2$. By the monotonicity pattern established above, this gives $P_j \leq P_t$ for all $2 \leq j \leq t$. Indeed, terms before the turning point are at most P_2, while terms after the turning point are nondecreasing and hence at most P_t. Therefore,

$$\begin{aligned}\sum_{s=1}^{t-1}(P_s - P_{s+1})Q_s &= -P_tQ_{t-1} + \sum_{s=1}^{t-2} P_{s+1}(Q_{s+1} - Q_s) \\ &\leq -P_tQ_{t-1} + P_t\sum_{s=1}^{t-2}(Q_{s+1} - Q_s) \\ &= -P_tQ_1 \leq 0.\end{aligned}$$

Thus, for $t \geq T_0$,

$$\sum_{s=1}^{t} \frac{\alpha_s}{2}\beta_1^{t-2s}\|\overline{\mathbf{v}}_s\|^2 \leq \beta_1^t P_tQ_t.$$

Part 3. It remains to bound the expectation of the terminal term. Expanding P_tQ_t,

$$\beta_1^t P_tQ_t = \frac{\sqrt{2\beta_2}\eta(1 - \beta_1)}{\beta_1\sqrt{1 - \beta_2}} \frac{\sqrt{1 - \beta_2^{t-1}}}{1 - \beta_1^{t-1}} \sqrt{\sum_{n=1}^{t} \beta_2^{t-n}\|\mathbf{v}_n\|^2}.$$

By Corollary 5.2,

$$\mathbb{E}\left[\sqrt{\sum_{n=1}^{t} \beta_2^{t-n}\|\mathbf{v}_n\|^2}\right] \leq \sqrt{\frac{1 - \beta_2^t}{1 - \beta_2}}\left(G + 2\sigma(1 - \beta_2)^{-\left(\frac{1}{p} - \frac{1}{2}\right)}\right).$$

Therefore,

$$\mathbb{E}\left[\sum_{s=1}^{t}\frac{\alpha_s}{2}\beta_1^{t-2s}\|\overline{\mathbf{v}}_s\|^2\right] \le \mathbb{E}[\beta_1^t P_t Q_t] \le \frac{\sqrt{2\beta_2}\eta(1-\beta_1)}{\beta_1(1-\beta_2)}\frac{\sqrt{(1-\beta_2^{t-1})(1-\beta_2^t)}}{1-\beta_1^{t-1}}\left(G+2\sigma(1-\beta_2)^{-\left(\frac{1}{p}-\frac{1}{2}\right)}\right)$$
$$\le \eta\frac{2\sqrt{\beta_2}(1-\beta_1)}{\beta_1(1-\beta_2)}\left(G+2\sigma(1-\beta_2)^{-\left(\frac{1}{p}-\frac{1}{2}\right)}\right).$$

where the last inequality uses (1) $1-\beta_2^t \le 2(1-\beta_2^{t-1})$ for $t \ge 2$; (2) $\frac{1-\beta_2^{t-1}}{1-\beta_1^{t-1}} \le 1$ for $\beta_1 \le \beta_2$.

This concludes the proof. □

Theorem 6.4. *Assume F satisfies Assumption 3.1 and fix any $\rho > 0$. Run Algorithm 1 with online learner Adam as in Algorithm 2. Then, for any $p \in (\frac{4}{3}, 2]$, there exists a random index $\tau \in [T]$ such that*

$$\mathbb{E}_\tau\left[\|\nabla F(\widetilde{\mathbf{x}}_\tau)\|^{[\rho]}\right] \le \left(8 + C(R) + 2C(R)C(\beta)^{\frac{1}{p}-\frac{1}{2}}\right)\epsilon,$$

where τ is a random index among $[T]$ such that $\mathbb{P}(\tau = t) = \begin{cases}\frac{1-\beta_1^t}{T} & \text{if} \quad t = 1, \cdots, T-1,\\ \frac{1-\beta_1^T}{(1-\beta_1)T} & \text{if} \quad t = T\end{cases}$. Moreover, the $\mathbf{y}_t$ is randomly distributed over $\{\mathbf{x}_s\}_{s=1}^t$ as $\mathbb{P}(\mathbf{y}_t = \mathbf{x}_s) = \frac{(1-\beta_1)\beta_1^{t-s}}{1-\beta_1^t}, s = 1, \ldots, t$.

In particular, it suffices to choose:

$\beta_1 = 1 - \left(\frac{\epsilon}{G+\sigma}\right)^{\frac{4p}{3p-4}}$, β_2 so that $1-\beta_1 = C(\beta)(1-\beta_2)$ for a fixed constant $C(\beta) \ge 1$, $D = \frac{(1-\beta_1)^{5/4}\epsilon^{1/2}}{C(\beta)\rho^{1/2}}$, and

$$T = \mathcal{O}\Big(\max\Big\{(C(R)G+\sigma)^{\frac{8p-4}{3p-4}}\epsilon^{-\frac{8p-4}{3p-4}}\log\tfrac{G+\sigma}{\epsilon},$$
$$\Delta\rho^{\frac{1}{2}}(C(R)G+\sigma)^{\frac{5p}{3p-4}}\epsilon^{-\left(\frac{5p}{3p-4}+\frac{3}{2}\right)}\Big\}\Big).$$

Moreover, in the special case $p = 2$, the bound further reduces to $T = \mathcal{O}\Big(\max\Big\{(C(R)G+\sigma)^6\epsilon^{-6}\log\frac{G+\sigma}{\epsilon}$,
$\Delta\rho^{\frac{1}{2}}(C(R)G+\sigma)^5\epsilon^{-\frac{13}{2}}\Big\}\Big)$.

Proof. Start from the conversion bound equation 1 (with $\beta = \beta_1$):

$$\mathbb{E}_\tau\|\nabla F(\widetilde{\mathbf{x}}_\tau)\|^{[\rho]} \le \frac{1}{DT}\left(\beta_1\mathbb{E}\left[\mathrm{Regret}_T^{[\beta]}(\mathbf{u}_T)\right] + (1-\beta_1)\sum_{t=1}^{T}\mathbb{E}\left[\mathrm{Regret}_t^{[\beta]}(\mathbf{u}_t)\right]\right)$$
$$+ 2\left(1-\beta_1+\frac{\beta_1}{T}\right)(1-\beta_1)^{-\frac{1}{p}}\sigma + \frac{\Delta}{DT} + \frac{2\rho\beta_1\sup_{t\in[T]}\mathbb{E}\left[\|\mathbf{z}_t\|^2\right]}{(1-\beta_1)^2}.$$

Let

$$T_0 := 2 + \frac{\log(1/(1-\beta_1))}{\log(\sqrt{\beta_2}/\beta_1)},$$

where T_0 is rounded up if necessary. We assume $\beta_1 < \sqrt{\beta_2}$, so that the denominator is positive.

For the first T_0 rounds, we use a crude bound inferred directly from the regret definition. Since

$$\|\mathbf{z}_s\| \le \frac{\sqrt{C(\beta)}D}{(1-\beta_1)^{1/4}}, \qquad \|\mathbf{u}_t\| \le D,$$

we have, for any $t < T_0$,

$$\mathbb{E}\left[\mathrm{Regret}_t^{[\beta]}(\mathbf{u}_t)\right] = \mathbb{E}\left[\sum_{s=1}^{t}\beta_1^{t-s}\langle\mathbf{v}_s, \mathbf{z}_s - \mathbf{u}_t\rangle\right]$$
$$\le D\left(1+\sqrt{C(\beta)}(1-\beta_1)^{-1/4}\right)\mathbb{E}\left[\sum_{s=1}^{t}\beta_1^{t-s}\|\mathbf{v}_s\|\right].$$

By Assumption 3.3,

$$\mathbb{E}\|\mathbf{v}_s\| \le \|\boldsymbol{\mu}_s\| + \mathbb{E}\|\boldsymbol{\xi}_s\| \le G + (\mathbb{E}\|\boldsymbol{\xi}_s\|^p)^{1/p} \le G + \sigma.$$

Therefore, summing the first T_0 rounds inside the conversion bound gives

$$\begin{aligned}
&\frac{1-\beta_1}{DT}\sum_{t<T_0}\mathbb{E}\left[\mathrm{Regret}_t^{[\beta]}(\mathbf{u}_t)\right] \\
&\quad\le \frac{1-\beta_1}{T}\left(1+\sqrt{C(\beta)}(1-\beta_1)^{-1/4}\right)\sum_{t<T_0}\sum_{s=1}^{t}\beta_1^{t-s}\mathbb{E}\|\mathbf{v}_s\| \\
&\quad= \frac{1}{T}\left(1+\sqrt{C(\beta)}(1-\beta_1)^{-1/4}\right)\sum_{s<T_0}\left((1-\beta_1)\sum_{t=s}^{T_0-1}\beta_1^{t-s}\right)\mathbb{E}\|\mathbf{v}_s\| \\
&\quad\le \frac{T_0}{T}\left(1+\sqrt{C(\beta)}(1-\beta_1)^{-1/4}\right)(G+\sigma).
\end{aligned}$$

For $t \ge T_0$, Theorem 6.2 gives

$$\mathbb{E}\left[\mathrm{Regret}_t^{[\beta]}(\mathbf{u}_t)\right] \le \frac{DC(R)\left(G+2\sigma(1-\beta_2)^{-\left(\frac{1}{p}-\frac{1}{2}\right)}\right)}{(1-\beta_1)^{3/4}}.$$

Thus, for $T \ge T_0$,

$$\begin{aligned}
&\frac{1}{DT}\left(\beta_1\mathbb{E}\left[\mathrm{Regret}_T^{[\beta]}(\mathbf{u}_T)\right] + (1-\beta_1)\sum_{t=1}^{T}\mathbb{E}\left[\mathrm{Regret}_t^{[\beta]}(\mathbf{u}_t)\right]\right) \\
&\quad\le \left(1-\beta_1+\frac{\beta_1}{T}\right)\frac{C(R)\left(G+2\sigma(1-\beta_2)^{-\left(\frac{1}{p}-\frac{1}{2}\right)}\right)}{(1-\beta_1)^{3/4}} \\
&\qquad+\frac{T_0}{T}\left(1+\sqrt{C(\beta)}(1-\beta_1)^{-1/4}\right)(G+\sigma).
\end{aligned}$$

Substituting the increment bound from Proposition 5.4 and $\eta = \frac{D}{(1-\beta_1)^{1/4}}$,

$$\|\mathbf{z}_t\| \le \sqrt{C(\beta)}\eta = \frac{\sqrt{C(\beta)}D}{(1-\beta_1)^{1/4}},$$

and choosing $D = \frac{(1-\beta_1)^{5/4}\epsilon^{1/2}}{C(\beta)\rho^{1/2}}$, we obtain

$$\begin{aligned}
\mathbb{E}_\tau\|\nabla F(\widetilde{\mathbf{x}}_\tau)\|^{[\rho]} \le{}& \left(1-\beta_1+\frac{\beta_1}{T}\right)\frac{C(R)\left(G+2\sigma(1-\beta_2)^{-\left(\frac{1}{p}-\frac{1}{2}\right)}\right)}{(1-\beta_1)^{3/4}} \\
&+\frac{T_0}{T}\left(1+\sqrt{C(\beta)}(1-\beta_1)^{-1/4}\right)(G+\sigma) \\
&+2\left(1-\beta_1+\frac{\beta_1}{T}\right)(1-\beta_1)^{-1/p}\sigma + \frac{\sqrt{C(\beta)}\rho^{1/2}\Delta}{(1-\beta_1)^{5/4}\epsilon^{1/2}T} + 2\epsilon.
\end{aligned}$$

Expanding the first and third terms gives

$$\begin{aligned}
\mathbb{E}_\tau\|\nabla F(\widetilde{\mathbf{x}}_\tau)\|^{[\rho]} \le{}& C(R)\left(G+2\sigma(1-\beta_2)^{-\left(\frac{1}{p}-\frac{1}{2}\right)}\right)(1-\beta_1)^{1/4} + 2\sigma(1-\beta_1)^{1-\frac{1}{p}} \\
&+\frac{C(R)\beta_1\left(G+2\sigma(1-\beta_2)^{-\left(\frac{1}{p}-\frac{1}{2}\right)}\right)}{(1-\beta_1)^{3/4}T} + \frac{2\beta_1\sigma}{(1-\beta_1)^{1/p}T} \\
&+\frac{\sqrt{C(\beta)}\rho^{1/2}\Delta}{(1-\beta_1)^{5/4}\epsilon^{1/2}T} + \frac{T_0}{T}\left(1+\sqrt{C(\beta)}(1-\beta_1)^{-1/4}\right)(G+\sigma) + 2\epsilon.
\end{aligned}$$

Hence, using $1-\beta_1 \le C(\beta)(1-\beta_2)$, we have

$$\begin{aligned}\mathbb{E}_\tau\|\nabla F(\widetilde{\mathbf{x}}_\tau)\|^{[\rho]} \le\ & C(R)G(1-\beta_1)^{1/4} + 2C(R)C(\beta)^{\frac{1}{p}-\frac{1}{2}}\sigma(1-\beta_1)^{\frac{3}{4}-\frac{1}{p}} + 2\sigma(1-\beta_1)^{1-\frac{1}{p}} \\ &+ \frac{C(R)\beta_1 G}{(1-\beta_1)^{3/4}T} + \frac{2C(R)C(\beta)^{\frac{1}{p}-\frac{1}{2}}\beta_1\sigma}{(1-\beta_1)^{\frac{1}{4}+\frac{1}{p}}T} + \frac{2\beta_1\sigma}{(1-\beta_1)^{1/p}T} \\ &+ \frac{\sqrt{C(\beta)}\rho^{1/2}\Delta}{(1-\beta_1)^{5/4}\epsilon^{1/2}T} + \frac{T_0}{T}\left(1+\sqrt{C(\beta)}(1-\beta_1)^{-1/4}\right)(G+\sigma) + 2\epsilon.\end{aligned}$$

Now assume $p > \frac{4}{3}$, $0 < \epsilon \le C(R)G + \sigma$, and $\beta_1 \le \beta_2 < \sqrt{\beta_2}$. Choose

$$\beta_1 = 1 - \left(\frac{\epsilon}{G+\sigma}\right)^{\frac{4p}{3p-4}}.$$

Then $(1-\beta_1)^{1/4} \le \frac{\epsilon}{G+\sigma}$, $(1-\beta_1)^{\frac{3}{4}-\frac{1}{p}} = \frac{\epsilon}{G+\sigma}$, and $(1-\beta_1)^{1-\frac{1}{p}} \le \frac{\epsilon}{G+\sigma}$. Thus, the first three non-T^{-1} terms are bounded by $C(R)\epsilon + 2C(R)C(\beta)^{\frac{1}{p}-\frac{1}{2}}\epsilon + 2\epsilon$. Therefore,

$$\begin{aligned}\mathbb{E}_\tau\|\nabla F(\widetilde{\mathbf{x}}_\tau)\|^{[\rho]} \le\ & \frac{C(R)\beta_1 G\,(C(R)G+\sigma)^{\frac{3p}{3p-4}}}{\epsilon^{\frac{3p}{3p-4}}T} + \frac{2C(R)C(\beta)^{\frac{1}{p}-\frac{1}{2}}\beta_1\sigma\,(C(R)G+\sigma)^{\frac{p+4}{3p-4}}}{\epsilon^{\frac{p+4}{3p-4}}T} \\ &+ \frac{2\beta_1\sigma\,(C(R)G+\sigma)^{\frac{4}{3p-4}}}{\epsilon^{\frac{4}{3p-4}}T} + \frac{\sqrt{C(\beta)}\rho^{1/2}\Delta\,(C(R)G+\sigma)^{\frac{5p}{3p-4}}}{\epsilon^{\frac{5p}{3p-4}+\frac{1}{2}}T} \\ &+ \frac{(G+\sigma)T_0}{T}\cdot\left(1+\sqrt{C(\beta)}\,(C(R)G+\sigma)^{\frac{p}{3p-4}}\,\epsilon^{-\frac{p}{3p-4}}\right) \\ &+ C(R)\epsilon + 2C(R)C(\beta)^{\frac{1}{p}-\frac{1}{2}}\epsilon + 4\epsilon.\end{aligned}$$

It is sufficient to choose

$$\begin{aligned}T \ge \max\Bigg\{ & 2\left(1+C(R)C(\beta)^{\frac{1}{p}-\frac{1}{2}}\right)(C(R)G+\sigma)^{\frac{4p}{3p-4}}\,\epsilon^{-\frac{4p}{3p-4}}, \\ & \sqrt{C(\beta)}\rho^{1/2}\Delta\,(C(R)G+\sigma)^{\frac{5p}{3p-4}}\,\epsilon^{-\left(\frac{5p}{3p-4}+\frac{3}{2}\right)}, \\ & (G+\sigma)\epsilon^{-1}T_0 + \sqrt{C(\beta)}\,(C(R)G+\sigma)^{\frac{4p-4}{3p-4}}\,\epsilon^{-\frac{4p-4}{3p-4}}T_0\Bigg\},\end{aligned}$$

where the last term can be further reformulated as $\mathcal{O}\left((C(R)G+\sigma)^{\frac{8p-4}{3p-4}}\,\epsilon^{-\frac{8p-4}{3p-4}}\log\frac{G+\sigma}{\epsilon}\right)$ due to $T_0 := 2+\frac{\log(1/(1-\beta_1))}{\log(\sqrt{\beta_2}/\beta_1)} = \mathcal{O}\left(\left(\frac{G+\sigma}{\epsilon}\right)^{\frac{4p}{3p-4}}\log\frac{G+\sigma}{\epsilon}\right)$.

The third candidate ensures both $T \ge T_0$ and that the burn-in term is at most ϵ. The first candidate controls the first three T^{-1} terms by at most 2ϵ, and the second candidate controls the Δ-term by at most ϵ. Consequently,

$$\mathbb{E}_\tau\|\nabla F(\widetilde{\mathbf{x}}_\tau)\|^{[\rho]} \le \left(8 + C(R) + 2C(R)C(\beta)^{\frac{1}{p}-\frac{1}{2}}\right)\epsilon.$$

Meanwhile, we observe that the second term in the selection of T dominates under fixed constants and nondegenerate Δ, ρ asymptotically. This proves the desired stationarity guarantee for $p > \frac{4}{3}$. □

Corollary 6.5. *Use the same notation and assumptions as in Theorem 6.4. Suppose the Adam online learner is run in constrained form over $\{\|\mathbf{z}\| \le D\}$, equivalently replacing $\mathbf{z}_t$ by $\mathrm{clip}_D(\mathbf{z}_t)$ in the conversion algorithm. Then, for any $p \in (1,2]$, the method guarantees a $(\rho, \mathcal{O}(\epsilon))$-stationary point when choosing*

$\beta_1 = 1 - \left(\frac{\epsilon}{G+\sigma}\right)^{\frac{p}{p-1}}$, *$\beta_2$ so that $1-\beta_1 = C(\beta)(1-\beta_2)$ for a fixed constant $C(\beta) \ge 1$, $D = \frac{(1-\beta_1)\epsilon^{1/2}}{\rho^{1/2}}$, and*

$T = \mathcal{O}\Big(\max\Big\{(G+\sigma)^{\frac{p}{p-1}+1}\,\epsilon^{-\left(\frac{p}{p-1}+1\right)}\log\frac{G+\sigma}{\epsilon},$
$\Delta\rho^{\frac{1}{2}}\,(C(R)G+\sigma)^{\frac{p}{p-1}}\,\epsilon^{-\left(\frac{p}{p-1}+\frac{3}{2}\right)}\Big\}\Big).$

Moreover, in the special case $p = 2$, the bound further reduces to $T = \mathcal{O}\Big(\max\Big\{ (G+\sigma)^3 \epsilon^{-3} \log \frac{G+\sigma}{\epsilon}, \Delta\rho^{\frac{1}{2}}(G+\sigma)^2\epsilon^{-\frac{7}{2}} \Big\}\Big)$.

Proof. Start from the unsimplified conversion bound equation 1 (with $\beta = \beta_1$):

$$\mathbb{E}_\tau \|\nabla F(\widetilde{\mathbf{x}}_\tau)\|^{[\rho]} \le \frac{1}{DT}\left(\beta_1 \mathbb{E}\left[\mathrm{Regret}_T^{[\beta]}(\mathbf{u}_T)\right] + (1-\beta_1)\sum_{t=1}^{T}\mathbb{E}\left[\mathrm{Regret}_t^{[\beta]}(\mathbf{u}_t)\right]\right) + 2\left(1-\beta_1+\frac{\beta_1}{T}\right)(1-\beta_1)^{-\frac{1}{p}}\sigma + \frac{\Delta}{DT} + \frac{2\rho\beta_1 \sup_{t\in[T]}\mathbb{E}\left[\|\mathbf{z}_t\|^2\right]}{(1-\beta_1)^2}.$$

Recall $T_0 := 2 + \frac{\log(1/(1-\beta_1))}{\log(\sqrt{\beta_2}/\beta_1)}$, where T_0 is rounded up if necessary.

For the first T_0 rounds, we use a crude bound inferred directly from the regret definition. Since

$$\|\mathbf{z}_s\| \le D, \qquad \|\mathbf{u}_t\| \le D,$$

we have, for any $t < T_0$,

$$\mathbb{E}\left[\mathrm{Regret}_t^{[\beta]}(\mathbf{u}_t)\right] = \mathbb{E}\left[\sum_{s=1}^{t}\beta_1^{t-s}\langle \mathbf{v}_s, \mathbf{z}_s - \mathbf{u}_t\rangle\right] \le 2D\mathbb{E}\left[\sum_{s=1}^{t}\beta_1^{t-s}\|\mathbf{v}_s\|\right].$$

By Assumption 3.3,

$$\mathbb{E}\|\mathbf{v}_s\| \le \|\boldsymbol{\mu}_s\| + \mathbb{E}\|\boldsymbol{\xi}_s\| \le G + (\mathbb{E}\|\boldsymbol{\xi}_s\|^p)^{1/p} \le G + \sigma.$$

Therefore, summing the first T_0 rounds inside the conversion bound gives

$$\begin{aligned}\frac{1-\beta_1}{DT}\sum_{t<T_0}\mathbb{E}\left[\mathrm{Regret}_t^{[\beta]}(\mathbf{u}_t)\right] &\le \frac{2(1-\beta_1)}{T}\sum_{t<T_0}\sum_{s=1}^{t}\beta_1^{t-s}\mathbb{E}\|\mathbf{v}_s\| \\ &= \frac{2}{T}\sum_{s<T_0}\left((1-\beta_1)\sum_{t=s}^{T_0-1}\beta_1^{t-s}\right)\mathbb{E}\|\mathbf{v}_s\| \\ &\le \frac{2T_0}{T}(G+\sigma).\end{aligned}$$

For $t \ge T_0$, Corollary 6.3 gives

$$\mathbb{E}\left[\mathrm{Regret}_t^{[\beta]}(\mathbf{u}_t)\right] \le \frac{DC(R)\left(G + 2\sigma(1-\beta_2)^{-\left(\frac{1}{p}-\frac{1}{2}\right)}\right)}{(1-\beta_1)^{1/2}}.$$

Thus, for $T \ge T_0$,

$$\begin{aligned}&\frac{1}{DT}\left(\beta_1\mathbb{E}\left[\mathrm{Regret}_T^{[\beta]}(\mathbf{u}_T)\right] + (1-\beta_1)\sum_{t=1}^{T}\mathbb{E}\left[\mathrm{Regret}_t^{[\beta]}(\mathbf{u}_t)\right]\right) \\ &\quad\le \left(1-\beta_1+\frac{\beta_1}{T}\right)\frac{C(R)\left(G + 2\sigma(1-\beta_2)^{-\left(\frac{1}{p}-\frac{1}{2}\right)}\right)}{(1-\beta_1)^{1/2}} + \frac{2T_0}{T}(G+\sigma).\end{aligned}$$

Substituting the clipped increment bound $\|\mathbf{z}_t\| \le D$, and choosing $D = \frac{(1-\beta_1)\epsilon^{1/2}}{\rho^{1/2}}$, we obtain

$$\begin{aligned}\mathbb{E}_\tau\|\nabla F(\widetilde{\mathbf{x}}_\tau)\|^{[\rho]} \le{}& \left(1-\beta_1+\frac{\beta_1}{T}\right)\frac{C(R)\left(G + 2\sigma(1-\beta_2)^{-\left(\frac{1}{p}-\frac{1}{2}\right)}\right)}{(1-\beta_1)^{1/2}} \\ &+ \frac{2T_0}{T}(G+\sigma) + 2\left(1-\beta_1+\frac{\beta_1}{T}\right)(1-\beta_1)^{-1/p}\sigma \\ &+ \frac{\rho^{1/2}\Delta}{(1-\beta_1)\epsilon^{1/2}T} + 2\epsilon.\end{aligned}$$

Expanding the first and third terms gives

$$\begin{aligned}\mathbb{E}_\tau\|\nabla F(\widetilde{\mathbf{x}}_\tau)\|^{[\rho]} \le C(R)\left(G+2\sigma(1-\beta_2)^{-\left(\frac{1}{p}-\frac{1}{2}\right)}\right)(1-\beta_1)^{1/2} + 2\sigma(1-\beta_1)^{1-\frac{1}{p}} \\ + \frac{C(R)\beta_1\left(G+2\sigma(1-\beta_2)^{-\left(\frac{1}{p}-\frac{1}{2}\right)}\right)}{(1-\beta_1)^{1/2}T} + \frac{2\beta_1\sigma}{(1-\beta_1)^{1/p}T} \\ + \frac{\rho^{1/2}\Delta}{(1-\beta_1)\epsilon^{1/2}T} + \frac{2T_0}{T}(G+\sigma) + 2\epsilon.\end{aligned}$$

Using $1-\beta_1 \le C(\beta)(1-\beta_2)$, we obtain

$$\begin{aligned}\mathbb{E}_\tau\|\nabla F(\widetilde{\mathbf{x}}_\tau)\|^{[\rho]} \le C(R)G(1-\beta_1)^{1/2} + 2C(R)C(\beta)^{\frac{1}{p}-\frac{1}{2}}\sigma(1-\beta_1)^{1-\frac{1}{p}} + 2\sigma(1-\beta_1)^{1-\frac{1}{p}} \\ + \frac{C(R)\beta_1 G}{(1-\beta_1)^{1/2}T} + \frac{2\left(1+C(R)C(\beta)^{\frac{1}{p}-\frac{1}{2}}\right)\beta_1\sigma}{(1-\beta_1)^{1/p}T} \\ + \frac{\rho^{1/2}\Delta}{(1-\beta_1)\epsilon^{1/2}T} + \frac{2T_0}{T}(G+\sigma) + 2\epsilon.\end{aligned}$$

Now assume $0 < \epsilon \le G+\sigma$, and choose $\beta_1 = 1 - \left(\frac{\epsilon}{G+\sigma}\right)^{\frac{p}{p-1}}$. Then $(1-\beta_1)^{1-\frac{1}{p}} = \frac{\epsilon}{G+\sigma}$ and $(1-\beta_1)^{1/2} \le \frac{\epsilon}{G+\sigma}$. Thus, the first three non-T^{-1} terms are bounded by $C(R)\epsilon + 2C(R)C(\beta)^{\frac{1}{p}-\frac{1}{2}}\epsilon + 2\epsilon$. Therefore,

$$\begin{aligned}\mathbb{E}_\tau\|\nabla F(\widetilde{\mathbf{x}}_\tau)\|^{[\rho]} \le \frac{C(R)\beta_1 G\left(C(R)G+\sigma\right)^{\frac{p}{2p-2}}}{\epsilon^{\frac{p}{2p-2}}T} + \frac{2\left(1+C(R)C(\beta)^{\frac{1}{p}-\frac{1}{2}}\right)\beta_1\sigma\left(C(R)G+\sigma\right)^{\frac{1}{p-1}}}{\epsilon^{\frac{1}{p-1}}T} \\ + \frac{\rho^{1/2}\Delta\left(C(R)G+\sigma\right)^{\frac{p}{p-1}}}{\epsilon^{\frac{p}{p-1}+\frac{1}{2}}T} + \frac{2(G+\sigma)T_0}{T} + 2C(R)C(\beta)^{\frac{1}{p}-\frac{1}{2}}\epsilon + 4\epsilon.\end{aligned}$$

It is sufficient to choose

$$\begin{aligned}T \ge \max\Bigg\{2\left(1+C(R)C(\beta)^{\frac{1}{p}-\frac{1}{2}}\right)\left(C(R)G+\sigma\right)^{\frac{p}{p-1}}\epsilon^{-\frac{p}{p-1}}, \\ \rho^{1/2}\Delta\left(C(R)G+\sigma\right)^{\frac{p}{p-1}}\epsilon^{-\left(\frac{p}{p-1}+\frac{3}{2}\right)}, \\ \left(1+2(G+\sigma)\epsilon^{-1}\right)T_0\Bigg\},\end{aligned}$$

where the last term can be further reformulated as $\mathcal{O}\left((G+\sigma)^{\frac{p}{p-1}+1}\epsilon^{-\left(\frac{p}{p-1}+1\right)}\log\frac{G+\sigma}{\epsilon}\right)$ due to $T_0 := 2 + \frac{\log(1/(1-\beta_1))}{\log(\sqrt{\beta_2}/\beta_1)} = \mathcal{O}\left(\left(\frac{G+\sigma}{\epsilon}\right)^{\frac{p}{p-1}}\log\frac{G+\sigma}{\epsilon}\right)$ when choosing $\beta_1 = 1 - \left(\frac{\epsilon}{G+\sigma}\right)^{\frac{p}{p-1}}$.

The third candidate ensures both $T \ge T_0$ and that the burn-in term is at most ϵ. The first candidate controls the first two T^{-1} terms by at most 2ϵ, and the second candidate controls the Δ-term by at most ϵ. Consequently,

$$\mathbb{E}_\tau\|\nabla F(\widetilde{\mathbf{x}}_\tau)\|^{[\rho]} \le \left(8 + C(R) + 2C(R)C(\beta)^{\frac{1}{p}-\frac{1}{2}}\right)\epsilon.$$

Meanwhile, we observe that the second term in the selection of T dominates w.r.t. ϵ under fixed constants and nondegenerate Δ, ρ asymptotically. This proves the clipped stationarity guarantee for $p \in (1, 2]$. □