CLOSING THE BIAS-CORRECTION GAP IN DISCOUNTED REGRET ANALYSIS OF ADAM

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent discounted-to-nonconvex analyses of Adam establish convergence guarantees by bounding the discounted regret of an associated online learner. However, these analyses typically remove Adam's original bias-correction factors by absorbing them into the learning-rate schedule or clipping radius, and therefore prove guarantees for a reparameterized update. To accurately understand its convergence behavior, we analyze clipped Adam while retaining the original bias correction in this paper. The key difficulty is that bias correction introduces nonmonotone time-dependent coefficients into the regret analysis: the relevant coefficient sequence first decreases and then increases, forming a single-valley structure. Under standard stochastic cost-vector assumptions and the mild parameter coupling, we prove that the effective step sizes remain automatically monotone and develop a summation argument showing that the transient valley cost is controlled by sufficiently many post-valley terms. The resulting discounted-regret guarantees match the known $1/\sqrt{1-\beta_1}$ dependence of analyses without retained bias correction, and hence preserve Adam's original bias-corrected parameterization without worsening the regret order.

1 INTRODUCTION

Adam (Kingma, 2014) and its variants are among the most popular optimizers for modern neural-network training Devlin (2018); Radford et al. (2019); Bommasani et al. (2021); Touvron et al. (2023); Team et al. (2023). Despite Adam's empirical dominance over stochastic gradient descent (SGD) across diverse applications, a complete theoretical understanding that reconciles Adam's practical implementation with its performance guarantees remains elusive (Li et al., 2024; Wang et al., 2024).

Recent theoretical advances in the discounted-to-nonconvex conversion (DTNC) framework (Cutkosky et al., 2023; Zhang & Cutkosky, 2024; Ahn & Cutkosky, 2024; Ahn et al., 2024b) established a principled connection between adaptive optimization and online learning. This line of work shows that effective nonconvex optimizers correspond to online learners that achieve small discounted regret, thereby providing a unified lens for analyzing optimization dynamics. This viewpoint explains why Adam's component design is effective from the perspective of classical FTRL-type online learners (Orabona & Pál, 2018), and the corresponding regret analysis yields optimal non-smooth non-convex guarantees for clipped Adam-type algorithms.

Classical convergence analyses that often obtain guarantees only after imposing restrictive hyperparameter schedules, such as exponentially decaying second-moment parameters, omitting or simplifying bias correction, or retaining the original correction terms at the cost of degraded rates (Reddi et al., 2019; Défossez et al., 2020; Zhang et al., 2022; Li et al., 2023; Wang et al., 2023a; Hong & Lin, 2024; Jin et al., 2024); existing regret analyses likewise either impose restrictive parameter coupling (Ahn & Cutkosky, 2024) or absorb the bias-correction factors into an effective scalar schedule (Nguyen, 2025; Ahn & Cutkosky, 2024). This absorption changes the schedule parameterization, so the analysis in fact applies to a nearby Adam variant rather than the exact bias-corrected update. To achieve a more accurate analysis and understand the behavior, we therefore ask whether bias correction can be retained explicitly in the standard clipped DTNC setting.

We answer this question by proving pointwise and cumulative discounted-regret bounds for clipped, bias-corrected Adam. Retaining bias correction introduces two challenges: the effective step sizes

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.

are not obviously monotone, and a key Abel-summation coefficient becomes a single valley rather than monotone. We first show that, despite the nonmonotone bias-correction factor, the effective step sizes remain monotone under mild parameter coupling. We then show that the remaining nonmonotonicity has a controlled structure: over a sufficiently long horizon, the post-valley terms absorb the the cost of passing this valley. Consequently, our regret theorem obtains the same $1/\sqrt{1-\beta_1}$ dependence as prior analyses that remove bias correction. Combined with the discounted-to-nonconvex conversion, our regret bounds yield the optimal nonconvex stationarity guarantee, and the standard vector-to-coordinate-wise conversion gives the corresponding coordinate-wise clipped Adam consequence.

2 Related Work

Convergence and adaptivity analyses of Adam. A large body of work studies Adam and its variants through direct descent, stability, or stationarity arguments in convex, smooth non-convex, and generalized-smooth settings (Reddi et al., 2019; Zhou et al., 2018; Défossez et al., 2020; Zhang et al., 2022; Li et al., 2023; Wang et al., 2023a; Hong & Lin, 2024; Jin et al., 2024). These analyses answer when Adam-type methods converge and what rates they obtain under their own assumptions. Many of them, however, simplify the practical recursion or use restrictive regimes, such as modified adaptive updates, time-varying second-moment schedules, omitted or absorbed bias correction, or correction terms that are retained only with worse rates. Several results also clarify why adaptive methods can be preferable to SGD rather than merely matching its worst-case rates. Gradient clipping, AdaGrad, Adam/RMSProp variants, and generalized sign methods have been shown to handle structured or generalized-smooth geometries in which fixed-schedule SGD may be slow or unstable (Zhang et al., 2019; Wang et al., 2023b; Zhang et al., 2024; Li et al., 2023; Crawshaw et al., 2022). Within the DTNC line, Ahn & Cutkosky (2024) further show that Adam-type adaptivity can yield better problem-dependent guarantees when relevant problem parameters are unknown.

Online learning and discounted regret. The closest line to this work connects non-convex optimization with online learning through discounted-to-nonconvex conversion (Cutkosky et al., 2023; Zhang & Cutkosky, 2024). In this framework, an optimizer is analyzed as an online learner, and discounted regret controls the stationarity guarantee after conversion. Adam-style updates arise naturally from FTRL-type online learners (Orabona & Pál, 2018; Ahn et al., 2024b), which makes the roles of momentum, discounting, and adaptive scaling explicit. Ahn & Cutkosky (2024) use this perspective to obtain optimal non-smooth non-convex guarantees for clipped Adam with model EMA, but their regret analysis relies on a restrictive momentum coupling and removes bias correction. Nguyen (2025) substantially relaxes the coupling requirement, but still absorbs bias correction into the scalar learning rate.

3 Background and Preliminaries

We first briefly introduce fundamental concepts of online learning and essential definitions of regret, which underpin our analysis, then formulate Adam as an online learner.

Online learning and discounted-to-nonconvex conversion. Online Linear Optimization (OLO) is the following sequential decision protocol: at round t out of T total rounds, the online learner selects an action $\mathbf{z}_t \in \mathbb{R}^d$ and then receives a linear loss $\ell_t(\cdot) := \langle \mathbf{v}_t, \cdot \rangle$, where $\mathbf{v}_t \in \mathbb{R}^d$ is a cost vector revealed by the environment. The learner aims to minimize static regret, defined as the cumulative difference between the learner's losses and those of an arbitrary comparator $\mathbf{u} \in \mathcal{D} \subseteq \mathbb{R}^d$: $\text{Regret}_T(\mathbf{u}) := \sum_{t=1}^T \langle \mathbf{v}_t, \mathbf{z}_t - \mathbf{u} \rangle$. In this work, we specifically focus on a bounded domain scenario, where $\mathcal{D}$ is the d-dimensional L_2-ball with radius D. Then, we introduce the concept of β-discounted static regret, essential to our framework. Given an algorithm that discounts the loss by β^{T-t} at round t out of T total rounds, i.e., $\ell_t^{[\beta]}(\cdot) = \beta^{T-t}\ell_t(\cdot)$, the associated β-discounted regret is defined as: $\text{Regret}_T^{[\beta]}(\mathbf{u}) := \sum_{t=1}^T \langle \beta^{T-t}\mathbf{v}_t, \mathbf{z}_t - \mathbf{u} \rangle$. For further details about online learning, readers may refer to Orabona (2019); Cutkosky et al. (2023).

A recent advance in optimization research is the discounted-to-nonconvex conversion framework (Ahn & Cutkosky (2024, Lemma 7) and Ahn et al. (2024a, Lemma 1)), which connects the dis-

counted regret of an associated online learner to stationarity guarantees in (non-smooth/smooth) non-convex optimization. At a high level, these results show that the expected norm of the gradient produced by the conversion algorithm can be controlled by a combination of optimization error terms and discounted regret. Concretely, one obtains bounds of the generic form $\mathbb{E}_{t\sim[T]}\big\|\mathbb{E}_{\mathbf{y}_t}\nabla F(\mathbf{y}_t)\big\| \lesssim \frac{1}{DT}\Big[\beta\,\mathbb{E}\big[\mathrm{Regret}_T^{[\beta]}(\mathbf{u}_T)\big] + (1-\beta)\sum_{t=1}^{T}\mathbb{E}\big[\mathrm{Regret}_t^{[\beta]}(\mathbf{u}_t)\big]\Big]$, where $\mathbf{y}_t$ is sampled from the parameter iterates (see Appendix I). Hence, a tight discounted regret bound translates directly into the main term in the non-convex convergence guarantee.

Given our focus on studying the stochastic optimization method Adam through the online learning techniques, regarding the cost vectors, we make assumptions corresponding to the standard assumptions in non-smooth non-convex stochastic optimization (Cutkosky et al., 2023; Ahn & Cutkosky, 2024; Zhang et al., 2020; Jordan et al., 2023), summarized as Assumption 3.1.

Assumption 3.1. For each $t \in [T]$, let $\mathbf{v}_t \in \mathbb{R}^d$ denote a stochastic cost vector. Let $\boldsymbol{\mu}_t := \mathbb{E}[\mathbf{v}_t]$. Assume the following conditions hold: (1) **Bounded mean:** The mean vector satisfies $\|\boldsymbol{\mu}_t\| \le G$. (2) **Bounded variance:** $\mathbb{E}\left[\|\mathbf{v}_t - \boldsymbol{\mu}_t\|^2\right] \le \sigma^2$.

The bounded mean corresponds, under the stochastic-optimization interpretation $\boldsymbol{\mu}_t = \nabla F(\mathbf{x}_t)$, to the standard bounded-gradient condition. For differentiable objectives, this is the Lipschitz-function assumption commonly used in non-smooth non-convex optimization. It is complementary to, and not implied by, the usual Lipschitz-gradient (smoothness) assumption.

3.1 Adam as an online learner

We present the clipped bias-corrected Adam update in the form of online learning. The increment $\mathbf{z}_t$ at step t is given by:

$$\mathbf{z}_t = \arg\min_{\mathbf{z}\in\mathcal{D}}\left\{\frac{\|\mathbf{z}\|^2}{2\alpha_t} + \sum_{s=1}^{t-1}\langle\beta_1^{-s}\mathbf{v}_s, \mathbf{z}\rangle\right\} = \mathrm{clip}_D\left(-\alpha_t\sum_{s=1}^{t-1}\beta_1^{-s}\mathbf{v}_s\right),$$

where $\alpha_t := \eta_t\frac{(1-\beta_1)\beta_1^{t-1}}{(1-\beta_1^{t-1})(\sqrt{b_t}+\epsilon)}$, $b_t := \frac{1-\beta_2}{1-\beta_2^{t-1}}\sum_{s=1}^{t-1}\beta_2^{t-1-s}\|\mathbf{v}_s\|^2$, and $\mathrm{clip}_D(\mathbf{z}) := \frac{\mathbf{z}}{\|\mathbf{z}\|}\min(D, \|\mathbf{z}\|)$ with $\mathrm{clip}_D(0) = 0$. At $t = 1$ the history is empty, so $\mathbf{z}_1 = \mathbf{0}$ regardless of the value assigned to α_1. Equivalently, before clipping, for $t \ge 2$ the increment is $-\alpha_t\sum_{s=1}^{t-1}\beta_1^{-s}\mathbf{v}_s = -\eta_t\frac{\frac{1-\beta_1}{1-\beta_1^{t-1}}\sum_{s=1}^{t-1}\beta_1^{t-1-s}\mathbf{v}_s}{\sqrt{\frac{1-\beta_2}{1-\beta_2^{t-1}}\sum_{s=1}^{t-1}\beta_2^{t-1-s}\|\mathbf{v}_s\|^2+\epsilon}}$. This is exactly the vector-valued Adam increment obtained by unrolling the classical EMA updates with zero initialization, where β_1 and β_2 are the EMA coefficients for the first- and second-moment accumulators, respectively; We call $\{\eta_t\}$ and $\{\alpha_t\}$ the learning rate sequence and effective step size sequence, respectively.

Remark 3.2 (Vector and coordinate-wise forms). Our main analysis treats Adam as a vector-valued online learner, with the primary focus on discounted regret over the L_2 ball. The corresponding Euclidean-norm stationarity consequence follows the discounted-to-nonconvex conversion, and the coordinate-wise result matching the coordinate-wise Adam, with the clipping required by our regret theorem, follows from the standard vector-to-coordinate-wise conversion pipeline used in Cutkosky et al. (2023); Ahn & Cutkosky (2024). The resulting stationarity and coordinate-wise consequences are stated after the main regret theorem (Section 7.1).

4 Regret Decomposition and Two Challenges

With Adam formulated as an online learner, we now turn to the discounted regret analysis.

4.1 Regret decomposition

Our analysis starts from the intermediate regret decomposition of Nguyen (2025), which splits the discounted regret bound into a comparator term plus a $\min$ of a quadratic-gradient term and a movement term (the exact form is adapted to suit the analysis in this work; proof in Appendix A).

Throughout the Adam regret analysis, the discount parameter in $\mathrm{Regret}^{[\beta]}$ is fixed to be $\beta = \beta_1$.

Theorem 4.1 (Discounted regret — intermediate form). *For any $t \in [T]$ and any comparator $\mathbf{u} \in \mathcal{D}$ with $\|\mathbf{u}\| \le D$, consider the clipped Adam/FTRL learner from Section 3.1 with step sizes $\eta_s \le \eta$, any $\epsilon > 0$, and non-increasing effective step sizes $\alpha_s \ge \alpha_{s+1}$. Then the discounted regret satisfies*

$$\mathrm{Regret}_t^{[\beta]}(\mathbf{u}) \;\le\; \underbrace{\frac{\beta_1^t \|\mathbf{u}\|^2}{2\alpha_{t+1}}}_{\textit{comparator}} \;+\; \sum_{s=1}^{t} \min\Big(\underbrace{\frac{\alpha_s}{2}\,\beta_1^{t-2s}\|\mathbf{v}_s\|^2}_{\textit{quadratic increment}},\; \underbrace{\beta_1^{t-s}\|\mathbf{z}_s - \mathbf{z}_{s+1}\|\|\mathbf{v}_s\|}_{\textit{movement increment}} \Big). \tag{1}$$

The summand-wise $\min$ in equation 1 pairs the past-only effective step size α_s (which depends only on $\mathbf{v}_1, \dots, \mathbf{v}_{s-1}$) with the current gradient $\mathbf{v}_s$, and admits a usable bound only via a per-index decision: at each s we route the summand to whichever of its two arguments can be controlled. We formalize this through a *selected-event* split, partitioning the indices $s \ge 2$ into the reverse-growth event $\mathcal{R}_s$ and its complement, the growth event $\mathcal{G}_s := \mathcal{R}_s^c$, both defined by a growth condition on the weighted gradient sequence $\{\beta_2^{-s}\|\mathbf{v}_s\|\}$:

- **Reverse-growth event $\mathcal{R}_s$:** $\beta_2^{-2s}\|\mathbf{v}_s\|^2 \le 2\sum_{n=1}^{s-1}\beta_2^{-2n}\|\mathbf{v}_n\|^2$ — the new weighted squared gradient is controlled by the cumulative weighted past up to a factor of two, so the *quadratic increment* admits an Abel-summation bound (Section 6).
- **Growth event $\mathcal{G}_s$:** $\beta_2^{-2s}\|\mathbf{v}_s\|^2 > 2\sum_{n=1}^{s-1}\beta_2^{-2n}\|\mathbf{v}_n\|^2$ — the new weighted gradient is a selected large jump with a factor-two gap, so the *movement increment* admits a packing-style geometric-growth bound under clipping (Section 5).

This indexing upgrades the $\min$ to an indicator decomposition:

$$\min\Big(\tfrac{\alpha_s}{2}\beta_1^{t-2s}\|\mathbf{v}_s\|^2,\; \beta_1^{t-s}\|\mathbf{z}_s - \mathbf{z}_{s+1}\|\,\|\mathbf{v}_s\|\Big) \le \tfrac{\alpha_s}{2}\beta_1^{t-2s}\|\mathbf{v}_s\|^2\,\mathbf{1}_{\mathcal{R}_s} + \beta_1^{t-s}\|\mathbf{z}_s - \mathbf{z}_{s+1}\|\,\|\mathbf{v}_s\|\,\mathbf{1}_{\mathcal{G}_s},$$

routing each summand to a single argument. We adopt this selected-event split throughout: Section 5 bounds the $\mathcal{G}_s$-side using the clipping step $\mathbf{z} \leftarrow \mathrm{clip}_D(\cdot)$ (matching prior work), and Section 6 bounds the $\mathcal{R}_s$-side under retained bias correction (the new contribution). The technical reasons why prior work dropped bias correction are the subject of the next subsection.

4.2 Two bias-correction challenges

We catalog the two challenges created by retaining bias correction.

Challenge 1: "non-monotonicity" of α_s. The standard FTRL regret analysis (Orabona, 2019) requires the regularizer weights $1/\alpha_s$ to be non-decreasing, which is equivalent to $\alpha_s \ge \alpha_{s+1}$. This non-increasing step-size requirement is explicitly assumed in both Ahn & Cutkosky (2024) and Nguyen (2025). Retaining bias correction complicates the previously simple step-size formula: the bias correction's non-monotone ratio enters α_s, disrupting the straightforward monotonicity that holds when bias correction is canceled.

We resolve this challenge in Section 4.3: under $\beta_1 \le \beta_2$, we prove that the effective step size α_s with constant learning rate η is automatically non-increasing (Proposition 4.3).

Challenge 2: non-monotonicity of the coefficient sequence in the quadratic term of equation 1. Even though α_s is automatically non-increasing (Section 4.3), the quadratic sum $\sum_s \frac{\alpha_s}{2}\beta_1^{t-2s}\|\mathbf{v}_s\|^2$ remains difficult. To bound this sum, one reformulates each summand as a product of a coefficient times an increment (Section 6). Without bias correction, the resulting coefficient sequence grows monotonically (Section 6.1), and the sum is thereby bounded by its terminal term. With bias correction, the ratio $\sqrt{1-\beta_2^{s-1}}/(1-\beta_1^{s-1})$ enters the coefficient. The result is a coefficient that first *decreases* and then *increases* — a behavior we call *single-valley structure* (formalized as Lemma 6.2 in Section 6). Unlike Challenge 1, which can be resolved under $\beta_1 \le \beta_2$ (Proposition 4.3), this challenge is intrinsic to the core component of the quadratic term.

Remark 4.2 (Why we shift the object of analysis). Challenge 2 has a concrete consequence. When each summand in the quadratic sum is reformulated as a coefficient times an increment (Section 6), bounding the sum by its terminal term requires the coefficient sequence to be non-decreasing. The

single-valley structure violates this: for any t below an explicit threshold T_0 (Proposition 6.5), the coefficient at the boundary has not yet recovered from the valley, so the clean per-iteration bound is unavailable from the Abel framework.

Our resolution: instead of bounding Regret_t for every t as in existing work (Ahn & Cutkosky, 2024; Nguyen, 2025), we separate the regret guarantees required by the analysis. We prove a clean pointwise bound for Regret_T once $T \geq T_0$, and we also prove a cumulative discounted-regret bound for $\sum_{t=1}^{T} \mathrm{Regret}_t$ once $T \geq T_1$. The cumulative result is a standalone discounted-regret guarantee showing that the single-valley transient can be amortized over time.

4.3 Automatic step-size monotonicity (resolving Challenge 1)

Initial-point convention. At $s = 1$ the closed form for α_s (Section 3.1) is indeterminate: the factor $(1 - \beta_1^0)^{-1}$ is undefined and $b_1 = 0$ (no past gradients). Because the FTRL iterate $\mathbf{z}_1 = \arg\min_{\mathbf{z}} \|\mathbf{z}\|^2/(2\alpha_1) = \mathbf{0}$ is independent of α_1, the value of α_1 may be fixed by convention without affecting the algorithm's trajectory; any choice satisfying $\alpha_1 \geq \alpha_s$ for every $s \geq 2$ is admissible for the analysis. We adopt $\alpha_1 := \alpha_2 = \frac{\eta\beta_1}{\sqrt{b_2}+\epsilon}$, the *minimal* admissible choice and the natural extension of the closed form at the boundary. With this convention, the following proposition shows that α_s is automatically non-increasing for all $s \geq 1$.

Proposition 4.3 (Automatic step-size monotonicity)**.** *Let $\beta_1, \beta_2 \in (0, 1)$ with $\beta_1 \leq \beta_2$ and $\epsilon > 0$. With constant learning rate $\eta_s = \eta$ and the initialization $\alpha_1 := \alpha_2$, the effective step size $\alpha_s = \frac{\eta(1-\beta_1)\beta_1^{s-1}}{(1-\beta_1^{s-1})(\sqrt{b_s}+\epsilon)}$, $s \geq 2$, is non-increasing: $\alpha_s \geq \alpha_{s+1}$ for all $s \geq 1$.*

The proof is given in Appendix B. The key mechanism is that the decay factor $(1 - \beta_1)\beta_1^{s-1}/(1 - \beta_1^{s-1})$ decreases at rate β_1, while the denominator factor $\sqrt{b_s}+\epsilon$ can decrease at most at rate $\sqrt{\beta_2} > \beta_1$.

5 Movement-Term Bound with Clipping

This section provide the necessary bound for the movement term in equation 1. The final regret argument uses the movement increment only on indices where the local growth event holds. We assign the boundary term $s = 1$ to the movement side and define the selected movement envelope

$$M_t^{\mathrm{sel}} := \beta_1^{t-1}\|\mathbf{z}_1 - \mathbf{z}_2\| \, \|\mathbf{v}_1\| + \textstyle\sum_{s=2}^{t} \beta_1^{t-s}\|\mathbf{z}_s - \mathbf{z}_{s+1}\| \, \|\mathbf{v}_s\| \mathbf{1}_{\mathcal{G}_s}.$$

Proposition 5.1 (Selected movement bound with clipping)**.** *Let $0 < \beta_1 \leq \beta_2 < 1$. Assume additionally that the iterates are clipped onto $\mathcal{D}$, so that $\|\mathbf{z}_s\| \leq D$ and therefore $\|\mathbf{z}_s - \mathbf{z}_{s+1}\| \leq 2D$ for all s. Then, define $C_M^{\mathrm{clip}} := \frac{2(3+\sqrt{2})}{\sqrt{1-\beta_1^2}}$, for every $T \geq 1$,*

$$\mathbb{E}\big[M_T^{\mathrm{sel}}\big] \leq C_M^{\mathrm{clip}} \, D \, (G + \sigma), \quad \textit{and} \quad \mathbb{E}\Big[\textstyle\sum_{t=1}^{T} M_t^{\mathrm{sel}}\Big] \leq C_M^{\mathrm{clip}} \, D \, (G + \sigma) \, T,$$

We work under the same clipped setting and follow the same proof strategy as Ahn & Cutkosky (2024); Nguyen (2025). The full proof is deferred to Appendix C.

6 Quadratic-Term Bound

This section resolves Challenge 2 from Section 4.2 for the reverse-growth summands selected by the local split in equation 1.

Proof sketch. The proof has three steps. First, each selected reverse-growth summand is rewritten as a coefficient H_s times an increment $I_s - I_{s-1}$. Second, Abel summation turns the resulting sum into a terminal term plus a remainder governed by the single-valley shape of $\{H_s\}$. Third, the pointwise and cumulative discounted quantities handle this valley remainder differently: the pointwise bound waits until H_T has recovered, while the cumulative bound shows that the descending-phase cost is amortized after a second explicit threshold. In particular, the first step is a local reformulation:

$$\frac{\alpha_s}{2}\beta_1^{-2s}\|\mathbf{v}_s\|^2 \mathbf{1}_{\mathcal{R}_s} \leq H_s(I_s - I_{s-1})\mathbf{1}_{\mathcal{R}_s} \leq H_s(I_s - I_{s-1}).$$

Thus the selected quadratic term is bounded by the Abel summation:

$$\beta_1^t \sum_{s=2}^{t} \frac{\alpha_s}{2} \beta_1^{-2s} \|\mathbf{v}_s\|^2 \mathbf{1}_{\mathcal{R}_s} \le \beta_1^t \sum_{s=1}^{t} H_s(I_s - I_{s-1}) = \beta_1^t \Big(H_t I_t + \sum_{s=1}^{t-1} (H_s - H_{s+1}) I_s \Big),$$

where $H_1 := 0$ and remainder $R(T) := \sum_{s=1}^{T-1}(H_s - H_{s+1})I_s$. The remaining analysis is the single-valley Abel argument: the per-iteration remainder is non-positive once $t \ge T_0$, and the cumulative remainder is absorbed once $T \ge T_1$.

6.1 Reformulation and properties of H_s

We start with the selected reformulation of each quadratic summand. See Appendix D for the proof.

Lemma 6.1 (Selected reformulation). *For $s \ge 2$, suppose $\eta_s \le \eta$ and define $\mathcal{R}_s := \Big\{ \beta_2^{-2s} \|\mathbf{v}_s\|^2 \le 2 \sum_{n=1}^{s-1} \beta_2^{-2n} \|\mathbf{v}_n\|^2 \Big\}$. Define*

$$H_s := \sqrt{3}\eta \, \frac{(1-\beta_1)\sqrt{1-\beta_2^{s-1}}}{(1-\beta_1^{s-1})\sqrt{1-\beta_2}} \, \frac{\beta_2^{s+1}}{\beta_1^{s+1}\beta_2^{s/2}}, \qquad I_s := \sqrt{\textstyle\sum_{n=1}^{s} \beta_2^{-n-1} \|\mathbf{v}_n\|^2}, \qquad I_0 := 0.$$

Then $\frac{\alpha_s}{2} \beta_1^{-2s} \|\mathbf{v}_s\|^2 \mathbf{1}_{\mathcal{R}_s} \le H_s \big(I_s - I_{s-1}\big) \mathbf{1}_{\mathcal{R}_s} \le H_s \big(I_s - I_{s-1}\big)$.

Note that H_s is exactly the coefficient discussed in Section 4.2: it contains the *bias-correction ratio*. Without bias correction, this ratio simplifies to $H_s/H_{s+1} = \beta_1/\sqrt{\beta_2} \le 1$ under $\beta_1 \le \beta_2$, so $\{H_s\}$ is non-decreasing and the Abel remainder is non-positive, considerably simplifying the analysis.

The following Lemma 6.2 confirms the single-valley structure of the coefficient sequence $\{H_s\}$ (Appendix E).

Lemma 6.2 (Single-valley structure). *Let $0 < \beta_1 < \beta_2 < 1$, and let $\{H_s\}_{s \ge 2}$ be defined as in Lemma 6.1. Then there exists a unique $s_\star \ge 2$ such that $H_{s+1} < H_s$ for $2 \le s < s_\star$ and $H_{s+1} \ge H_s$ for $s \ge s_\star$.*

Additionally, we further characterize H_s with the following Lemma 6.3 (Proof in Appendix F).

Lemma 6.3 (Properties of H_s). *Let $0 < \beta_1 \le \beta_2 < 1$. Write $\rho := \sqrt{\beta_2}/\beta_1 > 1$ and $C_L := \frac{\sqrt{3}\,\eta\,(1-\beta_1)\,\beta_2}{\beta_1}$. Define $C_{\mathrm{H}} := \frac{\sqrt{6\beta_2}\,(1-\beta_1)}{\beta_1(1-\beta_2)}$.* (i) ***Lower bound:*** *$H_s \ge C_L \rho^s$ for all $s \ge 2$.* (ii) ***Explicit H_2:*** *$H_2 = \frac{\sqrt{3}\,\eta\,\beta_2^2}{\beta_1^3} = \frac{C_L \rho^2}{1-\beta_1}$.* (iii) ***Boundary:*** *We set $H_1 := 0$ by convention to drop the $s = 1$ term.* (iv) ***Unconditional expectation bound:*** *Under Assumption 3.1, for every $s \ge 2$, $\mathbb{E}[\beta_1^s H_s I_s] \le C_{\mathrm{H}}\,\eta\,(G + \sigma)$.*

The constant C_{H} is kept explicit in the main regret bounds; under fixed proportional coupling $1 - \beta_2 = r(1-\beta_1)$ it contributes only to the corresponding $O_r(\cdot)$ constants.

6.2 Per-iteration limit of terminal Abel bounds

With the infrastructure above, we can formalize the *framework-level limit* of per-iteration Abel analysis with bias correction (as previewed in Remark 4.2). The limitation is not only that our proof produces a degraded upper bound: any argument that bounds the Abel sum solely by its terminal term must pay the valley depth of the coefficient sequence. This is a statement about the Abel envelope itself, not a lower bound on the optimizer's actual performance.

Proposition 6.4 (Per-iteration limit of the Abel framework). *Let $0 < \beta_1 < \beta_2 < 1$. Let H_s and I_s be defined as in Lemma 6.1, with $H_1 := 0$, and let $R(T) := \sum_{s=1}^{T-1}(H_s - H_{s+1})I_s$.*

(i) ***Threshold for the clean per-iteration Abel envelope.*** *If $T \ge T_0 := 2 + \left\lceil \frac{\log(1/(1-\beta_1))}{\log(\sqrt{\beta_2}/\beta_1)} \right\rceil$, then $R(T) \le 0$, and therefore $\beta_1^T \sum_{s=1}^{T} H_s(I_s - I_{s-1}) \le \beta_1^T H_T I_T$.*

(ii) ***Penalized upper bound for*** $T < T_0$**.** *In the pre-recovery regime* $2 \le T < T_0$*, combining Abel summation with the single-valley structure (Lemma 6.2) and the ratio* $H_2/H_s \le 1/(1-\beta_1)$ *(Lemma 6.3*(i)-(ii)*) yields the* penalized *pathwise envelope bound*

$$\beta_1^T \sum_{s=1}^{T} H_s(I_s - I_{s-1}) \;\le\; \frac{1}{1-\beta_1}\,\beta_1^T H_T I_T. \tag{2}$$

(iii) ***Sharpness and magnitude of terminal-term Abel bounds.*** *For any fixed* $T \ge 2$*, the smallest constant* Γ_T *for which* $\sum_{s=1}^{T} H_s(I_s - I_{s-1}) \le \Gamma_T H_T I_T$ *holds for every non-decreasing sequence* $0 = I_0 \le I_1 \le \cdots \le I_T$ *is* $\Gamma_T = \max_{1\le s\le T} \frac{H_s}{H_T}$*. Consequently, for every* $2 \le T \le s_\star$*,* $\Gamma_T = \frac{H_2}{H_T}$*, and at the bottom of the valley the necessary terminal-term constant is exactly* $H_2/H_{s_\star}$*. Moreover, if* $r \in (0,1)$ *is fixed and* $1-\beta_2 = r(1-\beta_1)$*, then, as* $\beta_1 \uparrow 1$*,* $\Gamma_{s_\star} = \frac{H_2}{H_{s_\star}} = \Theta_r\Big(\frac{1}{\sqrt{1-\beta_1}}\Big)$*. Here* $\Theta_r(\cdot)$ *hides constants that may depend on* r*.*

Part (ii) gives the convenient explicit upper bound, while Part (iii) shows that the sharp terminal-envelope valley loss has square-root size in the standard proportional-coupling regime. Thus, shifting away from the previous analysis target of bounding Regret_t for every t is not merely a cosmetic proof choice: rather than paying the pre-recovery valley depth in a per-iteration terminal bound, sufficiently many post-valley terms mitigate the valley's fixed cost. The proof is given in Appendix G.

6.3 Discounted selected-quadratic bounds (resolving Challenge 2)

This subsection resolves Challenge 2 for the selected reverse-growth side of the local split.

Proposition 6.5 (Selected reverse-growth quadratic bounds)**.** *Under Assumption 3.1, let* $\beta_1, \beta_2 \in (0,1)$ *with* $\beta_1 < \beta_2$ *and* $\eta_s \le \eta$*. Define* C_{H} *as in Lemma 6.3, for* $s \ge 2$*, define* $\mathcal{R}_s$ *as in Lemma 6.1, and for* $t \ge 1$ *define*

$$Q_t^{\mathrm{sel}} := \beta_1^t \sum_{s=2}^{t} \frac{\alpha_s}{2}\,\beta_1^{-2s}\|\mathbf{v}_s\|^2 \mathbf{1}_{\mathcal{R}_s} \qquad T_0 := 2 + \left\lceil \frac{\log(1/(1-\beta_1))}{\log(\sqrt{\beta_2}/\beta_1)} \right\rceil, \qquad T_1 := \left\lceil \frac{\beta_1(T_0-2)}{1-\beta_1} \right\rceil.$$

Part I (Pointwise). *For every* $T \ge T_0$*:*

$$\boxed{\mathbb{E}\Big[Q_T^{\mathrm{sel}}\Big] \le C_{\mathrm{H}}\,\eta\,(G+\sigma).} \tag{3}$$

Part II (Cumulative). *For every* $T \ge T_1$*:*

$$\boxed{\mathbb{E}\Big[\sum_{t=1}^{T} Q_t^{\mathrm{sel}}\Big] \le 2C_{\mathrm{H}}\,\eta\,(G+\sigma)\,T.} \tag{4}$$

Both thresholds T_0 and T_1 are explicit and depend only on β_1 and β_2.

The proof of Part I shows directly that the Abel-envelope remainder $R(T) \le 0$ for $T \ge T_0$ by exploiting the fact that H_T has recovered past the valley. For Part II, summing β_1^t times the Abel expansion over $t = 1, \ldots, T$ produces a double-sum quantity that appears with a *minus* sign. By the single-valley structure (Lemma 6.2), only the descending-phase terms ($s < s_\star$) can contribute negatively, yielding a penalty of at most $\frac{\beta_1(s_\star-2)}{1-\beta_1} \le \frac{\beta_1(T_0-2)}{1-\beta_1}$ terms; for $T \ge T_1$ this penalty is absorbed by the main sum, yielding the factor of 2. The full proof is given in Appendix H.

7 Final Discounted Regret Bounds

We now combine the rigorous selected reverse-growth analysis from Section 6 with the clipped movement bound from Section 5. Both ingredients are fully proved, so the final theorem is unconditional within the clipped bounded-domain setting.

Local increments and regimes. For $s \ge 1$, define $m_s := \frac{\alpha_s}{2}\beta_1^{-2s}\|\mathbf{v}_s\|^2$, $n_s := \beta_1^{-s}\|\mathbf{z}_s - \mathbf{z}_{s+1}\|\,\|\mathbf{v}_s\|$. We use the convention that empty sums are zero and assign the boundary term $s = 1$ to the movement side:

$$Q_t^{\mathrm{sel}} := \beta_1^t \sum_{s=2}^{t} m_s\,\mathbf{1}_{\mathcal{R}_s}, \quad \text{and} \quad M_t^{\mathrm{sel}} := \beta_1^t n_1 + \beta_1^t \sum_{s=2}^{t} n_s\,\mathbf{1}_{\mathcal{G}_s}.$$

The summand-wise regret decomposition gives, pathwise,

$$\beta_1^t \sum_{s=1}^{t} \min(m_s, n_s) \le Q_t^{\mathrm{sel}} + M_t^{\mathrm{sel}}, \tag{5}$$

because $\min(m_1, n_1) \le n_1$ and $\min(m_s, n_s) \le m_s \mathbf{1}_{\mathcal{R}_s} + n_s \mathbf{1}_{\mathcal{G}_s}$ for $s \ge 2$.

Bounding the comparator term. By Proposition 4.3, $\alpha_{t+1} = \frac{\eta(1-\beta_1)\beta_1^t}{(1-\beta_1^t)(\sqrt{b_{t+1}}+\epsilon)}$, so $\frac{\beta_1^t}{2\alpha_{t+1}} = \frac{(1-\beta_1^t)(\sqrt{b_{t+1}}+\epsilon)}{2\eta(1-\beta_1)} \le \frac{\sqrt{b_{t+1}}+\epsilon}{2\eta(1-\beta_1)}$. Since b_{t+1} is a nonnegative random variable with $\mathbb{E}[b_{t+1}] \le (G+\sigma)^2$, Jensen's inequality gives $\mathbb{E}[\sqrt{b_{t+1}}] \le \sqrt{\mathbb{E}[b_{t+1}]} \le G + \sigma$, so

$$\mathbb{E}\left[\frac{\beta_1^t \|\mathbf{u}\|^2}{2\alpha_{t+1}}\right] \;\le\; \frac{D^2(G+\sigma+\epsilon)}{2\eta(1-\beta_1)}\,. \tag{6}$$

Selected gradient terms. Combining equation 5 with Proposition 6.5 and Proposition 5.1 gives,

$$\text{for } T \ge T_0, \qquad \mathbb{E}\left[\beta_1^T \sum_{s=1}^{T} \min(m_s, n_s)\right] \le C_{\mathrm{H}}\,\eta\,(G+\sigma) + C_M^{\mathrm{clip}}\,D\,(G+\sigma), \tag{7}$$

$$\text{and, for } T \ge T_1, \quad \mathbb{E}\left[\sum_{t=1}^{T} \beta_1^t \sum_{s=1}^{t} \min(m_s, n_s)\right] \le \big[2C_{\mathrm{H}}\,\eta\,(G+\sigma) + C_M^{\mathrm{clip}}\,D\,(G+\sigma)\big]T. \tag{8}$$

Theorem 7.1 (Final discounted regret bounds with clipping)**.** *Let Assumption 3.1 hold, let* $0 < \beta_1 < \beta_2 < 1$*, and let* $\eta_s = \eta > 0$*. Consider the clipped Adam/FTRL update from Section 3.1, with* $\epsilon > 0$ *and* $\alpha_1 := \alpha_2$*. Then:*

Part I (Pointwise). *For every* $T \ge T_0$ *and every comparator* $\mathbf{u}_T \in \mathcal{D}$*:*

$$\boxed{\mathbb{E}\big[\mathrm{Regret}_T^{[\beta]}(\mathbf{u}_T)\big] \le \frac{D^2(G+\sigma+\epsilon)}{2\eta(1-\beta_1)} + C_{\mathrm{H}}\,\eta\,(G+\sigma) + C_M^{\mathrm{clip}}\,D\,(G+\sigma).}$$

Part II (Cumulative). *For* $T \ge \max(T_0, T_1)$ *and every comparator sequence* $\mathbf{u}_t \in \mathcal{D}$*:*

$$\boxed{\sum_{t=1}^{T} \mathbb{E}\big[\mathrm{Regret}_t^{[\beta]}(\mathbf{u}_t)\big] \le \left[\frac{D^2(G+\sigma+\epsilon)}{2\eta(1-\beta_1)} + 2C_{\mathrm{H}}\,\eta\,(G+\sigma) + C_M^{\mathrm{clip}}\,D\,(G+\sigma)\right] T.}$$

In particular, in the standard proportional-coupling regime $1-\beta_2 = r(1-\beta_1)$ *with fixed* $r \in (0,1)$*, choosing* $\eta = D\sqrt{\frac{G+\sigma+\epsilon}{4\,C_{\mathrm{H}}\,(1-\beta_1)\,(G+\sigma)}}$ *yields both bounds in the order* $O_r\Big(\frac{D\,(G+\sigma+\epsilon)}{\sqrt{1-\beta_1}}\Big)$*, matching the* $1/\sqrt{1-\beta_1}$ *rate of Ahn & Cutkosky (2024); Nguyen (2025) without bias correction. Hence retaining bias correction does not worsen the discounted-regret order. Outside this proportional-coupling simplification, the displayed bounds retain an explicit* $1/\sqrt{1-\beta_2}$ *factor through* C_{H}*.*

Proof. By Proposition 4.3, the effective step sizes are non-increasing under the stated initialization, so Theorem 4.1 applies. Taking expectation in equation 1 at $t = T$ and applying equation 6 and equation 7 gives the pointwise bound. The same argument summed over $t = 1, \dots, T$, using equation 6 uniformly in t and equation 8, gives the cumulative bound. For the stated step size, the common comparator-plus-quadratic bracket is

$$\frac{D^2(G+\sigma+\epsilon)}{2\eta(1-\beta_1)} + 2C_{\mathrm{H}}\eta(G+\sigma) = 2D\sqrt{\frac{C_{\mathrm{H}}(G+\sigma)(G+\sigma+\epsilon)}{1-\beta_1}} \le D(G+\sigma+\epsilon)\sqrt{\frac{4C_{\mathrm{H}}}{1-\beta_1}},$$

which has the claimed order. □

7.1 STATIONARITY AND COORDINATE-WISE CONSEQUENCES

For completeness, we follow the discounted-to-nonconvex conversion of Cutkosky et al. (2023); Ahn & Cutkosky (2024) to translate our regret bound into stationarity guarantees. The next theorem states the resulting Euclidean-norm stationarity guarantee; the stationarity measure, full assumptions, and proof are in Appendix I. For readability, the displayed lower bounds on T only keep the $\varepsilon_{\mathrm{stat}}$-dominant term; lower-order and logarithmic terms are given in the appendix.

Theorem 7.2 (Stationarity)**.** *Under Assumption I.1, let $A := G + \sigma + \epsilon$. For any $r \in (0,1)$, $\lambda > 0$, and $\varepsilon_{\mathrm{stat}} \in (0, A]$, there are choices of $\beta_1, \beta_2, D, \eta$ with $\beta_2 = 1 - r(1-\beta_1)$ such that, for $T = \Omega_r\left(\Delta \lambda^{1/2} A^2 \varepsilon_{\mathrm{stat}}^{-7/2}\right)$, the model-EMA output $\bar{\mathbf{x}}_t$ satisfies $\mathbb{E}_{t\sim[T]}\mathbb{E}\left[\|\nabla F(\bar{\mathbf{x}}_t)\|^{[\lambda]}\right] \le O(\varepsilon_{\mathrm{stat}})$.*

The proof of Theorem 7.2 combines the discounted-to-nonconvex conversion with Theorem 7.1 and the standard localization bound for clipped increments. Additionally, the cumulative part of Theorem 7.1 is a discounted-regret consequence of independent interest. For the stationarity theorem above, we only need the pointwise part for $t \ge T_0$; the first T_0 rounds are handled by a direct crude regret bound in Appendix I. Finally, its $\varepsilon_{\mathrm{stat}}^{-7/2}$ dominant iteration dependence matches the known lower-bound rate for the corresponding non-smooth non-convex stationarity notion (Zhang & Cutkosky, 2024).

The following corollary records the corresponding coordinate-wise clipped Adam consequence: coordinate aggregation gives pointwise and cumulative (coordinate-wise) discounted-regret bounds, and the L_1 conversion gives the associated stationarity guarantee.

Corollary 7.3 (Coordinate-wise clipped Adam: coordinate-wise regret and L_1 stationarity)**.** *Let $\delta := 1 - \beta_1$ and $A_1 := \|\boldsymbol{G} + \boldsymbol{\sigma} + \epsilon\mathbf{1}\|_1$. Under Assumption J.1, proportional coupling $\beta_2 = 1 - r\delta$, and the optimized step size from Theorem 7.1, the coordinate-wise clipped learner satisfies, for $T \ge \max(T_0, T_1)$ and any comparator sequence $\mathbf{u}_t \in [-D, D]^d$,*

$$\mathbb{E}[\mathrm{Regret}_T^{[\beta]}(\mathbf{u}_T)] \le O_r\left(\frac{DA_1}{\sqrt{\delta}}\right), \qquad \sum_{t=1}^{T} \mathbb{E}[\mathrm{Regret}_t^{[\beta]}(\mathbf{u}_t)] \quad \le O_r\left(\frac{TDA_1}{\sqrt{\delta}}\right).$$

Moreover, under Assumption J.2, for any $\lambda > 0$ and $\varepsilon_{\mathrm{stat}} \in (0, A_1]$, suitable choices of $\beta_1, \beta_2, D, \eta$ ensure that, for $T = \Omega_r\left(\Delta d^{1/2} \lambda^{1/2} A_1^2 \varepsilon_{\mathrm{stat}}^{-7/2}\right)$, we have $\mathbb{E}_{t\sim[T]}\mathbb{E}\left[\|\nabla F(\bar{\mathbf{x}}_t)\|_1^{[\lambda]}\right] \le O(\varepsilon_{\mathrm{stat}})$.

8 DISCUSSION

Our result shows that Adam's original bias correction creates a genuine proof obstruction in discounted-regret analysis, but not an asymptotic regret-order penalty. The obstruction is localized: the effective step sizes remain monotone, and the remaining Abel coefficient has a single-valley structure whose cost can be absorbed by the pointwise and cumulative regret quantities. Thus, within the clipped bounded-domain setting, existing guarantees can be made faithful to Adam's original bias-corrected parameterization.

Limitations. The framework-level penalty in Proposition 6.4 is intrinsic to terminal-only Abel envelopes, but whether it corresponds to a real per-iteration regret penalty for Adam is open. Two natural attack routes are (a) constructing an adversarial $\mathbf{v}_s$-sequence at the R_s-boundary (saturating $\|\mathbf{v}_s\|^2 = 2\beta_2^{2s}\sum_{n<s}\beta_2^{-2n}\|\mathbf{v}_n\|^2$), in the spirit of (Nguyen, 2025, Theorem 5), to lower-bound $\mathbb{E}[Q_T^{\mathrm{sel}}]$ in the pre-recovery regime; or (b) developing a non-Abel reformulation of the quadratic term that avoids paying the discrete valley depth altogether. We conjecture that the true per-iteration penalty, if present, is closer to the square-root terminal-envelope valley depth than to the crude $1/(1-\beta_1)$ factor used in our convenient upper bound.

Our final theorem still works in a clipped bounded domain, matching the movement-side assumption used in the existing discounted-regret analyses of Adam. Removing clipping while retaining bias correction remains an interesting future direction, but it is outside the scope of the present work.

Finally, the thresholds T_0 and T_1, while explicit and moderate, represent technical artifacts that are absent from analyses that cancel bias correction; for the default parameters $(\beta_1, \beta_2) = (0.9, 0.999)$, $T_0 = 24$ and $T_1 = 198$—negligible compared to typical training horizons. Whether both thresholds can be entirely eliminated remains open.

AI USE STATEMENT

Generative AI tools were used to assist with editing and polishing the writing of this paper for clarity and readability. All AI-assisted text was reviewed and revised by the authors, who take full responsibility for the final content of the paper.

REFERENCES

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.

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.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

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.

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*, pp. 6643–6670. PMLR, 2023.

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

Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Yusu Hong and Junhong Lin. On convergence of adam for stochastic optimization under relaxed assumptions. *Advances in Neural Information Processing Systems*, 37:10827–10877, 2024.

Ruinan Jin, Xiao Li, Yaoliang Yu, and Baoxiang Wang. A comprehensive framework for analyzing the convergence of adam: Bridging the gap with sgd. *arXiv preprint arXiv:2410.04458*, 2024.

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

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

Haochuan Li, Alexander Rakhlin, and Ali Jadbabaie. Convergence of adam under relaxed assumptions. *Advances in Neural Information Processing Systems*, 36:52166–52196, 2023.

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

Quan Nguyen. How to set β_1, β_2 in adam: An online learning perspective. *Proceedings of Machine Learning Research, ALT 2026*, 2025.

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

Francesco Orabona and Dávid Pál. Scale-free online learning. *Theoretical Computer Science*, 716: 50–69, 2018.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*, 2019.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Bohan Wang, Jingwen Fu, Huishuai Zhang, Nanning Zheng, and Wei Chen. Closing the gap between the upper bound and lower bound of adam's iteration complexity. *Advances in Neural Information Processing Systems*, 36:39006–39032, 2023a.

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

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

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

Qi Zhang, Yi Zhou, and Shaofeng Zou. Convergence guarantees for rmsprop and adam in generalized-smooth non-convex optimization with affine noise variance. *arXiv preprint arXiv:2404.01436*, 2024.

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

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

Dongruo Zhou, Jinghui Chen, Yuan Cao, Ziyan Yang, and Quanquan Gu. On the convergence of adaptive gradient methods for nonconvex optimization. *arXiv preprint arXiv:1808.05671*, 2018.

A Proof of Theorem 4.1 (Intermediate Regret)

Theorem 4.1 (Discounted regret — intermediate form). *For any $t \in [T]$ and any comparator $\mathbf{u} \in \mathcal{D}$ with $\|\mathbf{u}\| \le D$, consider the clipped Adam/FTRL learner from Section 3.1 with step sizes $\eta_s \le \eta$, any $\epsilon > 0$, and non-increasing effective step sizes $\alpha_s \ge \alpha_{s+1}$. Then the discounted regret satisfies*

$$\mathrm{Regret}_t^{[\beta]}(\mathbf{u}) \;\le\; \underbrace{\frac{\beta_1^t \|\mathbf{u}\|^2}{2\alpha_{t+1}}}_{\textit{comparator}} \;+\; \sum_{s=1}^{t} \min\Big(\underbrace{\frac{\alpha_s}{2}\,\beta_1^{t-2s}\|\mathbf{v}_s\|^2}_{\textit{quadratic increment}},\; \underbrace{\beta_1^{t-s}\|\mathbf{z}_s - \mathbf{z}_{s+1}\|\|\mathbf{v}_s\|}_{\textit{movement increment}} \Big). \tag{1}$$

Proof. 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

$$\mathbf{z}_s = \arg\min_{\mathbf{z}\in\mathcal{D}} F_s(\mathbf{z}) = \mathrm{clip}_D\left(-\alpha_s \sum_{n=1}^{s-1} \beta_1^{-n}\mathbf{v}_n \right)$$

is the clipped Adam/FTRL update.

Telescoping.

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

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 $F_{t+1}(\mathbf{z}_{t+1}) \le F_{t+1}(\mathbf{u})$.

Bounding Term B. Writing

$$\text{Term B} \le \underbrace{F_s(\mathbf{z}_s) - F_s(\mathbf{z}_{s+1})}_{\text{B.1}} + \underbrace{\langle \beta_1^{-s}\mathbf{v}_s, \mathbf{z}_s - \mathbf{z}_{s+1}\rangle}_{\text{B.2}}$$

(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 $\alpha_s \ge \alpha_{s+1}$):

Term B.1: $F_s(\mathbf{z}_s) - F_s(\mathbf{z}_{s+1}) \le -\frac{\|\mathbf{z}_s - \mathbf{z}_{s+1}\|^2}{2\alpha_s}$ by strong convexity and the variational inequality for the constrained minimizer $\mathbf{z}_s \in \arg\min_{\mathbf{z}\in\mathcal{D}} F_s(\mathbf{z})$. (R1)

Quadratic term (Young's inequality on B.2). $\langle \beta_1^{-s}\mathbf{v}_s, \mathbf{z}_s - \mathbf{z}_{s+1}\rangle \le \frac{\alpha_s}{2}\beta_1^{-2s}\|\mathbf{v}_s\|^2 + \frac{\|\mathbf{z}_s - \mathbf{z}_{s+1}\|^2}{2\alpha_s}$. (R2)

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

Movement term (Cauchy–Schwarz on B.2). $\langle \beta_1^{-s}\mathbf{v}_s, \mathbf{z}_s - \mathbf{z}_{s+1}\rangle \le \|\mathbf{z}_s - \mathbf{z}_{s+1}\| \cdot \beta_1^{-s}\|\mathbf{v}_s\|$. (R3)

Combining (R1)+(R3): Term B $\le \beta_1^{-s}\|\mathbf{z}_s - \mathbf{z}_{s+1}\|\|\mathbf{v}_s\|$.

Thus, for each s,

$$\text{Term B} \le \min\Big(\frac{\alpha_s}{2}\beta_1^{-2s}\|\mathbf{v}_s\|^2,\; \beta_1^{-s}\|\mathbf{z}_s - \mathbf{z}_{s+1}\|\|\mathbf{v}_s\| \Big).$$

Summing over s and multiplying by β_1^t gives the summand-wise bound equation 1. □

B Proof of Automatic Step-Size Monotonicity (Proposition 4.3)

Proposition B.1 (Automatic step-size monotonicity). *Let $\beta_1, \beta_2 \in (0,1)$ with $\beta_1 \le \beta_2$ and $\epsilon > 0$. With constant learning rate $\eta_s = \eta$ and the initialization $\alpha_1 := \alpha_2$, the effective step size $\alpha_s = \frac{\eta(1-\beta_1)\beta_1^{s-1}}{(1-\beta_1^{s-1})(\sqrt{b_s}+\epsilon)}$, $s \ge 2$, is non-increasing: $\alpha_s \ge \alpha_{s+1}$ for all $s \ge 1$.*

Proof. We begin with $c_s := \frac{(1-\beta_1)\beta_1^{s-1}}{(1-\beta_1^{s-1})(\sqrt{b_s}+\epsilon)}$. We show $c_s \le c_{s-1}$ for all $s \ge 2$; monotonicity of $\alpha_s = \eta\, c_s$ follows immediately.

Case $s = 2$. By the initial-point convention of Section 4.3, $c_1 := c_2$, so $c_2 \le c_1$ holds with equality.

Case $s \ge 3$. We need $c_s \le c_{s-1}$, equivalently

$$\frac{c_s}{c_{s-1}} = \frac{\beta_1(1-\beta_1^{s-2})}{1-\beta_1^{s-1}} \cdot \frac{\sqrt{b_{s-1}}+\epsilon}{\sqrt{b_s}+\epsilon} \;\le\; 1. \tag{9}$$

We bound each factor. Set $k = s-2 \ge 1$, $a = \beta_1^k$, $b = \beta_2^k$.

Bounding $\sqrt{b_s}$ from below. Unrolling the second-moment EMA and applying bias correction gives

$$b_s = \frac{1-\beta_2}{1-\beta_2^{s-1}} \sum_{n=1}^{s-1} \beta_2^{s-1-n} \|\mathbf{v}_n\|^2 = \frac{1-\beta_2}{1-\beta_2^{s-1}} \left(\beta_2 \sum_{n=1}^{s-2} \beta_2^{s-2-n} \|\mathbf{v}_n\|^2 + \|\mathbf{v}_{s-1}\|^2 \right).$$

Since

$$b_{s-1} = \frac{1-\beta_2}{1-\beta_2^{s-2}} \sum_{n=1}^{s-2} \beta_2^{s-2-n} \|\mathbf{v}_n\|^2,$$

we can rewrite this as the convex combination

$$b_s = w_s\, b_{s-1} + (1-w_s)\|\mathbf{v}_{s-1}\|^2, \qquad w_s = \frac{(1-\beta_2^{s-2})\beta_2}{1-\beta_2^{s-1}} \in (0,1),$$

where $1-w_s = (1-\beta_2)/(1-\beta_2^{s-1})$. Since $(1-w_s)\|\mathbf{v}_{s-1}\|^2 \ge 0$, we have $b_s \ge w_s\, b_{s-1}$, hence $\sqrt{b_s} \ge \sqrt{w_s}\,\sqrt{b_{s-1}}$.

Worst-case ratio. Substituting $\sqrt{b_s} \ge \sqrt{w_s}\,\sqrt{b_{s-1}}$ into equation 9:

$$\frac{c_s}{c_{s-1}} \;\le\; \frac{\beta_1(1-a)}{1-a\beta_1} \cdot \frac{\sqrt{b_{s-1}}+\epsilon}{\sqrt{w_s}\,\sqrt{b_{s-1}}+\epsilon},$$

where we used $\frac{1-\beta_1^{s-2}}{1-\beta_1^{s-1}} = \frac{1-a}{1-a\beta_1}$. The second factor $\frac{x+\epsilon}{\sqrt{w_s}\,x+\epsilon}$ (with $x = \sqrt{b_{s-1}}$) is increasing in x because $\sqrt{w_s} < 1$, so its supremum is at $x \to \infty$, giving $1/\sqrt{w_s}$.

Hence it suffices to prove the worst-case inequality

$$\frac{\beta_1^2(1-a)^2}{(1-a\beta_1)^2} \;\le\; w_s \;=\; \frac{(1-b)\beta_2}{1-b\beta_2}, \tag{10}$$

where $a = \beta_1^k$, $b = \beta_2^k$, and $k = s-2 \ge 1$.

Define

$$\phi_k(\beta) \;:=\; \frac{\beta(1-\beta^k)}{1-\beta^{k+1}} \;=\; \frac{\sum_{j=1}^{k}\beta^j}{\sum_{j=0}^{k}\beta^j} \;=\; 1 - \frac{1}{\sum_{j=0}^{k}\beta^j}.$$

Then the left side of equation 10 equals $\phi_k(\beta_1)^2$ and the right side equals $\phi_k(\beta_2)$, so it suffices to show $\phi_k(\beta_1)^2 \le \phi_k(\beta_2)$.

We use two elementary facts:

(i) $\phi_k(\beta) < 1$ for $\beta \in (0,1)$, since $\beta(1-\beta^k) < 1-\beta^{k+1} \iff \beta < 1$.

(ii) ϕ_k is strictly increasing on $(0,1)$, since $\sum_{j=0}^{k}\beta^j$ is strictly increasing (its derivative $\sum_{j=1}^{k} j\beta^{j-1}$ is positive) and $\phi_k = 1 - 1/\sum_{j=0}^{k}\beta^j$.

Combining, $\phi_k(\beta_1)^2 \overset{(i)}{\le} \phi_k(\beta_1) \overset{(ii)}{\le} \phi_k(\beta_2)$, where the second inequality uses $\beta_1 \le \beta_2$. This establishes equation 10 and hence $c_s \le c_{s-1}$ for all $s \ge 3$.

Combining with the $s = 2$ case, $\{c_s\}$ is non-increasing and $\alpha_s = \eta\, c_s$ is non-increasing. □

C Proof of Proposition 5.1

Proof. Let $\mathcal{S}_t := \{s \in \{2, \dots, t\} : \mathcal{G}_s\}$ and define $A_s := \beta_1^{-s}\|\mathbf{v}_s\|$. If $r < s$ are consecutive elements of $\mathcal{S}_t$, then $\mathcal{G}_s$ implies

$$\beta_2^{-2s}\|\mathbf{v}_s\|^2 > 2\sum_{n=1}^{s-1}\beta_2^{-2n}\|\mathbf{v}_n\|^2 \geq 2\beta_2^{-2r}\|\mathbf{v}_r\|^2,$$

hence $\beta_2^{-s}\|\mathbf{v}_s\| > \sqrt{2}\,\beta_2^{-r}\|\mathbf{v}_r\|$. Since $\beta_1 \leq \beta_2$, this yields

$$A_s = \Big(\frac{\beta_2}{\beta_1}\Big)^s \beta_2^{-s}\|\mathbf{v}_s\| > \sqrt{2}\,\Big(\frac{\beta_2}{\beta_1}\Big)^r \beta_2^{-r}\|\mathbf{v}_r\| = \sqrt{2}\,A_r.$$

Therefore the selected sequence $\{A_s : s \in \mathcal{S}_t\}$ grows by at least a factor $\sqrt{2}$, so

$$\sum_{s\in\mathcal{S}_t} A_s \leq (2+\sqrt{2})\max_{s\in\mathcal{S}_t} A_s,$$

with the convention that the sum and maximum are zero when $\mathcal{S}_t = \varnothing$. Using $\|\mathbf{z}_s - \mathbf{z}_{s+1}\| \leq 2D$,

$$\begin{aligned} M_t^{\mathrm{sel}} &\leq 2D\Big(\beta_1^{t-1}\|\mathbf{v}_1\| + \sum_{s\in\mathcal{S}_t}\beta_1^{t-s}\|\mathbf{v}_s\|\Big) \\ &= 2D\Big(\beta_1^t A_1 + \beta_1^t\sum_{s\in\mathcal{S}_t} A_s\Big) \leq 2(3+\sqrt{2})D\max_{1\leq s\leq t}\beta_1^{t-s}\|\mathbf{v}_s\|. \end{aligned}$$

Next,

$$\max_{1\leq s\leq t}\beta_1^{t-s}\|\mathbf{v}_s\| \leq \sqrt{\sum_{s=1}^{t}\beta_1^{2(t-s)}\|\mathbf{v}_s\|^2}.$$

Taking expectation and applying Jensen's inequality together with $\mathbb{E}\|\mathbf{v}_s\|^2 \leq (G+\sigma)^2$,

$$\mathbb{E}\Big[\max_{1\leq s\leq t}\beta_1^{t-s}\|\mathbf{v}_s\|\Big] \leq \sqrt{\sum_{s=1}^{t}\beta_1^{2(t-s)}\mathbb{E}\|\mathbf{v}_s\|^2} \leq \frac{G+\sigma}{\sqrt{1-\beta_1^2}}.$$

Combining the two displays yields the first bound in Proposition 5.1. Summing the same estimate over $t = 1, \dots, T$ gives the second bound. □

D Proof of Lemma 6.1 (Reformulation)

Lemma D.1 (Selected reformulation)**.** *For $s \geq 2$, suppose $\eta_s \leq \eta$ and define $\mathcal{R}_s := \Big\{\beta_2^{-2s}\|\mathbf{v}_s\|^2 \leq 2\sum_{n=1}^{s-1}\beta_2^{-2n}\|\mathbf{v}_n\|^2\Big\}$. Define*

$$H_s := \sqrt{3}\eta\,\frac{(1-\beta_1)\sqrt{1-\beta_2^{s-1}}}{(1-\beta_1^{s-1})\sqrt{1-\beta_2}}\,\frac{\beta_2^{s+1}}{\beta_1^{s+1}\beta_2^{s/2}}, \qquad I_s := \sqrt{\textstyle\sum_{n=1}^{s}\beta_2^{-n-1}\|\mathbf{v}_n\|^2}, \qquad I_0 := 0.$$

Then $\frac{\alpha_s}{2}\,\beta_1^{-2s}\|\mathbf{v}_s\|^2\mathbf{1}_{\mathcal{R}_s} \leq H_s\big(I_s - I_{s-1}\big)\mathbf{1}_{\mathcal{R}_s} \leq H_s\big(I_s - I_{s-1}\big)$.

Proof. Fix $s \geq 2$. If $\mathcal{R}_s$ does not hold, the claimed indicator inequality is trivial. On $\mathcal{R}_s$, the stronger weighted reverse condition implies the reverse condition needed below:

$$\beta_2^{-s-1}\|\mathbf{v}_s\|^2 = \beta_2^{s-1}\beta_2^{-2s}\|\mathbf{v}_s\|^2 \leq 2\sum_{n=1}^{s-1}\beta_2^{s-1-2n}\|\mathbf{v}_n\|^2 \leq 2\sum_{n=1}^{s-1}\beta_2^{-n-1}\|\mathbf{v}_n\|^2.$$

The last inequality uses $n \le s-1$ and $0 < \beta_2 < 1$. We then re-formulate $\frac{\alpha_s}{2}\beta_1^{-2s}\|\mathbf{v}_s\|^2$ as:

$$
\begin{aligned}
&\frac{\alpha_s}{2}\beta_1^{-2s}\|\mathbf{v}_s\|^2 \\
&\overset{(a)}{\le} \frac{\eta}{2}\frac{1-\beta_1}{1-\beta_1^{s-1}}\frac{\beta_1^{-s-1}}{\sqrt{b_s}}\|\mathbf{v}_s\|^2 \\
&= \frac{\eta}{2}\frac{1-\beta_1}{1-\beta_1^{s-1}}\frac{\sqrt{1-\beta_2^{s-1}}}{\sqrt{1-\beta_2}}\frac{1}{\sqrt{\beta_2^s}}\frac{\beta_1^{-s-1}\|\mathbf{v}_s\|^2}{\sqrt{\sum_{n=1}^{s-1}\beta_2^{-n-1}\|\mathbf{v}_n\|^2}} \\
&\overset{(b)}{\le} \frac{\eta}{2}\frac{1-\beta_1}{1-\beta_1^{s-1}}\frac{\sqrt{1-\beta_2^{s-1}}}{\sqrt{1-\beta_2}}\frac{\beta_1^{-s-1}}{\beta_2^{-s-1}}\frac{1}{\sqrt{\beta_2^s}}\frac{\sqrt{3}\,\beta_2^{-s-1}\|\mathbf{v}_s\|^2}{\sqrt{\beta_2^{-s-1}\|\mathbf{v}_s\|^2+\sum_{n=1}^{s-1}\beta_2^{-n-1}\|\mathbf{v}_n\|^2}} \\
&\overset{(c)}{\le} \frac{\eta}{2}\frac{1-\beta_1}{\sqrt{1-\beta_2}}\frac{\sqrt{1-\beta_2^{s-1}}}{1-\beta_1^{s-1}}\frac{\beta_1^{-s-1}}{\beta_2^{-s-1}}\frac{1}{\sqrt{\beta_2^s}}\cdot 2\sqrt{3}\underbrace{\left(\sqrt{\sum_{n=1}^{s}\beta_2^{-n-1}\|\mathbf{v}_n\|^2}-\sqrt{\sum_{n=1}^{s-1}\beta_2^{-n-1}\|\mathbf{v}_n\|^2}\right)}_{=I_s-I_{s-1}} \\
&= \underbrace{\sqrt{3}\eta\,\frac{(1-\beta_1)\sqrt{1-\beta_2^{s-1}}}{(1-\beta_1^{s-1})\sqrt{1-\beta_2}}\,\frac{\beta_2^{s/2+1}}{\beta_1^{s+1}}}_{=H_s}(I_s-I_{s-1})
\end{aligned}
$$

where (a) uses $\eta_s \le \eta$ and drops ϵ; (b) uses $\beta_2^{-s-1}\|\mathbf{v}_s\|^2 \le 2\sum_{n=1}^{s-1}\beta_2^{-n-1}\|\mathbf{v}_n\|^2$, which follows from $\mathcal{R}_s$ as shown above, and hence $\sum_{n=1}^{s-1}\beta_2^{-n-1}\|\mathbf{v}_n\|^2 \ge \frac{1}{2}\beta_2^{-s-1}\|\mathbf{v}_s\|^2$; (c) uses $\frac{x}{\sqrt{x+y}} \le 2(\sqrt{x+y}-\sqrt{y})$ for $x, y > 0$. Multiplying by $\mathbf{1}_{\mathcal{R}_s}$ gives the first displayed inequality, and dropping the indicator on the nonnegative right-hand side gives the second. □

E Proof of Lemma 6.2 (Single-Valley Structure)

Lemma E.1 (Single-valley structure)**.** *Let $0 < \beta_1 < \beta_2 < 1$, and let $\{H_s\}_{s\ge 2}$ be defined as in Lemma 6.1. Then there exists a unique $s_\star \ge 2$ such that $H_{s+1} < H_s$ for $2 \le s < s_\star$ and $H_{s+1} \ge H_s$ for $s \ge s_\star$.*

Proof. Set $\lambda_i := -\log\beta_i > 0$ and $\tau := \lambda_1/\lambda_2 > 1$ (strict since $\beta_1 < \beta_2$). Let $h(s) := \log H_s$ extended smoothly to $s \ge 2$. From the definition of H_s in Lemma 6.1, with $\log\rho = \lambda_1 - \lambda_2/2 = \lambda_2(\tau - \frac{1}{2})$:

$$h'(s) \;=\; \log\rho + \frac{\lambda_2/2}{e^{\lambda_2(s-1)}-1} - \frac{\lambda_1}{e^{\lambda_1(s-1)}-1} \;=\; \frac{\lambda_2}{2}\left[(2\tau-1) + \frac{1}{e^y-1} - \frac{2\tau}{e^{\tau y}-1}\right],$$

with $y := \lambda_2(s-1)$. Multiplying through by $2(e^y-1)(e^{\tau y}-1)/\lambda_2 > 0$, the sign of h' equals that of

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

We have $G(0) = 0$, $G(\infty) = +\infty$ (leading term $(2\tau-1)e^{(\tau+1)y}$, coefficient positive), and $G'(0) = -\tau < 0$.

G' has a unique positive zero. Factor $G'(y) = e^y\,Q(y)$ with $Q(y) := (2\tau-1)(\tau+1)e^{\tau y} - 2\tau(\tau-1)e^{(\tau-1)y} - (4\tau-1)$. Then $Q(0) = -\tau$, $Q(\infty) = +\infty$, and $Q'(y) = \tau e^{(\tau-1)y}[(2\tau-1)(\tau+1)e^y - 2(\tau-1)^2]$. Since $(2\tau-1)(\tau+1) - 2(\tau-1)^2 = 5\tau - 3 > 0$ for $\tau > 1$ and $e^y > 1$ for $y > 0$, $Q'(y) > 0$, so Q is strictly increasing with a unique positive zero, and therefore G' has a unique positive zero.

Combined with $G(0) = 0$, $G'(0) < 0$, and $G(\infty) = +\infty$: G decreases below zero, attains a minimum, then increases to $+\infty$, crossing zero at a unique $y_0 > 0$. With $s_0 := 1 + y_0/\lambda_2$, this gives $h'(s) < 0$ for $s < s_0$ and $h'(s) > 0$ for $s > s_0$.

Discrete lifting. For real $u \ge 2$, define $\Delta(u) := h(u+1) - h(u)$; the discrete increments are $\Delta_s = \Delta(s)$. If $u + 1 \le s_0$, then $h' < 0$ on $[u, u+1]$, so $\Delta(u) < 0$. If $u \ge s_0$, then $h' > 0$ on

$[u, u+1]$, so $\Delta(u) > 0$. The only remaining values of u lie in the transition window $(s_0 - 1, s_0)$, whose length is one. On this window,

$$\Delta'(u) = h'(u+1) - h'(u) > 0,$$

because $u < s_0 < u + 1$. Thus $\Delta(u)$ is strictly increasing through the transition window, and the integer sequence $\{\Delta_s\}_{s\geq 2}$ can change sign at most once. Since $\Delta_s \to \log \rho > 0$, the set $\{s \geq 2 : \Delta_s \geq 0\}$ is nonempty. Set $s_\star := \min\{s \geq 2 : \Delta_s \geq 0\}$. Then $\Delta_s < 0$ for $2 \leq s < s_\star$ and $\Delta_s \geq 0$ for $s \geq s_\star$, equivalently $H_{s+1} < H_s$ for $2 \leq s < s_\star$ and $H_{s+1} \geq H_s$ for $s \geq s_\star$. □

F Proof of Lemma 6.3 (Properties of H_s)

Lemma F.1 (Properties of H_s)**.** *Let $0 < \beta_1 \leq \beta_2 < 1$. Write $\rho := \sqrt{\beta_2}/\beta_1 > 1$ and $C_L := \frac{\sqrt{3}\,\eta\,(1-\beta_1)\,\beta_2}{\beta_1}$. Define $C_{\mathrm{H}} := \frac{\sqrt{6\beta_2}\,(1-\beta_1)}{\beta_1(1-\beta_2)}$.* (i) ***Lower bound:*** *$H_s \geq C_L \rho^s$ for all $s \geq 2$.* (ii) ***Explicit H_2:*** *$H_2 = \frac{\sqrt{3}\,\eta\,\beta_2^2}{\beta_1^3} = \frac{C_L \rho^2}{1-\beta_1}$.* (iii) ***Boundary:*** *We set $H_1 := 0$ by convention to drop the $s = 1$ term.* (iv) ***Unconditional expectation bound:*** *Under Assumption 3.1, for every $s \geq 2$, $\mathbb{E}[\beta_1^s H_s I_s] \leq C_{\mathrm{H}}\,\eta\,(G + \sigma)$.*

Proof. (i) $\frac{(1-\beta_1)\sqrt{1-\beta_2^{s-1}}}{(1-\beta_1^{s-1})\sqrt{1-\beta_2}} \geq 1 - \beta_1$ and $\frac{\beta_2^{s+1}}{\beta_1^{s+1}\beta_2^{s/2}} = \frac{\beta_2\,\rho^s}{\beta_1}$, hence $H_s \geq \sqrt{3}\eta(1-\beta_1)\beta_2\rho^s/\beta_1 = C_L\rho^s$.

(ii) At $s = 2$: $\sqrt{1-\beta_2}/\sqrt{1-\beta_2} = 1$, $(1-\beta_1)/(1-\beta_1) = 1$, so $H_2 = \sqrt{3}\eta\,\beta_2^3/(\beta_1^3\beta_2) = \sqrt{3}\eta\beta_2^2/\beta_1^3$.

(iii) The formula defining H_s is used for $s \geq 2$. We set $H_1 := 0$ by convention so that Abel summation may start at $s = 1$ without adding a boundary contribution.

(iv) We expand $\beta_1^s H_s I_s$ directly from the definitions (Lemma 6.1) and bound it without conditioning on the reverse-growth event.

Step 1 (expand $\beta_1^s H_s I_s$). From definitions of H_s in Lemma 6.1, multiplying H_s by β_1^s and using $\frac{\beta_2^{s+1}}{\beta_1^{s+1}\beta_2^{s/2}} = \frac{\beta_2^{s/2+1}}{\beta_1}$:

$$\beta_1^s H_s = \frac{\sqrt{3}\,\eta(1-\beta_1)}{\beta_1} \cdot \frac{\sqrt{1-\beta_2^{s-1}}}{(1-\beta_1^{s-1})\sqrt{1-\beta_2}} \cdot \beta_2^{s/2+1}\,.$$

Absorbing $\beta_2^{s/2+1}$ into I_s via $\beta_2^{s+2} \cdot \beta_2^{-n-1} = \beta_2^{s-n+1} = \beta_2 \cdot \beta_2^{s-n}$:

$$\beta_2^{s/2+1}\, I_s = \sqrt{\beta_2}\,\sqrt{\textstyle\sum_{n=1}^{s} \beta_2^{s-n} \|\mathbf{v}_n\|^2}\,.$$

Therefore

$$\beta_1^s H_s I_s = \frac{\sqrt{3\beta_2}\,\eta(1-\beta_1)}{\beta_1\sqrt{1-\beta_2}} \cdot \frac{\sqrt{1-\beta_2^{s-1}}}{1-\beta_1^{s-1}} \cdot \sqrt{\textstyle\sum_{n=1}^{s} \beta_2^{s-n} \|\mathbf{v}_n\|^2}\,. \tag{11}$$

Step 2 (expectation via Jensen). By Jensen's inequality and Assumption 3.1,

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

Taking expectations in equation 11 gives

$$\mathbb{E}[\beta_1^s H_s I_s] \leq \frac{\sqrt{3\beta_2}\,\eta(1-\beta_1)}{\beta_1(1-\beta_2)} \cdot \frac{\sqrt{(1-\beta_2^{s-1})(1-\beta_2^s)}}{1-\beta_1^{s-1}}\,(G + \sigma).$$

For $s \geq 2$, we have

$$1 - \beta_2^s \leq 2(1 - \beta_2^{s-1}),$$

equivalently $\sqrt{1-\beta_2^s} \leq \sqrt{2}\sqrt{1-\beta_2^{s-1}}$. Using also $\frac{1-\beta_2^{s-1}}{1-\beta_1^{s-1}} \leq 1$ (since $\beta_1 \leq \beta_2$), we obtain

$$\mathbb{E}[\beta_1^s H_s I_s] \;\leq\; C_{\mathrm{H}}\,\eta\,(G+\sigma)\,.$$

□

G Proof of Proposition 6.4 (Framework-Level Penalty)

Proposition G.1 (Per-iteration limit of the Abel framework)**.** *Let $0 < \beta_1 < \beta_2 < 1$. Let H_s and I_s be defined as in Lemma 6.1, with $H_1 := 0$, and let $R(T) := \sum_{s=1}^{T-1}(H_s - H_{s+1})I_s$.*

(i) ***Threshold for the clean per-iteration Abel envelope.*** *If $T \geq T_0 := 2 + \left\lceil \frac{\log(1/(1-\beta_1))}{\log(\sqrt{\beta_2}/\beta_1)} \right\rceil$, then $R(T) \leq 0$, and therefore $\beta_1^T \sum_{s=1}^{T} H_s(I_s - I_{s-1}) \leq \beta_1^T H_T I_T$.*

(ii) ***Penalized upper bound for $T < T_0$.*** *In the pre-recovery regime $2 \leq T < T_0$, combining Abel summation with the single-valley structure (Lemma 6.2) and the ratio $H_2/H_s \leq 1/(1-\beta_1)$ (Lemma 6.3*(i)-(ii)*) yields the* penalized *pathwise envelope bound*

$$\beta_1^T \sum_{s=1}^{T} H_s(I_s - I_{s-1}) \;\leq\; \frac{1}{1-\beta_1}\,\beta_1^T H_T I_T. \tag{2}$$

(iii) ***Sharpness and magnitude of terminal-term Abel bounds.*** *For any fixed $T \geq 2$, the smallest constant Γ_T for which $\sum_{s=1}^{T} H_s(I_s - I_{s-1}) \leq \Gamma_T H_T I_T$ holds for every non-decreasing sequence $0 = I_0 \leq I_1 \leq \cdots \leq I_T$ is $\Gamma_T = \max_{1 \leq s \leq T} \frac{H_s}{H_T}$. Consequently, for every $2 \leq T \leq s_\star$, $\Gamma_T = \frac{H_2}{H_T}$, and at the bottom of the valley the necessary terminal-term constant is exactly $H_2/H_{s_\star}$. Moreover, if $r \in (0,1)$ is fixed and $1 - \beta_2 = r(1-\beta_1)$, then, as $\beta_1 \uparrow 1$, $\Gamma_{s_\star} = \frac{H_2}{H_{s_\star}} = \Theta_r\Big(\frac{1}{\sqrt{1-\beta_1}}\Big)$. Here $\Theta_r(\cdot)$ hides constants that may depend on r.*

Proof. Write $\rho := \sqrt{\beta_2}/\beta_1 > 1$ and $R(T) := \sum_{s=1}^{T-1}(H_s - H_{s+1})I_s$.

(i) Threshold condition. Abel summation gives, for any $T \geq 1$,

$$\beta_1^T \sum_{s=1}^{T} H_s(I_s - I_{s-1}) \;=\; \beta_1^T\big[H_T I_T + R(T)\big]. \tag{12}$$

To show $R(T) \leq 0$ when $H_T \geq H_2$: with $\sum_{s=1}^{k}(H_s - H_{s+1}) = H_1 - H_{k+1} = -H_{k+1}$ (using $H_1 = 0$), Abel summation rewrites

$$R(T) \;=\; -H_T I_{T-1} \;+\; \sum_{s=1}^{T-2} H_{s+1}(I_{s+1} - I_s).$$

If $H_T \geq H_2$, single-valley structure gives $H_{s+1} \leq \max(H_2, H_{T-1}) \leq H_T$ for $s \in [1, T-2]$, hence

$$R(T) \;\leq\; -H_T I_{T-1} + H_T(I_{T-1} - I_1) \;=\; -H_T I_1 \;\leq\; 0.$$

By Lemma 6.3(i)–(ii), $H_T \geq C_L \rho^T$ and $H_2 = C_L\rho^2/(1-\beta_1)$, so $H_T \geq H_2$ follows from $\rho^{T-2} \geq 1/(1-\beta_1)$, i.e., $T \geq T_0 := 2 + \lceil \log(1/(1-\beta_1))/\log\rho \rceil$.

(ii) Penalized upper bound for $T < T_0$. Fix T with $2 \leq T < T_0$. We prove the penalized bound equation 2 by chaining three upper bounds: Abel summation, single-valley rearrangement, and the valley-depth upper bound $H_2/H_s \leq 1/(1-\beta_1)$ derived from Lemma 6.3*(i)-(ii)*. The result is a statement about the Abel upper envelope; any selected quadratic term bounded by this envelope may be strictly smaller for a given trajectory.

Step 1 (Abel). Applying Abel summation (with $I_0 = 0$), then multiplying by β_1^T, yields equation 12.

Step 2 (case split and single-valley rearrangement). If $H_T \geq H_2$, the argument from Part (i) gives $R(T) \leq 0$. Therefore

$$\beta_1^T \big(H_T I_T + R(T)\big) \leq \beta_1^T H_T I_T \leq \frac{\beta_1^T H_T I_T}{1-\beta_1},$$

so the desired penalized bound already holds. It remains to consider the case $H_T < H_2$. By Lemma 6.3(iii), $H_1 = 0$, so the $s = 1$ term of $R(T)$ equals $-H_2 I_1 \leq 0$ and can be dropped in any upper bound. By Lemma 6.2, $(H_s - H_{s+1}) \geq 0$ for $2 \leq s < s_\star$ and $(H_s - H_{s+1}) \leq 0$ for $s \geq s_\star$.

For $T \leq s_\star$, all non-dropped summands are non-negative and $I_s \leq I_T$:

$$R(T) \;\leq\; \sum_{s=2}^{T-1} (H_s - H_{s+1}) I_s \;\leq\; I_T \sum_{s=2}^{T-1} (H_s - H_{s+1}) \;=\; I_T (H_2 - H_T).$$

For $T > s_\star$, split the sum at $s_\star$. The descending part contributes at most $\sum_{s=2}^{s_\star - 1} (H_s - H_{s+1}) I_s \leq I_{s_\star}(H_2 - H_{s_\star})$. The ascending part has non-positive summands with $I_s \geq I_{s_\star}$, so replacing each I_s by $I_{s_\star}$ yields a larger (less negative) upper bound: $\sum_{s=s_\star}^{T-1} (H_s - H_{s+1}) I_s \leq I_{s_\star} \sum_{s=s_\star}^{T-1} (H_s - H_{s+1}) = I_{s_\star}(H_{s_\star} - H_T)$. Adding: $R(T) \leq I_{s_\star}(H_2 - H_T) \leq I_T (H_2 - H_T)$.

Step 3 (valley-depth conversion). By Lemma 6.3*(i)-(ii)*, $H_2 = C_L \rho^2/(1-\beta_1)$ and $H_T \geq C_L \rho^T$, so for $T \geq 2$: $H_2/H_T \leq \rho^{2-T}/(1-\beta_1) \leq 1/(1-\beta_1)$ (since $\rho > 1$). Hence $H_2 - H_T \leq H_T\big(\frac{1}{1-\beta_1} - 1\big) = \frac{\beta_1}{1-\beta_1} H_T$. Combining Steps 1–3:

$$\beta_1^T \big(H_T I_T + R(T)\big) \;\leq\; \beta_1^T \big(H_T I_T + (H_2 - H_T) I_T\big) \;\leq\; \beta_1^T H_T I_T \Big(1 + \tfrac{\beta_1}{1-\beta_1}\Big) \;=\; \frac{\beta_1^T H_T I_T}{1-\beta_1},$$

which is equation 2. The $1/(1-\beta_1)$ factor originates from Step 3, the only step that invokes the valley-depth bound.

(iii) Sharpness and magnitude of terminal-term Abel bounds. Fix $T \geq 2$ and define

$$\Gamma_T^\star := \max_{1 \leq s \leq T} \frac{H_s}{H_T}.$$

For any non-decreasing sequence $0 = I_0 \leq I_1 \leq \cdots \leq I_T$,

$$\sum_{s=1}^{T} H_s (I_s - I_{s-1}) \leq \Gamma_T^\star H_T \sum_{s=1}^{T} (I_s - I_{s-1}) = \Gamma_T^\star H_T I_T.$$

Therefore the optimal terminal-term constant satisfies $\Gamma_T \leq \Gamma_T^\star$.

For the reverse inequality, let $r \in \arg\max_{1 \leq s \leq T} H_s$ and choose the step sequence

$$I_s := \begin{cases} 0, & s < r, \\ 1, & s \geq r. \end{cases}$$

Then $\sum_{s=1}^{T} H_s (I_s - I_{s-1}) = H_r$ and $I_T = 1$, so any valid terminal-term bound with constant Γ must have

$$H_r \leq \Gamma H_T, \qquad \text{hence} \qquad \Gamma \geq \frac{H_r}{H_T} = \Gamma_T^\star.$$

Thus $\Gamma_T = \Gamma_T^\star$.

Finally, when $2 \leq T \leq s_\star$, the single-valley structure gives $H_2 \geq H_s$ for every $2 \leq s \leq T$ and $H_1 = 0$, so $\Gamma_T = H_2/H_T$. At $T = s_\star$, this is exactly the discrete valley depth $H_2/H_{s_\star}$.

Magnitude in the proportional-coupling regime. Write $a := 1 - \beta_1$ and fix $r \in (0,1)$ with $\beta_2 = 1 - ra$. The exact value in Lemma 6.3*(ii)* gives $H_2 = \sqrt{3}\eta\, \beta_2^2/\beta_1^3 = \Theta(\eta)$. For $s \geq 2$, with $x_s := a(s-1)$, the definition of H_s can be written as

$$H_s = \sqrt{3}\eta\, \frac{\beta_2}{\beta_1} \sqrt{\frac{a}{r}}\, \rho^s \frac{\sqrt{1-\beta_2^{s-1}}}{1-\beta_1^{s-1}}, \qquad \rho = \frac{\sqrt{\beta_2}}{\beta_1}.$$

The standard estimates, uniformly for x_s in compact subsets of $(0,\infty)$,

$$(1-a)^{x_s/a} = e^{-x_s+O(a)}, \qquad (1-ra)^{x_s/a} = e^{-rx_s+O(a)}, \qquad \rho^{x_s/a} = e^{(1-r/2)x_s+O(a)}$$

show that

$$\frac{H_s}{\eta\sqrt{a}} \asymp_r F_r(x_s), \qquad F_r(x) := e^{(1-r/2)x}\frac{\sqrt{1-e^{-rx}}}{\sqrt{r}\,(1-e^{-x})}.$$

The function F_r is continuous on $(0,\infty)$, satisfies $F_r(x) \asymp_r x^{-1/2}$ as $x \downarrow 0$ and $F_r(x) \asymp_r e^{(1-r/2)x}$ as $x \to \infty$. Hence F_r has a finite positive minimum on $(0,\infty)$. The endpoint estimates give a matching lower bound outside a fixed compact interval, while the grid spacing of $x_s = a(s-1)$ tends to zero and lets us choose an integer s within $O(a)$ of a minimizer of F_r. Therefore

$$H_{s_\star} = \min_{s\ge 2} H_s = \Theta_r(\eta\sqrt{a}).$$

Combining $H_2 = \Theta(\eta)$ with $H_{s_\star} = \Theta_r(\eta\sqrt{a})$ yields

$$\Gamma_{s_\star} = \frac{H_2}{H_{s_\star}} = \Theta_r(a^{-1/2}) = \Theta_r\left(\frac{1}{\sqrt{1-\beta_1}}\right).$$

□

H Proof of Proposition 6.5 (Discounted Gradient Bounds)

Proposition H.1 (Selected reverse-growth quadratic bounds)**.** *Under Assumption 3.1, let $\beta_1, \beta_2 \in (0,1)$ with $\beta_1 < \beta_2$ and $\eta_s \le \eta$. Define C_{H} as in Lemma 6.3, for $s \ge 2$, define $\mathcal{R}_s$ as in Lemma 6.1, and for $t \ge 1$ define*

$$Q_t^{\mathrm{sel}} := \beta_1^t \sum_{s=2}^{t} \frac{\alpha_s}{2}\,\beta_1^{-2s}\|\mathbf{v}_s\|^2 \mathbf{1}_{\mathcal{R}_s} \qquad T_0 := 2 + \left\lceil \frac{\log(1/(1-\beta_1))}{\log(\sqrt{\beta_2}/\beta_1)} \right\rceil, \qquad T_1 := \left\lceil \frac{\beta_1(T_0-2)}{1-\beta_1} \right\rceil.$$

Part I (Pointwise). *For every $T \ge T_0$:*

$$\boxed{\mathbb{E}\Big[Q_T^{\mathrm{sel}}\Big] \le C_{\mathrm{H}}\,\eta\,(G+\sigma).} \tag{3}$$

Part II (Cumulative). *For every $T \ge T_1$:*

$$\boxed{\mathbb{E}\Big[\sum_{t=1}^{T} Q_t^{\mathrm{sel}}\Big] \le 2C_{\mathrm{H}}\,\eta\,(G+\sigma)\,T.} \tag{4}$$

Proof. **Step 1 (Selected inner Abel envelope — shared).** By Lemma 6.1, for every $t \ge 1$,

$$\sum_{s=2}^{t} \frac{\alpha_s}{2}\beta_1^{-2s}\|\mathbf{v}_s\|^2\mathbf{1}_{\mathcal{R}_s} \le \sum_{s=1}^{t} H_s(I_s - I_{s-1}) = H_t I_t + \sum_{s=1}^{t-1}(H_s - H_{s+1})I_s\,, \tag{13}$$

where the equality is Abel summation with $I_0 = 0$ and $H_1 = 0$.

— Part I (Pointwise) —

Step 2 (Pathwise bound). Specializing equation 13 to $t = T$ and multiplying by β_1^T:

$$Q_T^{\mathrm{sel}} \;\le\; \beta_1^T\big[H_T I_T + R(T)\big], \quad R(T) := \sum_{s=1}^{T-1}(H_s - H_{s+1})I_s.$$

By Proposition 6.4*(i)*, for $T \ge T_0$ we have $R(T) \le 0$, and hence

$$Q_T^{\mathrm{sel}} \;\le\; \beta_1^T\, H_T\, I_T. \tag{14}$$

Step 3 (Expectation). Taking expectation in equation 14 and applying Lemma 6.3(iv), $\mathbb{E}[\beta_1^T H_T I_T] \le C_{\mathrm{H}}\,\eta\,(G+\sigma)$:

$$\mathbb{E}\big[Q_T^{\mathrm{sel}}\big] \;\le\; C_{\mathrm{H}}\,\eta\,(G+\sigma),$$

which is equation 3. □ (Part I)

— Part II (Cumulative) —

Fix $T \geq 2$ and define $\widetilde{W}_s := \beta_1^{s+1} - \beta_1^{T+1} > 0$ for $1 \leq s \leq T-1$.

Step 4 (Outer summation). Multiply equation 13 by β_1^t, sum over $t = 1, \ldots, T$, and exchange summation order:

$$\sum_{t=1}^{T} Q_t^{\text{sel}} \;\leq\; \sum_{t=1}^{T} \beta_1^t H_t I_t - \frac{\mathcal{A}}{1-\beta_1}\,, \tag{15}$$

where $\mathcal{A} := \sum_{s=1}^{T-1} (H_{s+1} - H_s) I_s \widetilde{W}_s$.

Step 5 (Absorbing the remainder). We bound the contribution of $-\mathcal{A}/(1-\beta_1)$ directly. By Lemma 6.3(iii), $H_1 = 0$, so the $s = 1$ term in $\mathcal{A}$ equals $H_2 I_1 \widetilde{W}_1 \geq 0$. By Lemma 6.2, $(H_{s+1} - H_s) \geq 0$ for $s \geq s_\star$. Thus the only negative contributions come from $2 \leq s < s_\star$:

$$-\mathcal{A} \;\leq \sum_{s=2}^{\min(T-1, s_\star - 1)} (H_s - H_{s+1})\, I_s\, \widetilde{W}_s.$$

Since $H_{s+1} \geq 0$, $H_s - H_{s+1} \leq H_s$, and $\widetilde{W}_s \leq \beta_1^{s+1}$, so

$$-\mathcal{A} \;\leq\; \beta_1 \sum_{s=2}^{s_\star - 1} \beta_1^s\, H_s\, I_s.$$

Dividing by $(1-\beta_1)$ and combining with equation 15:

$$\sum_{t=1}^{T} Q_t^{\text{sel}} \;\leq\; \sum_{t=1}^{T} \beta_1^t H_t I_t + \frac{\beta_1}{1-\beta_1} \sum_{s=2}^{s_\star - 1} \beta_1^s\, H_s\, I_s, \tag{16}$$

a pathwise bound valid for all $T \geq 1$.

Step 6 (Expectation and threshold). By Lemma 6.3(iv), $\mathbb{E}[\beta_1^t H_t I_t] \leq C_{\text{H}}\, \eta\, (G + \sigma)$ for every $t \geq 2$, while the $t = 1$ term is zero because $H_1 = 0$. The first sum in equation 16 has T terms and the second at most $s_\star - 2$ terms:

$$\mathbb{E}\left[\sum_{t=1}^{T} Q_t^{\text{sel}}\right] \;\leq\; C_{\text{H}}\, \eta\, (G+\sigma) \left(T + \frac{\beta_1 (s_\star - 2)}{1-\beta_1}\right).$$

Claim: $s_\star \leq T_0$. If $s_\star > T_0$, then H is strictly decreasing on $[2, s_\star] \supset [2, T_0]$, so $H_{T_0} < H_2$, contradicting $H_{T_0} \geq H_2$ (Proposition 6.4*(i)*). Hence $s_\star \leq T_0$, so $\frac{\beta_1(s_\star - 2)}{1-\beta_1} \leq \frac{\beta_1(T_0 - 2)}{1-\beta_1}$.

For $T \geq T_1 := \left\lceil \frac{\beta_1(T_0-2)}{1-\beta_1} \right\rceil$, this gives $\frac{\beta_1(s_\star - 2)}{1-\beta_1} \leq T$, yielding

$$\mathbb{E}\left[\sum_{t=1}^{T} Q_t^{\text{sel}}\right] \;\leq\; 2\, C_{\text{H}}\, \eta\, (G+\sigma)\, T,$$

which is equation 4. □ (Part II) □

I Stationarity Consequences

The assumptions, DTNC algorithm, and stationarity definition in this section follow the standard online-to-non-convex framework (Cutkosky et al., 2023; Zhang & Cutkosky, 2024; Ahn & Cutkosky, 2024); we restate them to fix notation and to make the conversion from our regret bound explicit.

Assumption I.1 (Function and stochastic-gradient oracle)**.** Let $F : \mathbb{R}^d \to \mathbb{R}$ be differentiable and let $\Delta := F(\mathbf{x}_0) - \inf_{\mathbf{x}} F(\mathbf{x}) < \infty$. Assume:

1. **Well-behavedness.** For all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,
$$F(\mathbf{y}) - F(\mathbf{x}) = \int_0^1 \langle \nabla F(\mathbf{x} + t(\mathbf{y} - \mathbf{x})),\, \mathbf{y} - \mathbf{x} \rangle \, dt.$$
2. **Bounded gradient.** $\|\nabla F(\mathbf{x})\| \le G$ for all $\mathbf{x}$.
3. **Unbiased stochastic gradients with bounded variance.** For $\mathbf{g} \leftarrow \mathrm{StoGrad}(\mathbf{x}, r)$,
$$\mathbb{E}[\mathbf{g}] = \nabla F(\mathbf{x}), \qquad \mathbb{E}[\|\mathbf{g} - \nabla F(\mathbf{x})\|^2] \le \sigma^2.$$

Algorithm 1 Discounted-to-non-convex conversion (DTNC) algorithm

Input: Initial point $\mathbf{x}_0$, horizon T, online learner $\mathcal{A}$ outputting increments $\mathbf{z}_t$, and discount factor $\beta_1 \in (0, 1)$
Initialize $\bar{\mathbf{x}}_0 \leftarrow \mathbf{x}_0$
for $t = 1$ **to** T **do**
  Receive $\mathbf{z}_t$ from $\mathcal{A}$
  Update $\mathbf{x}_t \leftarrow \mathbf{x}_{t-1} + \rho_t \mathbf{z}_t$, where $\rho_t \sim \mathrm{Exp}(1)$ independently
  Compute $\mathbf{g}_t \leftarrow \mathrm{StoGrad}(\mathbf{x}_t, r_t)$ with freshly sampled randomness r_t
  Send the linear loss $\ell_t^{[\beta_1]}(\mathbf{z}) := \langle \beta_1^{-t} \mathbf{g}_t, \mathbf{z} \rangle$ to $\mathcal{A}$
  Update $\bar{\mathbf{x}}_t \leftarrow \frac{\beta_1 - \beta_1^t}{1 - \beta_1^t} \bar{\mathbf{x}}_{t-1} + \frac{1-\beta_1}{1-\beta_1^t} \mathbf{x}_t$ {For output only}
end for
Return $\bar{\mathbf{x}}_\tau$ where $\tau \sim \mathrm{Unif}([T])$

The weighted loss in Algorithm 1 matches the discounted-regret convention used in the main text: multiplying the ordinary regret of $\mathcal{A}$ on losses $\ell_s^{[\beta_1]}$ by β_1^t gives $\sum_{s=1}^t \beta_1^{t-s} \langle \mathbf{g}_s, \mathbf{z}_s - \mathbf{u} \rangle = \mathrm{Regret}_t^{[\beta]}(\mathbf{u})$ with $\beta \equiv \beta_1$ and cost vectors $\mathbf{v}_s = \mathbf{g}_s$. Regarding the output $\bar{\mathbf{x}}_t$, $\bar{\mathbf{x}}_t = \mathbb{E}_{\mathbf{y}_t}[\mathbf{y}_t]$, where $\mathbf{y}_t$ is distributed over $\{\mathbf{x}_s\}_{s=1}^t$ with $\mathbb{P}(\mathbf{y}_t = \mathbf{x}_s) = \beta_1^{t-s} \frac{1-\beta_1}{1-\beta_1^t}$.

Definition I.2 ((λ, ε)-stationarity)**.** Suppose $F : \mathbb{R}^d \to \mathbb{R}$ is differentiable. For $\lambda > 0$, define
$$\|\nabla F(\mathbf{x})\|^{[\lambda]} := \inf_{\substack{p \in \mathcal{P}(\mathbb{R}^d) \\ \mathbb{E}_{\mathbf{y} \sim p}[\mathbf{y}] = \mathbf{x}}} \left\{ \|\mathbb{E}_{\mathbf{y} \sim p}[\nabla F(\mathbf{y})]\| + \lambda \, \mathbb{E}_{\mathbf{y} \sim p}[\|\mathbf{y} - \mathbf{x}\|^2] \right\}.$$
We say that $\mathbf{x}$ is a (λ, ε)-stationary point if $\|\nabla F(\mathbf{x})\|^{[\lambda]} \le \varepsilon$.

We now restate the Euclidean-norm stationarity consequence from Section 7.1 using the notation introduced above.

Theorem 7.2 (Stationarity)**.** *Under Assumption I.1, let $A := G + \sigma + \epsilon$. For any $r \in (0, 1)$, $\lambda > 0$, and $\varepsilon_{\mathrm{stat}} \in (0, A]$, there are choices of $\beta_1, \beta_2, D, \eta$ with $\beta_2 = 1 - r(1 - \beta_1)$ such that, for $T = \Omega_r\left(\Delta \lambda^{1/2} A^2 \varepsilon_{\mathrm{stat}}^{-7/2}\right)$, the model-EMA output $\bar{\mathbf{x}}_t$ satisfies $\mathbb{E}_{t \sim [T]} \mathbb{E}\left[\|\nabla F(\bar{\mathbf{x}}_t)\|^{[\lambda]}\right] \le O(\varepsilon_{\mathrm{stat}})$.*

Proof of Theorem 7.2. Let $\delta := 1 - \beta_1$, set $A := G + \sigma + \epsilon$, and $\bar{\mathbf{x}}_t = \mathbb{E}_{\mathbf{y}_t}[\mathbf{y}_t]$ denote the model EMA output in the discounted-to-nonconvex conversion. Choose
$$\delta = \Theta_r\left(\frac{\varepsilon_{\mathrm{stat}}^2}{A^2}\right), \qquad \beta_2 = 1 - r\delta, \qquad D = \Theta_r\left(\delta \lambda^{-1/2} \varepsilon_{\mathrm{stat}}^{1/2}\right),$$
and choose η as in the optimized proportional-coupling part of Theorem 7.1. The condition $\varepsilon_{\mathrm{stat}} \le A$ ensures that the chosen momentum gap $\delta = 1 - \beta_1$ can be taken in $(0, 1)$; it also keeps the logarithmic substitutions, such as $\log(A/\varepsilon_{\mathrm{stat}})$, in the stated high-accuracy regime. For this coupling, $\rho := \sqrt{\beta_2}/\beta_1$ satisfies $\log \rho = \Theta_r(\delta)$. Hence the thresholds in Theorem 7.1 obey $T_0 = O_r\left(\delta^{-1} \log \frac{1}{\delta}\right)$, which gives $T_0 = O_r(A^2 \varepsilon_{\mathrm{stat}}^{-2} \log(A/\varepsilon_{\mathrm{stat}}))$ after substituting the choice of δ.

By Definition I.2, since $\mathbf{y}_t$ is distributed over $\{\mathbf{x}_s\}_{s=1}^t$ with $\mathbb{P}(\mathbf{y}_t = \mathbf{x}_s) = \beta_1^{t-s} \frac{1-\beta_1}{1-\beta_1^t}$, we have
$$\mathbb{E}_{t \sim [T]} \mathbb{E}\left[\|\nabla F(\bar{\mathbf{x}}_t)\|^{[\lambda]}\right] \le \mathbb{E}_{t \sim [T]} \mathbb{E}\left[\|\mathbb{E}_{\mathbf{y}_t} \nabla F(\mathbf{y}_t)\| + \lambda \mathbb{E}_{\mathbf{y}_t} \|\mathbf{y}_t - \bar{\mathbf{x}}_t\|^2\right].$$

Localization term. By the variance/localization bound of Zhang & Cutkosky (2024, Lemma 3.2), equivalently in the form used by Ahn & Cutkosky (2024, Lemma 10),

$$\mathbb{E}\sum_{t=1}^{T}\mathbb{E}_{\mathbf{y}_t}\|\mathbf{y}_t-\bar{\mathbf{x}}_t\|^2\le\frac{12}{\delta^2}\mathbb{E}\sum_{t=1}^{T}\|\mathbf{z}_t\|^2.$$

Since the learner is clipped, $\|\mathbf{z}_t\|\le D$, so

$$\lambda\,\mathbb{E}_{t\sim[T]}\mathbb{E}_{\mathbf{y}_t}\|\mathbf{y}_t-\bar{\mathbf{x}}_t\|^2=O\!\left(\frac{\lambda D^2}{\delta^2}\right).$$

With $D=\Theta_r(\delta\lambda^{-1/2}\varepsilon_{\mathrm{stat}}^{1/2})$, this term is $O_r(\varepsilon_{\mathrm{stat}})$.

Averaged-gradient term. For each t, set

$$\mathbf{S}_t:=\sum_{s=1}^{t}\beta_1^{-s}\nabla F(\mathbf{x}_s),\qquad \mathbf{u}_t:=\begin{cases}-D\,\mathbf{S}_t/\|\mathbf{S}_t\|, & \mathbf{S}_t\neq\mathbf{0},\\ \mathbf{0}, & \mathbf{S}_t=\mathbf{0}.\end{cases}$$

The optimized proportional-coupling part of Theorem 7.1 gives the final pointwise regret bound

$$\mathbb{E}[\mathrm{Regret}_t^{[\beta]}(\mathbf{u}_t)]=O_r\!\left(\frac{DA}{\sqrt{\delta}}\right),\qquad t\ge T_0.$$

For the first T_0 rounds we use the following crude bound inferred directly from the regret definition. For any $t<T_0$,

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

Here the first inequality uses $\|\mathbf{z}_s\|\le D$ and $\|\mathbf{u}_t\|\le D$, while the second uses $\mathbb{E}\|\mathbf{g}_s\|\le G+\sigma\le A$. Thus

$$\mathbb{E}[\mathrm{Regret}_t^{[\beta]}(\mathbf{u}_t)]=O\!\left(\frac{DA}{\delta}\right),\qquad t<T_0,$$

for the initial segment. Therefore, for $T\ge T_0$,

$$\frac{1}{DT}\left(\beta_1\mathbb{E}[\mathrm{Regret}_T^{[\beta]}(\mathbf{u}_T)]+\delta\sum_{t=1}^{T}\mathbb{E}[\mathrm{Regret}_t^{[\beta]}(\mathbf{u}_t)]\right)=O_r\!\left(\frac{A}{T\sqrt{\delta}}+A\sqrt{\delta}+\frac{AT_0}{T}\right).$$

Lemma 7 of Ahn & Cutkosky (2024) then gives

$$\mathbb{E}_{t\sim[T]}\mathbb{E}[\|\mathbb{E}_{\mathbf{y}_t}\nabla F(\mathbf{y}_t)\|]=O_r\!\left(\frac{\Delta}{DT}+\frac{A}{\delta T}+\frac{A}{T\sqrt{\delta}}+\frac{AT_0}{T}+A\sqrt{\delta}\right).$$

Substituting the choices of δ and D, together with $T_0=O_r(A^2\varepsilon_{\mathrm{stat}}^{-2}\log(A/\varepsilon_{\mathrm{stat}}))$, gives

$$O_r\!\left(\frac{\Delta\lambda^{1/2}A^2}{\varepsilon_{\mathrm{stat}}^{5/2}T}+\frac{A^3}{\varepsilon_{\mathrm{stat}}^{2}T}\log\frac{A}{\varepsilon_{\mathrm{stat}}}+\frac{A^2}{\varepsilon_{\mathrm{stat}}T}+\varepsilon_{\mathrm{stat}}\right).$$

A sufficient detailed lower bound on T is

$$T\ge\Omega_r\!\left(\max\left\{\Delta\lambda^{1/2}A^2\varepsilon_{\mathrm{stat}}^{-7/2},A^3\varepsilon_{\mathrm{stat}}^{-3}\log\frac{A}{\varepsilon_{\mathrm{stat}}},A^2\varepsilon_{\mathrm{stat}}^{-2}\right\}\right).$$

Thus the target-accuracy term is dominated by the first displayed term, which is the $\varepsilon_{\mathrm{stat}}$-dominant condition stated in the theorem. Combining the averaged-gradient and localization bounds proves the claim. □

J From Vector to Coordinate-Wise Adam

The coordinate-wise assumptions and L_1 stationarity definition below are the standard coordinate-wise counterparts used in the vector-to-coordinate aggregation pipeline (Cutkosky et al., 2023; Ahn & Cutkosky, 2024); we restate them to make the conversion to coordinate-wise clipped Adam explicit.

Assumption J.1 (Coordinate-wise stochastic cost vectors)**.** For each $t \in [T]$ and coordinate $i \in [d]$, let $v_t[i]$ denote the i-th coordinate of the stochastic cost vector $\mathbf{v}_t$. There exist constants $G_i, \sigma_i \geq 0$ such that

$$|\mathbb{E}[v_t[i]]| \leq G_i, \qquad \mathbb{E}\big[(v_t[i] - \mathbb{E}[v_t[i]])^2\big] \leq \sigma_i^2.$$

Let $\boldsymbol{G} = (G_1, \ldots, G_d)$ and $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_d)$.

Assumption J.2 (Coordinate-wise function and stochastic-gradient oracle)**.** Assume the lower-boundedness and well-behavedness conditions in Assumption I.1. In addition, for each coordinate $i \in [d]$, there exist $G_i, \sigma_i \geq 0$ such that for all $\mathbf{x}$,

$$|\partial_i F(\mathbf{x})| \leq G_i, \qquad \mathbb{E}[g[i]] = \partial_i F(\mathbf{x}), \qquad \mathbb{E}[(g[i] - \partial_i F(\mathbf{x}))^2] \leq \sigma_i^2.$$

Let $\boldsymbol{G} = (G_1, \ldots, G_d)$ and $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_d)$.

Definition J.3 ((λ, ε)-L_1 stationarity)**.** For $\lambda > 0$, define

$$\|\nabla F(\mathbf{x})\|_1^{[\lambda]} := \inf_{\substack{p \in \mathcal{P}(\mathbb{R}^d) \\ \mathbb{E}_{\mathbf{y}\sim p}[\mathbf{y}] = \mathbf{x}}} \left\{ \|\mathbb{E}_{\mathbf{y}\sim p}[\nabla F(\mathbf{y})]\|_1 + \lambda\, \mathbb{E}_{\mathbf{y}\sim p}[\|\mathbf{y} - \mathbf{x}\|^2] \right\}.$$

We say that $\mathbf{x}$ is a (λ, ε)-L_1 stationary point if $\|\nabla F(\mathbf{x})\|_1^{[\lambda]} \leq \varepsilon$.

Corollary J.4 (Coordinate-wise clipped Adam: coordinate-wise regret and L_1 stationarity)**.** *Let $\delta := 1 - \beta_1$ and $A_1 := \|\boldsymbol{G} + \boldsymbol{\sigma} + \epsilon \mathbf{1}\|_1$. Under Assumption J.1, proportional coupling $\beta_2 = 1 - r\delta$, and the optimized step size from Theorem 7.1, the coordinate-wise clipped learner satisfies, for $T \geq \max(T_0, T_1)$ and any comparator sequence $\mathbf{u}_t \in [-D, D]^d$,*

$$\mathbb{E}[\mathrm{Regret}_T^{[\beta]}(\mathbf{u}_T)] \leq O_r\left(\frac{DA_1}{\sqrt{\delta}}\right), \qquad \sum_{t=1}^{T} \mathbb{E}[\mathrm{Regret}_t^{[\beta]}(\mathbf{u}_t)] \quad \leq O_r\left(\frac{TDA_1}{\sqrt{\delta}}\right).$$

Moreover, under Assumption J.2, for any $\lambda > 0$ and $\varepsilon_{\mathrm{stat}} \in (0, A_1]$, suitable choices of $\beta_1, \beta_2, D, \eta$ ensure that, for $T = \Omega_r\left(\Delta d^{1/2} \lambda^{1/2} A_1^2 \varepsilon_{\mathrm{stat}}^{-7/2}\right)$, we have $\mathbb{E}_{t\sim[T]} \mathbb{E}\left[\|\nabla F(\bar{\mathbf{x}}_t)\|_1^{[\lambda]}\right] \leq O(\varepsilon_{\mathrm{stat}})$.

Proof of Corollary 7.3. Let $\delta := 1 - \beta_1$ and $A_1 := \|\boldsymbol{G} + \boldsymbol{\sigma} + \epsilon \mathbf{1}\|_1 = \|\boldsymbol{G} + \boldsymbol{\sigma}\|_1 + d\epsilon$. For the regret statement, fix a coordinate i. The coordinate-wise clipped bias-corrected update is the one-dimensional learner

$$z_t[i] = \mathrm{clip}_D\left(-\alpha_t[i] \sum_{s=1}^{t-1} \beta_1^{-s} v_s[i]\right),$$

where $\mathrm{clip}_D(a) = \mathrm{sign}(a) \min(D, |a|)$ with $\mathrm{clip}_D(0) = 0$, and

$$\alpha_t[i] := \eta\, \frac{(1 - \beta_1)\beta_1^{t-1}}{(1 - \beta_1^{t-1})(\sqrt{b_t[i]} + \epsilon)}, \qquad b_t[i] := \frac{1 - \beta_2}{1 - \beta_2^{t-1}} \sum_{s=1}^{t-1} \beta_2^{t-1-s} v_s[i]^2.$$

With $\alpha_1[i] := \alpha_2[i]$, this is exactly the one-dimensional version of Theorem 7.1, with scalar cost sequence $v_t[i]$, radius D, and constants G_i, σ_i. Thus the optimized proportional-coupling form of Theorem 7.1 gives

$$\mathbb{E}[\mathrm{Regret}_{t,i}^{[\beta]}(u_t[i])] = O_r\left(\frac{D(G_i + \sigma_i + \epsilon)}{\sqrt{\delta}}\right)$$

pointwise for $t \geq T_0$, and the corresponding cumulative bound with an additional factor of T for $T \geq \max(T_0, T_1)$. The coordinate-wise regret decomposes as

$$\mathrm{Regret}_t^{[\beta]}(\mathbf{u}_t) = \sum_{i=1}^{d} \mathrm{Regret}_{t,i}^{[\beta]}(u_t[i]),$$

because the linear loss and the update are separable across coordinates. Summing the scalar pointwise and cumulative bounds over $i = 1, \dots, d$ proves the displayed coordinate-wise regret bounds because $\sum_{i=1}^{d}(G_i + \sigma_i + \epsilon) = A_1$.

For the stationarity statement, apply the proof of Theorem 7.2 verbatim after the following replacements: use the L_1 DTNC conversion of Ahn & Cutkosky (2024, Lemma 17), replace A by A_1, use the coordinate-wise comparator from that lemma, and replace the localization bound $\|\mathbf{z}_t\|^2 \le D^2$ by $\|\mathbf{z}_t\|^2 \le dD^2$. The regret input is exactly the coordinate-wise bound proved above, together with the same crude pre-T_0 estimate, now with A_1 in place of A. Thus the argument of Theorem 7.2 gives

$$\mathbb{E}_{t\sim[T]}\mathbb{E}\left[\|\nabla F(\bar{\mathbf{x}}_t)\|_1^{[\lambda]}\right] \le O_r\left(\frac{\Delta}{DT} + \frac{A_1}{\delta T} + \frac{A_1}{T\sqrt{\delta}} + \frac{A_1 T_0}{T} + A_1\sqrt{\delta} + \frac{\lambda d D^2}{\delta^2}\right).$$

Choose

$$\delta = \Theta_r\left(\frac{\varepsilon_{\mathrm{stat}}^2}{A_1^2}\right), \qquad \beta_2 = 1 - r\delta, \qquad D = \Theta_r\left(\delta d^{-1/2}\lambda^{-1/2}\varepsilon_{\mathrm{stat}}^{1/2}\right),$$

and choose η as in the optimized proportional-coupling part of Theorem 7.1. As in Theorem 7.2, this gives $O(\varepsilon_{\mathrm{stat}})$ whenever

$$T \ge \Omega_r\left(\max\left\{\Delta d^{1/2}\lambda^{1/2}A_1^2\varepsilon_{\mathrm{stat}}^{-7/2}, A_1^3\varepsilon_{\mathrm{stat}}^{-3}\log\frac{A_1}{\varepsilon_{\mathrm{stat}}}, A_1^2\varepsilon_{\mathrm{stat}}^{-2}\right\}\right),$$

where $\varepsilon_{\mathrm{stat}} \le A_1$ ensures $\delta \in (0, 1)$. The first term is the $\varepsilon_{\mathrm{stat}}$-dominant condition stated in the corollary. □